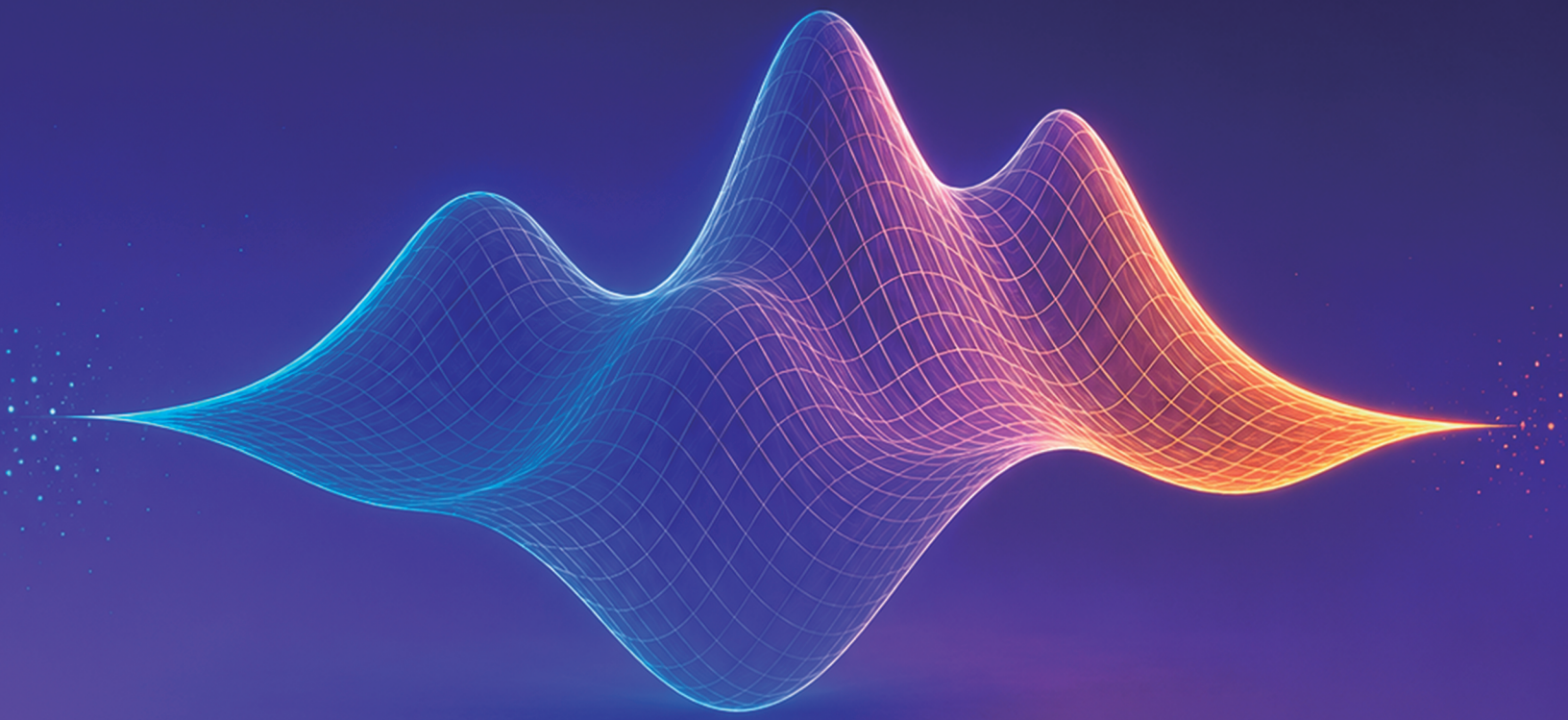
28

MATHEMATICAL MODELLING IN ENGINEERING & HUMAN BEHAVIOUR

CONFERENCE · MME&HB 2026



BOOK OF ABSTRACTS

im^2 Instituto Universitario de Matemática Multidisciplinar

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

II

Modelling for Engineering & Human Behaviour 2026

València, July 12th-15th, 2026

This book includes the extended abstracts of papers presented at XXVIII Edition of the Mathematical Modelling Conference Series at the Institute for Multidisciplinary Mathematics *Mathematical Modelling in Engineering & Human Behaviour 2026*.

ISBN: 978-84-09-89823-7

Edited by:

Clara Burgos Simón, Begoña Cantó Colomina, Alicia Cordero Barbero, Juan Carlos Cortés López, Elena López-Navarro, and María Dolores Roselló Ferragud.
I.U. de Matemàtica Multidisciplinar,
Universitat Politècnica de València.

Report any problems with this document to imm@imm.upv.es.
Version of July 27, 2026.

im^2
Instituto Universitario
de Matemática Multidisciplinar

UPV

Preface

July 2026

This book includes the works presented at the XXVIII Edition of the Mathematical Modelling Conference Series at the Institute for Multidisciplinary Mathematics, *the Mathematical Modelling in Engineering & Human Behaviour 2026 Conference*. We thank the ETS d'Enginyeria Informàtica of the Universitat Politècnica de València for hosting the conference in its comfortable building and generously providing access to its facilities from July 12 to July 15, 2026.

The objective of the Mathematical Modelling in Engineering and Human Behaviour conference is to develop an interdisciplinary forum in areas like Medicine, Sociology, Business and Engineering, where the latest mathematical techniques can be discussed in a common and understandable language with experts in cross-disciplinary areas. The conference is aimed at gathering researchers who need mathematics for the formulation and analysis of models.

This year, the following special sessions were organized:

1. Recent advances in Iterative Processes for Solving Nonlinear Problems.
   Organizers: Neus Garrido and Juan Ramón Torregrosa.
2. Mathematical Methods for Engineering Problems.
   Organizers: Damián Ginestar and Antoni Vidal.
3. Infectious Diseases and Mathematical Modeling in Public Health.
   Organizer: Clara Burgos and Almudena Sánchez.
4. Linear algebra, matrix analysis and applications.
   Organizer: José Mas and Alicia Roca.
5. Maths & physics.
   Organizer: Màrius Fullana.
6. Uncertainty Quantification and Modelling.
   Organizer: Dolors Roselló and Vicent Romero.
7. Mathematical models in population dynamics
   Organizers: Luis María Abia and Óscar Angulo.
8. Student's Projects
   A session for students of the double degrees of mathematics and the students of Data Science.

Organization

Mathematical Modelling in Engineering & Human Behaviour (MME&HB 2026) is organized by the Institute of Multidisciplinary Mathematics (IMM) of Universitat Politècnica de València.

Organizing Committee

The organizing committee of the MME&HB 2026 is composed by:

Begoña Cantó Colomina (Chair)	Universitat Politècnica de València
Clara Burgos Simón (Co-Chair)	Universitat Politècnica de València
Francisco Chicharro López	Universitat Politècnica de València
Juan Carlos Cortés López	Universitat Politècnica de València
Màrius Josep Fullana i Alfonso	Universitat Politècnica de València
Neus Garrido Saez	Universitat Politècnica de València
Elena López Navarro	Universitat Politècnica de València
Ana Navarro Quiles	Universitat Politècnica de València
Julia Real Herráiz	Universitat Politècnica de València
Alicia Roca Martínez	Universitat Politècnica de València
Cristina Santamaría Navarro	Universitat Politècnica de València
Antoni Vidal Ferràndiz	Universitat Politècnica de València

Sciencific Committee

The scientific committee of the MME&HB 2026 is composed by:

Alicia Cordero Barbero (Chair)	Universitat Politècnica de València
Dolors Roselló Ferragud (Co-Chair)	Universitat Politècnica de València
Tomás Caraballo Garrido	Universidad de Sevilla, Spain
Yun Sung Choi	National Institute for Mathematical Sciences, Korea
Regino Criado Herrero	Universidad Rey Juan Carlos, Spain
Hugo de la Cruz Cancino	FGV/EMAp, Brazil
Silvia Jerez Galiano	Centro de Investigación en Matemáticas, CIMAT, Mexico
Roberta Lima	Pontifícia Universidade Católica do Rio de Janeiro, Brazil
Carla M. A. Pinto	Polytechnic University of Porto, Portugal
Rubens Sampaio	Pontifícia Universidade Católica do Rio de Janeiro, Brazil
Eduardo Souza de Cursi	Institut National des Sciences Appliquées Rouen Normandie, France
Juan Ramón Torregrosa Sánchez	Universitat Politècnica de València, Spain
Fiza Zafar	Bahauddin Zakariya Univesity, Pakistan

Table of Contents

I Plenary Speakers

Taming Randomness: Robust Simulation of Stochastic Differential Systems
Hugo de la Cruz 3

Low-rank approximation of polynomial matrices
Andrii Dmytryshyn 4

From Solitons to Black Holes: New tools to study Black Hole resonances and scattering processes
Carlos F. Sopuerta 5

Mathematics in the Age of AI
Fiza Zafar 6

II General session

A contextual influence analysis incorporating expert weight variability for educational course design
S. Carpitella, B.M. Brentan, and J. Izquierdo 9

Optimal portfolio insurance: a new look
Alejandro Balbás, Beatriz Balbás and Raquel Balbás 17

Critical aspects for road maintenance prioritization
C.L. Parlapiano, S. Carpitella, A. Certa and J. Izquierdo 21

Out-uniform MAX, MIN–Boolean networks. Applications to social networks
Juan A. Aledo, Jose P. Llano, Cristina Santamaría and Jose C. Valverde . 28

Tackling the Optimal Bucket Order Problem through Integer Linear Programming
Juan A. Aledo, Concepción Domínguez, Juan de Dios Jaime-Alcántara and Mercedes Landete 32

High order non standard numerical schemes for a second order delay differential equation
Jean Marc Serrano, Francisco Rodríguez and María Ángeles Castro 35

Understanding Corporate Insolvency Across Regions and Sectors: A Macro-Financial and Machine Learning Perspective from Spain
Yu Zhang 38

Validated Cross-Ventilation Modeling of an occupied Indoor Environment Reveals Uneven Air Renewal After Window Opening

O.A. Shatarat, A. D'Ausilio, G. Aiello, S. La Grutta, V. Lazzara and V.Malizia 48

III Recent advances in Iterative Processes for Solving Nonlinear Problems

Fractal methods: Dynamics of the fractal Newton method
Jessica Paredes-Morales, Alicia Cordero and Juan R. Torregrosa 59

Dynamics of derivative-free methods for solving nonlinear equations with multiple roots
Jorge Rico, Alicia Cordero and Juan R. Torregrosa 69

Approximation of inverses and generalized inverses of complex matrices
Alicia Cordero, Eva G. Villalba, Juan R. Torregrosa and Paula Triguero-Navarro 76

Analysis and application of a parametric family of seventh-order iterative methods to the solution of a traffic management problem
Eva G. Villalba, Jesús Gallinal, Jorge D. Ávila, and Xènia Nicolau-Casademont 81

Efficiency and Stability of New Quasi-Newton Hybrid Algorithms with Higher-Order Singh-Type Correctors for Unconstrained Optimization
Javier García Maimó, Miguel A. Leonardo Sepúlveda, Antmel Rodríguez Cabral and Natanael Ureña Castillo 88

A Weighted Generalization of Singh's Method for Nonlinear Systems with Application to a Stationary Burgers Equation
Miguel A. Leonardo Sepúlveda, Javier G. Maimó, Antmel Rodríguez Cabral, and Natanael Ureña Castillo 94

Generalized Traub family for nonlinear systems: Fourth-order optimal method and dynamical analysis
Miguel A. Leonardo Sepúlveda, Alicia Cordero, Juan R. Torregrosa, Antmel Rodríguez Cabral, y María P. Vassileva 100

A Highly Efficient 2m-th Order Recursive Scheme for Nonlinear Systems under Increasing Uncertainty and Nonlinearity
Alicia Cordero, Renso V. Rojas-Hiciano and Juan R. Torregrosa 109

Parallel Methods for Multiple Roots via Predictor–Corrector Compositions
Francisco I. Chicharro, Neus Garrido and Julissa H. Jerezano 117

Halley-based method for simultaneous-root finding in nonlinear equations
Jesús Gallinal, Eva G. Villalba and Paula Triguero-Navarro 122

A Cubic-Order Iterative Scheme for Approximating the Inverse of a Non-Singular Matrix
Alicia Cordero, Bushra Rani and Juan Ramón Torregrosa 127

Extending a Class of Third-Order Iterative Methods for Solving Volterra Integral Equations
Rajinder Singh, Sukhjit Singh, Eulalia Martínez, M.A. Hernández 131

High-Order Multipoint Iterative Methods for Nonlinear Systems with Scalar Weight Functions
Maria Emilia Maldonado Machuca, Alicia Cordero, Juan Ramón Torregrosa and Paula Triguero-Navarro 136

Exploring Dynamical Planes in Nonlinear Equations and Systems: Derivative-Based versus Derivative-Free Iterative Schemes
Francisco I. Chicharro and Neus Garrido 143

IV Mathematical Methods for Engineering Problems

D-splitting. A new paradigm for low-storage RK methods
Sergio Blanes, Abutalib Elsaid Mohammed and Alejandro Escorihuela-Tomàs 153

Quasi-normal modes for exural waves on demand
Marta Gómez Pérez, Marc Martí-Sabaté and Vicente Romero-García 155

Mathematical Modelling of Acoustic Levitation Using Perturbative and k-Space Pseudospectral Methods. Demonstrator and Proof of Concept
Diego Carrión Perpiñá, and Vicente Romero-García 159

Multi-objective optimization of metamaterial sandwich panels
Vinícius F. Dal Poggetto, Mário Lázaro, Luis M. García-Raffia, and Vicent Romero-García 166

Small-Sample Prediction and Inverse Design of Three dimensional Topological Phononic Crystals
Xue-Qian Zhang, Vicent Romero-García, Tian-Xue Ma and Yue-Sheng Wang 173

Optimal Transport based models for Generative Design and Response Inference
Sergio Torregrosa, David Munoz, Hector Navarro, Charbel Farhat and Francisco Chinesta 177

Passive Vibration Control of a Clamped-Free Euler–Bernoulli Beam Using a Tuned Mass-Spring Absorber
Ana Pinto, Alain de Souza and Roberta Lima 179

Heuristic Methods for Almond Percentage Detection via Least Squares
Alejandro Arques, Ferran Verdú and Mónica Cortés 187

Human Action Recognition using Depth and Skeleton Data through V-JEPA2
J. Alberto Conejero and Lluïsa Gual-Vayà 196

A New Methodological Proposal: Application of Hybrid Digital Twins to NOx Prediction using Neural Networks

Marcos Llamazares, Juan Carlos Cortés, Rafael-J. Villanueva and Carlos Guardiola 201

Analisys of Gabor Filters for Defect Detection in Raw Fabric Images
T. Albero-Albero, B. Cantó, N. Garrido and R. Pérez 208

Development of a data-driven PINN to solve the neutron diffusion equation in nuclear reactors
E. Navarro-Gamón, A. Vidal-Ferràndiz, D. Ginestar, A. Carreño and G. Verdú 215

Proper orthogonal decomposition for neutron noise simulation and its application to small modular reactors
Antoni Vidal-Ferràndiz, Amanda Carreño, Damián Ginestar and Gumersindo Verdú 223

A Neural Network Approach for Inverse Kinematics in Arm Pose Reconstruction
Gloria Campos 231

Machine learning model for strength of high-performance concrete forecast by using the hybrid WOA optimized MARS technique
P.J. García–Nieto, E. García–Gonzalo, J.P. Paredes–Sánchez and L.A. Menéndez-García 239

Mathematical model for the thermoelectric response of a bipolar radiofrequency-based hemostatic device
Antonio Romero, Gabriel Ignacio Maureira and Macarena Trujillo 246

Aerodynamic Shape Optimization using Graph Neural Networks and Neural ODEs
Guillem Pérez, Victoria Muerza, Fernando Pascual and Emilio Larrodé 250

Explaining Couette Flows using AI
Clara Trillo-Yagüe, Andrés Cremades, Ricardo Vinuesa and Sergio Hoyas . 253

Entropy and Explainability: Exploring Information and Causality in Turbulent Flows
Belén Reverte Badillo, Andrés Cremades Botella and Sergio Hoyas Calvo .. 254

Physics-Based Mathematical Decomposition of Electric Vehicle Energy Consumption: Model Formulation, Python Implementation and Real-World Validation Against OBD Measurements
Paula Reguera Carrera, Emilio Larrodé Pellicer and Victoria Muerza Marín 256

V Infectious Diseases and Mathematical Modeling in Public Health

Predictive analysis of hospital mortality in patients admitted from the emergency department
R.de Ramón Frías, I. Barrachina Martínez and S. González-de Julián 263

A data-driven approach to cognitive performance profiling
Pedro Fernández de Córdoba, Álvaro González Cortés and Esperanza Navarro Pardo 271

Solving Inverse Problems in X-ray CT with Implicit Neural Representations
Hyoung Suk Park and Kiwan Jeon 273

Post-Pandemic Social Contact Patters in South Korea:A Nationwide Survey Dataset and Analysis
Min-Kyung Chae, Woo-Sik Son, Dong-Uk Hwang, Kyeongah Nah, Minsoo Kim, Jong-Hoon Kim and Jonggul Lee 275

Multi-scale RSV dynamics: SLIRS with ILI-informed calibration and time-series correction
Saebom Jeon, Almudena Sánchez-Sánchez and Rafael Jacinto Villanueva Micó 279

Predicting Patient and Graft Survival in Kidney Transplantation Using Machine Learning and Cox Regression
Laura Shalabayeva, Marc Romeu Ferras, Javier Díaz-Carnicero, María Ramos Cebrián, Alberto Budía Alba and David Vivas-Consuelo 284

Describing the Transmission Dynamics of Hepatitis A in Children Under One Year of Age: An Age-Structured Model with Uncertainty
Clara Burgos Simón, Cristina Valcárcel Nazco, Carmen Guirado Fuentes and Rafael Jacinto Villanueva Micó 292

Neural ODE-Based Longitudinal Forecasting of Pulmonary Disease Progression in Chest Radiographs
Gonzalo Atienza, Rocío Del Amor and Fernando López 297

Mathematical Modelling of the Atypical Epidemic Behaviour of Seasonal Influenza in the Valencian Community During the 2015–2016 Season
Rocío Bagán Navarro, Cristina-Luisovna Pérez and Rafael J. Villanueva Micó 305

Commol: A Rust-Powered Python Library for Compartment Dynamic Models
David Martínez-Rodríguez, Carlos Andreu-Vilarroig and Rafael Jacinto Villanueva Micó 307

Inverse Problems in Dental CBCT for Digital Dentistry
Hyoung Suk Park and Kiwan Jeon 313

Mathematical Modelling of Cognitive Load During Mental Arithmetic Tasks Using EEG Signals
Israr Ul Haq and Luis Acedo 315

A proposal to select the optimal immunization for Respiratory Syncytial Virus using a calibrated Agent-Based Model
J.Villanueva-Oller, R. J. Villanueva, J. C. Cortés, C. Pérez-Diukina 324

A mathematical model with bistability dynamics explaining malaria prevalence
Gilberto González-Parra, Sharmin Sultana, Luis Fernando Chaves and Daniela Lezcano 325

Impact of temperature on the life cycle and vectorial capacity of Aedes aegypti, Puerto Rico, 2022-2025
Juan Ángel Aledo, Juan C. Orengo, Mario Alberto Espinoza and Rafael-J. Villanueva 331

Memory Effects or Modeling Errors? (In)Consistency of Fractional-Order Systems in Epidemiology
Carla M. A. Pinto, Julia Calatayud and Marc Jornet 334

A framework for time-varying risk assessment of P. vivax transmission in Republic of Korea
Woo-Sik Son, Ah-Young Lim, Dong-Uk Hwang and Kyeongah Nah 335

Numerical validation of the quasi-stationary distribution in a stochastic SID coinfection Markov model
Carlos Andreu-Vilarroig, Sten Madec and Erida Gjini 337

VI Linear Algebra, Matrix Analysis and Applications

Optimal order-24 matrix exponential approximation with applications to generative AI flows
Jorge Sastre, José Miguel Alonso, Javier Ibáñez, Julio Andrés, Roger B. Dannenberg, Daniel Faronbi and Nuria Lloret 343

A new Taylor method for the matrix sign function
Emilio Defez, José Miguel Alonso, Javier Ibáñez, Jorge Sastre 356

The PageRank Vector with Memory: From Discrete Iterations to Continuous Dynamical Systems (Part I)
J. Flores, E. Primo, D. Rodríguez and M. Romance 364

The PageRank Vector with Memory: From Discrete Iterations to Continuous Dynamical Systems (Part II)
J. Flores, E. Primo, D. Rodríguez and M. Romance 365

Row completion of polynomial matrices with partial prescription of the singular structure
Agurtzane Amparan Itziar Baragaña, Silvia Marcaida and Alicia Roca 366

High relative accurate algorithms for Riordan arrays associated with Schr¨oder matrices
Jorge Ballarín, Jorge Delgado and Juan Manuel Peña 370

Preserving observability and structural decodability with output feedback transformations on ISO representations of convolutional codes
Noemí DeCastro-García, Lucía Mallo-Fernández and Miguel V. Carriegos . 372

VII Maths & Physics

Entangled photons detection
Guillem Gómez i Blanch, Màrius Josep Fullana i Alfonso and Roberto Dale Valdivia 377

More about harmonic oscillators on the lattice
Emilio Defez Candel, Michael M. Tung and Màrius Josep Fullana i Alfonso 383

Quantum Trajectories of Logistic Dynamics
Joan C. Micó 390

Data-Driven Scatter Correction in PET via Weighted Maximum Likelihood Expectation Maximization
Joan Prats-Climent, Alejandro Sanz-Sánchez and María José Rodríguez-Álvarez 398

Parameter Estimation of Sub-Solar-Mass Objects with Mini Extreme Mass-Ratio Inspirals
Qiming Song, Huaike Guo and Ju Chen 405

Unit Killing Initial Data
Salvador Mengual Sendra and Igor Khavkine 413

On the Einstein-Proca Equations in Spherical Symmetry
Alicia Herrero, Roberto Dale and Jesús Martínez 420

A Preliminary Study of CMB Polarisation Amplitude Statistics and Map Downgrading Techniques
R. Dale and J. M. Gandía 428

A geometric study on methods of temporary reparametrization in a two-body problem
José Antonio López Ortí, María José Martínez Usó, Francisco José Marco Castillo 436

VIII Uncertainty Quantification and Modelling

Nonuniform Polynomial Behavior for Random Discrete-Time Systems
Oana Albescu, Larisa Elena Biris, Traian Ceausu and Nicolae Marian Seimeanu 443

Probabilistic Analysis of Linear Delay Difference Equations with Uncertain Parameters
A. Navarro-Quiles, J.-V. Romero and M.-D. Roselló 445

A Probabilistic Framework for Heat Transfer Models with Implicit Solutions
E. Morellà, R.-J. Villanueva and J.-C. Cortés 450

Impact of Light-Cycle duration on Algal Biomass Production: A Random PSF Model
Juan-Carlos Cortés, Ana Navarro-Quiles, Stepan Papáček, Cristina Luisovna Pérez 456

Characterization of the Unstationnary Stochastic Response of Mechanical Systems Using a Hilbertian–Bayesian Approach
Carol Santos Almonte, Oscar Sanchez Jimenez, Eduardo Souza de Cursi and Emmanuel Pagnacco 462

Probabilistic Analysis of an RC Circuit for a Stochastic Delta-Correlated Process
Juan-Carlos Cortés, José-Vicente Romero, María-Dolores Roselló and Joaquin F. Valencia Sullca 463

A new type of multi-Sugeno-inspired aggregation functions
José Carlos R. Alcantud 468

On the use of the delta function for the computation of probability density functions in the random Burgers-Riemann problem
Marc Jornet 474

Mathematical modeling of crime data with stochastic differential equations
Julia Calatayud 480

Epistemic and Aleatoric Uncertainty Quantification in Bayesian Convolutional Neural Networks: Application to Food Image Classification
Raúl López-Romero, Joaquín Martínez-Minaya, Cristóbal Giménez-Devis, Fernando López-Mir, Adrián Colomer 483

Uncertainty Quantification in Inverse Identification Problem via Polynomial Chaos Expansions
Roberta Lima, Rubens Sampaio, Eduardo S. de Cursi and Emmanuel Pagnacco 490

Bayesian Inference of Contagion Parameters in Branching Processes under Additive Noise
João Pedro Freitas, Roberta Lima, Rubens Sampaio and Eduardo S. de Cursi 496

IX Mathematical models in population dynamics

On parameter estimation of differential equations with incomplete data
Benito Chen-Charpentier 505

Numerical integration of a nonlinear structured population model with two physiological structures
Luis M. Abia, Óscar Angulo, Juan C. López-Marcos and Miguel Á. López-Marcos 513

A general framework for the convergence analysis of Floquet multipliers of delay equations
Alessia Andò, Giusy Bosco, Dimitri Breda and Davide Liessi 515

On Runge-Kutta Methods for Stochastic Delay Differential Equations
Alessia Ando, Dimitri Breda, Davide Liessi, Stefano Maset and Faraz William 517

Population-Based Statistical Modeling and Lipschitz Reinforcement Learning for Response Times in Video Games
Ana Coronado Ferrer 519

A Hybrid Computational Framework for a Fisher-KPP Problem with Nonlocal Free Boundary
Vera Egorova, M.-C. Casabán and Rafael Company 526

On the existence and positivity of almost periodic solutions of a plankton model with seasonality
Abraham J. Arenas, Gilberto González-Parra and Myladis R. Cogollo 533

A Generic Model of CD8 T Cell Dynamics: Mathematical and Sensitivity Analysis
Oscar Angulo and Fabien Crauste 538

X Student's Projects

Parallel solution of the Schr¨odinger eigenvalue problem via real rational schemes
Sergio Blanes and José L. Ruiz-Benito 543

Numerical Methods for Solving the Black–Scholes Equation
Carlos Sevilla Cabedo, Manuel Sellés Armendia and Roger Rosado Santos . 546

A Reduced-Order Digital-Twin Framework for MOST-Assisted Thermal Buffering in Biological CubeSat Payloads
E. Vega-Fleitas, S. Moll-López, J. A. Moraño-Fernández, A. Herrero-Debón and O. Portero-Sáez 553

Part I

Plenary Speakers

Taming Randomness: Robust Simulation of Stochastic Differential Systems

Hugo de la Cruz[♭],[1]

(♭) FGV EMAp, Brazil.

1 Abstract

Stochastic Differential Systems are indispensable tools for modeling uncertainty across diverse fields, from quantitative finance and molecular dynamics to complex engineering. However, achieving robust simulations presents unique challenges, as traditional numerical methods often lack the stability required to handle the non-differentiable nature of the driven noise. These difficulties are amplified in systems where high dimensionality and stiffness lead to significant error accumulation and computational instability. This talk focuses on the development and application of high-fidelity numerical integrators for complex Stochastic Differential Systems. We will explore a range of techniques, from foundational methods to higher-order integrators designed for increased accuracy. Special emphasis will be placed on numerical stability and the structural properties of integrators, discussing how to select the optimal approach based on the system's requirements for strong versus weak convergence. The goal is to provide a reliable roadmap for "taming" randomness, translating complex stochastic systems into robust, ready-to-use numerical simulations.

[1] hugo.delacruz@fgv.br

Low-rank approximation of polynomial matrices

Sweta Das[♭], Andrii Dmytryshyn[♮],[1] Froilán M. Dopico[♯],
Mårten Gulliksson[♭] and Rakel Hellberg[♮]

(♭) School of Science and Technology, Örebro University, 70182 Örebro, Sweden

(♮) Department of Mathematical Sciences,
Chalmers University of Technology and University of Gothenburg, 41296 Gothenburg, Sweden.

(♯) Universidad Carlos III de Madrid, ROR: https://ror.org/03ths8210,
Departamento de Matemáticas, Avenida de la Universidad 30, 28911, Leganés, Spain.

Abstract

Finding low-rank approximations is a classical nearness problem in numerical linear algebra, with numerous applications across many areas of science and engineering. For constant matrices, this problem is well understood. However, the situation becomes considerably more challenging when the matrices depend on parameters.

In this talk, we study low-rank approximation of polynomial matrices $A(\lambda) \in \mathbb{C}[\lambda]^{m\times n}$. We present a new method for approximating a given polynomial matrix by another polynomial matrix of prescribed rank and degree bounded by d, where $d \geq 1$; see [1], [2]. The framework includes the well-known problem of computing the distance to singularity for a regular polynomial matrix.

The proposed method combines alternating-direction optimization techniques with recent advances in the theory of generic eigenstructures [3], [4] and minimal-rank factorizations [5] of polynomial matrices. We present numerical experiments illustrating the performance of the method and, whenever possible, comparisons with existing approaches.

References

[1] S. Das, A. Dmytryshyn, F. M. Dopico, and M. Gulliksson, "Generic factorizations and alternating direction method for fixed rank approximation of matrix polynomials," *In prep*, 2026.

[2] A. Dmytryshyn, F. M. Dopico, and R. Hellberg, "Distance to nearest skew-symmetric matrix polynomials of bounded rank," *arXiv preprint arXiv:2601.16676*, 2026.

[3] A. Dmytryshyn and F. M. Dopico, "Generic complete eigenstructures for sets of matrix polynomials with bounded rank and degree," *Linear Algebra Appl.*, vol. 535, pp. 213–230, 2017.

[4] F. De Terán, A. Dmytryshyn, and F. M. Dopico, "Even grade generic skew-symmetric matrix polynomials with bounded rank," *Linear Algebra Appl.*, vol. 702, pp. 218–239, 2024.

[5] A. Dmytryshyn, F. M. Dopico, and P. Van Dooren, "Minimal rank factorizations of polynomial matrices," *Linear Algebra Appl.*, vol. 721, pp. 736–774, 2025.

[1] andrii@chalmers.se

From Solitons to Black Holes: New tools to study Black Hole resonances and scattering processes

Carlos F. Sopuerta[♭,1]

(♭) Department of Nuclear, Gravitational Wave and Multi-Messenger Astrophysics, Institute of Space Sciences (ICE-CSIC and IEEC).

1 Abstract

The Korteweg-de Vries equation (KdV equation) was originally introduced to describe the propagation of shallow-water waves. In this context, it governs the dynamics of perturbations at the fluid surface, which can be directly related to the surface gravity. It was later discovered that this equation is completely integrable and also arises in many other physical settings, including plasma physics, nonlinear optics, and condensed-matter systems.

At the opposite end of the physical spectrum we have black holes, probably the most extreme prediction of the General Theory of Relativity proposed by Einstein more than a century ago. One of the main developments in black-hole physics is the study of their gravitational perturbations. Remarkably, the dynamics of these perturbations can be encoded in a set of master variables satisfying Schrödinger-type wave equations, whose solutions contain the complete physical information about the system.

In this talk, I will show that, despite describing fully generic gravitational dynamics, these master equations admit an infinite set of hidden symmetries. In particular, a distinguished subset of these symmetries is intimately related to the flow generated by the KdV equation and its associated integrable hierarchy. Exploiting this structure, one can derive balance laws that allow the computation of black-hole scattering probabilities directly from the underlying KdV symmetries and the associated conserved quantities.

I will conclude by discussing the mathematical and physical implications of this unexpected and surprising connection between integrable systems (in particular the KdV hierarchy) and black-hole physics, as well as several directions for future research.

[1] carlos.f.sopuerta@csic.es

Mathematics in the Age of AI

Fiza Zafar[b]

(b) Centre for Advanced Studies in Pure and Applied Mathematics, Bahauddin Zakariya University, Multan, Pakistan

Recent advances in artificial intelligence (AI) are opening new avenues for solving mathematical problems. AI-driven techniques, particularly hybrid neural network–based mathematical models, enable efficient exploration of complex, high-dimensional solution spaces and provide powerful tools for approximation, prediction, and optimization. Prompt engineering techniques are further enhancing mathematical problem solving and reasoning with large language models (LLMs), allowing AI systems to generate structured solutions, symbolic manipulations, and stepwise logical arguments. In parallel, computer-assisted proof systems such as Lean are improving the rigor and verification of mathematical results through formal proof development. Emerging agentic AI tools, including OptiMUS, are also demonstrating the potential of autonomous AI-assisted workflows for mathematical problems. While AI significantly accelerates experimentation, conjecture generation, and solution discovery, rigorous validation, abstraction, and theoretical understanding remain essential human contributions. This talk highlights the transformative role of hybrid AI methodologies, LLM-based reasoning, formal verification, and agentic AI systems in modern mathematical research, and discusses future directions for integrating intelligent tools into nonlinear analysis, optimization, and mathematical discovery.

References

[1] J. Avigad and P. Massot, *Mathematics in Lean*. 2025.

[2] F. Da Ros, M. Soprano, L.D. Gaspero, and K. Roitero, "Large Language Models for Combinatorial Optimization: A Systematic Review", *ACM Computing Surveys,* Vol. 58 no. 11, pp. 1-53, 2026.

[3] D. Shen, G. Wu, and H. I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, Vol. 19, pp. 221–248, 2017.

[4] A.A. Teshnizi, W. Gao, H. Brunborg, S. Talaei, and M. Udell, "OptiMUS-0.3: Using large language models to model and solve optimization problems at scale", 2024, arXiv:2407.19633. Retrieved from https://arxiv.org/abs/2407.19633.

[5] F. Zafar, A. Cordero, S. Mujtaba, and J.R. Torregrosa, "A hybrid Steffensen–Genetic algorithm for finding multi-roots of nonlinear equations and applications to biomedical engineering," *Algorithms*, Vol. 18 no. 9, Art. 582, 2025.

[1] fizazafar@bzu.edu.pk

Part II

General session

A contextual influence analysis incorporating expert weight variability for educational course design

S. Carpitella [♭], B.M. Brentan [♮], and J. Izquierdo[♢],[1]

(♭) California State University, Northridge, 18111 Nordhoff St, Northridge, California, US.
(♮) Federal University of Minas Gerais, Av. Presidente Antônio Carlos, Belo Horizonte, Brazil.
(♢) Universitat Politècnica de València, Cno. de Vera s/n, Valencia, Spain.

Abstract

Educational course improvement is a multidimensional process characterized by strong interdependencies among strategic configurations. This study applies the Decision Making Trial and Evaluation Laboratory (DEMATEL) method to examine mutual influence among four key configurations of course improvement, including digital integration, pedagogical innovation, assessment and feedback strategy, and industry alignment. Judgments were collected from three domain experts representing complementary perspectives in educational pedagogy, educational technology, and industry engagement. This study introduces a contextual expert weighting approach to investigate how variations in expert influence affect the resulting causal structure. The findings highlight the importance of explicitly addressing epistemic dominance in multi expert decision models and suggest that strategic educational planning should account for contextual differences in expertise.

1 Introduction and Literature Review

Universities operate in an environment characterized by rapid technological advancement, evolving labor market demands, and increasing societal expectations [1]. Nowhere is this transformation more visible than in engineering education. Emerging technologies such as artificial intelligence, automation, advanced manufacturing, sustainable systems, and digital infrastructure are continuously reshaping professional practice [2]. As a result, engineering graduates are expected not only to possess strong technical foundations but also to demonstrate adaptability, interdisciplinary thinking and problem solving capacity in engaging with complex socio technical systems [3].

Traditional course structures, often designed around stable disciplinary boundaries and static knowledge bases, struggle to keep pace with these dynamic conditions. Engineering knowledge evolves quickly, and competencies that were considered advanced only a few years ago may become foundational or even obsolete. Consequently, periodic and systematic course redesign is no longer optional, but essential to maintaining academic relevance, graduate employability, and institutional competitiveness. Course redesign at the university level involves more than content updating. It requires rethinking pedagogical approaches [4], integrating digital learning technologies [5], aligning assessment methods with competency development, and strengthening connections with industry and professional bodies [6].

[1] silvia.carpitella@csun.edu

Engineering education in particular demands experiential learning opportunities, collaborative problem solving environments, and exposure to real world challenges. These requirements place pressure on curriculum designers to adopt holistic and flexible improvement strategies. Furthermore, the increasing diversity of student populations and the expansion of digital delivery modes introduce additional complexity. Learners expect interactive environments, timely feedback, and opportunities for personalized learning pathways. At the same time, accreditation frameworks and quality assurance mechanisms impose structured standards that must be integrated into course development processes [7,8]. Balancing innovation with compliance becomes a central managerial and academic challenge. Given these conditions, course redesign should be approached as a systemic process characterized by interdependent decisions rather than isolated adjustments [9]. Changes in pedagogical design may require new assessment strategies. Digital integration may influence student engagement patterns and resource allocation. Industry alignment may reshape learning outcomes and competency frameworks. Understanding these mutual influences is critical for informed decision making. Therefore, there is a growing need for structured analytical approaches that can model the complex relationships among course improvement configurations in engineering education [10]. By capturing the interdependencies among key strategic dimensions, universities can better prioritize interventions and allocate resources effectively. In rapidly evolving engineering environments, such evidence based and context sensitive redesign processes are fundamental to sustaining educational quality and long term institutional relevance.

The goal of this research is to analyze the mutual influences among key course improvement configurations in engineering education by applying the Decision Making Trial and Evaluation Laboratory (DEMATEL) method [11] within a multi expert framework. Specifically, the study aims to identify the causal structure linking digital integration, pedagogical innovation, assessment and feedback strategy, and industry alignment, while examining how variations in the relative weight assigned to different experts affect the resulting influence network. By introducing and testing a contextual expert weighting approach, this research provides a context aware foundation for strategic decision making in university level engineering education.

The methodological details of DEMATEL are detailed in Section 2, Section 3 reports the practical application by discussing redesign implementation, and Section 4 reports the conclusions of this work with potential future development lines.

2 Methodological Approach

The DEMATEL method is well suited to this context because it models complex systems with interdependent and feedback relationships among factors. Unlike hierarchical methods that assume independence, DEMATEL identifies cause and effect structures within an influence network [12], which is essential for analyzing mutually interacting course improvement configurations. It is also advantageous in multi expert settings, as it enables the aggregation of different expert judgments and supports weighting schemes [13], allowing the analysis to reflect diverse perspectives and assess their impact on the resulting causal structure.

The implementation of the DEMATEL method involves the following steps [14,15].

1. Definition of the system and factor set. The procedure is initialized by defining a finite set of factors representing the components of the system under investigation as:

$$F = \{f_1, f_2, \ldots, f_n\}. \tag{1}$$

2. Elicitation of direct influence assessments. Experts evaluate the degree of direct influence of factor f_i on factor f_j using a predefined scale. The evaluations form the input matrix:

$$\mathbf{A} = [a_{ij}]_{n \times n} \tag{2}$$

where $a_{ij} \geq 0$ and $a_{ii} = 0$ for all i.

3. Aggregation of expert judgments. In a multi expert context, individual direct relation matrices are aggregated, typically using a weighted average:

$$\mathbf{A} = \sum_{k=1}^{m} w_k \mathbf{A}^{(k)} \tag{3}$$

where $\mathbf{A}^{(k)}$ is the matrix provided by expert k, w_k is the weight assigned to expert k, and

$$\sum_{k=1}^{m} w_k = 1. \tag{4}$$

4. Normalization of the direct relation matrix. The aggregated matrix $\mathbf{A}$ is normalized to obtain the matrix $\mathbf{D}$:

$$\mathbf{D} = \frac{1}{\max_i \sum_{j=1}^{n} a_{ij}} \mathbf{A} \tag{5}$$

ensuring that the maximum row sum equals one and guaranteeing convergence in subsequent calculations.

5. Computation of the total relation matrix. The total relation matrix captures both direct and indirect effects among factors and is calculated as follows:

$$\mathbf{T} = \mathbf{D}(\mathbf{I} - \mathbf{D})^{-1} \tag{6}$$

where $\mathbf{I}$ is the identity matrix.

6. Derivation of prominence and relation indices. Let

$$D_i = \sum_{j=1}^{n} t_{ij} \quad \text{and} \quad R_i = \sum_{j=1}^{n} t_{ji} \tag{7}$$

where t_{ij} denotes the elements of matrix $\mathbf{T}$. The indices are defined as:

$$\text{Prominence}_i = D_i + R_i \tag{8}$$

$$\text{Relation}_i = D_i - R_i \tag{9}$$

$D_i + R_i$ indicates the overall involvement of factor i in the system, while $D_i - R_i$ distinguishes net cause factors (positive values) from net effect factors (negative values).

7. Construction of the impact relation map. The causal structure of the system is visualized by plotting each factor using the coordinates

$$(D_i + R_i,\ D_i - R_i) \tag{10}$$

which enables identification of driving and dependent elements within the system.

3 Case Study

This case study is in the context of graduate engineering education, where programs are expected to provide advanced technical expertise while simultaneously fostering research capability, innovation capacity, and industry readiness. Graduate courses operate in a highly dynamic environment characterized by rapid technological development, increasing interdisciplinarity, and strong expectations from industry, accreditation bodies, and research stakeholders. As a result, course redesign at the graduate level requires a strategic alignment between advanced knowledge delivery, research integration, and professional relevance.

The institution under consideration aims to evaluate four key configurations for improving graduate engineering courses. These configurations represent strategic dimensions that collectively influence the quality, relevance, and impact of postgraduate education.

- F_1 Advanced Digital and Research Infrastructure Integration. This factor refers to the incorporation of advanced digital tools and research oriented technologies into graduate teaching. It includes simulation platforms, high performance computing access, virtual laboratories, data analytics tools, and collaborative digital research environments. The objective is to enhance analytical depth, experimentation capability, and research based learning.

- F_2 Research Driven and Inquiry Based Pedagogy. This configuration captures the extent to which graduate courses are structured around research problems, open ended inquiry, and project driven learning. It emphasizes critical analysis, methodological rigor, scientific reasoning, and the development of independent research competencies.

- F_3 Competency Aligned Assessment and Scholarly Feedback. This factor concerns the design of evaluation systems that measure advanced competencies such as research design, innovation capacity, technical communication, and interdisciplinary integration. It includes continuous evaluation, publication oriented outputs, peer review processes, and structured academic mentoring feedback.

- F_4 Industry, Innovation and Professional Integration. This configuration represents the degree to which graduate courses align with industrial research challenges, technological innovation needs, and professional standards. It includes collaboration with industry laboratories, co supervised theses, applied research projects, and integration of entrepreneurial or technology transfer perspectives.

To ensure a comprehensive evaluation of these configurations, three experts with complementary graduate level expertise were involved in the study.

- E_1 Expert in Graduate Engineering Education. A senior faculty member with extensive experience in postgraduate curriculum design, supervision of master and doctoral students, and accreditation processes for advanced engineering programs.

- E_2 Expert in Digital and Research Technologies. A specialist in advanced computational tools, digital research platforms, and technology enabled learning environments relevant to graduate engineering training.

- E_3 Industry and Research Collaboration Expert. A professional engaged in industry research partnerships and innovation projects, with experience in supervising applied graduate research and defining advanced workforce competencies.

Expert judgments were collected individually from the three experts in order to avoid mutual influence and ensure independent evaluation. Each expert assessed the degree of direct influence of one configuration over another using a discrete scale ranging from 0 to 5, where 0 indicates no influence and 5 indicates maximum influence. The pairwise evaluations provided by each expert

were structured into direct relation input matrices. These matrices represent the perceived intensity of influence among the four configurations of graduate course improvement and constitute the primary data for subsequent analysis. The individual input matrices are reported in Tables 1, 2, 3.

Table 1: Input matrix provided by Expert 1, Graduate Engineering Education Specialist

$\mathbf{E}_1$	$\mathbf{F}_1$	$\mathbf{F}_2$	$\mathbf{F}_3$	$\mathbf{F}_4$
$\mathbf{F}_1$	0	2	3	2
$\mathbf{F}_2$	2	0	1	2
$\mathbf{F}_3$	3	3	0	4
$\mathbf{F}_4$	2	2	1	0

Table 2: Input matrix provided by Expert 2, Digital and Research Technologies Specialist

$\mathbf{E}_2$	$\mathbf{F}_1$	$\mathbf{F}_2$	$\mathbf{F}_3$	$\mathbf{F}_4$
$\mathbf{F}_1$	0	4	3	4
$\mathbf{F}_2$	3	0	4	5
$\mathbf{F}_3$	4	4	0	4
$\mathbf{F}_4$	5	4	5	0

Table 3: Input matrix provided by Expert 3, Industry and Research Collaboration Expert

$\mathbf{E}_3$	$\mathbf{F}_1$	$\mathbf{F}_2$	$\mathbf{F}_3$	$\mathbf{F}_4$
$\mathbf{F}_1$	0	4	2	3
$\mathbf{F}_2$	3	0	3	2
$\mathbf{F}_3$	3	3	0	4
$\mathbf{F}_4$	4	2	3	0

The divergence in influence intensity across experts reflects their epistemic positioning within the graduate education ecosystem. While the academic expert emphasizes assessment as a structuring mechanism, the technology expert perceives systemic interdependence driven by infrastructure, and the industry expert prioritizes professional integration as a causal driver.

For the initial iteration, differentiated weights were assigned to reflect the strategic orientation of the graduate program under analysis. Specifically, Expert 1 was assigned a weight of 15%, Expert 2 a weight of 45%, and Expert 3 a weight of 40%. This configuration prioritizes technological infrastructure and industry integration, consistent with a scenario in which the institution is pursuing a strategy focused on digital transformation and strengthening research collaboration with external stakeholders. Under this perspective, expertise related to advanced digital environments and industry engagement is considered to be more influential in shaping course redesign decisions.

The aggregation process was repeated under an alternative weighting scheme in which Expert 1 was assigned a weight of 50%, while Experts 2 and 3 were each assigned 25%. This second configuration reflects a scenario in which academic and pedagogical considerations are of primary importance in guiding graduate course redesign. By contrasting these two weighting distributions, the analysis captures how shifts in institutional priorities may alter the perceived hierarchy and causal relationships among the course improvement configurations.

The individual input matrices were aggregated under each of the two weighting scenarios described above, generating two distinct collective direct relation matrices. The aggregation results have been visualized and reported in Figures 1, 2. These figures illustrate the weighted integration of expert judgments under the technology and industry oriented scenario and the academically driven scenario, respectively. The aggregated matrices obtained from each configuration serve as the basis for the subsequent analytical phase. Specifically, they are used to conduct two separate iterations of the DEMATEL procedure, allowing for a comparative examination of how different expert weighting schemes affect the causal structure of the course improvement configurations.

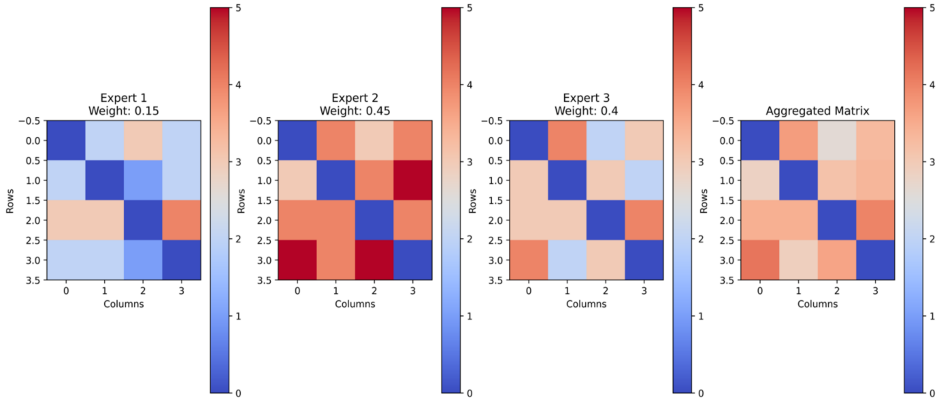


Figure 1: Aggregated matrix under technology and industry oriented weights

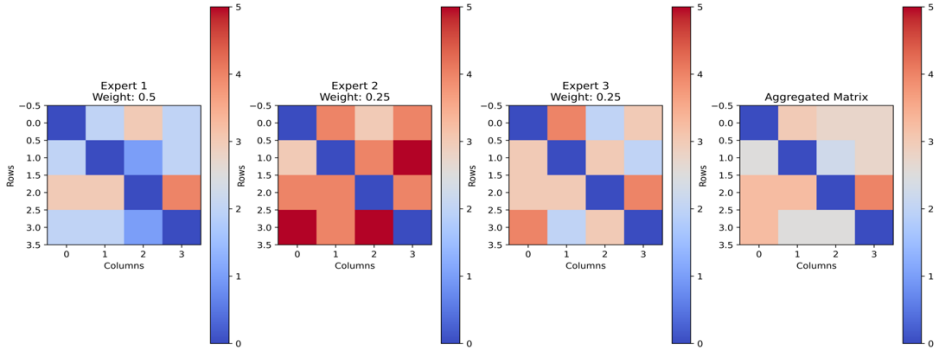


Figure 2: Aggregated matrix under academically oriented weights

The results of the DEMATEL analysis under the two expert weighting scenarios are reported in Tables 4 and 5, which present the total relation matrices together with the corresponding row sums, column sums, prominence values and relation indices for each factor. The magnitude of prominence variation across scenarios confirms the structural sensitivity of the system. The graphical representation of the influence structures is illustrated in Figure 3, which depicts the impact relation maps associated with the two weighting scenarios, enabling visual comparison of the changes in causal positioning resulting from changes in expert weighting.

Table 4: Total Relation Matrix and results under technology and industry oriented weights

	F_1	F_2	F_3	F_4	D_i	R_i	$D_i + R_i$	$D_i - R_i$
F_1	2.9966	3.1547	2.9413	3.2589	12.3515	13.2673	25.6188	-0.9158
F_2	3.1638	2.8593	2.9336	3.2205	12.1772	12.8399	25.0172	-0.6627
F_3	3.5658	3.4565	3.0474	3.6272	13.6969	12.1594	25.8565	1.5375
F_4	3.5411	3.3695	3.2371	3.2956	13.4433	13.4022	26.8455	0.0411

Table 5: Total Relation Matrix and results under academically oriented weights

	F_1	F_2	F_3	F_4	D_i	R_i	$D_i + R_i$	$D_i - R_i$
F_1	1.0150	1.2121	1.0873	1.2594	4.5738	4.7894	9.3633	-0.2156
F_2	1.1094	0.8930	0.9716	1.1565	4.1305	4.6682	8.7986	-0.5377
F_3	1.4355	1.4030	1.0385	1.5200	5.3970	4.1506	9.5476	1.2464
F_4	1.2296	1.1600	1.0532	1.0271	4.4699	4.9630	9.4328	-0.4931

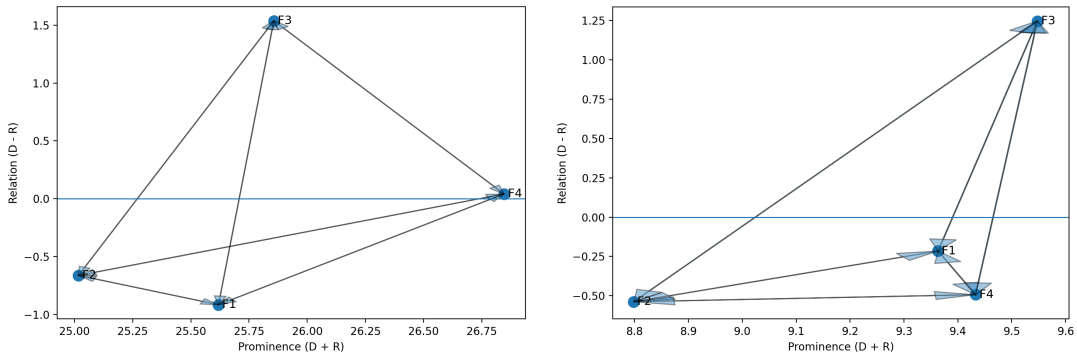


Figure 3: Impact relation maps under (a) technology and industry oriented weighting and (b) academically oriented weighting

The results reveal a clear shift in the dominant causal configuration across the two weighting scenarios. Under the technology and industry oriented weighting scheme, F_4 Industry, Innovation, and Professional Integration exhibits the highest prominence and moves toward the cause group, indicating that when technological and external collaboration perspectives dominate, industry integration becomes structurally central in shaping graduate course redesign. In contrast, in the academically oriented weighting scenario, the Competency Aligned Assessment and Scholarly Feedback F_3 emerges as the strongest cause factor, assuming a clear driving role within the system. In this configuration, the assessment logic becomes the primary structuring mechanism that influences the other dimensions. The remaining factors, namely Advanced Digital and Research Infrastructure Integration and Research Driven Pedagogy, display more context dependent and generally reactive roles, shifting between secondary influence and effect positions depending on the prevailing expert weighting.

4 Conclusions and Future Research

This study analyzed the mutual influence among four strategic configurations for graduate engineering course redesign using a weighted multi expert DEMATEL framework. In particular, the analysis revealed a clear shift in the dominant driving factor across the two weighting scenarios. When technological and industry expertise was prioritized, Industry, Innovation and Professional Integration emerged as the structurally central configuration. In contrast, when academic and pedagogical expertise assumed greater weight, Competency Aligned Assessment and Scholarly Feedback became the primary causal driver. More importantly, the results demonstrate that graduate course redesign should not be interpreted as a competition among strategic priorities, but as a structurally interdependent system in which assessment logic, technological infrastructure, pedagogical organization, and industry integration co-evolve. Industry integration cannot be effective without appropriate assessment mechanisms capable of evaluating the results of research and innovation. Similarly, advanced digital infrastructure becomes meaningful only when embedded within inquiry driven pedagogy and supported by coherent competency frameworks. Even when one configuration appears to assume a leading role, its influence propagates through a network of reciprocal dependencies that require collaborative calibration.

Future research may involve additional stakeholders, such as students and program coordinators, to broaden the perspective of course redesign. The integration of DEMATEL with complementary methods could further validate interdependencies among factors. Applying the framework across institutions or over time would also improve the robustness and generalizability of the findings.

References

[1] Rahmadi, I. F., Research on digital transformation in higher education: Present concerns and future endeavours. *TechTrends*, 68(4), 647–660, 2024.

[2] James, J., Malkin, T., Nusche, D., Trends Shaping Education 2025. *OECD Publishing*, 2025.

[3] Pandya, B., Ruhi, U., Patterson, L., Preparing the future workforce for 2030: The role of higher education institutions. *Frontiers in Education*, 8, 1295249, 2023.

[4] Law, M. Y., A review of curriculum change and innovation for higher education. *Journal of Education and Training Studies*, 10(2), 16, 2022.

[5] Børte, K., Lillejord, S., Learning to teach: Aligning pedagogy and technology in a learning design tool. *Teaching and Teacher Education*, 148, 104693, 2024.

[6] Bari, L., Connecting academia and industry: Change management in curriculum co-design. *Industry and Higher Education*, 39(4), 395–407, 2025.

[7] Duarte, N., Vardasca, R., Literature review of accreditation systems in higher education. *Education Sciences*, 13(6), 582, 2023.

[8] Puerta-Guardo, F. A., Cantillo-Orozco, A. S., Castillo-Loaiza, J. L., Narváez-Grisales, J. A., Molina-Guerrero, C. J., Quality in Higher Education Institutions as a Transversal Tool in Institutional Accreditation: A Bibliometric Review. *European Journal of Educational Research*, 15(1), 19–38, 2026.

[9] Buitrago-Florez, F., Sanchez, M., Pérez Romanello, V., Hernandez, C., Hernández Hoyos, M., A systematic approach for curriculum redesign of introductory courses in engineering: a programming course case study. *Kybernetes*, 52(10), 3904–3917, 2023.

[10] De Silva, L. M. H., Rodríguez-Triana, M. J., Chounta, I.-A., Pishtari, G., Curriculum analytics in higher education institutions: a systematic literature review. *Journal of Computing in Higher Education*, 37(3), 898–944, 2025.

[11] Si, S. L., You, X. Y., Liu, H. C., Zhang, P., DEMATEL technique: a systematic review of the state-of-the-art literature on methodologies and applications. *Mathematical Problems in Engineering*, 2018(1), 3696457, 2018.

[12] John, R., Singh, A. K., A DEMATEL approach for analysing the interdependence among the efficiency barriers in the agri-fresh produce supply chains. *Supply Chain Analytics*, 10, 100106, 2025.

[13] Chen, W., Li, W., Shao, L., Zhang, T., Wang, X., Large-scale group-hierarchical DEMATEL method for complex systems. *PLOS ONE*, 18(12), e0288326, 2023.

[14] Gabus, A., Fontela, E., World Problems, an Invitation to Further Thought within the Framework of DEMATEL. *Geneva Research Centre, Battelle Memorial Institute*, Geneva, 1972.

[15] Gabus, A., Fontela, E., Perceptions of the World Problematique: Communication Procedure, Communicating with Those Bearing Collective Responsibility. *Battelle Geneva Research Centre*, 1973.

Optimal portfolio insurance: a new look

Alejandro Balbás[♭, 1], Beatriz Balbás[♮] and Raquel Balbás[♮]

(♭) University Carlos III of Madrid,
C/ Madrid, 126. 28903 Getafe (Madrid, Spain). alejandro.balbas@uc3m.es. 0000-0003-1678-0550.
(♮) University Complutense of Madrid,
Somosaguas, 28223 Pozuelo de Alarcón (Madrid, Spain). bebalbas@ucm.es. 0000-0002-4725-752X,
raquel.balbas@ccee.ucm.es. 0000-0003-1744-3038

1 Introduction

The use of monetary risk measures and performance measures in portfolio selection is becoming increasingly popular in the literature (Mansini et al., 2007, Stoyanov et al., 2007, Balbás et al., 2010 and 2022, Scutellà and Recchia, 2013, Kapsos et al., 2014, Lejeune and Shen, 2016, Bernard et al., 2019, and many others). Nevertheless, its use in portfolio insurance appears to be extremely limited, despite the fact that portfolio insurance may play a critical role in risk management problems, at least if the financial market is going through turbulent times. The wealth protection when facing market a turmoil was always a major question for many managers (Leland and Rubistern, 1976, Bertrand and Prigent, 2003, Annaert et al., 2009, Zagst and Kraus, 2011, Nangolo et al., 2023, etc.). This paper deals with optimal portfolio insurance problems involving three major objectives, namely, the monetary risk measures value at risk and conditional value at risk (henceforth V@R and CV@R, Rockafellar and Uryasev, 2000) and the expected wealth. Both V@R and CV@R may critically depend on a selected confidence level. If the confidence level is modified, then almost nothing is known about the risk level of the selected insurance strategy. For that reason the first objective of the proposed vector optimization problem will be the minimization of the difference CV@R-V@R for a specific confidence level. Recent literature has shown that the difference above allows us to give a good upper bound of both V@R and CV@R beyond any confidence level.

The presentation outline is as follows. The first part is devoted to summarizing some background that will be later very important. In particular, the notions of stochastic discount factor (SDF),V@R and CV@R are recalled, as well as their main properties. Theorem 2 borrows a (non new) polynomial upper bound of V@R and VC@R with coefficients ϑ and V which are constant parameters. Since ϑ is proportional to the difference CV@R-V@R for a specific (but arbitrary) confidence level, for the rest of confidence levels the CV@R-value may be accurately controlled.

The second part is devoted to presenting the manager problem. Suppose that the random variable U represents the manager wealth at a future horizon T. The manager wonders how to add a protection $f(U)$ so as to minimize the risk (V@R and CV@R) of the total strategy $U+f(U)$ and simultaneously maximize the expected value of $U+f(U)$. It will be shown that the optimization of the three objectives above in the sense of Pareto is a particular case of the simultaneous optimization of CV@R-V@R, V@R and

expected wealth, but replacing CV@R with CV@R-V@R has the aforementioned advantage. Theorem 7 is a very important result of this part because it enables us to transform the original problem into an equivalent linear one.

The third part provides us with the problem solution in Theorem 10. The generality of this theorem is significant, as it yields a valid solution for static and dynamic frameworks, and for every asset valuation model, provided that it is free of arbitrage opportunities. A very important consequence of Theorem 10 is that the optimal portfolio insurance problem is always solved by a strategy *f(U)* saturating the imposed constraints. In other words, a partial hedging is never optimal, despite the fact that managers often select a partial hedging strategy in practice.

The fourth part deals with the Black-Scholes-Merton (BSM) pricing model (Duffie, 1988) and particularizes the general solution of Theorem 10. It is usual to consider the BSM model when studying portfolio insurance problems (Leland and Rubistern, 1976, or Nangolo et al., 2023). Although other models may be also appropriate (Bertrand and Prigent, 2003), BSM only requires the calibration of two parameters (drift and volatility), and therefore it is less dependent on errors made in the estimation and calibration processes. As will be seen, BSM and Theorem 10 imply that the optimal portfolio insurance strategy rarely equals the sale of future contracts or the purchase of put options, despite the fact that both alternatives are frequently used in practice (Joossens and Schoutens, 2008, Annaert et al., 2009, etc.).

2 Conclusion

The portfolio insurance problem is addressed as a vector optimization problem involving three objectives, namely, the conditional value at risk, the value at risk and the expected wealth. Furthermore, the first objective is slightly modified so as to involve more confidence levels that may become also interesting to the manager. The problem is linearized and totally solved, providing us with a general solution applying for every arbitrage free buy and hold or dynamic pricing model. The general solution is particularized for the Black-Scholes-Merton model, and several numerical experiments are implemented. These experiments show that classical strategies such as the sale of futures or the purchase of European puts are rarely optimal. In contrast, the sale of OTM European calls becomes very frequently optimal, even when facing a market turmoil.

Acknowledgements

This research is partially supported by the Spanish Ministry of Science and Innovation (Project PID2021-125133NB-I00) and the University Carlos III of Madrid (Project 2009/00445/003)..

References

[1] Allaj, E., M.E. Mancino and S. Sanfelici (2025). Identifying the number of latent factors of stochastic volatility models. Decisions in Economics and Finance, 48, 571-602.

[2] Annaert, J., S.V. Osselaer and V Bert (2009). Performance evaluation of portfolio insurance strategies using stochastic dominance criteria. Journal of Banking and Finance, 33, 272-280.

[3] Anderson, E.J. and P. Nash (1987). Linear programming in infinite dimensional spaces. John Wiley & Sons.

[4] Balbás A., B. Balbás and R. Balbás (2010). CAPM and APT-like models with risk measures. Journal of Banking and Finance, 34, 1166-1174.

[5] Balbás A., B. Balbás and R. Balbás (2019).Golden options in financial mathematics. Mathematics and Financial Economics, 13, 637-659.

[6] Balbás A., B. Balbás and R. Balbás (2022). Pareto efficient buy and hold investment strategies under order book linked constraints. Annals of Operations Research, 311, 945-965.

[7] Balbás, A., B. Balbás, R. Balbás and J.P. Charron (2023a). Bidual representation of expectiles. Risks, 11, 220.

[8] Balbás, A., B. Balbás, R. Balbás and A. Heras (2023b). Actuarial pricing with financial methods. Scandinavian Actuarial Journal, 2023 (5), 450-476.

[9] Balbás, A., B. Balbás, R. Balbás and A. Heras (2026). Optimal reinsurance: the role of the difference CVaR-VaR. Preprint.

[10] Bernard, C. , S. Vanduffel and J. Ye (2019). Optimal strategies under omega ratio. European Journal of Operational Research, 275 , 755-767.

[11] Bertrand, P. and J.L. Prigent (2003). Portfolio insurance strategies: a comparison of standard methods when the volatility of the stock is stochastic. International Journal of Business, 8, 461-472.

[12] Dhaene, J., M. Denuit, M. Goovaerts, R. Kaas, and D. Vyncke (2002). The concept of comonotonicity in actuarial science and finance: theory. Insurance: Mathematics and Economics, 31, 3-33.

[13] Duffie, D. (1988). Security markets: stochastic models. Academic Press

[14] Hamada, M. and M. Sherris (2003). Contingent claim pricing using probability distortion operators: method from insurance risk pricing and their relationship to financial theory. Applied Mathematical Finance, 10, 19-47.

[15] Herdegen, M. and C. Munari (2023). An elementary proof of the dual representation of expected shortfall. Mathematics and Financial Economics, 17, 655-662.

[16] Hull, J.C. (2021). Options, futures, and other derivatives, global edition. Pearson.

[17] Joossens, E and W. Schoutens (2008). An overview of portfolio insurances: CPPI and CPDO. JCR Scientific and Technical Reports. European Commission. Joint Research Centre. Institute for the Protection and Security of the Citizen.

[18] Kapsos, M. , N. Christofides and B. Rustem (2014). Worst-case robust omega ratio. European Journal of Operational Research, 234, 499-507 .

[19] Kopp, P.E. (1984). Martingales and stochastic integrals. Cambridge University Press.

[20] Lejeune, M.A. and S. Shen (2016). Multi-objective probabilistically constrained programs with variable risk: Models for multi-portfolio financial optimization. European Journal of Operational Research, 252, 522-39.

[21] Leland, H and M. Rubistern (1976). The evolution of portfolio insurance. Wiley.

[22] Mansini, R.,W. Ogryczak and M.G. Speranza (2007).Conditional value at risk and related linear programming models for portfolio optimization. Annals of Operations Research, 152, 227--256.

[23] Nangolo, P., E.R. Offen and O. Basmanebothe (2023). The performance of option-based portfolio insurance on a dividend payment stock. Journal of Mathematical Finance,13, 180-190.

[24] Overhaus, M., A. Bermudez, H. Buehler, A. Ferraris, C. Jordinson, and A. Lamnouar (2007). Equity hybrid derivatives. John Wiley & Sons.

[25] Rockafellar, R.T. and S. Uryasev (2000). Optimization of conditional-value-at-risk. The Journal of Risk, 2, 3, 21- 42.

[26] Rockafellar, R.T., S. Uryasev and M. Zabarankin (2006). Generalized deviations in risk analysis. Finance & Stochastics, 10, 51-74.

[27] Scutellà, M. G. and R. Recchia (2013). Robust portfolio asset allocation and risk measures. Annals of Operations Research, 204, 145-169.

[28] Stoyanov, S.V., S.T. Rachev and F.J. Fabozzi (2007). Optimal financial portfolios. Applied Mathematical Finance,14, 401-36.

[29] Wang S. S. (2000). A class of distortion operators for pricing financial and insurance risks. Journal of Risk and Insurance, 67, 15--36.

[30] Zagst, R. and J. Kraus (2011) Stochastic dominance of portfolio insurance strategies. Annals of Operations Research, 185, 75-103.

[31] Zeidler, E. (1995). Applied functional analysis: Main principles and their applications. Springer.

Critical aspects for road maintenance prioritization

C.L. Parlapiano[♭],[1] S. Carpitella[♮], A. Certa[♭] and J. Izquierdo[♢]

(♭) Department of Engineering, University of Palermo; Viale delle Scienze, 90128, Italy

(♮) Department of Manufacturing Systems Engineering and Management, California State University, Northridge; 18111 Nordhoff St, CA 91330, US

(♢) Universitat Politècnica de València, Cno. de Vera s/n, Valencia, Spain.

1 General Background

Road infrastructure maintenance represents a critical challenge for transportation agencies, as limited financial resources must be allocated across extensive and often deteriorating networks. In this context, the prioritization of maintenance interventions has been widely addressed in the literature as a Multi-Criteria Decision-Making (MCDM) problem, requiring the simultaneous consideration of such criteria as technical, economic, social, and operational factors. Despite these advancements, the literature does not converge toward a standardized or universally accepted set of evaluation criteria. Instead, existing approaches often rely on case-specific selections, which may limit the transparency, comparability, and transferability of results. This lack of consistency reveals a gap in the systematic identification of criteria that are grounded in the literature and suitable for structured decision-making methods. To address this issue, the present study aims to define a coherent and literature-supported set of criteria for the prioritization of road maintenance interventions. Building on this foundation, the *Technique for Order Preference by Similarity to Ideal Solution* (TOPSIS) is adopted to rank alternative interventions. By combining a literature-based set of criteria with a well-established MCDM method, the study provides a transparent and replicable framework for maintenance prioritization.

2 Literature Review

In the context of this research, maintenance prioritization requires the simultaneous consideration of multiple criteria, often conflicting with each other. As previously discussed, among the various dimensions that can be assessed within this context, technical aspects play a central role, with pavement-related indicators consistently emerging as some of the most frequently adopted ones. Huang *et al.* [1] identified such maintenance indicators as pothole count, the International Roughness Index (IRI), and the Pavement Condition Index (PCI) as the most critical aspects in a two-stage pavement maintenance strategy. Similarly, Hasan *et al.* [2] included potholes and cracking among the road-condition sub-criteria, while Khichad *et al.* [3] extended the technical set to shoulder condition, drainage, skid resistance, and camber. A similar emphasis on condition-based indicators can also be found in other studies [4], [5], [6].

[1] costanzalaura.parlapiano@unipa.it

Along the same line, Ouma *et al.* [7] relate maintenance prioritization to pavement distress manifestations such as cracking, potholes, raveling, and patching, while identifying road safety and pavement surface preservation as the dominant maintenance objectives. Nevertheless, several studies suggest that pavement-condition indicators alone are not sufficient for road maintenance prioritization, and that they should be complemented by additional criteria reflecting, for example, traffic conditions, safety concerns, socio-economic relevance and other significant institutional factors. Consistent with this, the literature does not support the existence of a universal set of evaluation criteria. Instead, it suggests that criteria selection should be tailored to the specific objectives, constraints, and context of the analysis. In this perspective, pavement-condition indicators are frequently integrated with variables related to traffic demand, strategic relevance within the network, economic efficiency, accident-related impacts and associated social costs, as well as land use [2], [5], [8]. This multidimensional view is clearly reflected in the work carried out by Hasan *et al.* [2], where road condition emerges as the highest-weighted main criterion, while institutional and urban dimensions are ranked immediately afterwards. However, although road condition dominates at the aggregate level, some sub-criteria belonging to lower-weighted dimensions emerge as highly influential when considered individually. This suggests that dimensions which appear secondary in the overall hierarchy may still embed highly influential factors for decision-making. Accordingly, "Inclusion of the road into strategic planning" (IA_1), "Adding and building new services" (UR_1) and "Cost" (EC1) emerge as the highest-weighted sub-criteria according to the study [2].

Such evidence indicates that, although technical performance remains central in identifying maintenance needs, prioritization frameworks are expected to capture a broader set of factors in order to support decision-making. In light of this evidence, the present study defines a set of relevant criteria that reflects the main dimensions identified in the literature by means of which interventions can be prioritized by using the *Technique for Order Preference by Similarity to Ideal Solution* (TOPSIS). In particular, the main aspects considered are briefly described below.

- **C_1, pavement condition.** This criterion reflects the current physical and functional condition of the road surface and represents the primary indicator of maintenance need [9]. It can be quantified through metrics such as PCI, IRI, pothole density, and severity of cracking or other surface distresses. Poor pavement condition directly affects ride quality, vehicle operating costs, and long-term infrastructure deterioration.

- **C_2, traffic demand and usage intensity.** This criterion captures the level of utilization of a road segment, typically measured through traffic volume, average daily traffic, and the proportion of heavy vehicles [10]. Roads with higher traffic loads, especially freight traffic, experience faster deterioration and have greater functional importance, thus requiring higher prioritization.

- **C_3, road safety.** This criterion accounts for the safety performance of the road segment, considering indicators such as accident frequency and severity, skid resistance, visibility conditions, and the presence of hazardous locations [11]. Maintenance interventions can significantly reduce accident risks, making safety a key factor in prioritization.

- **C_4, economic efficiency.** This criterion evaluates the financial implications of maintenance interventions, including direct intervention costs as well as broader cost–benefit considerations [8]. It may incorporate lifecycle costs, user cost savings, and the economic return of maintenance investments, ensuring that limited resources are allocated efficiently.

- **C_5, strategic importance within the network.** This criterion reflects the role of a road segment in the overall transportation system, including its contribution to connectivity, accessibility, and regional mobility [2]. Roads that are part of key corridors, support economic activities, or are included in strategic development plans are typically assigned higher priority.

- **C_6, socio-environmental impact.** This criterion considers the broader effects of road conditions and maintenance on society and the environment [12]. It includes factors such as access to essential services (such as hospitals and schools, among others), land use served, the social costs of poor infrastructure, and potential environmental impacts. This ensures that prioritization accounts for equity and sustainability considerations.

This set was chosen for three reasons. First, all criteria are recurrent in the literature or are directly supported by broader maintenance-prioritization frameworks. Second, they provide a balanced representation of the problem by covering critical issues such as pavement deterioration, serviceability, safety, network use, and public relevance. Third, they are suitable for TOPSIS application because they can be translated into measurable indicators and directly compared across alternatives within a unified decision framework.

Having defined the set of criteria, the set of alternatives is now formalized. The concept of alternatives in road maintenance decision models depends on the specific decision problem addressed. They may represent either where to intervene (road sections, pavement units, projects) or how to intervene (maintenance interventions). For example, Khichad *et al.* [3] treated the five sections of the Jaipur Ring Road as the alternatives in a FAHP–TOPSIS–VIKOR framework, while Hasan *et al.* [2] compared six road projects for maintenance prioritization. Similarly, Huang *et al.* [1] applied their two-stage maintenance strategy at the pavement-unit level, using 100 m pavement management units as the basic units for classification and ranking. Other studies instead consider the alternatives as maintenance options. For example, Liu *et al.* [13] compared four preventive maintenance strategies that are, respectively, slurry seal, micro-surfacing, chip seal, and ultra-thin overlays. Likewise, Ouma *et al.* [7] evaluated alternative maintenance treatments, including fog seal, crack sealing, chip seal, thin cold mix seals such as slurry seal and micro-surfacing, thin hot-mix asphalt overlays, and resurfacing. Finally, Borghetti *et al.* [8] considered the alternatives as maintenance interventions to be prioritized across a road network.

In the present study, alternatives are defined as maintenance interventions, as the primary objective is to rank and prioritize intervention options for efficient resource allocation. The following interventions represent different levels of maintenance intensity, ranging from preventive to structural actions, and are typically applied according to the severity of pavement deterioration.

- **A_1, routine monitoring.** No immediate intervention is applied, the pavement is monitored as a baseline or deferred maintenance option and only minor routine activities are performed.
- **A_2, crack sealing.** Sealing of surface cracks to prevent water infiltration and slow down pavement deterioration.
- **A_3, micro-surfacing.** Thin surface treatments used to seal minor distresses, improve skid resistance, and extend pavement service life.
- **A_4, chip seal.** A bituminous binder with aggregate is applied to provide waterproofing and improve surface texture, typically for low to medium-traffic roads.
- **A_5, thin asphalt overlay.** Placement of a thin hot-mix asphalt layer to restore ride quality and address moderate surface deterioration.
- **A_6, structural overlay.** Application of thicker layers or structural interventions to restore load-bearing capacity in more severely deteriorated pavements.

This choice is consistent with the reviewed literature, where several studies frame alternatives as maintenance strategies rather than spatial units, particularly when the objective is to support decision-making in the selection of intervention types.

While some approaches focus on prioritizing road segments or projects, others emphasize the comparison of feasible maintenance interventions to identify the most effective treatment under given conditions. This definition allows the analysis to directly address the allocation of resources across competing treatment options.

3 TOPSIS method

The *Technique for Order Preference by Similarity to Ideal Solution* (TOPSIS) is a MCDM method developed by Hwang and Yoon in 1981. Its objective is to identify the alternative characterized by the shortest geometric distance from the positive ideal solution and the farthest distance from the negative ideal solution [14]. The input data necessary to implement the method are:

- a decision matrix $G_{ij} = [g_{ij}]_{nm}$ in which the n rows are the alternatives a_i, $(i = 1, \ldots, n)$ and the m columns are the evaluation criteria c_j $(j = 1, \ldots, m)$. The generic element g_{ij} of this matrix represents the score of the alternative a_i with respect to the c_j criterion.
- the criteria weights.

The first step requires the **normalized decision matrix** by using the following formula:

$$z_{ij} = \frac{g_{ij}}{\sqrt{\sum_{i=1}^{n} g_{ij}^2}} \qquad i = 1, \ldots, n;\ j = 1, \ldots, k \qquad \text{(sum by column)} \tag{1}$$

Then, it is necessary to construct the **weighted normalized decision matrix** by using the following formula:

$$u_{ij} = w_j \cdot z_{ij} \qquad i = 1, \ldots, n;\ j = 1, \ldots, k \tag{2}$$

The following step aims to identify the positive (**Azimut** A^*) and the negative (**Nadir** A^-) ideal solutions:

$$A^* = \begin{cases} \max\limits_i u_{ij} & j \in I' \\ \min\limits_i u_{ij} & j \in I'' \end{cases} \qquad A^- = \begin{cases} \min\limits_i u_{ij} & j \in I' \\ \max\limits_i u_{ij} & j \in I'' \end{cases} \tag{3}$$

where I' and I'' are benefit and cost criteria respectively.
Finally, the calculation of relative distances:

$$S_i^* = \sqrt{\sum_{j=1}^{k} (u_{ij} - u_j^*)^2} \qquad \text{(sum by row)} \tag{4}$$

$$S_i^- = \sqrt{\sum_{j=1}^{k} (u_{ij} - u_j^-)^2} \qquad \text{(sum by row)} \tag{5}$$

and the calculation of the relative coefficient closeness:

$$C_i = \frac{S_i^-}{S_i^- + S_i^*} \tag{6}$$

As a result, the ranking of the alternatives based on the C_i is obtained.

4 Case Study

The proposed framework is applied to a case study in Southern California, a region characterized by an extensive and heavily utilized road network subject to significant traffic loads and environmental stressors. In particular, the analysis focuses on urban and peri-urban road segments in the Los Angeles metropolitan area, where high traffic demand, a substantial share of heavy vehicles, and climate-related factors such as temperature fluctuations and surface aging contribute to accelerated pavement deterioration. These conditions, combined with budget constraints and the strategic importance of maintaining network functionality in one of the most densely populated regions in the United States, make road maintenance prioritization a critical and complex task.

The evaluation of maintenance alternatives with respect to the selected criteria is based on expert judgment. In particular, the assessment was carried out through consultation with a transportation engineer with experience in pavement management systems and road maintenance planning. The expert was asked to evaluate the relative performance of each intervention under each criterion by using a discrete scale ranging from 1 to 10, where higher values correspond to more favorable outcomes. In other terms, all criteria are treated as benefit criteria (to be maximized from the TOPSIS perspective). Moreover, in the present study, all criteria are assumed to have equal importance. This assumption is motivated by the exploratory nature of the analysis, as adopting equal weights ensures neutrality and avoids introducing additional subjectivity into the model, while still allowing the evaluation to reflect the multidimensional structure of the problem.

The input data required from the TOPSIS application, along with the resulting C_i values for each alternative are reported in Table 1. Results are visualized in Figure 1.

Table 1: Decision matrix and TOPSIS results

Alternative	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_3$	$\mathbf{C}_4$	$\mathbf{C}_5$	$\mathbf{C}_6$	$\mathbf{C}i$
A_1 Routine monitoring	2	3	2	9	4	3	0.311506
A_2 Crack sealing	5	5	5	8	5	5	0.499750
A_3 Micro-surfacing	7	6	6	7	6	6	0.658631
A_4 Chip seal	6	6	6	7	5	5	0.569207
A_5 Thin asphalt overlay	8	7	7	5	7	6	0.698514
A_6 Structural overlay	9	8	8	3	8	7	0.688494

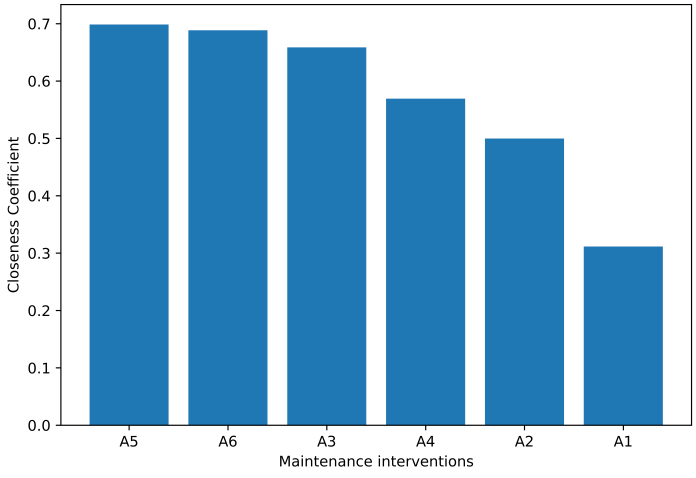


Figure 1: Report of results

The results show that the most suitable maintenance options in the analyzed context are thin asphalt overlay (A_5) and structural overlay (A_6), with closeness coefficients of 0.698514 and 0.688494, respectively. This means that these two options are favorable overall in terms of balance across the selected criteria. On the other hand, routine monitoring (A_1) is the least preferable alternative. It is evident in Table 1 that structural overlay achieves the highest scores in terms of pavement conditions, but its lower economic efficiency penalizes its final ranking. This specific result demonstrates that the preferred intervention is not simply the most technically effective one, but the one that best combines technical effectiveness with broader decision criteria.

The intermediate position of micro-surfacing (A_3) suggests that medium intensity treatments may represent a good compromise when decision-makers aim to balance performance and resource efficiency. By contrast, chip seal (A_4) and crack sealing (A_2) show lower overall suitability, in the present case.

Therefore, the results support the idea that pavement-related and operational criteria remain central in maintenance prioritization, but they also show that economic and other aspects, such as traffic demand, socio-environmental impact, strategic importance or road safety, are called to affect the final ranking, whenever assuming equal or different weights for all criteria and evaluation based on one or more expert judgments. In this sense, the analysis confirms the usefulness of a multi-criteria framework for road maintenance prioritization.

5 Conclusions

This study shows that a TOPSIS-based approach can support road maintenance prioritization by comparing alternative maintenance interventions with a coherent set of technical, operational, economic, strategic, and socio-environmental criteria. In the case analyzed, thin asphalt overlay emerges as the most preferable option, followed closely by structural overlay, whereas routine monitoring ranks as the least preferable alternative.

The main contribution of the study therefore lies in the development of a transparent and replicable decision support framework capable of moving beyond a single-criterion evaluation.

Nevertheless, the analysis is affected by some limitations. First, the assumption regarding all criteria having equal importance is suitable for an exploratory application but does not fully capture actual decision priorities. Second, the evaluation is based on the judgment of a single expert, which inevitably influences the resulting ranking. Third, the set of alternatives is restricted to conventional maintenance interventions. Accordingly, the results should be interpreted as a structured and transparent exploratory outcome rather than as a universally applicable prescription for road maintenance decision-making.

The findings highlight that the most preferable intervention is the one that achieves the most balanced overall performance across multiple criteria. Building on these results, future research could reinforce the framework by introducing a more structured weighting method, involving a larger panel of experts, and assessing the robustness of the ranking through sensitivity analyses on both criteria weights and alternative evaluations. In addition, the set of alternatives could be expanded to include more innovative solutions, particularly those based on predictive maintenance technologies.

References

[1] L. L. Huang, J. D. Lin, W. H. Huang, C. H. Kuo, Y. S. Chiou, and M. Y. Huang, "Developing pavement maintenance strategies and implementing management systems," *Infrastructures*, vol. 9, no. 7, p. 101, 2024. DOI: 10.3390/infrastructures9070101.

[2] A. E. Hasan and F. K. Jaber, "Prioritizing road maintenance: A framework integrating fuzzy best-worst method and vikor for multi-criteria decision making," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 13 990–13 997, 2024. DOI: 10.48084/etasr.7056.

[3] J. S. Khichad, R. J. Vishwakarma, A. Gaur, and A. Sain, "Optimization of highway performance and safety by integrated multi-criteria decision-making techniques," *International Journal of Pavement Research and Technology*, vol. 19, pp. 182–198, 2026. DOI: 10.1007/s42947-024-00452-w.

[4] R. Ceriani, V. Vignali, D. Chiola, M. Pazzini, M. Pettinari, and C. Lantieri, "Exploring the effectiveness of road maintenance interventions on iri value using crowdsourced connected vehicle data," *Sensors*, vol. 25, p. 3091, 2025. DOI: 10.3390/s25103091.

[5] H. Siswanto, B. Supriyanto, Pranoto, R. P. Prihatditya, and M. A. Friansa, "District road maintenance priority using analytical hierarchy process," *AIP Conference Proceedings*, vol. 2114, p. 060 019, 2019. DOI: 10.1063/1.5112490.

[6] P. K. Agarwal, A. B. Khan, and S. Choudhary, "A rational strategy for resource allocation for rural road maintenance," *Transportation Research Procedia*, vol. 25, pp. 2195–2207, 2017. DOI: 10.1016/j.trpro.2017.05.422.

[7] Y. O. Ouma, J. Opudo, and S. Nyambenya, "Comparison of fuzzy ahp and fuzzy topsis for road pavement maintenance prioritization: Methodological exposition and case study," *Advances in Civil Engineering*, vol. 2015, p. 140 189, 2015. DOI: 10.1155/2015/140189.

[8] F. Borghetti, G. Beretta, N. Bongiorno, and M. De Padova, "Road infrastructure maintenance: Operative method for interventions' ranking," *Transportation Research Interdisciplinary Perspectives*, vol. 25, p. 101 100, 2024. DOI: 10.1016/j.trip.2024.101100.

[9] S. Ahmed, P. Vedagiri, and K. K. Rao, "Prioritization of pavement maintenance sections using objective based Analytic Hierarchy Process," *International Journal of Pavement Research and Technology*, vol. 10, no. 2, pp. 158–170, 2017. DOI: 10.1016/j.ijprt.2017.01.001.

[10] G. Hadjidemetriou, M. Tsangaris, and S. Christodoulou, "Pavement condition and traffic indices for prioritizing road maintenance," in *EC3 Conference 2019*, European Council on Computing in Construction, vol. 1, 2019, pp. 213–221. DOI: 10.35490/EC3.2019.239.

[11] K. H. Janstrup, M. Møller, and N. Pilegaard, "A clustering approach to integrate traffic safety in road maintenance prioritization," *Traffic injury prevention*, vol. 20, no. 4, pp. 442–448, 2019. DOI: 10.1080/15389588.2019.1580700.

[12] S. A. Dabous, W. Zeiada, T. Zayed, and R. Al-Ruzouq, "Sustainability-informed multi-criteria decision support framework for ranking and prioritization of pavement sections," *Journal of Cleaner Production*, vol. 244, p. 118 755, 2020. DOI: 10.1016/j.jclepro.2019.118755.

[13] Y. Liu et al., "Multi-criteria decision-making framework (ahp-topsis): Pavement preventive maintenance case study for ordinary national trunk highways," *Buildings*, vol. 14, p. 3048, 2024. DOI: 10.3390/buildings14103048.

[14] G.-H. Tzeng and J.-J. Huang, *Multiple Attribute Decision Making: Methods and Applications*. Boca Raton, FL: CRC Press, 2011. DOI: 10.1201/b11032.

Out-uniform $\{\mathrm{MAX}, \mathrm{MIN}\}$–Boolean networks. Applications to social networks.

Juan A. Aledo[♭], Jose P. Llano[♯], Cristina Santamaría[♮],[1] and Jose C. Valverde[♯]

(♭) Departamento de Matemáticas,
Universidad de Castilla-La Mancha, Albacete, 02071, Spain.

(♮) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, 46022, Spain.

(♯) Instituto de Matemática Aplicada a las Ciencias y la Ingeniería
Albacete, 02071, Spain.

1 Introduction

Boolean networks (BN) constitute a fundamental modeling framework for the analysis of complex discrete dynamical systems, with applications ranging from gene regulatory networks and neural systems to distributed computation and social dynamics. Their appeal lies in the balance they provide between expressive power and mathematical tractability, allowing rich global behaviors to emerge from simple local interaction rules. In particular, understanding the long-term dynamics of Boolean networks, such as the structure of their attractors, remains a central problem, as these features often correspond to stable patterns, recurrent behaviors, or consensus configurations in the underlying red modeled system.

In this work, we focus on a class of Boolean networks defined by synchronous updates and out-uniform local updating functions, where the underlying dependency graph is undirected and includes a self-loop at every node. The local updating rules are given by Boolean functions of maxterm or minterm type; that is, each function computes either the maximum or the minimum of its input variables, allowing each variable input to appear in either positive or negated form. More specifically, the maxterm and minterm Boolean operators of n Boolean variables are given by

$$\mathrm{MAX}(x_1, \ldots, x_n) = z_1 \vee \cdots \vee z_n, \quad \text{and} \quad \mathrm{MIN}(x_1, \ldots, x_n) = z_1 \wedge \cdots \wedge z_n,$$

where $z_i = x_i$ or z_i' for every $1 \leq i \leq n$. Here, $\vee$, $\wedge$, and $'$ denote the usual Boolean operators OR, AND, and NOT, respectively. The out-uniform property means that the variable associated with each vertex i appears either always in direct form or always in complemented form in the local functions of all the neighbors of i.

This class of networks, which we call *out-uniform* $\{\mathrm{MAX}, \mathrm{MIN}\}$–BN [1], naturally generalizes and unifies several well-studied families in the literature, including homogeneous Boolean networks defined by maxterm or minterm functions [2], as well as AND-OR Boolean networks [3], [4], both of which have attracted considerable attention in recent years due to their analytical tractability and wide range of applications.

[1] crisanna@imm.upv.es

As an example of this kind of dynamical systems, let us consider the out-uniform $\{\text{MAX}, \text{MIN}\}$–BN whose dependency graph is displayed in Figure 1 with local updating functions

$$\begin{cases} F_1(x_1, x_2, x_3, x_4) = x_1' \wedge x_2 \wedge x_4' \\ F_2(x_1, x_2, x_3, x_4) = x_1' \vee x_2 \vee x_3 \\ F_3(x_1, x_2, x_3, x_4) = x_2 \wedge x_3 \wedge x_4' \\ F_4(x_1, x_2, x_3, x_4) = x_1' \vee x_3 \vee x_4' \end{cases}$$

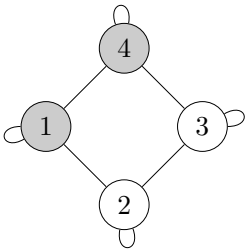


Figure 1: Local updating functions and associated interaction graph of the out-uniform $\{\text{MAX}, \text{MIN}\}$–BN. Nodes with negative variables are colored in gray.

It is easy to verify that $\{(0,1,0,0),(1,1,0,1)\}$ is a 2-periodic orbit and $(0,1,0,1)$ is a fixed point.

The state space of the system is partitioned into two basins of attraction, corresponding to the fixed point and the 2-periodic orbit, as illustrated in Figure 2. Initial configurations that are structurally consistent with the stabilizing interactions converge to the fixed point $(0,1,0,1)$. In contrast, configurations exhibiting conflicting influences evolve towards the 2-periodic orbit $\{(0,1,0,0),(1,1,0,1)\}$. This behavior highlights how the underlying network structure determines the global organization of the dynamics into stable and oscillatory regions.

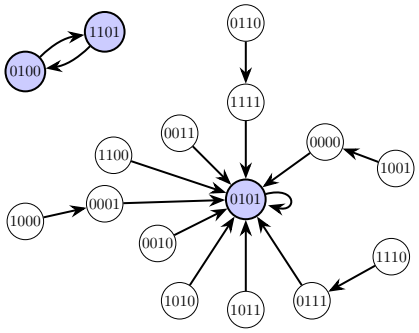


Figure 2: Transition diagram and basins of attraction of the Boolean network. The periodic orbits of the system are colored in blue.

This example shows that, even in simple out-uniform $\{\text{MAX}, \text{MIN}\}$–BN, the orbital structure can be richer than in previously studied classes. In particular, unlike homogeneous Boolean networks—where fixed points and 2-periodic orbits cannot coexist—and AND-OR networks—where only fixed points arise—both types of attractors may appear simultaneously.

2 Main Results

The main goal of this work is to provide a detailed description of the asymptotic behavior of out-uniform $\{\text{MAX}, \text{MIN}\}$–BN. We show that their periodic dynamics is highly constrained: every trajectory eventually converges either to a fixed point or to a periodic orbit of length two. Moreover, these two types of attractors may coexist within the same system, in contrast with previously studied subclasses discussed above.

Building on this result, we obtain a complete structural characterization of the networks that exhibit exclusively fixed points and those in which only 2-periodic orbits arise. Furthermore, in the case where fixed points exist, we provide an explicit description of their number and structural properties in terms of the underlying interaction graph.

As part of our analysis, we also establish a general Fixed Point Theorem, which provides sufficient conditions under which the system admits a unique fixed point and no other periodic attractors. This result may be of independent interest in the study of dynamical systems.

3 Applications

Beyond their theoretical interest, the results obtained in this work have direct implications in several applied contexts. BN are a well-established tool in systems biology, where they are commonly used to model gene regulatory networks in a qualitative yet informative way. In these models, nodes represent genes and edges encode regulatory interactions such as activation or inhibition. This framework allows the study of how complex interaction patterns give rise to global dynamical behaviors, including stable states and oscillations. In particular, fixed points are often associated with distinct cellular phenotypes, while periodic dynamics may correspond to recurring biological processes such as cell cycles or circadian rhythms ([5], [6]).

It is widely recognized that both the topology of the regulatory network and the nature of the local update rules play crucial roles in determining the resulting dynamical regimes. For this reason, special attention has been given to BN with structurally constrained local functions, such as canalizing or unate functions, due to their ability to capture biologically meaningful interactions while remaining analytically tractable ([7]). In this context, the local rules in out-uniform $\{\mathrm{MAX}, \mathrm{MIN}\}$–BN belong to the class of unate Boolean functions and can be viewed as simple instances of canalizing dynamics.

In this work, we also consider an application to consensus dynamics in social networks, where individuals update their opinions based on local interactions. In this setting, each node represents an agent and its Boolean state encodes a binary decision or opinion. The update rules capture how individuals are influenced by their neighbors. A MAX-type rule models agents who tend to adopt a position whenever at least one neighbor supports it, reflecting imitation or reinforcement behavior. In contrast, a MIN-type rule represents agents who abandon a position as soon as a conflicting signal appears, capturing a more cautious or veto-based attitude.

Within this framework, fixed points correspond to stable consensus configurations, while 2-periodic orbits can be interpreted as persistent disagreement or oscillatory opinion patterns. Our results show that the structural properties of the network determine whether consensus is achievable or whether oscillatory behavior inevitably arises, providing insight into how local interaction rules shape collective outcomes in social systems.

4 Conclusions

In this work, we have analyzed the dynamical behavior of out-uniform $\{\mathrm{MAX}, \mathrm{MIN}\}$–Boolean networks on undirected graphs with self-loops. Our results show that, despite allowing both activating and inhibitory interactions, the global dynamics remain highly structured: every trajectory converges to either a fixed point or a cycle of period two, and, in contrast with previously studied classes, both types of attractors may coexist within the same system.

This characterization highlights the role of structural constraints in shaping the long-term behavior of Boolean networks. In particular, our results show that the interplay between the sign of interactions and the underlying graph topology determines whether the system stabilizes or exhibits persistent oscillations. This provides a unified framework that bridges previously disconnected

models and offers new insight into the emergence of complex collective dynamics from simple local rules.

From an applied perspective, these findings contribute to the understanding of agreement and disagreement phenomena in networked systems. Beyond social consensus models, the results may be relevant in contexts such as voting dynamics, collective decision-making, or coordinated behaviors in economic and financial systems, where local interactions can lead to either stabilization or persistent fluctuations.

Finally, this work opens several directions for future research, including the extension of the analysis to directed interaction graphs, non-uniform influence structures, and networks without self-loops, where a richer and less constrained dynamical behavior is expected.

Acknowledgments

J. A. Aledo was supported by PID2022-139293NB-C32 (MICIU /AEI /10.13039/501100011033 and ERDF, EU), and by 2025-GRIN-38476 (Universidad de Castilla-La Mancha/ERDF). J.P. Llano and J.C. Valverde were supported by the Universidad de Castilla-La Mancha and the "ERDF A way of making Europe" within the Operational Program 2021-2027 through the project 2025-GRIN-38587.

References

[1] J. A. Aledo, J. P. Llano, C. Santamaria, and J. C. Valverde, "Limit cycles of AND-OR-NOT Boolean networks and applications to consensus in social networks," *Preprint*, 2026.

[2] J. A. Aledo, S. Martinez, F. L. Pelayo, and J. C. Valverde, "Parallel discrete dynamical systems on maxterm and minterm Boolean functions," *Mathematical and Computer Modelling*, vol. 55, pp. 666–671, 2012.

[3] E. Goles and G. Hernandez, "Dynamical behavior of Kauffmann networks with AND-OR gates," *Journal of Biological Systems*, vol. 8, pp. 151–175, 2000.

[4] J. Aracena, A. Richard, and L. Salinas, "Maximum number of fixed points in AND–OR–NOT networks," *Journal of Computer and System Sciences*, vol. 80, pp. 1175–1190, 2014.

[5] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of Theoretical Biology*, vol. 22, pp. 437–467, 1969.

[6] S. Huang, "Reprogramming cell fates: Reconciling rarity with robustness," *BioEssays*, vol. 31, pp. 546–560, 2009.

[7] D. Murrugarra, A. Veliz-Cuba, B. Aguilar, and R. Laubenbacher, "Identification of control targets in Boolean molecular network models via computational algebra," *BMC Systems Biology*, vol. 10, p. 94, 2016.

Tackling the Optimal Bucket Order Problem through Integer Linear Programming

Juan A. Aledo[♭1], Concepción Domínguez[♮], Juan de Dios Jaime-Alcántara[♯] and Mercedes Landete[♯]

(♭) Departmento de Matemáticas,
Universidad de Castilla-La Mancha, Albacete, Spain.

(♮) Departamento de Estadística e Investigación Operativa,
Universidad de Murcia, Spain.

(♯) Departamento de Estadística, Matemáticas e Informática.
Instituto Centro de Investigación Operativa,
Universidad Miguel Hernández de Elche, Spain.

1 Introduction

Rank aggregation focuses on deriving a single consensus ordering from several rankings that may disagree with each other. This setting commonly appears in applications where multiple agents or systems provide ordinal evaluations, such as recommendation engines, information retrieval, or data-driven scientific analysis. The challenge lies in producing a final ranking that is reliable, interpretable, and computationally manageable, while also accounting for consistency and fairness considerations [1], [2].

Different modeling approaches have been proposed depending on how discrepancies between rankings are measured and on the characteristics of the input data. Common strategies rely on distance-based formulations, where the goal is to minimize the disagreement between the consensus and the original rankings. These formulations must also accommodate diverse types of input, including incomplete or weighted rankings [3], [4].

Optimization-based methods, particularly those grounded in integer linear programming, provide a flexible and transparent framework for addressing these problems [5], [6]. However, most existing contributions focus on strict rankings, whereas scenarios allowing ties, where items may share the same position, have been less extensively studied, despite their practical relevance in many decision-making contexts.

2 Results

This work introduces a unified integer linear programming framework for rank aggregation in which the consensus takes the form of a weak ordering, allowing ties between elements. The proposed approach enables the explicit incorporation of structural constraints, such as limiting the number of groups or controlling their sizes, as well as additional requirements related to fairness across predefined subsets [7].

[1] JuanAngel.Aledo@uclm.es

Several model extensions are developed within this framework. First, formulations are proposed for cases where the number of groups is fixed, covering both uniform and non-uniform configurations. Second, a top-focused variant is presented, in which the highest-ranked elements are ordered in detail while the remaining items are grouped into a single category.

The framework also integrates fairness constraints, allowing the user to regulate how different groups are represented in the final ordering. A key contribution is the development of an exact optimization model for a well-known bucket-order aggregation problem that had previously only been addressed using heuristic approaches. Specifically, our formulation enables the identification of optimal solutions for Optimal Bucket Order Problem (OBOP) instances [8], thereby allowing a precise assessment of which previously reported heuristic solutions are indeed optimal and which are not [9].

The effectiveness of the proposed models is demonstrated through computational experiments on both benchmark datasets and real-world instances. The results highlight their scalability and their ability to analyze the trade-offs introduced by structural and fairness constraints.

3 Conclusions

The findings show that integer programming provides a powerful and adaptable tool for handling rank aggregation problems with ties. The proposed framework offers a coherent way to model a wide range of problem variants while maintaining interpretability and control over the solution structure.

Moreover, the availability of exact formulations makes it possible to assess the performance of existing heuristic methods. Overall, this work extends current approaches to rank aggregation by delivering a flexible and practically applicable methodology suited to scenarios where non-strict rankings are more appropriate.

Acknowledgments

Concepción Domínguez and Mercedes Landete were supported by PID2021-122344NB-I00 (MCIN/AEI/ERDF "A way of making Europe"), and Concepción Domínguez also by PID2024-156594NB-C21 (MCIU/AEI/ERDF/EU) and Generalitat Valenciana (Emergent CIGE/2024/132). Juan de Dios Jaime-Alcántara and Mercedes Landete were also supported by CIPROM/2024/34 (Conselleria de Educación, Generalitat Valenciana). Juan A. Aledo was supported by SBPLY/21/180225/000062 and PID2022-139293NB-C32 (Junta de Comunidades de Castilla-La Mancha/ERDF and ERDF, EU) and 2025-GRIN-38476 (Universidad de Castilla-La Mancha/ERDF).

References

[1] J. A. Aledo, J. A. Gámez, and A. Rosete, "A consensus set for the aggregation of partial rankings: The case of the optimal set of bucket orders problem," *Information Fusion*, vol. 124, p. 103 353, 2025.

[2] C. Kuhlman and E. Rundensteiner, "Rank aggregation algorithms for fair consensus," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, 2020.

[3] S. Lin, "Rank aggregation methods," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 5, pp. 555–570, 2010.

[4] J. Alcaraz, M. Landete, and J. F. Monge, "Rank aggregation: Models and algorithms," in *The Palgrave Handbook of Operations Research*, Springer, 2022, pp. 153–178.

[5] B. Brancotte, B. Yang, G. Blin, S. Cohen-Boulakia, A. Denise, and S. Hamel, "Rank aggregation with ties: Experiments and analysis," in *Proceedings of the VLDB Endowment (PVLDB)*, ACM-Association for Computing Machinery, vol. 8, 2015, pp. 1202–1213.

[6] Y. Yoo and A. R. Escobedo, "A new binary programming formulation and social choice property for kemeny rank aggregation," *Decision Analysis*, vol. 18, no. 4, pp. 296–320, 2021.

[7] J. A. Aledo, C. Domínguez, J. de Dios Jaime-Alcántara, and M. Landete, "An optimization-based approach to ranking aggregation with weak order outputs," *Preprint*, 2026.

[8] A. Ukkonen, K. Puolamäki, A. Gionis, and H. Mannila, "A randomized approximation algorithm for computing bucket orders," *Information Processing Letters*, vol. 109, no. 7, pp. 356–359, 2009.

[9] J. A. Aledo, J. A. Gámez, and A. Rosete, "Approaching rank aggregation problems by using evolution strategies: The case of the optimal bucket order problem," *European Journal of Operational Research*, vol. 270, no. 3, pp. 982–998, 2018.

High order non standard numerical schemes for a second order delay differential equation

Jean Marc Serrano[♭], Francisco Rodríguez[♮] and María Ángeles Castro[♮1]

(♭) Dpto. de Fundamentos del Análisis Económico, Universidad de Alicante
Ctra. San Vicente del Raspeig s/n, Alicante, Spain.

(♮) Dpto. de Matemática Aplicada e Ingeniería Aeroespacial, Universidad de Alicante
Ctra. San Vicente del Raspeig s/n, Alicante, Spain.

1 Introduction

Delay differential equations (DDE) constitute basic tools to model processes with time lags or memory effects [1]. Nonstandard finite difference (NSFD) numerical schemes [2], [3] have been described for different types of DDE. In particular, families of NSFD schemes of increasing orders, derived from an exact numerical solution, were proposed in [4] for the scalar linear initial value delay problem

$$\begin{aligned} x'(t) &= \alpha x(t) + \beta x(t-\tau), \qquad t > 0, && (1)\\ x(t) &= f(t), \qquad -\tau \le t \le 0. && (2) \end{aligned}$$

These results were extended in [5] to coupled linear delay systems,

$$\begin{aligned} X'(t) &= AX(t) + BX(t-\tau), \qquad t > 0, && (3)\\ X(t) &= F(t), \qquad -\tau \le t \le 0, && (4) \end{aligned}$$

where $X(t)$ and $F(t)$ are d-dimensional vector functions, and A and B are $d \times d$ commuting matrices.

In a subsequent work, [6], an exact numerical solution was obtained for problem (3)-(4) with, in general, non-commuting matrix coefficients, requiring only A and $C = I + A^{-1}B$ being non-singular matrices. This expression was used in [7] to construct exact numerical solutions and NSFD difference schemes for the second orden DDE

$$\begin{aligned} x''(t) &= \alpha x(t) + \beta x(t-\tau), \qquad t > 0, && (5)\\ x(t) &= f(t), \qquad -\tau \le t \le 0. && (6) \end{aligned}$$

In this communication, we deal with the construction of high order NSFD schemes for the second order initial value DDE problem

$$\begin{aligned} x''(t) &= \alpha x'(t) + \beta x(t-\tau), \qquad t > 0, && (7)\\ x(t) &= f(t), \qquad -\tau \le t \le 0. && (8) \end{aligned}$$

[1] ma.castro@ua.es

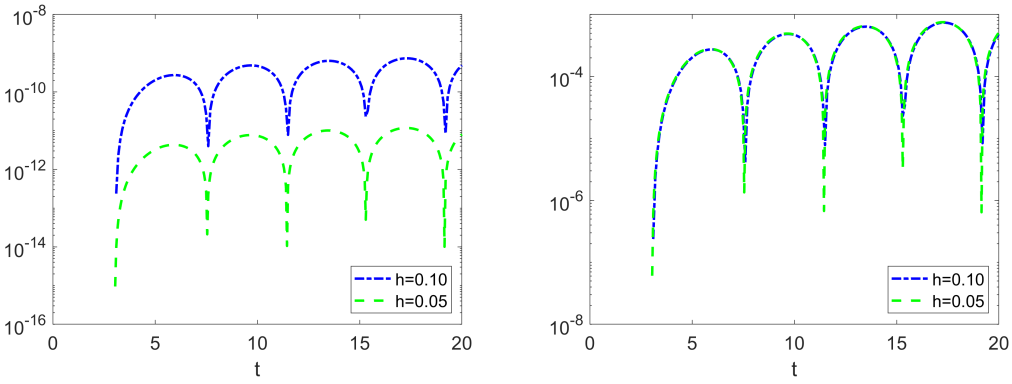


Figure 1: Absolute errors (log-scale) of the numerical solution for problem (7)-(8), with parameters $\alpha = \beta = -1$ and $\tau = 1$, computed using the NSFD scheme with $M = 3$, for two different mesh sizes (left). Errors divided by h^6, that is, relative to the expected order of the method (right).

2 Methods and Results

The general approach used in this work is similar to that used in [7] for the problem (5)-(6), consisting in transforming the second order scalar problem (7)-(8) into the first order vector problem (3)-(4). However, the resulting matrix coefficient A in this case turns out to be singular, so that a necessary first step is to extend the result presented in [6, Theorem 2] avoiding the requirement of A being invertible.

Thus, a first result of this work consists in providing an exact numerical solution of problem (3)-(4) under the milder condition of $A + B$ being invertible, which would be a natural condition for the resulting vector difussion problem when there is no delay.

An expplicit expression for the exact numerical solution of problem (7)-(8) is obtained in terms of spetial functions. As in previous works, the exact numerical solution includes an integral term in the form of a convolution with the initial function. A family of NSFD schemes of increasing orders can be defined by computing the exact solution in the first M intervals of τ amplitude and dropping the integral term in the following intervals.

It can be proved that the resulting NSFD methods are of order $2M$. An example of application is presented in Figure 1, showing the expected order (Fig. 1, right).

3 Conclusions

The exact numerical solution obtained in this work for the vector problem (3)-(4) allows to tackle more general problems than those dealt to in previous works, by not requiring the matrix coefficient A to be invertible.

In particular, it can be used to obtain exact numerical solutions for higher order scalar DDE problems like the one presented in this communication, where the matrix A resulting from the transformation into a first order vector problem is singular.

The family of NSFD methods derived from the exact numerical solution for problem (3)-(4) possess the same high order, $2M$, as those described in [7] for problem (3)-(4), allowing the efficient computation of numerical approximations for large time intervals.

Work in progress will assess the dynamical consistency properties of the proposed NSFD methods, and in particular the delay dependent stability, which seems to hold from numerical experiments.

References

[1] V. Kolmanovskii and A. Myshkis, *Introduction to the Theory and Applications of Functional Differential Equations* (Mathematics and Its Applications). Dordrecht, The Netherlands: Kluwer Academic Publishers, 1999, vol. 463.

[2] R. E. Mickens, *Nonstandard Finite Difference Models of Differential Equations*. World Scientific, 1993. DOI: 10.1142/2081

[3] R. E. Mickens, *Nonstandard Finite Difference Schemes: Methodology and applications*. World Scientific, 2020. DOI: 10.1142/11891

[4] M. García, M. Castro, J. Martín, and F. Rodríguez, "Exact and nonstandard numerical schemes for linear delay differential models," *Applied Mathematics and Computation*, vol. 338, pp. 337–345, 2018. DOI: 10.1016/j.amc.2018.06.029

[5] M. Á. Castro, M. A. García, J. A. Martín, and F. Rodríguez, "Exact and nonstandard finite difference schemes for coupled linear delay differential systems," *Mathematics*, vol. 7, no. 11, 2019. DOI: 10.3390/math7111038

[6] M. Á. Castro, A. Sirvent, and F. Rodríguez, "Nonstandard finite difference schemes for general linear delay differential systems," *Mathematical Methods in the Applied Sciences*, vol. 44, no. 5, pp. 3985–3999, 2021. DOI: 10.1002/mma.7002

[7] M. Á. Castro, C. J. Mayorga, A. Sirvent, and F. Rodríguez, "Exact numerical solutions and high order nonstandard difference schemes for a second order delay differential equation," *Mathematical Methods in the Applied Sciences*, vol. 46, no. 17, pp. 17 962–17 979, 2023. DOI: 10.1002/mma.9540

Understanding Corporate Insolvency Across Regions and Sectors: A Macro-Financial and Machine Learning Perspective from Spain

Yu Zhang[♭,1]

(♭) Programa de Doctorado en Matemáticas, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

Corporate insolvency is a phenomenon whose implications extend well beyond the microeconomic sphere. Although financial deterioration typically unfolds gradually, identifying the point at which a firm becomes critically vulnerable remains challenging, and formal insolvency proceedings often occur with substantial delay. Consequently, insolvency dynamics reflect not only firm-level conditions but also broader macroeconomic, financial, and institutional forces. When insolvencies rise simultaneously across firms, the associated destruction of employment, productive capacity, and fiscal revenues tends to amplify recessionary phases of the business cycle. Empirical evidence consistently documents the countercyclical nature of insolvencies, which increase during economic downturns and decline in expansionary periods [1], [2], [3].

From an analytical perspective, the study of insolvency presents important structural challenges. Insolvency constitutes a legally defined event rather than a direct measure of economic failure, and its occurrence depends on institutional frameworks while being strongly influenced by sectoral and territorial heterogeneity. These characteristics complicate the identification of common determinants and obscure the mechanisms through which aggregate shocks translate into observed insolvency outcomes. Understanding these dynamics therefore requires an approach capable of linking macro-financial conditions with heterogeneous regional and sectoral responses.

While insolvency prediction has a long tradition grounded in firm-level financial indicators [4], [5], [6], much of the literature remains focused on microeconomic determinants. More recent contributions incorporate macroeconomic and financial variables, yet the transmission of aggregate shocks to corporate insolvency—and the extent to which this transmission varies across regions and sectors—remains comparatively underexplored. This gap limits our understanding of how systemic stress propagates through the productive structure of an economy.

This study addresses these issues through three interrelated research questions. First, how do aggregate macro-financial shocks influence regional and sectoral insolvency rates? Second, is this transmission homogeneous, or does it vary systematically according to structural characteristics across territories and industries? Third, can an integrated empirical framework combining econometric methods with machine learning techniques improve both structural interpretation and the early identification of high-risk insolvency episodes?

[1] yzhang2@upv.edu.es

The contribution of the paper is threefold. First, it provides new empirical evidence on regional and sectoral heterogeneity in corporate insolvency in Spain over the period 2008–2024, using a harmonized panel linking insolvency outcomes with macro-financial conditions. Second, it proposes an integrated methodological framework that combines two-way fixed-effects models with Driscoll–Kraay inference and supervised learning algorithms, allowing structural interpretation alongside predictive flexibility. Third, the results yield policy-relevant insights by identifying territories and sectors systematically more exposed to macro-financial stress, contributing to the development of targeted policies and early-warning systems.

Spain offers a particularly informative setting for analysing these dynamics. Its business structure is dominated by micro and small firms, which tend to be more fragile and heavily dependent on bank financing, while sectors such as construction, tourism, hospitality, and retail trade exhibit strong cyclical sensitivity. After reaching historical peaks during the global financial crisis, insolvency filings declined following legal reforms and fell to exceptionally low levels in 2020 due to pandemic-related moratoria. Once these measures were withdrawn, insolvency filings rose sharply, returning to levels close to those observed prior to COVID-19 [7], [8], [9], [10], [11]. Recent assessments further indicate that the post-pandemic increase in Spain has been more pronounced than in several other European economies [12].

In this study, insolvency is measured through formal entries into bankruptcy proceedings recorded in official statistics. This definition captures legally observable events and reflects both underlying economic vulnerability and institutional thresholds, a distinction that is central for interpreting the empirical results.

2 Methodology

The empirical strategy pursues two complementary objectives. First, it seeks to characterise the intensity of aggregate macro-financial stress affecting the Spanish economy over time. Second, it evaluates how this common disturbance is transmitted heterogeneously across regions and sectors in terms of corporate insolvency outcomes. Rather than aiming at strict causal identification, the analysis focuses on relative exposure and differential sensitivity to shared macro-financial conditions. To this end, a structural panel framework is combined with non-linear extensions and an out-of-sample predictive validation layer.

The analysis relies on an annual panel defined at the region–sector level $g = (r, s)$ covering the period 2008–2024. The dependent variable is the insolvency rate,

$$\mathrm{IR}_{g,t} = \frac{\text{Insolvent Firms}_{g,t}}{\text{Active Firms}_{g,t}},$$

which is transformed as

$$y_{g,t} = \log(1 + \mathrm{IR}_{g,t}),$$

to reduce skewness and stabilise variance. Estimation is conducted using weighted least squares, assigning each observation a weight proportional to the number of active firms within the corresponding region–sector group.

Aggregate financial stress is summarised through a *Market Risk Index* (MRI) constructed from monthly returns and volatilities of stock market indices, Spanish listed banks, and the euro–US dollar exchange rate. For each asset k and month m,

$$\mathrm{RV}_{k,m} = \text{return}_{k,m} \times \text{volatility}_{k,m}.$$

After standardising each series and averaging across assets, the monthly index is aggregated to annual frequency and normalised:

$$\text{Shock}_t = \frac{\mathrm{MRI}_t - \mu_{\mathrm{MRI}}}{\sigma_{\mathrm{MRI}}}.$$

In addition, standardized univariate shocks are constructed for individual macroeconomic and financial indicators. These variables are introduced separately to compare relative sensitivities and to attribute transmission channels.

Heterogeneous transmission of the aggregate shock is estimated using a two-way fixed-effects model:

$$y_{g,t} = \alpha_g + \tau_t + \beta_g \, \text{Shock}_t + \varepsilon_{g,t},$$

where α_g controls for time-invariant region–sector heterogeneity and τ_t absorbs common macroeconomic, regulatory, and policy factors. The coefficient β_g measures the relative sensitivity of group g to a one-standard-deviation increase in the aggregate shock. Statistical inference relies on Driscoll–Kraay standard errors, robust to heteroskedasticity, serial correlation, and cross-sectional dependence.

As an extension, dimensionality reduction is applied to the financial block using principal component analysis with Varimax rotation, yielding two orthogonal factors summarising market conditions. To capture potential thresholds and asymmetric responses, these factors are complemented with smooth non-linear transformations based on logistic fuzzy membership functions. Interactions with regional and sectoral identifiers allow heterogeneous response profiles to be characterised while preserving the interpretability of the fixed-effects framework.

The structural analysis is complemented by a supervised learning exercise serving as an external validation layer. Annual observations are classified according to upper quantiles of the insolvency distribution, using financial indicators, non-linear transformations, macroeconomic variables, simple dynamic terms, and regional and sectoral identifiers as predictors. Temporal validation follows an expanding-window scheme. Gradient boosting classifiers (CatBoost, XGBoost and LightGBM) are trained, and ensemble predictions are constructed using mean and median combinations. Model performance is evaluated through standard classification metrics together with a targeting exercise designed to assess the operational relevance of risk rankings.

Figure 1 summarises the integrated architecture of the empirical strategy.

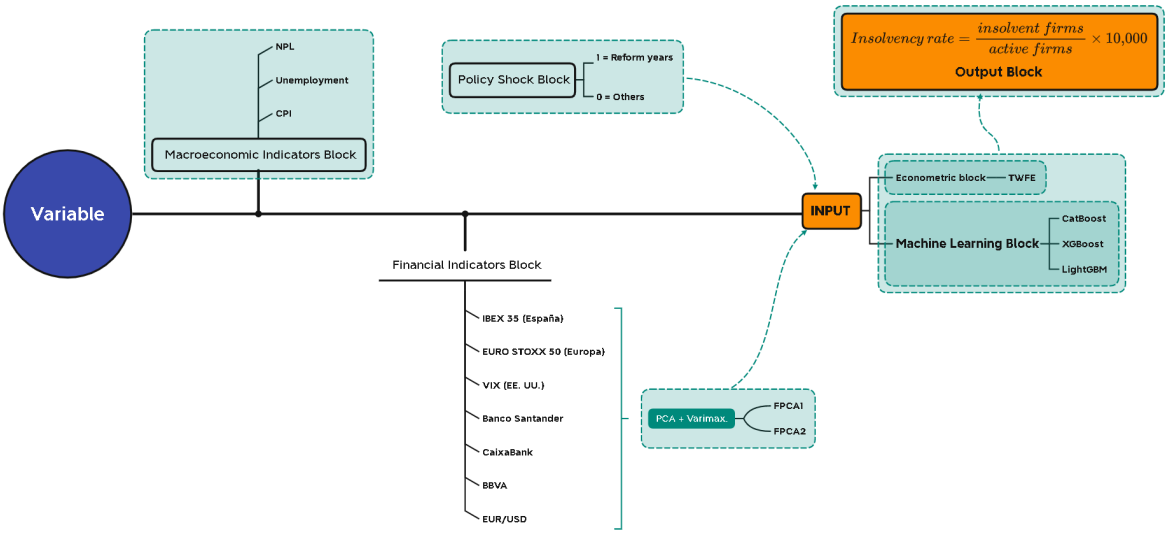


Figure 1: Integrated Modelling Framework

3 Results

3.1 Sectoral and Regional Heterogeneity under Multiple Shocks

Benchmark: heterogeneity under an aggregated MRI shock (TWFE–DK).

Table 1: Significant differential sensitivities (MRI; TWFE with DK).

Scope	Group	β	DK-p	95% CI (low)	95% CI (high)
Sectors (base = Admin. & Support)					
Sector	Real estate, financial and insurance activities	-0.115	0.0017	-0.187	-0.044
Sector	Construction	-0.066	0.0515	-0.132	0.000
Sector	Information and communication	0.075	0.0697	-0.006	0.157
Regions (base = Andalusia)					
Region	Madrid	0.066	¡0.001	0.043	0.089
Region	Valencian Community	0.074	¡0.001	0.041	0.106
Region	Cantabria	0.095	0.0002	0.046	0.145
Region	Aragon	0.088	0.0071	0.024	0.152
Region	Canary Islands	0.073	0.0302	0.007	0.139
Region	Region of Murcia	0.076	0.0106	0.018	0.135
Region	Catalonia	0.065	0.0390	0.003	0.127
Region	Asturias	0.084	0.0520	-0.001	0.168
Region	Balearic Islands	-0.111	0.0014	-0.180	-0.043

Table 1 reports a benchmark based on a single aggregated MRI shock. Coefficients are semi-elasticities relative to the baseline sector (Administrative and Support Services) and region (Andalusia). A homogeneity test rejects equal slopes ($p < 0.001$), but fit is modest (within-R^2: sectors 0.0128; regions 0.0444), motivating a decomposition into multiple orthogonal shocks.

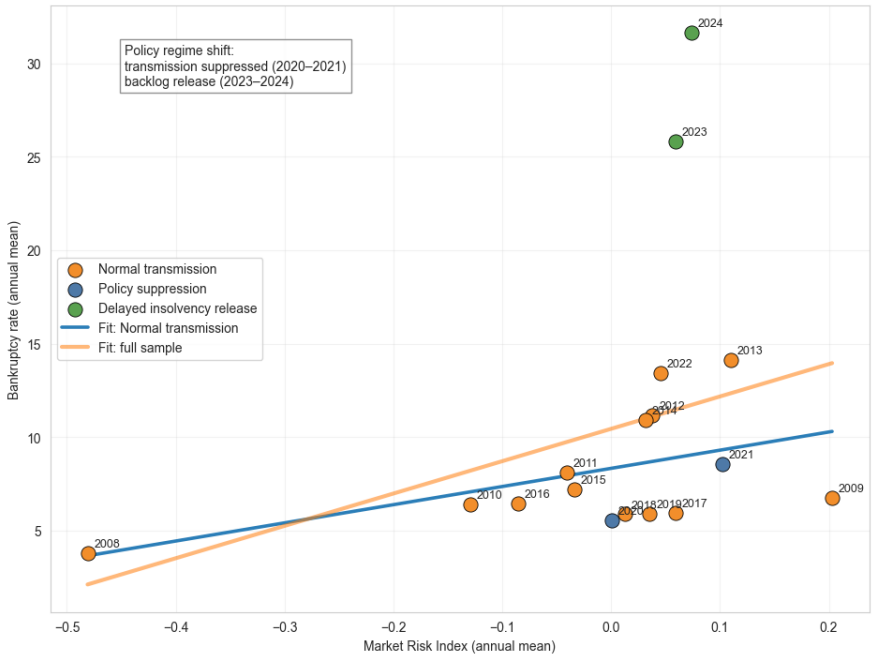
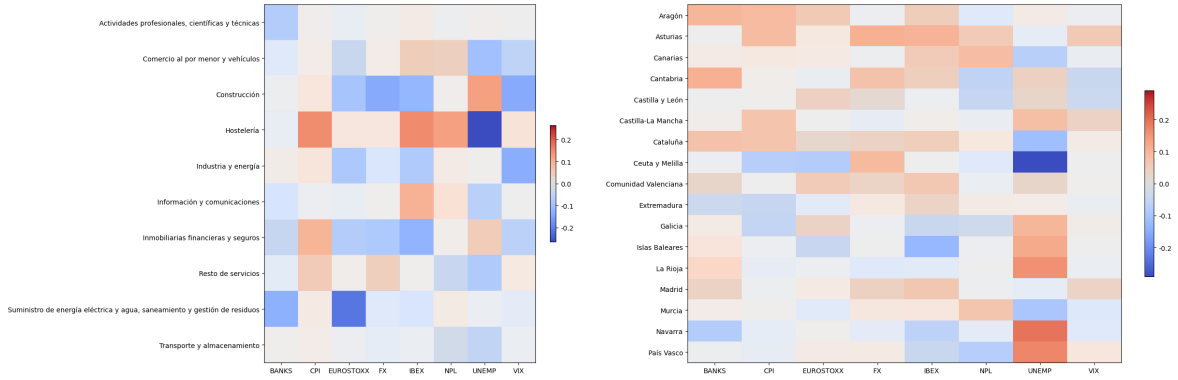


Figure 2: Financial Stress–Insolvency Relationship with Policy-Induced Regime Shifts



(a) Sector × shock (cells with $p > 0.10$ faded)

(b) Region × shock (cells with $p > 0.10$ faded)

Figure 3: Standardised sensitivities β to orthogonalised shocks (order: VIX, IBEX, BANKS, EUROSTOXX, FX, UNEMP, CPI, NPL). TWFE with entity and year fixed effects; Driscoll–Kraay errors (Parzen, bw=2). Cells with $p > 0.10$ are displayed with reduced intensity.

Figure 2 indicates that the stress–insolvency link is regime-dependent. During COVID–19 (2020–2021), extraordinary interventions (moratoria and liquidity support) weakened contemporaneous transmission despite elevated stress. In 2023–2024, insolvencies rise sharply without a proportional increase in financial stress, consistent with delayed materialisation after policy withdrawal. These patterns suggest a nonlinear, state-contingent mapping from financial stress to firm failure.

Main specification: multivariate orthogonalised shocks (TWFE–DK).

Table 2: Model specifications and within-R^2 comparison

Specification	Scope	Within R^2
Single aggregated shock × Sector	Sectors	0.0128
Multiple orthogonalised shocks × Sector	Sectors	**0.1661**
Single aggregated shock × Region	Regions	0.0444
Multiple orthogonalised shocks × Region	Regions	**0.0631**

Table 2 shows that orthogonalising stress into multiple shocks materially improves fit, especially for sectoral heterogeneity (within-R^2 from 0.013 to 0.166), consistent with differentiated transmission across industries.

Figures 3a–3b summarise heterogeneous slopes by shock. Financial shocks (VIX, IBEX, BANKS, EUROSTOXX, FX) drive the strongest sectoral dispersion, while macro shocks (UNEMP, CPI, NPL) add further differentiation linked to labour-market and credit conditions. Regionally, economically larger regions tend to exhibit stronger responses across equity, banking and macro shocks than lower-density areas.

Overall, the evidence points to a nonlinear and state-dependent transmission of financial stress, while policy effects are absorbed by year fixed effects and examined separately in robustness and machine-learning analyses.

Table 3: Rotated Loadings for Financial PCA (Varimax)

	FINPC1	FINPC2
vix_rv	-0.465	0.462
ibex35_rv	0.522	0.147
stoxx50e_rv	0.488	0.249
eurusd_rv	0.005	-0.837
bank_rv	0.522	0.040

3.2 Detection of Anomalous Bankruptcy Episodes

A principal component analysis (PCA) is applied to VIX, IBEX35, EUROSTOXX50, EUR/USD, and banking volatility (`bank_rv`), followed by Varimax rotation (Table 3). Two orthogonal factors explain about 88% of total variance. FINPC1 loads positively on IBEX, EUROSTOXX and `bank_rv`, and negatively on VIX, capturing domestic equity–banking pressure relative to global risk aversion. FINPC2 loads strongly on EUR/USD (negative) and on VIX (positive), forming a VIX–FX axis associated with global/exchange-rate shocks. These factors are used as orthogonal shocks in the panel estimations and in non-linear (fuzzy) experiments for extreme episodes.

Table 4 documents heterogeneous transmission of financial stress across sectors and regions under TWFE–DK estimation.

For the dominant linear component (FINPC1), most sectors exhibit significantly positive differential sensitivities, indicating that tighter financial conditions systematically increase insolvency exposure in economically cyclical activities such as construction, retail, hospitality and professional services. By contrast, electricity, water and waste management display a large negative coefficient, suggesting weaker alignment with market-driven stress, likely reflecting regulatory pricing structures and cost pass-through mechanisms.

The second component (FINPC2) plays a more limited role under normal conditions. Only a small subset of sectors shows statistically relevant responses, implying that this dimension captures a secondary channel whose average linear impact is less pervasive across industries.

Nonlinear effects emerge once attention shifts to extreme-stress regimes. The interaction FINPC2Mu×Region highlights significant responses for Madrid and Aragón, with additional suggestive evidence for several interior regions. Appendix results show that directional decompositions reverse sign across regimes, indicating threshold-dependent transmission: once latent financial stress exceeds critical levels, the relationship between global–FX shocks and insolvency outcomes changes qualitatively.

Taken together, the results support a state-dependent view of insolvency dynamics. Sectoral heterogeneity is primarily governed by the main financial component, whereas regional differentiation becomes visible mainly during extreme financial conditions.

Table 4: Sectoral and Regional Heterogeneity in Financial Stress Transmission

Group	β	SE	p	CI Low	CI High	Sig.
Sector heterogeneity — FINPC1 (linear, 1σ)						
Actividades profesionales, científicas y técnicas	1.615	0.638	0.011	0.364	2.866	*
Construcción	1.839	0.759	0.015	0.352	3.326	*
Información y comunicaciones	1.443	0.615	0.019	0.238	2.648	*

Table 4: Sectoral and Regional Heterogeneity in Financial Stress Transmission (continued)

Group	β	SE	p	CI Low	CI High	Sig.
Inmobiliarias financieras y seguros	1.298	0.597	0.030	0.128	2.468	*
Comercio al por menor y vehículos	1.602	0.745	0.032	0.142	3.063	*
Hostelería	1.367	0.638	0.032	0.117	2.617	*
Suministro de energía eléctrica y agua, saneamiento y gestión de residuos	-14.679	6.964	0.035	-28.328	-1.030	*
Actividades administrativas y servicios auxiliares	1.393	0.756	0.065	-0.089	2.875	†
Resto de servicios	1.719	0.936	0.066	-0.116	3.554	†
Industria y energía	1.469	0.811	0.070	-0.121	3.059	†
Sector heterogeneity — FINPC2 (linear, 1σ)						
Resto de servicios	1.246	0.619	0.044	0.033	2.458	*
Construcción	0.938	0.479	0.050	-0.002	1.877	†
Información y comunicaciones	0.990	0.593	0.095	-0.173	2.153	†
Regional heterogeneity — FINPC2Mu × Region (nonlinear channel)						
Madrid	-401.233	163.712	0.014	-722.110	-80.357	*
Aragón	-368.270	172.725	0.033	-706.811	-29.729	*
Comunidad Valenciana	-329.657	187.151	0.078	-696.473	37.160	†
Castilla-La Mancha	-317.053	184.480	0.086	-678.633	44.527	†
La Rioja	-304.921	181.756	0.094	-661.162	51.320	†
Asturias	-282.183	169.792	0.097	-614.977	50.610	†

Notes: Coefficients correspond to differential slopes relative to the effect-coded reference group. Significance levels: * $p < 0.05$, † $p < 0.10$.

Table 5: Sectoral Heterogeneity in OOS Predictive Performance (Top-10 by AUC)

		CatBoost		XGBoost (monotone)		LightGBM			
Rk	Sector	AUC	AP	AUC	AP	AUC	AP	n	Pos.%
1	Transport and storage	0.946	0.518	0.686	0.343	0.799	0.204	108	5.556
2	Retail trade and vehicle repair	0.794	0.587	0.853	0.649	0.785	0.519	108	21.296
3	Industry and energy	0.785	0.883	0.840	0.932	0.558	0.786	108	69.444
4	Construction	0.743	0.446	0.783	0.465	0.679	0.348	108	20.370
5	Hospitality	0.727	0.560	0.728	0.603	0.747	0.623	108	29.630
6	Electricity, water and waste management	0.718	0.391	0.579	0.315	0.627	0.305	108	20.370
7	Administrative and support services	0.608	0.282	0.718	0.221	0.774	0.499	108	10.185
8	Information and communication	0.581	0.512	0.587	0.536	0.599	0.556	108	44.444
9	Professional, scientific and technical activities	0.121	0.011	0.411	0.016	0.818	0.048	108	0.926
10	Real estate, financial and insurance activities	—	—	—	—	—	—	108	0.000

Notes: AUC and Average Precision (AP) are computed from rolling out-of-sample predictions. "—" indicates undefined metrics due to absence of positive observations in OOS validation (Pos.%=0).

Based on Table 5, four operational findings emerge:

- **Consistency among leading sectors.** Across CatBoost and XGB models, the highest discrimination is consistently observed in *Retail trade and vehicle repair*, *Industry and energy*, *Construction*, and *Hospitality*. The financial signal captured by FINPC1, together with nonlinear effects from FINPC2, appears particularly informative in cyclical and credit-intensive activities.
- **Transport and storage** achieves the highest AUC under CatBoost (≈ 0.95), although with very low prevalence (pos% $\approx 5.6\%$). In such rare-event settings, ROC-based metrics may be optimistic; Average Precision and Recall@TopK provide more operational measures of screening performance.
- **Industry and energy** combines strong ROC discrimination with high Average Precision, indicating effective ranking of high-risk observations when the positive class is sufficiently represented.
- **Professional, scientific and technical activities** exhibit extremely low prevalence ($\approx 0.9\%$), explaining the unstable AUC and low AP values. Under rare-event conditions, PR-based metrics are more informative than ROC measures.

Special case (undefined metrics). For *Real estate, financial and insurance activities*, AUC and AP are undefined because no observations fall within the top 20% of the national target distribution during OOS validation (pos%=0). Consequently, ROC and PR curves cannot be computed. Substantively, this indicates that the sector does not concentrate extreme outcome episodes in relative country–year terms.

3.3 Fuzzy Segmentation through Response Profiles

Sectors and regions are grouped using response profiles built from: (i) group-level moments of y (mean, dispersion, trend), (ii) correlations $\mathrm{corr}(y, \cdot)$ with FINPCs and macro/policy variables, and (iii) out-of-sample classifier metrics by group (AUC, AP, and mean predicted probabilities from CatBoost, XGBoost and LightGBM). After standardisation, Fuzzy C-Means is applied with automatic K selection using the silhouette criterion (ties broken by Calinski–Harabasz) and a fixed seed. PCA is used only for visualisation.

Table 6: Cluster composition by sector and region (FCM, $K = 3$)

Panel A: Sectors	
Cluster	Sectors
0	Construction; Industry and Energy; Information and Communications
1	Administrative and Support Services; Professional, Scientific and Technical Activities; Real Estate, Financial and Insurance Activities; Electricity, Water Supply, Sanitation and Waste Management; Transport and Storage
2	Retail Trade and Vehicles; Hospitality; Other Services
Panel B: Regions	
Cluster	Regions
0	Cantabria; Castile and León; Castile–La Mancha; Galicia; Navarra; Basque Country
1	Andalusia; Canary Islands; Ceuta and Melilla; Extremadura; Balearic Islands; La Rioja
2	Aragon; Asturias; Catalonia; Valencian Community; Madrid; Murcia

With $m = 2.0$, the algorithm selects $K = 3$ for both sectors and regions (Figures 4–5; Table 6). Results are stable when varying m (e.g., 1.6 and 2.4): the selected K, cluster composition, and dominant features of cluster centres remain essentially unchanged. The membership matrix shows little overlap, implying that the partition is nearly crisp in practice; membership values can nonetheless be retained as an uncertainty measure for borderline cases.

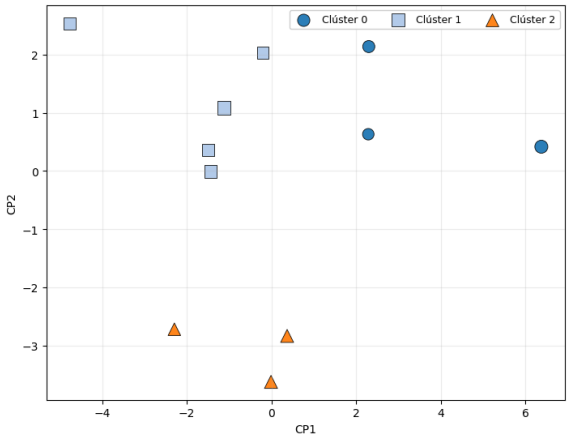


Figure 4: PCA projection of sectoral response profiles (FCM).

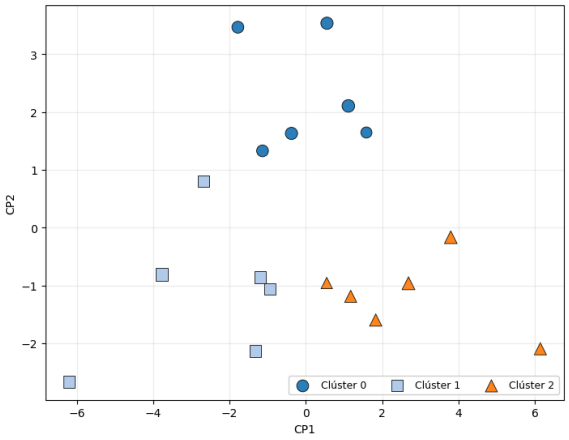


Figure 5: PCA projection of regional response profiles (FCM).

4 Conclusions

This paper analyses the macro-financial determinants of regional corporate insolvency in Spain during 2008–2024. Instead of firm-level characteristics, the study focuses on how aggregate financial conditions are transmitted heterogeneously across regions and sectors. The evidence indicates that insolvency dynamics are driven by a common macro-financial environment whose effects depend on local economic structures.

A benchmark specification based on an aggregated market risk indicator confirms the presence of systemic disturbances, while a decomposition into orthogonal shocks substantially improves explanatory power, particularly at the sectoral level. This suggests that heterogeneity arises mainly from differentiated transmission channels rather than from variations in overall stress intensity.

PCA–Varimax results reveal a two-dimensional structure of financial stress. The first component (FINPC1) captures domestic equity and banking pressures and constitutes the dominant transmission channel, whereas the second component (FINPC2), linked to global uncertainty and exchange-rate risk, operates primarily through asymmetric and nonlinear regional responses. Together, these dimensions highlight the multidimensional financial nature of macroeconomic fragility in Spain.

Machine-learning validation supports the robustness of these findings, showing stable predictive performance across models and validation windows and suggesting that the identified relationships reflect persistent economic mechanisms.

Limitations include the absence of firm-level information, the reduced-form nature of the framework, and the partial treatment of spatial spillovers and policy interactions. Future research should integrate spatial and multi-level approaches incorporating sectoral composition and credit-market structures.

Overall, regional corporate insolvency emerges as a multidimensional macro-financial transmission process, linking systemic financial stress, heterogeneous regional exposure, and predictive early-warning modelling.

Acknowledgments

The author would like to thank the supervisors and colleagues at the Universitat Politècnica de València for valuable discussions and constructive feedback throughout the development of this

research. All remaining errors are the author's responsibility.

References

[1] W. Cui and L. Kaas, "Default cycles," *Journal of Monetary Economics*, vol. 117, pp. 377–394, 2021.

[2] S. Hol, "The influence of the business cycle on bankruptcy probability," *International transactions in operational research*, vol. 14, no. 1, pp. 75–90, 2007.

[3] K. Giesecke, F. A. Longstaff, S. Schaefer, and I. Strebulaev, "Corporate bond default risk: A 150-year perspective," *Journal of financial Economics*, vol. 102, no. 2, pp. 233–250, 2011.

[4] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *Journal of Finance*, vol. 23, no. 4, pp. 589–609, 1968.

[5] W. H. Beaver, "Financial ratios as predictors of failure," *Journal of Accounting Research*, vol. 4, no. 3, pp. 71–111, 1966.

[6] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of Accounting Research*, vol. 18, no. 1, pp. 109–131, 1980.

[7] B. M. y. M. d. E. Colegio de Registradores de la Propiedad, *Anuario concursal 2013*, Disponible en: `https://www.registradores.org`, 2014.

[8] I. B. Aldana, P. A. Alonso, and J. S. Espuch, *The spanish insolvency reform: A focus on the new pre-insolvency system*, Disponible en: `https://www.chasecambria.com`, 2023.

[9] B. M. y. M. d. E. Colegio de Registradores de la Propiedad, *Anuario concursal 2018*, Disponible en: `https://www.registradores.org`, 2019.

[10] B. M. y. M. d. E. Colegio de Registradores de la Propiedad, *Anuario concursal 2020*, Disponible en: `https://www.registradores.org`, 2021.

[11] B. M. y. M. d. E. Colegio de Registradores de la Propiedad, *Anuario concursal 2021*, Disponible en: `https://www.registradores.org`, 2022.

[12] A. de Pastors, V. Kouro, J. L. Luna Alpizar, and S. Otero, *Moody's edfx foresaw 82% of spanish sme bankruptcies*, `https://edfx.com/wp-content/uploads/2024/11/Spanish-SME-bankruptcies.pdf`, This article is a product of Moody's Analytics, a division of Moody's Corp. separate from Moody's Ratings. The analysis and viewpoints expressed herein are solely those of Moody's Analytics. Data Source: Eurostat, Nov. 2024.

Validated Cross-Ventilation Modeling of an occupied Indoor Environment Reveals Uneven Air Renewal After Window Opening

O.A. Shatarat[♭],[1] A. D'Ausilio[♮], G. Aiello[♭], S. La Grutta[♢], V. Lazzara[♢] and V.Malizia[♢]

(♭) Department of Engineering, University of Palermo; Viale delle Scienze, 90128, Italy

(♮) Suez ARIANET srl, 20159 Milano, Via Benigno Crespi 52, Italy

(♢) Consiglio Nazionale delle Ricerche Area Territoriale di Ricerca di Palermo, Via Ugo La Malfa n. 153, Italy.

1 Abstract

Ventilation in dense indoor environments is often assessed using CO_2 concentration value, although this may conceal local weak renewal zones. This study uses a validated CFD model to assess cross ventilation in a real dense indoor environment in Palermo, Italy, through surface based CO_2 and Age of Air indicators. A transient CO_2 decay experiment, initiated by opening two windows and one door, was reproduced in `OpenFOAM v2312` using a URANS k-ε formulation with buoyancy effects. The model reproduced the measured decay with acceptable agreement, with MBE = -19.7 ppm and nRMSE = 8.6%. The results show that cross ventilation improved indoor air renewal, but unevenly. The strongest contrast was vertical, with faster renewal and younger air in the lower region than in the upper region. The study shows that validated CFD can reveal spatial ventilation deficiencies that may be missed by single-point or room average CO_2 measurements.

2 Introduction

Indoor air quality in densely occupied spaces is a relevant engineering concern because ventilation controls the renewal of indoor air and the dilution of occupant generated pollutants. Inadequate ventilation can lead to the accumulation of re-breathed air and poorer indoor conditions over prolonged exposure periods. Within this context, carbon dioxide (CO_2), although not an overall indicator of indoor air quality, is widely used as a practical proxy for ventilation performance when interpreted together with occupancy and air-distribution conditions [1], [2]. In such environments, the CO_2 concentration is commonly evaluated against thresholds around 1200 ppm, above which ventilation is generally considered insufficient [3].

In practice, however, ventilation is often assessed through room average CO_2 values or measurements taken at a single sensor location. This provides only a bulk description of the indoor environment and may not represent how ventilation quality varies across the room. ASHRAE notes that the interpretation of indoor CO_2 measurements must account for both spatial and temporal variability and that sensor location is critical [2]. Recent studies on naturally ventilated indoor

[1] omarahmadmohd.shatarat@unipa.it

environments support this concern, showing that indoor air distribution can be highly non-uniform, with local stagnation regions and marked differences in ventilation conditions across the occupied zone [4], [5], [6], [7]. As a result, bulk measurements alone may mask regions that remain weakly renewed even when the overall room condition appears acceptable.

Computational fluid dynamics (CFD) is well suited to address this limitation because it can resolve airflow and pollutant transport throughout the room, rather than only at isolated measurement points. When validated against experimental data, CFD can provide spatially distributed information on ventilation behavior with useful fidelity [8], [9]. This is particularly important in naturally ventilated indoor spaces, where airflow structure, buoyancy effects, and opening configuration can produce uneven mixing and localized differences in air renewal. In this context, the Age of Air (AoA) is a useful complementary indicator because it describes how long air remains in the space and helps distinguish well renewed from weakly renewed regions [10], [11].

Against this background, the present study uses a CFD model validated with measurements to evaluate cross ventilation in a real densely environment. The study compares stale air and open window conditions through surface-based CO_2 and Age-of-Air indicators in order to quantify spatial heterogeneity in ventilation quality and identify regions that remain weakly renewed after opening. The paper therefore addresses a specific gap in ventilation assessment; bulk CO_2 measurements can indicate whether ventilation is generally adequate, but they do not show how ventilation performance is distributed across the room.

3 Materials and Methods

The study was conducted on a real, densely occupied indoor environment in Palermo, Italy, within the PNR MISSION project. The site was located in a dense urban area influenced mainly by local traffic and nearby residential activity. The investigated space had a volume of about 200 m^3 and dimensions of approximately 7.6 m $\times$ 6.0 m $\times$ 4.4 m, with two windows and one door available for natural ventilation. The occupied configuration was kept fixed throughout the campaign. The monitored case included 24 seated occupants plus one instructor, representing a real use condition, with activity levels consistent with about 1.2 MET. Environmental monitoring included CO_2, temperature, relative humidity, and particulate matter, measured using a GrayWolf DirectSense II while outdoor meteorological data were collected to define the external boundary conditions of the CFD model. The wind velocity recorded at a nearby meteorological station was used as input for a preliminary analysis to estimate the flow conditions at the window openings under the imposed atmospheric boundary layer (ABL). The results indicate a representative normal velocity of 0.15 $m\,s^{-1}$ at the windows. The main indoor sensor was placed near the middle of the room at 1.2 m height, corresponding to breathing level in the occupied zone, and its measurements were used as the reference for validation of the cross ventilation case.

The analysis presented in this paper focuses on a single validated cross ventilation case under real occupied conditions. To generate a reproducible transient decay, the indoor environment was first kept closed, with all windows and the door shut, so that CO_2 released by the occupants could accumulate indoors. Once the indoor concentration reached the target initial level used for validation, the two windows and the door were opened simultaneously, initiating cross ventilation and allowing dilution with outdoor air. The resulting transient CO_2 decay was recorded at the central monitoring point and used as the reference dataset for comparison with the CFD results. For this specific case, the initial indoor CO_2 concentration was set at about 1562 ppm before the openings were activated.

These experimental conditions were then reproduced in the CFD simulations, enabling direct comparison between measured and modeled pollutant dynamics. The simulations were performed with `OpenFOAM v2312` using the unsteady Reynolds-averaged Navier–Stokes (URANS) formulation

solved with `buoyantBoussinesqPimpleFoam` Buoyancy was modeled with the Boussinesq approximation, appropriate for small indoor temperature differences. Turbulence was represented with the the k-ε model. The governing equations included incompressible continuity, momentum, energy, and passive scalar transport for CO_2 and AoA.

The governing equations for the airflow field are reported below. The conservation of mass is given by

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}) = 0 \tag{1}$$

where ρ is the air density and $\boldsymbol{u}$ is the velocity vector.

The momentum equation is solved in the following form:

$$\frac{\partial (\rho \boldsymbol{u})}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}\boldsymbol{u}) = -\nabla p + \rho \boldsymbol{g} + \nabla \cdot (2\mu_{\text{eff}} \boldsymbol{D}(\boldsymbol{u})) - \nabla \left(\frac{2}{3} \mu_{\text{eff}} \nabla \cdot \boldsymbol{u} \right) \tag{2}$$

where p is the static pressure, $\boldsymbol{g}$ is the gravitational acceleration, μ_{eff} is the effective viscosity, and $\boldsymbol{D}(\boldsymbol{u})$ is the strain-rate tensor.

To reproduce the transient dilution of occupant-generated pollutants, CO_2 is treated as a passive scalar governed by a convection–diffusion equation:

$$\frac{\partial C}{\partial t} + \nabla \cdot (\boldsymbol{u} C) = \nabla \cdot (D_{\text{eff}} \nabla C) + S_C \tag{3}$$

where C is the CO_2 concentration, D_{eff} is the effective diffusivity, and S_C is the source term. In the present case, CO_2 is introduced only through the breathing patches and the openings, while no volumetric source term is imposed within the domain.

In addition, the age of air (AoA) is solved in parallel to complement the CO_2 analysis and to identify regions of delayed air renewal.

Breathing was modelled by imposing a sinusoidal velocity boundary condition at the mouth patches:

$$u_n(t) = U_{\max} \sin(\omega t) \tag{4}$$

where $u_n(t)$ is the velocity normal to the mouth surface, $U_{\max} = 0.4\,\mathrm{m\,s^{-1}}$ is the velocity amplitude, and $\omega = 1.45\,\mathrm{s^{-1}}$ is the angular frequency, corresponding to a breathing period of approximately 4.3 s, or about 14 breaths $\mathrm{min^{-1}}$. The exhaled air temperature was fixed at 310 K, whereas the ambient air temperature was set to 293 K. The CO_2 concentration was prescribed as 27 200 ppm during exhalation and 800 ppm during inhalation.

The computational domain reproduced the real occupied indoor environment monitored during the campaign. The mesh was generated with `snappyHexMesh` and contained about 1.3 million cells, with local refinement near openings, occupants, and mouth patches. The validated transient simulation reproduced the cross ventilation decay after opening the windows and door from an initial indoor CO_2 concentration of about 1562 ppm. The main boundary and initial conditions applied in the model are summarized in Table 1

Model validation was carried out by comparing the simulated and measured CO_2 concentration at the central monitoring point during the transient cross ventilation decay after the openings were activated. The agreement between simulation and measurements was evaluated using MAPE (Mean Absolute Percentage Error), nRMSE (normalized root mean square error) and MBE (mean bias error). Validation is presented through one time series comparison and one compact table summarizing the error metrics.

After validation, the CFD results were analyzed using a surface based post processing method. Sets of planes normal to the x-, y-, and z-directions were used to sample the room and extract the fraction of area above 1200 ppm CO_2, while Age of Air (AoA) was evaluated as a complementary indicator of air renewal. The comparison at selected times of the transient decay was used to

Table 1: Boundary conditions adopted for the main patches and regions of the computational domain.

Patch / region	**Velocity condition**	**Pressure condition**	**Scalar / thermal note**
Windows	`flowRateInletVelocity`	`zeroGradient`	Outdoor air imposed
Door	`pressureInletOutletVelocity`	`fixedValue` (1×10^5 Pa)	Outlet and reference pressure
Mouth patches	Sinusoidal breathing velocity	`zeroGradient`	Transient exhalation/inhalation of CO_2
Solid surfaces	`noSlip`	`zeroGradient`	Impermeable walls
Internal field	Initialized from measured state	Initialized	CO_2 = 1562 ppm

identify directional gradients and delayed renewal zones. Where persistent trends were observed on the surfaces, candidate slab volumes were defined as a secondary step to confirm weakly renewed regions, avoiding a full 3D partition of the room.

4 Results

The reliability of the CFD model was first assessed by comparing the simulated CO_2 decay with the experimental measurements at the probe position. Figure 1 shows that the model reproduces the main behavior of the measured decay well. The agreement is good and the statistical indicators confirm this result, where the mean bias error (MBE), which shows whether the model tends to overestimate or underestimate the measurements, was -19.7 ppm, indicating a slight overall underestimation. The mean absolute error (MAE), which represents the average absolute difference between measured and simulated values, was 68.4 ppm, while the root mean square error (RMSE), which gives more weight to larger deviations, was 90.3 ppm. When normalized, the RMSE was 0.086, corresponding to about 8.6%, which is acceptable for this type of indoor CFD application. Taken together, these values show that the model captures the measured CO_2 evolution with good accuracy.

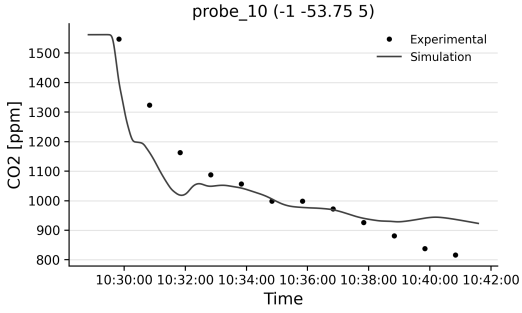


Figure 1: Measured vs simulated CO_2 decay during cross ventilation

On this basis, the model was considered sufficiently reliable for the following analysis of airflow structure and spatial ventilation heterogeneity. The remaining differences between measurements and simulation are limited and do not affect the use of the model as a tool to interpret the spatial behavior of the ventilated room.

To interpret the spatial distribution of ventilation, the airflow field was examined through the streamline pattern shown in Figure 2. The figure indicates that the incoming air enters mainly through the window openings and then penetrates the occupied zone rather than spreading uniformly across the room. As the fresh air moves inward, it is affected by the heat released by the

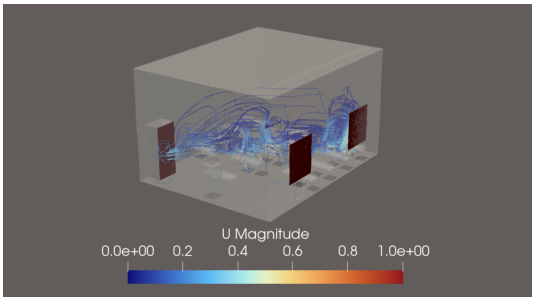
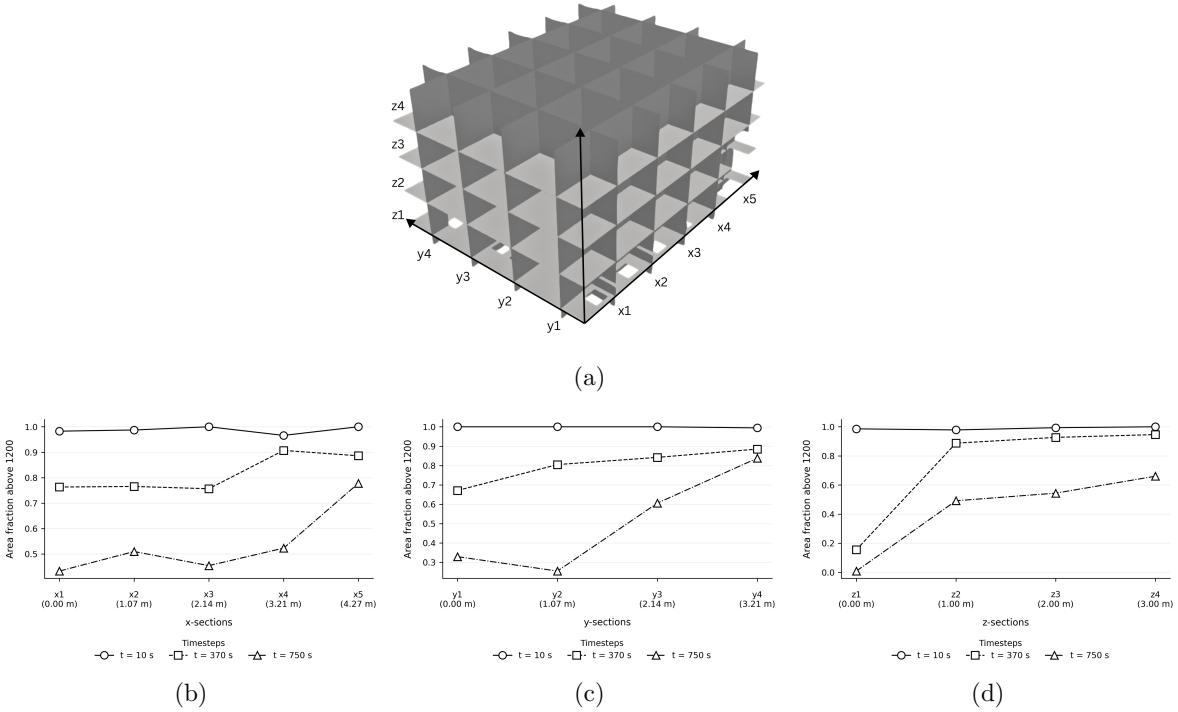


Figure 2: Streamlines of the airflow field under cross ventilation, colored by velocity magnitude.



(a) (b) (c) (d)

Figure 3: Surface based screening of ventilation heterogeneity. (a) Definition of the x-, y-, and z-sections used for the analysis. (b–d) Area fraction above 1200 ppm along the x-, y-, and z-directions at 10 s, 370 s, and 750 s.

occupants, which lifts part of the flow upward and promotes circulation in the upper part of the room. This produces a non-uniform ventilation pattern, with marked recirculation above the occupied zone and a more direct path from the window side toward the outlet. The result is that some regions are renewed rapidly, while others remain more weakly ventilated for longer times. This uneven flow structure is important for interpreting the scalar fields reported next, because it explains why the room does not behave as a perfectly mixed volume and why local differences in CO_2 concentration and age of air persist even under cross ventilation.

Surface oriented analysis was used to screen how the ventilation response varied across the room. The area fraction above 1200 ppm gave a clearer indication of where poorly renewed air persisted and allowed a direct comparison among the x-, y-, and z-directions. The resulting trends are shown in Figure 3.

Table 2: Volume-based comparison of the lower and upper slabs at the representative times.

Region	**Mean CO_2 (ppm)**	**Fraction $>$ 1200 ppm**	**Mean AoA (s)**	**CO_2 SD (ppm)**
t = 360 s				
Lower slab (floor–z2)	1148.3	0.420	219.6	267.5
Upper slab (z2–top)	1431.0	0.945	301.1	141.7
t = 750 s				
Lower slab (floor–z2)	1019.1	0.179	354.8	212.1
Upper slab (z2–top)	1266.2	0.647	495.3	158.6

The clearest separation appears in the vertical direction. At both representative times, 370 s and 750 s, the lowest level (z1) remained markedly less affected than the sections above it, while z2, z3, and z4 retained a much larger fraction of area above 1200 ppm. This indicates that the strongest spatial contrast is between the lower and upper parts of the room, rather than along a horizontal division of the plan. A secondary trend appears in the y-direction, where the far side of the room becomes progressively worse at later times, especially at 750 s. By contrast, the x-direction shows a weaker and less regular pattern. In all three directions, the rise observed toward the terminal sections is consistent with low velocity recirculation regions and weak renewal corners, where the local residence time increases and CO_2 removal becomes less effective.

The age-of-air results support the same interpretation. In the vertical direction, the lowest section consistently had the youngest air, with mean values close to 180 s at 370 s and about 295 s at 750 s, while the upper sections remained distinctly older, reaching about 285–310 s at 370 s and about 450–510 s at 750 s. The y-direction also showed a moderate increase toward the far side. Taken together, these results show that cross ventilation improves the room, but not uniformly; renewal is fastest in the lower zone, weaker in the upper zone, and locally poorer toward the far y-side.

Based on the surface evidence, the room was divided into two candidate slabs: a lower slab extending from the floor to z2, and an upper slab extending from z2 to the top of the room. This split follows the clearest directional result of the surface analysis, namely the persistent vertical separation between the lower and upper parts of the room. Table 2 compares the two slabs.

At both representative times, the lower slab performed better than the upper slab, with lower CO_2 levels, a smaller high CO_2 fraction, and younger air. Although both regions improved with time, the same vertical separation remained, confirming that air renewal was consistently more effective below z2 than above it. These results support the slab based interpretation and show that the main spatial contrast is vertical, while any refinement toward the far y-side remains secondary.

Therefore, the results confirm that cross ventilation improves indoor air quality, but also show that the improvement is spatially uneven, with a clear vertical contrast and secondary side differences that affect the local ventilation response.

5 Conclusion

The validated CFD model reproduced the measured CO_2 decay with acceptable agreement, which supports its use for spatial interpretation of the ventilation response. The main finding is that cross ventilation improved indoor air renewal, but the improvement was spatially uneven. Rather than behaving as a well-mixed volume, the classroom showed a persistent vertical separation, with the lower part renewing faster and the upper part remaining more weakly renewed over time. A secondary worsening trend was also observed toward the far y-side. These results indicate that

the classroom did not behave as a well mixed volume after opening, and that local weak-renewal regions persisted even when the overall room condition improved.

This interpretation is consistent with previous ventilation studies, which have shown that naturally ventilated rooms may present strong spatial variability and that single-point CO_2 measurements do not always represent the full indoor distribution. In this sense, the present work adds a validated case in which that heterogeneity is quantified through surface-based CO_2 and Age-of-Air indicators, with a slab-based interpretation guided by the strongest directional evidence. The main limitation is that the analysis refers to one occupied room and one validated cross-ventilation case, with simplified boundary conditions and one main validation point. Even with such limitations, the study provides a useful basis for future digital-twin applications in which sensor inputs, occupancy, and outdoor conditions can be linked to a predictive loop between monitoring and control, supporting ventilation decisions and other managerial actions before indoor conditions deteriorate.

Acknowledgments

This research was conducted as part of the PNR MISSION project (Monitoraggio abbattimento rischi sanitari inquinamento indoor, PREV-A-2022-12377010). The authors acknowledge ISCRA for granting access to computational resources on the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking and hosted by CINECA in Italy.

References

[1] World Health Organization, "Roadmap to improve and ensure good indoor ventilation in the context of covid-19," World Health Organization, Technical report, 2021.

[2] ASHRAE, *Ashrae position document on indoor carbon dioxide*, Approved 2022; revised 2025, 2025.

[3] E. Ding, D. Zhang, and P. M. Bluyssen, "Ventilation regimes of school classrooms against airborne transmission of infectious respiratory droplets: A review," *Building and Environment*, vol. 207, p. 108 484, 2022. DOI: 10.1016/j.buildenv.2021.108484.

[4] C. V. M. Vouriot, M. van Reeuwijk, and H. C. Burridge, "Robustness of point measurements of carbon dioxide concentration for the inference of ventilation rates in a wintertime classroom," *Indoor Environments*, vol. 1, p. 100 004, 2024. DOI: 10.1016/j.indenv.2024.100004.

[5] D. Zhang, E. Ding, and P. M. Bluyssen, "Guidance to assess ventilation performance of a classroom based on co2 monitoring," *Indoor and Built Environment*, vol. 31, no. 4, pp. 1107–1126, 2022. DOI: 10.1177/1420326X211058743.

[6] N. Mahyuddin and E. A. Essah, "Spatial distribution of co2 impact on the indoor air quality of classrooms within a university," *Journal of Building Engineering*, vol. 89, p. 109 246, 2024. DOI: 10.1016/j.jobe.2024.109246.

[7] A. Muelas et al., "Analysis of different ventilation strategies and co2 distribution in a naturally ventilated classroom," *Atmospheric Environment*, vol. 283, p. 119 176, 2022. DOI: 10.1016/j.atmosenv.2022.119176.

[8] S. Gilani, H. Montazeri, and B. Blocken, "Cfd simulation of stratified indoor environment in displacement ventilation: Validation and sensitivity analysis," *Building and Environment*, vol. 95, pp. 299–313, 2016. DOI: 10.1016/j.buildenv.2015.09.010.

[9] A. Fikry and A. Elsayed, "Improving the indoor air quality (iaq) in naturally ventilated lecture hall with a single facade by solar chimneys," *Journal of Engineering and Applied Science*, vol. 68, p. 29, 2021. DOI: 10.1186/s44147-021-00027-7.

[10] S. F. Díaz-Calderón, J. A. Castillo, and G. Huelsz, "Natural cross-ventilation configurations: Comparison using different evaluation parameters," *Journal of Building Engineering*, vol. 98, p. 111 130, 2024. DOI: 10.1016/j.jobe.2024.111130.

[11] C. Buratti, R. Mariani, and E. Moretti, "Mean age of air in a naturally ventilated office: Experimental data and simulations," *Energy and Buildings*, vol. 43, pp. 2021–2027, 2011. DOI: 10.1016/j.enbuild.2011.04.015.

Part III

Recent advances in Iterative Processes for Solving Nonlinear Problems

Fractal methods: Dynamics of the fractal Newton method

Jessica Paredes-Morales[♭],[1] Alicia Cordero[♭] and Juan R. Torregrosa[♭]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

Iterative methods for solving nonlinear equations are studied from two complementary perspectives: the local convergence analysis [1], which describes the behavior of the sequence when the initial point is close to a root, and the dynamical analysis, which studies the iteration operator as a discrete system [2], [3]. In the case of the classical Newton method, this second perspective leads to the study of rational maps and their basins of attraction in the complex plane [4], [5], [6], [7].

In this work a fractal derivative, also known in the literature as the Hausdorff derivative, is used, following the formulation introduced by Chen in the context of anomalous diffusion [8] and subsequently developed within modeling methodologies with fractal and fractional derivatives [9]. It is important to distinguish this local derivative with respect to the fractal coordinate x^α from nonlocal fractional derivatives, although both families of operators appear related in models on non-Euclidean or fractal media [10].

Definition 1 (Domain and branch of powers)**.** *Let $0 < \alpha \leq 1$. In real dynamics, unless otherwise indicated, one works on an interval $I \subset (0, +\infty)$ so that x^α and $x^{\alpha-1}$ are well defined as regular real functions. In complex dynamics, the principal branch is fixed*

$$z^\alpha = e^{\left(\alpha \operatorname{Log} z\right)},$$

so the origin and the branch cut must be treated as singular points of the iteration.

On the other hand, fractal calculus makes it possible to reinterpret classical iterative methods for nonlinear equations through a nonstandard derivation, such as this derivative. This derivative, unlike the classical derivative, allows the derivative of nondifferentiable functions to be calculated. In the same way, it has served as a tool for constructing Newton-type schemes together with their local convergence analysis and numerical tests.

Definition 2 (Fractal derivative of order α)**.** *Let $f : I \subset (0, +\infty) \to \mathbb{R}$. The fractal derivative of f at $x_0 \in I$ is defined by*

$$D_F^\alpha f(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x^\alpha - x_0^\alpha}, \qquad 0 < \alpha \leq 1, \tag{1}$$

if the limit exists. If, in addition, the function $x \mapsto D_F^\alpha f(x)$ is continuous on I, then f is said to be α-differentiable or fractally differentiable on I.

[1]jessica.paredes@upv.es

In the same way, the fractal derivative has been defined in relation to the classical derivative. Therefore, when f is differentiable in the classical sense and $x > 0$, the following relation is obtained:

$$D_F^\alpha f(x) = \frac{f'(x)}{\alpha x^{\alpha-1}}, \qquad 0 < \alpha \leq 1, \tag{2}$$

which shows how the parameter α modifies the local scaling. For $\alpha = 1$, the fractal derivative coincides with the classical derivative.

From the first-order truncated fractal Taylor expansion,

$$f(x) \approx f(x_0) + D_F^\alpha f(x_0)(x^\alpha - x_0^\alpha), \tag{3}$$

the fractal Newton method is obtained.

Definition 3 (Fractal Newton method)**.** *Let $f : I \subset (0, +\infty) \to \mathbb{R}$ be an α-differentiable function and let $D_F^\alpha f(x_k) \neq 0$. The fractal Newton method is defined by*

$$x_{k+1} = \left(x_k^\alpha - \frac{f(x_k)}{D_F^\alpha f(x_k)} \right)^{1/\alpha}, \qquad k = 0, 1, \ldots. \tag{4}$$

If (2) is used, then

$$x_{k+1} = \left(x_k^\alpha - \alpha x_k^{\alpha-1} \frac{f(x_k)}{f'(x_k)} \right)^{1/\alpha},$$

provided that $x_k > 0$ and $f'(x_k) \neq 0$.

For $\alpha = 1$ the classical Newton method is recovered. For $0 < \alpha < 1$, the fractional power and the factor $x^{\alpha-1}$ introduce an additional dynamical dependence that requires the domain and the branch of the power to be fixed precisely.

2 Local convergence and numerical tests

Definition 4 (Order of convergence)**.** *Let $\{x_k\}$ be a sequence that converges to a root $\bar{x}$ and let $e_k = x_k - \bar{x}$. The sequence is said to converge with order $p \geq 1$ and asymptotic constant $C > 0$ if*

$$\lim_{k \to \infty} \frac{|e_{k+1}|}{|e_k|^p} = C, \qquad 0 < C < +\infty.$$

In particular, $p = 1$ corresponds to linear convergence and $p = 2$ to quadratic convergence.

Theorem 1 (Local convergence of the fractal Newton method [11])**.** *Let $f : [a, b] \subset [0, +\infty] \to [0, +\infty]$ be continuous on an interval containing a root $\bar{x}$ of f. Let $D_F^\alpha f(x)$ be the fractal derivative of f of order $\alpha \in (0, 1]$. If $D_F^\alpha f$ is continuous and nonzero at $\bar{x}$, and if x_0 is sufficiently close to $\bar{x}$, then method (4) converges with order two and its error equation is*

$$e_{k+1} = \alpha \bar{x}^{\alpha-1} C_2 e_k^2 + O(e_k^3), \qquad C_j = \frac{1}{j!} \frac{D_F^{\alpha(j)} f(\bar{x})}{D_F^\alpha f(\bar{x})}, \tag{5}$$

where $D_F^{\alpha(j)}$ denotes the successive application of the fractal derivative with respect to the coordinate x^α.

Remark 1. *The hypothesis $\bar{x} > 0$ is essential to use (2) without singularities. Therefore, experiments with roots at $\bar{x} = 0$ are not directly covered by the theorem for $0 < \alpha < 1$; in those cases, the degeneration of $D_F^\alpha f$ and the behavior of the chosen branch must be analyzed.*

The numerical tests were carried out in Matlab R2020a with double precision, stopping criterion

$$|f(x_{k+1})| + |x_{k+1} - x_k| < 10^{-8},$$

a maximum of 500 iterations, and order estimation by means of the approximate computational order of convergence, ACOC [12],

$$\rho = \frac{\ln\left(|x_{k+1} - x_k|/|x_k - x_{k-1}|\right)}{\ln\left(|x_k - x_{k-1}|/|x_{k-1} - x_{k-2}|\right)}. \tag{6}$$

Table 1 reports the results of the numerical tests performed for the fractal Newton method. In $f(x) = x^2 + x$, the root $\bar{x} = 0$ is simple for the classical Newton method, but it becomes degenerate for the fractal method because $D_F^\alpha f(x)$ tends to 0 when $x \to 0^+$ and $0 < \alpha < 1$. This explains why the observed ACOC is close to 1, although the theorem predicts order two only under regular hypotheses. In contrast, for $f(x) = (x-1)^3 - 1$ and real root $\bar{x} = 2$, order two is observed for all tested values of α.

Table 1: Representative numerical evidence for the fractal Newton method.

Function	x_0	Values of α	Iterations	ACOC
$x^2 + x$	0.5	$0.1, \ldots, 0.9$	19 to 9	1
$x^2 + x$	0.5	1	6	2
$x^{1/3}$	1	$0.1, 0.2, 0.5$	$16, 13, 40$	1
$x^{1/3}$	1	$0.6, 0.7, 1$	500	–
$(x-1)^3 - 1$	3	$0.1, 0.3, 0.5, 0.7, 0.9, 1$	7	2

Table 2: Results for $f(x) = x^{1/3}$ with $x_0 = 1$.

α	$\bar{x}$	$\lvert f(\bar{x}_{k+1})\rvert$	$\lvert\bar{x}_{k+1} - \bar{x}_k\rvert$	Iterations
0.1	1.6432×10^{-25}	5.4772×10^{-9}	5.6528×10^{-24}	16
0.2	1.3611×10^{-26}	2.3876×10^{-9}	1.3156×10^{-24}	13
0.3	2.1544×10^{-27}	1.2915×10^{-9}	4.6394×10^{-24}	8
0.4	3.4360×10^{-25}	7.0041×10^{-9}	1.9211×10^{-23}	14
0.5	8.2718×10^{-25}	9.3871×10^{-9}	2.4815×10^{-24}	40
0.6	-8.5904×10^{-25}	9.5061×10^{-9}	1.1046×10^{-24}	149
0.7	$-3.3183 \times 10^{29} + 1.5980 \times 10^{29}i$	7.1681×10^{9}	5.4004×10^{29}	500
0.8	$1.5323 \times 10^{77} + 2.1381 \times 10^{91}i$	2.7755×10^{30}	3.2845×10^{91}	500
0.9	$8.1552 \times 10^{127} + 6.8430 \times 10^{127}i$	4.7394×10^{42}	1.6319×10^{128}	500
1	$3.2734 \times 10^{150} + 7.7087 \times 10^{134}i$	1.4848×10^{50}	4.9101×10^{150}	500

In Table 2, for the equation $f(x) = x^{1/3}$, convergence toward the root $\bar{x} = 0$ is observed when α takes small values, but the ACOC remains close to 1. This is because this root does not satisfy the regular hypotheses of convergence theorem 1: the function is not fractally differentiable at 0 and, for $0 < \alpha < 1$, the fractal derivative becomes singular or degenerate in a neighborhood of the root. Therefore, the numerical estimate reflects effective linear convergence, not the quadratic convergence predicted for regular roots with $D_F^\alpha f(\bar{x}) \neq 0$. Starting from $\alpha = 0.7$, the iteration reaches the maximum number of iterations and moves away from the root, which shows loss of stability for this function that is not fractally differentiable at 0.

Table 3: Results for $f(x) = x^2 - 3x + 2$ with $x_0 = 3$.

α	$\bar{x}$	$\lvert f(\bar{x}_{k+1})\rvert$	$\lvert \bar{x}_{k+1} - \bar{x}_k\rvert$	Iterations
0.1	2	8.8818×10^{-16}	0	7
0.2	2	0	0	7
0.3	2	0	0	7
0.4	2	0	6.5977×10^{-9}	6
0.5	2	0	4.0165×10^{-9}	6
0.6	2	8.8818×10^{-16}	2.3915×10^{-9}	6
0.7	2	0	1.3906×10^{-9}	6
0.8	2	0	7.8826×10^{-10}	6
0.9	2	0	4.3477×10^{-10}	6
1	2	0	2.3283×10^{-10}	6

In Table 3, for the equation $f(x) = x^2 - 3x + 2$, it is observed that all values of α converge to the root $\bar{x} = 2$ with few iterations. In addition, the ACOC computed in these experiments remains close to 2 for all values of α, which numerically confirms the quadratic convergence predicted by the theorem when the root is regular and $D_F^{\alpha} f(\bar{x}) \neq 0$.

3 Dynamics of the method

The local convergence analysis describes what happens when the initial approximation is already sufficiently close to a root. However, to understand the global behavior of the method it is necessary to study the associated fixed-point operator. In this section the polynomial

$$p(x) = (x-1)(x-2) = x^2 - 3x + 2, \tag{7}$$

whose roots are $x = 1$ and $x = 2$, is considered. For $x > 0$, $x \neq \frac{3}{2}$, the fractal Newton method is represented by

$$N_{\alpha}(x) = \left(x^{\alpha} - \frac{\alpha x^{\alpha-1}(x-1)(x-2)}{2x-3} \right)^{1/\alpha}, \qquad 0 < \alpha \leq 1. \tag{8}$$

When $\alpha = 1$, operator (8) reduces to the classical Newton operator. For $0 < \alpha < 1$, the fractional powers modify the geometry of the basins of attraction and require distinguishing between the positive real branch and the principal complex branch.

Moreover, to carry out the study of the dynamics of the fractal Newton method, it is necessary to study the associated fixed-point operator. For example, to prove that the simple roots of $f(x) = 0$, $\bar{x}$, are critical points of the fixed-point operator for the Newton method

$$g(x) = x - \frac{f(x)}{f'(x)}, \tag{9}$$

the following is performed:

$$g'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}. \tag{10}$$

Then $g'(\bar{x}) = 0$ since $f(\bar{x}) = 0$ and $f'(\bar{x}) \neq 0$ because $\bar{x}$ is a simple root. In the same way, the following theorem is presented.

Theorem 2. *Let f be a fractally differentiable function and let $\bar{x}$ be a simple root of f, that is,*

$$f(\bar{x}) = 0, \qquad f'(\bar{x}) \neq 0.$$

Consider the fixed-point operator associated with the fractal Newton method

$$N_\alpha(x) = \left(x^\alpha - \frac{f(x)}{D_F^\alpha(f)(x)} \right)^{1/\alpha},$$

where $D_F^\alpha(f)(x)$ denotes the fractal derivative of order α of f. Then $\bar{x}$ is a fixed point and a critical point of N_α, that is,

$$N_\alpha(\bar{x}) = \bar{x} \qquad \textit{and} \qquad N'_\alpha(\bar{x}) = 0.$$

3.1 Fixed points and stability

Definition 5 (Fixed points, multiplier, and basins)**.** *A point x^* is fixed for N_α if $N_\alpha(x^*) = x^*$. If N_α is differentiable at x^*, the multiplier is $\lambda = (N_\alpha)'(x^*)$. The fixed point is attracting if $|\lambda| < 1$, repelling if $|\lambda| > 1$, neutral or parabolic if $|\lambda| = 1$, and superattracting if $\lambda = 0$. The basin of attraction of a root r is*

$$\mathcal{B}(r) = \{x_0 : \ N_\alpha^k(x_0) \to r \ \textit{when} \ k \to \infty\}.$$

The only regular solutions of $N_\alpha(x) = x$ on the positive real branch, excluding the pole $x = 3/2$ and the points where the fractional power does not define a regular iteration, are the roots $x = 1$ and $x = 2$. Moreover, since $p(1) = p(2) = 0$, both roots remain invariant under the iteration. At those roots one has

$$(N_\alpha)'(1) = (N_\alpha)'(2) = 0,$$

so they are superattracting fixed points of the considered branch. The value $x = 0$ must be treated as a singular point or branch point when $0 < \alpha < 1$, due to the factor $x^{\alpha-1}$, and not as a root or as a regular fixed point of the method.

To analyze the behavior at infinity, the local chart $u = 1/x$ is introduced and the conjugate operator is defined

$$\widetilde{N}_\alpha(u) = \frac{1}{N_\alpha(1/u)}. \tag{11}$$

Thus, infinity is studied as the point $u = 0$ of the conjugate map. In the considered real range, infinity should only be classified dynamically when it acts as a fixed point; otherwise, it is interpreted as nonconvergence or singularity.

3.2 Critical points

Critical points organize the dynamics because their orbits largely determine the shape of the basins of attraction [5], [6], [7]. Writing

$$N_\alpha(x) = G_\alpha(x)^{1/\alpha}, \qquad G_\alpha(x) = x^\alpha - \frac{\alpha x^{\alpha-1}(x-1)(x-2)}{2x-3},$$

one has

$$(N_\alpha)'(x) = \frac{1}{\alpha} G_\alpha(x)^{1/\alpha-1} G'_\alpha(x),$$

and, in the regular domain,

$$G'_\alpha(x) = -\frac{\alpha x^{\alpha-2}(x-1)(x-2)\,(2(\alpha-2)x + 3(1-\alpha))}{(2x-3)^2}. \tag{12}$$

Therefore, the critical points are separated into two types. First, the roots of the polynomial are critical points linked to the attractors:

$$r_1 = 1, \qquad r_2 = 2. \tag{13}$$

Second, three free critical points appear:

$$
\begin{aligned}
c_\alpha^{(1)} &= \frac{3(\alpha-1)}{2(\alpha-2)}, \\
c_\alpha^{(2)} &= \frac{3-2\sqrt{\frac{9-2\alpha+\alpha^2}{(\alpha-2)^2}}+\alpha\left(-3+\sqrt{\frac{9-2\alpha+\alpha^2}{(\alpha-2)^2}}\right)}{4-2\alpha}, \\
c_\alpha^{(3)} &= \frac{-3-2\sqrt{\frac{9-2\alpha+\alpha^2}{(\alpha-2)^2}}+\alpha\left(3+\sqrt{\frac{9-2\alpha+\alpha^2}{(\alpha-2)^2}}\right)}{2(\alpha-2)}, \qquad 0<\alpha<1.
\end{aligned} \tag{14}
$$

The points $r_1 = 1$ and $r_2 = 2$ are associated with the roots of the polynomial and are critical points linked to attractors. In contrast, the free critical points $c_\alpha^{(1)}$, $c_\alpha^{(2)}$, and $c_\alpha^{(3)}$ are important for constructing parameter lines: as α varies, their orbits allow convergence toward a root, falling into a singularity, or the appearance of nonconvergent regions to be detected. For $\alpha = 1$, the limiting value of $c_\alpha^{(1)}$ should not be interpreted as a free critical point of the classical Newton method, but rather as a point lying outside the regular domain used for $0 < \alpha < 1$.

3.3 Dynamical lines and parameter line

Definition 6 (Dynamical line, parameter line, and convergence plane)**.** *A dynamical line fixes the parameter α and represents the destination of the orbits as the initial point x_0 varies. A parameter line fixes an initialization rule depending on α, in this case one of the free critical points $c_\alpha^{(j)}$, and represents the destination of its orbit as the parameter varies. A convergence plane simultaneously encodes the destination of the orbits for pairs (x_0, α) [13], [14].*

To visualize the real dynamics, values of α are fixed and operator (8) is iterated for initial points $x_0 \in [0, 3]$. The colors of the dynamical line indicate the destination of each orbit: convergence to root 1, convergence to root 2, or singularity. Figure 1 shows the dynamical lines obtained for $\alpha = 0.1$, $\alpha = 0.5$, $\alpha = 0.7$, and $\alpha = 1$.

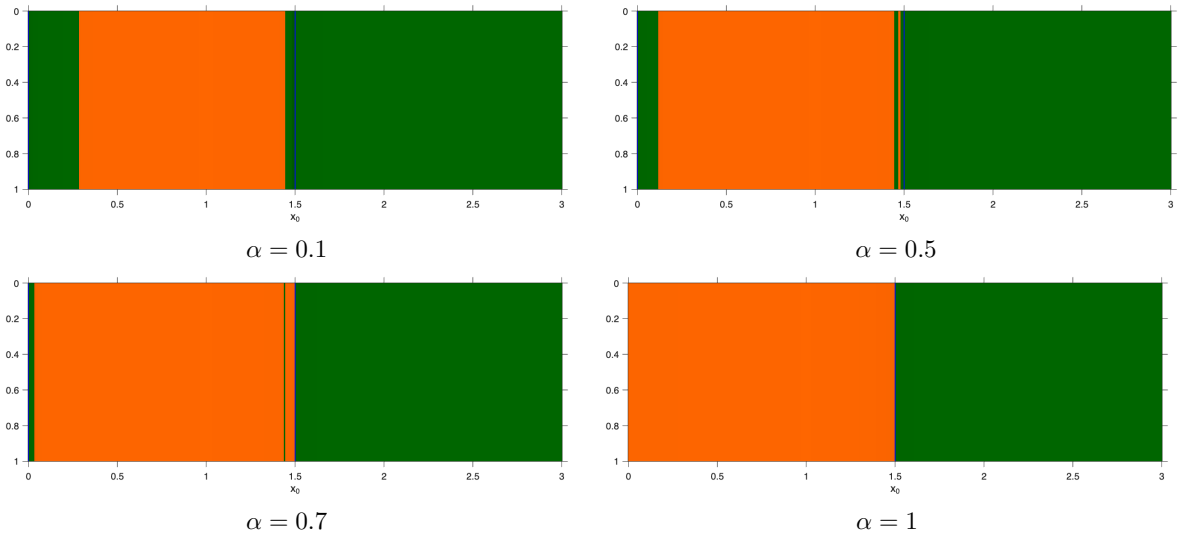


Figure 1: Dynamical lines of the fractal Newton method for $p(x) = x^2 - 3x + 2$ and different values of α.

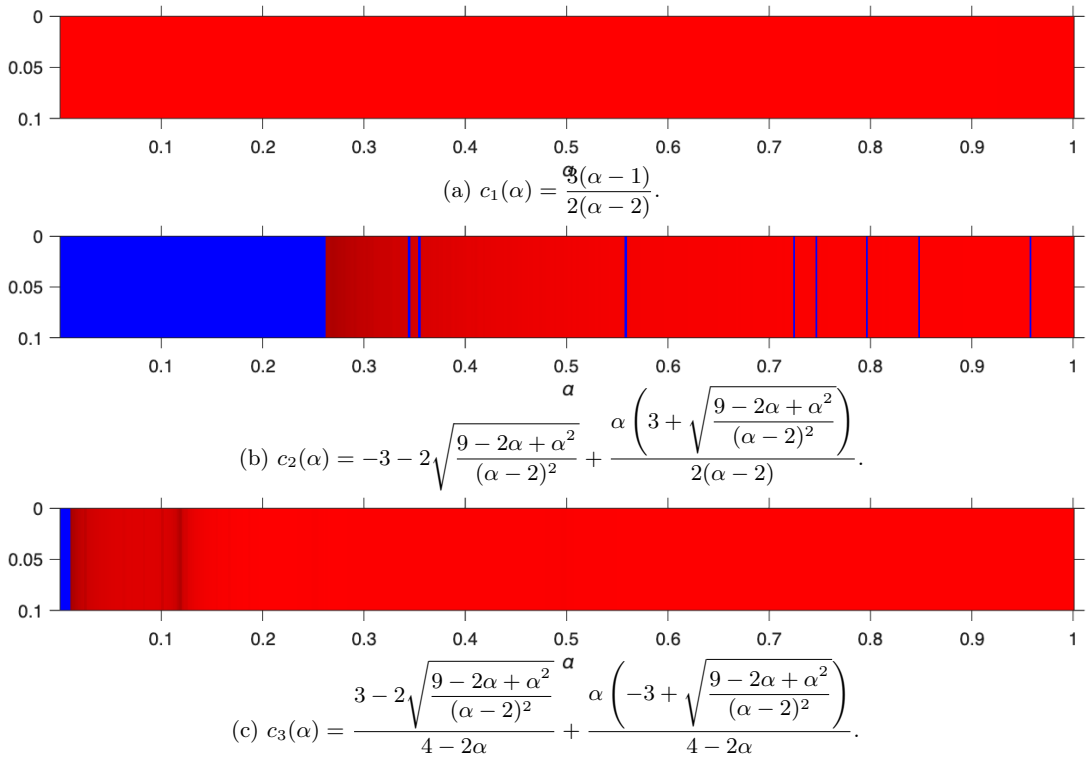
In the dynamical lines it is observed that the variation of α modifies the distribution of the initial points that converge to each root. The point $x = 3/2$ acts as a natural boundary of the

dynamics because it cancels the denominator $2x - 3$; for $0 < \alpha < 1$ the singularity associated with the origin also has an influence. For $\alpha = 0.1$ a small convergence region appears near the origin, which indicates greater sensitivity of the orbits close to the edge of the real branch. For $\alpha = 0.5$ this region shrinks and approaches more closely the separation induced by $x = 1.5$. For $\alpha = 0.7$ the boundary between basins becomes more concentrated near $x = 1.5$, although there is some instability near the origin. Finally, for $\alpha = 1$ the behavior of the classical Newton method is recovered; that is, the line is divided at $x = 1.5$, with convergence toward root 1 on the left and toward root 2 on the right.

From the parameter viewpoint, these lines indicate that α does not act only as a parameter of local order, but as a dynamical parameter that changes the destination of the orbits and the relative size of the basins.

Therefore, a parameter line must be constructed by following the orbit of each free critical point $c_\alpha^{(j)}$ in (14); if that orbit converges to a root, the parameter is considered stable; if it falls into a singularity or does not converge within the maximum number of iterations, an instability region is identified. Thus, numerical convergence is connected with the global structure of the operator, in a graphical and numerical comparison of the values of α that preserve dynamics close to the classical Newton method against those that introduce regions more sensitive to the choice of x_0 [15].

Figure 2 complements this analysis by following the orbit of the free critical points as $\alpha \in (0, 1]$ varies. In these lines, the red color represents convergence toward one of the roots 1 or 2, while blue identifies parameter values for which the orbit falls into a singularity.



(a) $c_1(\alpha) = \dfrac{3(\alpha-1)}{2(\alpha-2)}$.

(b) $c_2(\alpha) = -3 - 2\sqrt{\dfrac{9-2\alpha+\alpha^2}{(\alpha-2)^2}} + \dfrac{\alpha\left(3+\sqrt{\dfrac{9-2\alpha+\alpha^2}{(\alpha-2)^2}}\right)}{2(\alpha-2)}$.

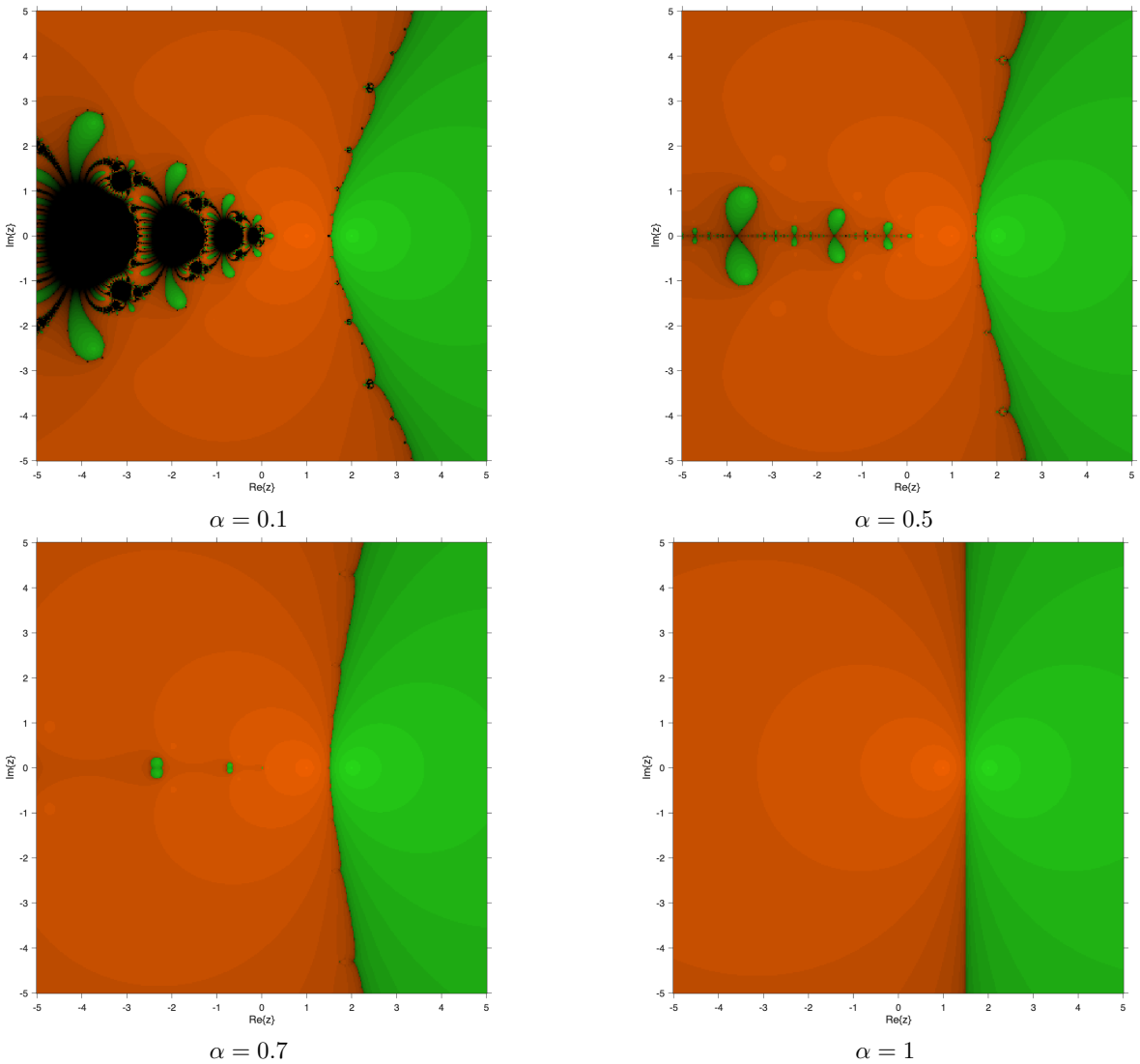
(c) $c_3(\alpha) = \dfrac{3-2\sqrt{\dfrac{9-2\alpha+\alpha^2}{(\alpha-2)^2}}}{4-2\alpha} + \dfrac{\alpha\left(-3+\sqrt{\dfrac{9-2\alpha+\alpha^2}{(\alpha-2)^2}}\right)}{4-2\alpha}$.

Figure 2: Parameter lines associated with the fractal Newton method for $p(x) = x^2 - 3x + 2$, obtained by varying α and following the orbit of the free critical points: (a) $c_\alpha^{(1)}$, (b) $c_\alpha^{(2)}$, and (c) $c_\alpha^{(3)}$.

In Figure 2(a), associated with the critical point $c_\alpha^{(1)} = \frac{1}{2}\frac{3(\alpha-1)}{\alpha-2}$, a completely red band is observed, which indicates that, for the critical orbit considered, the method converges to one of

the roots throughout the represented parameter interval. Figure 2(b) shows a blue region for small values of α and several narrow blue windows within a mostly stable region; this suggests that, although the method recovers stability for intermediate and large values of α, there are isolated parameters where the critical orbit approaches singularities of the operator. In Figure 2(c) the instability is concentrated near $\alpha = 0$, while convergence dominates over the rest of the interval. Taken together, these lines describe how the parameter α defines the global stability of the critical orbits.

The study is completed with dynamical planes in the complex domain. In them, operator (8) is iterated on a grid of initial points $z_0 \in \mathbb{C}$; the orange color represents convergence toward root 1, green convergence toward root 2, and the dark regions indicate singularity or lack of convergence within the maximum number of iterations. Figure 3 shows the planes obtained for $\alpha = 0.1$, $\alpha = 0.5$, $\alpha = 0.7$, and $\alpha = 1$.



$\alpha = 0.1$ $\alpha = 0.5$

$\alpha = 0.7$ $\alpha = 1$

Figure 3: Dynamical planes of the fractal Newton method for $p(z) = z^2 - 3z + 2$ and different values of α.

The dynamical planes allow the global structure of the basins to be observed in greater detail than the real lines. For $\alpha = 0.1$ a very irregular boundary appears, with numerous dark regions and

green components inserted into the orange basin; this indicates that for small values of the parameter there is greater sensitivity with respect to the complex initial condition. For $\alpha = 0.5$ the boundary simplifies, although small convergence and nonconvergence sets are still visible near the real axis. For $\alpha = 0.7$ the basins are more regular and the separation between the roots is concentrated near the region where the denominator $2x - 3$ vanishes. Finally, for $\alpha = 1$ the dynamical plane of the classical Newton method is recovered, where the boundary between basins is much simpler and the partition of the plane remains between the two roots. Thus, as α tends to 1, the dynamics tends to stabilize and the fractal complexity of the boundaries is reduced.

Finally, a convergence plane is considered in which the real initial point x_0 and the parameter α are varied simultaneously. In Figure 4, the orange color indicates convergence toward root 1, blue convergence toward root 2, and the dark regions correspond to values for which the iteration does not converge within the maximum number of iterations considered. If the plane includes values $x_0 < 0$, the interpretation corresponds to the principal complex branch of the fractional power; in strictly real dynamics those points must be excluded.

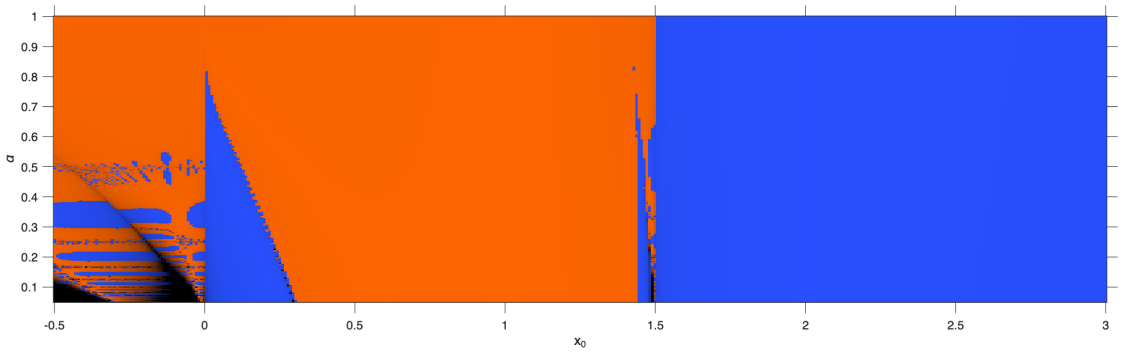


Figure 4: Convergence plane of the fractal Newton method for $p(x) = x^2 - 3x + 2$, varying x_0 and α.

Figure 4 shows that convergence depends not only on the closeness of x_0 to a root, but also on the value of α. For small values of α, dark bands and color alternation appear, especially when $x_0 < 0$, reflecting more sensitive dynamics and the presence of regions where the iteration may lose stability. However, for values of α close to 1, a wide orange region is observed around the initial points that converge to root 1, while for sufficiently large x_0 convergence toward root 2 dominates. The almost vertical boundary around $x_0 = 1.5$ coincides with the singular role of the denominator $2x - 3$ and clearly separates the two main basins. This plane therefore summarizes how the global stability of the method changes when the choice of x_0 is coupled with the fractal order α.

4 Conclusions

The fractal Newton method preserves order two under regular hypotheses: root $\bar{x} > 0$, nonzero fractal derivative $D_F^\alpha f(\bar{x}) \neq 0$, local continuity, and consistent choice of the branch of x^α. The numerical tests show that, when the root becomes degenerate for D_F^α, as occurs in examples with a root at the origin, the ACOC may drop to linear order or cease to be interpretable if the sequence does not converge. When the root is regular, quadratic order is recovered.

In the dynamical example $p(x) = (x - 1)(x - 2)$, the operator presents the roots 1 and 2 as superattracting fixed points, shows no strange admissible fixed points on the positive real branch, has free critical points $c_\alpha^{(1)}$, $c_\alpha^{(2)}$, and $c_\alpha^{(3)}$ for $0 < \alpha < 1$, and requires careful treatment of the singularities at $x = 0$, $x = 3/2$, and the behavior at infinity. These elements justify incorporating dynamical lines, parameter lines, dynamical planes, and convergence planes as tools to understand

the global stability of the method and to detect instability windows associated with specific values of α.

Acknowledgments

This work is part of the development of the doctoral thesis *Fractal iterative methods for solving nonlinear equations* at Universitat Politècnica de València.

References

[1] J. F. Traub, *Iterative Methods for the Solution of Equations*. Englewood Cliffs, NJ: Prentice-Hall, 1964.

[2] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*. New York: Academic Press, 1970. DOI: 10.1016/C2013-0-11263-9.

[3] A. M. Ostrowski, *Solution of Equations and Systems of Equations*, 2nd ed. New York: Academic Press, 1966.

[4] P. Blanchard, "Complex analytic dynamics on the Riemann sphere," *Bulletin of the American Mathematical Society*, vol. 11, no. 1, pp. 85–141, 1984. DOI: 10.1090/S0273-0979-1984-15240-6.

[5] J. H. Curry, L. Garnett, and D. Sullivan, "On the iteration of a rational function: Computer experiments with Newton's method," *Communications in Mathematical Physics*, vol. 91, pp. 267–277, 1983. DOI: 10.1007/BF01211162.

[6] J. Milnor, *Dynamics in One Complex Variable* (Annals of Mathematics Studies 160), 3rd ed. Princeton, NJ: Princeton University Press, 2006.

[7] A. F. Beardon, *Iteration of Rational Functions: Complex Analytic Dynamical Systems* (Graduate Texts in Mathematics 132). New York: Springer-Verlag, 1991. DOI: 10.1007/978-1-4612-4422-6.

[8] W. Chen, "Time–space fabric underlying anomalous diffusion," *Chaos, Solitons & Fractals*, vol. 28, no. 4, pp. 923–929, 2006. DOI: 10.1016/j.chaos.2005.08.199.

[9] W. Chen and Y. Liang, "New methodologies in fractional and fractal derivatives modeling," *Chaos, Solitons & Fractals*, vol. 102, pp. 72–77, 2017. DOI: 10.1016/j.chaos.2017.03.066.

[10] A. Deppman, E. Megías, and R. Pasechnik, "Fractal derivatives, fractional derivatives and q-deformed calculus," *Entropy*, vol. 25, no. 7, p. 1008, 2023. DOI: 10.3390/e25071008.

[11] J. Paredes Morales, "Métodos iterativos fractales para resolver ecuaciones no lineales," Trabajo Fin de Máster, Universitat de València, València, Spain, 2025.

[12] A. Cordero and J. R. Torregrosa, "Variants of Newton's method using fifth-order quadrature formulas," *Applied Mathematics and Computation*, vol. 190, no. 1, pp. 686–698, 2007. DOI: 10.1016/j.amc.2007.01.062.

[13] F. I. Chicharro, A. Cordero, and J. R. Torregrosa, "Drawing dynamical and parameters planes of iterative families and methods," *The Scientific World Journal*, vol. 2013, p. 780 153, 2013. DOI: 10.1155/2013/780153.

[14] Á. A. Magreñán, "A new tool to study real dynamics: The convergence plane," *Applied Mathematics and Computation*, vol. 248, pp. 215–224, 2014. DOI: 10.1016/j.amc.2014.09.061.

[15] J. L. Varona, "Graphic and numerical comparison between iterative methods," *The Mathematical Intelligencer*, vol. 24, no. 1, pp. 37–46, 2002. DOI: 10.1007/BF03025310.

Dynamics of derivative-free methods for solving nonlinear equations with multiple roots

Jorge Rico [♭],[1] Alicia Cordero [♭♮] and Juan R. Torregrosa [♭♮]

(♭) DAMRES Research Group - IMM, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

(♮) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

Iterative methods for approximating the roots of nonlinear equations play a fundamental role in engineering, from hydraulics to structural analysis. These methods enable solutions to be found for equations of the form $f(x) = 0$ when it is not possible to obtain an exact analytical solution. They start with an initial approximation (which may be more or less accurate) and, by iterating over a fixed-point function $G(x)$, another approximation is obtained.

However, when a root has a multiplicity, m, greater than one, conventional iterative methods may converge slowly or even fail. The multiplicity of a root x^* is defined as the smallest positive integer m such that $f(x^*) = 0$ and $f'(x^*) = f''(x^*) = \ldots = f^{(m-1)}(x^*) = 0$, but $f^{(m)}(x^*) \neq 0$. For example, Newton's well-known method has quadratic order for single roots ($m = 1$), however, it becomes linear-order when applied to multiple roots ($m > 1$). One modification on Newton's method was proposed by Schröder [1] in 1870, where the convergence remains quadratic by using the following iterative expression,

$$x_{k+1} = x_k - m\frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, \ldots \tag{1}$$

Over the years, a vast number of iterative methods have been devised to solve this problem, and the vast majority of them explicitly require the multiplicity of the root to be found [2], [3], [4].

Avoiding the need to know the multiplicity, m, is undoubtedly an advantage when dealing with the most complex problems. The first method in the literature designed for multiple roots that does not rely on their multiplicity is based on Newton's method and maintains the second order, it is also known as the Schröder scheme [1], whose iterative expression is

$$x_{k+1} = x_k - \frac{f(x_k)f'(x_k)}{f'(x_k)^2 - f(x_k)f''(x_k)}, \quad k = 0, 1, 2, \ldots \tag{2}$$

Recent methods have also been developed with the same aim of not knowing the multiplicity a priori, such as a method based on the classical secant scheme [5], and another modifying Kurchatov's

[1] jricest@upv.edu.es

method [6] denoted as *KM*, which follows this scheme:

$$x_{k+1} = x_k - \frac{g(x_k)}{g[2x_k - x_{k-1},\, x_{k-1}]}, \quad k = 0, 1, 2, \ldots \tag{3}$$

where $g(x) = \dfrac{f(x)}{f'(x)}$ and the expression of the divided difference operator is $f[y, z](y - z) = f(y) - f(z)$, in the above case (where $y = 2x_k - x_{k-1}$ and $z = x_{k-1}$) is called the divided difference of Kurchatov.

The most basic and widely used approach to avoid relying on the derivative is to use divided differences to approximate it. The last two works mentioned also explore this possibility including also the design of a family of methods [5]. In Cordero et al. [6], the *KMD* method was developed using this technique with forward divided differences applied to *KM* method, as shown below,

$$x_{k+1} = x_k - \frac{g(x_k)}{g[2x_k - x_{k-1},\, x_{k-1}]}, \quad k = 0, 1, 2, \ldots \tag{4}$$

where $g(x) = \dfrac{f(x)}{f[x + f(x),\, x]}$.

In this paper, we propose a new derivative-free iterative scheme for finding multiple roots in nonlinear equations without prior knowledge of the multiplicity. The design process is outlined in Section 2. Section 3 comprises two subsections providing a comprehensive analysis of the dynamics of the proposed method and the *KMD* method when applied to the same polynomials. Finally, the conclusions of this paper are presented in Section 4.

2 Design of a new method

Based on the same principle as the *KMD* method, which approximates the derivative of the nonlinear equation, *f*, using forward divided differences, the scheme designed in this work employs central divided differences in the construction of the function *g*, as shown below. This method will be denoted as *KMDs*:

$$x_{k+1} = x_k - \frac{g(x_k)}{g[2x_k - x_{k-1},\, x_{k-1}]}, \quad k = 0, 1, 2, \ldots \tag{5}$$

where $g(x) = \dfrac{f(x)}{f[x + f(x),\, x - f(x)]}$.

Using Taylor's expansion this new scheme holds the second convergence order of the original method and has the following error equation,

$$e_{k+1} = -\frac{1}{m} C_1 e_k^2 + \mathcal{O}_3(e_k, e_{k-1}), \tag{6}$$

where $C_1 = \frac{m!}{(m+1)!} \frac{f^{(m+1)}(\alpha)}{f^{(m)}(\alpha)}$ and $\mathcal{O}_3(e_k, e_{k-1})$ denotes all the terms whose sum of the exponents of e_k and e_{k-1} is at least 3.

3 Dynamical Analysis

In this section, based on real multidimensional dynamics, we will analyze the stability of the two iterative methods, *KMD* and *KMDs*. The analysis will conclude with the display of various dynamic planes to visualize this behavior across different polynomials. But first, we will introduce

the necessary concepts.

Since those methods use two initial estimations, the standard iteration function $\psi : \mathbb{R}^2 \to \mathbb{R}$ cannot possess fixed points, an auxiliary discrete dynamical system $G : \mathbb{R}^2 \to \mathbb{R}^2$, defined as $G(x_{k-1}, x_k) = (x_k, \psi(x_{k-1}, x_k))$, is introduced. Within this system, a fixed point (z, x) must satisfy $z = x$ and $x = \psi(z, x)$, with *strange fixed points* being those that do not correspond to the roots of the original equation $f(x) = 0$.

The stability of these fixed points is evaluated through the eigenvalues (λ_1, λ_2) of the Jacobian matrix of the operator G evaluated at the fixed point. According to [7], a fixed point is classified as an *attractor* if the magnitudes of all eigenvalues are strictly less than one ($|\lambda_j| < 1$), a *repulsor* if all are greater than one ($|\lambda_j| > 1$), and *unstable* (acting as a repulsor or saddle) if at least one eigenvalue exceeds one. Additionally, critical points are identified as those where the eigenvalues of the Jacobian matrix are zero.

Finally, the global dynamical behavior is characterized by the basin of attraction, defined as $\mathcal{A}(\hat{x}) = \{y \in \mathbb{R}^n : G^r(y) \to \hat{x}, r \to \infty\}$, which represents the set of all initial estimations that eventually converge to the fixed point $\hat{x}$. To represent them, the dynamical planes are generated. Using a 1000×1000 point mesh, mapping the initial iterations x_1 and x_0 to the abscissa and ordinate axes, respectively. An initial point is considered to converge to a root if the distance between the iteration and the root falls below 10^{-3} within a maximum of 60 iterations. The convergence behavior is then color-coded: orange and blue indicates convergence to one of the roots, and black is reserved for non-convergent points. Roots are marked with an asterisk, fixed points with a circle and critical points with squares.

Therefore, in the following subsections for each method *KMD* (4) and *KMDs* (5), we will determine the operator G for the polynomial $p(x) = (x+1)(x-1)^m$, $m \in \mathbb{Z}_{\geq 1}$, its fixed points and their nature, its critical points, and the dynamical planes.

3.1 *KMD*

Now, we calculate the vectorial operator for the *KMD* method where $z = x_{k-1}$ and $x = x_k$.

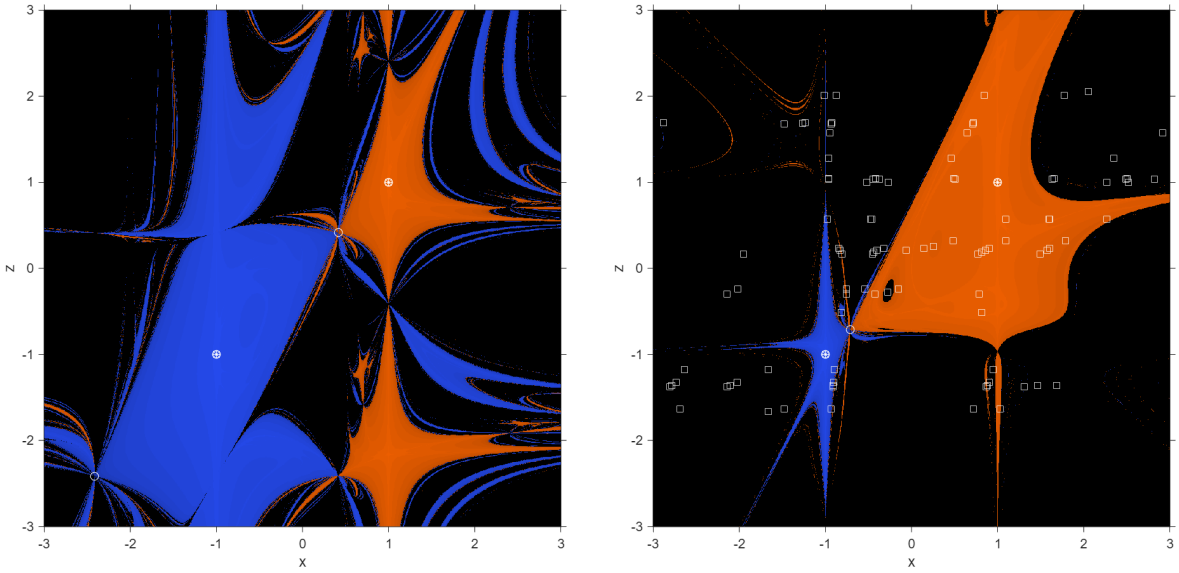
$$G_{KMD}(z, x) = \left(x, \quad x - \frac{A(x)2(x-z)}{A(2x-z) - A(z)} \right),$$

where $A(y) = \dfrac{(y-1)^{2m}(y+1)}{-(y-1)^m + (1+(y-1)^m)(y-1+(y-1)^m(y+1))^m}$.

In this expression, we cannot directly substitute $z = x$ because that would result in a division by zero; while the expression could be expanded, doing so is not practical. That is why, for the analysis, we will consider multiplicities $m = 1, 2, 3, 4$ and evaluate them independently. In this case, with a known multiplicity, we can simplify the operator and avoid division by zero when calculating the fixed points.

For the case $m = 1$ there are four fixed points, $z^* = \pm 1$ that are superattracting, and the strange fixed points $z^F = -1 \pm \sqrt{2}$ that are saddle points. As we can see in the respective dynamical plane (Figure 1 (a)) the basins of attraction of both attractive points are asymmetrical. The only critical points in this case are the roots 1 and -1.

In case $m = 2$, the importance of the method for solving multiple roots begins to become apparent, since $z^* = 1$ has multiplicity 2; as can be seen on the dynamical plane, its basin of attraction is considerably increased compared with the single root. In this case, a third fixed point appears, an strange fixed point, also a saddle point, at $z^F \approx -0.710434$. In this case, the actual critical points (which have been found by approximation) are 130 pairs (z, x) and eight real roots of a polynomial of degree 16 in x and variables in z. Many of these critical points are shown in Figure 1 (b).



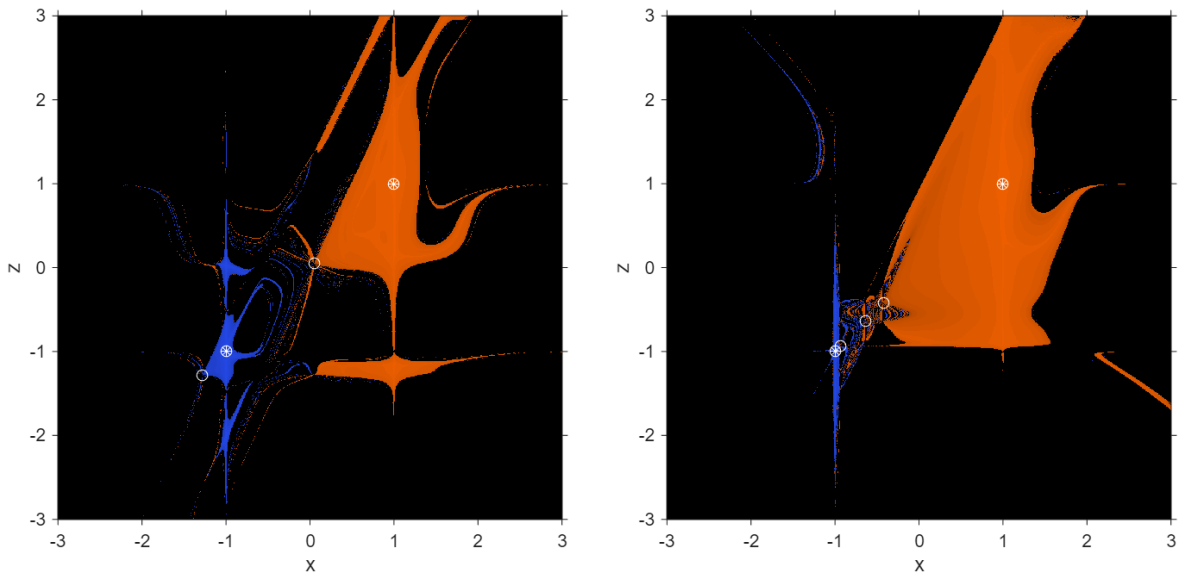
(a) $m = 1$ (b) $m = 2$

Figure 1: *KMD* dynamical planes in $p(x) = (x + 1)(x - 1)^m$.

Finally, for cases where $m = 3, 4$, the analysis is similar. In Table 1 are the fixed points for each case, along with their asymptotic behavior. The dynamic plots (Figure 2) show a tendency for the basin of attraction of the simple root to be reduced, and therefore for the dark region of initial approximations that do not converge to any fixed point to increase. In these cases, the calculation of critical points becomes computationally complex and will not be addressed in this work.

Table 1: Fixed points of *KMD* operator applied to $p(x) = (x + 1)(x - 1)^m$

Case m	z^F	behavior
	-1.2794200	saddle point
3	-1	superattractive
	0.0470807	saddle point
	1	superattractive
	-1	superattractive
	-0.931863	saddle point
4	-0.637965	saddle point
	-0.425848	saddle point
	1	superattractive



(a) $m = 3$ (b) $m = 4$

Figure 2: *KMD* dynamical planes in $p(x) = (x+1)(x-1)^m$.

3.2 *KMDs*

The operator for the *KMDs* method is as follow

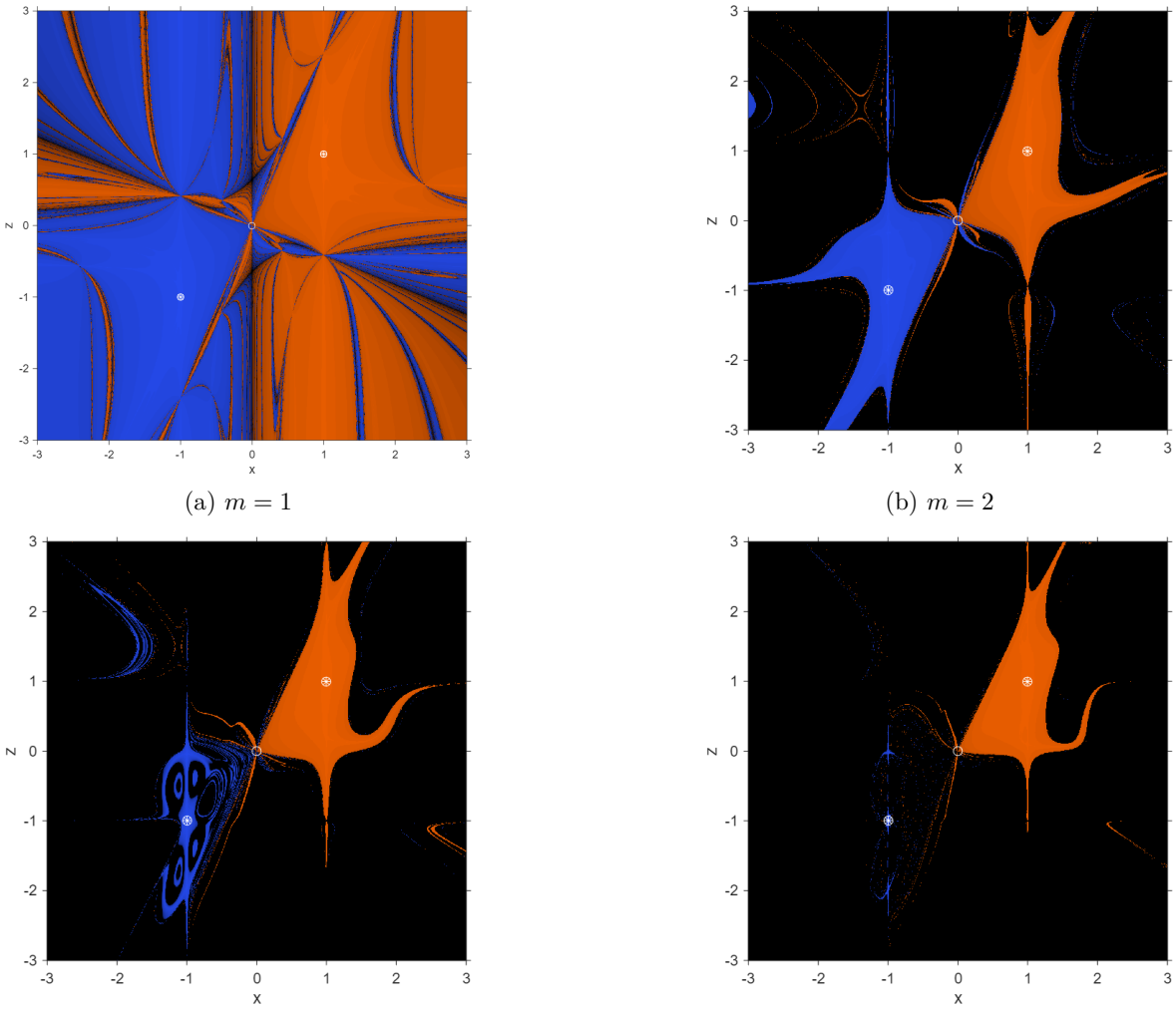
$$G_{KMDs}(z,x) = \left(x, \quad x - \frac{B(x)2(x-z)}{B(2x-z) - B(z)}\right),$$

where $B(y) = \dfrac{2(y-1)^{2m}(y+1)}{(-1+(y-1)^m)(y-1-(y-1)^m(y+1))^m + (1+(y-1)^m)(y-1+(y-1)^m(y+1))^m}$.

As with the previous method, due to Kurchatov's divided difference, an indeterminacy appears when calculating the fixed points. We shall therefore proceed in the same way, analyzing the behavior of the method for four multiplicities of the polynomial.

For all the four cases $m = 1, 2, 3, 4$ there are three fixed points, $z^* = -1$ and $z^* = 1$ that are superattracting, and $z^F = 0$ that is a saddle point. The similar behavior observed in all the cases analyzed could be explained by the symmetry employed in the central difference used in the method. In this method, the critical points have only been found for the case where $m = 1$, which are the roots 1 and –1 of the polynomial.

In this case, the dynamic planes (Figure 3) also exhibit the same behavior. For the polynomial without multiple roots, we can observe extended, symmetrical basins of attraction without dark regions, which gives a good indication that *KMDs* works well at finding simple roots. As the multiplicity of the root $z^* = 1$ increases, the attractor basin of the simple root decreases until it is almost zero for $m = 4$, but it can be seen that the basin of the multiple root remains similar in all three cases.



(a) $m = 1$ (b) $m = 2$

(c) $m = 3$ (d) $m = 4$

Figure 3: *KMDs* dynamical planes in $p(x) = (x+1)(x-1)^m$.

4 Conclusions

In this paper, a comparison has been carried out between two similar methods for solving nonlinear equations with multiple roots without using derivatives or multiplicity. One of the methods is already known, whilst the other has been newly designed. This involves the use of concepts and results from multidimensional real analysis applied to a polynomial with two roots: one simple and the other of variable multiplicity. Thanks to this analysis, greater regularity can be observed in the newly designed *KMDs* method (5) in terms of: a) identical behavior of fixed points, and b) greater stability in the basin of attraction of the multiple root. It is also observed that the newly designed method performs better for single roots than the *KMD* method (4), as the basins of attraction are wider.

Acknowledgments

This study has been developed during the "Extracurricular external internships at the Institute of Multidisciplinary Mathematics (IMM)", in the summer of 2025.

References

[1] E. Schröder, "Ueber unendlich viele Algorithmen zur Auflösung der Gleichungen," *Mathematische Annalen*, vol. 2, no. 2, pp. 317–365, Jun. 1870, ISSN: 1432-1807. DOI: 10.1007/BF01444024.

[2] S. Li, L. Cheng, and B. Neta, "Some fourth-order nonlinear solvers with closed formulae for multiple roots," *Computers & Mathematics with Applications*, vol. 59, no. 1, pp. 126–135, Jan. 2010, ISSN: 08981221. DOI: 10.1016/j.camwa.2009.08.066.

[3] F. I. Chicharro, R. A. Contreras, and N. Garrido, "A Family of Multiple-Root Finding Iterative Methods Based on Weight Functions," *Mathematics*, vol. 8, no. 12, p. 2194, Dec. 2020, ISSN: 2227-7390. DOI: 10.3390/math8122194.

[4] R. Behl, A. Cordero, and J. R. Torregrosa, "A new higher-order optimal derivative free scheme for multiple roots," *Journal of Computational and Applied Mathematics*, vol. 404, p. 113 773, Apr. 2022, ISSN: 03770427. DOI: 10.1016/j.cam.2021.113773.

[5] F. I. Chicharro, N. Garrido-Saez, and J. H. Jerezano, "Secant-Type Iterative Classes for Nonlinear Equations with Multiple Roots," *Algorithms*, vol. 18, no. 9, p. 568, Sep. 2025, ISSN: 1999-4893. DOI: 10.3390/a18090568.

[6] A. Cordero, N. Garrido, J. R. Torregrosa, and P. Triguero-Navarro, "Modifying Kurchatov's method to find multiple roots of nonlinear equations," *Applied Numerical Mathematics*, vol. 198, pp. 11–21, Apr. 2024, ISSN: 0168-9274. DOI: 10.1016/j.apnum.2023.12.009.

[7] A. Cordero and J. R. Torregrosa, "Variants of Newton's Method using fifth-order quadrature formulas," *Applied Mathematics and Computation*, vol. 190, no. 1, pp. 686–698, Jul. 2007, ISSN: 00963003. DOI: 10.1016/j.amc.2007.01.062.

Approximation of inverses and generalized inverses of complex matrices

Alicia Cordero [♭], Eva G. Villalba [♭], Juan R. Torregrosa [♭] and Paula Triguero-Navarro [♭ 1]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

Introduction

The problem of computing the inverse of a non-singular complex matrix $A \in \mathbb{C}^{n\times n}$ arises in many areas of applied mathematics, computational physics, and data science. Although obtaining A^{-1} via LU or QR decomposition is standard practice, in situations involving large matrices, it is desirable to have *iterative methods* that allow the inverse to be approximated without performing explicit factorizations.

A classic approach involves reformulating the problem as the solution to a nonlinear matrix equation:

$$F(X) = X^{-1} - A = 0.$$

In this way, it is possible to extend iterative techniques for nonlinear systems to the matrix problem.

Furthermore, in contexts where the matrix is non-square or singular, the use of *generalized inverses* is required, the most notable being the Moore–Penrose inverse $A^{\dagger}$; and the Drazin inverse A^{D}, which is useful for singular matrices of finite rank.

Our goal is to present a new high-order iterative method for computing A^{-1};, to extend this method to the estimation of pseudoinverses $A^{\dagger}$ and Drazin inverses A^{D};, and to compare its performance with classical and recent methods.

Preliminaries

The convergence of iterative algorithms depends largely on the magnitude of the matrix error,

$$E_k = I - AX_k.$$

Under appropriate conditions, if $\|E_0\| < 1$, the convergence of methods based on

$$X_{k+1} = X_k P(E_k),$$

where P is a polynomial, is usually guaranteed.

On the other hand, the pseudoinverse or Moore–Penrose inverse is the only matrix X such that:

$$AXA = A, \quad XAX = X, \quad (AX)^* = AX, \quad (XA)^* = XA.$$

It is also known [1] that the Drazin inverse A^{D}, where $k = \mathrm{ind}(A)$, is the only matrix that satisfies:

$$A^{k+1}X = A^k, \qquad XAX = X, \qquad AX = XA.$$

[1] ptrinav@mat.upv.es

Proposed New Method

The key to the design lies in the introduction of scalar accelerators into Newton-type schemes. First, a Newton–Schulz-type predictor step is applied:

$$Y_k = X_k(2I - AX_k).$$

Next, a cubic corrector (derived from a scheme with vector accelerators constructed in [2]) is used, resulting in a fifth-order method:

$$X_{k+1} = Y_k\,[2I - AY_k + (I - AY_k)^2].$$

This method can be written as a single-step scheme:

$$X_{k+1} = X_k(2I - AX_k)\,[3I + AX_k(-2I + AX_k)(3I + AX_k(-2I + AX_k))]\,. \tag{1}$$

The final scheme, denoted by MP, requires five matrix products per iteration and has a convergence order of five for the estimation of inverses and six for generalized inverses.

Below, we present the convergence results obtained for the proposed method regarding the estimation of inverses, pseudoinverses, and Drazin inverses. In these results, the assumptions regarding the initial estimate used are of particular relevance.

Theorem 1. *Let $A \in \mathbb{C}^{n\times n}$ be nonsingular. Taking an initial estimate X_0 such that $\|I-AX_0\| < 1$, then the sequence $\{X_k\}$ obtained by the method (1) converges to A^{-1} with order of convergence 6.*

Theorem 2. *Let $A \in \mathbb{C}^{m\times n}$ and $A^\dagger$ be its Moore–Penrose pseudoinverse. Consider an initial estimate $X_0 = \beta A^*$, where $\beta < \dfrac{1}{\sigma_1^2}$ is a constant and σ_1 is the largest singular value of A. Then, the sequence $\{X_k\}$ obtained by the method (1) converges to $A^\dagger$ with order 6.*

Theorem 3. *Let $A \in \mathbb{C}^{n\times n}$ and A^D its Drazin inverse. Take an initial estimate $X_0 = \alpha A^l$ where $l \geq ind(A)$ and α such that $\|I - AX_0\| < 1$. Then the sequence $\{X_k\}$ obtained by the method (1) converges to A^D with order 6.*

Numerical results

We include three reference methods that we use for numerical comparison with our iterative scheme. First, the Newton–Schulz scheme [3], whose iterative expression is

$$X_{k+1} = X_k(2I - AX_k).$$

We also use the well-known Chebyshev method [4],

$$X_{k+1} = X_k(3I - AX_k(3I - AX_k)),$$

the family of Kansal et al. [5]:

$$X_{k+1} = X_k\Big((3+\beta)I + (-3-3\beta)AX_k + (1+3\beta)(AX_k)^2 - \beta(AX_k)^3\Big),$$

of order 3 for any β, and of order 4 if $\beta = 1$. and the parametric class of Cordero et al. [6]:

$$X_{k+1} = X_k\Big[(4+\alpha)I + AX_k(-2(3+2\alpha)I + AX_k(2(2+3\alpha)I + (AX_k)(-(1+4\alpha)I + \alpha AX_k)))\Big],$$

of order 4 for any α, order 5 if $\alpha = 1$.

The numerical experiments were performed in `MATLAB R2024b`, with the stopping criterion

$$\|X_{k+1} - X_k\|_2 + \|F(X_{k+1})\|_2 < 10^{-5}.$$

The initial estimate used is:

$$X_0 = 0.5\,\frac{A^T}{\|A\|^2}.$$

Example 1: Inverse Estimation

The results presented in Table 1 show that the proposed method (MP) finds the inverse with fewer iterations than the compared methods, including Newton-Schulz, Chebyshev, and CSTV. Furthermore, the final error and the ACOC confirm the high accuracy and the expected order of convergence. For larger matrices, the method maintains competitive runtime, highlighting its efficiency compared to lower-order methods.

Table 1: Calculation of the inverse for random matrices of different sizes

Method	n	*iter*	$\|X_{k+1} - X_k\|$	$\|F(X_{k+1})\|$	ACOC	Time(s)
Newton-Schulz	10	20	5.1948e-07	4.9275e-16	2.0054	1.0469e+00
Chebyshev	10	15	3.1427e-08	1.9837e-18	3.0021	1.0430e+00
CSTV $\alpha = 0.5$	10	11	7.8109e-18	1.7949e-21	4.0103	1.2969e+00
CSTV $\alpha = 1$	10	10	9.2125e-11	9.9601e-22	5.0180	2.2500e+00
MP	10	9	1.5707e-07	6.0398e-22	6.1798	1.0312e+00
Newton-Schulz	50	28	1.6636e-07	5.2424e-18	2.0027	9.8188e+01
Chebyshev	50	20	1.4057e-07	7.3527e-20	3.0031	7.0234e+01
CSTV $\alpha = 0.5$	50	14	8.8430e-07	3.3614e-20	4.1308	1.4302e+02
CSTV $\alpha = 1$	50	14	7.6475e-20	3.5772e-20	2.9211	2.4323e+02
MP	50	12	1.1921e-07	9.7802e-21	6.2161	2.4141e+01

Example 2: Moore–Penrose inverse

As shown in Table 2, the proposed method satisfies the properties of the pseudoinverse with an extremely small error under all test conditions. Compared to other methods, MP requires fewer iterations to reach convergence and maintains a similar or shorter runtime, demonstrating its robustness and applicability to non-square matrices.

Table 2: Calculation of the pseudoinverse for random matrices of different sizes

Method	m	n	*iter*	$\|X_{k+1} - X_k\|$	ACOC	Time(s)
Newton-Schulz	10	20	13	2.8047e-07	2.0103	2.0469e+00
Chebyshev	10	20	10	1.2368e-07	3.0002	1.8724e+00
CSTV $\alpha = 0.5$	10	20	7	4.4033e-06	4.3370	7.0027e+01
CSTV $\alpha = 1$	10	20	7	7.8005e-11	5.0299	1.1875e+00
MP	10	20	7	3.4405e-26	6.0002	1.2344e+00
Newton-Schulz	50	60	20	1.0796e-08	2.0028	1.2842e+02
Chebyshev	50	60	15	1.0834e-07	3.0012	1.0713e+02
CSTV $\alpha = 0.5$	50	60	11	1.4231e-20	3.9830	7.0027e+02
CSTV $\alpha = 1$	50	60	10	6.6396e-13	5.0095	7.8438e+01
MP	50	60	9	2.8193e-09	6.1138	6.0453e+01

Example 3: Drazin inverse

Let A be a 6×6 matrix,

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 1 & -1 & 0 & 0 \\ -1 & -1 & -1 & 1 & 0 & 0 \\ -1 & -1 & -1 & 0 & 2 & -1 \\ -1 & -1 & 0 & -1 & -1 & 2 \end{pmatrix}$$

whose index is $ind(A) = 2$. Table 3 shows that, even though the condition $|I - AX_0| < 1$ is not strictly satisfied, the proposed method converges reliably to the Drazin inverse. The number of iterations is lower than that of the comparative methods, and the errors in the equations defining A^D are practically zero, indicating stable and accurate behavior.

Table 3: Results for the Drazin inverse

Method	*iter*	Ex	e_1	e_2	e_3	ACOC	Time(s)
Newton-Schulz	11	6.4029e-11	2.8990e-21	2.8990e-21	1.0513e-37	2.0008	1.8141e-01
Chebyshev	8	1.0271e-10	2.0347e-25	2.0864e-25	1.2371e-37	3.0002	1.6134e-01
CSTV $\alpha = 0.5$	6	5.3462e-8	1.4441e-30	1.4441e-30	1.9849e-37	4.1557	1.2437e-01
CSTV $\alpha = 1$	6	3.3373e-13	1.9643e-36	2.5382e-37	1.5281e-37	5.0103	1.2187e-01
MP	6	9.2667e-27	2.1027e-37	1.7632e-38	1.0513e-37	6.0001	1.1703e-01

Conclusions

This paper presents a new high-order iterative method for approximating the inverse, the Moore–Penrose inverse, and the Drazin inverse of complex matrices. The theoretical analysis has shown that the proposed scheme improves the order of convergence compared to classical and recent methods in the literature, while maintaining a moderate computational cost. Numerical experiments corroborate these properties: the method reaches the solution in fewer iterations, yields smaller final errors, and performs particularly efficiently in the computation of generalized inverses. Taken together, the results indicate that the new method constitutes a robust and competitive alternative, suitable for large-scale problems and situations where stability and convergence speed are critical.

Funding

This conference was supported by Universitat Politècnica de València. Funded with IMM Research Funding Call 2026 – Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València.

References

[1] A. Ben-Israel and T. N. Greville, *Generalized inverses: theory and applications*. Springer, 2003.

[2] A. Cordero, R. V. Rojas-Hiciano, J. R. Torregrosa, and M. P. Vassileva, "A highly efficient class of optimal fourth-order methods for solving nonlinear systems," *Numerical Algorithms*, vol. 95, pp. 1879–1904, 2024.

[3] N. J. Higham, *Functions of matrices: theory and computation*. SIAM, 2008.

[4] S. Amat, J. A. Ezquerro, and M. A. Hernández-Verón, "Approximation of inverse operators by a new family of high-order iterative methods," *Numerical Linear Algebra with Applications*, vol. 21, pp. 629–644, 2014.

[5] M. Kansal, M. Kaur, L. Rani, and L. Jäntschi, "A cubic class of iterative procedures for finding the generalized inverses," *Mathematics*, vol. 11, p. 3031, 2023.

[6] A. Cordero, E. Segura, J. R. Torregrosa, and M. P. Vassileva, "New iterative procedures for approximating different types of inverse matrices: Convergence and stability," *Mathematical Methods in the Applied Sciences*, vol. 48, no. 8, pp. 8806–8817, 2025.

Analysis and application of a parametric family of seventh-order iterative methods to the solution of a traffic management problem

Eva G. Villalba[♭],[1] Jesús Gallinal[♭],[2] Jorge D. Ávila[♮], and Xènia Nicolau-Casademont[♯]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

(♮) EnergyRIS Research & Intelligence Solutions and Universidad de La Laguna
Canary Island, Spain.

(♯) Department of Signal Theory and Communications, Carlos III University of Madrid
Leganés, Madrid, Spain.

1 Introduction

The development of iterative methods for solving non-linear equations and systems is an essential tool in numerical analysis, as it enables us to approach problems in applied mathematics which, due to their complexity or large scale, cannot be solved analytically.

Within this field of research, Newton's method is one of the leading approaches due to its simplicity, applicability and quadratic order of convergence. This method has inspired many others in the literature to develop improved methods, whether by composing the method with others, introducing weight functions, making it derivative-free, etc.

In this work, we focus on the study of an iterative family inspired by composing Newton's method with itself three times, freezing the derivative and modifying it using two weight functions. This yields a three-step family requiring five functional evaluations:

$$FN_a: \begin{cases} y_k = x_k - \dfrac{f(x_k)}{f'(x_k)}, \\ z_k = y_k - \dfrac{f'(y_k)+3f'(x_k)}{5f'(y_k)-f'(x_k)}\dfrac{f(y_k)}{f'(x_k)}, \\ x_{k+1} = z_k - \dfrac{af'(x_k)-f'(y_k)}{(a-2)f'(y_k)+f'(x_k)}\dfrac{f(z_k)}{f'(x_k)}, \end{cases} \tag{1}$$

where $a \in \mathbb{R} \setminus \{1\}$, with 1 excluded because the final step would be of the Traub type and, as we see in Theorem 1, cancels out the denominator of the error equation.

[1] egarvil@mat.upv.es
[2] jagalmor@doctor.upv.es

2 Convergence analysis

This section prove the order of convergence of the FN_a methods from the parametric family described in (1).

Theorem 1 *Let $f : D \subset \mathbb{R} \to \mathbb{R}$ be a function that is sufficiently differentiable in a neighbourhood D of α, such that $f(\alpha) = 0$. Assuming that $f'(\alpha) \neq 0$, if x_0 is sufficiently close to α, then the sequence $\{x_k\}_{k\geq 0}$ generated by the family FN_a converges to α with order 7 for all $a \in \mathbb{R} \setminus \{1, -1\}$ and order 8 for $a = -1$, where (2) is the error equation:*

$$e_{k+1} = \frac{(1+a)\, c_2^4\, (3c_2^2 - 2c_3)}{-1+a}\, e_k^7 + \mathcal{O}(e_k^8), \tag{2}$$

where $\Gamma = f'(\alpha)$, $c_k = \frac{1}{k!}\frac{f^{(k)}(\alpha)}{f'(\alpha)}$, $k = 2, 3, 4, \ldots$, and $e_k = x_k - \alpha$ is the error in the k-th iteration.

3 Analysis of complex dynamics

In this section, we study the complex dynamics or accessibility of the uniparametric family FN_a, defined in (1), applied to any polynomial of the form $p(z) = (z - r)(z - s)$, where $z, r, s \in \mathbb{C}$. To simplify this analysis, we conjugate the operator $R_{p,a}$ using the Möbius transformation defined in Theorem 2. This conjugation allows us to obtain an operator whose stability is equivalent to that of $R_{p,a}$, but which does not depend on the parameters r and s. Indeed, thanks to the Scaling Theorem, it is possible to compare the dynamical behaviour of both operators, since this result guarantees dynamical equivalence between operators conjugate by affine transformations.

Theorem 2 (Scaling Theorem) *Let $f(z)$ be an analytical function on the Riemann sphere, and let $T(z) = \alpha z + \beta$, where $\alpha \neq 0$, be an affine function. We define $g(z) = \lambda(f \circ T)(z)$, where $\lambda \in \mathbb{C} \setminus \{0\}$. Let $R_{f,a}$ and $R_{g,a}$ be the fixed-point operators of the family FN_a associated with f and g, respectively. Then, $R_{f,a}$ is analytically conjugate to $R_{g,a}$ via T, that is, $(T \circ R_{g,a} \circ T^{-1}) = R_{f,a}$. Upon applying the Möbius transformation $h(z) = \frac{z-r}{z-s}$, the resulting transformed rational operator is as follows:*

$$\begin{aligned} O_a(z) &= (h \circ R_{p,a} \circ h^{-1})(z) \\ &= \frac{z^7(2z+3)\left(a(z+1)^2\left(2z^4+3z^3+4z^2+6z+4\right)-2z^6-3z^5+2z^4+3z^3+6z+4\right)}{(3z+2)\left(a(z+1)^2\left(4z^4+6z^3+4z^2+3z+2\right)+4z^6+6z^5+3z^3+2z^2-3z-2\right)}. \end{aligned} \tag{3}$$

Next, we examine the fixed points of the operator O_a, that is, the values of $z \in \mathbb{C}$ such that $O_a(z) = z$, analysing their existence and nature as a function of the parameter a.

Theorem 3 *Let $a \in \mathbb{C} \setminus \{1\}$. The operator O_a in equation (3) has the following fixed points:*

1. *$z = 0$ and $z = \infty$ are superattractor fixed points for all $a \neq 1$.*
2. *$z = 1$ is a strange fixed point for $a \neq -\frac{5}{38}$.*
3. *The roots of the polynomial:*

$$\begin{aligned} K(z) = {} & z^{12}(4a-4) + z^{11}(24a-16) + z^{10}(69a-21) + z^9(139a-9) + 230z^8a \\ & + z^7(318a+12) + z^6(356a+26) + z^5(318a+12) + 230z^4a + z^3(139a-9) \\ & + z^2(69a-21) + z(24a-16) + 4a - 4, \end{aligned} \tag{4}$$

which we denote by E_i^a, $i = 1, \ldots, 12$, *are strange fixed points if* $E_i^a \neq -\frac{2}{3}$ *or*

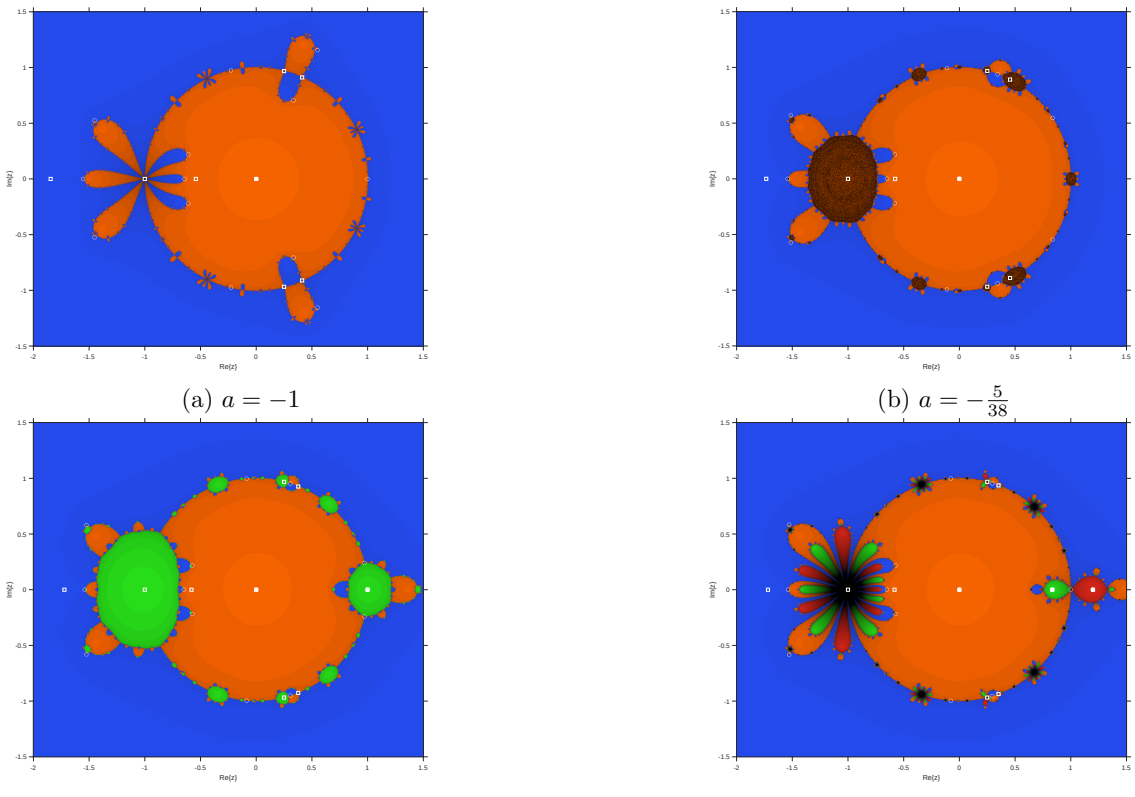
$$a \neq \frac{2 + 3E_i^a - 2(E_i^a)^2 - 3(E_i^a)^3 - 6(E_i^a)^5 - 4(E_i^a)^6}{(1 + E_i^a)^2(2 + 3E_i^a + 4(E_i^a)^2 + 6(E_i^a)^3 + 4(E_i^a)^4}.$$

In these cases, the denominator of O_a *vanishes.*

It is obvious that the points $z = 0$ and $z = \infty$ are fixed points that are superattractors. We now examine the stability of the point $z = 1$ as a function of a.

Theorem 4 *Let* $a \in \mathbb{C} \setminus \{1, -\frac{5}{38}\}$, *then the fixed point* $z = 1$ *is:*

1. *A superattractor if* $a = 0$.
2. *A repulsor if* $Re(a) \in \left(-\infty, -\frac{25}{1342}\right) \cup \left(\frac{25}{962}, \infty\right)$.
3. *A repulsor if* $Re(a) \in \left[-\frac{25}{1342}, \frac{25}{962}\right]$ *and* $|Im(a)| > \frac{\sqrt{625+9500Re(a)-1291004Re(a)^2}}{2\sqrt{322751}}$.
4. *Parabolic if* $Re(a) \in \left[-\frac{25}{1342}, \frac{25}{962}\right]$ *and* $|Im(a)| = \frac{\sqrt{625+9500Re(a)-1291004Re(a)^2}}{2\sqrt{322751}}$.
5. *Attractor in any other case.*



(a) $a = -1$ (b) $a = -\frac{5}{38}$

(c) $a = 0$ (d) $a = 0.03982$

Figure 1: Basins of attraction of methods associated with different values of a.

As regards the stability of the strange fixed points E_i^a, it has not been possible to carry out a theoretical study as a function of a, since these points are the roots of the twelfth-degree polynomial (4), which is symmetric and can be reduced to a sixth-degree polynomial, but this cannot be further reduced to a lower degree. We therefore carried out an in-depth study for specific values of a. In this text, we present the conclusions solely through a number of dynamical planes in Figure 1.

4 Parametric family for the multidimensional case

This section deals with the extension to the multidimensional case of the parametric family defined in (1) which, in the one-dimensional case, as prove in Section 2, exhibits a order of convergence of seven, increasing to order 8 when $a = -1$. Its iterative scheme is given by the expression:

$$FN_a^s: \quad \begin{cases} y_k = x_k - [\boldsymbol{F}'(x_k)]^{-1}\boldsymbol{F}(x_k), \\ z_k = y_k - [5\boldsymbol{F}'(y_k) - \boldsymbol{F}'(x_k)]^{-1}[\left(\boldsymbol{F}'(y_k) + 3\boldsymbol{F}'(x_k)\right)\left([\boldsymbol{F}'(x_k)]^{-1}\boldsymbol{F}(y_k)\right)], \\ x_{k+1} = z_k - \left[(a-2)\boldsymbol{F}'(y_k) + \boldsymbol{F}'(x_k)\right]^{-1}\left[(a\boldsymbol{F}'(x_k) - \boldsymbol{F}'(y_k))\left([\boldsymbol{F}'(x_k)]^{-1}\boldsymbol{F}(z_k)\right)\right]. \end{cases} \tag{5}$$

where $\boldsymbol{F}'(x_k)$ must be non-singular and $a \in \mathbb{R} \setminus \{1\}$.

The following result shows that the multidimensional iterative family retains the same order of convergence as the one-dimensional family given by (1).

Theorem 5 *Let $\boldsymbol{F} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ be a function that is sufficiently differentiable in an open, convex neighbourhood D of α, such that $\boldsymbol{F}(\alpha) = 0$. Assuming that the Jacobian matrix $\boldsymbol{F}'(x)$ is continuous and nonsingular at α, if $x_0 \in \mathbb{R}^n$ is sufficiently close to α, then the sequence $\{x_k\}_{k\geq 0}$ generated by the multidimensional family (5) converges to α with order 7 for all $a \in \mathbb{R} \setminus \{1, -1\}$, and order 8 for $a = -1$, being the error equation:*

$$e_{k+1} = \frac{1}{-1+a}\,(1+a)\,e_k^7\left(3\boldsymbol{C}_2^6 - 3\boldsymbol{C}_2^2\boldsymbol{C}_3\boldsymbol{C}_2^2 + \boldsymbol{C}_2^3\boldsymbol{C}_3\boldsymbol{C}_2\right) + \mathcal{O}(e_k^8), \tag{6}$$

where $\boldsymbol{\Gamma} = \boldsymbol{F}'(\alpha)$, $\boldsymbol{C}_k = \frac{1}{k!}\left[\boldsymbol{F}'(\alpha)\right]^{-1}\boldsymbol{F}_k(\alpha)$, $k = 2, 3, 4, \ldots$, $e_k = x_k - \alpha$ y $\boldsymbol{F}_k(\alpha)$ es la derivada k-ésima de $\boldsymbol{F}$ en α.

5 Numerical experiments for non-linear systems

This chapter presents a numerical comparison of the methods that make up the FN_a^s family, focusing on the cases $a = -1$ and $a = 2$.

The numerical tests are carried out using Matlab R2024b, with floating-point arithmetic configured to 5000 significant digits. Each method is applied iteratively, starting from an initial point x_0, until a maximum of 50 iterations is reached or the following stopping criterion is satisfied:

$$\|x_{k+1} - x_k\|_2 + \|\boldsymbol{F}(x_{k+1})\|_2 < 10^{-250}.$$

The results are presented in tables with the same seven columns, corresponding respectively to the value of the parameter a used, the value of $\|x_{k+1} - x_k\|_2$, the value of $\|\boldsymbol{F}(x_{k+1})\|_2$, the number of iterations performed, the execution time in seconds, and the calculated value of ACOC, defined in [1].

5.1 Academic example of large dimensions

Let us solve the following system of non-linear equations:

$$x_i = \sum_{j=1, j\neq i}^{n} e^{x_j}, \quad i = 1, \ldots, n, \tag{7}$$

where n is the size of the system to be solved. To compare the methods, we study their behaviour for $n = 10$, $n = 50$, $n = 100$, using, in all cases, the initial approximation $x_0 = (\frac{1}{9}, \ldots, \frac{1}{9})$. Table 1 shows the numerical results obtained.

Table 1: Results for the equation (7)

n	a	$\|x_{k+1} - x_k\|_2$	$\|\boldsymbol{F}(x_{k+1})\|_2$	Iteraciones	Tiempo (s)	ACOC
10	-1	2.07271e-568	5.56127e-4550	4	1.6875	8.000000
	2	9.2394e-355	5.84183e-2485	4	1.5312	6.999999
50	-1	4.63472e-568	6.77038e-4549	4	68.4635	8.000000
	2	2.06599e-354	7.11193e-2484	4	68.0938	6.999999
100	-1	6.55449e-568	1.93449e-4548	4	494.6562	8.000000
	2	2.92176e-354	2.03208e-2483	4	492.5781	6.999999
$\alpha \approx 0.5671432904 \cdot (1, \ldots, 1)^T$						

The results confirm that the family of parameterised methods performs well even on large-scale problems, producing highly accurate solutions to systems of a hundred equations within a reasonable time given the required tolerance and number of significant digits.

5.2 Traffic management

In this section, we use the parametric family to approximate the solution to a partial differential equation that models the evolution of traffic in cities as a function of time. Furthermore, we compare the numerical results with the 8th-order Wang-Yang-Qin method [2], denoted by WYQ^s.

In [3], the modelling of this problem can be found via the following initial value problem:

$$\begin{aligned} u_t(x,t) &= -v_{max}\left(1 - \frac{2u}{u_{max}}\right) \cdot u_x, \quad x \in (c,d),\ t > 0, \\ u(x,0) &= u_0(x), \quad x \in (c,d), \end{aligned}$$

where $u(x,t)$ is the vehicle density on the road $[a,b]$, $[c,d] \subset [a,b]$, $u_0(x)$ is the traffic density at the initial time of the study, and v_{max} and u_{max} are the maximum speed and density of the road, respectively.

In our example, for simplicity, we assume that $v_{max} = u_{max} = 1$, that the interval is $[0,1]$, that the traffic density at the initial time is $u(x,0) = 1 - x$ and that we wish to determine the traffic density at $t = 1$. Next, to solve the equation numerically, we discretise the domain, taking the nodes $x_i = i \cdot h$, $t_j = j \cdot k$, $i = 0, \ldots, 10$, $j = 0, \ldots, 20$ with $h = 0.1$ and $k = 0.05$.

Now, we take backward differences to approximate the time derivative, central differences to approximate the spatial derivative at the intermediate nodes, and forward and backward differences to approximate the spatial derivative at the first and last nodes respectively. Carrying out this process and denoting by $u_{i,j} = u(x_i, t_j)$, we have the following set of equations:

$$\begin{cases} u_{0,j} + \dfrac{v_{max}k}{h}\left(1 - \dfrac{2u_{0,j}}{u_{max}}\right)(u_{1,j} - u_{0,j}) - u_{0,j-1} = 0, \\ u_{i,j} + \dfrac{v_{max}k}{2h}\left(1 - \dfrac{2u_{i,j}}{u_{max}}\right)(u_{i+1,j} - u_{i-1,j}) - u_{i,j-1} = 0, \quad i = 1, \ldots, 9, \\ u_{10,j} + \dfrac{v_{max}k}{h}\left(1 - \dfrac{2u_{10,j}}{u_{max}}\right)(u_{10,j} - u_{9,j}) - u_{10,j-1} = 0, \end{cases} \tag{8}$$

for $j = 1, \ldots, 20$. Thus, we have, for each time t_j, a non-linear system of 11 unknowns and 11 equations; therefore, we can use the schemes to solve this system at each time step with a tolerance of 10^{-250} and using, as the initial point, the solution to the system of equations corresponding to the previous time step, except for the first system, where we use the initial condition.

In Table 2 we can see the numerical results. Except for the column referring to execution time, which indicates the total execution time, all columns represent the average of the results obtained by solving the 20 systems of equations presented. Furthermore, Figure 2 shows the solution obtained at the final time.

Table 2: Results for traffic management

Método	$\|x_{k+1} - x_k\|_2$	$\|\boldsymbol{F}(x_{k+1})\|_2$	Iteraciones	Tiempo (s)	ACOC
FN^s_{-1}	3.15538e-1068	7.75988e-5009	4	51.2392	7.999999
FN^s_2	2.72243e-685	2.06833e-4790	4	50.9425	6.999999
WYQ^s	4.19649e-926	8.77651e-5009	4	54.1345	7.999999

Both methods belonging to FN^s_a converge in exactly 4 iterations for each of the systems, exhibiting extremely low errors in both the norm of the difference between consecutive iterations and the residual norm. On the other hand, it is observed that the WYQ^s method is slower and has a higher error for the same ACOC, that is, when compared with our method for $a = -1$.

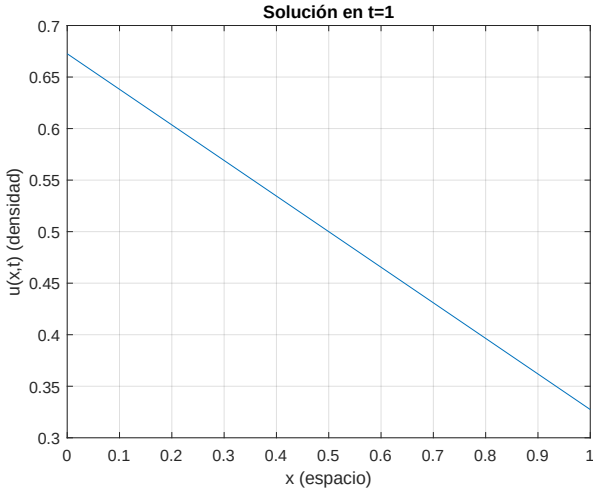


Figure 2: Solution to the traffic problem at the final time $t = 1$.

6 Conclusions

Throughout this work, a uniparametric family of iterative methods has been developed and analysed for solving non-linear equations and systems of equations, with the aim of improving stability, convergence rate, efficiency and applicability compared to existing classical methods. To this end, the scheme has been formulated, its order of convergence studied, and its stability analysed through an examination of its complex dynamics. Subsequently, in the case of non-linear systems of equations, the order of convergence has been theoretically proved, and numerical experiments have been carried out which corroborate its robustness compared to other known methods that exhibit the same order of convergence.

References

[1] A. Cordero and J. R. Torregrosa, "Variants of Newton's method using fifth-order quadrature formulas.," *Appl. Math. Comput.*, vol. 190, pp. 686–698, 2007.

[2] Y. Y. e. Y. Q. X. Wang, "Semilocal convergence analysis of an eighth order iterative method for solving nonlinear systems.," *AIMS Mathematics*, vol. 8, pp. 22 371–22 384, 2023.

[3] J. M. Lobo., *Trabajo Fin de Grado: Desarrollo de un modelo matemático para simular el flujo de tráfico en las ciudades.* Universidad Politécnica de Madrid, 2023.

Efficiency and Stability of New Quasi-Newton Hybrid Algorithms with Higher-Order Singh-Type Correctors for Unconstrained Optimization

Javier García Maimó[1], Miguel A. Leonardo Sepúlveda[2,3], Antmel Rodríguez Cabral[4], and Natanael Ureña Castillo[1,4,*]

[1] Ciencias Básicas y Ambientales (CBA), Instituto Tecnológico de Santo Domingo (INTEC), Avda. Los Próceres, Santo Domingo 10602, Dominican Republic; natanael.urena@intec.edu.do

[2] Departamento de Matemática, Instituto Superior de Formación Docente Salomé Ureña (ISFODOSU), Av. Caonabo, Santo Domingo 10114, Dominican Republic; miguel.leonardo@isfodosu.edu.do

[3] Departamento de Matemática, Universidad APEC (UNAPEC), Avenida Máximo Gómez No. 72, Sector El Vergel, Santo Domingo 10107, Dominican Republic; mleonardos@unapec.edu.do

[4] Escuela de Matemáticas, Universidad Autónoma de Santo Domingo (UASD), Ciudad Universitaria, Av. Alma Mater, Santo Domingo 10105, Dominican Republic; arodriguez05@uasd.edu.do

*Correspondence: natanael.urena@intec.edu.do

1 Introduction

Numerical optimization is a fundamental area of numerical analysis and applied mathematics because of its role in solving complex problems in science, engineering, economics, and machine learning. Within this area, unconstrained optimization is especially relevant, since many practical models can be written as the problem of minimizing a sufficiently smooth function without explicit constraints. Newtonian and quasi-Newtonian methods are among the most widely used strategies for this purpose, owing to their ability to approximate minimizers rapidly when the local geometry of the objective function is favorable [1].

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable, and let $\nabla f(x)$ and $\nabla^2 f(x)$ denote its gradient and Hessian, respectively. The classical Newton method generates a sequence $\{x^{(k)}\}$ by

$$d^{(k)} = -\left[\nabla^2 f\left(x^{(k)}\right)\right]^{-1} \nabla f\left(x^{(k)}\right), \qquad x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}, \tag{1}$$

where $d^{(k)}$ is the Newton direction and $\alpha^{(k)} > 0$ is a step length obtained through a line-search procedure. Although Newton's method is locally efficient, it is computationally demanding for large-scale problems, since it requires computing, storing, and inverting the Hessian matrix at each iteration.

Quasi-Newton methods reduce this cost by replacing the exact inverse Hessian with a sequence of matrix approximations. In general, they compute

$$\begin{aligned} d^{(k)} &= -H^{(k)} \nabla f\left(x^{(k)}\right), \qquad x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}, \\ s^{(k)} &= x^{(k+1)} - x^{(k)}, \qquad y^{(k)} = \nabla f\left(x^{(k+1)}\right) - \nabla f\left(x^{(k)}\right), \end{aligned} \tag{2}$$

where $H^{(k)}$ approximates $[\nabla^2 f(x^{(k)})]^{-1}$. Classical update formulas include Davidon–Fletcher–Powell (DFP) [2], Broyden–Fletcher–Goldfarb–Shanno (BFGS) [3], symmetric rank-one (SR1) [4], and Powell symmetric Broyden (PSB) [5]. These updates preserve secant information and avoid repeated exact Hessian inversions. However, for non-convex functions with several local minima, narrow valleys, or flat regions, their efficiency and robustness may decrease.

These limitations have led to the development of hybrid schemes that combine quasi-Newton updates with high-order correction mechanisms. In this paper, we propose a new family of quasi-Newton hybrid methods of convergence order four, called HOQN-4, in which a Newton-type predictor is coupled with a Sharma-type corrector defined by scalar weight functions. The correction stage is inspired by high-order schemes for nonlinear systems [6, 7, 8, 9, 10], while the optimization algorithm replaces the exact inverses of the Jacobian or Hessian with quasi-Newton approximations. This provides a flexible framework for designing efficient methods that maintain a low cost per iteration and exhibit competitive numerical and dynamic behavior.

2 Higher-Order Correctors and the HOQN-4 Algorithm

The proposed framework uses two line-search stages: one after the predictor and one after the high-order corrector. The line searches can be interpreted under standard Armijo–Wolfe assumptions, which guarantee sufficient decrease and suitable curvature information. In the numerical implementation, however, the step sizes are computed by bounded one-dimensional minimization using MATLAB's `fminbnd` routine over $[0, 10]$.

We first recall the high-order correction family that motivates the method. Let $F : \Omega \subset \mathbb{R}^m \to \mathbb{R}^m$ be sufficiently Fréchet differentiable in a neighborhood of a simple root ξ, with $F(\xi) = 0$ and $F'(\xi)$ nonsingular. Consider

$$\begin{cases} y^{(k)} = x^{(k)} - \left[F'\left(x^{(k)}\right)\right]^{-1} F\left(x^{(k)}\right), \\ x^{(k+1)} = y^{(k)} - \left[F'\left(y^{(k)}\right)\right]^{-1} \left(H\left(\nu_k\right) F\left(y^{(k)}\right) + G\left(\nu_k\right) F\left(x^{(k)}\right)\right), \end{cases} \tag{3}$$

where

$$\nu_k = \frac{\left\|F\left(y^{(k)}\right)\right\|^2}{\left\|F\left(x^{(k)}\right)\right\|^2}, \tag{4}$$

and H and G are analytic scalar weight functions near the origin. If

$$H(0) = 1, \qquad H'(0) = 1, \qquad G(0) = 0, \qquad G'(0) = 0, \tag{5}$$

then the exact correction scheme has local order five. In the multidimensional case, the operators arising in the local error equation are generally noncommutative, so product ordering must be preserved in any rigorous expansion. This exact result is the analytical starting point for the quasi-Newton hybridization developed here. The quasi-Newton part is controlled by the Dennis–Moré condition, which characterizes when inverse Hessian approximations are sufficiently accurate to preserve fast local convergence [11].

The general structure of one HOQN-4 iteration is summarized in Algorithm 1. The variable $z^{(k)}$ is used for the predictor point, while $y^{(k)}$ is reserved for the gradient difference in the quasi-Newton update.

Algorithm 1 HOQN-4 iteration with two line searches

Inputs: $x^{(k)}$, inverse Hessian approximation $H^{(k)}$, objective f, gradient ∇f

$g^{(k)} \leftarrow \nabla f\left(x^{(k)}\right)$

$d_N^{(k)} \leftarrow -H^{(k)} g^{(k)}$

$\alpha_N^{(k)} \leftarrow \arg\min_{\alpha \in (0,\alpha_{\max}]} f\left(x^{(k)} + \alpha d_N^{(k)}\right)$

$z^{(k)} \leftarrow x^{(k)} + \alpha_N^{(k)} d_N^{(k)}$

$g_z^{(k)} \leftarrow \nabla f\left(z^{(k)}\right)$

$\nu_k \leftarrow \|g_z^{(k)}\|^2 / \|g^{(k)}\|^2$

Compute $\widetilde{H}^{(k)} \approx \left[\nabla^2 f\left(z^{(k)}\right)\right]^{-1}$

$d_{HO}^{(k)} \leftarrow -\widetilde{H}^{(k)}\left(H(\nu_k) g_z^{(k)} + G(\nu_k) g^{(k)}\right)$

$\alpha_{HO}^{(k)} \leftarrow \arg\min_{\alpha \in (0,\alpha_{\max}]} f\left(z^{(k)} + \alpha d_{HO}^{(k)}\right)$

$x^{(k+1)} \leftarrow z^{(k)} + \alpha_{HO}^{(k)} d_{HO}^{(k)}$

$g^{(k+1)} \leftarrow \nabla f\left(x^{(k+1)}\right)$

$s^{(k)} \leftarrow x^{(k+1)} - x^{(k)}; \quad y^{(k)} \leftarrow g^{(k+1)} - g^{(k)}$

$H^{(k+1)} \leftarrow \text{DFP update}\left(H^{(k)}, s^{(k)}, y^{(k)}\right)$

Output: $x^{(k+1)}, H^{(k+1)}$

If an intermediate quasi–Newton update is incorporated, that is, if an additional update is performed after the Newton step in Algorithm 1, for instance using a BFGS update, the following variants are obtained: BA_1D, BA_2D and BA_3D. Under (5), simple polynomial weights are especially attractive in high-dimensional optimization, since they are inexpensive to evaluate and less prone to numerical instability than rational or transcendental alternatives. The following representative choices are considered:

$$\begin{aligned} &\text{(i)} \quad H(\nu) = 1 + \nu, \quad G(\nu) = 0, && \text{(ii)} \quad H(\nu) = 1 + \nu + \alpha\nu^2, \quad G(\nu) = 0, \\ &\text{(iii)} \quad H(\nu) = 1 + \nu, \quad G(\nu) = \gamma\nu^2, && \text{(iv)} \quad H(\nu) = 1 + \nu + \alpha\nu^2, \quad G(\nu) = \beta\nu^2. \end{aligned} \tag{6}$$

where α, β, and γ are real parameters.

2.1 First members of the high-order correction family

By particularizing $H(\nu_k)$ and $G(\nu_k)$ in (3), one obtains four members that can be embedded into the proposed hybrid quasi-Newton optimization methods. In the optimization setting, with $F(x) = \nabla f(x)$, they are written as

$$\begin{aligned} A_1: \quad & x^{(k+1)} = z^{(k)} - (1+\nu_k)\left[\nabla^2 f\left(z^{(k)}\right)\right]^{-1} \nabla f\left(z^{(k)}\right), \\ A_2: \quad & x^{(k+1)} = z^{(k)} - \left(1 + \nu_k + \alpha\nu_k^2\right)\left[\nabla^2 f\left(z^{(k)}\right)\right]^{-1} \nabla f\left(z^{(k)}\right), \\ A_3: \quad & x^{(k+1)} = z^{(k)} - \left[\nabla^2 f\left(z^{(k)}\right)\right]^{-1}\left((1+\nu_k)\nabla f\left(z^{(k)}\right) + \gamma\nu_k^2 \nabla f\left(x^{(k)}\right)\right), \\ A_4: \quad & x^{(k+1)} = z^{(k)} - \left[\nabla^2 f\left(z^{(k)}\right)\right]^{-1}\left((1+\nu_k+\alpha\nu_k^2)\nabla f\left(z^{(k)}\right) + \beta\nu_k^2 \nabla f\left(x^{(k)}\right)\right), \end{aligned}$$

where $\nu_k = \dfrac{\left\|\nabla f\left(z^{(k)}\right)\right\|^2}{\left\|\nabla f\left(x^{(k)}\right)\right\|^2}$. In practical HOQN-4 implementations, the exact inverse Hessian in these correctors is replaced by the quasi-Newton approximation $\widetilde{H}^{(k)}$.

3 Numerical tests

We compare the proposed variants, BA_iD, with the classical quasi-Newton methods BFGS, DFP, and SR1. All numerical experiments were carried out in MATLAB R2024. The iterative process was terminated when the gradient norm satisfied the stopping criterion $\left\|\nabla f\left(x^{(k)}\right)\right\| < 10^{-6}$. The test functions used in this work are the Booth function (f_B), the Himmelblau function (f_H), and the Styblinski–Tang function (f_{ST}) [12]:

- $f_B(x_1, x_2) = \left(x_1 + 2x_2 - 7\right)^2 + \left(2x_1 + x_2 - 5\right)^2$,
- $f_H(x_1, x_2) = \left(x_1^2 + x_2 - 11\right)^2 + \left(x_1 + x_2^2 - 7\right)^2$,
- $f_{ST}(x_1, x_2) = \frac{1}{2}\left(x_1^4 - 16x_1^2 + 5x_1 + x_2^4 - 16x_2^2 + 5x_2\right)$.

The reported order ρ is the approximate computational order of convergence estimated from the asymptotic error sequence.

Table 1: Numerical tests for the Booth function with two initial estimates.

Booth function (f_B), $\xi = (1,3)^T$				
$x^{(0)} = (-3.3, 1.99)^T$				
Method	Iter	ρ	$\|\nabla f(x^{(k+1)})\|$	CT (s)
BFGS	2	–	5.5495×10^{-14}	0.005531
DFP	2	–	5.8593×10^{-14}	0.001652
SR1	2	–	3.1776×10^{-14}	0.001878
BA_1D	1	–	5.7286×10^{-14}	0.000809
BA_2D	1	–	4.0194×10^{-14}	0.000850
BA_3D	2	–	1.0049×10^{-14}	0.000878
BA_4D	2	–	0.0000×10^{0}	0.000997
$x^{(0)} = (3.45, 1)^T$				
BFGS	2	–	2.8422×10^{-14}	0.000819
DFP	2	–	1.4211×10^{-14}	0.000781
SR1	2	–	4.9738×10^{-14}	0.000719
BA_1D	1	–	4.3221×10^{-14}	0.000833
BA_2D	1	–	5.6843×10^{-14}	0.000970
BA_3D	2	–	1.0049×10^{-14}	0.000864
BA_4D	2	–	0.0000×10^{0}	0.004021

Table 2: Numerical tests for the Himmelblau function with two initial estimates.

Himmelblau function (f_H), $\xi = (3,2)^T$				
$x^{(0)} = (2.55, 5)^T$				
Method	Iter	ρ	$\|\nabla f(x^{(k+1)})\|$	CT (s)
BFGS	6	1.168635	3.9843×10^{-12}	0.000564
DFP	6	1.173697	3.7157×10^{-12}	0.007404
SR1	6	1.163372	3.9498×10^{-12}	0.003250
BA_1D	3	3.775415	4.1768×10^{-12}	0.000912
BA_2D	3	3.775388	4.1669×10^{-12}	0.008120
BA_3D	3	3.891668	3.8373×10^{-11}	0.005558
BA_4D	3	3.891628	3.8273×10^{-11}	0.006511
$x^{(0)} = (2.35, 3.75)^T$				
BFGS	6	1.693516	1.7257×10^{-12}	0.002988
DFP	6	1.705731	1.7597×10^{-12}	0.003639
SR1	6	1.706193	1.7549×10^{-12}	0.006608
BA_1D	3	3.658757	1.6656×10^{-12}	0.006213
BA_2D	3	3.658791	1.6493×10^{-12}	0.003571
BA_3D	3	3.642714	1.8134×10^{-12}	0.004709
BA_4D	3	3.642736	1.8619×10^{-12}	0.003596

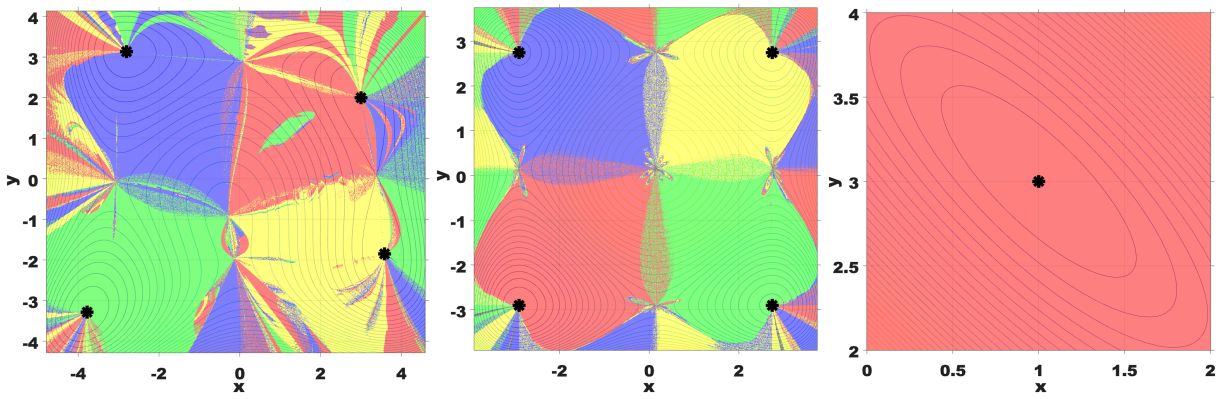
Table 3: Numerical tests for the Styblinski–Tang function with two initial estimates.

Styblinski–Tang function (f_{ST})				
$x^{(0)} = (2.35, 1.15)^T$, $\xi = (2.7468, 2.7468)^T$				
Method	Iter	ρ	$\|\nabla f(x^{(k+1)})\|$	CT (s)
BFGS	6	1.657788	9.7826×10^{-8}	0.004086
DFP	6	1.659922	1.7920×10^{-7}	0.003214
SR1	6	1.660730	1.0906×10^{-8}	0.004252
BA_1D	3	3.781905	7.0287×10^{-8}	0.004457
BA_2D	3	3.799194	2.0879×10^{-7}	0.003744
BA_3D	3	3.986936	1.1202×10^{-7}	0.004740
BA_4D	3	3.926117	5.7107×10^{-9}	0.004520
$x^{(0)} = (-3.3, 1.99)^T$, $\xi = (-2.9035, 2.7468)^T$				
BFGS	5	1.037820	2.9408×10^{-7}	0.001038
DFP	5	1.038411	9.4827×10^{-7}	0.008468
SR1	5	1.038439	3.3923×10^{-7}	0.003422
BA_1D	3	3.500874	6.5147×10^{-7}	0.002186
BA_2D	3	3.496297	3.2010×10^{-7}	0.003572
BA_3D	3	3.508161	7.2931×10^{-7}	0.007635
BA_4D	3	3.501838	2.1085×10^{-8}	0.004534

The numerical results show that the proposed hybrid variants, BA_iD, generally require fewer iterations than the classical quasi-Newton methods in each of the problems studied and often achieve smaller final gradient norms. Moreover, the obtained computational orders of convergence corroborate the expected theoretical behavior of the HOQN-4 scheme.

3.1 Dynamical planes

Dynamical planes complement the numerical tables by describing the sensitivity of each method with respect to the initial estimate. They are especially useful for non-convex functions, where the basin structure may be fragmented and several local minimizers may coexist.



(a) BA_2D. **(b)** BA_3D. **(c)** Booth function.

Figure 1: Dynamical planes of hybrid methods with two quasi-Newton updates for the Styblinski–Tang function.

For the non-convex cases, the hybrid methods display structured basins of attraction and competitive convergence regions. For strictly convex functions, the planes are more regular, as expected, and the methods exhibit a stable convergence pattern over a broad region of initial points. These observations support the numerical evidence that the HOQN-4 strategy is effective for both convex and non-convex test landscapes.

4 Conclusions

A new family of fourth-order quasi-Newton hybrid algorithms for unconstrained optimization has been proposed. The iterative scheme combines a Newton-type predictor, a Singh-type weighted corrector, two line-search stages, and quasi-Newton updates of the inverse Hessian approximation. The use of simple polynomial weight functions allows the derivation of practical corrector members A_1, A_2, A_3, and A_4, associated with an underlying family of fifth-order convergence. When incorporated into the correction stage, these members contribute to a significant improvement in the performance of classical quasi-Newton methods.

Preliminary numerical experiments show that the HOQN-4 variants achieve competitive performance compared with the classical BFGS, DFP, and SR1 schemes, often requiring fewer iterations and producing comparable or lower final gradient norms. Furthermore, the dynamical planes reveal a favorable structure of the attraction basins and suggest good stability with respect to the initial estimate, particularly when two quasi-Newton updates are incorporated. Future work will focus on rigorously extending the analysis to limited-memory implementations, studying adaptive strategies for selecting the weight functions, and validating the proposed framework in large-scale machine-learning and scientific computing problems.

Acknowledgments

This research was financed by Fondo Nacional de Innovación y Desarrollo Científico y Tecnológico (FONDOCyT) of Ministerio de Educación Superior Ciencia y Tecnología de la República Dominicana (MESCyT), grant number FONDOCyT 2023-1-1D2-0537.

References

[1] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York: Springer, 1999.

[2] R. Fletcher, "A new approach to variable metric algorithms," *The Computer Journal*, vol. 13, no. 3, pp. 317–322, 1970.

[3] C. G. Broyden, "The convergence of a class of double-rank minimization algorithms," *IMA Journal of Applied Mathematics*, vol. 6, no. 1, pp. 76–90, 1970.

[4] R. H. Byrd, H. F. Khalfan, and R. B. Schnabel, "Analysis of a symmetric rank-one trust region method," *SIAM Journal on Optimization*, vol. 6, no. 4, pp. 1025–1039, 1996.

[5] M. J. D. Powell, "A new algorithm for unconstrained optimization," in *Nonlinear Programming*, pp. 31–65, 1970.

[6] J. F. Traub, *Iterative Methods for the Solution of Equations*. Providence, RI: American Mathematical Society, 1982.

[7] C. Chun, "Construction of Newton-like iteration methods for solving nonlinear equations," *Numerische Mathematik*, vol. 104, pp. 297–315, 2006.

[8] A. M. Ostrowski, *Solution of Equations and Systems of Equations*. Amsterdam: Elsevier, 2016.

[9] A. Cordero, R. V. Rojas-Hiciano, J. R. Torregrosa, and M. P. Vassileva, "A highly efficient class of optimal fourth-order methods for solving nonlinear systems," *Numerical Algorithms*, vol. 95, no. 4, pp. 1879–1904, 2024.

[10] H. Singh, J. R. Sharma, and S. Kumar, "A simple yet efficient two-step fifth-order weighted Newton method for nonlinear models," *Numerical Algorithms*, vol. 93, no. 1, pp. 203–225, 2023.

[11] J. E. Dennis and J. J. Moré, "A characterization of superlinear convergence and its application to quasi-Newton methods," *Mathematics of Computation*, vol. 28, no. 126, pp. 549–560, 1974.

[12] J. M. Dieterich and B. Hartke, "Empirical review of standard benchmark functions using evolutionary global optimization," *arXiv preprint arXiv:1207.4318*, 2012.

A Weighted Generalization of Singh's Method for Nonlinear Systems with Application to a Stationary Burgers Equation

Miguel A. Leonardo Sepúlveda[♭,♮,1] Javier G. Maimó[♯], Antmel Rodríguez Cabral[♢], and Natanael Ureña Castillo[♯]

(♭) Departamento de Matemática, Instituto Superior de Formación Docente Salomé Ureña, Recinto Félix Evaristo Mejía, Av. Caonabo, Santo Domingo 10114, Dominican Republic.

(♮) Universidad APEC (UNAPEC), Avenida Máximo Gómez No. 72, Santo Domingo 10107, Dominican Republic.

(♯) Àrea de Ciencias Básicas y Ambientales, Instituto Tecnológico de Santo Domingo (INTEC), Av. Los Próceres, Santo Domingo 10602, Dominican Republic.

(♢) Escuela de Matemáticas, Universidad Autónoma de Santo Domingo (UASD), Ciudad Universitaria, Av. Alma Mater, Santo Domingo 10105, Dominican Republic.

1 Introduction

The numerical solution of nonlinear systems

$$F(x) = 0, \qquad F : \Omega \subset \mathbb{R}^m \to \mathbb{R}^m, \tag{1}$$

is a fundamental task in scientific computing, since many models arising in mechanics, engineering, transport phenomena, and applied sciences lead, after discretization, to algebraic equations of the form (1). Newton's method remains the classical reference scheme, but for medium- and large-scale problems its repeated Jacobian evaluations and matrix factorizations may represent a significant portion of the total computational cost [1, 2]. This motivates the design of multipoint procedures that preserve a high local convergence order while keeping the most expensive matrix operations under control.

Among the most effective design strategies, weight functions provide additional flexibility in the correction stage without modifying the general structure of the method. Instead of prescribing a fixed correction, one introduces scalar functions depending on an accelerator parameter, so that lower-order terms in the error equation can be cancelled and a family of admissible schemes can be generated. In the scalar setting this idea is well known, and several multidimensional extensions have confirmed that weighted corrections also constitute a useful tool for nonlinear systems [3, 4, 5].

In this contribution, we present a weighted family of fifth-order methods obtained as a generalization of the efficient Singh–Sharma scheme [6]. The proposed class retains the convergence order of the original method, recovers it as a particular case, and enlarges the admissible correction stage through simple conditions on the weight functions. The numerical part of this extended abstract is restricted to

[1]Corresponding author: mleonardos@unapec.edu.do

selected nonlinear systems, with special attention to large-scale benchmarks where the proposed family exhibits a clear computational advantage over Newton's method and several high-order competitors. In addition, the same framework can also be applied to a nonlinear ordinary differential equation of Burgers type, although those details are omitted here.

2 Weighted family and local convergence

Let $\xi \in \Omega$ be a simple zero of F, that is,

$$F(\xi) = 0, \qquad F'(\xi) \text{ is nonsingular.}$$

The weighted generalization of Singh's method is defined by

$$\begin{cases} y^{(k)} = x^{(k)} - \left[F'(x^{(k)})\right]^{-1} F(x^{(k)}), \\ x^{(k+1)} = y^{(k)} - \left[F'(y^{(k)})\right]^{-1} \Big(H(\nu_k) F(y^{(k)}) + G(\nu_k) F(x^{(k)}) \Big), \end{cases} \tag{2}$$

where

$$\nu_k = \frac{\|F(y^{(k)})\|_2^2}{\|F(x^{(k)})\|_2^2}, \tag{3}$$

and H and G are analytic scalar weight functions in a neighbourhood of the origin. The first step is Newton's predictor, whereas the second one introduces the weighted correction that extends Singh's original construction.

Theorem 2.1 *Let $F : \Omega \subset \mathbb{R}^m \to \mathbb{R}^m$ be sufficiently Fréchet differentiable in a neighbourhood of a simple zero $\xi \in \Omega$, that is,*

$$F(\xi) = 0, \qquad F'(\xi) \text{ is nonsingular.}$$

Consider the family (2)*, where H and G are analytic scalar weight functions in a neighbourhood of the origin. Let $e^{(k)} = x^{(k)} - \xi$ and define*

$$C_q = \frac{1}{q!} [F'(\xi)]^{-1} F^{(q)}(\xi), \qquad q \geq 2.$$

Assume that

$$H(0) = 1, \qquad H'(0) = 1, \qquad G(0) = 0, \qquad G'(0) = 0. \tag{4}$$

Then the family (2) *converges locally to ξ with order five. Moreover, its error equation is*

$$e^{(k+1)} = \left[\left(2 - \frac{1}{2} G''(0) \right) C_2^4 - 2 C_3 C_2^2 + P^{-1} Q C_2^3 \right] e^{(k)^5} + \mathcal{O}\left(e^{(k)^6} \right), \tag{5}$$

where

$$P = (F'(\xi))^T F'(\xi), \qquad Q = \sum_{i=1}^{m} (R_i^T G_i + G_i^T R_i),$$

with $R_i = f_i'(\xi)$ and $G_i = \frac{1}{2} f_i''(\xi)$, for $i = 1, \ldots, m$.

Consequently, the coefficients of orders $e^{(k)}$, $e^{(k)^2}$, $e^{(k)^3}$, and $e^{(k)^4}$ vanish under (4)*, and the first nonzero term appears at order five.*

The previous theorem is useful from both the analytical and computational viewpoints. On the one hand, it provides a simple set of verifiable conditions on H and G ensuring fifth-order convergence. On the other hand, it shows that the second derivative $G''(0)$ enters explicitly into the leading error constant, so different admissible weight functions may preserve the same order while exhibiting different asymptotic behaviour in practice. In particular, when $G(\nu) \equiv 0$ and $H(\nu) = 1 + \nu$, the family reduces to the original Singh–Sharma method [6].

2.1 Selected members of the family

The complete article includes several admissible choices of the weight functions. In the present extended abstract we retain the five representatives listed in Table 1, which were used throughout the numerical experiments.

Table 1: Representative members of the weighted family.

Method	$H(\nu)$	$G(\nu)$
NMA1	$1+\nu$	0
NMA2	$1+\nu+\alpha\nu^2$, $\alpha=\frac{1}{2}$	0
NMA3	$1+\nu$	$\gamma\nu^2$, $\gamma=-1$
NMA4	$1+\nu+\alpha\nu^2$, $\alpha=\frac{1}{2}$	$\beta\nu^2$, $\beta=\frac{1}{4}$
NMA5	$1+\nu+a\nu^3$, $a=\frac{1}{2}$	0

These five variants satisfy (4) and therefore share the same fifth-order local convergence. Their practical relevance lies in the fact that methods with identical formal order may still display different runtimes, residual histories, and sensitivity to the problem dimension.

3 Selected nonlinear systems and computational performance

The complete numerical study in the full paper contains nine nonlinear systems with dimensions ranging from 25 to 1499. Here we retain only the most representative cases, namely those that better illustrate the computational behaviour of the proposed family in comparison with Newton's method and with the high-order reference schemes Newton–Jarratt, Vassileva, SHM5, and O6.

Among the retained examples, two large cyclic models are especially demanding. The first one is the logarithmic system

$$\begin{cases} x_1+\log(2+x_1+x_2)=0,\\ x_2+\log(2+x_2+x_3)=0,\\ \quad\vdots\\ x_{998}+\log(2+x_{998}+x_{999})=0,\\ x_{999}+\log(2+x_{999}+x_1)=0, \end{cases} \tag{6}$$

whose solution is approximately $(-0.3149230578,\ldots,-0.3149230578)^T$. The second one is the exponential-product model

$$\begin{cases} x_1x_2-e^{-x_1}-e^{-x_2}=0,\\ x_2x_3-e^{-x_2}-e^{-x_3}=0,\\ \quad\vdots\\ x_{1498}x_{1499}-e^{-x_{1498}}-e^{-x_{1499}}=0,\\ x_{1499}x_1-e^{-x_{1499}}-e^{-x_1}=0, \end{cases} \tag{7}$$

with solution $(0.9012010317,\ldots,0.9012010317)^T$. Both systems are large, cyclic, and strongly coupled, which makes them appropriate benchmarks for assessing the efficiency of high-order methods.

Table 2 summarizes the most relevant timing data for the selected examples. In Examples 1–3, every member of the proposed family is faster than Newton and also faster than the other high-order competitors included in the comparison. In Example 4, the best variants of the proposed family remain

ahead of Newton and Vassileva. In Example 9, all five proposed members outperform the full set of competing methods in runtime.

Table 2: Selected benchmark systems from Examples 1, 2, 3, 4, and 9 in the article, emphasizing runtime advantages of the proposed family. Times are given in seconds.

Example in the article	n	Proposed runtime	Newton runtime	Best other high-order runtime
1	40	NMA range: 3.1067–3.2024	3.5744	Vassileva: 4.4415
2	30	NMA range: 1.6932–1.7588	2.0140	Vassileva: 2.4733
3	25	NMA range: 0.8422–0.8618	0.9769	Vassileva: 1.2044
4	999	Best NMA4: 49.0768	57.0281	Vassileva: 72.7727
9	1499	NMA range: 84.4928–102.1299	109.3066	Vassileva: 144.9235

The information in Table 2 shows that the weighted family is not only asymptotically correct but also practically effective. In the first three examples, the five proposed methods keep values of the approximated computational order close to five while requiring the lowest execution times in the comparison. This trend becomes even more relevant when competing schemes need one additional iteration, since the extra Jacobian evaluation and matrix factorization have an immediate impact on the total runtime.

A more detailed comparison on the two largest systems retained in this abstract is given in Table 3. In Example 4, NMA4 and NMA5 are the fastest proposed variants, whereas in Example 9 all five members of the family remain faster than Newton and clearly faster than Newton–Jarratt, Vassileva, SHM5, and O6. These results indicate that the freedom introduced by the weight functions may improve not only the asymptotic error constant but also the practical balance between accuracy and execution time.

Table 3: Detailed comparison on the two large-scale systems (6) and (7).

System	Method	Iter	$\|x^{(k+1)} - x^{(k)}\|$	$\|F(x^{(k+1)})\|$	e-time (s)	p
Example 4, $n = 999$	NMA4	3	1.78288×10^{-19}	1.16174×10^{-101}	49.0768	4.9278
	NMA5	3	5.57978×10^{-20}	2.31525×10^{-104}	50.0233	4.9622
	Newton	7	1.31659×10^{-51}	5.84264×10^{-104}	57.0281	2.0000
	Vassileva	3	2.04247×10^{-20}	2.30225×10^{-106}	72.7727	4.8779
Example 9, $n = 1499$	NMA1	3	7.71284×10^{-26}	1.68563×10^{-134}	84.4928	5.0030
	NMA3	3	2.12069×10^{-26}	1.98452×10^{-137}	85.7993	5.0116
	Newton	7	1.37639×10^{-73}	2.90608×10^{-148}	109.3066	2.0000
	Vassileva	3	3.53469×10^{-29}	8.49020×10^{-152}	144.9235	5.0142

From an algorithmic viewpoint, the observed behaviour can be understood through the correction stage in (2). The Newton predictor moves the iteration into the asymptotic regime, whereas the weighted correction incorporates the scalar information encoded in ν_k without changing the overall matrix structure of the method. As a result, the family preserves the same formal convergence order while allowing practical improvements in runtime and residual decay, particularly for large cyclic problems.

4 Brief remark on the differential application

Besides the nonlinear algebraic benchmarks considered above, the same weighted family can also be employed in the numerical treatment of a nonlinear ordinary differential equation associated with a stationary Burgers-type model. Since the purpose of the present extended abstract is to emphasize nonlinear systems and their computational performance, the differential formulation and its numerical details are omitted here.

5 Conclusions

The weighted generalization considered in this work preserves the fifth-order convergence of the Singh–Sharma method under the simple conditions (4), while enlarging the class of admissible corrections through the functions H and G. The resulting framework contains the original method as a particular case and, at the same time, provides enough freedom to build variants with different practical behaviour.

The numerical evidence reported here highlights three points. First, the proposed family is theoretically well founded, with a fifth-order local convergence result that is easy to verify through the values of the weight functions at the origin. Second, in several representative nonlinear systems, including large cyclic examples of dimensions 999 and 1499, the proposed variants are faster than Newton and also faster than the remaining high-order competitors considered in this study. Third, the flexibility introduced by the weight functions is relevant in practice, since methods of the same formal order may display noticeably different computational costs.

Overall, the proposed weighted family offers a useful balance between analytical simplicity, fifth-order convergence, and competitive CPU times for nonlinear systems. This makes it a promising alternative for large-scale algebraic problems and a suitable basis for further developments involving other families of weight functions and additional nonlinear models.

Acknowledgments

The authors gratefully acknowledge the institutional support received from ISFODOSU, UNAPEC, UASD and INTEC.

References

[1] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, 1970.

[2] C. T. Kelley, *Solving Nonlinear Equations with Newton's Method*, SIAM, 2003.

[3] S. Artidiello, A. Cordero, J. R. Torregrosa, and M. P. Vassileva, Multidimensional generalization of iterative methods for solving nonlinear problems by means of weight-function procedure, *Applied Mathematics and Computation*, 268 (2015), 1064–1071.

[4] J. R. Sharma and H. Arora, On efficient weighted-Newton methods for solving systems of nonlinear equations, *Applied Mathematics and Computation*, 222 (2013), 497–506.

[5] R. R. Capdevila, A. Cordero, and J. R. Torregrosa, A new three-step class of iterative methods for solving nonlinear systems, *Mathematics*, 7(12) (2019), 1221.

[6] H. Singh and J. R. Sharma, Simple and efficient fifth order solvers for systems of nonlinear problems, *Mathematical Modelling and Analysis*, 28(1) (2023), 1–22.

Generalized Traub family for nonlinear systems: Fourth-order optimal method and dynamical analysis

Miguel A. Leonardo Sepúlveda[♭,♯,1] Alicia Cordero[♮], Juan R. Torregrosa[♮], Antmel Rodríguez Cabral[⋄], y María P. Vassileva[♣]

(♭) Departamento de Matemática, Instituto Superior de Formación Docente Salomé Ureña (ISFODOSU)
Recinto Félix Evaristo Mejía, Santo Domingo, Dominican Republic.

(♯) Universidad APEC (UNAPEC)
Santo Domingo, Dominican Republic.

(♮) Instituto de Matemática Multidisciplinar, Universitat Politècnica de València
València, Spain.

(♣) Ciencias Básicas y Ambientales (CBA), Instituto Tecnológico de Santo Domingo (INTEC)
Santo Domingo, Dominican Republic.

(⋄) School of Mathematics, Autonomous University of Santo Domingo (UASD)
Santo Domingo, Dominican Republic.

1 Introduction

The numerical solution of nonlinear systems remains a central topic in scientific computing. Newton's method continues to be the basic benchmark because of its quadratic order and its balanced computational structure. However, its order is limited to two, and therefore there is a permanent interest in constructing higher-order schemes that preserve a favorable cost distribution between evaluations of the vector function F and its Jacobian F'.

In the scalar setting, the Kung–Traub conjecture provides a natural notion of optimality for multipoint methods without memory. Its analogue for systems was formulated by Cordero and Torregrosa, who established that if a method uses k_1 Jacobian evaluations and k_2 evaluations of the vector function per iteration, with $k_1 \leq k_2$, then its order cannot exceed

$$2^{k_1+k_2-1}.$$

Under this point of view, Newton's method is optimal of order two, and the next relevant target is to obtain fourth-order methods with the cost pair $(k_1, k_2) = (1, 2)$.

The article develops a new family, called *Generalized Traub family*, inspired by Traub-type predictor–corrector ideas and by weighted corrections based on norm quotients. The construction uses a single Jacobian matrix and two evaluations of F, while solving two linear systems sharing the same coefficient matrix. This feature lowers the linear algebra burden and makes the method

[1] mleonardos@unapec.edu.do

adaptable to large-scale settings. The central result is that the family is third order for generic values of the parameter γ, but the member corresponding to $\gamma = 2$ becomes fourth order and attains the Cordero–Torregrosa bound.

Besides the convergence theory, the paper includes a detailed real dynamical analysis of the associated parametric rational operator. This qualitative viewpoint helps describe the stability properties of the family, identify favorable regions of the parameter, and explain the behavior observed in the numerical experiments.

2 Description of the family and cost structure

Let $F : \mathcal{D} \subset \mathbb{R}^m \to \mathbb{R}^m$ be sufficiently smooth and let $\bar{x}$ be a simple zero of the nonlinear system

$$F(x) = 0.$$

The proposed family is defined by

$$\begin{cases} y^{(k)} = x^{(k)} - \left[F'(x^{(k)})\right]^{-1} F(x^{(k)}), \\ z^{(k)} = y^{(k)} - \left[F'(x^{(k)})\right]^{-1} F(y^{(k)}), \\ x^{(k+1)} = z^{(k)} + \gamma\, \mu_k \left(x^{(k)} - z^{(k)}\right), \end{cases} \tag{1}$$

where

$$\mu_k = \frac{\|F(z^{(k)})\|^2}{\|F(x^{(k)})\|^2}, \qquad \gamma \in \mathbb{R}.$$

The method uses exactly:

- one Jacobian evaluation,
- two evaluations of the vector function F,
- two linear solves with the same matrix $F'(x^{(k)})$.

This computational pattern is particularly useful because the LU factorization of the Jacobian matrix can be reused in both solves. Therefore, from an operational perspective, the family preserves a moderate cost while allowing a gain in convergence order.

The paper compares the new family with classical schemes such as Ostrowski-, Chun-, and Jarratt-type methods, emphasizing that not every scalar optimal extension remains optimal in the vectorial setting. This makes the construction of the new Traub-based family mathematically meaningful and computationally relevant.

3 Main convergence result

The local convergence analysis is based on Taylor expansions of the vector function and of the inverse Jacobian around a simple root $\bar{x}$. Let

$$\varepsilon^{(k)} = x^{(k)} - \bar{x}$$

be the error vector. The paper derives an error equation of the form

$$\varepsilon^{(k+1)} = (2-\gamma)D_2^2\big(\varepsilon^{(k)}\big)^3 + \Big((-9+12\gamma)D_2^3 + (7-4\gamma)D_2D_3 - 2\gamma D_3D_2\Big)\big(\varepsilon^{(k)}\big)^4 + \mathcal{O}\Big((\varepsilon^{(k)})^5\Big),$$

where D_2, D_3 are multilinear coefficients associated with derivatives of F at $\bar{x}$.

This directly implies the central theorem of the article.

Theorem 1 *For the family* (1)*:*

- *if* $\gamma \neq 2$*, the method has local order of convergence three;*
- *if* $\gamma = 2$*, the cubic term vanishes and the method has local order four.*

Moreover, since the cost pair is $(k_1, k_2) = (1, 2)$*, the member* $\gamma = 2$ *reaches the Cordero–Torregrosa upper bound and is therefore optimal.*

This result is one of the main achievements of the paper: it exhibits a fourth-order vectorial Traub-type method that is optimal in the sense of the systems conjecture, while keeping a favorable computational structure.

4 Dynamical analysis of the parametric family

A second major contribution of the paper is the multidimensional real dynamical study of the rational operator associated with the family. The qualitative analysis is performed by applying the method to low-degree polynomial systems and studying the resulting rational map.

The dynamical investigation includes:

- fixed points and strange fixed points,
- periodic points,
- stability classification via Jacobian eigenvalues,
- dynamical planes and parameter-dependent behavior.

This approach is valuable because it reveals information that pure local convergence theory cannot capture. In particular, the paper shows how the parameter γ influences stability, basin geometry, and the appearance of undesirable behaviors such as divergence or nonconvergent regions.

Figure 1 illustrates one representative dynamical portrait of the parametric family.

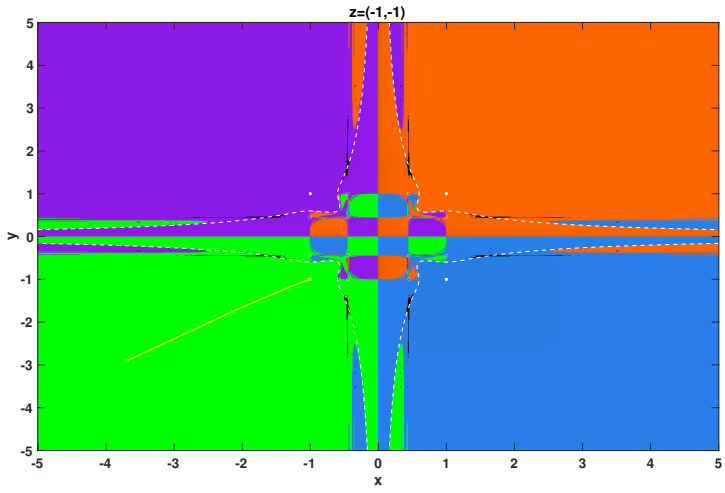


Figure 1: Representative dynamical plane for the Generalized Traub family. The figure illustrates the basin structure associated with different roots and the influence of the parameter on the global organization of the iteration.

The article emphasizes that the best-performing member, $\gamma = 2$, is not only optimal in order but also shows favorable stability features in the dynamical study. This reinforces the choice of that parameter from both the analytical and the practical viewpoints.

The dynamical study supports the analytical conclusion that the parameter value $\gamma = 2$ provides the most favorable balance between convergence order and stability.

5 Results

The proposed family is tested on benchmark nonlinear systems and on large-scale examples. The numerical evidence confirms the theoretical order predicted by the error equation and shows that the method behaves robustly in several scenarios.

The paper reports standard indicators such as:

- number of iterations,
- approximate computational order of convergence (ACOC),
- last iterate error,
- norm of the residual,
- reached solution.

Two representative experiments are especially noteworthy:

1. a nonlinear system involving exponential and trigonometric terms, with several exact solutions;
2. a n-dimensional system solved from a distant initial guess.

To illustrate the practical performance of the proposed family, we include one representative numerical experiment extracted from the full study. The goal is to show that the theoretical conclusions obtained in the local convergence analysis are consistent with the computational behavior observed in practice.

We consider the nonlinear system composed of m equations

$$f_i(x) = x_i - 1.5 \sin\left(\sum_{\substack{j=1 \\ j\neq i}}^{m} x_j\right) = 0, \qquad i = 1, 2, \ldots, m. \tag{2}$$

For the numerical experiment, we take $m = 20$ and the initial approximation

$$x^{(0)} = \left(\frac{1}{10}, \frac{1}{10}, \ldots, \frac{1}{10}\right)^T.$$

The iterative schemes converge, when successful, to the symmetric approximate solution

$$x^* \approx (0.15973, 0.15973, \ldots, 0.15973)^T.$$

Table 1: Representative numerical results for the nonlinear system (2) with $m = 20$.

Method	Iter	$\|x^{(k+1)} - x^{(k)}\|$	$\|F(x^{(k+1)})\|$	e-time	ACOC
Generalized Traub	6	1.8694×10^{-366}	$1.22054 \times 10^{-1826}$	0.778765	4.0
Generalized Ostrowski	5	2.39383×10^{-115}	6.20273×10^{-458}	9.006553	4.0
Chun's method	6	3.79955×10^{-234}	2.12822×10^{-932}	10.464925	4.0
Classical Ostrowski method	6	2.93871×10^{-250}	1.42037×10^{-997}	10.123275	4.0
Sharma's fourth-order method	7	7.28143×10^{-268}	$4.82567 \times 10^{-1007}$	1.691032	4.0
Jarratt–Cordero method	5	1.43332×10^{-160}	3.18463×10^{-799}	1.061132	5.0
Singh's method	5	4.22607×10^{-155}	1.14672×10^{-771}	0.607800	5.0
Vassileva's method	50	7.33782×10^{-2}	1.03868×10^{1}	7.748143	–
Newton's method	50	2.42758×10^{1}	1.09698×10^{2}	3.169261	–

The following table summarizes the behavior of the proposed Generalized Traub method and several classical competitors. The reported quantities are the number of iterations, the norm of the difference between two consecutive iterates, the residual norm, the execution time, and the approximate computational order of convergence (ACOC).

The data in Table 1 show that the proposed Generalized Traub method is highly competitive. In this example, it converges in only six iterations, reaches an extremely small residual norm, and does so with a moderate execution time. Although some fifth-order methods also perform well, the proposed scheme attains a very favorable balance between order, stability, and computational cost. In contrast, Newton's method and Vassileva's method fail to achieve satisfactory convergence within the prescribed number of iterations, which further highlights the reliability of the proposed family. The results also show that the parameter choice $\gamma = 2$ consistently yields fourth-order convergence and often reduces the number of iterations compared with classical fourth-order competitors. In some cases, the family also exhibits greater sensitivity to the initial estimate.

Table 2 summarizes the main theoretical and numerical conclusions.

Table 2: Summary of the main results for the Generalized Traub family.

Parameter choice	Order	Optimality	Main qualitative feature
General γ	3	No	Competitive cubic member
$\gamma = 2$	4	Yes	Best balance between order and cost
Dynamical analysis	—	—	Stability improves around preferred parameter values
Numerical tests	—	—	Strong performance on benchmark and large systems

The paper also compares efficiency indices and computational behavior. The charts extracted from the article illustrate that the optimal member enjoys a favorable balance between theoretical efficiency and practical performance.

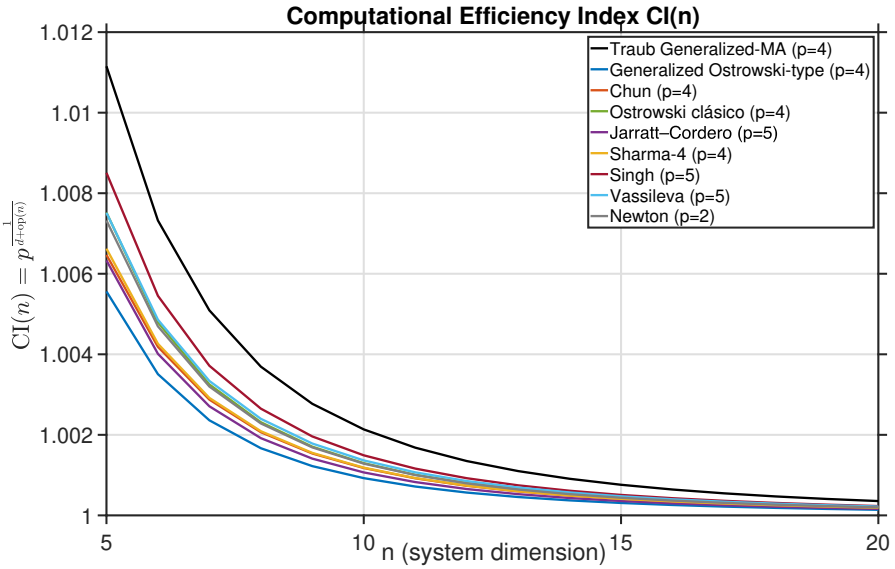


Figure 2: Efficiency comparison.

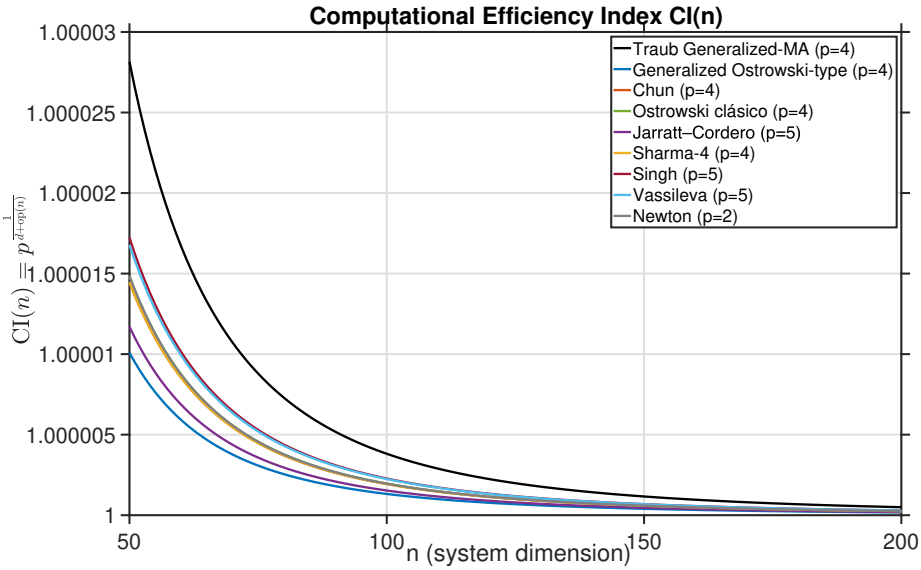


Figure 3: Additional computational comparison.

6 Most relevant contributions

The article provides several important contributions to the theory of iterative methods for nonlinear systems:

- It constructs a new Traub-type vectorial family with a simple and efficient predictor–corrector structure.
- It proves that the member $\gamma = 2$ is fourth order and optimal under the Cordero–Torregrosa conjecture.
- It develops a detailed error analysis based on multilinear expansions.
- It includes a multidimensional real dynamical study of the associated rational operator.
- It confirms, through numerical experiments, that the proposed optimal member is competitive with classical methods of equal order.

These points make the contribution significant both from the theoretical side and from the viewpoint of practical computation.

7 Conclusions

The paper shows that the Generalized Traub family is a mathematically sound and computationally effective contribution to the literature on nonlinear systems. The family is simple to implement, exploits a single Jacobian matrix efficiently, and contains an optimal fourth-order member when $\gamma = 2$.

From a broader perspective, the article highlights the usefulness of combining three complementary layers of analysis:

1. local convergence theory,
2. multidimensional real dynamics,
3. numerical experimentation.

This combination yields a more complete understanding of the proposed methods, since it not only certifies the theoretical order but also clarifies stability, qualitative behavior and practical robustness.

In summary, the most important message of the work is that the parameter $\gamma = 2$ produces a fourth-order optimal Generalized Traub method that is both theoretically justified and numerically competitive. The paper therefore opens the door to further developments, including new weighted variants, broader dynamical studies, and extensions to derivative-free or memory-based settings.

Acknowledgments

This extended abstract is based on the results presented in the paper *Generalized Traub family for nonlinear systems: Fourth-order optimal method and dynamical analysis.*

References

[1] A. AuthorSurname, *Title of the book*. Editorial, 2026.

[2] A. AuthorSurname, "Title of the article," *Name of the Journal*, vol. 1, pp. 1–15, 2026.

[3] R. Behl, A. Cordero, J. R. Torregrosa, and S. Bhalla, "A new high-order jacobian-free iterative method with memory for solving nonlinear systems," *Mathematics*, vol. 9, no. 17, 2021, ISSN: 2227-7390. DOI: 10.3390/math9172122. [Online]. Available: https://www.mdpi.com/2227-7390/9/17/2122.

[4] X. Wang and X. Chen, "Derivative-free kurchatov-type accelerating iterative method for solving nonlinear systems: Dynamics and applications," *Fractal and Fractional*, vol. 6, no. 2, p. 59, 2022.

[5] J. Steffensen, "Remarks on iteration.," *Scandinavian Actuarial Journal*, vol. 1933, no. 1, pp. 64–72, 1933.

[6] J. F. Traub, *Iterative methods for the solution of equations*. American Mathematical Soc., 1982, vol. 312.

[7] J. F. Traub, *Iterative methods for the solution of equations*. American Mathematical Soc., 1982, vol. 312.

[8] I. Argyros, S. Shakhno, and Y. Shunkin, "On an iterative moser-kurchatov method for solving systems of nonlinear equations," *Matematychni Studii*, vol. 63, no. 1, pp. 88–97, 2025.

[9] S. Wang, H. Xian, T. Liu, and S. Shateyi, "Solving nonlinear equation systems via a steffensen-type higher-order method with memory," *Mathematics*, vol. 12, no. 23, p. 3655, 2024.

[10] A. Cordero and J. R. Torregrosa, "Variants of newton's method using fifth-order quadrature formulas," *Applied Mathematics and Computation*, vol. 190, no. 1, pp. 686–698, 2007.

[11] A. Ostrowski, "Solution of equations and systems of equations, acad," *Press, New York*, 1966.

[12] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*. SIAM, 2000.

[13] H. Singh, J. R. Sharma, and S. Kumar, "A simple yet efficient two-step fifth-order weighted-newton method for nonlinear models," *Numerical Algorithms*, vol. 93, no. 1, pp. 203–225, 2023.

[14] J. R. Sharma, R. K. Guha, and R. Sharma, "An efficient fourth order weighted-newton method for systems of nonlinear equations," *Numerical Algorithms*, vol. 62, no. 2, pp. 307–323, 2013.

[15] V. Arroyo, A. Cordero, J. R. Torregrosa, and M. Vassileva, *Artificial satellites preliminary orbit determination by modified high-order gauss methods (cmmse10)*, 2010.

[16] A. Cordero, J. L. Hueso, E. Martínez, and J. R. Torregrosa, "A modified newton-jarratt's composition," *Numerical Algorithms*, vol. 55, no. 1, pp. 87–99, 2010.

[17] Y. H. Geum, Y. I. Kim, and B. Neta, "A sixth-order family of three-point modified newton-like multiple-root finders and the dynamics behind their extraneous fixed points," *Applied Mathematics and Computation*, vol. 283, pp. 120–140, 2016.

[18] R. R. Clark, "An introduction to dynamical systems: Continuous and discrete," *American Mathematical Soc*, 2012.

[19] R. R. Capdevila, A. Cordero, and J. R. Torregrosa, "A new three-step class of iterative methods for solving nonlinear systems," *Mathematics*, vol. 7, no. 12, p. 1221, 2019.

[20] R. Devaney, *An introduction to chaotic dynamical systems*. CRC press, 2018.

[21] J. Gaston, "Mémoire sur l'iteration des fonctions rationnelles," *J. Mat. Pur. Appl*, vol. 8, pp. 47–245, 1918.

[22] P. Fatou, "Sur les équations fonctionnelles premier mémoire," *Bulletin de la Société mathématique de France*, vol. 47, pp. 161–271, 1919.

[23] P. Fatou, "Sur les équations fonctionnelles troisieme mémoire," *Bulletin de la Société mathématique de France*, vol. 48, pp. 208–314, 1920.

A Highly Efficient $2m$-th Order Recursive Scheme for Nonlinear Systems under Increasing Uncertainty and Nonlinearity

Alicia Cordero[♭], Renso V. Rojas-Hiciano[♮] [1] and Juan R. Torregrosa[♭]

(♭) Multidisciplinary Mathematics Institute, Universitat Politècnica de València, Camí de Vera s/n, València, Spain.

(♮) Escuela de Ciencias Naturales y Exactas, Pontificia Universidad Católica Madre y Maestra, Sarasota Ave. 10109, Santo Domingo, R.D.

1 Introduction

Modeling problems in Economics, Physics, Chemistry, Engineering, and Life Sciences induces the need to solve systems of nonlinear equations of the form $F(x) = 0$, where $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$, with $n \in \mathbb{N}$, is a function defined on an open convex domain D.

Iterative methods are often the tools used to solve such problems. The most commonly used is Newton's scheme, which we denote by N2S:

$$x^{(k+1)} = x^{(k)} - F'\left(x^{(k)}\right)^{-1} F\left(x^{(k)}\right), \, k = 0, 1, \ldots, \tag{1}$$

where $F'(x) \in \mathscr{L}\left(\mathbb{R}^n, \mathbb{R}^n\right)$ is a linear operator representing the Jacobian matrix of F.

We can analyze and compare different iterative methods using the order of convergence, which indicates how quickly the sequence $\{x^{(k)}\}_{k\geq 0}$ generated by the iterative method approaches the root ξ. We take $e^{(k)} = x^{(k)} - \xi$ as the measure of the local error at the k^{th} iteration. The order of convergence p of an iterative method can be obtained from the error equation

$$e^{(k+1)} = Le^{(k)^p} + O\left(e^{(k)^{p+1}}\right), \, k = 0, 1, \ldots,$$

where $L \in \mathscr{L}(\overbrace{\mathbb{R}^n \cdots \times \mathbb{R}^n}^{p- \text{ times}}, \mathbb{R}^n)$ is a p linear operator and $\mathscr{L}$ is the set of bounded linear operators. In fact, N2S is a quadratic convergence order scheme.

Many recent studies focus on improving the efficiency of Newton-type methods, without drastically increasing the computational cost. For example, Wang in [1] presented a three-step eighth-order scheme (W8S):

$$\begin{aligned} v^{(k)} &= y^{(k)} - \left[I + \left(I + \frac{5}{4}E^{(k)}\right)E^{(k)}\right][F'(x^{(k)})]^{-1}F(y^{(k)}), \\ x^{(k+1)} &= v^{(k)} - \left[I + \left(I + \frac{3}{2}E^{(k)}\right)E^{(k)}\right][F'(x^{(k)})]^{-1}F(v^{(k)}), \end{aligned} \tag{2}$$

where $E^{(k)} = F'(x^{(k)})^{-1}(F'(x^{(k)}) - F'(y^{(k)}))$ and $y^{(k)}$ is the iteration of N2S.

[1] rv.rojas@ce.pucmm.edu.do

Singh et al. in [2] suggested a fifth-order scheme (S5S) using a scalar weight function $\gamma_k = 1 + \frac{\|F(z^{(k)})\|^2}{\|F(x^{(k)})\|^2}$. Based on this, Zhanlav and Otgondorj in [3] proposed a recursive method of order $3m-1$. For $m=3$ steps and $z_1^{(k)}$ as the iteration of N2S, they obtained the ZO8S scheme:

$$\begin{aligned} z_2^{(k)} &= z_1^{(k)} - \gamma_k [F'(z_1^{(k)})]^{-1} F(z_1^{(k)}) k = 0, 1, \ldots, & (3) \\ x^{(k+1)} &= z_2^{(k)} - (2\gamma_k - 1)\,[F'(z_1^{(k)})]^{-1} F(z_2^{(k)}). & (4) \end{aligned}$$

Erfanifar-Hajarian in [4] developed a multi-step scheme with a frozen Jacobian matrix and first step $w_1^{(k)}$ is the iteration of N2S. Its order is $2m+1$. We denote its ninth-order version as EH9S:

$$\begin{aligned} w_{i+1}^{(k)} &= w_i^{(k)} - 2[F'(x^{(k)})]^{-1}[F(w_i^{(k)} + [F'(x^{(k)})]^{-1} F(w_i^{(k)})) & (5) \\ &\quad + \frac{3}{2} F(w_i^{(k)} - [F'(x^{(k)})]^{-1} F(w_i^{(k)})) - \frac{3}{2} F(w_i^{(k)})], k = 0, 1, \ldots,\ i = 1, 2, 3, & (6) \\ x^{(k+1)} &= w_4^{(k)}. & (7) \end{aligned}$$

Cordero and Torregrosa proposed a generalization of the Kung-Traub bound for non-linear systems in [5]: *'An iterative scheme using k_1 evaluations of the Jacobian matrix and k_2 of the non-linear function is optimal in the vectorial sense if its order reaches $p = 2^{k_1+k_2-1}$, with $k_2 \geq k_1$.'*

Recently, Cordero et al. presented in [6] the first family of optimal vector schemes of higher order than Newton's in the literature, demonstrating robustness and efficiency in solving first-order partial differential equations (PDEs) [7].

In pursuit of optimality, Zhanlav and Otgondorj developed the ESS8 scheme in [8]. This eighth-order optimal method follows the following architecture:

$$\begin{aligned} z_k &= y_k - \bar{\tau}_k \cdot [F'(x_k)]^{-1} F(y_k), \\ x_{k+1} &= z_k - \alpha_k \cdot [F'(x_k)]^{-1} F(z_k), \end{aligned} \tag{8}$$

where $y^{(k)}$ is the iteration of N2S and the weight functions $\bar{\tau}_k$ and α_k use component-wise vector operations: $u/v = [u_1/v_1, \ldots, u_n/v_n]^T$, $\quad u \cdot v = [u_1 v_1, \ldots, u_n v_n]^T$.

Specifically, for $\Theta_k = F(y_k)/F(x_k)$, $t_k = F(z_k)/F(x_k)$, and $s_k = F(z_k)/F(y_k)$, the functions are:

$$\bar{\tau}_k = \frac{\mathbf{1}}{\mathbf{1} - 2\Theta_k}, \quad \alpha_k = \left(\frac{\mathbf{1} - \Theta_k}{\mathbf{1} - 2\Theta_k}\right)^2 \left(\frac{\mathbf{1} + t_k}{(\mathbf{1} - t_k)(\mathbf{1} - s_k)}\right). \tag{9}$$

To compare these methods, we use several metrics. The vectorial extension of Ostrowski's efficiency index is $I = p^{1/d}$, where $= d = k_1 n^2 + k_2 n$. We also use the computational efficiency index [9] $CI = p^{1/(d+op)}$. The necessary number of product-quotients to calculate an iterate, including total effort expended in solving the linear systems, is denoted by *op*. Recently, the authors proposed new performance indicators in [10] and [11]:

$$E_\Phi = 1 - \frac{\ln 2}{\ln p} \frac{d}{n(n+1)}, \quad T_\Phi = 1 - \frac{\ln 2}{\ln p} \frac{3\ op}{n^3 + 3n^2 - n}, \quad C_\Phi = \frac{1}{2}(E_\Phi + T_\Phi), \tag{10}$$

that are used in Section 4 to compare the proposed scheme with the existing ones.

Finally, we define a method as Least-Evaluative Cost (LEC) if the amount of evaluated Jacobian matrices in the iterative expression $k_1 = 1$, it uses only scalar weight functions, and the times F is evaluated, k_2, equals the number of steps. If two methods have the same order, the LEC scheme is the most efficient one.

2 Design and analysis of new LEC weight methods

We propose a new simple and LEC iterative scheme with an arbitrary $m \geq 2$ number of steps :

$$\begin{aligned} \mathcal{U}_i^{(k)} &= \mathcal{U}_{i-1}^{(k)} - \beta_k [F'(x^{(k)})]^{-1} F(\mathcal{U}_{i-1}^{(k)}),\ i = 2, \ldots, m, \\ x^{(k+1)} &= \mathcal{U}_m, \end{aligned} \tag{11}$$

where $\mathcal{U}_1^{(k)}$ is Newton step, and $\beta_k = 1 + 2\frac{F(x^{(k)})^T F(\mathcal{U}_1^{(k)})}{\|F(x^{(k)})\|_2^2}$, which we call Super-Traub. Its order of convergence is twice the number of steps, as shown below.

Theorem 1. *Let $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a sufficiently differentiable function on a convex neighborhood D of $\bar{x}$, solution of the non-linear system $F(x) = 0$. Suppose $F'(x)$ is non-singular in $\bar{x}$ and $x^{(0)}$ is an initial estimate close to $\bar{x}$. Then, the sequence $\{x^{(k)}\}_{k \geqslant 0}$ obtained by (11) with $m \geq 2$ steps, converges to $\bar{x}$ with convergence order $2m$ and error equation*

$$e^{(k+1)} = \left(M_3 + 2P^{-1}\Lambda M_2\right)^{m-2} \left(M_3 + 2P^{-1}\Lambda M_2 - M_2^2\right) M_2 e^{(k)^{2m}} + O\left(e^{(k)^{2m+1}}\right), \tag{12}$$

where $e^{(k)} = x^{(k)} - \bar{x}$, $M_i = \frac{1}{i!}\left[F'(\bar{x})\right]^{-1} F^{(i)}(\bar{x})$, $P = \sum_{i=1}^{n} R_i^T R_i$, $\Lambda = \sum_{i=1}^{n} R_i^T H_i$, being $R_i = \nabla f_i(\bar{x}) = \left(\frac{\partial f_i}{\partial x_1}(\bar{x}), \frac{\partial f_i}{\partial x_2}(\bar{x}), \ldots, \frac{\partial f_i}{\partial x_n}(\bar{x})\right)$ and $H_i = Hf_i(\bar{x})$ the Hessian matrix associated to the coordinate function of F, f_i with $i = 1, 2, \ldots, n$.

Theorem 1 allows us to design iterative schemes that reach very high order. This without incurring excessive computational costs, with respect to very efficient schemes existing in the literature.

New LEC Eighth-Order Scheme

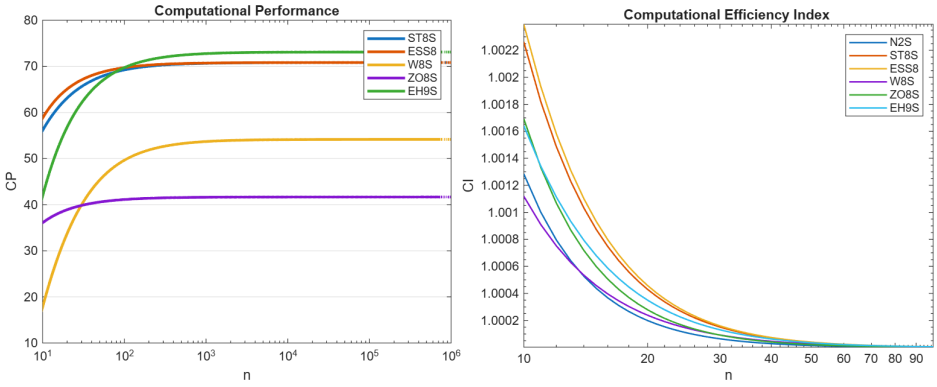
The present paper puts forward a novel method, named ST8S, which is based on the damped Super-Traub scheme with $m = 4$. This scheme, by Theorem 1, has order of convergence 8 and is characterised by its minimal evaluation and computational cost. In the subsequent section, it is demonstrated that this approach exhibits superior evaluative and computational efficiency and performance in comparison to recent known methods with the same or higher convergence order.

3 Efficiency Analysis

The computational efficiency of iterative methods is inherently linked to their operational cost. Indeed, the computational efficiency of schemes of the same order is higher in the one with the lowest cost. As shown in Figure 1 (derived from the data in Table 1), the new schemes achieve a superior balance between the convergence order p_j and the functional/arithmetic effort (d_j and

Table 1: Computational Permance and Efficiency Index Comparison

j	***Method***	$\boldsymbol{p_j}$	$\boldsymbol{d_j}$	$\boldsymbol{op_j}$	***CP***	***CI***
1	N2S	2	n^2+n	$\frac{1}{3}(n^3-n)+n^2$	$1-\frac{1}{n\,p_1}\left(\frac{d_1}{n+1}+\frac{3\,op_1}{n^2+3n-1}\right)$	$2^{\frac{1}{d_1+op_1}}$
2	ST8S	8	n^2+4n	$\frac{1}{3}(n^3-n)+4n^2+5n$	$1-\frac{1}{n\,p_2}\left(\frac{d_2}{n+1}+\frac{3\,op_2}{n^2+3n-1}\right)$	$8^{\frac{1}{d_2+op_2}}$
3	ESS8	8	n^2+3n	$\frac{1}{3}(n^3-n)+3n^2+11n$	$1-\frac{1}{n\,p_3}\left(\frac{d_3}{n+1}+\frac{3\,op_3}{n^2+3n-1}\right)$	$8^{\frac{1}{d_3+op_3}}$
4	W8S	8	$2n^2+3n$	$\frac{1}{3}(n^3-n)+13n^2$	$1-\frac{1}{n\,p_4}\left(\frac{d_4}{n+1}+\frac{3\,op_4}{n^2+3n-1}\right)$	$8^{\frac{1}{d_4+op_4}}$
5	ZO8S	8	$2n^2+3n$	$\frac{2}{3}(n^3-n)+3n^2+4n$	$1-\frac{1}{n\,p_5}\left(\frac{d_5}{n+1}+\frac{3\,op_5}{n^2+3n-1}\right)$	$8^{\frac{1}{d_5+op_5}}$
6	EH9S	9	n^2+10n	$\frac{1}{3}(n^3-n)+7n^2+11n$	$1-\frac{1}{n\,p_6}\left(\frac{d_6}{n+1}+\frac{3\,op_6}{n^2+3n-1}\right)$	$9^{\frac{1}{d_6+op_6}}$



(a) Computational Performance (%) up to n=1,000,000

(b) Computational Index from n=10 to 100

Figure 1: Computational Performance and Efficiency Comparison

op_j). Consequently, the Computational Performance (CP) and Efficiency Index (CI) of our recursive proposal not only compete with but often outperform other high-order schemes in the literature.

The new ST8S scheme exhibits high computational performance, well above 70%, outperforming W8S, which has just over 60%, and ZO8S which is around 50%, and its computational strength of the new ST8S scheme which, similarly to optimal ESS8 of order 8, maintains a computational index always higher than the rest of the schemes, as shown in Figure 1.

4 Basins of attraction

The present section investigates the stability of the novel method, ST8S, when applied to the solution of a standard non-linear problem. In order to analyse the basins of attraction associated with the proposed method, it is necessary to study the non-linear system $\{x_1 = 1, x_2 = 1\}$, whose solution set consists of four ordered pairs, namely $\{(1,1)^T, (1,-1)^T, (-1,1)^T, (-1,-1)^T\}$.

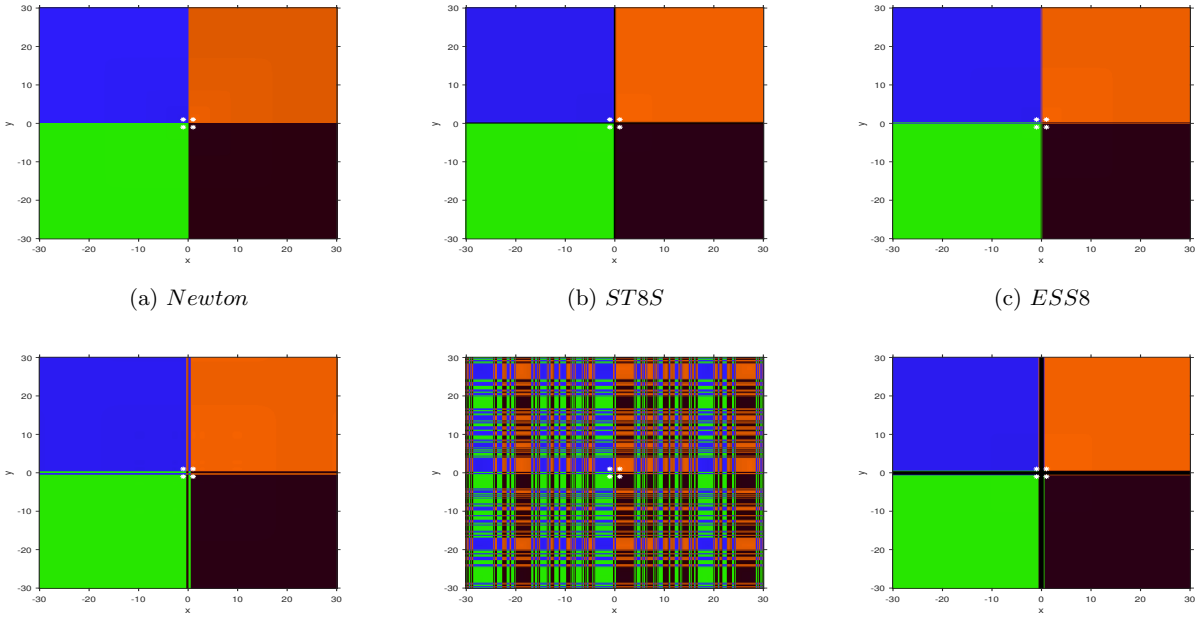
We analyze the basins of attraction for the ST8S method and its competitors by mapping convergence regions over a mesh of initial guesses. Points converging to a specific root are color-coded (blue, orange, green, or purple), while non-convergent points are marked in black. Figure2 illustrate these basins across three increasingly wider regions. White dots represent the system solutions, allowing for a direct comparison of stability and the extent of the attraction domains as the distance from the roots increases.

As shown in Figure 2, it is evident that the ST8S convergence basins for each solution are independent (one connected component per basin), while those of ESS8, W8S, ZO8S, and EH9S are not. It is evident that the proposed scheme provides extensive coverage, thus competing with other methods that have been proven to be highly stable.

5 Numerical Performance

We evaluate the numerical performance of the proposed ST8S method in comparison with the N2S (Newton), W8S, ZO8S, EH9S, and ESS8 schemes. The experimental analysis is structured into two stages:

First, we examine the sensitivity of each method to initial guesses by solving the 1D Bratu problem. This assessment is conducted using double-precision floating-point arithmetic to determine the practical reliability of the schemes. Second, we perform a rigorous verification of the theoretical convergence order through high-precision arithmetic, ensuring that the asymptotic behavior of the recursive architecture is numerically consistent with the established theory.



(a) *Newton* (b) *ST8S* (c) *ESS8*

(d) *W8S* (e) *ZO8S* (f) *EH9S*

Figure 2: *Attraction basins for quadratic systems in a* 60×60 *region.* The plots illustrate the stability of: (a) Newton, (b) ST8S, (c) ESS8, (d) W8S, (e) ZO8S, and (f) EH9S. Pure colors represent stable convergence to the roots, while the fractal-like patterns in (d), (e) and (f) indicate higher sensitivity to initial conditions compared to the proposed ST8S.

All calculations were performed in Matlab R2022b. For the high-precision tests, we used a mantissa of 1000 digits to minimize rounding errors. Our stopping criterion is $\|x^{(k)} - x^{(k-1)}\| + \|F(x^{(k)})\| < tol$ or a maximum of 50 iterations. In the elastic string problem, we set $tol = 10^{-40}$ for a system of size 500. In this work, we use $\breve{\rho} = \frac{\ln|x^{(k+1)}-x^{(k)}|}{\ln|x^{(k)}-x^{(k-1)}|}$. as an approximation of the computational order of convergence p.

5.1 The 1D Bratu Problem: Stability and Robustness Analysis

The 1D Bratu problem is a classic non-linear boundary value problem defined as:

$$u''(x) + \lambda e^{u(x)} = 0, \quad x \in [0, 1], \tag{13}$$

with boundary conditions $u(0) = u(1) = 0$. For this test, we discretize the domain using finite differences with $n = 150$ degrees of freedom (DoF). The numerical difficulty increases as the parameter λ approaches the critical limit $\lambda_c \approx 3.5138$. To provide a rigorous evaluation, we perform an extensive stability analysis using $M = 100$ Monte Carlo trials per grid point. The parameter space covers initial uncertainty $\|u_0 - u^*\|_\infty \in [1, 10]$ and non-linearity values $\lambda \in [0.5, 3.5]$.

The results in Figure 3 reveal a significant finding: the proposed eighth-order ST8S scheme exhibits a larger convergence success rate than the second-order Newton method. This superiority is further confirmed in the 3D landscape (Figure 4), where the success surface of ST8S consistently encompasses and exceeds the Newton baseline. The obtained results by ESS8 and EH9S show worse results, including areas of no convergence and widest regions of convergence below 30%.

Finally, Figure 5 presents a survival analysis at the critical limit $\lambda = 3.5$. Under this extreme stress regime, ST8S remains highly reliable, even outperforming Newton's method as uncertainty

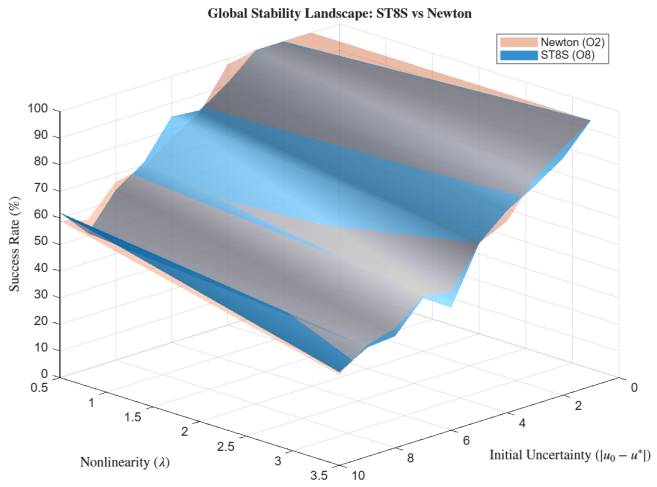


Figure 3: Global stability maps for the 1D Bratu problem ($n = 150$). The maps represent $M = 100$ Monte Carlo simulations per grid point.

increases. In contrast, the ESS8 and EH9S schemes show a near-null survival rate, proving their lack of robustness in highly non-linear environments.

5.2 High-precision analysis: Elastic string problem

Let us consider the boundary-value problem

$$\ddot{x}(t) = -1 - a^2\dot{x}^2, \quad t \in [0, 1]$$

with $x(0) = 0$, $x(1) = 0, a \in \mathbb{R}$. This problem arises in the study of finite deflections of an elastic string under a transverse load and its exact solution is $x(t) = \ln\left(\frac{\cos(a(t-1/2))}{\cos a/2}\right)\frac{1}{a^2}$.

In order to get an estimation of this solution, we also use the second order finite differences method and, therefore, the following non-linear system provides us an estimation of the solution of the original boundary value problem:

$$4(x_{i+1} - 2x_i + x_{i-1}) + 4h^2 + a^2(x_{i+1} - x_{i-1})^2 = 0, \quad i = 1, 2, 3, \ldots, n, \tag{14}$$

where $x_i \approx x\,(t_i)$ for each i and $h = \frac{1}{n+1}$.

We use $n = 500$. The numerical tests were performed using multi-precision arithmetic with 1000 digits of accuracy. To ensure the reliability of the performance metrics, we used four distinct initial estimations: $x_1^{(0)} = \frac{13}{25000}(1, 2, 3, \ldots, 249, 250, 249, \ldots, 3, 2, 1)^T$, $x_2^{(0)} = \frac{1}{500000}(499 \cdot 1, 498 \cdot 2, 497 \cdot 3, \ldots, 2 \cdot 498, 1 \cdot 499)^T$, $x_3^{(0)} = \frac{1}{500}(1, 2, 3, \ldots, 249, 250, 249, \ldots, 3, 2, 1)^T$, $x_4^{(0)} = \frac{1}{2}(1, 1, \ldots, 1, 1)^T$ and $a = 1$ are analyzed.

In Table 2 we show, for each initial estimate $x_i^{(0)}$ $i = 1, 2, 3, 4$, Time_i (in seconds), the quantity of iterations (It_i), the values of the logarithm of the approximate error (E_i) rounded to unity and of the residual error (R_i), and the approximation of the theoretical order $\breve{\rho}_i$. All implementations with the same system size $n = 500$. The best time outline is placed in red color.

The robustness of the data obtained indicates that among the 4 initial estimations, the ones that allow a faster convergence in all the methods are the first two. This is evidenced by the number of iterations that each scheme needed with the different initial points.

In all numerical tests, our ST8S scheme outperforms the others, because it converges at lowest time. For the estimate $x_3^{(0)}$, it converges as fast as the best ones, and for the least good one, $x_4^{(0)}$, it is much better than the others. Also, the second best performance of the methods corresponds to EH9S, in all cases. Similarly, it is observed that the approximate order of the schemes is not identical to the theoretical order. But despite all suffering from the same flaw, our proposal remains one of the most competitive and the one that required the least amount of time in all cases.

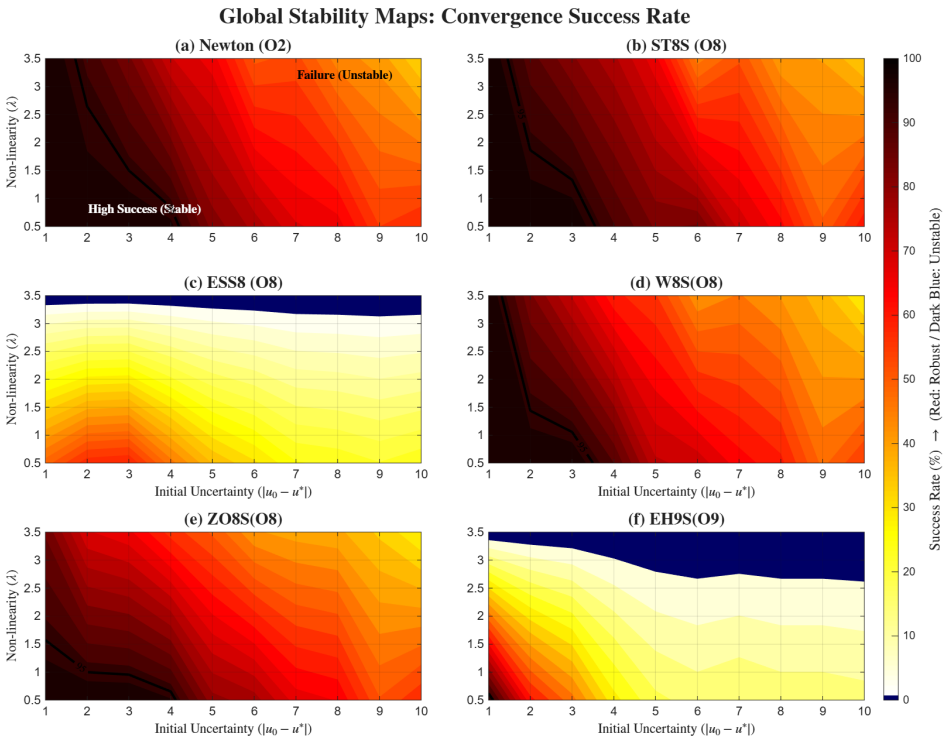


Figure 4: Global Stability Landscape: ST8S vs. Newton, ESS8, W8S, ZO8S and EH9S ($n = 150$).

Table 2: Numerical results for the elastic string problem ($n = 500$ and 4 different initial estimations ($x_i^{(0)}$)).

Method	$x_1^{(0)}$					$x_2^{(0)}$					$x_3^{(0)}$					$x_4^{(0)}$				
	$Time_1$	It_1	E_1	R_1	$\check{\rho}_1$	$Time_2$	It_2	E_2	R_2	$\check{\rho}_2$	$Time_3$	It_3	E_3	R_3	$\check{\rho}_3$	$Time_4$	It_4	E_4	R_4	$\check{\rho}_4$
N2S	11.56	5	-23	-49	2.04	9.26	4	-22	-45	2.05	13.71	6	-22	-45	2.05	13.89	6	-26	-52	2.05
ST8S	3.87	2	-7	-41	4.23	5.52	2	-11	-63	4.52	6.42	3	-13	-74	4.62	8.25	3	-18	-97	5.09
ESS8	24.18	10	25	48	1.02	8.09	3	-15	-48	1.93	24.47	10	7	14	-1.16	26.06	10	15	30	0.76
W8S	41.03	2	-8	-54	5.63	42.90	2	-17	-103	7.37	79.65	3	-26	-162	6.56	59.75	3	-26	-212	8.57
ZO8S	9.94	2	-8	-52	5.54	9.99	2	-15	-92	6.63	20.75	3	-22	-134	6.46	14.28	3	-31	-193	7.30
EH9S	7.11	2	-12	-100	8.19	7.22	2	-22	-177	9.63	9.92	2	-4	-41	10.79	10.58	3	-38	-346	9.33

6 Conclusions

High-order schemes for nonlinear systems are frequently avoided in practical applications due to their prohibitive computational cost or inherent numerical instability. This work overcomes these limitations by introducing a highly efficient recursive family of methods based on damped Traub-type steps and scalar weight functions. The proposed architecture achieves a convergence order of $2m$ for m stages, maintaining a *Least-Evaluative Cost* (LEC) by reusing a single Jacobian matrix and its LU decomposition across all recursive steps.

Comparative analysis against state-of-the-art $O(8)$ and $O(9)$ schemes demonstrates that our proposal equals or surpasses existing high-performance methods in terms of functional, operational, and computational efficiency. From a dynamical perspective, the ST8S scheme provides a superior balance of connected convergence regions and wide attraction basins, effectively eliminating the chaotic behavior typically associated with high-order iterations.

Notably, Monte Carlo simulations on the Bratu problem reveal that the proposed method can exceed the stability of the optimal second-order Newton method under increasing uncertainty. Finally, numerical tests on the elastic string problem with large-scale systems ($n = 500$) and high-precision arithmetic confirm the reliability and CPU-time advantages of this approach. This work

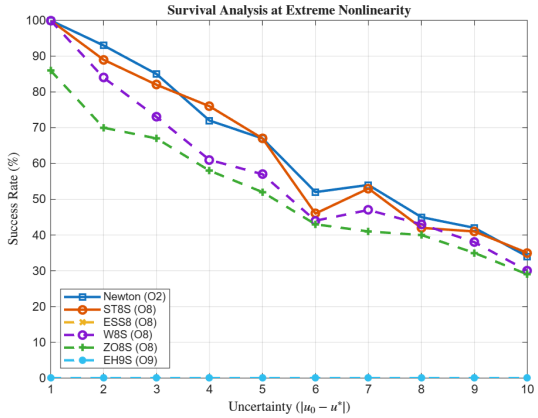


Figure 5: Survival analysis at extreme non-linearity ($\lambda = 3.5$). Success rate as a function of the initial uncertainty.

establishes a new benchmark for stable, low-cost, high-order solvers, contributing a robust tool for the treatment of complex nonlinear problems in science and engineering.

References

[1] Wang, X.: Fixed-point iterative method with eighth-order constructed by undetermined parameter technique for solving nonlinear systems, Symmetry, 13(5), 863 (2021).

[2] Singh, H., Sharma, J.R., Kumar, S.: A simple yet efficient two-step fifth-order weighted-Newton method for nonlinear models, Numerical Algorithms, 93(1), 203-225 (2023).

[3] Zhanlav, T., Otgondorj, K.: High Efficient Iterative Methods with Scalar Parameter Coefficients for Systems of Nonlinear Equations, Journal of Mathematical Sciences, 279(6), 866-875 (2024).

[4] Erfanifar, R., Hajarian, M.: A new multi-step method for solving nonlinear systems with high efficiency indices, Numerical Algorithms, 97(2), 959-984 (2024).

[5] Arroyo, V., Cordero, A., Torregrosa, J.R.: Approximation of artificial satellites' preliminary orbits: the efficiency challenge, Mathematical and Computer Modelling, 54(7-8), 1802-1807 (2011).

[6] Cordero, A., Rojas-Hiciano, R.V., Torregrosa, J.R., Vassileva, M.P.: A highly efficient class of optimal fourth-order methods for solving nonlinear systems, Numerical Algorithms, 95(4), 1879-1904 (2024).

[7] Cordero, A., Rojas-Hiciano, R. V., Torregrosa, J. R., Vassileva, M. P.: High-efficiency implicit scheme for solving first-order partial differential equations, Results in Applied Mathematics, 24, 100507 (2024).

[8] Zhanlav, T., Otgondorj, K.: Optimal eight-order three-step iterative methods for solving systems of nonlinear equations, Discrete and Continuous Models and Applied Computational Science, 33(4), 389-403 (2025).

[9] Cordero, A., Hueso, J.L., Martínez, E., Torregrosa, J.R.: A modified Newton-Jarratt's composition, Numer Algor 55, 87-99 (2010).

[10] Cordero, A., Rojas-Hiciano, R. V., Torregrosa, J. R., Vassileva, M. P.: Maximally efficient damped composed Newton-type methods to solve nonlinear systems of equations, Applied Mathematics and Computation, 492, 129231 (2025).

[11] Cordero, A., Rojas-Hiciano, R. V., Torregrosa, J. R., Vassileva, M. P.: High-level convergence order accelerators of iterative methods for nonlinear problems, Applied Numerical Mathematics, 217, 390-411 (2025).

Parallel Methods for Multiple Roots via Predictor–Corrector Compositions

Francisco I. Chicharro[♭], Neus Garrido[♭], and Julissa H. Jerezano[§,♮] [1]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain.

(§) PhD program in Mathematics, Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain.

(♮) Departmento de Matemática, Universidad Nacional Autónoma de Honduras,
21102–San Pedro Sula, Honduras.

1 Introduction

Many problems across fields such as science, engineering, economics, and other disciplines require solving nonlinear equations of the form $f(x) = 0$, $f \subseteq D \subset \mathbb{R} \longrightarrow \mathbb{R}$ [1, 2]. Unlike linear equations, these types of equations generally do not admit exact closed form analytical solutions, so iterative methods are required to approximate them.

These methods consist of numerical procedures that, given an initial guess $x^{[0]}$, generate successive approximations to the roots of the equation. However, when a root has multiplicity greater than one, meaning that the function $f(x)$ and several of its derivatives vanish at that point, standard iterative methods may lose efficiency, exhibiting slow convergence or even failing in some cases. In particular, Newton's method exhibits quadratic convergence when the root is simple; however, for roots with multiplicity greater than one ($m > 1$), the convergence deteriorates to linear. This reduction in the convergence order arises because the error equation at each iteration is influenced by the multiplicity of the root, thereby decreasing the overall efficiency of the method. Therefore, it is necessary to use iterative methods specifically designed for finding multiple roots. In this context, Schröder [3] proposed a modification of Newton's method in 1870, in which quadratic convergence is preserved through the iterative expression

$$x^{[k+1]} = x^{[k]} - m\frac{f\left(x^{[k]}\right)}{f'\left(x^{[k]}\right)}, \quad k = 0, 1, \ldots$$

Numerous methods with different orders of convergence have been proposed in the literature for finding multiple roots of the equation $f(x) = 0$. However, only a few of them are designed for the simultaneous determination of roots of nonlinear equations, particularly when some of these roots are multiple. Related studies can be found in [4, 5, 6].

Ehrlich [7] proposed an iterative method with cubic order of convergence for the simultaneous approximation of the n roots of a polynomial $p(x)$. According to this method, each root α_i, $i = 1, 2, \ldots, n$, is approximated using the following iterative expression

[1] jhjerull@doctor.upv.es

$$x_i^{[k+1]} = x_i^{[k]} - \frac{p\left(x_i^{[k]}\right)}{p'\left(x_i^{[k]}\right) - p\left(x_i^{[k]}\right)\sum\limits_{\substack{j=1\\ j\neq i}}^{n} \frac{1}{x_i^{[k]} - x_j^{[k]}}}, \quad k = 0, 1, \ldots$$

Cordero et al. designed in [8] and [9] simultaneous root-finding schemes of order 2ρ, given a fixed-point iterative scheme of order ρ and using Ehrlich-type methods, to approximate simple roots of nonlinear equations. Furthermore, a fourth-order scheme was also proposed for the approximation of multiple roots in [9].

Following the ideas in [8, 9], this work presents iterative procedures for solving nonlinear equations with multiple roots simultaneously. Section 2 describes the design procedure of the methods. Section 3 presents numerical experiments involving three nonlinear functions. Finally, Section 4 provides the main conclusions.

2 Methods

We consider the nonlinear equation $f(x) = 0$, which has n multiple roots denoted by α_i, for $i = 1, \ldots, n$, where each α_i has multiplicity $m_i \in \mathbb{N}$. Taking into account Ehrlich's method, we propose the following two-step iterative scheme for the simultaneous approximation of the multiple roots of $f(x) = 0$:

$$\begin{aligned} y_i^{[k]} &= x_i^{[k]} - m_i \frac{f\left(x_i^{[k]}\right)}{f'\left(x_i^{[k]}\right)}, \\ x_i^{[k+1]} &= y_i^{[k]} - \frac{m_i}{\dfrac{f'\left(y_i^{[k]}\right)}{f\left(y_i^{[k]}\right)} - \sigma_i^{[k]}}, \quad k = 0, 1, \ldots \end{aligned} \tag{1}$$

for $i = 1, \ldots, n$. Method NY replaces $\sigma_i^{[k]}$ by

$$\sigma_i^{[k]} = \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{1}{y_i^{[k]} - y_j^{[k]}},$$

while method NX replaces it by

$$\sigma_i^{[k]} = \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{1}{y_i^{[k]} - x_j^{[k]}}.$$

It is worth noting that the first step, $y_i^{[k]}$, corresponds to a modified Newton method [3], whereas the second step applies a modified Ehrlich method [7]. The resulting iterative scheme involves a predictor-corrector method. It can be shown that the iterative scheme (1) has order of convergence equal to four.

3 Numerical experiments

The proposed iterative schemes are evaluated for solving the nonlinear equations

- $f_1(x) = (x-1)(x-2)(x-4)$.

- $f_2(x) = (x-2)^2(x+1)$.
- $f_3(x) = (x+1)^3\left(e^x-1\right)^2$.

The numerical experiments were carried out using MATLAB R2023a on a computer equipped with an Intel Core i7-14700 processor at 2.10 GHz and 16 GB of RAM. The computations were performed using variable precision arithmetic with a 200 digit mantissa to ensure that the stopping criterion $\|f(x^{[k]})\|_2 < 10^{-25}$ is satisfied without numerical issues such as division by zero. In addition, a maximum of 50 iterations is used as the stopping condition.

The results are presented in Tables 1, 2, and 3. These tables include the initial approximation $x^{[0]}$, the approximations of the solutions α_i, the number of iterations k required to reach convergence, the difference of iterates in the last iteration $\mathcal{D} := \log_{10}\left(\|x^{[k+1]} - x^{[k]}\|_2\right)$, the residual in the last iteration $\mathcal{R} := \log_{10}\left(\|f(x^{[k]})\|_2\right)$, and the approximate computational order of convergence ACOC [10].

Table 1 shows the performance of the proposed methods when applied to a polynomial with three simple roots. In this setting, issues related to different root multiplicities are deliberately excluded, since all roots are simple. The purpose of this experiment is to isolate and analyze the effect of the simultaneous approximation of the roots, focusing exclusively on the interaction inherent to parallel root-finding schemes.

Table 1: Results for equation $f_1(x) = 0$.

Method	$x^{[0]}$	α_1	α_2	α_3	k	$\mathcal{D}$	$\mathcal{R}$	ACOC
NY	$\begin{pmatrix} 0.5 \\ -1 \\ 3 \end{pmatrix}$	2	4	1	6	-5.2355	-32.9980	4.6056
NX	$\begin{pmatrix} 0.5 \\ -1 \\ 3 \end{pmatrix}$	2	4	1	6	-19.4780	-101.6200	5.0613

The theoretical order of the proposed methods for approximating multiple roots is four; however, in this case, slightly higher values of the ACOC are observed. In particular, the NX method exhibits better performance in terms of the difference between the last two iterates and the residual.

Table 2 presents the results obtained by applying the iterative procedures to a polynomial whose roots have different multiplicities, namely one simple root and one double root. In this case, the methods are employed in a combined setting that addresses both multiple roots of multiplicity one and two, while simultaneously approximating all roots of the polynomial. This allows us to assess the performance of the proposed schemes not only in the presence of heterogeneous multiplicities, but also in the context of parallel root-finding, where all roots are computed concurrently.

Table 2: Results for equation $f_2(x) = 0$.

Method	$x^{[0]}$	α_1	α_2	k	$\mathcal{D}$	$\mathcal{R}$	ACOC
NY	$\begin{pmatrix} 1 \\ -0.5 \end{pmatrix}$	2	-1	3	-6.7947	-30.2280	3.4517
NX	$\begin{pmatrix} 1 \\ -0.5 \end{pmatrix}$	2	-1	3	-7.8225	-28.8300	3.4006

It can be concluded that the proposed iterative procedures successfully compute both roots simultaneously, despite their different multiplicities, and achieve convergence within a very small number of iterations. In particular, the methods accurately approximate the simple and double roots in parallel, confirming their effectiveness in the presence of mixed multiplicities. The reduced number of iterations is explained by the early satisfaction of the stopping criterion, as reflected in the behavior of the residual $\mathcal{R}$. As a consequence, the computation terminates before the asymptotic regime is fully attained, preventing the ACOC from stabilizing at a constant value. Nevertheless, the observed ACOC values remain close to the theoretical convergence order of four, providing further numerical evidence of the high efficiency and rapid convergence of the proposed schemes.

Table 3 shows the performance of the proposed methods when applied to an arbitrary nonlinear function possessing multiple roots. Simultaneous root-finding methods are traditionally designed and analyzed within the framework of polynomial equations, where the problem structure naturally supports parallel approximation strategies. However, in this experiment, we deliberately move beyond the polynomial case and test the behavior of the methods on general nonlinear functions, containing roots of different multiplicities. This setting allows us to assess the robustness and versatility of the proposed schemes when confronted with more general nonlinear problems, demonstrating their capability to simultaneously approximate multiple roots of varying multiplicities even in the absence of an underlying polynomial structure.

Table 3: Results for equation $f_3(x) = 0$.

Method	$x^{[0]}$	α_1	α_2	k	$\mathcal{D}$	$\mathcal{R}$	ACOC
NY	$\begin{pmatrix} \text{-0.5} \\ 0.5 \end{pmatrix}$	-1	$8.8170 \cdot 10^{-18}$	3	-4.4582	-34.1090	2.8766
NX	$\begin{pmatrix} \text{-0.5} \\ 0.5 \end{pmatrix}$	-1	$1.8023 \cdot 10^{-19}$	3	-4.8806	-33.4190	2.8579

The results reported in Table 3 confirm that the proposed method is capable of successfully computing the roots of arbitrary nonlinear functions with different multiplicities. In particular, the method can simultaneously approximate roots of heterogeneous multiplicities, thereby extending its applicability beyond the classical polynomial framework typically associated with simultaneous root-finding techniques. As observed previously for the function $f_2(x)$ in Table 2, the stopping criterion is satisfied after a very small number of iterations. This early termination prevents the ACOC from stabilizing.

4 Conclusions

In this work, we have investigated the performance of parallel iterative methods for the simultaneous computation of multiple roots of nonlinear equations without requiring prior knowledge of their multiplicities. The study has focused on the numerical behavior of the proposed schemes when applied to different problem settings, including polynomials and arbitrary nonlinear functions with roots of heterogeneous multiplicities.

The numerical experiments demonstrate that the methods can successfully approximate several roots simultaneously, even when the roots have different multiplicities. In particular, the results confirm that the proposed approach remains effective beyond the classical polynomial framework, thereby extending its applicability to general nonlinear functions. This highlights the versatility and robustness of parallel root-finding strategies in more general contexts.

Overall, the obtained results confirm that the considered parallel iterative methods provide an effective and reliable tool for the simultaneous approximation of multiple roots, including cases involving arbitrary nonlinear functions and mixed root multiplicities.

Acknowledgments

Funded with IMM Research Funding Call 2026 – Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València

References

[1] C. Argyros, M. I. Argyros, I. K. Argyros, Á. A. Magreñán, and Í. Sarría, "Local and semilocal convergence for chebyshev two point like methods with applications in different fields," *Journal of Computational and Applied Mathematics*, vol. 426, p. 115 072, 2023.

[2] Bhavna and S. Bhatia, "Convergence analysis of optimal iterative family for multiple roots and its applications," *Journal of Mathematical Chemistry*, vol. 62, no. 8, pp. 2007–2038, 2024.

[3] E. Schröder, "Über unendlich viele algorithmen zur auflösung der gleichungen," *Mathematische Annalen*, vol. 2, no. 2, pp. 317–365, 1870.

[4] G. H. Nedzhibov, "Iterative methods for simultaneous computing arbitrary number of multiple zeros of nonlinear equations," *International Journal of Computer Mathematics*, vol. 90, no. 5, pp. 994–1007, 2013.

[5] M. S. Petković, L. D. Petković, and J. Džunić, "On an efficient method for the simultaneous approximation of polynomial multiple roots," *Applicable Analysis and Discrete Mathematics*, pp. 73–94, 2014.

[6] M. Shams and B. Carpentieri, "A high-order parallel framework for simultaneous root-finding in nonlinear systems with multiple solutions," *AppliedMath*, vol. 6, no. 3, p. 43, 2026.

[7] L. W. Ehrlich, "A modified newton method for polynomials," *Communications of the ACM*, vol. 10, pp. 107–108, 1967. DOI: 10.1145/363067.363115.

[8] A. Cordero, N. Garrido, J. R. Torregrosa, and P. Triguero-Navarro, "Iterative schemes for finding all roots simultaneously of nonlinear equations," *Applied Mathematics Letters*, vol. 134, p. 108 325, 2022.

[9] A. Cordero, N. Garrido, J. R. Torregrosa, and P. Triguero-Navarro, "An iterative scheme to obtain multiple solutions simultaneously," *Appl. Math. Lett.*, vol. 145, p. 108 738, 2023.

[10] A. Cordero and J. Torregrosa, "Variants of Newton's method using fifth–order quadrature formulas," *Appl. Math. Comput.*, vol. 190, pp. 686–698, 2007.

Halley-based method for simultaneous-root finding in nonlinear equations

Jesús Gallinal[♮1], Eva G. Villalba[♭] Paula Triguero-Navarro[♭]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València,
Camí de Vera s/n, 46022 València, Spain.
(♮) PhD program in Mathematics, Universitat Politècnica de València,
Camí de Vera s/n, 46022 València, Spain.

1 Introduction

Many phenomena arising in nature can be modeled by means of a nonlinear equation of the form $f(x) = 0$, where $f : D \subseteq \mathbb{R} \longrightarrow \mathbb{R}$. Iterative methods constitute an effective tool for obtaining approximate solutions to this equation when analytical techniques fail to provide a closed-form solution. Certain models do not admit a unique solution, but rather exhibit several values for which the function vanishes. Iterative methods for solving nonlinear equations with simultaneous roots, also referred to as parallel methods, allow for the computation of the complete set of r solutions. Starting from initial approximations, simultaneous schemes produce sequences that converge to all roots of the equation simultaneously under the appropriate conditions [1]–[6].

Since the introduction of Ehrlich's method [7] for solving polynomials $p(x) = 0$

$$x_i^{(k+1)} = x_i^{(k)} - \frac{1}{\dfrac{p'\left(x_i^{(k)}\right)}{p\left(x_i^{(k)}\right)} - \displaystyle\sum_{\substack{j=1 \\ j\neq i}}^{r} \dfrac{1}{x_i^{(k)} - x_j^{(k)}}}, \quad k = 0, 1, \ldots, \quad i = 1, 2, \ldots, r, \tag{1}$$

numerous contributions in recent years have addressed this class of iterative techniques, either through modifications of existing schemes or through the composition of a known method acting as a predictor together with Ehrlich's method serving as a corrector. In [8], the authors show the applicability of Ehrlich's method to solve non-polynomial functions.

Halley's method [9]

$$x_{k+1} = x_k - \frac{1}{\dfrac{f'(x_k)}{f(x_k)} - \dfrac{1}{2}\dfrac{f''(x_k)}{f'(x_k)}}, \quad k = 0, 1, \ldots, \tag{2}$$

has cubic order of convergence. It is originally devoted to solve nonlinear equations starting with a single initial guess for the estimation of simultaneous roots.

In this work, we aim to adapt (2) to the simultaneous-root finding procedure. This manuscript is structured as follows: Section 2 presents the methodology, Section 3 provides numerical results, and Section 4 discusses their implications.

[1] jagalmor@doctor.upv.es

2 Methods

Following the approach proposed by Ehrlich [7], if the estimation x_i remains in a neighborhood of one of the r roots α_i of $f(x) = 0$, then these roots are zeros of f, so we estimate f and f' as

$$f(x) \approx P(x) = \prod_{i=1}^{r}(x - \alpha_i), \tag{3}$$

$$f'(x) \approx P'(x) = \sum_{j=1}^{r}\prod_{\substack{i=1\\ i\neq j}}^{r}(x - \alpha_i), \tag{4}$$

and equations (3) and (4) lead to the approximation

$$\frac{f'(x)}{f(x)} \approx \frac{P'(x)}{P(x)} = \sum_{i=1}^{r}(x - \alpha_i)^{-1}.$$

Defining

$$s_0(x) = \frac{P'(x)}{P(x)} = \sum_{i=1}^{r}(x - \alpha_i)^{-1}, \tag{5}$$

we can construct the simultaneous-root finding version of Halley's method, namely HS, whose iterative expression is

$$x_i^{(k+1)} = x_i^{(k)} - \frac{1}{\dfrac{f'\left(x_i^{(k)}\right)}{f\left(x_i^{(k)}\right)} - s_0\left(x_i^{(k)}\right) - \dfrac{1}{2}\dfrac{f''\left(x_i^{(k)}\right)}{f'\left(x_i^{(k)}\right)}}, \quad k = 0, 1, \ldots, \quad i = 1, 2, \ldots, r. \tag{6}$$

Theorem 2.1. *Let $f : D \subset \mathbb{R} \to \mathbb{R}$ be a sufficiently differentiable function in a neighborhood D of α_i such that $f(\alpha_i) = 0$ and $f'(\alpha_i) \neq 0$ for $i = 1, 2, \ldots, r$. If $x_i^{(0)}$ is sufficiently close to α_i, then the sequence $\left\{x_i^{(k)}\right\}_{k\geq 0}$ generated by the simultaneous Halley's method* (6) *converges to α_i with order 2, for $i = 1, 2, \ldots, r$, where the error equation is*

$$e_i^{(k+1)} = -s_0\left(x_i^{(k)}\right)\left(e_i^{(k)}\right)^2 + \left(c_{2,i}^2 - c_{3,i} - s_0^2\left(x_i^{(k)}\right)\right)\left(e_i^{(k)}\right)^3 + \mathcal{O}\left(\left(e_i^{(k)}\right)^4\right), \tag{7}$$

where $c_{n,i} = \frac{1}{n!}\frac{f^{(n)}(\alpha_i)}{f'(\alpha_i)}$, $n > 1$, and $e_i^{(k)} = x_i^{(k)} - \alpha_i$ is the error at the k-th iteration.

3 Numerical tests

The simultaneous Halley's method (6) performance is compared against Ehrlich's method (1). Both of them are applied starting from an initial estimate $x^{(0)} = \left(x_1^{(0)}, \ldots, x_r^{(0)}\right)^T$ of the solution $\alpha = (\alpha_1, \ldots, \alpha_r)^T$, until a maximum of 100 iterations are reached or the following stopping criterion is satisfied

$$\left\|x^{(k+1)} - x^{(k)}\right\|_2 + \left\|f\left(x^{(k)}\right)\right\|_2 < 10^{-250}.$$

The numerical tests are performed using Matlab R2026a, with variable-precision arithmetic set to 2000 significant digits. Simulations are carreid out on a computer equipped with an Intel(R)

Core(TM) i7-4790S CPU @ 3.20GHz, 16 GB of RAM, and the Windows 10 Professional Edition operating system.

The results are reported in Tables 1 and 2, which indicate for different initial conditions $x^{(0)}$ the difference between the two last iteratites $\left\|x^{(k+1)} - x^{(k)}\right\|_2$, the residual $\left\|f\left(x^{(k)}\right)\right\|_2$, the number of iterations, and the approximated computational order of convergence (ACOC), defined in [10].

The polynomial function under study is

$$p(x) = x\left(x - \tfrac{1}{4}\right)\left(x - \tfrac{3}{4}\right)\left(x - \tfrac{3}{2}\right)\left(x - \tfrac{5}{2}\right),$$

whose zeros define $\alpha = \left(0,\ \frac{1}{4},\ \frac{3}{4},\ \frac{3}{2},\ \frac{5}{2}\right)^T$.

Table 1: Results applying HS (6) for solving $p(x)$.

$x^{(0)}$	$\left\|x^{(k+1)} - x^{(k)}\right\|_2$	$\left\|f\left(x^{(k)}\right)\right\|_2$	Iterations	ACOC
$\alpha + (1,1,1,1,1)^T \cdot 10^{-2}$	$6.94388 \cdot 10^{-308}$	$2.16979 \cdot 10^{-614}$	9	2.00000
$\alpha + (2,2,2,2,2)^T \cdot 10^{-2}$	$1.54141 \cdot 10^{-460}$	$1.06917 \cdot 10^{-919}$	10	2.00000
$\alpha + (3,3,3,3,3)^T \cdot 10^{-2}$	$1.16429 \cdot 10^{-369}$	$6.10002 \cdot 10^{-738}$	10	2.00000
$\alpha + (4,4,4,4,4)^T \cdot 10^{-2}$	$1.11738 \cdot 10^{-303}$	$5.61846 \cdot 10^{-606}$	10	2.00000
$\alpha + (5,5,5,5,5)^T \cdot 10^{-2}$	$1.35300 \cdot 10^{-499}$	$8.23776 \cdot 10^{-998}$	11	2.00000
$\alpha + (6,6,6,6,6)^T \cdot 10^{-2}$	$4.58093 \cdot 10^{-408}$	$9.44321 \cdot 10^{-815}$	11	2.00000
$\alpha + (7,7,7,7,7)^T \cdot 10^{-2}$	$6.14087 \cdot 10^{-325}$	$1.69696 \cdot 10^{-648}$	11	2.00000
$\alpha + (8,8,8,8,8)^T \cdot 10^{-2}$	$4.94452 \cdot 10^{-490}$	$1.10017 \cdot 10^{-978}$	12	2.00000
$\alpha + (9,9,9,9,9)^T \cdot 10^{-2}$	$1.10974 \cdot 10^{-330}$	$5.5418 \cdot 10^{-660}$	12	2.00000

Table 2: Results with Ehrlich's method (1) for solving $p(x)$.

$x^{(0)}$	$\left\|x^{(k+1)} - x^{(k)}\right\|_2$	$\left\|f\left(x^{(k)}\right)\right\|_2$	Iterations	ACOC
$\alpha + (1,1,1,1,1)^T \cdot 10^{-2}$	$5.21796 \cdot 10^{-332}$	$9.76037 \cdot 10^{-994}$	6	2.99576
$\alpha + (2,2,2,2,2)^T \cdot 10^{-2}$	$1.55747 \cdot 10^{-258}$	$2.33822 \cdot 10^{-773}$	6	2.99889
$\alpha + (3,3,3,3,3)^T \cdot 10^{-2}$	$5.63547 \cdot 10^{-643}$	$1.21359 \cdot 10^{-1926}$	7	2.99966
$\alpha + (4,4,4,4,4)^T \cdot 10^{-2}$	$1.77841 \cdot 10^{-549}$	$3.99932 \cdot 10^{-1646}$	7	2.99999
$\alpha + (5,5,5,5,5)^T \cdot 10^{-2}$	$2.91301 \cdot 10^{-475}$	$3.12697 \cdot 10^{-1423}$	7	2.99927
$\alpha + (6,6,6,6,6)^T \cdot 10^{-2}$	$3.65053 \cdot 10^{-413}$	$6.28375 \cdot 10^{-1237}$	7	2.99777
$\alpha + (7,7,7,7,7)^T \cdot 10^{-2}$	$3.41671 \cdot 10^{-359}$	$5.15520 \cdot 10^{-1075}$	7	2.99602
$\alpha + (8,8,8,8,8)^T \cdot 10^{-2}$	$1.04474 \cdot 10^{-312}$	$1.45817 \cdot 10^{-935}$	7	2.99811
$\alpha + (9,9,9,9,9)^T \cdot 10^{-2}$	$1.25559 \cdot 10^{-268}$	$2.54921 \cdot 10^{-803}$	7	2.99756

The numerical results in Table 1 demonstrate that the simultaneous Halley's method (6) achieves an ACOC of 2, thus validating the theoretical order established in theorem 2.1. For $\delta \geq 18.889$ the initial condition $x^{(0)} = \alpha + (\delta, \delta, \delta, \delta, \delta)^T \cdot 10^{-2}$ causes the method to not converge.

However, the numerical results in Table 2 demonstrate that the Ehrlich method achieves ACOC ≈ 3. No restriction could be placed on δ under the initial condition $x^{(0)} = \alpha + (\delta, \delta, \delta, \delta, \delta)^T \cdot 10^{-2}$, which implies that the method always converges for every δ.

While most methods are tailored for polynomials and often fail to converge for other types of functions, we also employ the Ehrlich and HS methods to handle non-polynomial cases. The following functions will be analyzed:

- $f_1(x) = e^{x^2} - x$,
- $f_2(x) = \cos(x^2) - x\sin(x)$,
- $f_3(x) = e^{-x}\ln(x) + \cos(x)$.

Table 3: Results with Ehrlich's (1) and HS (6) methods for solving $f_j(x) = 0$, $j = 1, 2, 3$.

Method	Function	$x^{(0)}$	$\left\|x^{(k+1)} - x^{(k)}\right\|_2$	$\left\|f\left(x^{(k)}\right)\right\|_2$	Iterations	ACOC
Ehrlich	f_1	$(i, -i)$	$2.64947 \cdot 10^{-371}$	$9.92111 \cdot 10^{-742}$	12	2.0
HS			$5.68733 \cdot 10^{-452}$	$3.43163 \cdot 10^{-903}$	11	2.0
Ehrlich	f_2	$(1, -1, 2, -2)$	$2.19085 \cdot 10^{-389}$	$9.29158 \cdot 10^{-774}$	30	2.0
HS			$7.49575 \cdot 10^{-251}$	$3.04907 \cdot 10^{-500}$	23	2.0
Ehrlich	f_3	$(0.5, 1.5, 2.5, 3.5)$	$4.25329 \cdot 10^{-289}$	$2.57995 \cdot 10^{-577}$	13	2.0
HS			$1.32651 \cdot 10^{-257}$	$1.57162 \cdot 10^{-513}$	21	2.0

Table 3 shows the numerical results obtained for the three selected non-linear functions. It is important to highlight that the initial approximations were intentionally chosen to differ from the actual solutions, testing the robustness of each approach.

The data reveals that both methods show strong performance for non-polynomial functions. In two of the analyzed cases, the HS method defined in this study proves to be more efficient, requiring fewer iterations than Ehrlich's method to achieve the chosen tolerance.

4 Conclusions

The extension of Halley's method for the simultaneous approximation of roots of nonlinear equations has been proposed, following the methodology introduced by Ehrlich. This results in the simultaneous Halley's method HS (6), which served as the basis for the numerical tests that confirm the order of convergence established in Theorem 2.1. The same tests were conducted using Ehrlich's method (1) to provide a comparative analysis. The results show that the simultaneous Halley method is more limited than Ehrlich's methods for polynomials, but for others non-polynomial functions both methods have comparable performance.

Additional tests, using initial guesses constructed from equally spaced points within intervals containing the roots, suggest that convergence may occur for certain points even when the simultaneous Halley method does not converge globally. Furthermore, the nature of convergence for a given set of initial points depends on all of them collectively, as any subset thereof does not converge in the same manner.

All of this suggests that studying this simultaneous Halley method is a promising avenue for exploring alternatives to established simultaneous schemes and contributing new insights to this field.

References

[1] M. Petković, L. Petković, and J. Džunić, "On an efficient simultaneous method for finding polynomial zeros," *Applied Mathematics Letters*, vol. 28, pp. 60–65, 2014, ISSN: 0893-9659. DOI: https://doi.org/10.1016/j.aml.2013.09.011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089396591300284X.

[2] P. Ivan and R. L. Z., "Computational geometry as a tool for studying root-finding methods," *Filomat*, vol. 33, pp. 1019–1027, 4 2019. DOI: 10.2298/FIL1904019P. [Online]. Available: https://www.pmf.ni.ac.rs/filomat-content/2019/33-4/33-4-2-8769.pdf.

[3] P. D. Proinov and M. T. Vasileva, "On a family of Weierstrass-type root-finding methods with accelerated convergence," *Applied Mathematics and Computation*, vol. 273, pp. 957–968, 2016, ISSN: 0096-3003. DOI: https://doi.org/10.1016/j.amc.2015.10.048. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0096300315013995.

[4] P. D. Proinov, "On the local convergence of Ehrlich method for numerical computation of polynomial zeros," *Calcolo*, vol. 53, pp. 413–426, 3 2016. DOI: 10.1007/s10092-015-0155-y. [Online]. Available: https://doi.org/10.1007/s10092-015-0155-y.

[5] P. D. Proinov and M. T. Vasileva, "On the convergence of high-order Gargantini–Farmer–Loizou type iterative methods for simultaneous approximation of polynomial zeros," *Applied Mathematics and Computation*, vol. 361, pp. 202–214, 2019, ISSN: 0096-3003. DOI: https://doi.org/10.1016/j.amc.2019.05.026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0096300319304199.

[6] P. D. Proinov and S. I. Ivanov, "Convergence analysis of Sakurai–Torii–Sugiura iterative method for simultaneous approximation of polynomial zeros," *Journal of Computational and Applied Mathematics*, vol. 357, pp. 56–70, 2019, ISSN: 0377-0427. DOI: https://doi.org/10.1016/j.cam.2019.02.021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377042719300949.

[7] L. W. Ehrlich, "A modified Newton method for polynomials," *Communications of the ACM*, vol. 10, pp. 107–108, 2 1967. DOI: 10.1145/363067.363115. [Online]. Available: https://doi.org/10.1145/363067.363115.

[8] A. Cordero, N. Garrido, J. Torregrosa, and P. Triguero-Navarro, "Iterative schemes for finding all roots simultaneously of nonlinear equations," *Applied Mathematics Letters*, vol. 134, p. 108 325, 2022. DOI: https://doi.org/10.1016/j.aml.2022.108325.

[9] E. Halley, "A new exact and easy method for finding the roots of equations generally and without any previous reduction," *Philosophical Transactions of the Royal Society of London*, vol. 18, pp. 136–147, 210 1694. DOI: 10.1098/rstl.1694.0029. [Online]. Available: http://dx.doi.org/10.1098/rstl.1694.0029.

[10] A. Cordero and J. R. Torregrosa, "Variants of Newton's method using fifth-order quadrature formulas," *Applied Mathematics and Computation*, vol. 190, pp. 686–698, 1 2007. DOI: 10.1016/j.amc.2007.01.062. [Online]. Available: https://doi.org/10.1016/j.amc.2007.01.062.

A Cubic-Order Iterative Scheme for Approximating the Inverse of a Non-Singular Matrix

Alicia Cordero[♭], Bushra Rani[♭1] and Juan Ramón Torregrosa[♭]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

In numerical linear algebra, iterative methods for finding the inverse of a matrix play an important role, particularly when direct methods fail while dealing with large-scale and sparse systems. These iterative methods avoid the *fill-in* phenomenon, which destroys sparsity and leads to prohibitive memory requirements that often occur when applying direct methods. Additionally, iterative methods can significantly reduce computational costs.

Many iterative methods have been developed to approximate the inverse of a matrix (see [1]). Several root-finding iterative schemes used to solve scalar nonlinear equations of the form $f(x) = 0$ can be extended to matrix equations for computing matrix inverses, namely

$$F(X) = X^{-1} - A = 0, \tag{1}$$

where A is a nonsingular matrix. Among these schemes, the Newton–Schulz and Homeier schemes are two of the most well-known iterative techniques that allow the computation of A^{-1} without explicitly using the inverse operator in their formulations.

The Newton–Schulz method is one of the most simple inversion-free iterative method whose mathematical expression is given by

$$X_{n+1} = X_n(2I - AX_n), \tag{2}$$

which converges quadratically (see [2]). In other words, at each iteration of this method, the residual $E_n = I - AX_n$ satisfies

$$\|E_{n+1}\| \leq \|E_n\|^2, \quad \text{for all } n.$$

This method is an extension of the classical Newton–Raphson method.

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad \text{for all } n. \tag{3}$$

On the other hand, a variant of the Newton–Raphson method that improves the speed of convergence was introduced by Homeier H.H.H in [3]. Its scalar formulation is given as

[1] brani@upv.edu.es

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)}{f'(x_n)} \\ x_{n+1} &= x_n - \frac{f(x_n)}{2}\left[\frac{1}{f'(x_n)} + \frac{1}{f'(y_n)}\right] \end{aligned} \tag{4}$$

which converges cubically.
Extending this root-finding method to matrix equations leads to an iterative scheme for approximating the inverse of a matrix studied in [4]. The corresponding matrix iteration is given by

$$X_{n+1} = X_n\left[I + \frac{1}{2}(I - AX_n)\left(I + (2I - AX_n)^2\right)\right], \quad n = 1, 2, 3, \ldots \tag{5}$$

Therefore, the objective of this paper is to use a cubic-order root-finding iterative method and convert it into an inversion-free iterative method of order three for approximating the inverse of a non-singular matrix.

2 Proposed Method

A two-step root-finding iterative method for solving the nonlinear equation $f(x) = 0$:

$$\begin{aligned} y_n &= x_n - \frac{2}{3}\frac{f(x_n)}{f'(x_n)} \\ x_{n+1} &= x_n - \frac{f(x_n)}{4}\left[\frac{1}{f'(x_n)} + \frac{3}{f'(y_n)}\right] \end{aligned} \tag{6}$$

called Heun formula was studied in [5] ,whose error equation is given by

$$e_{n+1} = \frac{2}{3}c_2^2 e_n^3 + \left(\frac{5}{3}c_2c_3 - \frac{13}{9}c_3^2 + \frac{1}{9}c_4\right)e_n^4 + O(e_n^5).$$

This shows that Heun's method has a cubic order of convergence. For the matrix equation $F(X) = X^{-1} - A$, its iterative expression corresponding to the previous Heun method is given by

$$X_{n+1} = -\frac{1}{3}X_n\left[-10I + AX_n\left(12I + AX_n\left(-6I + AX_n\right)\right)\right].$$

where I is the identity matrix. For $B_n = AX_n$, the equation can be simplified as

$$X_{n+1} = -\frac{1}{3}X_n\left[-10I + B_n\left(12I + B_n\left(-6I + B_n\right)\right)\right]. \tag{9}$$

Theorem 1. Let $A \in \mathbb{C}^{n\times n}$ be a non-singular matrix, and let X_0 be an initial approximation such that $\|I - AX_0\| < 1$. Then, the sequence $\{X_n\}$ obtained by (9), converges to A^{-1} with convergence order $p = 3$. In this case, the error equation $e_n = X_n - A^{-1}$ satisfies

$$\|e_{n+1}\| \leq \|A^2\|\,\|e_n\|^3. \tag{7}$$

3 Computational Cost Analysis

To compare the efficiency of the proposed method with the Newton–Schulz method, we count the number of matrix multiplications required per iteration.The efficiency index (see [6]) is defined as

$$EI = p^{\frac{1}{d}}$$

where p and d are the rate of convergence and the total cost per cycle,respectively. Hence the efficiency index of different methods is given by

- Newton–Schulz Method: 2 matrix multiplications per iteration ; $EI = 2^{\frac{1}{2}} \approx 1.4142$
- Proposed Method: 4 matrix multiplications per iteration ; $EI = 3^{\frac{1}{4}} \approx 1.3160$

Although the proposed method requires slightly more matrix multiplications per iteration, its cubic order of convergence reduces the total number of iterations, which may imply a lower overall computational cost.

4 Numerical Results

Numerical experiments were carried out on random square matrices of sizes $n = 50$, $n = 100$, and $n = 500$ to approximate the inverse of a non-singular matrix A. The initial approximation was chosen as $X_0 = \frac{A^T}{\|A\|_2^2}$, which ensures that each method satisfies the convergence condition $\|I - AX_0\|_2 < 1$. The stopping criterion used in all computations was $\|F(X_{n+1})\|_2 = \|I - AX_{n+1}\|_2 < 10^{-9}$.To verify the numerical convergence of each method, Jay in [7] introduced the computational order of convergence (COC), which is defined as

$$\text{COC} \approx \frac{\ln\left(\|F(X_{n+1})\|/\|F(X_n)\|\right)}{\ln\left(\|F(X_n)\|/\|F(X_{n-1})\|\right)}. \tag{8}$$

The numerical results confirm that the Heun method exhibits cubic order of convergence. Moreover, the proposed method requires fewer iterations to obtain the inverse of a nonsingular matrix compared to the Newton–Schulz method, which has quadratic order of convergence.These results were carried out in MATLAB R2025a with a processor Intel Core i3-N305 processor running at 1.80 GHz with 8 GB RAM.

Table 1: Results obtained by the approximation of inverse of a matrix of size $n = 50$

Method	n	Iter.	$\|I - AX_{n+1}\|_2$	$\|X_{n+1} - X_n\|_2$	COC
Newton Schulz	50	25	5.8166×10^{-7}	0.0472	2.0000
Homeier	50	14	3.5172×10^{-7}	0.5488	3.0630
Huen	50	15	1.6537×10^{-11}	0.0181	3.0066

Table 2: Results obtained by the aproximation of inverse of a matrix of size $n = 100$

Method	n	Iter.	$\|I - AX_{n+1}\|_2$	$\|X_{n+1} - X_n\|_2$	COC
Newton Schulz	100	32	$1.6963 \times 10^{-}12$	0.0003	1.9997
Homeier	100	18	4.5380×10^{-13}	0.0135	2.7474
Huen	100	18	6.7292×10^{-7}	2.5059	3.0339

Table 3: Results obtained by the aproximation of inverse of a matrix of size $n = 500$

Method	n	Iter.	$\|I - AX_{n+1}\|_2$	$\|X_{n+1} - X_n\|_2$	COC
Newton Schulz	500	37	$4.3824 \times 10^{-}7$	0.2559	2.0000
Homeier	500	21	1.4682×10^{-10}	0.2571	3.0198
Huen	500	22	1.8888×10^{-12}	0.0396	2.8488

5 Conclusions

In this paper, we propose an iterative method without memory for approximating the inverse of a non-singular matrix. The convergence analysis of the proposed method was carried out using both theoretical and numerical results.
The obtained results demonstrate that although the Newton-Schulz method is computationally efficient due to its simple structure, it requires more iterations to converge to the desired solution because of its quadratic order of convergence. In contrast, the proposed Heun method, having a cubic order of convergence, achieves the solution in fewer iterations and thus provides an effective alternative for computing matrix inverses, especially for large-scale problems.
In future work, We can also study the theoretical and computational convergence analysis of the proposed method for finding the generalized inverse.

References

[1] N. J. Higham, *Functions of Matrices*. Society for Industrial and Applied Mathematics, 2008.

[2] G. Schulz, "Iterative Berechung der reziproken Matrix," *Zeitschrift Angewandte Mathematik und Mechanik*, vol. 13, no. 1, pp. 57–59, Jan. 1933.

[3] H. Homeier, "On newton-type methods with cubic convergence," *Journal of Computational and Applied Mathematics*, vol. 176, no. 2, pp. 425–432, 2005.

[4] H.-B. Li, T.-Z. Huang, Y. Zhang, X.-P. Liu, and T.-X. Gu, "Chebyshev-type methods and preconditioning techniques," *Applied Mathematics and Computation*, vol. 218, no. 2, pp. 260–270, 2011.

[5] A. Singh and J. P. Jaiswal, "Several new third-order and fourth-order iterative methods for solving nonlinear equations," *International Journal of Engineering*, vol. 2014, pp. 1–11, 2014.

[6] A. Ostrowski, *Solutions of the Equations and Systems of Equations*. Prentice-Hall: England Cliffs, NJ, USA; New York, NY,USA, 1964.

[7] L. Jay, "A note on q-order of convergence," *BIT Numerical Mathematics*, 2001.

Extending a Class of Third-Order Iterative Methods for Solving Volterra Integral Equations

Rajinder Singh[a], Sukhjit Singh[a], Eulalia Martínez[b], M.A. Hernández[c]

(a) Department of Mathematics and Computing,
Dr B. R. Ambedkar National Institute of Technology, Jalandhar, India.

(b) Instituto Universitario de Matemática Multidisciplinar,
Universitat Politècnica de València. València, Spain.

(c) Department of Mathematics and Computation,
University of La Rioja, Spain.

E-mail: rajinders.ma.25@nitj.ac.in, kundalss@nitj.ac.in, eumarti@mat.upv.es, mahernan@unirioja.es.

1 Introduction

Nonlinear Volterra integral equations have garnered considerable interest in recent years, owing to their intricate mathematical structure and their extensive applications across diverse areas of science and engineering. [1–3]. These equations naturally arise in diverse applications, including the modeling of viscoelastic materials, population dynamics, and predator–prey interactions. Furthermore, they play a significant role in describing systems with memory effects, which are commonly encountered in biological, physical, and engineering processes. Despite their importance, obtaining exact analytical solutions is often challenging in practical scenarios. Consequently, numerical methods are widely adopted to obtain the approximate solutions.

In this study, we consider Volterra integral equations of the second kind given by

$$z(s) = F(s) - \eta \int_b^s X(s,t)\, \mathcal{Q}(z)(t)\, dt, \quad \eta \in \mathbb{R}, \quad s \in [b,c], \tag{1}$$

where, $-\infty < b < s \leq c < +\infty$, $f \in C([b,c])$, the kernel $X(s,t)$ is continuous in both arguments with the Nemytskii operator such that $\mathcal{Q} : C([b,c]) \to C([b,c])$ with $(\mathcal{Q}z)(t) = Q(z(t))$, where, $\mathcal{Q} : \mathbb{R} \to \mathbb{R}$ is differentiable and z is the unknown function to be determined.
For this, we consider the operator $\mathcal{R} : \Omega \subseteq C([b,c]) \to C([b,c])$ such that

$$[\mathcal{R}(z)](s) = F(s) - \eta \int_b^s X(s,t)\, \mathcal{Q}(z(t))\, dt, \tag{2}$$

It is observed that $z^* \in \Omega$ is a fixed point of the operator $\mathcal{R}$ if it satisfies $\mathcal{R}(z^*) = z^*$. This condition can equivalently be written as $(I - \mathcal{R})(z^*) = 0$, which shows that z^* is a solution of equation (1), and conversely, any solution of (1) corresponds to a fixed point of $\mathcal{R}$.

In [4], Ezquerro et al. studied the existence and approximation of solution of the equation (1) using method of successive approximations, which has linear convergence. Further, Ezquerro and

Hernández [5] extended this approach using Newton's method to approximate a fixed point of the operator $\mathcal{R}$, given by

$$z_{j+1} = z_j - [I - \mathcal{R}'(z_j)]^{-1}(I - \mathcal{R})(z_j), \quad j \geq 0, \tag{3}$$

where, $\mathcal{R}'(z_j)$ denotes the Fréchet derivative of the operator $\mathcal{R}$ at point z_j given by

$$[\mathcal{R}'(z)v](s) = \eta \int_b^s X(s,t)\, \mathcal{Q}'(z(t))\, v(t)\, dt. \tag{4}$$

Motivated by these studies, we consider a family of third-order iterative methods [6] to approximate the fixed point of the operator $\mathcal{R}$, which is presented in the form of the following algorithm,

$$\begin{cases} f_0 \text{ given in } \Omega, \\ g_j = f_j - [I - \mathcal{R}'(f_j)]^{-1}(I - \mathcal{R})(f_j), \\ h_j = f_j + \tau(g_j - f_j), \qquad\qquad \tau \in (0,1], \\ f_{j+1} = g_j - \frac{1}{\tau^2}[I - \mathcal{R}'(f_j)]^{-1}((\tau - 1)(I - \mathcal{R})(f_j) + (I - \mathcal{R})(h_j)), \quad j \geq 0. \end{cases} \tag{5}$$

This family depends on a parameter $\tau \in (0,1]$. For $\tau = 1$, it reduces to the two-step Newton's method, which is more efficient than existing methods such as successive approximation and Newton's method. Moreover, there exist several nonlinear integral equations which does not satisfy the Lipschitz continuity conditions. Therefore, we extend and relax the assumptions by establishing the convergence of a family (5) under an ω-continuity condition:

$$||\mathcal{Q}'(f) - \mathcal{Q}'(h)|| \leq \omega(\|f - h\|), \qquad f, h \in \Omega. \tag{6}$$

Here, $\omega(t)$ denotes a continuous, non-decreasing real function for $t > 0$ with $\omega(0) \geq 0$.
Consider an integral equation given by,

$$[\mathcal{R}(z)](s) = \frac{\cos(2s) + 32\cos(s) - 9}{24} - \frac{1}{4}\int_0^s \sin(s-t)\left(z(t)^{7/5} + \frac{z(t)^2}{10}\right) dt,$$

where, $\mathcal{Q}(z)(t) = \left(z(t)^{7/5} + \frac{z(t)^2}{10}\right)$ is the Nemytskii operator. Clearly, $\mathcal{Q}'$ does not satisfy the Lipschitz as well as Hölder continuity conditions. Motivated by these considerations, we study the method (5) under weaker continuity conditions on $\mathcal{W}'$.

$$||\mathcal{Q}'(f) - \mathcal{Q}'(h)|| \leq \left(\frac{7}{5}\|f - h\|^{\frac{2}{5}} + \frac{1}{5}\|f - h\|\right),$$

In particular, if $\omega(t) = Kt$ or $\omega(t) = Kt^{\mu}$, condition (6) reduces to the Lipschitz or Hölder continuity of $\mathcal{Q}$, respectively. It is to be noted that the Lipschitz condition and Hölder condition are special cases of the ω - continuity condition.
The key contributions of this work are listed as follows:

- This study focuses on establishing the existence and location of a fixed point of the operator $\mathcal{R}$, and hence the corresponding solution of the integral equation (1).
- Global convergence analysis of a family of third-order iterative method (5) is presented under weaker continuity condition.
- Recurrence relations are established to obtain the convergence domains for the solution.
- We establish the existence and uniqueness theorem that enables us to identify the domains of existence and uniqueness for the solutions.
- The applicability and effectiveness of the presented approach are further established by validating the derived theoretical results on nonlinear Volterra-type integral equations.

2 Convergence Analysis

In this section, we construct the recurrence relations and derive the existence and uniqueness domains to establish the global convergence analysis of method (5) under the following assumptions.

(CI) Let $\tilde{f} \in \Omega$ be such that the operator $D^{-1} = [I - \mathcal{R}'(\tilde{f})]^{-1} \in \mathcal{L}(C([b,c]), C([b,c]))$ exists, $\|D^{-1}\| \leq \gamma$ and $\|B\| = \|(I - \mathcal{R})(\tilde{f})\| \leq \Gamma$.

(CII) There exists a continuous, real, and non-decreasing function $\omega(y)$, $y > 0$, $\omega(0) \geq 0$, $\omega(ty) \leq t^{\mu}\omega(y)$ for $t \in [0,1]$, $y \in [0,\infty)$, $\mu \in [0,1]$, such that

$$||\mathcal{Q}'(f) - \mathcal{Q}'(h)|| \leq \omega(\|f - h\|), \qquad f, h \in \Omega.$$

Now, to establish the global convergence of the family of iterative methods given by (5), we first present some lemmas that are useful in determining the convergence domains.

Lemma 2.1. *Let $\tilde{x} \in \Omega$ and assume that*

$$|\eta|UV < 1, \tag{7}$$

where, $U = \max_{s\in[b,c]} \int_b^s |X(s,t)|\, dt$, and $V = \max_{t\in[b,c]} \|\mathcal{R}'(\tilde{f}(t))\|$. Then, the operator $C^{-1} = [I - \mathcal{R}'(\tilde{f})]^{-1}$ exists and

$$\|C^{-1}\| \leq \frac{1}{1 - |\eta|UV} = \gamma.$$

Furthermore, define $D = \tilde{f} - \mathcal{R}(\tilde{f}) = (I - \mathcal{R})(\tilde{f})$. Then the operator D is bounded and satisfies $\|D\| \leq \|\tilde{f} - f\| + |\eta|UW = \Gamma$, where $W = \max_{t\in[b,c]} \left|\mathcal{R}(\tilde{f}(t))\right|$.

Lemma 2.2. *Assume* (CI)–(CII) *and $\overline{B(\widetilde{f},\sigma)} \subset \Omega$. If $f \in \overline{B(\widetilde{f},\sigma)}$ then there exists $[I - \mathcal{R}'(f)]^{-1}$,*

$$\|[I - \mathcal{R}'(f)]^{-1}\| \leq \frac{\gamma}{1 - \gamma|\eta|U\omega(\sigma)}, \quad \text{and} \quad \|[I - \mathcal{R}'(f)]^{-1}C\| \leq \frac{1}{1 - \gamma|\eta|U\omega(\sigma)},$$

where,

$$\gamma|\eta|U\omega(\sigma) < 1. \tag{8}$$

Lemma 2.3. *Under condition (CI)-(CII), following are satisfied for $j \geq 0$, we have*

(i) $(I - \mathcal{R})(g_j) = (I - \mathcal{R})(f_j) + [I - \mathcal{R}'(f_j)](g_j - f_j) + \int_{f_j}^{g_j} \left[(I - \mathcal{R})'(f) - (I - \mathcal{R})'(f_j)\right] df.$

(ii) $(I - \mathcal{R})(f_0) = (I - \mathcal{R})(\tilde{f}) + \left[I - \mathcal{R}'(\tilde{f})\right](f_0 - \tilde{f}) + \int_{\tilde{f}}^{f_0} \left[(I - \mathcal{R})'(f) - (I - \mathcal{R})'(\tilde{f})\right] df,$
with $f_0 \in \Omega$.

(iii) $(I - \mathcal{R})(h_j) = (1 - \tau)(I - \mathcal{R})(f_j) + \int_{f_j}^{h_j} \left[(I - \mathcal{R})'(f) - (I - \mathcal{R})'(f_j)\right] df,$
with $f_j, h_j \in \Omega$.

(iv) For $f_k, h_k \in \Omega$, it follows

$$(I - \mathcal{R})(f_{j+1}) = (I - \mathcal{R})(g_j) + \left[I - \mathcal{R}'(g_j)\right](f_{j+1} - g_j) + \int_{g_j}^{f_{j+1}} \left[(I - \mathcal{R})'(f) - (I - \mathcal{R})'(g_j)\right] df.$$

(vi) For $f_j \in \Omega$, it follows

$$(I-\mathcal{R})(f_j) + \left[I - \mathcal{R}'(f_j)\right](\tilde{f} - f_j) = (I-\mathcal{R})(\tilde{f}) - \int_{f_j}^{\tilde{f}} \left[(I-\mathcal{R})'(f) - (I-\mathcal{R})'(f_j)\right] df.$$

Let $a_0 = \dfrac{\gamma|\eta|U\omega(\zeta)}{1-\gamma|\eta|U\omega(\sigma)\tau^{1-\mu}}$, *and define a real positive sequence* $\{a_j\}$ *such that*

$$a_{j+1} = G(a_j)^{\mu} a_j, \quad j \geq 0, \tag{9}$$

$$G(x) = 2a_j + 2a_j^2, \tag{10}$$

Lemma 2.4. *The real-valued functions $G(t)$, defined in (10), are monotonically increasing for all $t > 0$. Moreover, if $a_0 < 0.3660$, then $f(a_0) < 1$, which ensures that the sequence $\{a_n\}$ is monotonically decreasing.*

Lemma 2.5. *Let $G(x)$ be real-valued functions defined by (10), respectively. Suppose there exists $\sigma > 0$ such that condition (7),(8) and*

$$\frac{a_0\zeta}{1+\mu} + \frac{(1+\mu)\gamma\Gamma + \gamma|\eta|U\omega(\sigma)\sigma}{(1+\mu)(1-\gamma|\eta|\omega(\sigma))} \leq \sigma \tag{11}$$

are satisfied, then the following recurrence relation holds true for every $j \geq 1$:

(i) $\|g_j - f_j\| \leq f(a_{j-1})\|g_{j-1} - f_{j-1}\|$,

(ii) $\|g_j - \tilde{f}\| \leq \frac{(1+\mu)\gamma\Gamma+\gamma|\eta|U\omega(\sigma)\sigma}{(1+\mu)(1-\gamma|\eta|U\omega(\sigma))}$,

(iii) $\|h_j - f_j\| \leq \tau G(a_{j-1})\|g_{j-1} - f_{j-1}\|$,

(iv) $\|f_{j+1} - f_j\| \leq \left(1 + \frac{a_j}{1+\mu}\right)\|g_j - f_j\|$,

(v) $\|f_{j+1} - \tilde{f}\| \leq \frac{(1+\mu)\gamma\Gamma+\gamma\Gamma U\omega(\sigma)\sigma}{(1+\mu)(1-\gamma\Gamma U\omega(\sigma))} + \frac{a_0}{1+\mu}\left(\prod_{i=0}^{n-1} f(a_i)^{1+u}\right)\zeta$.

Theorem 1. *Assume that conditions (CI) and (CII) hold. Let $\widetilde{f} \in \Omega$ and suppose there exists a radius $\sigma > 0$ such that the conditions (7),(8) and (11) are satisfied, and $\overline{B(\widetilde{f},\sigma)} \subset \Omega$. If $a_0 < 0.3660$, then the operator $\mathcal{R}$ admits a fixed point $f^* \in \overline{B(\widetilde{f},\sigma)}$. Moreover, for any initial point $f_0 \in \overline{B(\widetilde{f},\sigma)}$, the iterative process generated by σ remains within the ball and converges to the fixed point f^*.*

Theorem 2. *Under the assumptions of Theorem 1, the fixed point f^* of the operator $\mathcal{R}$ is unique in the ball $\overline{B(\widetilde{f},\sigma)}$.*

The derived theoretical results are further validated on nonlinear Volterra-type integral equations to demonstrate the applicability and effectiveness of the presented approach.

Acknowledgements

This work was supported by the Anusandhan National Research Foundation (ANRF), Government of India, under Grant [ANRF/IRG/2024/001327/MS-G].

References

[1] H. Brunner, *Volterra integral equations: an introduction to theory and applications,* Cambridge, Cambridge University Press, 2017

[2] C. Corduneanu, I. Sandberg, *Volterra equations and applications,* Boca Raton, CRC Press, 2000

[3] A. Unterreiter, *Volterra integral equation models for semiconductor devices,* Mathematical Methods in the Applied Sciences, 19(6), 425–450, 1996

[4] J.A. Ezquerro, M.A. Hernández, Á. A. Magreñán, A. Moysi, *On the existence and the approximation of solutions of Volterra integral equations of the second kind,* Applied Mathematics and Computation, 478, 128829, 2024

[5] J.A. Ezquerro, M.A. Hernández, *A quadratic approximation for solving nonlinear integral equations of Volterra type,* Journal of Computational and Applied Mathematics, 117366, 2026

[6] J.A. Ezquerro, M.A. Hernández, *An optimization of Chebyshev's method,* Journal of Complexity, 25(4), 343–361, 2009

High-Order Multipoint Iterative Methods for Nonlinear Systems with Scalar Weight Functions

Maria Emilia Maldonado Machuca[1], Alicia Cordero, Juan Ramón Torregrosa and Paula Triguero-Navarro

I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

The solution of systems of nonlinear equations

$$F(x) = 0, \qquad F : \mathbb{R}^n \to \mathbb{R}^n,$$

constitutes one of the fundamental problems in modern numerical analysis, due to its wide presence in mathematical models arising from engineering, physics, and the analysis of complex systems. Unlike the scalar case, the coupled nature of the equations, the possible singularity of the Jacobian matrix, and the sensitivity to the initial approximation significantly complicate their resolution.

The generalized Newton method constitutes the reference scheme in this context, thanks to its local quadratic convergence. However, its global efficiency is limited by several factors, including the high computational cost associated with the evaluation and factorization of the Jacobian matrix, its strong dependence on an initial approximation sufficiently close to the solution, and the impossibility of surpassing quadratic order without increasing the number of functional evaluations. These limitations have motivated the development of multipoint iterative methods, whose objective is to increase the order of convergence while keeping the computational cost under control.

In this context, the optimality principle established in the Kung and Traub conjecture [1] for nonlinear equations, and subsequently generalized to the vectorial case by Cordero et al. [2], provides a fundamental theoretical framework. This principle establishes that, for a multipoint iterative Newton-type method requiring $d = k_1 + k_2$ functional evaluations per iteration, where k_1 corresponds to the number of Jacobian matrix evaluations and k_2 to the number of functional evaluations, the optimal order is given by:

$$2^{k_1+k_2-1}, \quad \text{if } k_1 \leq k_2.$$

Furthermore, Weerakoon and Fernando [3] introduce the computational order of convergence (COC) to evaluate the convergence order of an iterative method, defined by

$$p \approx COC = \frac{\ln\left|\frac{x_{n+1}-\alpha}{x_n-\alpha}\right|}{\ln\left|\frac{x_n-\alpha}{x_{n-1}-\alpha}\right|}, \quad n = 1, 2, \ldots,$$

[1] memalmac@posgrado.upv.es

Since the exact value of the root α is unknown in practice, the approximate computational order of convergence (ACOC), introduced in 2007 by Cordero and Torregrosa [4], is used, defined as

$$p \approx ACOC = \frac{\ln\left|\frac{x_{n+1}-x_n}{x_n-x_{n-1}}\right|}{\ln\left|\frac{x_n-x_{n-1}}{x_{n-1}-x_{n-2}}\right|}, \quad n = 2, 3, \ldots.$$

This definition takes into account the last three approximations of the root to estimate the convergence order of the method.

Additionally, in computational implementations, a simplified version based on consecutive increments is frequently used. This is known as the Petković Computational Local Order of Convergence (pCLOC), introduced in [5], and defined by

$$p \approx \text{pCLOC} = \frac{\ln|incr_{n+1}|}{\ln|incr_n|}, \quad n = 1, 2, \ldots,$$

where $incr_n = |x_n - x_{n-1}|$. This expression is derived under the asymptotic assumption that the error satisfies $e_{n+1} \approx e_n^p$, and it allows the order of convergence to be estimated directly from successive differences.

On the other hand, the evaluation of the performance of iterative methods is based on the use of efficiency indices. In particular, the Ostrowski efficiency index [6], defined as

$$IE = p^{\frac{1}{d}},$$

and the computational efficiency index [7], given by

$$IC = p^{\frac{1}{d+op}},$$

where p is the order of convergence of the method, d is the number of function evaluations per iteration, and op is the number of products/quotients per iteration. This makes it possible to rigorously quantify the relationship between the method's order of convergence and its computational cost.

Recently, the classical concept of optimality has been extended through the principle of maximum efficiency introduced in [8]. Within this framework, it is established that the minimum number of evaluations in a multistep Newton-type method must satisfy:

$$k_1 + k_2 \geq m + 1,$$

where m is the number of steps of the method.

From this condition, new classifications of iterative methods are introduced:

- Low-eval-cost methods, which achieve the minimum possible number of evaluations.
- Least-eval-cost methods, which additionally employ a frozen Jacobian and perform exactly one functional evaluation per step.

Furthermore, it is established that, among methods with the same order of convergence and number of steps, the most efficient are those of least-eval-cost type. This result provides a key constructive criterion for the design of new iterative methods.

Additionally, it is shown that the maximum attainable order by a damped composed Newton-type method with m steps is bounded by 2^m, which implies that the minimum number of steps required to reach an order P is:

$$\left\lceil \frac{\ln P}{\ln 2} \right\rceil.$$

Finally, the concept of maximum efficiency is introduced, defining as such those methods for which no other method exists with the same order of convergence and a higher efficiency index. This new theoretical framework significantly extends the classical notion of optimality, providing more precise tools for both the comparison and the design of high-order iterative methods with low computational cost.

In this line, recent works such as [9], [8], and [10] have proposed new strategies to increase the order of convergence without increasing the evaluation cost. In particular, in [9], the use of scalar weight functions constructed from ratios of norms of the form

$$\frac{\|F(y^{(k)})\|^2}{\|F(x^{(k)})\|^2},$$

is introduced, which are incorporated into the correction step of the method. This strategy allows increasing the order of convergence by exploiting previously computed evaluations, so that the increase in computational cost is practically negligible, as it only involves scalar operations.

Motivated by these ideas, in the present research a new multipoint predictor-corrector type iterative method incorporating scalar weight functions is developed, with the aim of simultaneously improving the order of convergence and computational efficiency.This method will be compared with the multipoint methods MOSC4 [5], MOSG4 [11], and MSH4 [12], all of convergence order 4; with MSFS5 [13], [14] and MSSH5 [9], of order 5; and with MSCT6 [13], [15], of order 6. Likewise, the Newton method, of order 2, will be included.

2 Methods

In this work, a new multipoint predictor-corrector type iterative method for solving nonlinear systems is proposed, based on the incorporation of carefully designed scalar weight functions.

The iterative scheme MN14 consists of a Newton-type predictor step

$$y^{(k)} = x^{(k)} - F'(x^{(k)})^{-1}F(x^{(k)}), \tag{1}$$

followed by the corrector step

$$x^{(k+1)} = y^{(k)} - F'(x^{(k)})^{-1}\big(H(\mu_k)F(x^{(k)}) + G(\nu_k)F(y^{(k)})\big), \tag{2}$$

where the scalar parameters are defined by

$$\mu_k = \frac{\|F(y^{(k)})\|^2}{\|F(x^{(k)})\|^2}, \qquad \nu_k = \frac{F(x^{(k)})^T F(y^{(k)})}{\|F(x^{(k)})\|^2},$$

with $\|\cdot\|$ denoting the Euclidean norm.

Let α be a solution of $F(x) = 0$. Under the standard assumptions $F \in C^2$ in a neighborhood of α and $F'(\alpha)$ nonsingular, the local error equation of the method can be written as

$$\begin{aligned} e_{k+1} = \Big((2G'(0) - 6 - \tfrac{1}{2}G''(0))C_2^3 + (12 - 7G'(0))C_2C_3 \\ + 2P^{-1}QC_2^2 + (8 - 4G'(0))C_3^2\Big)e_k^4 + O(e_k^5), \end{aligned}$$

where $e_k = x^{(k)} - \alpha$. This expression shows that the scheme achieves fourth-order convergence provided that suitable weight functions G and H are chosen.

In this work, we consider the following family of weight functions:

$$G(\nu) = 1 + a\nu, \qquad H(\mu) = (2 - a)\mu,$$

where $a \in \mathbb{R}$ is a free parameter.

From a computational point of view, the method requires:

$$d = n^2 + 2n,$$

functional evaluations per iteration, and

$$op = \frac{1}{3}n^3 + 2n^2 + \frac{8}{3}n + 2,$$

algebraic operations.

Consequently, the efficiency indices are given by:

$$IE = 4^{\frac{1}{n^2+2n}}, \qquad IC = 4^{\frac{1}{\frac{1}{3}n^3+3n^2+\frac{14}{3}n+2}}.$$

These values place the method within the class of minimum evaluation cost methods, aligned with the principles of maximum efficiency.

3 Results

To analyze the performance of the proposed method, both the classical efficiency index and the computational efficiency index have been graphically represented. The results are shown in Figure 1 and Figure 2, respectively.

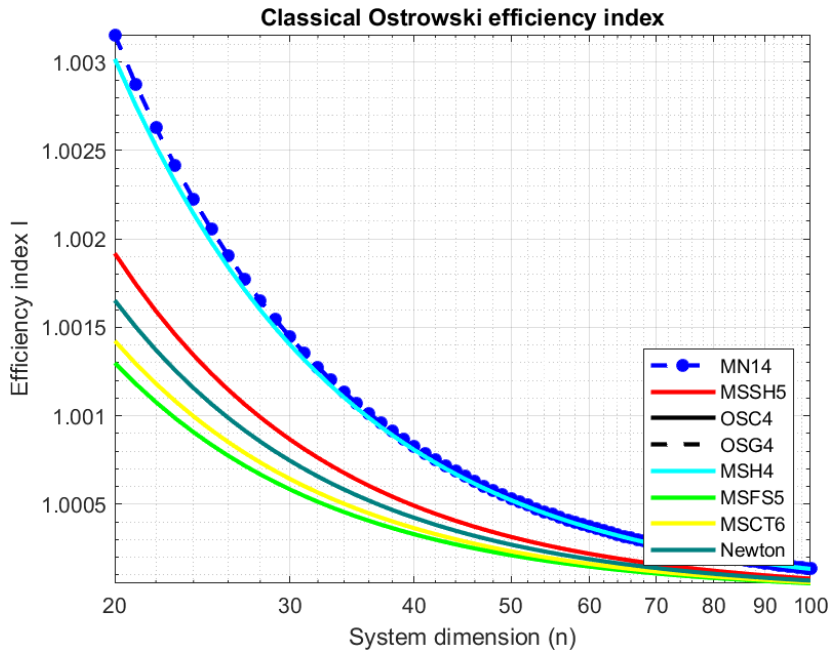


Figure 1: Ostrowski efficiency index of the considered iterative methods.

From the analysis of both figures, it is observed that the proposed method exhibits superior performance in terms of efficiency compared to other methods with four, five and six order of convergence considered in the literature. This result is justified by the structure of the method, which satisfies the minimum evaluation cost criterion.

In particular, the developed scheme employs only two functional evaluations per iteration, efficiently combined in the corrector step, together with the use of a frozen Jacobian obtained in

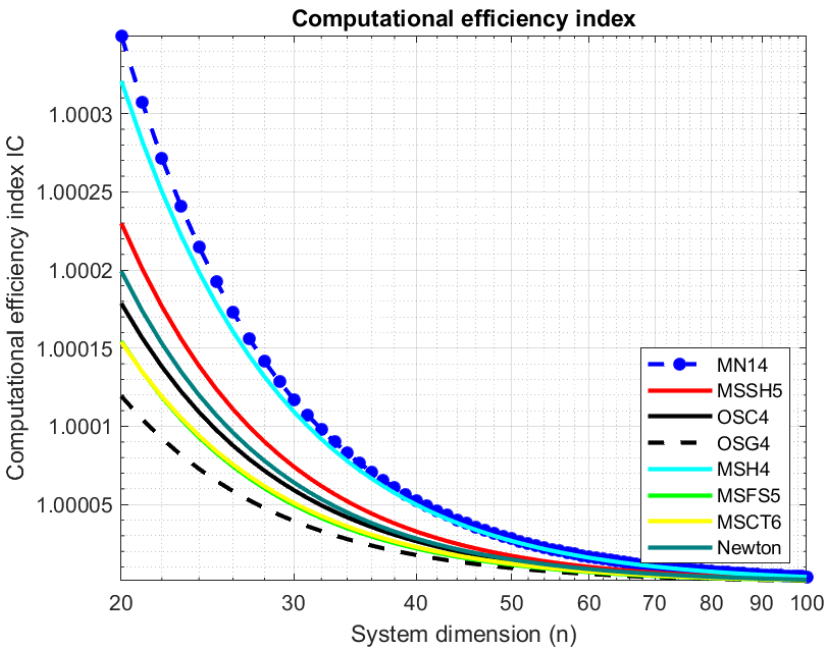


Figure 2: Computational efficiency index of the considered iterative methods.

the predictor step. This strategy significantly reduces the computational cost without sacrificing the order of convergence, which results in higher values for both the classical efficiency index and the computational efficiency index.

To further validate the behavior of the proposed method, numerical experiments were conducted on a nonlinear system of dimension $n = 20$, defined by:

$$F_i(x_1, \ldots, x_{20}) = x_i - \frac{3}{2} \sin\left(\sum_{\substack{j=1 \\ j \neq i}}^{20} x_j \right), \quad i = 1, \ldots, 20.$$

The initial approximation was chosen as $x^{(0)} = (1, 1, \ldots, 1)^T \in \mathbb{R}^{20}$.

The simulation uses the stopping criterion

$$\texttt{incr} + \texttt{incr2} > \texttt{tol},$$

where $\texttt{incr} = \|x^{(k+1)} - x^{(k)}\|$ measures the variation between consecutive iterations, and $\texttt{incr2} = \|F(x^{(k+1)})\|$ quantifies the system residual. Therefore, this condition implies that the iterative process continues as long as the sum of both errors is greater than the prescribed tolerance, simultaneously guaranteeing convergence in both the variable and the function.

Additionally, the restriction

$$\texttt{iter} < \texttt{maxiter},$$

is imposed, which limits the maximum number of allowed iterations, acting as a safety mechanism to prevent infinite loops in case the method does not converge.

The simulations were performed using a stopping tolerance $\text{tol} = 10^{-80}$ and a maximum number of iterations $k_{\max} = 50$, which ensures an accurate analysis of the asymptotic behavior of the methods under consideration. The obtained results are shown in Table 1.

Method	Iterations	$\|x^{(k+1)} - x^{(k)}\|$	$\|F(x^{(k+1)})\|$	ACOC	pCLOC	Time
MN14	5	3.3987e-140	1.4202e-555	4.000	4.016	8.715
Newton	8	2.2352e-97	2.0811e-192	2.0000	2.0102	4.3380
MOSC4	4	2.4226e-37	2.6338e-89	3.540	2.465	61.518
MOSG4	4	1.1005e-39	1.6784e-93	3.224	2.438	65.178
MSH4	5	5.2428e-158	5.6378e-627	4.000	4.015	11.676
MSSH5	5	4.9327e-372	6.2392e-1854	5.000	5.009	8.835
MSFS5	10	1.6018e-133	4.1038e-264	2.000	2.003	23.729
MSCMT6	4	2.2605e-131	1.1147e-780	5.9999	6.030	12.491

Table 1: Numerical results for a nonlinear system of 20 equations

The obtained results show that the method systematically achieves the theoretical order $p = 4$, with experimental values very close to it, thus confirming the validity of the theoretical analysis. In comparison with Newton's method, a reduction in the number of iterations required to reach the desired accuracy is observed, which compensates for the moderate increase in the cost per iteration.

Likewise, the method exhibits robust and stable behavior compared to other multipoint methods in the literature. In particular, it is observed that, although higher-order methods exist, they require a greater number of iterations or present a significantly higher computational cost.

4 Conclusions

In this work, a new fourth-order multipoint iterative method for solving systems of nonlinear equations has been developed, based on the incorporation of scalar weight functions within a predictor-corrector scheme.

The proposed method improves the performance of Newton's method by increasing the order of convergence without increasing the number of evaluations of the Jacobian matrix, which allows maintaining a controlled computational cost. From a theoretical point of view, its local convergence has been proven and its behavior has been characterized through the corresponding error equation.

Moreover, the efficiency analysis, both in terms of the Ostrowski index and the computational efficiency index, shows that the method exhibits superior performance compared to other schemes of similar order considered in the literature. This improvement is due to its low evaluation cost structure and the use of a frozen Jacobian, which allows optimizing the balance between accuracy and computational cost.

On the other hand, the numerical results confirm the validity of the theoretical analysis, showing that the method achieves the expected order of convergence, as well as robust and stable behavior in problems of moderate-high dimension. In comparison with classical and higher-order methods, the proposed scheme stands out for its ability to reduce the number of iterations without incurring a significant increase in the cost per iteration.

Overall, this work contributes to the development of highly efficient iterative methods, providing a systematic framework for the design of new schemes that satisfy optimality and maximum efficiency criteria in the solution of nonlinear systems.

References

[1] H.-T. Kung and J. F. Traub, "Optimal order of one-point and multipoint iteration," *Journal of the ACM*, vol. 21, no. 4, pp. 643–651, 1974.

[2] V. Arroyo, A. Cordero, and J. R. Torregrosa, "Approximation of artificial satellites' preliminary orbits: The efficiency challenge," *Mathematical and Computer Modelling*, vol. 54, no. 7-8, pp. 1802–1807, 2011.

[3] S. Weerakoon and T. Fernando, "A variant of newton's method with accelerated third-order convergence," *Applied Mathematics Letters*, vol. 13, no. 8, pp. 87–93, 2000.

[4] A. Cordero and J. R. Torregrosa, "Variants of newton's method using fifth-order quadrature formulas," *Applied Mathematics and Computation*, vol. 190, no. 1, pp. 686–698, 2007.

[5] M. Grau-Sánchez, Á. Grau, and M. Noguera, "Ostrowski type methods for solving systems of nonlinear equations," *Applied Mathematics and Computation*, vol. 218, no. 6, pp. 2377–2385, 2011.

[6] J. F. Traub, *Iterative methods for the solution of equations*. American Mathematical Soc., 1982, vol. 312.

[7] A. Cordero, J. L. Hueso, E. Martínez, and J. R. Torregrosa, "A modified newton-jarratt's composition," *Numerical Algorithms*, vol. 55, no. 1, pp. 87–99, 2010.

[8] A. Cordero, R. V. Rojas-Hiciano, J. R. Torregrosa, and M. P. Vassileva, "Maximally efficient damped composed newton-type methods to solve nonlinear systems of equations," *Applied Mathematics and Computation*, vol. 492, p. 129 231, 2025.

[9] H. Singh, J. R. Sharma, and S. Kumar, "A simple yet efficient two-step fifth-order weighted-newton method for nonlinear models," *Numerical Algorithms*, vol. 93, no. 1, pp. 203–225, 2023.

[10] A. Cordero, J. R. Torregrosa, and P. Triguero-Navarro, "First optimal vectorial eighth-order iterative scheme for solving non-linear systems," *Applied Mathematics and Computation*, vol. 498, p. 129 401, 2025, ISSN: 0096-3003.

[11] S. Artidiello, "Diseño, implementación y convergencia de métodos iterativos para resolver ecuaciones y sistemas no lineales utilizando funciones peso," PhD thesis, Universitat Politècnica de València, València, Spain, 2014.

[12] J. L. Hueso, E. Martínez, and J. R. Torregrosa, "Third and fourth order iterative methods free from second derivative for nonlinear systems," *Applied mathematics and Computation*, vol. 211, no. 1, pp. 190–197, 2009.

[13] A. Cordero, J. L. Hueso, E. Martínez, and J. R. Torregrosa, "Increasing the convergence order of an iterative method for nonlinear systems," *Applied Mathematics Letters*, vol. 25, no. 12, pp. 2369–2374, 2012, ISSN: 0893-9659.

[14] M. Frontini and E. Sormani, "Some variant of newton's method with third-order convergence," *Applied Mathematics and Computation*, vol. 140, no. 2-3, pp. 419–426, 2003.

[15] A. Cordero, E. Martínez, and J. R. Torregrosa, "Iterative methods of order four and five for systems of nonlinear equations," *Journal of Computational and Applied Mathematics*, vol. 231, no. 2, pp. 541–551, 2009.

Exploring Dynamical Planes in Nonlinear Equations and Systems: Derivative-Based versus Derivative-Free Iterative Schemes

Francisco I. Chicharro and Neus Garrido[1]

I.U. Matemática Multidisciplinar, Universitat Politècnica de València, Camí de Vera s/n, 46022 - València, Spain.

1 Introduction

Many natural processes are modeled by nonlinear equations of the form $f : D \subseteq \mathbb{C} \to \mathbb{C}$. In numerous practical scenarios, these equations do not admit exact solutions or cannot be expressed in closed form. A robust approach to approximating a solution α is the use of iterative methods, which generate a sequence that converges toward the desired root with arbitrary precision.

The most prominent iterative procedure is Newton's method, defined by

$$z^{(k+1)} = z^{(k)} - \left[F'\left(z^{(k)}\right)\right]^{-1} F\left(z^{(k)}\right), \quad k = 0, 1, \ldots \tag{1}$$

Newton's method is characterized by quadratic convergence and exhibits significant stability depending on the choice of the initial approximation x_0.

Standard convergence theorems for iterative schemes typically require the initial estimate x_0 to be sufficiently close to the solution α. To analyze the global behavior and identify all initial approximations that lead to convergence, one can employ a dynamical analysis of the rational operator associated with the method when applied to a polynomial. This analysis is grounded in the theory of complex dynamical systems [7, 8].

The literature contains extensive stability analyses for various numerical scenarios, including derivative-free methods, methods with memory [2], techniques for multiple roots, and simultaneous root-finding methods [9], as well as their extensions for solving systems of nonlinear equations [5].

From this perspective, dynamical planes provide a visual tool to identify all initial estimates that converge to the expected solutions by depicting their basins of attraction. Furthermore, the existence of strange fixed points and free critical points significantly influences the convergence topology and the reliability of initial guesses.

In this work, we contrast the effects of derivative-based versus derivative-free methods, as well as the impact of multi-step compositions. We analyze the complex dynamics of iterative schemes involving derivatives and compare them with counterparts where derivatives are replaced by divided differences or similar approximations.

[1] neugarsa@mat.upv.es

2 Derivative-Based versus Derivative-Free Dynamical Comparison

In this comparative study, we forgo the use of the scaling theorem [3, 4], as the derivative-free methods considered here do not satisfy its underlying assumptions. Consequently, we perform a direct comparative analysis by applying the iterative methods to the quadratic polynomial $p(z) = z^2 - 1$, with roots at ± 1.

The interpretation of the resulting dynamical planes focuses on the identification of strange fixed points and free critical points, defined as follows.

Definition 1 (Fixed and critical points) *Let $R : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ be a rational function on the Riemann sphere.*

- *A point $z_0 \in \hat{\mathbb{C}}$ is a **fixed point** if $R(z_0) = z_0$. Fixed points that do not coincide with the roots of the polynomial are termed **strange fixed points**. A fixed point is classified as attracting if $|R'(z_0)| < 1$, repelling if $|R'(z_0)| > 1$, or neutral if $|R'(z_0)| = 1$.*
- *A point $z_0 \in \hat{\mathbb{C}}$ is a **critical point** if $R'(z_0) = 0$. Critical points that are not roots of the polynomial are referred to as **free critical points**.*

Once identified, we focus on the points that compromise stability: strange attracting fixed points and free critical points. Graphical representations follow the conventions established in [**?**]. The stopping criterion is set at $\|z_k - z^*\| < 10^{-6}$, with a maximum of 40 iterations; otherwise, the method is deemed to have failed. Basins of attraction for 1 and -1 are colored orange and blue, respectively, while non-convergence is shown in black. The domain of initial guesses is the square $[-5, 5] \times [-5, 5]$. Roots are marked with white stars, strange fixed points with white squares, and free critical points with white circles. All simulations were performed using MATLAB R2026a.

2.1 One-step optimal methods for nonlinear equations

We compare Newton's method (1) and Steffensen's method:

$$z_{k+1} = z_k - \frac{f(z_k)}{f[z_k + f(z_k), z_k]}, \quad k = 0, 1, \ldots \tag{2}$$

where the derivative in Newton's operator is approximated by a divided difference.

While Newton's method lacks points that interfere with convergence for this polynomial, Steffensen's rational operator $S(z) = \frac{z^3+z^2-z+1}{z^2+2z-1}$ possesses critical points at $z = -2 \pm \sqrt{3}$.

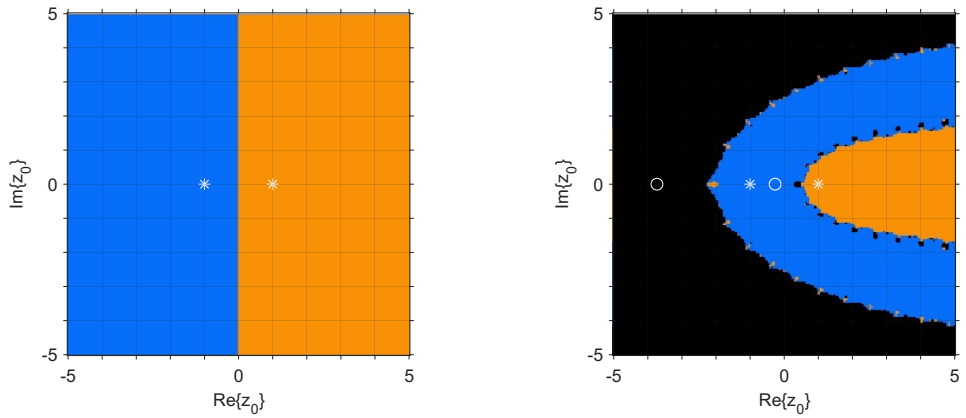
Figure 1 illustrates the basins of attraction for both schemes. In the Newton plane, every initial guess converges to its nearest root. Conversely, Steffensen's method exhibits a significant region of non-convergence. Notably, the free critical point $z = -2 - \sqrt{3}$ generates its own basin of attraction, leading to numerical instability.

2.2 Two-step optimal methods for nonlinear equations

We further analyze the composition of these methods, testing all four combinations of Newton and Steffensen over two iterative steps.

Method N2 is the composition of Newton's method with itself, resulting in

$$y_k = z_k - \frac{f(z_k)}{f'(z_k)}, \quad z_{k+1} = y_k - \frac{f(y_k)}{f'(y_k)}, \quad k = 0, 1, \ldots \tag{3}$$



(a) (b)

Figure 1: Dynamical planes associated with (a) Newton's method, and (b) Steffensen's method.

Method NS is the composition of Newton's method in the first step with Steffensen's method, resulting in

$$y_k = z_k - \frac{f(z_k)}{f'(z_k)}, \quad z_{k+1} = y_k - \frac{f(y_k)}{f[y_k + f(y_k), y_k]}, \quad k = 0, 1, \ldots \tag{4}$$

Method SN swaps the order of method NS, resulting in

$$y_k = z_k - \frac{f(z_k)}{f[z_k + f(z_k), z_k]}, \quad z_{k+1} = y_k - \frac{f(y_k)}{f'(y_k)}, \quad k = 0, 1, \ldots \tag{5}$$

Finally, method S2 is the composition of Steffensen's method with itself, resulting in

$$y_k = z_k - \frac{f(z_k)}{f[z_k + f(z_k), z_k]}, \quad z_{k+1} = y_k - \frac{f(y_k)}{f[y_k + f(y_k), y_k]}, \quad k = 0, 1, \ldots \tag{6}$$

For these compositions, every method exhibits strange fixed points, and several possess free critical points.

The operator $N2(z) = \frac{z^4+6z^2+1}{4z(z^2+1)}$ has strange fixed points at $z = \pm i\sqrt{\frac{1}{3}}$, which are repelling since $N2'\left(\pm i\sqrt{\frac{1}{3}}\right) = 4 > 1$.

The operator $NS(z) = \frac{z^6+2z^5-z^4+12z^3-z^2+2z+1}{2z(z^4+4z^3-2z^2+4z+1)}$ has the strange fixed points $z^4 + 6z^3 - 2z^2 + 2z + 1 = 0$, that are computed numerically resulting in two real values and one pair of complex conjugated values. The only attracting point is $z \approx -6.36$. There are four critical points: three of them at the basins of the roots and one of them at the basin of $z \approx -6.36$.

The operator $SN(z) = \frac{z^6+2z^5+4z^3+5z^2-6z+2}{2(z^5+3z^4-2z^2+3z-1)}$ has two real values and one pair of complex conjugated values. The only attracting strange fixed point is $z \approx -3.26$. There are four critical points: three of them at the basins of the roots and one of them at the basin of $z \approx -3.26$.

The operator $S2(z) = \frac{z^9+4z^8+3z^7-4z^6+3z^5+18z^4-3z^3-16z^2+12z-2}{z^8+6z^7+11z^6-15z^4+6z^3+17z^2-12z+2}$ has two real values and two pairs of complex conjugated values, but no one is attracting as in one-step Steffensen's method. However, it presents ten free critical points, four of which belong to the basin of attraction of infinity.

Figure 2 confirms that N2 maintains the high stability of the one-step Newton method. For the mixed compositions, using Newton in the first step (NS) yields superior convergence compared to starting with Steffensen (SN). The S2 method shows the most chaotic behavior due to the high density of free critical points.

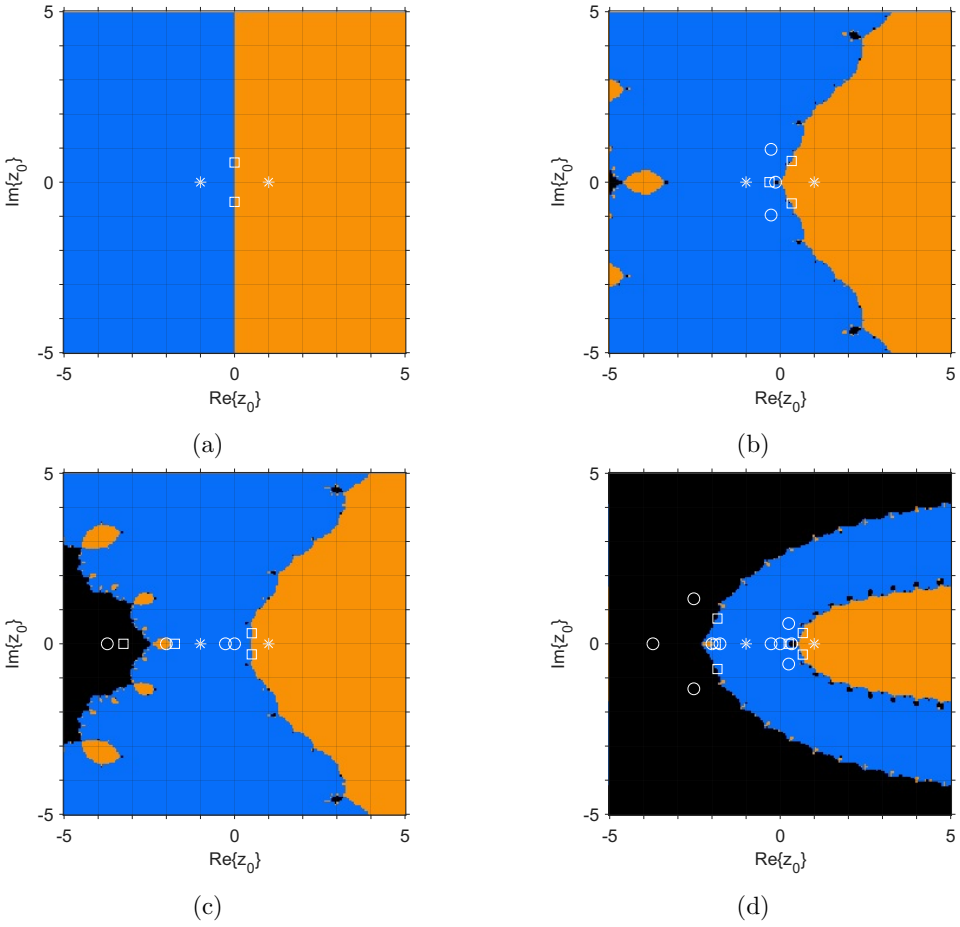


Figure 2: Dynamical planes associated with (a) N2, (b) NS, (c) SN, and (d) S2.

2.3 Methods with memory for nonlinear equations

Methods with memory are evaluated via their real dynamics [5]. Consequently, the dynamical planes are plotted with x_1 and x_0 on the axes. We compare the derivative-based method (MD) from [1, 2] with its derivative-free version (MF).

Expression of method MD is

$$x_{k+1} = x_k - \frac{1}{\lambda_k + \frac{f'(x_k)}{f(x_k)}}, \quad k = 1, 2, \ldots \tag{7}$$

where $\lambda_k = \dfrac{f'(x_{k-1}) - f'(x_k)}{2f'(x_{k-1})(x_k - x_{k-1})}$. Its error equation is

$$e_{k+1} = \left(2c_2^2 - \frac{3}{2}c_3\right) e_k^2 e_{k-1}^2 + \mathcal{O}_4(e_k, e_{k-1}),$$

yielding to order of convergence $1 + \sqrt{2}$, where $c_k = \frac{1}{k!}\frac{f^{(k)}(\alpha)}{f'(\alpha)}$ and $e_{k+q} = x_{k+q} - \alpha, q = -1, 0, 1$. Method MF replaces the derivatives of (7) by divided differences, having the iterative expression

$$x_{k+1} = x_k - \frac{1}{\mu_k + \frac{f[x_k + f^2(x_k), x_k]}{f(x_k)}}, \quad k = 1, 2, \ldots \tag{8}$$

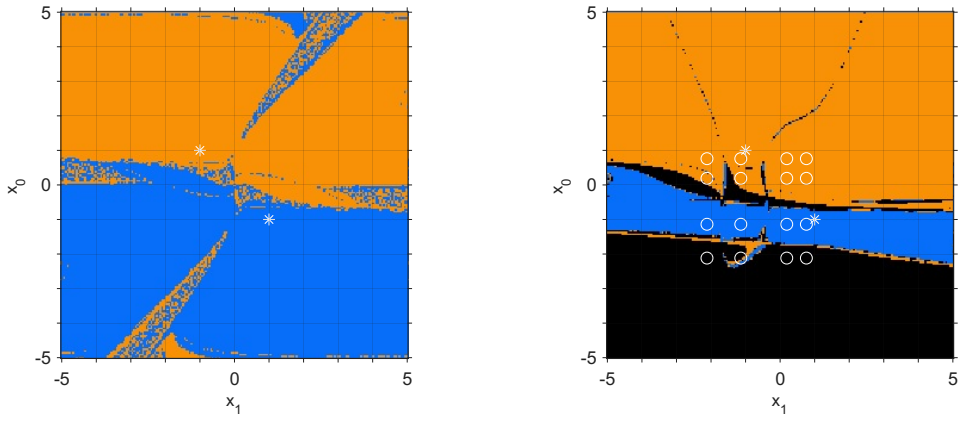
where $\mu_k = \dfrac{f\left[x_{k-1}+f^2(x_{k-1}),x_{k-1}\right]-f\left[x_k+f^2\left(x_k\right),x_k\right]}{2f\left[x_{k-1}+f^2(x_{k-1}),x_{k-1}\right]\left(x_k-x_{k-1}\right)}$. Its error equation is

$$e_{k+1}=\left(-\frac{c_2}{2}+2c_2^2-\frac{3}{2}c_3\right)e_k^2e_{k-1}^2+\mathcal{O}_4\left(e_k,e_{k-1}\right),$$

yielding to order of convergence $1+\sqrt{2}$, where $c_k=\frac{1}{k!}\frac{f^{(k)}(\alpha)}{f'(\alpha)}$ and $e_{k+q}=x_{k+q}-\alpha, q=-1,0,1$.

The one-dimensional operator of MD, $\tilde{MD}(x)=\frac{x^2+1}{2x}$, is free of strange fixed or free critical points. In contrast, the one-dimensional operator of MF, $\tilde{MF}(x)=\frac{x^5-2x^3+x^2+x+1}{x^4-2x^2+2x+1}$, possesses four critical points that are the four real roots of the polynomial $x^6-3x^4+6x^3+3x^2-6x+1$.

Figure 3 represents the basins of attraction of the iterative methods with memory. The limit for convergence has been increaded to 300 iterations. The dynamical plane of MD shows a high-stability behaviour. The absence of points that affect convergence is the key point. However, method MF show black regions motivated by divergence to infinity. In fact, one of the four free critical points (and its combinations) lies in the black region, that diverges to infinity.



(a) (b)

Figure 3: Dynamical planes associated with (a) MD, and (b) MF.

2.4 Methods for systems of nonlinear equations

The dynamics of multidimensional methods can also be analyzed in a similar way to methods with memory. To analyze the stability of fixed points in vectorial rational functions associated with schemes for solving nonlinear systems, tools from multidimensional real dynamical systems are also employed [6]. One way to construct dynamical planes for n-dimensional methods is to specialize to the case $n=2$, so that one variable can be represented along each axis of the plane. Nevertheless, the analytical results obtained from the dynamical study can be generalized to higher dimensions. The fixed points of the vectorial operator $R:\mathbb{R}^2\longrightarrow\mathbb{R}^2$ are obtained in the same way as the solution to the system $R(x_1,x_2)=(x_1,x_2)$. The strange fixed points and free critical points appear again with the same meaning as in the scalar case. However, the eigenvalues of the Jacobian matrix evaluated on each fixed point are what determine the asymptotic behavior of them according to [10].

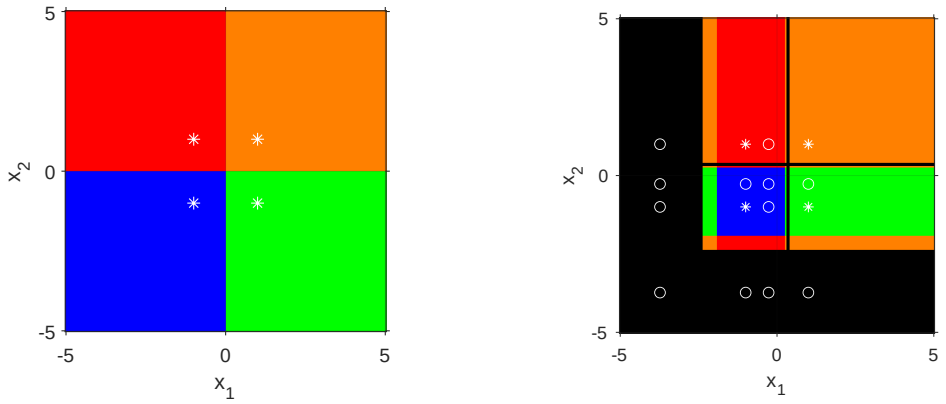
We will now examine how the stability of the scalar methods of Newton and Steffensen, as well as their two-step scheme combinations, is affected when they are adapted for solving multidimensional problems. To perform an analogous process, we will apply the methods to the polynomial system

$p(x_1, x_2) = \begin{cases} x_1^2 - 1 \\ x_2^2 - 1 \end{cases}$, which has four real roots $(\pm 1, \pm 1)$. The implementation of the dynamical planes follows the same structure and rules as in Figures 1 and 2, but we now use the following colour coding for the basins of attraction of the four roots of $p(x)$: orange, blue, green and red, respectively, for $(1,1)$, $(-1,-1)$, $(1,-1)$ and $(-1,1)$.

We first apply the multidimensional Newton and Steffensen methods to the polynomial system under consideration. The resulting operators are

$$NN(x_1, x_2) = \begin{pmatrix} \frac{x_1^2+1}{2x_1} \\ \frac{x_2^2+1}{2x_2} \end{pmatrix}, \qquad SS(x_1, x_2) = \begin{pmatrix} \frac{x_1^3+x_1^2-x_1+1}{x_1^2+2x_1-1} \\ \frac{x_2^3+x_2^2-x_2+1}{x_2^2+2x_2-1} \end{pmatrix}.$$

In neither case are any strange fixed points observed, and the roots of $p(x)$ are the only superattracting fixed points. However, we observe 12 free critical points for Steffensen's scheme: $(-2 \pm \sqrt{3}, \pm 1)$, $(\pm 1, -2 \pm \sqrt{3})$, and $(-2 \pm \sqrt{3}, 2 \pm \sqrt{3})$. They are plotted in Figure 4 as white circles, and it can be seen that seven of them have divergent basins of attraction, resulting in large black areas. Conversely, in Newton's method, only convergence to the roots is observed.



(a) (b)

Figure 4: Dynamical planes associated with (a) multidimensional Newton's method, and (b) multidimensional Steffensen's method.

In the composition of Newton with itself, although the behavior is similar to that of the scalar case, unlike the latter, the vectorial operator has no strange real fixed points. The dynamical plot shown in Figure 5(a) is the same as that obtained using the multidimensional Newton's scheme. The two-dimensional operators for $N2$ and NS methods, denoted respectively $NN2$ and $NNSS$, are the following:

$$NN2(x_1, x_2) = \begin{pmatrix} \frac{x_1^4+6x_1^2+1}{4(x_1^3+x_1)} \\ \frac{x_2^4+6x_2^2+1}{4(x_2^3+x_2)} \end{pmatrix}, \qquad NNSS(x_1, x_2) = \begin{pmatrix} \frac{x_1^6+2x_1^5-x_1^4+12x_1^3-x_1^2+2x_1+1}{2x_1^5+8x_1^4-4x_1^3+8x_1^2+2x_1} \\ \frac{x_2^6+2x_2^5-x_2^4+12x_2^3-x_2^2+2x_2+1}{2x_2^5+8x_2^4-4x_2^3+8x_2^2+2x_2} \end{pmatrix}.$$

We obtain for $NNSS$ operator 12 strange fixed points, being 5 of which 5 are attracting points, 5 are repelling points, and the other two behave as saddle points. In addition, we obtain 12 free critical points.

The dynamical analysis of the $SSNN$ and $SS2$ methods is, as might be expected, more chaotic than the results obtained using Newton's method in the first step of the composition. However, the roots of the polynomial system are always superattracting fixed points. In addition to the

roots of the polynomial system, $SSNN$'s scheme presents 12 strange fixed points, seven of them in the basins of attraction of infinity. The others belong to the boundaries between the basins of attraction. Among them, five are attracting points, one is a repelling point and six are saddle points. Moreover, there is presence of 32 free critical points, but only some of them belong to the basin of attraction of the roots. As a consequence, we can see large areas represented in black on the dynamical plane of Figure 5(c). Finally, the stability of the $SS2$ method is very similar to that of the $SSNN$ scheme, as can be seen in the dynamical planes of Figures 5(c) and 5(d). The two-dimensional operator of the $SS2$'s method also has 12 strange fixed points and 5 free critical points, in which both belong to divergent basins of attraction. In both cases, the basins of attraction of the roots are much smaller than those of the $NN2$ and $NNSS$ methods, indicating less stable behavior.

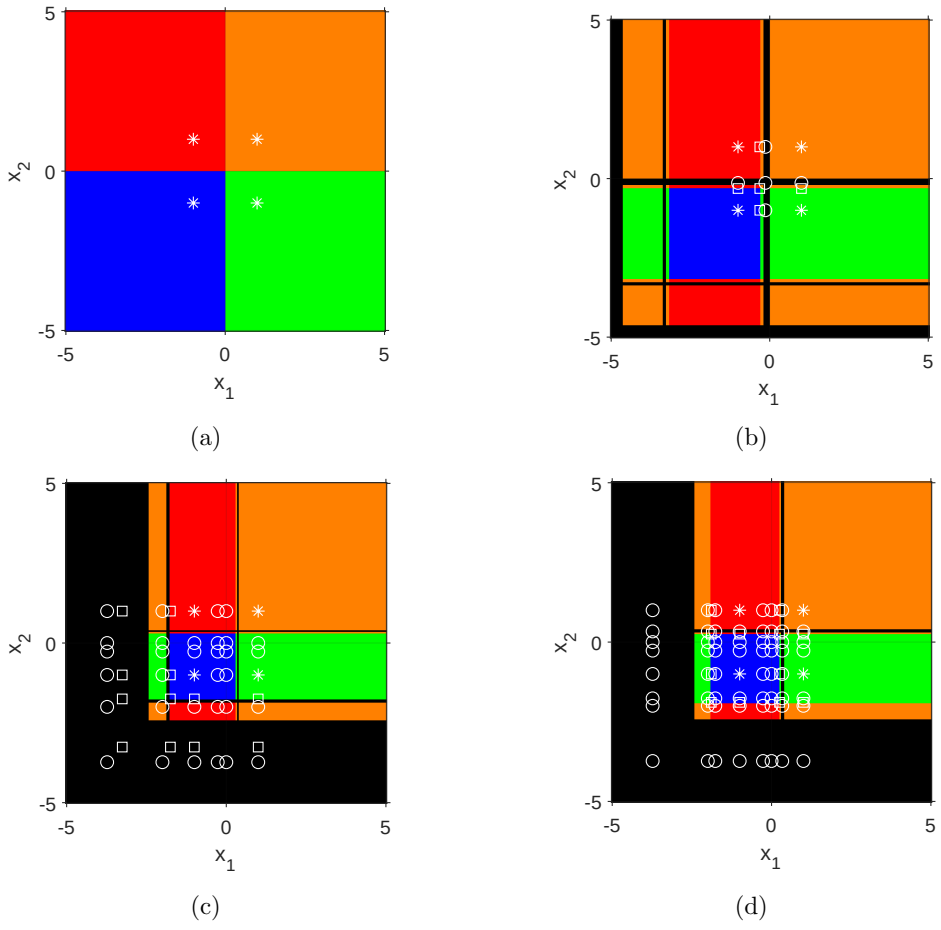


Figure 5: Dynamical planes associated with (a) NN2, (b) NNSS, (c) SSNN, and (d) SS2.

3 Conclusions

The transition from derivative-based methods to derivative-free schemes generally introduces greater dynamical complexity. While Newton's method exhibits high stability and no points affecting convergence for the test polynomial, its derivative-free counterpart, Steffensen's method, has free critical points that lead to significant regions of non-convergence.

In two-step methods, the order of composition is critical for stability. Composing Newton's method with itself (N2) preserves the high stability of the one-step version. However, in mixed compositions, using Newton's method in the first step followed by Steffensen's method (NS) produces larger convergence regions and better performance than the inverse order (SN).

Derivative-based methods with memory (MD) exhibit superior stability because they lack strange and free critical points. In contrast, the derivative-free version (MF) introduces real free critical points that facilitate divergence to infinity, resulting in unstable black regions within the dynamical plane.

The composition of multiple derivative-free steps (S2) significantly hinders convergence. The high density of free critical points in such schemes creates a more chaotic dynamical plane, making the choice of initial approximations far more restrictive.

In summary, while derivative-free methods are valuable when derivatives are unavailable or costly to compute, they inherently carry a higher risk of numerical instability. The presence of free critical points and strange attracting fixed points serves as a primary indicator of reduced reliability in these iterative schemes.

Acknowledgments

Funded with IMM Research Funding Call 2026 – Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València

References

[1] Wu, X., A new continuation Newton-like method and its deformation *Appl. Math. Comput.* 112:75–78, 2000.

[2] Wang, X., Zhang, T., A new family of Newton-type iterative methods with and without memory for solving nonlinear equations *Calcolo* 51:1–15, 2014.

[3] Curry, J., Garnet, L., Sullivan, D., On the iteration of a rational function: computer experiments with Newton's method *Comm. Math. Phys.* 91:267–277, 1983.

[4] Amat, S., Busquier, S., Plaza, S., Review of some iterative root-finding methods from a dynamical point of view *Sci. Ser. A Math. Sci.* 10:3–35, 2004.

[5] Campos, B., Cordero, A., Torregrosa, J.R., Vindel, P., A multidimensional dynamical approach to iterative methods with memory *Appl. Math. Comput.* 271:701–715, 2015.

[6] Campos, B., Soleymani, F., Torregrosa, J.R., Dynamical analysis of iterative methods for nonlinear systems or how to deal with the dimension? *Appl. Math. Comput.* 244:398–412, 2014.

[7] Blanchard, P., Complex analytic dynamics on the Riemann sphere *Bulletin of the American Mathematical Society* 11:85–141, 1984.

[8] Devaney, R.L. An introduction to chaotic dynamical systems. Addison-Wesley, 1989.

[9] Reinke, B., Diverging orbits for the Ehrlich–Aberth and the Weierstrass root finders *Proceedings of the AMS* 150:1287–1300, 2022.

[10] Robinson, R.C. An introduction to dynamical systems: continuous and discrete. American Mathematical Society, 2012.

Part IV

Mathematical Methods for Engineering Problems

D-splitting. A new paradigm for low-storage RK methods.

Sergio Blanes[♭], Abutalib Elsaid Mohammed[♭], and Alejandro Escorihuela-Tomàs[♮1]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

(♮) Departament de Matemàtiques and IMAC, Universitat Jaume I
Av. Sos Baynat s/n, Castelló de la Plana, Spain.

1 Introduction

Runge-Kutta methods are widely known and used for the numerical integration of differential equations of the form $\dot{x} = f(x)$. When advancing one step, these methods use several intermediate values of x, denoted by k_i, and in general there are as many k_i as stages in the method. This can be a drawback for problems where the dimension of x is very large, as it may lead to excessive memory usage. For this reason, Runge-Kutta methods have been developed that can be implemented using only two or three registers per iteration. These are known as low-storage Runge-Kutta methods. In this presentation, we introduce a new approach in which splitting methods can be used as low-storage methods, and we discuss the advantages they offer compared to classical low-storage Runge-Kutta methods.

2 Splitting methods

Splitting methods can be applied to problems of the type $\dot{x} = f(x)$, where the field $f(x)$ can be separated into two or more parts, $f(x) = f^{[A]}(x) + f^{[B]}(x)$. Each subproblem $\dot{x} = f^{[i]}(x)$ (with $i = A, B$) must be exactly integrable, with $x(t) = \varphi_t^{[i]}(x_0)$ being the exact solution. By fulfilling these premises and composing the exact solutions of each subproblem, we can obtain an approximate solution to the complete problem.

The best-know splitting method is the Strang splitting:

$$S_h = \varphi_{\frac{h}{2}}^{[A]} \circ \varphi_h^{[B]} \circ \varphi_{\frac{h}{2}}^{[A]},$$

which produces a second-order numerical approximation to the solution. However, splitting methods can be constructed with further function compositions:

$$\Phi_h = \varphi_{a_1 h}^{[A]} \circ \varphi_{b_1 h}^{[B]} \circ \varphi_{a_2 h}^{[A]} \circ \varphi_{b_2 h}^{[B]} \circ \cdots \circ \varphi_{a_s h}^{[A]} \circ \varphi_{b_s h}^{[B]},$$

where the coefficients a_i, b_i can be chosen such that we obtain a numerical solution with less error compared to the exact solution.

[1] alescori@uji.es

3 D-splitting methods

When a problem cannot be separated into two parts, we can force the separation by doubling the phase space:

$$y' = f(z),$$
$$z' = f(y),$$

which is equivalent to

$$\frac{d}{dt}\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ f(y) \end{pmatrix} + \begin{pmatrix} f(z) \\ 0 \end{pmatrix} = f^{[A]}(y) + f^{[B]}(z),$$

with the initial conditions $y(t_0) = z(t_0) = x_0$. Splitting methods specifically designed for when we double the phase space are called D-splitting methods.

4 Conclusions

Splitting methods offer the advantage of only requiring two variables in memory at each integration step, unlike general RK methods that need to store the k_i vectors as mentioned in the introduction. For problems where the dimension of the vector x is very large, RK methods have been designed that only require 2 or 3 intermediate variables per step. These methods are known as low-storage RK methods. Splitting methods in the doubled phase space can be viewed as low-storage RK methods, which provides the advantage that there are already many existing splitting methods in the literature that can be applied.

Acknowledgments

The work of both authors is supported by Ministerio de Ciencia e Innovación (Spain) through project PID2022- 136585NB-C21, MCIN/AEI/10.13039/501100011033/FEDER, UE. AE-T acknowledge financial support by Universitat Jaume I through project UJI-2025-06.

Bibliography

[1] S. Blanes and F. Casas, A Concise Introduction to Geometric Numerical Integration, Second revised ed. CRC Press, Boca Raton, 2025.

[2] S. Blanes, F. Casas, A. Murua, Splitting methods for differential equations, Acta Numerica (204), 1-161.

[3] S.Blanes, A. Escorihuela-Tomàs, D-splitting methods: $2N$-storage embedded explicit pseudo-geometric Runge-Kutta methods using splitting methods,. arXiv preprint (2026), arXiv:2604.03457.

[4] M.H. Carpenter and C.A. Kennedy, Fourth-order 2N-storage Runge-Kutta schemes, Technical Report NASA-TM-109112, NASA (1994)

[5] C.A. Kennedy, M.H. Carpenter, and R.M. Lewis, Low-storage, explicit Runge–Kutta schemes for the compressible Navier–Stokes equations, Appl. Num. Math. 35 (2000), 177-219.

Quasi-normal modes for flexural waves on demand

Marta Gómez Pérez[♭], Marc Martí-Sabaté[♭] and Vicente Romero-García[♭1]

(♭) I.U. de Matemática Pura y Aplicada (IUMPA), Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

The vibration characteristics of elastic plates are traditionally described in terms of normal modes, obtained by solving a self-adjoint eigenvalue problem under bounded domain conditions [1]. In such settings, the governing operator is Hermitian, leading to real eigenfrequencies and orthogonal mode shapes that form a complete basis for the structural response. For thin plates, the transverse displacement field is governed by the biharmonic equation

$$D\nabla^4 w(\mathbf{x}) - \rho h\omega^2 w(\mathbf{x}) = 0, \tag{1}$$

where $w(\mathbf{x})$ is the out-of-plane displacement, D is the bending stiffness, ρ the material density, and h the plate thickness. However, this framework ceases to apply when the system departs from energy-conserving and/or closed configurations. The governing operator becomes non-Hermitian whenever energy is allowed to leave the domain or be dissipated, as occurs under radiation (outgoing wave) boundary conditions, coupling to an unbounded surrounding medium, the presence of damping, or more generally any mechanism that breaks time-reversal symmetry. In these cases, the modal structure fundamentally changes, and the dynamics are more appropriately described by quasi-normal modes, characterized by complex frequencies that encode both oscillation and decay due to the non-Hermiticity of the governing operator.

This work explores the deliberate design of spatially distributed masses on a plate to engineer its quasi-normal modal spectrum. Specifically, we address the inverse problem of determining both the position and magnitude of discrete masses such that the system exhibits a prescribed quasi-normal mode at a target complex frequency. By introducing localized inertial perturbations, the originally conservative system is transformed into an effectively non-Hermitian one, where energy leakage and modal damping emerge naturally. This provides a controllable pathway to bridge the classical theory of normal modes in bounded plates with the richer framework of quasi-normal modes in open systems.

[1] virogar1@mat.upv.es

2 Methods

2.1 Scattering matrix and quasi-normal modes

By applying a Multiple Scattering framework [2], an infinite system of equations is obtained that can be written in the following shape

$$\begin{pmatrix} 1 & -T_2G_0(\mathbf{R}_1-\mathbf{R}_2) & \cdots & -T_nG_0(\mathbf{R}_1-\mathbf{R}_n) \\ -T_1G_0(\mathbf{R}_2-\mathbf{R}_1) & 1 & \cdots & -T_nG_0(\mathbf{R}_2-\mathbf{R}_n) \\ \vdots & \vdots & \ddots & \vdots \\ -T_1G_0(\mathbf{R}_3-\mathbf{R}_1) & -T_2G_0(\mathbf{R}_3-\mathbf{R}_2) & \cdots & 1 \end{pmatrix} \begin{pmatrix} \psi_e(\mathbf{R}_1) \\ \psi_e(\mathbf{R}_2) \\ \vdots \\ \psi_e(\mathbf{R}_n) \end{pmatrix} = \begin{pmatrix} \psi_0(\mathbf{R}_1) \\ \psi_0(\mathbf{R}_2) \\ \vdots \\ \psi_0(\mathbf{R}_n) \end{pmatrix}. \quad (2)$$

where T_i encapsulates the properties of the i-th scatterer, $\mathbf{R}_i$ represents the spatial position of the i-th scatterer and G_0 represents the Green function. By truncating the system, a numerical solution can be obtained following standard methodologies [3]. The solution can be used to obtain the displacement in a given point by using

$$W_1(\mathbf{r}) = \psi_0(\mathbf{r}) + \sum_\alpha T_\alpha \psi_e(\mathbf{R}_\alpha) G_0(\mathbf{r}-\mathbf{R}_\alpha). \quad (3)$$

Definition 2.1 ("Scattering" matrix)**.** We define the scattering matrix as the matrix that relates the scattered amplitudes, $T_\beta\psi_e(\mathbf{R}_\beta)$, to the incident ones $\psi_0(\mathbf{R}_\alpha)$, and we denote it by $\mathcal{M}$. Its coefficients are given by

$$\mathcal{M}_{\alpha\beta} = [\delta_{\alpha,\beta} t_\beta^{-1} - G_0(\mathbf{R}_\alpha - \mathbf{R}_\beta)]T_\beta. \quad (4)$$

One the plate where the scatterers is known, the scattering matrix depends only on the properties of the scatterers, their spatial distribution, and the frequency.

Definition 2.2 (Quasi-Normal Modes)**.** The resonant modes or quasi-normal modes (QNMs) of a passive scattering system are defined as the non-trivial solutions of the homogeneous problem with outgoing boundary conditions, corresponding to complex frequencies ω_n such that

$$\det\big[\mathcal{M}(\omega_n)\big] = 0. \quad (5)$$

Each QNM with frequency $\omega_n = \omega_n' - i\gamma_n$ is characterized by a resonance frequency ω_n' and a decay rate γ_n. In practice, QNMs can be obtained numerically by searching for the zeros of $\det[\mathcal{M}(\omega)]$ in the complex ω plane [4]. These modes correspond to the natural modes of the complete structure. Physically, they manifest themselves as resonances; mathematically, they appear as zeros of $\det(\mathcal{M}(\omega))$ in the complex frequency plane.

2.2 Optimization

Consider a system composed of N point masses whose positions in the plane are $\mathbf{R}_\alpha = (x_\alpha, y_\alpha)$, $\alpha = 1, \ldots, N$. Each mass m_α can take continuous values in the interval $[m_0, m_N]$, and their positions are restricted to the spatial domain $x_\alpha \in [a, b]$, $y_\alpha \in [c, d]$.

Let $F(\omega, \mathbf{m}, \mathbf{r})$ denote the characteristic function of the system (for instance $F(\omega) = \det[\mathcal{S}(\omega)]$), whose zeros in the complex plane determine the *quasi-normal modes*. The goal is to find the spatial distribution of masses such that the system exhibits a quasi-normal mode at the target complex frequency $\omega_p = \omega_0 + i\gamma_0$. In our case, since we use the Fourier convention $e^{-\imath\omega t}$, we have $\omega_0 > 0$ and $\gamma_0 \leq 0$.

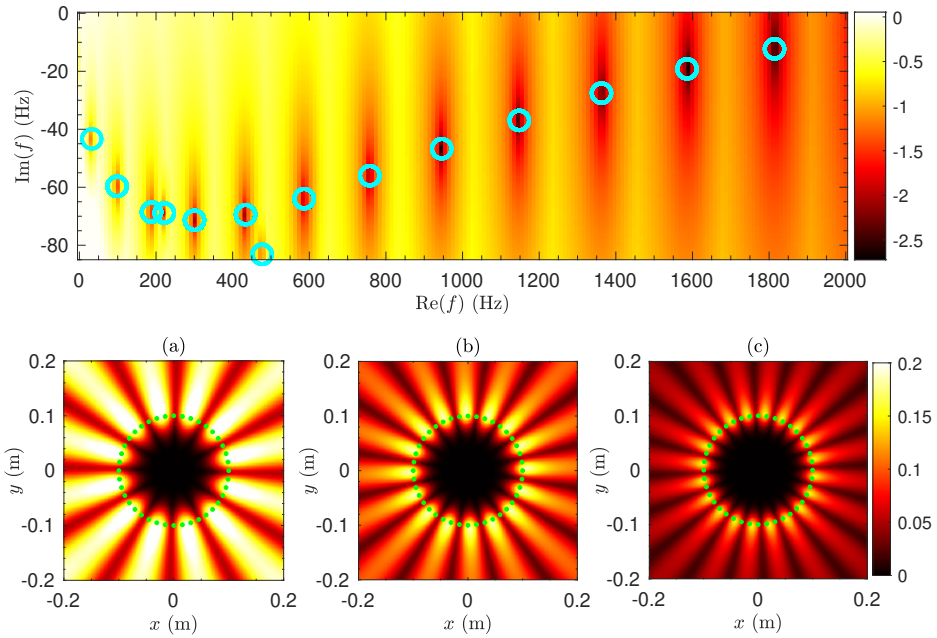


Figure 1: Analysis of a system made of a distribution of 40 point masses of $m = 0.1$ kg, uniformly distributed along a radius $R = 0.1$ m. Upper panel: $\det[\mathcal{M}(\omega)]$, cyan circles represents the zeroth of $\det[\mathcal{M}(\omega)]$ obtained by the Muller algorithm. Lower pannel: Eigenfunction for (a) $\omega_p = 2\pi(586.6 - 63.9\imath)$, (b) $\omega_p = 2\pi(945.6 - 46.7\imath)$, (c) $\omega_p = 2\pi(1362.4 - 27.5\imath)$

The problem can be formulated as the following optimization problem

$$\min_{\{m_i,x_i,y_i\}_{i=1}^N} \; J(\mathbf{m},\mathbf{r}), = |F(\omega_p,\mathbf{m},\mathbf{r})|^2 \tag{6}$$

subject to

$$m_0 \leq m_i \leq m_N, \qquad i = 1,\ldots,N, \tag{7}$$

$$a \leq x_i \leq b, \qquad c \leq y_i \leq d, \qquad i = 1,\ldots,N. \tag{8}$$

We impose a minimum separation constraint between masses

$$\|\mathbf{r}_i - \mathbf{r}_j\|_2 \geq \delta, \qquad \forall\, i \neq j,\ i,j = 1,\ldots,N. \tag{9}$$

The vector of design variables is $\mathbf{p} = (m_1,\ldots,m_N,x_1,\ldots,x_N,y_1,\ldots,y_N)$, and the optimal solution $\mathbf{p}^*$ provides the distribution of masses and their positions in the spatial domain such that the complex frequency ω_p corresponds (approximately) to a zero of the characteristic function of the system.

3 Results

Upper panel in Figure **??** shows the quasi-normal modal spectrum of a system made of a distribution of 40 point masses of $m = 0.1$ kg, uniformly distributed along a radius $R = 0.1$ m. Lower

panel shows the eigenfunction, i.e., the modal shapes at for (a) $\omega_p = 2\pi(586.6 - 63.9\imath)$, (b) $\omega_p = 2\pi(945.6 - 46.7\imath)$, (c) $\omega_p = 2\pi(1362.4 - 27.5\imath)$, representing three zeros of $\det[\mathcal{M}(\omega)]$. The bigger the imaginary part the bigger the outgoing radiation energy, i.e., the leakage.

4 Conclusions

The proposed approach highlights how tailored mass distributions can be used not only to shift eigenfrequencies but to induce and shape modal decay, thereby extending modal design into the complex frequency plane. Establishing this connection enables the use of well-understood concepts from self-adjoint spectral theory as a foundation for addressing non-Hermitian problems, and opens new possibilities for vibration control, wave manipulation, and structural optimization in systems where radiation and dissipation cannot be neglected.

Acknowledgments

V.R.-G and M.M.-S. acknowledge the support from grant PID2023-146237NB-I00 funded by MICIU/AEI/ 10.13039/501100011033.

References

[1] K. F. Graff, *Wave Motion in Elastic Solids*. Courier Corporation, 2012.

[2] P. Martin, *Multiple Scattering. Interaction of Time-Harmonic Waves with N Obstcles*. Cambirdge University Press, UK, 2006.

[3] P. W. Barber and S. C. Hill, *Light Scattering By Particles: Computational Methods*. World Scientific Publishing Co. Pte. Ltd., London, 1990.

[4] B. Vial, M. Martı Sabaté, R. Wiltshaw, S. Guenneau, and R. V. Craster, "Quasinormal mode expansion in thin elastic plates," *Phys. Rev. B*, vol. 110, p. 174 305, 17 Nov. 2024. DOI: 10.1103/PhysRevB.110.174305. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.110.174305.

Mathematical Modelling of Acoustic Levitation Using Perturbative and k-Space Pseudospectral Methods. Demonstrator and Proof of Concept

Diego Carrión Perpiñá[♭], and Vicente Romero-García[♭1]

(♭) I.U. de Matemática Pura y Aplicada (IUMPA), Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

Acoustic levitation is the phenomenon by which objects are suspended in a fluid medium, typically air, through the action of high-intensity sound waves [1], [2]. In standing wave configurations, pressure nodes and antinodes generate stable trapping regions through non linear effects where small particles can be levitated.

Mathematical modeling this phenomenon is nontrivial due to the coupling between wave propagation, nonlinear acoustic effects, and particle dynamics. Analytical solutions are limited to simplified cases, motivating the use of both approximate analytical and numerical approaches.

In this work, we use perturbative methods for analytical insight into acoustic radiation forces and k-space pseudospectral methods for accurate numerical simulation of wave propagation. A demonstrator is designed to validate the theoretical predictions and establish a proof of concept.

2 Methods

2.1 Constitutive equations

Let us consider an ideal, inviscid and compressible fluid. The governing equations[1], [3] are the continuity equation and the Euler equation[2]

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, \tag{1}$$

$$\rho \left(\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} \right) = -\nabla p, \tag{2}$$

where ρ is the fluid density, $\mathbf{v}$ the particle velocity, and p the pressure.

[1] virogar1@mat.upv.es

[2] Navier–Stokes equation neglecting viscosity terms. In the case of linear acoustics, the first-order Euler equation is considered, $\rho \frac{\partial \mathbf{v}}{\partial t} = -\nabla p$, where nonlinear terms are neglected. In our case, we retain nonlinear terms, namely the convective term $(\mathbf{v} \cdot \nabla)\mathbf{v}$.

2.2 Perturbative Approach

To describe an acoustic field, a perturbation expansion around the equilibrium state is introduced:

$$p = p_0 + p_1 + p_2 + \mathcal{O}(3), \tag{3}$$
$$\rho = \rho_0 + \rho_1 + \rho_2 + \mathcal{O}(3), \tag{4}$$
$$\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2 + \mathcal{O}(3), \tag{5}$$

where subscripts indicate the order of the perturbation. The term p_1 describes the linear acoustic field, while p_2 accounts for nonlinear effects responsible for radiation forces. The equilibrium state satisfies

$$\nabla p_0 = 0, \quad \mathbf{v}_0 = 0.$$

This corresponds to a fluid with no acoustic waves, i.e., constant pressure and zero velocity.

First-order equations

Substituting the expansions into (1) and (2), and retaining only first-order terms, yields

$$\frac{\partial \rho_1}{\partial t} + \rho_0 \nabla \cdot \mathbf{v}_1 = 0, \tag{6}$$

$$\rho_0 \frac{\partial \mathbf{v}_1}{\partial t} = -\nabla p_1. \tag{7}$$

These equations describe the linear acoustic field, responsible for harmonic oscillations of pressure and velocity. They lead to the classical wave equation for the first-order acoustic pressure:

$$\nabla^2 p_1 - \frac{1}{c_0^2} \frac{\partial^2 p_1}{\partial t^2} = 0. \tag{8}$$

The field p_1 oscillates harmonically and has zero time average, and therefore cannot exert a net force on an object.

Second-order equations

To analyze nonlinear effects responsible for radiation forces, second-order terms are retained. The Euler equation at second order becomes

$$\rho_0 \frac{\partial \mathbf{v}_2}{\partial t} + \rho_1 \frac{\partial \mathbf{v}_1}{\partial t} + \rho_0 (\mathbf{v}_1 \cdot \nabla)\mathbf{v}_1 = -\nabla p_2. \tag{9}$$

In the time-harmonic regime, after transient effects, any first-order field can be written as $f(\mathbf{r}, t) = f(\mathbf{r})e^{\imath\omega t}$. We define the time average over one period T as

$$\langle f \rangle = \frac{1}{T} \int_0^T f(t)\, dt. \tag{10}$$

First-order terms, being purely oscillatory, have zero time average. Applying the time average to the second-order equation yields

$$\rho_0 \left\langle \frac{\partial \mathbf{v}_2}{\partial t} \right\rangle + \rho_0 \left\langle (\mathbf{v}_1 \cdot \nabla)\mathbf{v}_1 \right\rangle + \left\langle \rho_1 \frac{\partial \mathbf{v}_1}{\partial t} \right\rangle = -\nabla \langle p_2 \rangle. \tag{11}$$

It is important to note that the physically meaningful quantity is the real part of the time average. The time average of the product of two harmonic fields f and g can be written using the real-part rule:

$$\langle fg \rangle = \frac{1}{2}\mathrm{Re}[fg^*].$$

Thus, second-order terms formed by products of first-order fields are generally non-zero. However, the first term vanishes because it is not a product of first-order quantities, and the time average of a periodic time derivative is zero:

$$\left\langle \frac{\partial \mathbf{v}_2}{\partial t} \right\rangle = 0. \tag{12}$$

Rearranging the remaining quadratic terms in $\mathbf{v}_1$ and p_2, we obtain

$$\rho_0 \left\langle (\mathbf{v}_1 \cdot \nabla)\mathbf{v}_1 \right\rangle + \left\langle \rho_1 \frac{\partial \mathbf{v}_1}{\partial t} \right\rangle = -\nabla \langle p_2 \rangle. \tag{13}$$

Using the equation of state $p_1 = c_0^2 \rho_1$ and the first-order Euler equation, we finally obtain

$$\nabla \langle p_2 \rangle = \frac{1}{\rho_0 c_0^2} \left\langle p_1 \nabla p_1 \right\rangle - \rho_0 \left\langle (\mathbf{v}_1 \cdot \nabla)\mathbf{v}_1 \right\rangle. \tag{14}$$

The right-hand side represents a **mean volumetric force** exerted by the acoustic field on the fluid[3]

$$\mathbf{f}_{\mathrm{rad}} = \frac{1}{\rho_0 c_0^2} \left\langle p_1 \nabla p_1 \right\rangle - \rho_0 \left\langle (\mathbf{v}_1 \cdot \nabla)\mathbf{v}_1 \right\rangle = \nabla \left[\frac{1}{2\rho_0 c_0^2} \left\langle p_1^2 \right\rangle - \frac{\rho_0}{2} \left\langle v_1^2 \right\rangle \right], \tag{15}$$

which can be expressed as the gradient of a **radiation pressure**, $\langle p_2 \rangle$,

$$\langle p_2 \rangle = \frac{1}{2\rho_0 c_0^2} \left\langle p_1^2 \right\rangle - \frac{\rho_0}{2} \left\langle v_1^2 \right\rangle. \tag{16}$$

This volumetric force is nonlinear and non-oscillatory, and is responsible for phenomena such as acoustic radiation forces and levitation. Although the acoustic pressure p_1 oscillates around zero, the quadratic terms have a non-zero mean value. Once p_1 and v_1 are known, the volumetric force can be computed. The emergence of this second-order effective pressure generates a net average force capable of producing stable equilibrium points, which is an essential condition for acoustic levitation.

In the previous calculation, the presence of a particle has not been considered, and therefore the acoustic field is not influenced by any scatterer. If a particle is introduced, a scattered field is generated that interferes with the incident field. To compute the radiation force, this effect must be taken into account. Let us consider a particle in an incident acoustic field characterized by p_1^{in} and $\mathbf{v}_1^{in}$, which gives rise to scattered waves characterized by p_1^{sc} and $\mathbf{v}_1^{sc}$. Using the total pressure and velocity fields, $p_1 = p_1^{in} + p_1^{sc}$ and $\mathbf{v}_1 = \mathbf{v}_1^{in} + \mathbf{v}_1^{sc}$, we compute the time-averaged acoustic radiation pressure using Eq. (16), and then use

$$\mathbf{F}_{rad} = -\iint_{S_0} \langle p_2 \rangle \mathbf{n}\, dS, \tag{17}$$

to determine the acoustic radiation force acting on the particle, whose surface is S_0. If it is not yet clear, carrying out this procedure is far from trivial. Applying the divergence theorem to Eq. (17), we obtain

$$\mathbf{F}_{rad} = -\iint_{S_0} \langle p_2 \rangle \mathbf{n}\, dS = -\iiint_{V} \nabla \langle p_2 \rangle\, dV.$$

[3]In the last term of Eq. (15) we have used the vector identities: $\nabla(p_1^2) = 2p_1\nabla p_1$ and $\nabla(\mathbf{v}_1 \cdot \mathbf{v}_1) = \nabla(v_1^2) = 2(\mathbf{v}_1 \cdot \nabla)\mathbf{v}_1$.

If the particle is very small, $R \ll \lambda$, two assumptions can be made: first, that $\langle p_2 \rangle$ is approximately constant within the volume of the sphere, and second, that the field scattered by the sphere is sufficiently small compared to the incident field and can be neglected. Under these assumptions, we obtain

$$\mathbf{F}_{rad} = -\frac{4}{3}\pi R^3 \nabla \langle p_2 \rangle = -\nabla U, \tag{18}$$

or equivalently, using Eq. (16) with only the incident field,

$$U = \frac{4}{3}\pi R^3 \left[\frac{1}{2\rho_0 c_0^2} \left\langle p_1^2 \right\rangle - \frac{\rho_0}{2} \left\langle v_1^2 \right\rangle \right]. \tag{19}$$

In the Rayleigh regime, when the particle radius R is much smaller than the wavelength ($R \ll \lambda$), the radiation force can be expressed as the gradient of a scalar potential U,

$$\mathbf{F}_{rad} = -\nabla U. \tag{20}$$

The Gor'kov potential, which accounts for both the incident field and the field scattered by the particle, is given by

$$U = \frac{4\pi R^3}{3} \left[\frac{1}{2} f_1 \frac{\langle p_1^2 \rangle}{\rho_0 c_0^2} - \frac{3}{4} f_2 \rho_0 \langle v_1^2 \rangle \right], \tag{21}$$

where f_1 and f_2 depend on the relative mechanical properties of the particle and the fluid,

$$f_1 = 1 - \frac{\rho_0 c_0^2}{2\rho_p c_p^2}, \tag{22}$$

$$f_2 = 2\left(\frac{\rho_p - \rho_0}{2\rho_p + \rho_0} \right). \tag{23}$$

The resulting force is therefore proportional to the spatial gradient of the acoustic energy. Note that in most cases the particle is made of a solid, and typically $\rho_p \gg \rho_0$ and $c_p \gg c_0$, so that we can approximate $f_1 \simeq f_2 \simeq 1$. The difference between Eq. (19) and Eq. (21) lies in the fact that, in the latter, the scattered field produced by the particle has been taken into account.

2.3 k-Space Pseudospectral Method

To approach to realistic geometries, we numerically solve the first-order acoustic equations Eqs. (1,2) together with the following equation of state

$$p = c_0^2 \left(\rho + \frac{B}{2A} \frac{\rho^2}{\rho_0} \right), \tag{24}$$

where B/A is the parameter governing acoustic nonlinearity[1]. To do that we use the k-Wave toolbox open-source from MATLAB[4]–[6] performing time-domain simulations of acoustic wave fields in homogeneous and heterogeneous media. It is particularly well suited for modeling wave propagation in fluids and soft tissues, including nonlinear effects, absorption, and dispersion. The numerical core of k-wave is based on a k-space pseudo-spectral method, which combines Fourier collocation for spatial derivatives with finite-difference time integration. This approach allows for high accuracy with relatively coarse spatial grids compared to conventional finite-difference schemes.

To account for acoustic absorption and dispersion, k-Wave incorporates a frequency power-law model using fractional Laplacian operators, enabling accurate representation of realistic media. Spatial gradients are computed in the Fourier domain, which reduces numerical dispersion, while

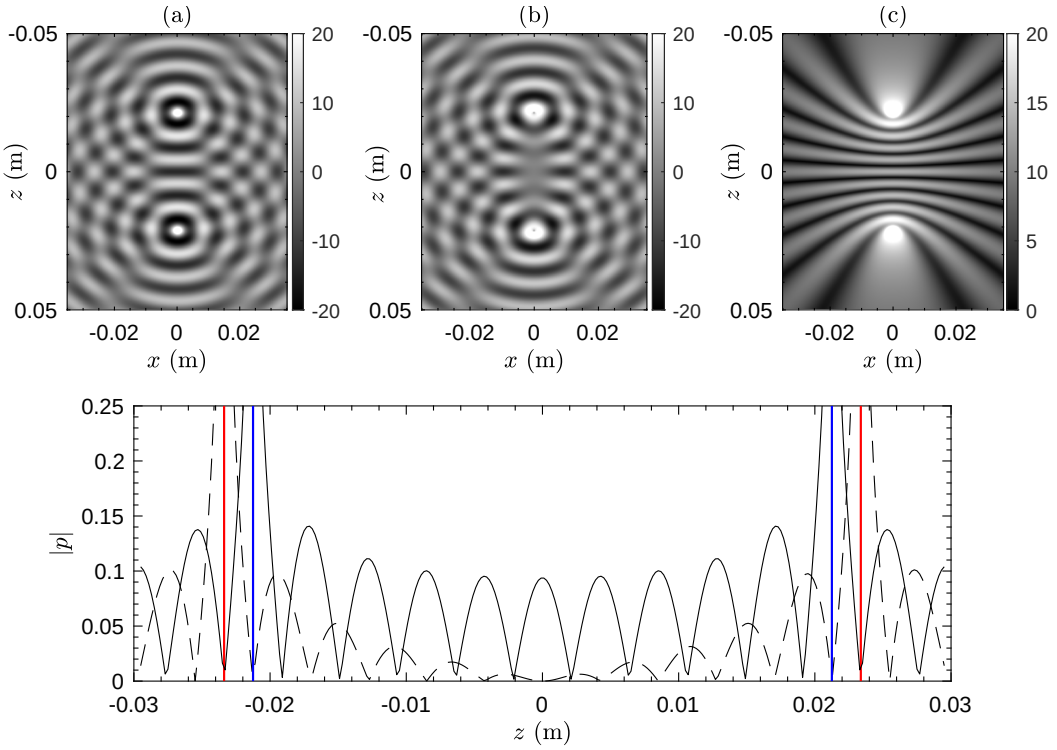


Figure 1: Upper panel: (a), (b) and (c) shows the real, imaginary and absolute value of the acoustic field $p(\mathbf{r})$ respectively produced by the interference between two point sources. Lower panel: Cut of the absolute value of the $p(\mathbf{r})$ along z-axis for two different configurations where the sources are placed at positions determined by the red and blue lines.

a k-space correction is applied to the time-stepping scheme to improve stability and accuracy for larger time steps.

Overall, k-Wave provides an efficient and flexible framework for simulating linear and nonlinear acoustic wave propagation, making it widely used in applications such as ultrasound imaging, photoacoustics, and wave-structure interaction problems.

3 Results

3.1 Perturbative Predictions

The perturbative model provides a clear physical description of acoustic levitation. It predicts that particles can be trapped at pressure nodes or antinodes depending on their material properties, with the radiation force being proportional to the acoustic intensity. Furthermore, the equilibrium positions correspond to the extrema of the acoustic potential, satisfying the condition $\nabla U = 0$. This framework highlights the dependence of the levitation behavior on both the acoustic field and the intrinsic properties of the particle, offering an intuitive understanding of the mechanisms governing trapping and stability.

Numerical simulations based on the k-space pseudospectral method complement this analytical approach by accurately resolving the acoustic field (see Fig. 1). The simulations reveal the formation of standing wave patterns with node spacing equal to $\lambda/2$, as expected theoretically, and show consistent pressure amplitudes throughout the domain. Importantly, the computed fields allow

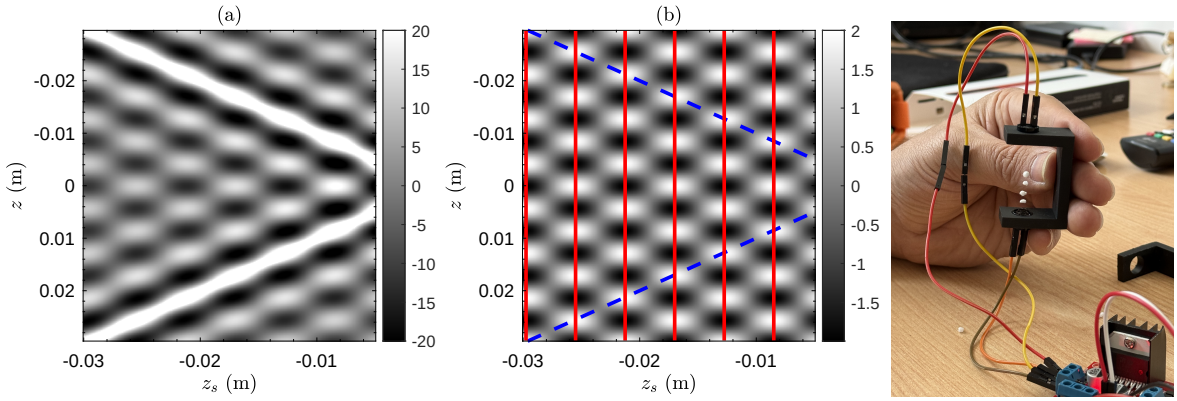


Figure 2: Interference pattern analysis at $x = 0$ as a function of the distance between the point sources, z_s. The absolute value of the pressure is shown for 40 kHz. Panel (a) depicts the interference pattern $\tilde{p}(0, z)$. Panel (b) shows the analysis of $\tilde{p}(0, z)$ after removing singularities for different values of z_s. For guidance, blue dashed lines indicate the phase origins of the plane waves, i.e., the positions of the sources. Picture represents the preliminary demonstrator showing the acoustic levitation of 4 particles.

for the direct evaluation of radiation forces through time averaging, providing a quantitative link between the acoustic field and the resulting forces.

Experimental validation using a demonstrator confirms these predictions (see Fig. 2), with particles observed to be trapped at the expected node locations. The stability of the trapping is found to depend on factors such as alignment and excitation amplitude, while small discrepancies between theory and experiment can be attributed to air turbulence, non-ideal transducer behavior, and boundary imperfections.

4 Conclusions

In conclusion, perturbative analysis and k-space pseudospectral simulations provides robust frameworks for the mathematical modeling of acoustic levitation. The perturbative approach offers valuable physical insight into the origin of radiation forces, while the numerical method delivers accurate predictions of the acoustic field and resulting particle dynamics. Experimental results show good agreement with theoretical expectations, validating the overall approach. Future work may focus on extending the model to include higher-order nonlinear effects, investigating multi-particle interactions, and developing adaptive strategies for real-time control of acoustic fields.

Acknowledgments

V.R.-G. acknowledge the support from grant PID2023-146237NB-I00 funded by MICIU/AEI/ 10.13039/501100011033.

References

[1] M. Hamilton and D. Blackstock, *Nonlinear Acoustics*. Academic Press, 1998, ISBN: 9780123218605. [Online]. Available: `https://books.google.es/books?id=k-kWozCE0jIC`.

[2] D. P. Jackson and M.-H. Chang, "Acoustic levitation and the acoustic radiation force," *American Journal of Physics*, vol. 89, no. 4, pp. 383–392, Apr. 2021, ISSN: 0002-9505. DOI: 10.1119/10.0002764. eprint: https://pubs.aip.org/aapt/ajp/article-pdf/89/4/383/20097030/383_1_10.0002764.pdf. [Online]. Available: https://doi.org/10.1119/10.0002764.

[3] L. Kinsler, *Fundamentals of Acoustics*. Wiley, 1982, ISBN: 9780471029335. [Online]. Available: https://books.google.es/books?id=ibTvAAAAMAAJ.

[4] B. E. Treeby and B. T. Cox, "k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields," *Journal of Biomedical Optics*, vol. 15, no. 2, p. 021 314, 2010. DOI: 10.1117/1.3360308. [Online]. Available: https://doi.org/10.1117/1.3360308.

[5] B. E. Treeby, J. Jaros, A. P. Rendell, and B. T. Cox, "Modeling nonlinear ultrasound propagation in heterogeneous media with power law absorption using a k-space pseudospectral method," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4324–4336, Jun. 2012, ISSN: 0001-4966. DOI: 10.1121/1.4712021. eprint: https://pubs.aip.org/asa/jasa/article-pdf/131/6/4324/15298882/4324_1_online.pdf. [Online]. Available: https://doi.org/10.1121/1.4712021.

[6] B. E. Treeby and B. T. Cox, "Modeling power law absorption and dispersion for acoustic propagation using the fractional laplacian," *The Journal of the Acoustical Society of America*, vol. 127, no. 5, pp. 2741–2748, May 2010, ISSN: 0001-4966. DOI: 10.1121/1.3377056. eprint: https://pubs.aip.org/asa/jasa/article-pdf/127/5/2741/14776711/2741_1_online.pdf. [Online]. Available: https://doi.org/10.1121/1.3377056.

Multi-objective optimization of metamaterial sandwich panels

Vinícius F. Dal Poggetto[a],[1] Mário Lázaro[a],
Luis M. García-Raffi[a], and Vicent Romero-García[a],[2]

(a) Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

The scientific and technological development of composite materials has significantly expanded the range of engineering applications across multiple domains [1–3], particularly in aerospace [4–6] and automotive industries [7–9]. From a mathematical modeling perspective, composites present a rich framework in which heterogeneous material properties can be homogenized, parameterized, and tailored to meet specific performance objectives [10, 11]. By combining constituents with contrasting physical properties, composite materials enable the formulation of design spaces where structural responses can be systematically tuned, often leading to superior performance compared to conventional materials. In particular, their high strength-to-weight ratio makes them especially attractive for transportation systems, where weight reduction directly impacts fuel efficiency and system-level optimization.

Among the various configurations of composite structures, sandwich panels represent a canonical example for the application of mathematical modeling and structural optimization [12, 13]. These structures consist of two thin, stiff outer faces separated by a lightweight core, leading to a hierarchical mechanical system that can be described through multi-scale models. The faces primarily carry in-plane and bending loads, while the core contributes to shear stiffness, stability, and buckling resistance [14, 15]. This separation of roles enables the derivation of reduced-order models that capture the global structural response while retaining key physical parameters, making sandwich panels particularly suitable for optimization-based design. As a result, they offer excellent mechanical efficiency and multifunctional performance, combining stiffness, strength, and energy absorption [16].

Despite these advantages, sandwich panels introduce several challenges that complicate their mathematical modeling and optimization. Phenomena such as buckling [17], interfacial debonding [18], and failure at joints [19] require the incorporation of nonlinearities, discontinuities, and multi-physics coupling into the governing equations. Additionally, environmental effects such as moisture absorption can degrade material properties over time [20], introducing uncertainty and variability into the model parameters. These aspects significantly increase the complexity of the design space and motivate the development of robust and computationally efficient optimization strategies. In this context, simplified or monolithic structural configurations are often explored to reduce model complexity and manufacturing constraints while preserving performance.

[1] vfondal@upv.es
[2] virogar1@mat.upv.es

Modern engineering applications impose stringent and often conflicting requirements on structural components. Beyond static performance metrics such as stiffness and strength, structures must also satisfy dynamic constraints, including vibration attenuation, fatigue resistance, and frequency-dependent behavior. From a modeling standpoint, this necessitates the integration of both static and dynamic analyses, often leading to coupled systems of partial differential equations or frequency-domain formulations. Consequently, the design of composite structures—and sandwich panels in particular—naturally leads to multi-objective optimization problems, where competing criteria must be balanced within a constrained design space.

In recent years, mechanical metamaterials have emerged as a powerful paradigm for expanding the achievable design space [21–23]. These materials are characterized by engineered periodic microstructures, whose effective properties can be derived through homogenization techniques and dispersion analysis [24, 25]. From a mathematical perspective, they enable the design of structures with tailored wave propagation characteristics, such as band gaps and negative effective parameters. Examples include auxetic behavior [26] and selective attenuation of vibrations within prescribed frequency ranges [27]. Incorporating such metamaterial concepts into sandwich panel design introduces additional degrees of freedom, which can be exploited through optimization to achieve multifunctional performance.

Although numerous studies have focused on optimizing sandwich panels for either static or dynamic performance, relatively few have addressed the simultaneous integration of both aspects within a unified mathematical framework. This limitation arises from the inherently conflicting nature of the objectives, where improvements in stiffness or strength may adversely affect dynamic behavior, and vice versa. Therefore, the design of multifunctional sandwich panels requires advanced optimization methodologies capable of handling multi-objective, constrained, and potentially nonlinear problems. Such approaches must effectively couple high-fidelity mathematical models with efficient numerical optimization techniques to explore the design space and identify optimal or Pareto-efficient solutions.

In this work, we propose a topology-based optimization approach to determine the optimal material distribution in corrugated profiles of metamaterial-based sandwich panels. The sandwich cores are generated to ensure simple connectivity and smooth profiles, reducing stress concentrators and ensuring connectivity between the opposing faces. The performance of these panels is optimized considering (i) bending stiffness and (ii) flexural wave attenuation capabilities. We propose specifically designed objective functions in combination with genetic algorithms to enable the exploitation of a wide space of solutions. We obtain distinct structures by considering either isolated objective functions or their combination, demonstrating the versatility of the method.

2 Methods

We begin the design of our sandwich panel by considering a core region discretized as $n_q = N_x \times N_y$ pixels, whose description is given by a binary vector $\mathbf{q} \in \{0,1\}^{n_q}$ (Fig. 1**a**, left panel: material in purple, voids in grey). The resulting distribution of pixels is then treated using a Gaussian filter on subdivided pixels, typically yielding regions with smooth curvatures to avoid stress concentrators (Fig. 1**a**, center panel). The resulting geometry is then reflected and assigned to a structure of length a and internal thickness h_i. Two layers of homogeneous material with thickness h_e are added to the top and bottom faces of this internal structure (Fig. 1**a**, right panel), yielding a total thickness $h = h_i + 2h_e$. This unit cell is then infinitely repeated along the length direction (x), obtaining an infinitely periodic sandwich panel (Fig. 1**b**).

To apply the optimization procedure, we consider two distinct objectives for the optimization problem: (i) minimizing the compliance of the rectangular super cell with respect to pure bending loads (static behavior) and (ii) maximizing the normalized band gap width (dynamic behavior).

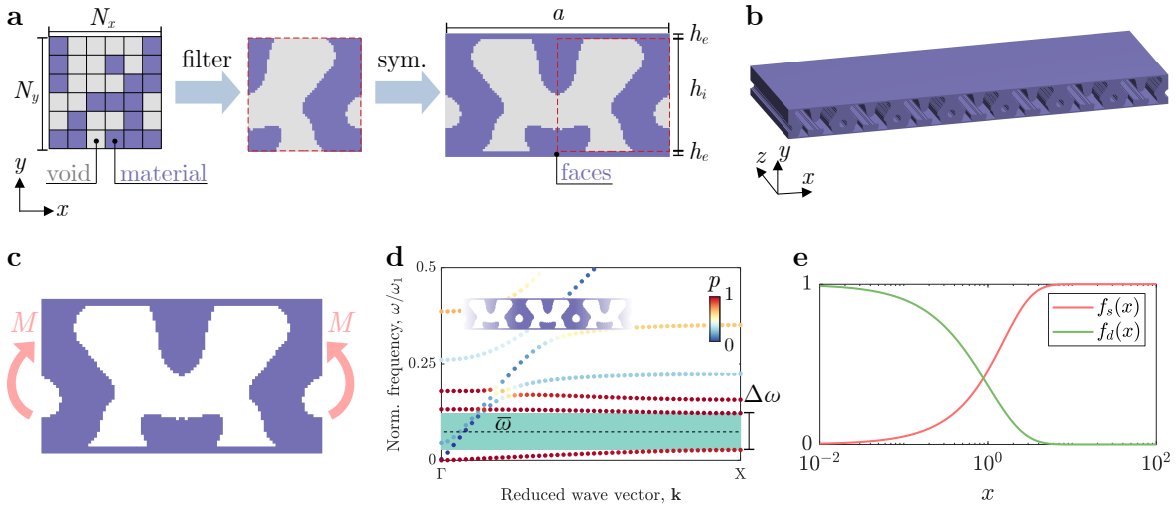


Figure 1: **Sandwich panel design method**. **a**. Design region (left) which is smoothened (center) and used to obtain a periodic unit cell (right). **b**. Resulting sandwich panel. **c**. Unit cell under pure bending. **d**. Dispersion relation showing a flexural band gap. **e**. Normalized objective functions.

For the first objective (static), we consider the objective function to be minimized as

$$f_s(\mathbf{q}) = \frac{1 - \exp(-\Delta U(\mathbf{q}))}{1 + \exp(-\Delta U(\mathbf{q}))}, \tag{1}$$

where $\Delta U(\mathbf{q}) = \frac{U(\mathbf{q}) - U_0}{U_0}$ is the normalized strain energy of the unit cell under unit bending loads (Fig. 1**c**) and U_0 is the reference strain energy relative to a full unit cell ($\mathbf{q} = \{1\}^{n_q}$). The total strain energy is defined as $U(\mathbf{q}) = \int_\Omega \frac{1}{2}\boldsymbol{\epsilon}^T\boldsymbol{\sigma}\, d\Omega$, where Ω is the domain of the unit cell, and $\boldsymbol{\sigma}$ and $\boldsymbol{\epsilon}$ are, respectively, the states of stresses and strains, represented in vector form. As we consider U_0 to be the best solution (minimum strain energy), this yields $f_s(\mathbf{q}) \in [0, 1]$. Considering a finite element implementation, this expression is numerically equivalent to the total work done by external forces on the unit cell, expressed as

$$U(\mathbf{q}) = \frac{1}{2}\mathbf{u}^T\mathbf{K}(\mathbf{q})\mathbf{u}, \tag{2}$$

where $\mathbf{K}$ is the stiffness matrix of the unit cell and $\mathbf{u}$ is the nodal vector of displacements obtained from a static bending test. According to the objective function $f_s(\mathbf{q}) = 0$ corresponds to a structure with equal strain energy as a unit cell filled with material and $f_s(\mathbf{q}) = 1$ corresponds to an infinitely compliant structure.

The second considered objective function (dynamic) to be minimized is

$$f_d(\mathbf{q}) = \exp\left(-\frac{\Delta\omega(\mathbf{q})}{\overline{\omega}(\mathbf{q})}\right), \tag{3}$$

where $\Delta\omega(\mathbf{q})$ is the band gap width and $\overline{\omega}(\mathbf{q})$ is the band gap central frequency for a given pixel distribution $\mathbf{q}$ (Fig. 1**d**). For the case of no band gaps present in the dispersion relation ($\Delta\omega(\mathbf{q}) = 0$), we obtain $f_d(\mathbf{q}) = 1$; otherwise, for band gaps with a wide normalized width ($\Delta\omega(\mathbf{q})/\overline{\omega}(\mathbf{q}) \gg 1$), this yields $f_d(\mathbf{q}) \to 0$. It is important to notice that the usual objective of maximizing the band gap width is here formulated as a minimization problem to be consistent with the other optimization objective, ensuring $f_d \in [0, 1]$. The relation between these objective functions is shown in Fig. 1**e**, where $x = -\Delta U(\mathbf{q})$ for f_s and $x = -\Delta\omega(\mathbf{q})/\overline{\omega}(\mathbf{q})$ for f_d.

3 Results

We then implement the optimization problem in MATLAB using an internally developed code for finite element computations. Material properties are considered as those of a plastic material, with Young's modulus $E = 3.4$ GPa, Poisson's coefficient $\nu = 0.3$, and specific mass density $\rho = 1400$ kg/m^3. We consider an internal thickness of 20 mm, with outer layers of 1 mm each. The total thickness of the unit cell is then given as $h = 22$ mm, and the unit cell length is considered as $a = 2h_i = 40$ mm. These dimensions were chosen to allow the 3D printing of samples for experimental validation, as the first Bragg flexural band gap opens at $\omega_1/2\pi = 9.7$ kHz.

We then use a genetic algorithm (GA toolbox in Matlab) to perform the optimization, considering different combinations of the objective functions given by Eqs. (1) and (3). For the optimization algorithm, we consider a 32×32 design space. In order to obtain only lightweight designs, we restrict the filling fraction (total area occupied by material in the unit cell) to 50% of the material.

We begin by testing our optimization procedure considering the sole objective functions concerning the static (f_s) and dynamic (f_d), behaviors. This initial run allows us to validate the performance of the method, as the expected solutions are commonly known. The optimized solution considering only the static objective function is shown in Fig. 2**a** (case I, leftmost unit cell), resembling a truss-like structure, well-known for its bending load performance. Its corresponding normalized bending stiffness, computed as U_0/U_I, is shown in Fig. 2**b** (left panel), indicating a decrease of nearly 20% with respect to the reference case (in grey). We note that the decrease in bending stiffness is largely compensated by the reduction in mass, resulting in a higher value of bending stiffness per unit mass. Its corresponding band diagram is represented in Fig. 2**c** (I), with longitudinal waves in blue ($p = 0$) and flexural waves in red ($p = 1$), exhibiting only a residual band gap owed to band folding at $k = \pi/a$.

Next, we analyze the optimized solution considering only the dynamic objective function. The resulting unit cell is shown in Fig. 2**a** (case V, rightmost unit cell). In this case, the majority of mass is dislocated to the center of the unit cell, while the thickness of the outer layers is reduced to a minimum. Its corresponding normalized bending stiffness, computed as U_0/U_V, is shown in Fig. 2**b** (left panel), indicating a decrease of nearly 70% with respect to the reference case (in grey). This result was also expected, as in this case, the static performance of the unit cell is not considered in the optimization objective. Finally, its corresponding band diagram is represented in Fig. 2**c** (V), exhibiting a normalized band gap width of nearly 160% (Fig. 2**b**, right panel), an extremely wide value in the sub-Bragg regime. This is mostly owed to the largest impedance modulation given by the central region with high localized stiffness.

Next, we perform the same optimization considering a linear combination between both objective functions, namely: cases (II) $0.75f_s + 0.25f_d$, (III) $0.5f_s + 0.5f_d$, and (IV) $0.25f_s + 0.75f_d$. The corresponding optimized unit cells for each case are shown in Fig. 2**a**, showing a dislocation of material from the initial configuration at an angle and evenly distributed (case I), to practically vertical at the center (case V). Interestingly, the material distribution obtained in case III presents regions of concentrated material at the edges (which are equivalent to concentrations at the center due to the unit cell's periodicity), however, with thicker outer layers.

The corresponding normalized bending stiffness (with corresponding objective function f_s) and normalized band gap width (with corresponding objective function f_d) are shown in Fig. 2**b** on the left and right panels, respectively. We note that, from left to right, each case objective function demonstrates a monotonic behavior, with the best static objective function evaluation at case I and the best dynamic objective function evaluation at case V. This demonstrates that the proposed objective functions are well-suited to capture subtle differences between distinct designs, and their combination can be chosen to tune design parameters according to the requirements of specific applications.

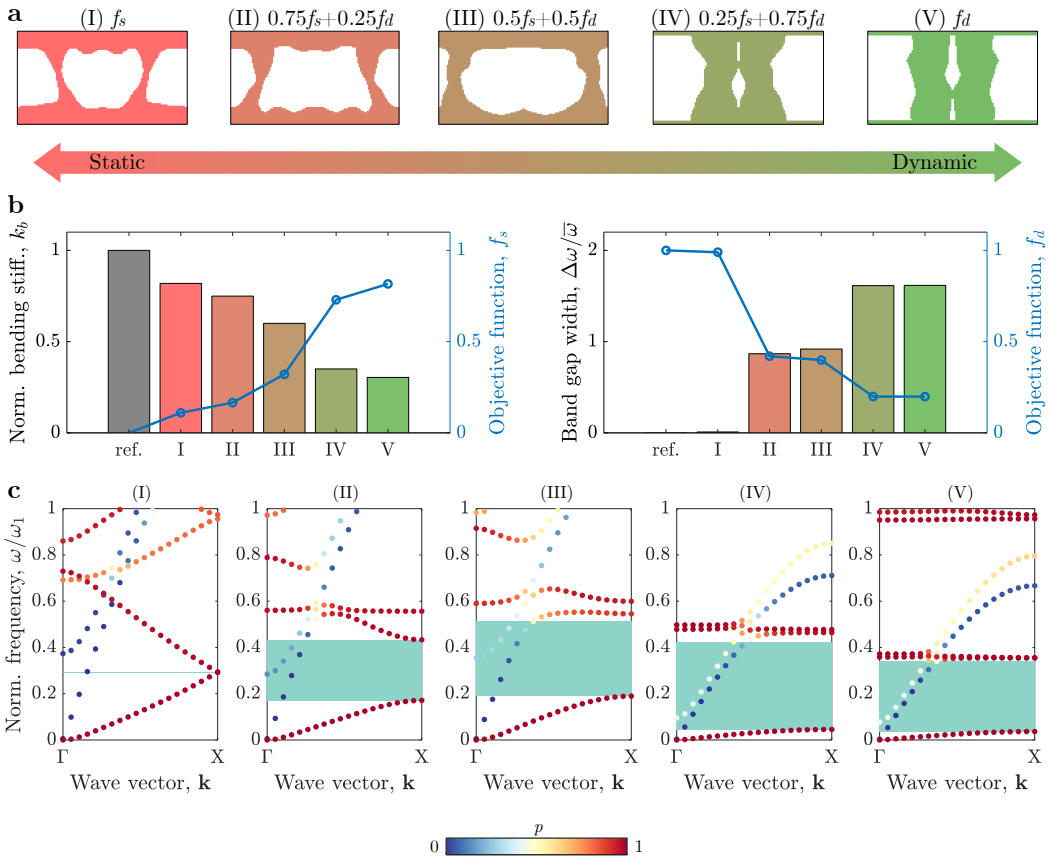


Figure 2: **Optimization results for different combinations of objective functions**. **a**. Optimized unit cells for different combinations of static (f_s) and dynamic (f_d) objective functions. **b**. Normalized bending with respect to reference case (left) and normalized band gap width (right). **c**. Dispersion relations for the distinct unit cells.

4 Conclusions

In conclusion, in this work we demonstrate the effectiveness of a topology-based multi-objective optimization framework for the design of metamaterial-based sandwich panels to simultaneously address (i) static and (ii) dynamic performance requirements. By combining tailored objective functions for (i) bending stiffness and (ii) flexural wave attenuation with genetic algorithms, the proposed methodology can generate a wide range of optimized core geometries. Our results highlight the inherent trade-offs between bending stiffness and vibration attenuation, while also showing that balanced, multifunctional designs can be achieved through an appropriate combination of objectives. The approach provides a flexible and robust tool to design sandwich structures, enabling engineers to tailor material distributions to meet specific application demands in fields such as aerospace and automotive engineering.

Acknowledgments

VFDP and VRG are funded by the European Union's Horizon programme in the framework of the MSCA "META-SONIC" project under grant agreement No. 101202648.

References

[1] W. J. Cantwell and J. Morton. "The impact resistance of composite materials — a review". In: *Composites* 22.5 (1991), pp. 347–362.

[2] S. Ramakrishna et al. "Biomedical applications of polymer-composite materials: a review". In: *Composites science and technology* 61.9 (2001), pp. 1189–1224.

[3] J. Chróścielewski et al. "A novel sandwich footbridge - Practical application of laminated composites in bridge design and in situ measurements of static response". In: *Composites Part B: Engineering* 126 (2017), pp. 153–161.

[4] P. D. Mangalgiri. "Composite materials for aerospace applications". In: *Bulletin of Materials Science* 22.3 (1999), pp. 657–664.

[5] S. Rana and R. Fangueiro. *Advanced composite materials for aerospace engineering: processing, properties and applications*. Woodhead Publishing, 2016.

[6] P. E. Irving and C. Soutis. *Polymer composites in the aerospace industry*. Woodhead Publishing, 2019.

[7] Y. Kurihara. "Polymer matrix composite materials in automobile industries". In: *Advanced Composite Materials* 4.3 (1995), pp. 209–219.

[8] A. Wazeer et al. "Composites for electric vehicles and automotive sector: A review". In: *Green Energy and Intelligent Transportation* 2.1 (2023), p. 100043.

[9] F. Khan et al. "Advances of composite materials in automobile applications – A review". In: *Journal of Engineering Research* (2024).

[10] S. W. Tsai. *Introduction to composite materials*. Routledge, 2018.

[11] T. W. Clyne and D. Hull. *An introduction to composite materials*. Cambridge University Press, 2019.

[12] J. M. Davies. "Sandwich panels". In: *Thin-walled Structures* 16.1-4 (1993), pp. 179–198.

[13] N. J. Hoff and S. E. Mautner. "Sandwich construction". In: *Aeronautical Engineering Review,* 3.8 (1998), pp. 29–43.

[14] D. Zenkert. *The handbook of sandwich construction*. Engineering Materials Advisory Services, 1997.

[15] J. R. Vinson. "Sandwich structures". In: *Applied Mechanics Reviews* 54.3 (2001), pp. 201–214.

[16] F. Zhu et al. "Plastic deformation, failure and energy absorption of sandwich structures with metallic cellular cores". In: *International Journal of Protective Structures* 1.4 (2010), pp. 507–541.

[17] W. Miller, C. W. Smith, and K. E. Evans. "Honeycomb cores with enhanced buckling strength". In: *Composite Structures* 93.3 (2011), pp. 1072–1077.

[18] I. Odessa, Y. Frostig, and O. Rabinovitch. "Modeling of interfacial debonding propagation in sandwich panels". In: *International Journal of Solids and Structures* 148 (2018), pp. 67–78.

[19] S. Heimbs and M. Pein. "Failure behaviour of honeycomb sandwich corner joints and inserts". In: *Composite Structures* 89.4 (2009), pp. 575–588.

[20] E. Kececi and R. Asmatulu. "Effects of moisture ingressions on mechanical properties of honeycomb-structured fiber composites for aerospace applications". In: *The International Journal of Advanced Manufacturing Technology* 88.1 (2017), pp. 459–470.

[21] Z. Liu et al. "Locally resonant sonic materials". In: *Science* 289.5485 (2000), pp. 1734–1736.

[22] P. A. Deymier. *Acoustic metamaterials and phononic crystals*. Vol. 173. Springer Science & Business Media, 2013.

[23] A. Khelif and A. Adibi. "Phononic crystals". In: *Berlin, Germany, Springer* 10 (2015), pp. 978–1.

[24] R. V. Craster et al. "Mechanical metamaterials". In: *Reports on Progress in Physics* (2023).

[25] M. Kadic et al. "3D Metamaterials". In: *Nature reviews physics* 1.3 (2019), pp. 198–210.

[26] M. Morvaridi et al. "Tunable topologically protected waveguiding in auxetic nonlinear metamaterials". In: *Physical Review Applied* 21.3 (2024), p. 034024.

[27] V. F. Dal Poggetto et al. "Band gap enhancement in periodic frames using hierarchical structures". In: *International Journal of Solids and Structures* 216 (2021), pp. 68–82.

Small-Sample Prediction and Inverse Design of Three-dimensional Topological Phononic Crystals

Xue-Qian Zhang[♭,♮,1] Vicent Romero-García[♭] Tian-Xue Ma[♮] and Yue-Sheng Wang[♮]

(♭) Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

(♮) Department of Mechanics, School of Physical Science and Engineering, Beijing Jiaotong University
Beijing, 100044, China.

1 Introduction

Phononic crystals (PnCs) are a class of artificial functional materials formed by periodically arranging one material within another medium. They possess a unique ability to control the propagation of elastic or acoustic waves. One of their key features is the existence of band gaps (BGs), within which wave propagation is prohibited. With the incorporation of topological concepts, two PnCs sharing the same BG range but exhibiting different topological properties can lead to the emergence of topological interface states (TISs) at their interface. These interface states are robust against defects and free from backscattering, offering great potential for wave manipulation. However, the design of PnCs with target BG typically relies on empirical approaches and extensive numerical simulations, resulting in high computational costs. This challenge becomes even more pronounced for three-dimensional (3D) PnCs, where the computational burden increases substantially. With the rapid development of artificial intelligence, deep learning (DL) has emerged as a promising approach to address these challenges. As a subfield of machine learning, DL has achieved remarkable success in areas such as computer vision [1] and natural language processing [2]. By learning intrinsic patterns from data, DL models can uncover complex relationships between input and output variables. Recent studies have demonstrated that DL can effectively map structural parameters to wave-related properties, including optical responses [3], [4], topological characteristics [5], and other physical behaviors. In this work, we establish a mapping between the geometric configurations of 3D PnCs and their elastic wave BGs using a DL framework. To address the high computational cost associated with 3D structures, a small-sample training strategy is adopted. The proposed model successfully achieves both the prediction and inverse design of low-order odd BGs, and further enables the realization of TISs.

[1] xzhan19x@upvnet.upv.es

2 Methods

2.1 Geometry and band structure

In this work, we focus on 3D PnCs with geometric configurations satisfying C_{4v} symmetry, and investigate their flexural wave propagation characteristics. The unit cell with lattice constant a = 50 mm is described by a 32 × 32 binary image, where the blue regions in Fig. 1(a) represent the protruded parts of the structure. The structure is fabricated using photosensitive resin, with geometric parameters set as $h_1 = 0.05a$ and $h_2 = 0.8a$. Based on the above structural model, the flexural wave band structure of the unit cell is calculated using the finite element method (FEM) implemented in COMSOL Multiphysics, as shown in Fig. 1(b). The blue shaded region indicates the seventh-order BG, within which wave propagation is prohibited. Unit cell 2 is obtained by translating unit cell 1 by half a lattice constant along both the horizontal and vertical directions, as illustrated in Fig. 1(a). Although these two unit cells share identical band structures, their topological properties of the odd-order BGs are different. As a result, when PnCs composed of these two unit cells are combined, localized TISs emerge at the interface, as shown in Fig. 1(c). At the corresponding frequency of 2283.2 Hz, the vibration energy is clearly localized at the interface region, confirming the existence of the TIS.

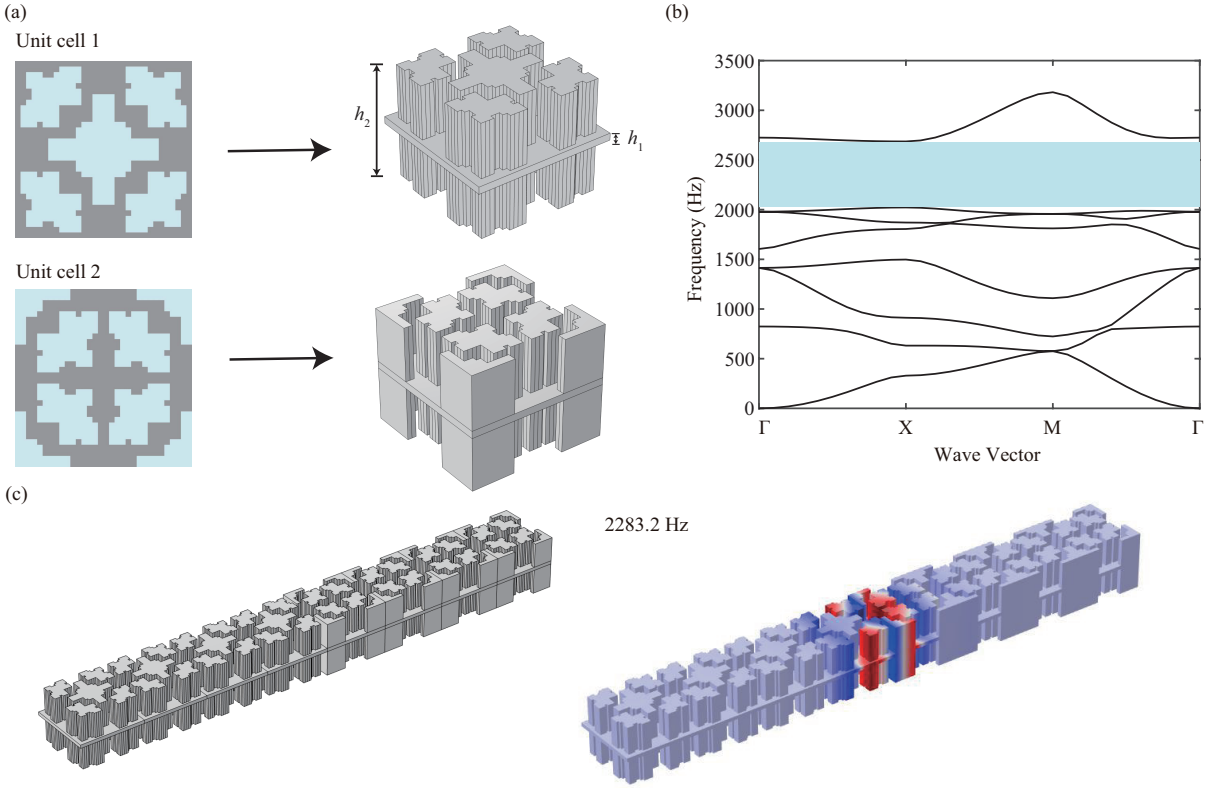


Figure 1: (a) Two unit cells with C_{4v} symmetry and their corresponding 3D geometry. (b) Flexural wave band structure, where the shaded region indicates the seventh-order BG. (c) TIS formed by combining the two unit cells, showing energy localization at the interface (2283.2 Hz).

2.2 Deep learning models

The variational autoencoder (VAE) is a generative artificial intelligence model that consists of two main components: an encoder and a decoder. The encoder extracts features from high-dimensional data and maps them into a low-dimensional latent space, while the decoder reconstructs the data from the latent space back to the original input space. In this work, the VAE is employed to encode unit cells into a low-dimensional latent vector of size 1×5, as shown in Fig. 2(a).

A multilayer perceptron (MLP) is a type of artificial neural network consisting of an input layer, multiple hidden layers, and an output layer. Each layer contains multiple neurons that perform a weighted summation of the inputs from the previous layer, followed by a nonlinear transformation through an activation function. This enables the MLP to approximate complex nonlinear functions. In this work, MLPs are employed to construct both the forward network and the generative network, as shown in Fig. 2(b). The forward network is used to predict the BGs of PnCs, while the generative network is designed to generate structures with target BGs.

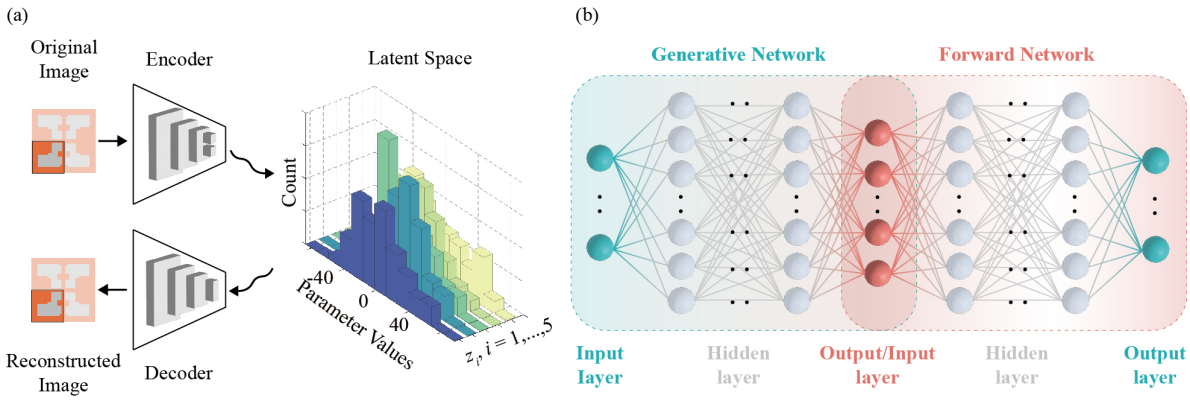


Figure 2: Schematic diagrams of the VAE (a) and the inverse design framework (b). The inverse design framework is comprised of a generative network and a trained forward network (i.e., the MLP).

3 Results

For dimensionality reduction, the VAE achieves a reconstruction accuracy of approximately 97%, enabling efficient prediction and design directly in the latent space. For the forward model, high prediction accuracy is obtained using only 973 training samples, demonstrating the effectiveness of the proposed approach under small-sample conditions. The predicted low-order BGs are compared with the ground truth in Fig. 3(a), showing strong agreement and indicating that the model successfully captures the intrinsic relationship between structural features and wave properties. For inverse design, the training dataset consists of 380 samples. The generative network is capable of producing latent vectors corresponding to target BGs. The generated unit cells are decoded by the VAE and subsequently validated using finite element simulations. As shown in Fig. 3(b), the designed structures exhibit BG characteristics consistent with the target values, demonstrating the effectiveness of the proposed inverse design strategy. Furthermore, the topological properties of the PnC can be tuned by translating the unit cell by half a lattice constant. By combining two unit cells with identical band structures but different topological properties, TISs are successfully realized. These results demonstrate that the proposed framework not only enables accurate BG prediction and inverse design, but also facilitates the realization of topological wave phenomena.

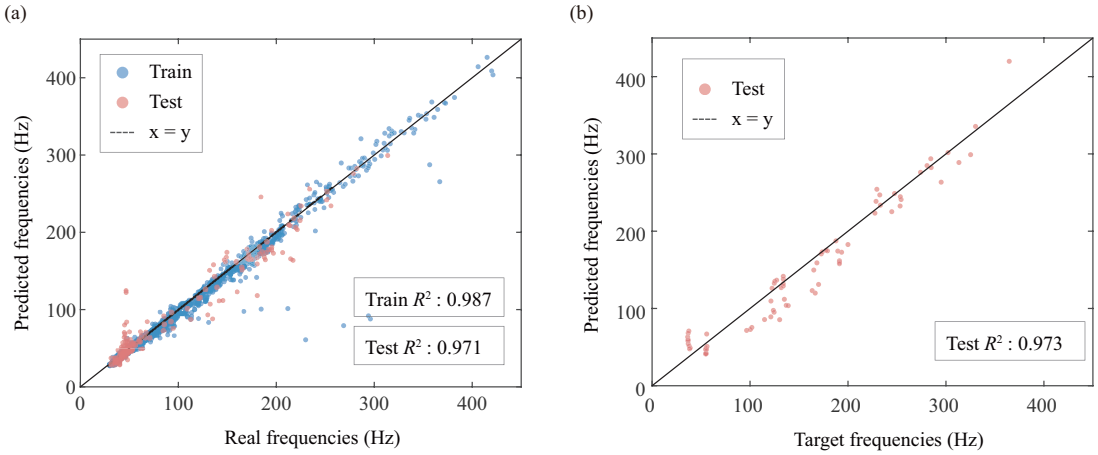


Figure 3: (a) Forward prediction results. (b) Inverse design results.

4 Conclusions

In this work, a deep learning-based framework is proposed for the efficient design of 3D PnCs with targeted BG. By integrating a VAE and MLPs, a mapping between unit cell images and elastic wave BGs is established under small-sample conditions. The proposed method achieves accurate forward prediction and effective inverse design of low-order odd BGs. Moreover, by exploiting the topological difference between translated unit cells, TISs are successfully realized and verified through numerical simulations. Overall, this work provides a data-efficient and robust approach for the design of topological PnCs, especially for computationally expensive 3D systems. The proposed framework offers a promising pathway for accelerating the development of advanced wave functional materials.

References

[1] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, p. 7 068 349, 2018.

[2] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE transactions on neural networks and learning systems*, vol. 32, pp. 604–624, 2020.

[3] W. Ma, F. Cheng, and Y. Liu, "Deep-learning-enabled on-demand design of chiral metamaterials," *ACS nano*, vol. 12, pp. 6326–6334, 2018.

[4] W. Ma, F. Cheng, Y. Xu, Q. Wen, and Y. Liu, "Probabilistic representation and inverse design of metamaterials based on a deep generative model with semi-supervised learning strategy," *Advanced Materials*, vol. 31, p. 1 901 111, 2019.

[5] K. Chen, X. Dong, P. Gao, Q. Chen, Z. Peng, and G. Meng, "Physics-informed neural networks for topological metamaterial design and mechanical applications," *International Journal of Mechanical Sciences*, vol. 301, p. 110 489, 2025.

Optimal Transport based models for Generative Design and Response Inference

Sergio Torregrosa[♭,1], David Munoz[♭], Hector Navarro[♭],
Charbel Farhat[†] and Francisco Chinesta[♮,♭]

(♭) ENSAM Institute of Technology, PIMM, UMR CNRS 8006,
151 Bv. de l'Hôpital, 75013 Paris, France.
(†) Stanford University, Department of Aeronautics and Astronautics,
Stanford, CA 94305-4035, USA.
(♮) CNRS@CREATE Ltd.,
1 CREATE Way, 08-01, CREATE Tower, 138602 Singapore.

Generative Design (GD) combines artificial intelligence (AI), physics-based modeling, and multi-objective optimization to autonomously explore and refine engineering designs [1]. Despite its promise in aerospace, automotive, and other high-performance applications, current GD methods face critical challenges: AI approaches require large datasets and often struggle to generalize; topology optimization is computationally intensive and difficult to extend to multiphysics problems; and model order reduction for evolving geometries remains underdeveloped [2, 3, 4]. To address these challenges, we introduce a unified, structure-preserving framework for GD based on Optimal Transport (OT) [5], enabling simultaneous interpolation of complex geometries and their associated physical solution fields across evolving design spaces, even with non-matching meshes and substantial shape changes. This capability leverages Gaussian splatting to provide a continuous, mesh-independent representation of the solution and Wasserstein barycenters [6] to enable smooth, mathematically "mass"-preserving blending of geometries, offering a major advance over surrogate models tied to static meshes. Our framework efficiently interpolates positive scalar fields across arbitrarily shaped, evolving geometries without requiring identical mesh topology or dimensionality. OT also naturally preserves localized physical features -- such as stress concentrations or sharp gradients -- by conserving the spatial distribution of quantities, interpreted as "mass" in a mathematical sense, rather than averaging them, avoiding artificial smoothing []. Representative test cases demonstrate enhanced efficiency, adaptability, and physical fidelity, establishing a foundation for future foundation-model-powered generative design workflows.

This work establishes a foundational framework for a novel approach to GD, elegantly leveraging OT theory to tackle the challenge of obtaining parametric solutions for engineering problems across new geometrical domains of arbitrarily different shapes, particularly when existing solution data for diverse geometries is limited. Recognizing the intricate nature of industrial design data, our proposed methodology first employs a re-parameterization of the geometry based on signed distance functions to achieve a unified representation of geometrical boundaries. Subsequently, to further unify the representation of corresponding solutions, our methodology decomposes both the geometry and its associated solutions into Gaussian functions -- a technique known as Gaussian splatting in computer graphics -- thereby circumventing the complexities arising from disparate geometrical parameterizations and solution dimensions.

[1] sergio.torregrosa_jordan@ensam.eu

To effectively manage the constraints of limited data and the underlying physical principles governing the solutions, we harness the power of OT as a robust solution inference tool. To mitigate the data requirements for accurate geometry blending, we strategically rely on the notion of Wasserstein barycenters for geometry interpolation. This proves particularly advantageous in shape optimization scenarios, allowing for the synthesis of plausible intermediate geometries from a sparse set of examples.

Addressing the inherent computational demands of traditional OT, we introduce a streamlined methodology predicated on discretizing fields into a collection of weighted particles, each associated with a Gaussian kernel. This particle-based representation elegantly reformulates the OT problem into a computationally tractable optimal assignment problem, drastically reducing its complexity and enhancing its practical applicability in design workflows [8].

The synergistic combination of our OT-based geometry and solution interpolation techniques with classical regression approaches empowers the creation of entirely new geometrical domains from existing ones and the subsequent inference of their corresponding parametric solutions. The observed accuracy of the method in different use cases, coupled with the near real-time responsiveness of the methodology in exploring novel geometrical domains and their associated parametric solutions, positions it as a potentially transformative asset within the engineering design process, enabling rapid exploration and evaluation of design alternatives.

References

[1] L. Regenwetter, A. Nobari, F. Ahmed, "Deep generative models in engineering design: A review", *Journal of Mechanical Design*, Vol. 114, pp. 71—704, 2022.

[2] S. Deshpande, A. Purwar, "Computational creativity via assisted variational synthesis of mechanisms using deep generative models", *Journal of Mechanical Design*, Vol. 141, 2019.

[3] S. Yoo, S. Lee, S. Kim, K. Hwang, J. Park, N. Kang, "Integrating deep learning into CAD/CAE system: generative design and evaluation of 3D conceptual wheel", *Structural and multidisciplinary optimization*, Vol. 64, pp. 2725—2747, 2021.

[4] E. Yilmaz, B. German, "Conditional generative adversarial network framework for airfoil inverse design", *AIAA aviation 2020 forum*, pp. 31—85, 2020.

[5] C. Villani, et al., "Optimal transport: old and new". Springer, 2008.

[6] E. Anderes, S. Borgwardt, J. Miller, "Discrete Wasserstein barycenters: optimal transport for discrete data", *Mathematical Methods of Operations Research*, Vol. 84, pp. 389—409, 2016.

[7] G. Peyré, M. Cuturi, "Computational Optimal Transport". *Foundations and Trends in Machine Learning*, Vol. 11, pp.355—607, 2019.

[8] S. Torregrosa, V. Champaney, A. Ammar, V. Herbert, F. Chinesta, "Surrogate parametric metamodel based on Optimal Transport". *Mathematics and Computers in Simulation*, Vol. 194, pp. 36—63, 2022.

Passive Vibration Control of a Clamped-Free Euler–Bernoulli Beam Using a Tuned Mass-Spring Absorber

Ana Pinto[♭,♯,1] Alain de Souza[♯] and Roberta Lima[♮]

(♭) Portuguese Air Force Academy
Sintra, Portugal.
(♯) Instituto Superior Técnico
Lisbon, Portugal.
(♮) Pontifícia Universidade Católica do Rio de Janeiro,
Rio de Janeiro, Brazil.

1 Introduction

Lightweight structures are widely used in aerospace applications, especially in satellite panels, solar arrays, and appendages [1]. Although these structures are efficient in terms of mass, they are also more sensitive to dynamic disturbances [2]. Small excitations may lead to significant vibrations [3], which can affect stability, pointing accuracy, and structural performance.

In space applications, such vibrations can degrade mission performance, particularly in high-precision instruments such as antennas or optical payloads, where strict pointing requirements must be satisfied [1], [2]. Furthermore, due to constraints in power, reliability, and system complexity, passive vibration control solutions are often preferred over active systems [4]. In this context, the use of tuned mass absorbers provides an effective and robust approach to mitigate vibrations without requiring external energy input [5].

A simplified way to study this problem is to analyse a slender beam with clamped-free boundary conditions [6]. Even though this is not a complete satellite model, it captures the essential behaviour of a flexible appendage subjected to external excitation [2]. For this reason, the Euler–Bernoulli beam model is adopted in the present work. The main objective of this study is to investigate how passive vibration control can be achieved through the addition of a mass-spring absorber. The beam is excited at different positions, and the absorber is also placed at different positions along the structure. The purpose is to identify which absorber location provides the best vibration attenuation for each excitation case. The analysis is carried out using the finite element method (FEM) [4]. Natural frequencies, mode shapes, frequency response functions (FRFs), time responses, and mechanical energy are evaluated. In addition, an efficiency map is built in order to compare the performance of the absorber for different combinations of forcing position and absorber position.

This study provides a simple but physically meaningful approach in order to understand how a satellite-like flexible panel may respond to perturbations and how passive control devices can reduce vibration levels. The results indicate that the tuned mass-spring absorber is highly effective at reducing vibration amplitudes at the target frequency, leading to very high efficiency values. However, these results correspond to an idealised numerical model without experimental validation. It is also observed that the absorber position has a strong influence on the performance, with locations closer to the free end providing the best results, while the forcing position has a comparatively smaller effect.

[1] ana.carolina.pinto@tecnico.ulisboa.pt

2 Methods

2.1 Beam model

The structure is modelled as an Euler–Bernoulli clamped-free beam [4]. The transverse vibration of an Euler–Bernoulli beam is governed by

$$EI\frac{\partial^4 w(x,t)}{\partial x^4} + \rho A\frac{\partial^2 w(x,t)}{\partial t^2} = 0, \tag{1}$$

where $w(x,t)$ is the transverse displacement, E is the Young's modulus of the material, ρ is the material density, A is the cross-sectional area of the beam, I is the second moment of area of the cross-section, and x is the spatial coordinate along the beam length. This partial differential equation is transformed into a discrete system using the finite element method. Each node has two degrees of freedom: transverse displacement and rotation. After finite element discretisation [7], the global equation of motion is written as

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{K}\mathbf{q} = \mathbf{f}(t), \tag{2}$$

where $\mathbf{M}$ is the mass matrix, $\mathbf{K}$ is the stiffness matrix, and $\mathbf{f}(t)$ is the external force vector. In this work, the beam is clamped at one end and free at the other. The fixed-end displacement and rotation are removed from the system in order to impose the boundary conditions [8].

2.2 Finite element discretisation and modal analysis

The beam is divided into Euler–Bernoulli finite elements [8]. A convergence study is first performed using a very fine FEM mesh as reference. This step is important to verify that the numerical model predicts the natural frequencies within a predefined accuracy and that the chosen number of elements is sufficient for the dynamic analyses.

Once the model is verified, approximations to the natural frequencies and mode shapes of the beam are obtained from the generalised eigenvalue problem [9]

$$\mathbf{K}\boldsymbol{\phi} = \omega^2\mathbf{M}\boldsymbol{\phi}, \tag{3}$$

where ω represents an approximation to the natural circular frequency and $\boldsymbol{\phi}$ an approximation to the corresponding mode shape.

2.3 External forcing

The beam is subjected to harmonic forcing applied at different locations along its span. In the present study, seven excitation positions are considered along the beam, namely $x = 0.16$, 0.32, 0.64, 1.92, 4.00, 6.00, and 8.00 m, as illustrated in Fig. 1. These positions are denoted by f_1 to f_7 in the figure. The goal is to understand how the excitation position changes the structural response and how it affects the performance of the passive control device.

2.4 Mass-spring absorber

A tuned mass-spring absorber operates by introducing an additional degree of freedom that interacts with the primary structure. When properly tuned, the absorber creates a dynamic interaction that splits the original resonance peak into two smaller peaks, leading to a reduction of the response at the target frequency.

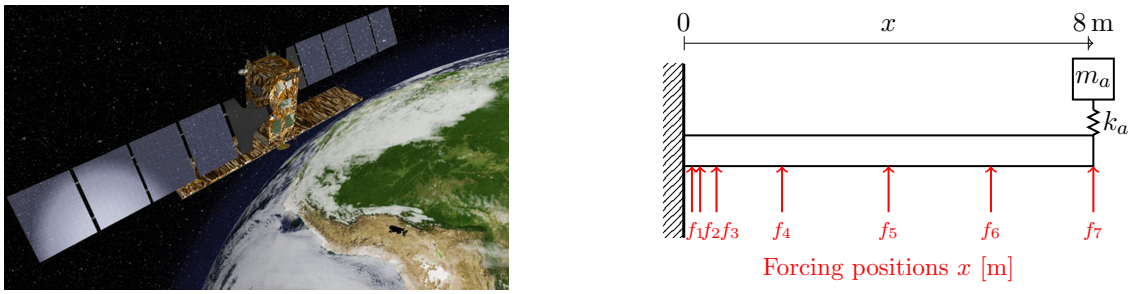


Figure 1: Satellite structure [2] and its simplified beam representation with a tip-mounted mass-spring absorber.

This phenomenon is associated with the creation of an anti-resonance condition in the frequency response. By the reference [5], the augmented equations of motion become

$$\mathbf{M}_a\ddot{\mathbf{x}}_a + \mathbf{K}_a\mathbf{x}_a = \mathbf{F}(t), \tag{4}$$

where the subscript a denotes the augmented system including the beam and the absorber. The absorber stiffness [5] is selected according to the tuning relation

$$k_a = m_a\omega_a^2, \tag{5}$$

where m_a is the absorber mass and ω_a is the target tuning frequency. In this work, the absorber is tuned to the first natural frequency of the original beam. Different absorber positions are tested along the beam in order to determine where the passive control action is most effective.

2.5 Frequency and time-domain response

The vibration response is analysed in both the frequency and time domains for two configurations: the original beam and the beam with an attached absorber. In the frequency domain, the frequency response function (FRF) of the original system [10] is given by

$$\mathbf{H}(\omega) = \left(\mathbf{K} - \omega^2\mathbf{M}\right)^{-1}, \tag{6}$$

where $\mathbf{K}$ and $\mathbf{M}$ are the global stiffness and mass matrices of the beam. When the absorber is included, the system is augmented to account for the additional degree of freedom associated with the absorber. The corresponding FRF becomes $\mathbf{H}_a(\omega) = \left(\mathbf{K}_a - \omega^2\mathbf{M}_a\right)^{-1}$, where $\mathbf{K}_a$ and $\mathbf{M}_a$ are the augmented stiffness and mass matrices including the absorber contributions. The FRF is used to evaluate how the absorber modifies the resonance peaks near the target frequency.

In the time domain, the dynamic response of both systems is obtained using the Newmark-β method under harmonic excitation [11]. The displacement at the beam tip is used as the main response quantity, since this point is typically the most flexible region of a clamped-free structure. The mechanical energy of the system is also evaluated. For the original beam [9], it is defined as

$$E(t) = \frac{1}{2}\dot{\mathbf{q}}^T\mathbf{M}\dot{\mathbf{q}} + \frac{1}{2}\mathbf{q}^T\mathbf{K}\mathbf{q}, \tag{7}$$

while for the system with an absorber, the energy becomes $E_a(t) = (1/2)(\dot{\mathbf{q}}^T\mathbf{M}_a\dot{\mathbf{q}} + \mathbf{q}^T\mathbf{K}_a\mathbf{q})$.

2.6 Efficiency metric

To compare all cases in a consistent manner, an absorber efficiency metric is defined based on the reduction of the FRF amplitude at the target frequency, a standard quantity in vibration analysis [12],

$$\eta_{\mathrm{FRF}} = 100\left(1 - \frac{|H_{\mathrm{abs}}(\omega_{\mathrm{target}})|}{|H_{\mathrm{orig}}(\omega_{\mathrm{target}})|}\right). \tag{8}$$

3 Results and Discussion

3.1 Model validation and modal analysis of the original beam

The convergence analysis shows that the FEM model provides stable and accurate results as the number of elements increases, as illustrated in Fig. 2. The first approximating modes converge quickly, while the higher modes require a finer mesh, as expected for bending-dominated systems. This confirms that the numerical implementation is suitable for the subsequent vibration and control analyses. The simulation parameters adopted in this study are summarised in Table 1.

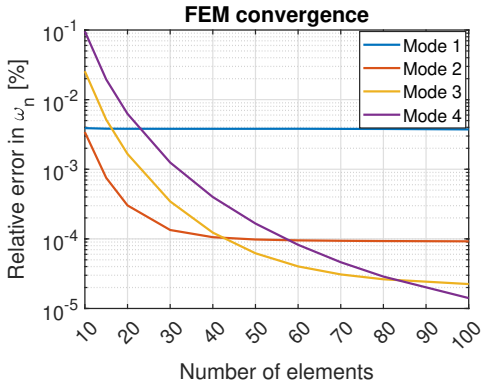


Figure 2: Relative error of the approximations of the first four natural frequencies.

Table 1: Simulation parameters.

Parameter	Value
Length L	8.0 m
Area A	6.25×10^{-4} m^2
Second moment I	3.4×10^{-7} m^4
Density ρ	1600 kg/m^3
Young's modulus E	30 GPa
Boundary condition	Clamped-Free

A mesh with $N = 50$ elements was adopted, as the natural frequencies of interest had already reached a converged regime, as shown in Fig. 2. The relative errors of the first modes were sufficiently small, and further refinement led to negligible improvements. The natural frequencies and mode shapes obtained for the clamped-free beam are consistent with the expected behaviour of an Euler–Bernoulli cantilever. The first mode is dominated by a smooth bending shape, while the higher modes present additional curvature and nodal regions. The mode shape comparison for different meshes also shows that the numerical approximation improves with refinement, especially for the higher-order modes. As the number of elements increases, the finite element solution converges towards a stable shape and frequency, with noticeable differences between coarse and fine meshes in the higher modes. This is evident in Fig. 3, where solutions with a larger number of elements overlap closely, indicating increased accuracy and convergence of the discretisation.

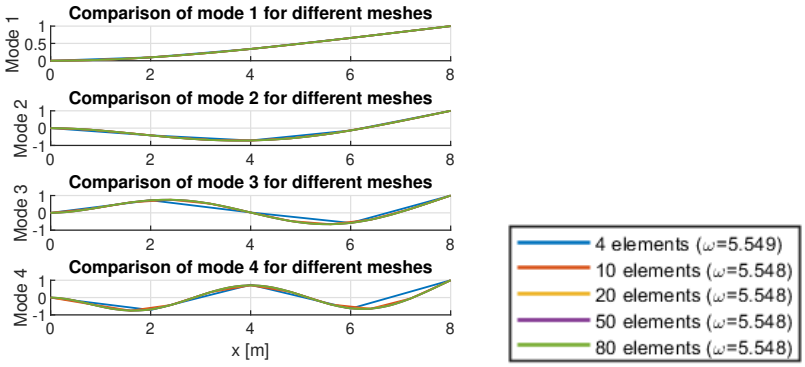


Figure 3: Convergence of the first four beam mode shapes with increasing number of finite elements.

3.2 Effect of the absorber on the FRF

The absorber is attached at the free end of the beam, while the external harmonic excitation is applied at $x = 1.92$ m, corresponding to f_4 in Fig. 1. The results obtained clearly demonstrate the effectiveness of the tuned mass–spring absorber. As shown in Fig. 4, the FRF of the original system exhibits a pronounced resonance peak at the target frequency, corresponding to the first natural frequency. After introducing the absorber, this peak is significantly reduced and split into two smaller neighbouring peaks, indicating the classical behaviour of a tuned vibration absorber and the creation of an anti-resonance region [4]. The time-domain response further confirms this behaviour. As illustrated in Fig. 5, the displacement amplitude of the original beam grows progressively due to resonance, whereas in the presence of the absorber the response remains bounded and significantly attenuated over time. Finally, the mechanical energy evolution, shown in Fig. 6, highlights a substantial reduction in the total energy of the system when the absorber is present. While the original system accumulates energy continuously due to resonance, the controlled system exhibits a much lower energy level, confirming that part of the vibrational energy is effectively transferred to the absorber. Overall, these results demonstrate that placing the absorber at the free end is highly effective for reducing vibrations at the target frequency.

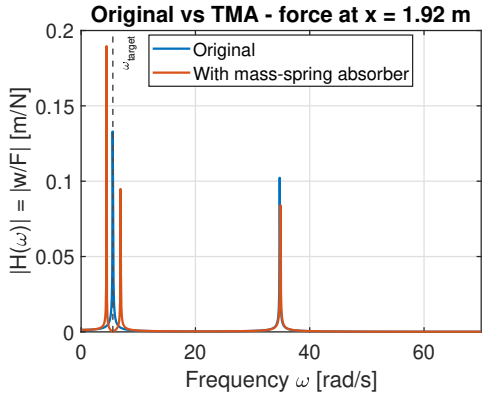


Figure 4: FRF at the beam tip (force at $x = 1.92$ m): original vs. absorber.

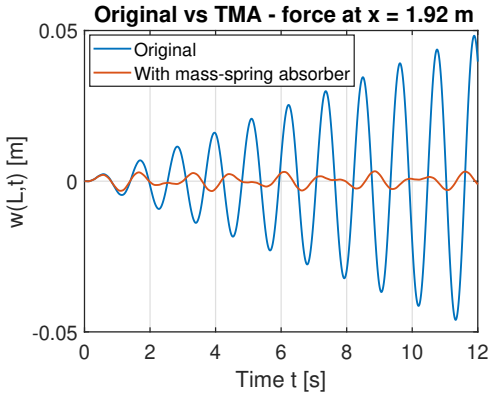


Figure 5: Time response at the beam tip (force at $x = 1.92$ m).

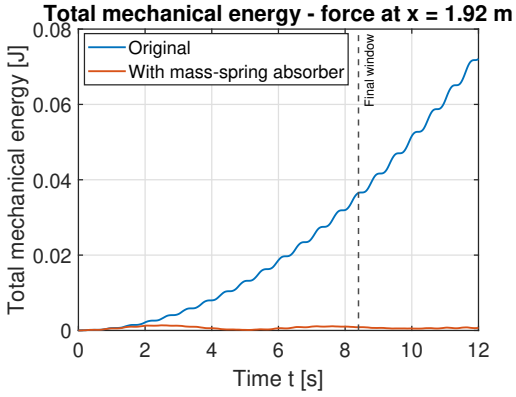


Figure 6: Mechanical energy: original vs. absorber.

3.3 Effect of forcing position

To analyse the influence of the forcing position on the efficiency of the absorber, the absorber location is kept fixed at the free end, while the excitation is applied at seven positions along the beam, namely $x = 0.16$, 0.32, 0.64, 1.92, 4.00, 6.00, and 8.00 m. For each forcing position, the efficiency is evaluated, allowing a systematic assessment of how the excitation location affects the vibration mitigation performance. The response of the beam depends on the location of the applied force. When the excitation is applied in regions of higher modal participation, the vibration amplitude becomes more significant. For a clamped-free beam, the first mode presents its maximum displacement near the free end, making this region more sensitive to external excitation. However, although the forcing position affects the absolute magnitude of the frequency response, its influence on the absorber efficiency is negligible. This behaviour can be explained by the definition of efficiency, which is based on the relative reduction of the frequency response function at the target frequency, as given in Eq. 8. At this specific frequency, the response is dominated by the corresponding mode shape. Since the absorber is tuned to this mode and placed near its maximum displacement, it effectively suppresses the response regardless of the forcing position. As a result, both the original and controlled responses scale similarly with the forcing location, leading to nearly constant efficiency values, as observed in Fig. 7. Therefore, within the present configuration and at that target frequency, the absorber efficiency is independent of the forcing location.

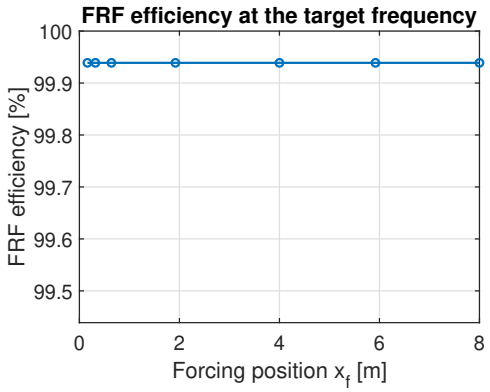


Figure 7: Effect of forcing position on FRF-based efficiency.

3.4 Effect of absorber position

The position of the absorber is a key parameter influencing the control performance. The results show that the absorber efficiency increases significantly as it is placed closer to the free end of the beam, as illustrated in Fig. 8. For positions near the clamped end, the efficiency is noticeably lower, whereas for positions closer to the tip, the efficiency approaches 100%. Beyond mid-span, the efficiency shows only marginal variation with further increases in the absorber position. This behaviour is consistent with the modal characteristics of the first bending mode, which exhibits its maximum displacement near the free end. When the absorber is placed in regions of higher modal displacement, the relative motion between the beam and the absorber mass is maximised, leading to a more effective reduction of the structural response. However, it should be noted that the nearly uniform high efficiency observed for positions beyond mid-span may also be influenced by the simplifying assumptions of the model, including the absence of damping and the use of a single-mode tuning approach. In more realistic structures, a stronger dependence on the absorber position could be expected.

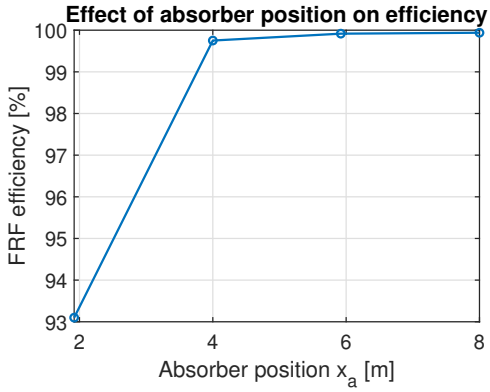


Figure 8: FRF-based efficiency as a function of absorber and forcing positions.

3.5 Global efficiency analysis

A global efficiency study was performed by combining several forcing positions with several absorber positions. As shown in Fig. 9, the efficiency maps indicate that the absorber does not perform equally well in all configurations. Instead, its effectiveness depends on the interaction between the excitation point and the absorber location. Among the tested cases, the best results are obtained when the absorber is placed closer to the free end of the beam. In these configurations, the FRF efficiency reaches values close to 100% at the target frequency, indicating a significant reduction in vibration amplitude. This result is physically reasonable because the first bending mode has its largest displacement near the beam tip. Therefore, placing the absorber in that region allows it to interact more effectively with the dominant vibration pattern.

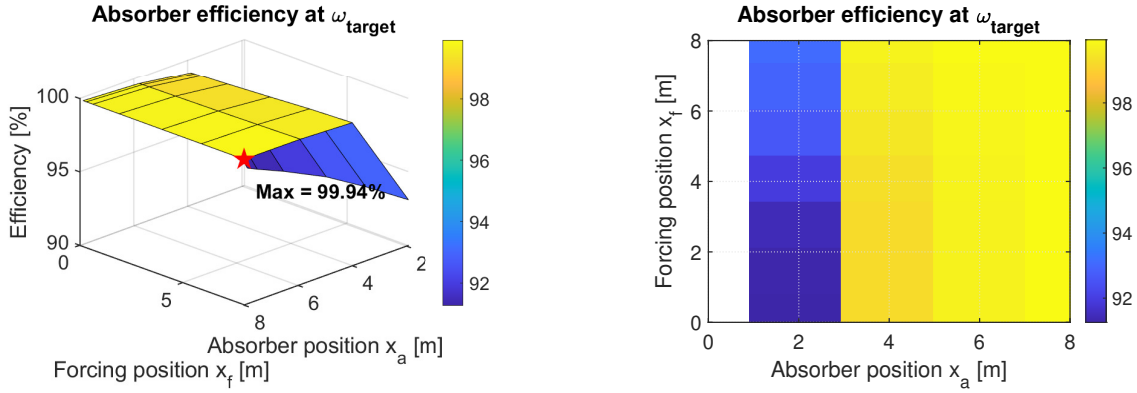


Figure 9: FRF-based efficiency maps as a function of forcing and absorber positions: (left) 3D representation and (right) 2D representation.

3.6 Interpretation for satellite structures

Although the model is simplified, the main conclusions are relevant for flexible aerospace structures such as satellite panels. Disturbances acting at different points may excite the structure in different ways, and the location of a passive control device is critical for obtaining good performance. The present beam model, therefore, provides a useful first approximation for understanding the dynamic behaviour of flexible appendages and for guiding the placement of passive absorbers in preliminary design stages.

4 Conclusions

This work studied the vibration behaviour of a clamped-free Euler–Bernoulli beam subjected to harmonic excitation, as well as the effect of introducing a passive mass-spring absorber. Additionally, the simplified beam model offers valuable insight into the vibration control of flexible structures such as satellite panels. The finite element model was successfully verified through a convergence study, ensuring the reliability of the numerical results. The computed natural frequencies and mode shapes are consistent with the expected behaviour of a clamped-free flexible structure. The addition of the mass-spring absorber leads to a significant reduction of the vibration response at the target frequency. The results show that the absorber position plays a crucial role in the control of the vibration, with the highest efficiency achieved when the device is placed near the free end of the beam, where the modal displacement is largest. In contrast, the position of the applied force has a negligible influence on the absorber efficiency, as it primarily affects the amplitude of the response rather than the relative reduction achieved by the control device. Overall, the study demonstrates that passive vibration control using a tuned mass-spring absorber can be highly effective at a specific target frequency, provided that the absorber is properly tuned and appropriately positioned along the structure. Future work may consider other target frequencies, including higher-order modes, to assess how the influence of forcing and absorber positions varies with modal shape.

References

[1] L. Celentano, "An approach to model and control a flexible spacecraft," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 11 551–11 557, 2023, ISSN: 2405-8963. DOI: 10.1016/j.ifacol.2023.10.448.

[2] L. Mazzini, *Flexible Spacecraft Dynamics, Control and Guidance*. Springer, 2015, ISBN: 978-3-319-16108-1. DOI: 10.1007/978-3-319-16109-8.

[3] R. Sampaio, P. Almeida, and T. Ritto, *Vibrações Mecânicas: Dinâmica de Estruturas Flexíveis*. Rio de Janeiro, 2007. DOI: 10.13140/RG.2.1.1775.0241.

[4] D. J. Inman, *Engineering Vibration*, 4th ed. Upper Saddle River, NJ: Pearson Education, 2014, ISBN: 9780132871693.

[5] S. Krenk and J. Høgsberg, "Tuned mass absorber on a flexible structure," *Journal of Sound and Vibration*, vol. 333, no. 6, pp. 1577–1595, 2014. DOI: 10.1016/j.jsv.2013.11.015.

[6] J. Wei, W. Liu, J. Liu, and T. Yu, "Dynamic modeling and analysis of spacecraft with multiple large flexible structures," *Actuators*, vol. 12, no. 7, p. 286, 2023. DOI: 10.3390/act12070286.

[7] R. R. Craig and A. J. Kurdila, *Fundamentals of Structural Dynamics*, 2nd ed. Hoboken, NJ, USA: John Wiley & Sons, 2006, ISBN: 978-0-471-43044-5.

[8] A. C. Matutti and R. Sampaio, *Introdução ao método dos elementos finitos (mef)*, Departamento de Engenharia Mecânica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brazil.

[9] L. Meirovitch, *Fundamentals of Vibrations*, 1st ed. Long Grove, IL, USA: Waveland Press, 2001, Reissued in 2010 by Waveland Press, ISBN: 978-1-57766-691-2.

[10] D. J. Ewins, *Modal Testing: Theory, Practice and Application*, 2nd ed. Baldock, Hertfordshire, England: Research Studies Press, 2000, ISBN: 0-86380-218-4.

[11] J. Angeles, *Dynamic Response of Linear Mechanical Systems: Modeling, Analysis and Simulation*. Springer, 2011, ISBN: 978-1-4419-1026-4. DOI: 10.1007/978-1-4419-1027-1.

[12] S. G. Kelly, *Schaum's Outline of Theory and Problems of Mechanical Vibrations*, 1st ed. New York: McGraw-Hill, 1996, ISBN: 0-07-035048-5.

Heuristic Methods for Almond Percentage Detection via Least Squares

Alejandro Arques[♭],[1] Ferran Verdú[♮] Mónica Cortés [♮]

(♭) Colegio El Valle, Alicante
(♮) D. de Matemática Aplicada e Ingeniería Aeroespacial, Univ. de Alicante
Crtra. de S. Vicente s/n, Alicante, Spain.

1 Introduction

One of the great beauties of science is its connection to the real world. Human beings need the creation of mathematical models to attempt to predict certain characteristics of interest with a high probability of success. In regression analysis, the goal is to find relationships between variables and understand how one variable depends on the others. Linear regression is a well-studied type of regression that can provide a reasonably accurate model of the relationship between variables. However, we may wonder: is there a functional relationship, with other types of functions, that can improve our linear prediction?

Nonlinear regression, also known as non-linear least squares, is not as extensively studied as its linear counterpart due to various reasons, including the vast number of functions and different cases that can be considered. In this study, we focus on nonlinear regression models with separable variables, developing a Python algorithm that helps us find the best relationships between variables using the Gauss–Newton method. Furthermore, a heuristic is employed to attempt to determine which choice of the starting point is better or more efficient than the others. Finally, we explore the practical application of this algorithm in the determination of ingredients in agri-food products such as Jijona nougat.

1.1 Preliminary Concepts and definitions

In regression models, the goal is to find a functional relationship or estimate between certain variables considered independent with respect to another variable to be predicted (the dependent variable).

Definition 1.1. A regression model can be written mathematically as:

$$Y \approx f(X, \theta) \tag{1}$$

where X represents the vector of p independent or explanatory variables used to predict the variable Y, f is the function that defines the type of relationship between variables, and θ is the vector of unknown parameters.

[1] alexixona2000@gmail.com, ferran.verdu@ua.es , monica.cortes@ua.es

Definition 1.2. In some cases it is useful to define a regression model as:

$$Y = f(X, \theta) + \varepsilon \tag{2}$$

where ε represents the error committed in the estimation.

The objective of regression is to assign a value to each of the parameters in θ so that the approximation is as good as possible. To this end, a large number of samples from the population are collected to fit the functional relationship well.

Definition 1.3. A regression model is said to be **linear** if, and only if, the function is linear in each of the parameters. Equivalently:

The model in (1) is linear if, and only if, $\dfrac{\partial}{\partial \theta_i} f(X, \theta)$ does not depend on θ_i, for all i. Otherwise, we say we face a **nonlinear regression model**.

Definition 1.4 (Separable Variable Model)**.** Suppose that in a regression model we observe p explanatory variables $(x_1, x_2, \ldots, x_p)$ and wish to study the behaviour of a dependent variable y. Consider a nonlinear model of the form:

$$y \approx \sum_{j=1}^{p} f_j(\theta, x_j) \tag{3}$$

Models that consist of a sum of different functions, each depending on only one variable, are called **separable variable models**.

Nonlinear regression models with separable variables will be the main focus of this study. Before continuing with the central development, however, we first examine linear regression models in more depth, [1]. Although they are not the object of study here, understanding their solution method will be very useful when dealing with nonlinear models.

2 Python Implementation

In order to find the best model approximating a given dataset, we require a Python algorithm that analyses all possible function types and applies the Gauss–Newton method to each one, seeking a sufficiently good solution. An implementation of this algorithm can be found in [2], [3], where an iterative search over different models is performed by applying the Gauss-Newton sub-algorithm.

First, a data-reading module is needed to collect and organise the variable data and the different function types provided by the user.

Before starting the algorithm, the different functions to be analysed must be generated. The user creates a text file containing the candidate function types, and the module then generates all possible combinations following the separable variable model. The structure of the module is as follows.

Suppose the user provides k different functions: $f_1, f_2, \ldots, f_k$, and the regression problem has p independent variables: $x_1, x_2, \ldots, x_p$. We organise the information as in Table 1.

Variable x_1	Variable x_2	$\cdots$	Variable x_p
$f_1(x_1, \theta)$	$f_1(x_2, \theta)$	$\cdots$	$f_1(x_p, \theta)$
$f_2(x_1, \theta)$	$f_2(x_2, \theta)$	$\cdots$	$f_2(x_p, \theta)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$f_k(x_1, \theta)$	$f_k(x_2, \theta)$	$\cdots$	$f_k(x_p, \theta)$

Table 1: Function generation table

Selecting one function per column and summing them yields the complete model. Each complete function is determined by a p-tuple. Let $S_k = \{1, 2, \ldots, k\}$. The total number of functions that can be created equals the number of variations with repetition of k elements taken in groups of p: $V_{k,p} = k^p$.

Example 2.1. Suppose two base function types are provided:

$$f_1 = A \cdot \sin(Bx + C), \qquad f_2 = A \cdot \ln(Bx + C) \tag{4}$$

with three explanatory variables (x_1, x_2, x_3). The generation table is shown in Table 2.

	Variable x_1	Variable x_2	Variable x_3
1	$A_0 \sin(A_1 x_1 + A_2)$	$A_3 \sin(A_4 x_2 + A_5)$	$A_6 \sin(A_7 x_3 + A_8)$
2	$A_9 \ln(A_{10} x_1 + A_{11})$	$A_{12} \ln(A_{13} x_2 + A_{14})$	$A_{15} \ln(A_{16} x_3 + A_{17})$

Table 2: Example function table

Therefore, each complete function is uniquely determined by a 3-tuple. In this case, we only have two types of basis functions to consider: $S = \{1, 2\}$. The total number of functions is $V_{2,3} = 2^3 = 8$. These are listed in Table 3.

#	Function
1	$A_0 \sin(A_1 x_1 + A_2) + A_3 \sin(A_4 x_2 + A_5) + A_6 \sin(A_7 x_3 + A_8)$
2	$A_0 \sin(A_1 x_1 + A_2) + A_3 \sin(A_4 x_2 + A_5) + A_{15} \ln(A_{16} x_3 + A_{17})$
3	$A_0 \sin(A_1 x_1 + A_2) + A_{12} \ln(A_{13} x_2 + A_{14}) + A_6 \sin(A_7 x_3 + A_8)$
4	$A_0 \sin(A_1 x_1 + A_2) + A_{12} \ln(A_{13} x_2 + A_{14}) + A_{15} \ln(A_{16} x_3 + A_{17})$
5	$A_9 \ln(A_{10} x_1 + A_{11}) + A_3 \sin(A_4 x_2 + A_5) + A_6 \sin(A_7 x_3 + A_8)$
6	$A_9 \ln(A_{10} x_1 + A_{11}) + A_3 \sin(A_4 x_2 + A_5) + A_{15} \ln(A_{16} x_3 + A_{17})$
7	$A_9 \ln(A_{10} x_1 + A_{11}) + A_{12} \ln(A_{13} x_2 + A_{14}) + A_6 \sin(A_7 x_3 + A_8)$
8	$A_9 \ln(A_{10} x_1 + A_{11}) + A_{12} \ln(A_{13} x_2 + A_{14}) + A_{15} \ln(A_{16} x_3 + A_{17})$

Table 3: Generated functions

The central part of the program is the Gauss–Newton algorithm.

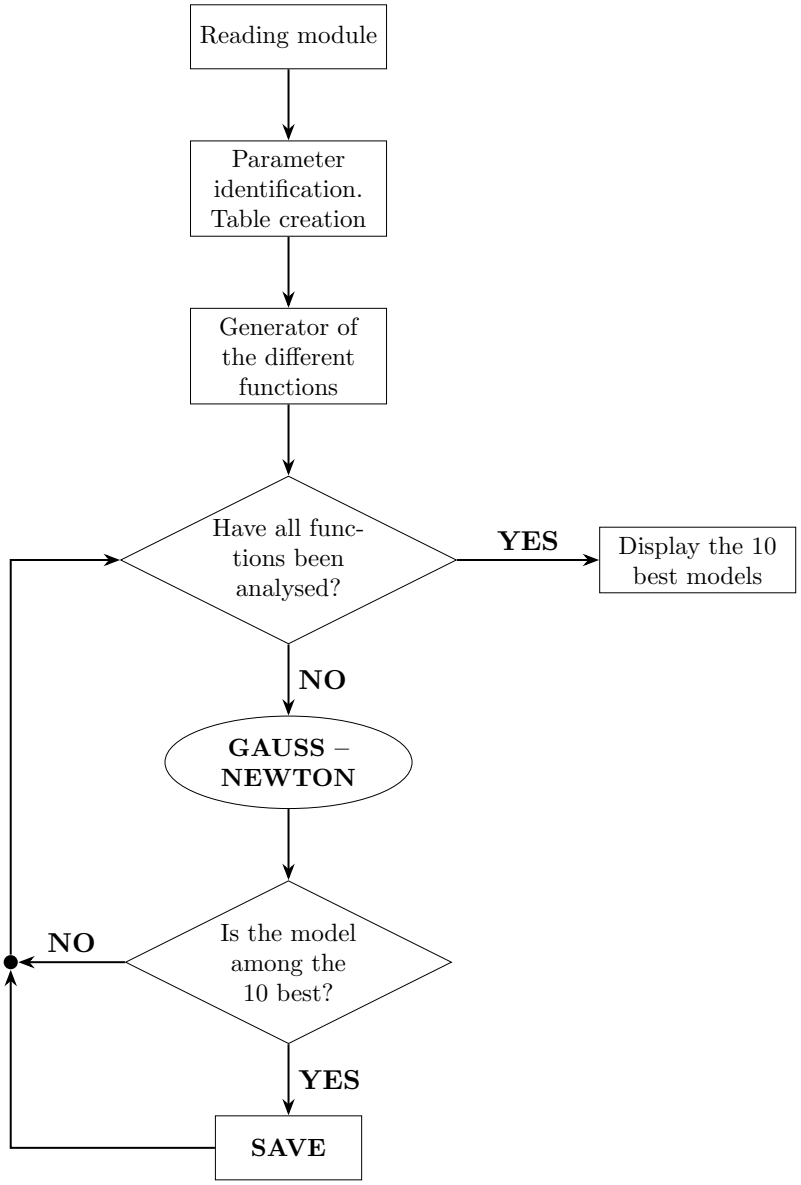


Figure 1: Diagram of the algorithm

- **Method 1 (Fixed starting point):** All parameters are initialised to $A_i = 1$ at the start of each Gauss–Newton run.

- **Method 2 (Recovering best values):** Parameters are initialised to 1 for the first run. After convergence for each model, the parameters that yielded the best result are identified and stored. These become the starting point for subsequent runs.

In summary, in Method 1 the starting point is always the same, while in Method 2 the parameter values that have given the best approximations so far are recovered to initialise the algorithm.

3 Almond Percentage Detection in Agri-Food Products

In this section the algorithm is applied and results are presented for for the nougat industry application.

Jijona and Alicante nougats (*turrones*) are the quintessential Christmas sweets in Spain, made with a high proportion of almonds; the best variety for their production is the Marcona almond, due to its high oil content. The *Consejo Regulador del Turrón* (Nougat Regulatory Council) is the independent body responsible for analysing the quality of each product and awarding the Protected Geographical Indication (PGI). Two fundamental requirements must be met:

- The product must be manufactured within the municipality of Jijona.
- Traditional processes must be followed with respect to ingredient proportions and production methods.

The second requirement is where the algorithm developed in this work finds a practical application: one of the main conditions is a minimum almond content. For the *Calidad Suprema* (Supreme Quality) grade analysed here, this minimum is 64%.

The main difference between Jijona and Alicante nougat lies in the production process, since the ingredients are exactly the same. In Alicante nougat, the white paste made from egg white, honey, and sugar is mixed directly with whole almonds; in Jijona nougat, everything (paste and toasted almonds) is ground in a stone mill and then emulsified in the *boixet*. This difference makes it impossible to determine the quantity of almond integrated into Jijona nougat by simple inspection.

A total of 127 nougat samples from the production campaigns between 2020 and 2022 were collected. Five relevant variables were measured analytically:

$$X_0 \equiv \text{Moisture } X_1 \equiv \text{Fibre } X_2 \equiv \text{Protein } X_3 \equiv \text{Fat } X_4 \equiv \text{Total Sucrose}$$

The objective is to obtain a good model approximating the dependent variable:

$$Y \equiv \text{Almond percentage}$$

Full regulatory information can be found in [4] and in the Official State Gazette (B.O.E.) [5].

The data were provided by the Product Characterisation Service of the Association of Nougat, Derivatives and Chocolates Manufacturers of the Valencian Community (T.D.C.) (Data in [6]), the independent official body for the accreditation of Protected Designations of Origin.

With five independent variables and the seven base functions of Table **??**, the program would need to analyse $7^5 = 16807$ models. To reduce execution time, only the five most influential base functions are retained (Table 4), giving $5^5 = 3125$ candidate functions.

Functions
$A(Bx + C)^2$
$A(Bx + C)^3$
$A\sin(Bx + C)$
$A\tan(Bx + C)$
$A\arctan(Bx + C)$

Table 4: Functions provided for the nougat analysis

The three best models from Method 2 (parameter recovery) are shown in Table 5.

#	Model	$\|\varepsilon\|$	It.
1	$10^{-8}(-107.7x_0 + 217.7)^3 + 1.75(0.64x_1 - 3.7)^3 - 0.89\sin(0.89x_2 + 2.97)$ $-3.08\sin(-0.135x_3 + 48.415) + 0.032(0.2x_4 - 52.79)^2$	2.8679	22
2	$-0.15\arctan(2.6 \times 10^5 x_0 - 1.27 \times 10^6) + 1.74(0.63x_1 - 3.66)^3$ $-4.6 \times 10^{-5}(37.7x_2 - 760)^2 - 1.9\sin(-0.2x_3 + 51.67) + 0.03(0.19x_4 - 52.8)^2$	2.8700	19
3	$5.97(-118.8x_0 + 210.7)^3 + 1.3(0.59x_1 - 3.2)^3 - 0.8\sin(0.95x_2 + 2.07)$ $-3.7(-0.003x_3 - 2.87)^3 + 2.88 \times 10^{-6}(-1.5x_4 - 159.32)^3$	2.8819	20

Table 5: Models obtained with parameter recovery (Method 2): nougat data

The three best models from Method 1 (fixed initial point) are shown in Table 6.

#	Model	$\|\varepsilon\|$	It.
1	$-0.007\tan(-383.34x_0 + 1428.84) + 1.87(0.59x_1 - 3.2)^3 - 1.8(0.2x_2 - 4.1)^2$ $+0.0082(0.19x_3 + 87.7)^2 - 34.8(0.2x_4 - 0.25)^2$	2.8301	71
2	$0.02\tan(-28.05x_0 + 108.96) + 3.07(-0.6x_1 + 3.97)^2 - 1.07\sin(-0.75x_2 + 29.5)$ $-0.2(-0.3x_3 + 14.5)^2 + 3.44(0.017x_4 - 5.12)^2$	2.8449	55
3	$2.12 \times 10^{-6}(-18.14x_0 + 36.5)^3 + 1.87(0.6x_1 - 3.4)^3 - 0.87\sin(0.92x_2 + 2.52)$ $-2.76(0.085x_3 - 3.89)^2 + 140.19(0.003x_4 - 0.81)^2$	2.8682	38

Table 6: Models obtained with fixed initial point (Method 1): nougat data

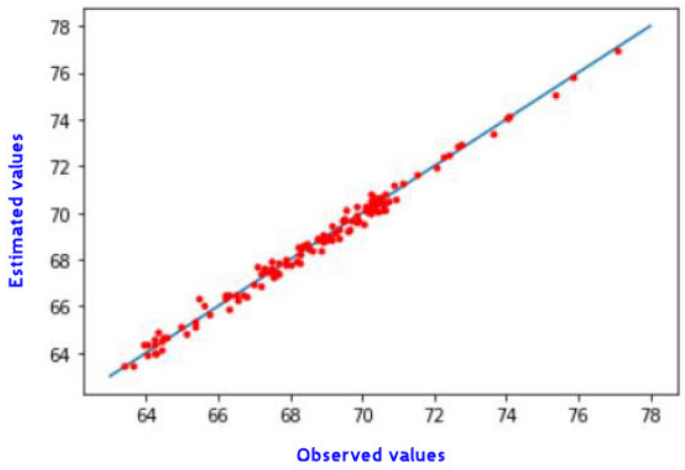


Figure 2: Observated vs estimated: Method of fixed starting point

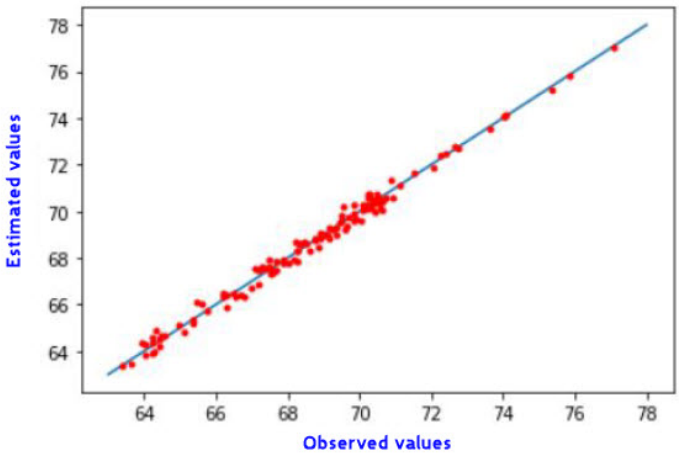


Figure 3: Observated vs estimated: Mehod of recovering best values

4 Conclusions

In this section we present a critical analysis of the results obtained. The main objective was to compare both criteria for choosing the starting point when applying the Gauss–Newton method.

We begin by comparing the number of iterations required to obtain each of the six best models.

File	Method 1 (init. at 1)	Method 2 (best values)	Max. It.
`torrosuprema.txt`	(35, 35, 20, 20, 35, 35)	(5, 3, 5, 6, 6, 5)	35

Table 7: Number of iterations for the six best models

The method incorporating parameter recovery is clearly more efficient, substantially reducing the number of iterations required for convergence. Table 8 compares execution times.

File	Method 1 (init. at 1)	Method 2 (best values)
`torrosuprema.txt`	21438.15 s	15431.49 s

Table 8: Program execution times

From Table 8 it can be seen that Method 2 is more efficient overall, particularly when a larger number of functions needs to be analysed. The difference is minimal or negligible when the total time is between 1 and 20 minutes.

Finally, Table 9 compares the ℓ_2 norms of the residual vector for the best and worst of the ten models generated.

	Method 1		Method 2	
File	Best	Worst	Best	Worst
`torrosuprema.txt`	2.8301	2.8969	2.8679	2.9098

Table 9: ℓ_2 norm of residual vectors

Although the difference between the best models of each method is not always large, there is a notable improvement in the tenth-best model when using Method 2, meaning the program generates a greater number of models that are comparable to the best one found. The only exception is the almond-percentage prediction, where Method 2 performs marginally worse; however, the differences are negligible, and given the time savings, it is highly advisable to use the parameter-recovery method.

In conclusion, using the variable starting-point method (Method 2) is optimal compared to the fixed starting-point method (Method 1). In the majority of cases, it achieves a better fit, a reduction in the total number of iterations, and therefore a reduction in execution time.

5 Future Research

This work opens up several lines of future investigation:

- Establish pruning criteria in the function space to discard functions that cause convergence errors in a sufficient number of cases, and to stop testing functions that consistently fail to fit the data.
- Change the order in which different functions are tested, as well as the order in which the sum of functions is generated.
- Introduce a step-size factor into the algorithm.
- Test other search algorithms besides Gauss–Newton, including the gradient-descent algorithm.
- Design an assistant program.
- Attempt to estimate the total order of iterations for each method.

Funding

his research has been funded by the research project "Mathematical Modelling of Systems" through the grants for established research groups AICO-2025 granted by Secretaría Autonómica de Universidades (Generalitat Valenciana, Spain) (CIAICO/2024/130)

References

[1] M. Cortés, Y. Villacampa, J. Mateu, and J. L. Usó, "A new methodology for modelling highly structured systems," *Environmental Modelling & Software*, vol. 15, pp. 461–470, Apr. 2000. DOI: 10.1016/S1364-8152(00)00024-4.

[2] D. Bates and D. G. Watts, *Nonlinear Regression Analysis and Its Applications*. John Wiley & Sons, 2007.

[3] F. Verdú and Y. Villacampa, "A computational algorithm for the multiple generation of non-linear mathematical models and stability study," *Advances in Engineering Software*, vol. 39, pp. 430–437, May 2008. DOI: 10.1016/j.advengsoft.2007.03.004.

[4] Consejo Regulador de Jijona y del Turrón de Alicante. "El turrón," Accessed: Jul. 1, 2023. [Online]. Available: https://jijona.com/el-turron/.

[5] *Real decreto 1787/1982, de 14 de mayo, por el que se aprueba la reglamentación técnico-sanitaria para la elaboración y venta de turrones y mazapanes*, *BOE* núm. 183, de 2 de agosto de 1982, 1982.

[6] A. Arques., *Nougat data*, Accessed: April 30, 2026. [Online]. Available: `https://personal.ua.es/va/ferran.verdu/dades-torro/data-nougat.pdf`.

Human Action Recognition using Depth and Skeleton Data through V-JEPA2

J. Alberto Conejero[♭],[1] and Lluïsa Gual-Vayà[♭,♮]

(♭) Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

(♮) Department of Mathematics, Universitat Jaume I,
Avda. Sos Baynat s/n, Castelló, Spain.

1 Introduction

Human Action Recognition (HAR) from video sequences plays a key role in many practical applications, such as monitoring environments, analyzing sports performance, supporting healthcare, improving security systems, and enabling more natural interaction between humans and technology. Traditionally, the field has evolved through various data modalities, including RGB, skeletal joints, point clouds, and depth maps, transitioning from classical pipelines based on handcrafted feature engineering to sophisticated end-to-end deep learning architectures [7].

In the RGB and depth domains, early methods relied on spatio-temporal interest points and motion descriptors. The advent of deep learning introduced Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and 3D-CNNs capable of modeling complex spatio-temporal dynamics directly from video volumes. Similarly, skeleton-based recognition has shifted from kinematic descriptors to Graph Convolutional Networks (GCNs) that explicitly model the relational structure of human body joints. Meanwhile, point cloud-based approaches have harnessed 4D-CNNs and Transformers to manage the irregular nature of 3D surface geometry.

Despite their strong empirical performance, most state-of-the-art models, such as hybrid CNN-GCN-Transformer architectures, rely heavily on carefully designed spatio-temporal inductive biases. These models often require predefined skeletal topologies or explicit modularization to separate spatial and temporal information. While effective, these designs are highly specialized and tightly coupled to specific input representations.

In contrast, the Video Joint Embedding Predictive Architecture (V-JEPA2) [1] introduces a modality-agnostic predictive representation learning framework. Unlike traditional models, V-JEPA2 does not rely on handcrafted structural assumptions about the human body. Instead, it learns to predict future latent representations, encouraging the emergence of high-level semantic and dynamical features without explicit structural priors. This raises a fundamental research question: *can a general predictive embedding model achieve competitive HAR performance without highly specialized architectural adaptations?*

To investigate this hypothesis, we have used the MSR Action3D dataset [4, 6], a classial benchmark in the field. Since its introduction, this dataset has been a primary gold standard, with

[1] aconejero@upv.es

its usage growing exponentially to test everything from skeletal geometry [8] to advanced spatial-temporal fusion [5, 10]. Recent high-performing methods often employ multi-modal fusion and complex feature extraction pipelines by integrating depth motion maps and skeletal data[2] or through dynamic spatial modeling of body regions [3].

Motivated by the robustness of general predictive models, we propose using V-JEPA2 for human action classification under three distinct input representations:

1. Sequences of raw depth maps.
2. Skeletal data consisting of 20 joint positions with their inherent structural relationships.
3. The same 20 joint coordinates treated independently, without explicitly encoding connectivity.

Our experimental results demonstrate that relying exclusively on depth map sequences through V-JEPA2 achieves a competitive accuracy of 92.35%, which is competitive with the aforementioned methods. This is particularly noteworthy as our model achieves these results without the complex feature engineering or the multi-modal fusion utilized in other leading works, underscoring the potential of general predictive architectures in modeling human behavior and engineering intelligent recognition systems.

1.1 Data Modalities and Representations

The MSR Action3D dataset contains 20 action categories performed by 10 subjects. We process this data using three distinct input representations to assess the robustness of the V-JEPA framework:

1. **Depth Maps (DM):** Video sequences of 640×480 depth frames captured at 15 fps. This representation exploits the full geometric structure and distance information of the scene.
2. **Skeleton-Based (SB):** Temporal sequences of 20 anatomical joints. Each joint is defined by coordinates (u, v, d) and a confidence score c. V-JEPA is applied to these structured joint configurations to model body dynamics.
3. **Point-Based (PB):** Derived from skeletal data, this representation ignores explicit connectivity (segments). A rigid rotation is applied to normalize orientation, ensuring the head-to-foot axis is consistently aligned. This isolates the spatio-temporal evolution of joint positions from topological priors.

1.2 The V-JEPA Framework

V-JEPA is a self-supervised architecture that learns latent representations by predicting missing or future video segments in an embedding space. Formally, given a sequence $X = \{x_1, x_2, \ldots, x_T\}$, the model learns a mapping $f_\theta : X \to \mathbb{R}^d$ that produces a compact feature vector $z = f_\theta(X)$.

By training f_θ to predict future embeddings z_{t+k} from a context z_t, the model captures: (i) the 3D spatial geometry of the subject, (ii) the temporal dynamics of the movement, and (iii) the semantic patterns of the human actions. This predictive mechanism allows for motion and shape understanding that is invariant to texture and illumination.

1.3 Action Classification and Experimental Protocols

We have evaluated the quality of the learned embeddings z through two standardized experimental settings:

- **Experimental Setting 1 [6]:** The dataset is divided into three subsets (AS1, AS2, AS3) of eight actions each (see Table 1). For each subset, 75% of the samples are used for training and 25% for testing.
- **Experimental Setting 2 [9]:** A more challenging protocol where all 20 action classes are considered simultaneously, maintaining the 75%/25% training-test split.

In both protocols, V-JEPA first learns the embeddings in a self-supervised manner. Subsequently, a classification head is trained on these fixed representations to map z to the corresponding action labels, testing the hypothesis that general predictive embeddings can surpass specialized, hand-crafted descriptors in HAR tasks.

Table 1: Action subsets for Experimental Setting 1.

AS1	**AS2**	**AS3**
Horizontal wave	High wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw x	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Two hand wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pickup throw	Side boxing	Pickup throw

2 Results

The embeddings generated by the V-JEPA model were utilized as high-dimensional features for supervised action classification. We evaluated five distinct classifiers: Support Vector Machine (SVM) with RBF kernel, k-Nearest Neighbors ($k = 5$), Gradient Boosting, Random Forest, and L2-regularized Logistic Regression.

Following Experimental Setting 2 (all 20 actions), Logistic Regression consistently achieved the highest performance across all input representations (Table 2). This suggests that V-JEPA predictive pre-training successfully linearizes the manifold of human actions, placing different categories into nearly linearly separable regions of the latent space.

Table 2: Overall classification accuracy (%) on MSR Action3D (Setting 2).

Classifier	**Depth Maps**	**Skeleton**	**Points**
Logistic Regression	**89.01**	**76.44**	**75.39**

The depth map representation yielded the best results, as it preserves the full geometric and volumetric detail of the motion. While the skeleton and point-cloud modalities showed a slight decrease in accuracy, they remain highly competitive, demonstrating V-JEPA's robustness even when structural connectivity is omitted (point-cloud case).

For Experimental Setting 1 (subgroup-based), the depth map modality achieved a global accuracy of **92.35%**, with peak performance in subset AS3 (94.59%). The detailed breakdown in Table 3 highlights that the model benefits from the richness of depth information across all subsets.

The consistent superiority of Logistic Regression over non-linear models (e.g., Random Forest or Gradient Boosting) can be attributed to the nature of the V-JEPA embeddings. The predictive learning objective encourages the emergence of semantically rich, compact, and convex class

Table 3: Per-subgroup accuracy (%) using Logistic Regression (Setting 1).

Group	Depth Maps	Skeleton	Points
AS1	91.78	80.26	84.21
AS2	90.67	76.92	75.64
AS3	94.59	93.42	88.16
Total	**92.35**	**83.53**	**82.67**

clusters. In this high-dimensional latent space ($d \gg N$), complex non-linear models are prone to overfitting spurious correlations within the relatively small MSR Action3D dataset. Conversely, L2-regularized Logistic Regression effectively manages model capacity, providing better generalization and reliable decision boundaries that align with the smooth geometric variations of human motion.

3 Conclusions

In this work, we have successfully demonstrated the application of the V-JEPA2 for the task of human action recognition using the MSR Action3D dataset. By evaluating the model across three distinct data representations, depth maps, skeletal sequences, and independent joint points, we have established several key conclusions regarding the robustness of general predictive embeddings.

Our results indicate that self-supervised predictive representation learning provides a powerful and stable alternative to handcrafted inductive biases. This approach simplifies the engineering pipeline while maintaining high accuracy, offering a promising direction for future research in general-purpose spatio-temporal modelling of human behaviour.

References

[1] Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al. (2025). V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985.* https://doi.org/10.48550/arXiv.2506.09985

[2] Chao, X., Ji, G., and Qi, X. (2024). Multi-view key information representation and multi-modal fusion for single-subject routine action recognition. *Applied Intelligence*, 54, 3222–3244. https://doi.org/10.1007/s10489-024-05319-y

[3] Du, Y., Hou, Z., Lin, E., Li, X., Liang, J., and Zhou, X. (2025). PRG-Net: Point Relationship-Guided Network for 3D human action recognition. *Neurocomputing*, 635, 130015. https://doi.org/10.1016/j.neucom.2025.130015

[4] Li, W. (2026). *MSR Action3D Dataset.* https://sites.google.com/view/wanqingli/data-sets/msr-action3d, (Accessed: 2026-02-30).

[5] Li, X., Huang, Q., Wang, Z., Yang, T., Hou, Z., and Miao, Z. (2023). Real-Time 3-D Human Action Recognition Based on Hyperpoint Sequence. *IEEE Transactions on Industrial Informatics*, 19(8), 8933–8942. https://doi.org/10.1109/TII.2022.3223225

[6] Li, W., Zhang, Z., and Liu, Z. (2010). Action recognition based on a bag of 3D points. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 9–14. https://doi.org/10.1109/CVPRW.2010.5543273

[7] Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., and Liu, J. (2023). Human Action Recognition From Various Data Modalities: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3), 3200–3225. https://doi.org/10.1109/TPAMI.2022.3183112

[8] Vemulapalli, R., Arrate, F., and Chellappa, R. (2014). Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 588–595. https://doi.org/10.1109/CVPR.2014.82

[9] Wang, C., Wang, Y., and Yuille, A. L. (2016). Mining 3D Key-Pose-Motifs for Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2639–2647. https://doi.org/10.1109/CVPR.2016.289

[10] Xin, C., Kim, S., Cho, Y., and Park, K. S. (2024). Enhancing Human Action Recognition with 3D Skeleton Data: A Comprehensive Study of Deep Learning and Data Augmentation. *Electronics*, 13, 747. https://doi.org/10.3390/electronics13040747

A New Methodological Proposal: Application of Hybrid Digital Twins to NOx Prediction using Neural Networks

Marcos Llamazares [♭],[1] Juan Carlos Cortés[♭], Rafael-J. Villanueva[♭], and Carlos Guardiola[†]
(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.
(†) Máquinas y Motores Térmicos, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction and State of the Art

The reduction of nitrogen oxides (NOx) remains one of the most critical challenges in the automotive industry, especially under the increasingly stringent regulatory framework of Euro 6d and the forthcoming Euro 7 standards. Selective Catalytic Reduction (SCR) systems have established themselves as the leading technology for neutralizing these pollutants in diesel engines. However, ensuring compliance in real-world driving conditions—beyond controlled laboratory cycles—requires a paradigm shift in how we monitor and predict tailpipe emissions. The emergence of On-Board Monitoring (OBM) demands models that are not only accurate in real-time but also robust against the inherent variability of road traffic, ambient conditions, and driver behavior.

Traditionally, SCR modeling has relied on physico-chemical frameworks based on systems of Ordinary Differential Equations (ODEs) that describe the kinetics of adsorption, desorption, and reduction of NH_3. While these models provide high interpretability and adhere to mass balance principles, they often exhibit systematic discrepancies. These errors are particularly evident during cold starts, thermal transients, or catalyst saturation phases, where the simplified physics of the model cannot capture the high-dimensional non-linearities of the system.

Our methodology for addressing this problem has evolved systematically across our prior research. Initially, we explored Grammatical Evolution (GE) as a standalone data-driven technique [1]. However, this approach relied heavily on extensive data preprocessing and the manual isolation of short time-windows under highly specific operational conditions. To improve robustness and incorporate physical constraints, our subsequent work integrated GE within a Hybrid Digital Twin (HDT) framework [2]. The objective was to maintain the physical base while learning an explicit analytical function to correct the residual error. While the use of GE provided the significant advantage of Explainable AI (XAI) through auditable symbolic expressions, it still presented inherent limitations when applied to complete, raw experimental trips. Real-world datasets—characterized by sensor noise, signal delays, and massive data volumes (hundreds of thousands of points per trip)—pose a fundamental scalability challenge for symbolic regression techniques. They struggle to generalize across a wide spectrum of operating modes without leading to overly complex and

[1] marllalo@upvnet.upv.es

uninterpretable equations. Consequently, to overcome these bottlenecks and effectively model full-scale experimental runs, our current methodology evolves by replacing GE with Neural Networks within the HDT architecture.

2 Experimental Setup and Dataset

The experimental data for this study were collected using a mild-hybrid Euro 6d-Temp diesel passenger car, equipped with a 2.0 L turbocharged engine. The exhaust after-treatment system includes a Selective Catalytic Reduction (SCR) unit for Nitrogen Oxides emission reduction.

The database comprises 66 everyday trips under diverse ambient conditions and real-world driving styles. The dataset was partitioned into training (46), validation (10) and test (10) sets. Unlike previous studies limited to short, curated sequences, this work utilizes full-trip data, encompassing thousands of time-steps.

While the analysis is based on continuous full-trip data, specific experiments were filtered out of the dataset if their behavior was predominantly driven by engine start-up transients or SCR saturation, to avoid skewing the evaluation metrics with non-representative dynamics.

Data synchronization was performed at a high-resolution sampling rate, with data attained every 10 ms. To develop the Hybrid Digital Twin, a specific set of variables was selected as inputs. Engine speed (X_1), fuel mass flow (X_2), boost pressure (X_3), intake air temperature (X_4), EGR flow (X_5), SCR inlet temperature (X_6), exhaust flow (X_7), AdBlue dosing (X_8), and load torque (X_9).

3 Methodology

3.1 Physico-Chemical Base Model

The foundation of the system is a control-oriented physico-chemical model describing the SCR dynamics. It consists of a system of 11 Ordinary Differential Equations (ODEs) and 14 calibration parameters that account for NH_3 storage, oxidation, and the standard, fast, and NO_2 reduction reactions. The model parameters were calibrated using an asynchronous random Particle Swarm Optimization (arPSO) algorithm, aimed at minimizing the discrepancy between predicted and measured tailpipe NOx under nominal conditions. Specifically, this refers to the optimal physical (temperature, pressure, air mass flow) and chemical conditions required for engine and SCR operation. This model serves as the baseline for the Hybrid Digital Twin methodology.

3.2 Hybrid Digital Twin Framework

To enhance predictive accuracy without increasing the complexity of the ODE system, we adopt an HDT formulation [3]. The final output $y(t)$ is defined as the superposition of the physics-based mathematical model prediction $y_{mm}(t)$ and a data-driven correction term $E(t)$, which captures the unmodeled residuals:

$$y(t) = y_{mm}(t) + E(t) \tag{1.1}$$

In this work, $E(t)$ is modeled using a deep feedforward neural network that leverages temporal lag expansion to capture the complex transient dynamics of the physical system.

3.3 Feedforward Neural Network for Error Prediction

The core of the data-driven correction is a feedforward deep neural network designed to estimate the scalar residual $\hat{\varepsilon}$ (in ppm), which represents the bias between the mechanistic SCR simulator (S_1) and the actual measured NOx concentration. The complete architecture is illustrated in Figure 1.

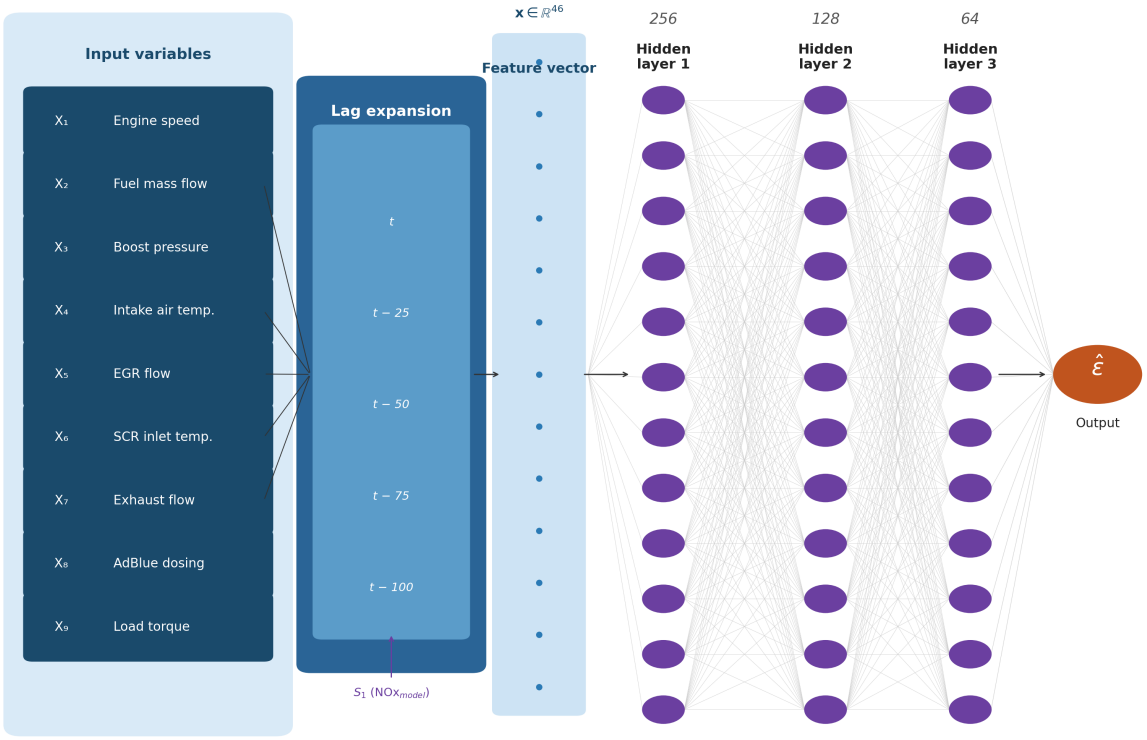


Figure 1: Feedforward neural network architecture with temporal lag expansion designed for SCR model error prediction.

The model takes nine primary process variables defined in Section 2. These signals are smoothed using a 500-sample moving average to mitigate high-frequency sensor noise. For each of these nine variables, the network receives the value at the current time step t along with four temporal lags at $\tau \in \{25, 50, 75, 100\}$ samples. This expansion yields 45 temporal features, which, combined with the real-time mathematical model simulator output S_1, forms a comprehensive 46-dimensional feature vector ($x \in \mathbb{R}^{46}$). All input features are subsequently Interquartile Range (IQR) normalized before being fed into the network.

The feature vector is processed through three fully connected hidden layers, comprising 256, 128, and 64 neurons, respectively. Each hidden layer utilizes a Rectified Linear Unit (ReLU) activation function.

Once the residual $\hat{\varepsilon}$ is predicted, the final corrected emission estimate, $\hat{S}_1$, is computed enforcing the physical constraint of non-negative emissions:

$$\hat{S}_1 = \max(0,\ S_1 - \hat{\varepsilon}) \tag{1.2}$$

4 Results and Discussion

The proposed HDT framework was evaluated across all 66 experimental trips, covering training (46), validation (10), and test (10) sets. Results are presented both qualitatively, through representative time-series comparisons, and quantitatively, through per-experiment RMSE analysis.

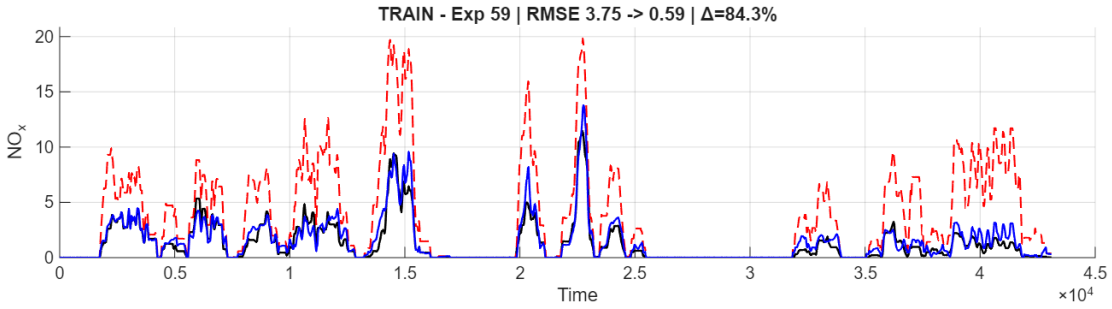


Figure 2: Representative time series from the training set. The Hybrid model (blue) closely tracks the experimental measurements (black), substantially reducing the systematic overestimation of the base physico-chemical model (red dashed).

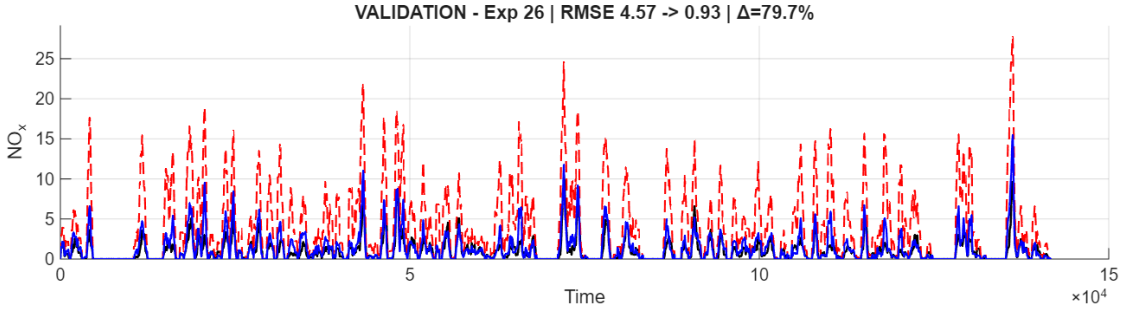


Figure 3: Representative time series from the validation set.

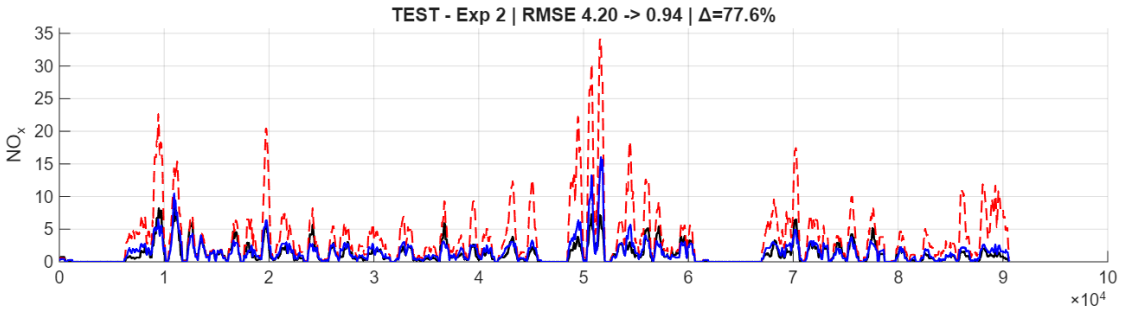


Figure 4: Representative time series from the test set.

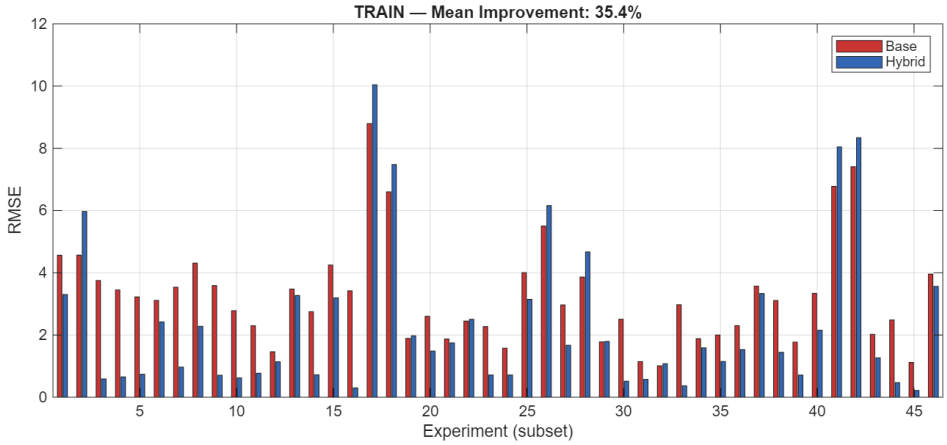
Figures 2–4 show representative trips from each data partition. In all cases, the base physico-chemical model exhibits a characteristic systematic overestimation of NOx concentration, particularly during high-load transients and warm engine phases. The feedforward neural network correction effectively suppresses these residuals, aligning the HDT output closely with the measured signal.

Figure 5 displays the per-experiment RMSE for both the base physico-chemical model and the proposed HDT. While the HDT achieves substantial reductions across the vast majority of experiments in all three data partitions, it is worth noting that a minority of cases—most visibly in the test set (e.g., Experiments 2 and 5 in Figure 5c)—show limited improvement or even marginal degradation. These outlier cases correspond to trips with atypical dynamics that fall outside the nominal operating conditions assumed in this study. This behavior is attributed to the presence of different engine operating modes within the same experiment, a common occurrence given the inherent difficulty of isolating individual operating regimes in real-world testing. This limitation is consistent with the nature of a single global network and motivates future work on regime-aware correction strategies. Finally, the aggregate performance is summarized in Table 1.1.

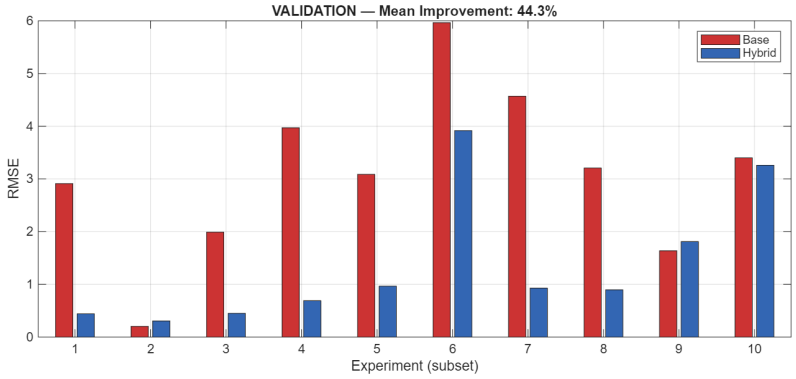
Set	$RMSE_{\text{base}}$	$RMSE_{\text{HDT}}$	Improvement (%)
Train	3.263	2.350	35.4
Validation	3.094	1.367	44.3
Test	3.993	2.479	44.7

Table 1.1: Average RMSE (ppm) comparison between the baseline physico-chemical model and the proposed HDT across all data partitions.

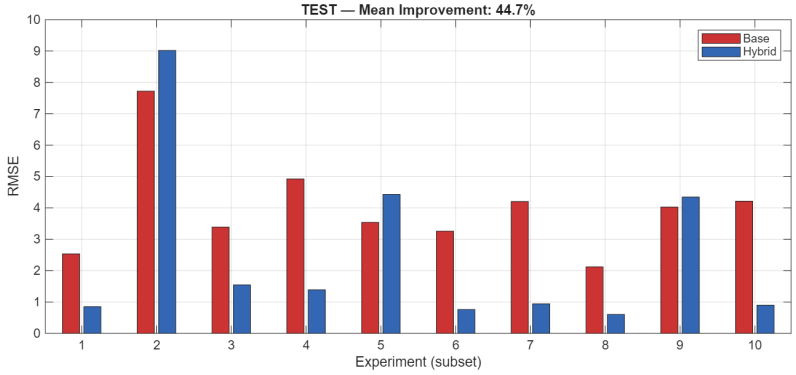
The HDT consistently reduces RMSE by approximately 35–45% relative to the base model across all sets. The comparatively lower improvement observed on the training set is not indicative of underfitting, but rather reflects the heterogeneity inherent to real-world driving data. Since trip assignment to each partition was performed randomly, the training set happened to encompass a higher proportion of experiments with challenging dynamics that are intrinsically harder to correct regardless of model capacity. The improvements on validation and test sets (44.3% and 44.7%, respectively) confirm that the learned correction generalizes effectively to unseen trips.



(a) Train (mean: −35.4%)



(b) Validation (mean: −44.3%)



(c) Test (mean: −44.7%)

Figure 5: Per-experiment RMSE comparison between the base model and the proposed HDT across training, validation, and test sets.

5 Conclusions

This work has presented an evolution of the Hybrid Digital Twin (HDT) framework for NOx prediction, replacing the previously employed Grammatical Evolution with a feedforward neural network trained on temporal lag-expanded features. The results are promising: a relatively simple neural network architecture achieves substantial improvements over the calibrated physico-chemical model across all data partitions, with average RMSE reductions of 35.4%, 44.3%, and 44.7% on

training, validation, and test sets, respectively. Notably, the proposed HDT attains a test accuracy comparable to prior window-based GE models while operating on complete real-world trips.

Despite these encouraging results, a subset of experiments reveals cases where the neural network correction fails to improve or marginally degrades the base model prediction. These cases are predominantly associated with specific engine operating modes, such as cold-start transients or SCR saturation. Future work will therefore focus on extending the methodology to explicitly account for different engine operating conditions, developing mode-aware correction strategies capable of handling the full diversity of real-world driving scenarios and addressing the limitations identified in this study.

References

[1] Llamazares López, M., Parra, D., Velasco Cabo, J.M., Garnica, Ó., Villanueva Micó, R.J., and Hidalgo, J.I. *Unveiling the dynamics of NOx pollution in internal combustion engines by Structured Grammatical Evolution*. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '25)*, pages 1388–1396. ACM, 2025. https://doi.org/10.1145/3712256.3726421

[2] Llamazares, M., Cortés López, J.C., Villanueva, R.J., and Guardiola, C. *Hybrid Digital Twins Enhanced with Evolutionary Grammars for Real-World Nitrogen Oxides Emission Prediction in Euro 6 Hybrid Diesel Engines*. SSRN Preprint, 2026. http://dx.doi.org/10.2139/ssrn.6568981

[3] Chinesta, F., Cueto, E., Abisset-Chavanne, E., Duval, J.L., and El Khaldi, F. *Virtual, Digital and Hybrid Twins: A New Paradigm in Data-Based Engineering and Engineered Data*. *Archives of Computational Methods in Engineering*, 2018. https://doi.org/10.1007/s11831-018-9301-4

Analisys of Gabor Filters for Defect Detection in Raw Fabric Images

T. Albero-Albero♯ B. Cantó♭ N. Garrido♭ and R. Pérez♯,[1]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

(♯) I.U. Mixto de Tecnología Informática, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

Textile defect detection can be formulated as a structured anomaly detection problem on quasi-periodic visual patterns, where the objective is to identify statistically relevant deviations from the natural variability of woven textures, particularly under non-ideal acquisition conditions. Traditional image processing techniques have been applied to textile defect detection. Methods such as edge detection, thresholding, and morphological operations can identify defects including missing threads, stains and holes [1]. Among these methods, Gabor filters, defined as Gaussian-modulated sinusoids, provide a well-established framework for textile inspection because they capture spatial and frequency information in a way that is similar to human visual perception [2].

Previous work has shown that Gabor-based methods can be efficiently optimized for ARM (Advanced RISC Machines) architectures. Under controlled conditions, a reduced set of carefully tuned filters combined with thresholding and morphological processing achieves reliable performance. In particular, high sensitivity has been reported on the AFID dataset, which features stable illumination and limited variability [3].

However, the applicability of these models strongly depends on the validity of their assumptions. Fabrics such as cotton exhibit pronounced natural variability in yarn thickness, density and reflectance, resulting in non-stationary textures whose spectral content cannot be adequately captured by a fixed filter bank. In these cases, filter responses associated with normal fabric variability may overlap with those produced by true defects. Consequently, detection performance becomes highly sensitive to parameter selection, and classical pipelines are forced into an unfavorable trade-off between precision (false positive suppression) and recall (defect sensitivity).

This limitation becomes particularly evident when working with photometrically raw images, where illumination gradients and contrast variations are not artificially compensated. Under such conditions, threshold-based decisions and morphological operations act more as heuristic corrections than as principled solutions, often transferring error from one stage of the pipeline to another.

In this work, we investigate the limitations of classical Gabor-based defect detection under realistic industrial conditions. We evaluate the performance of the RAW-FABRID dataset, which contains cotton samples acquired from different origins with different characteristics. We hypothesize that classical filtering approaches reach a performance ceiling in such environments due to

[1] ruperez@disca.upv.es

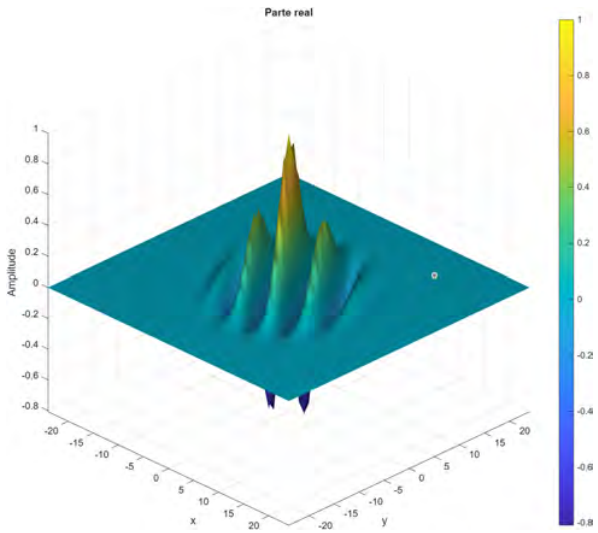


Figure 1: Gabor filter representation

their limited modeling capacity.

2 Methodology

The proposed system is based on a classical image processing pipeline using two-dimensional Gabor filters for texture analysis. Gabor filters are particularly well suited to woven textiles because they combine spatial localization with frequency selectivity, which makes them effective for capturing the directional patterns generated by periodic yarn structures. A Gabor filter consists of a Gaussian kernel modulated by a sinusoidal function. This allows the filter to respond to texture patterns at a specific scale and orientation. Figure 1 shows a three-dimensional representation of a Gabor filter, and its analytical expression is given by:.

$$g(x,y) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda} + \psi\right),$$

where $x' = x\cos\theta + y\sin\theta$ and $y' = -x\sin\theta + y\cos\theta$. The parameter λ represents the wavelength of the sinusoidal component and defines the spatial frequency to which the filter is tuned. In textile inspection, this parameter must be adjusted according to the average distance between consecutive warp or weft yarns. The parameter σ denotes the standard deviation of the Gaussian envelope and controls the size of the receptive field. The orientation θ defines the directional selectivity of the filter, while the phase offset ϕ and the spatial aspect ratio γ determine the symmetry and geometric elongation of the kernel. The algorithm is implemented in $C++$ using OpenCV and adapted to industrial constraints. Texture information is extracted from both the real (cosine) and imaginary (sine) components of the filter response. In particular, the detection is based on three complementary technical strategies. First, we use a double spatial verification, when a candidate contour is validated only if a consistent response is observed in both the real (cosine) and imaginary (sine) phases of the Gabor filtering. Second, we adopt a column based thresholding strategy to mitigate illumination gradients caused by the physical limitations of LED bars over fabric widths of approximately 50 cm. The image is processed in vertical strips of fixed width (255 pixels), allowing threshold parameters to adapt locally to illumination variations without modifying the raw intensity distribution. Finally, we apply an 11×11 elliptical kernel to reconstruct defects such

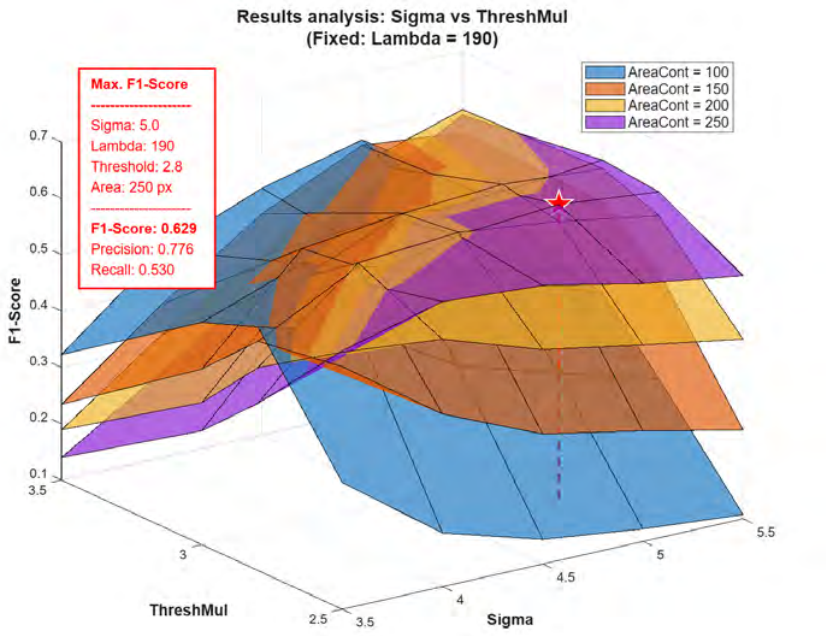


Figure 2: Gabor filter battery tests results

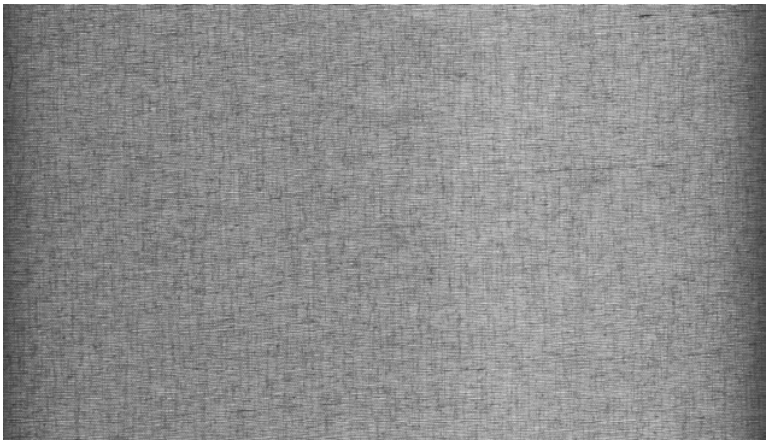
as broken threads that may appear fragmented after filtering and thresholding.

In contrast to the tests given in [4], where 245 images of size 4096×256 pixels were acquired under highly controlled and homogeneous conditions, the present study uses the RAW-FABRID dataset [5] which includes 709 high-resolution images of size 1792×1024 pixels, acquired using a line-scan camera at a spatial resolution of 4 pixels per millimeter from 7 different manufacturers to ensure robustness. This introduce variability in texture, thread density, and weave pattern. All images are photometrically raw, with no preprocessing applied.

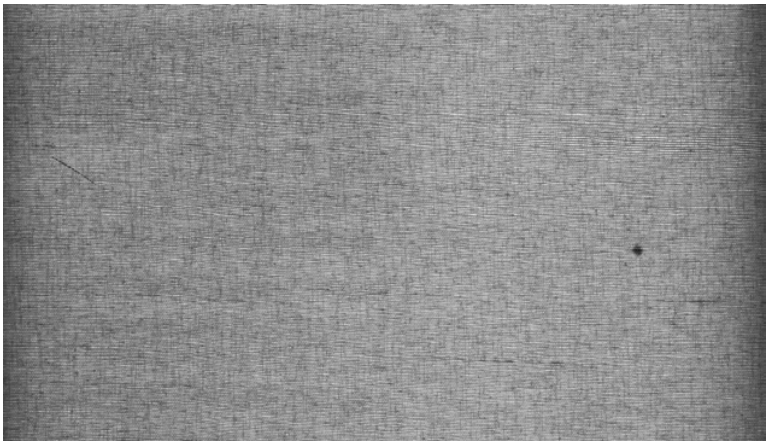
For the evaluation of the optimized $C++$ Gabor-based algorithm, we use fully images with the original width of 1792 pixels in the dataset, which are cropped from 2024 raw camera acquisitions to avoid lateral shadow effects typically present near fabric edges in the demonstrator machine used for acquisition, and allow the analysis to better reflect realistic industrial inspection conditions.

Empirical evaluation on cotton fabrics led to the identification of an optimal parameter configuration, referred to as the best configuration. The analysis included Gaussian standard deviation values σ ranging from 3.5 to 5.5, wavelength values λ between 175 and 220, threshold multipliers from 2.5 to 3.5, and minimum contour area filters ranging from 100 to 250 pixels. The kernel size was fixed at 23 pixels throughout all experiments. The best F1-Score was obtained using $\sigma = 5.0$, a wavelength of $\lambda = 190$, a threshold multiplier of 2.8, and a minimum contour area of 250 pixels. The complete parameter exploration and corresponding performance results are summarized in Figure 2.

A sample of RAW-FABRID images with a defect-free image, defective samples, and their corresponding pixel-level ground truth masks are presented in Figure 3. These examples illustrate the variability in fabric textures as well as the pixel-level annotation quality provided in that dataset.



(a) Defect-free fabric image



(b) Fabric image with defect

(c) Corresponding binary ground truth mask

Figure 3: Representative high-resolution images from the RAW-FABRID dataset: defect-free fabric, defective fabric sample, and corresponding pixel-level ground truth mask.

3 Results and discussion

Initial experiments on the AFID dataset show stable performance under controlled conditions, achieving a Sensitivity (Detection Rate) of 86.6%, a Precision of 84.5%, and an F1-Score of 85.5%. These results confirm that Gabor-based methods perform well when assumptions about texture regularity and illumination hold. However, when the same configuration is applied to newly captured real-time images after fine adjustment of the best Gabor parameters, the Sensitivity decreased to 53.0%, a Precision of 77.6% and an F1-Score of 62.9% due to variability in fabrics.

A more demanding evaluation was then performed using the RAW-FABRID dataset, due to variability in the manufacturers of fabrics. In contrast to AFID, this dataset lacks that variability in texture geometry, yarn density, and illumination, providing a poor realistic benchmark for more different real industrial fabrics.

Several intermediate approaches have been evaluated. The use of contrast enhancement techniques such as CLAHE significantly increase recall but introduce excessive noise, resulting in a large number of false detections. Similarly, geometric filtering based on aspect ratio removes noise but also eliminates valid defects.

Phases 1 and 2 focus on threshold adjustments. Initial configurations prioritized precision, achieving high precision values (0.84) but an unacceptably low recall (0.32). Subsequent use of more aggressive thresholds increased recall, but caused False Positives (FP), crashing the F1-Score to drop to 0.28.

CLAHE and Geometry Filters are also tested. These failed enhancements attempt to fix lighting gradients using Contrast Limited Adaptive Histogram Equalization (CLAHE), but failed completely. Although recall improved to 0.67, the enhancement strongly amplified the natural noise of the normal warp and weft grid, generating more than 1700 False Positives. Additionally, implementing a geometry filter based on aspect ratio (greater than 3.0) to eliminate circular noise erroneously deleted real diagonal and curved defects, as their bounding boxes form almost perfect squares.

The optimal solution for the final algorithm required adapting the mathematical processing to the physical characteristics of the loom. The final architecture implemented a Double Spatial Verification, where a contour is only accepted if it resonates in both the real (cosine) and imaginary (sine) phases of the filter. To combat illumination gradients without using CLAHE, cylindrical thresholding was applied by processing the image in vertical strips of 255 pixels. Finally, an 11×11 elliptical morphological dilation was used to fuse valid fragments and reconstruct broken threads.

Thus, the best performance has been achieved using this algorithm configuration, which combined double spatial verification, cylindrical thresholding, and morphological reconstruction. You can see the block diagram of the complete algorithm in Figure 4. The data confirmed that the minimum area constraint is the primary lever for eliminating stochastic noise.

A grid search analysis reveals that the minimum contour area is the most influential parameter for reducing false positives. The best configuration is obtained using a standard deviation $\sigma = 5.0$, a wavelength $\lambda = 190$, a threshold multiplier of 2.8, and a minimum area of 250 pixels. Under this configuration, the system has detected 287 True Positives, 83 False Positives, and 255 False Negatives, yielding a maximum F1-Score of 62.9.

These results highlight three main limitations of classical approaches. First, the method is highly sensitive to intrinsic fabric variability, which leads to confusion between normal texture and defects. Second, geometric filtering lacks semantic understanding, which does not lead to distinguish between benign texture irregularities and structurally critical defects. Third, illumination gradients remain a critical source of instability, and the fundamentally static nature of Gabor filters remains highly vulnerable to uncorrected illumination variability.

Overall, these results indicate that the observed performance ceiling reflects intrinsic limitations

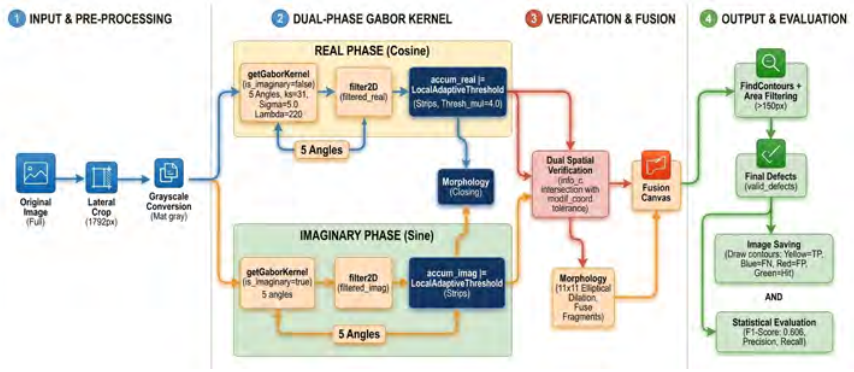
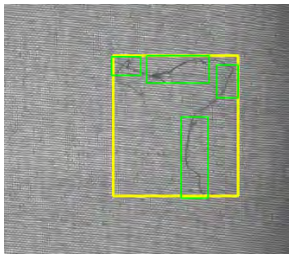
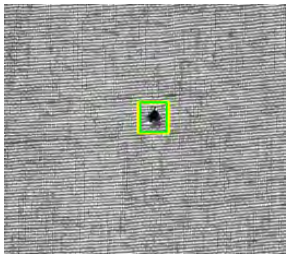


Figure 4: Algorithm block diagram

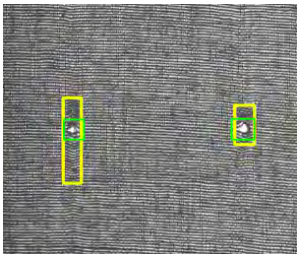
of analytically defined filters when applied to highly variable real-world data. Figure 5 presents representative examples of detected true defects (first row in green/yellow), not detected defects (second row in blue) and false detected defects (third row in red).



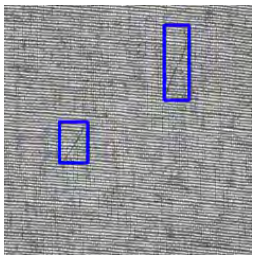
(a) True positive



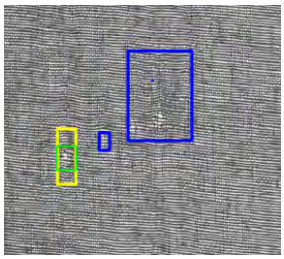
(b) True positive



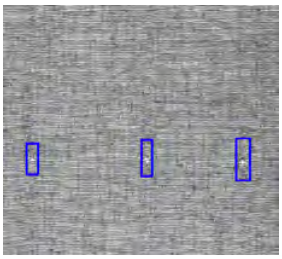
(c) True positive



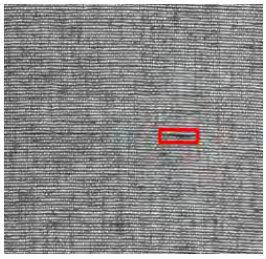
(d) False negative



(e) False negative

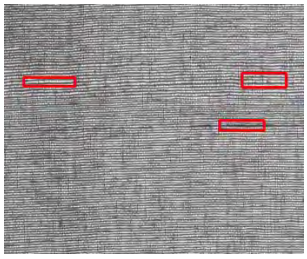


(f) False negative



(g) False positive

(h) False positive



(i) False positive

Figure 5: Samples of the resulting detection algorithm.

4 Conclusions and future work

This work shows that the natural variability of textile materials, together with variation in illumination conditions due to the angle from yards to the camera impose fundamental limitations on traditional image processing approaches such as Gabor filtering.

The experimental analysis demonstrates that Gabor filters optimized for embedded and ARM -based architectures provide cost-effective and reliable defect detection when acquisition conditions are stable, as observed on the AFID dataset. However, when applied to the highly variable RAW-FABRID cotton samples, performance degrades significantly. Despite the incorporation of advanced engineering strategies—including double spatial verification, cylindrical thresholding, and morphological reconstruction—the best achievable performance was limited to an F1-Score of 0.608. The results reveal a pronounced sensitivity to intrinsic fabric noise and raw illumination gradients, highlighting a clear performance ceiling for analytically predefined filters in unconstrained industrial scenarios.

Convolutional Neural Networks (CNNs) provide a more flexible framework by learning hierarchical feature representations directly from data. This enables improved robustness to variability without manual parameter tuning.

Future work will focus on leveraging the RAW-FABRID dataset to develop and evaluate CNN-based models for textile defect detection, with the goal of overcoming the limitations identified in this study.

References

[1] N. Ingale, A. Ansari, S. Yerunkar, S. Thakur, and S. Sayyed, "Fabric defect detection," *Int. J. of Creative Research Thoughts*, vol. 11(3), pp. 1–5, 2023. [Online]. Available: `https://api.semanticscholar.org/CorpusID:261956391`.

[2] A. Bodnarova, Bennamoun, and S. J. Latham, "Optimal gabor filters for textile flaw detection," *Pattern Recognit.*, vol. 35, pp. 2973–2991, 2002. [Online]. Available: `https://api.semanticscholar.org/CorpusID:25740107`.

[3] B. Canto-Colomina, N. Garrido-Sanz, R. Perez-Llorensa, and J. Silvestre-Blanes, "Optimization of defect detection and classification in textiles using arm architectures," *Proceedings MME-HB*, vol. 1, pp. 93–98, 2025.

[4] J. Silvestre-Blanes, T. Albero-Albero, I. Miralles, R. Perez-Llorens, and J. Moreno, "A public fabric database for defect detection methods and results," *Autex Research Journal*, vol. 19, pp. 363–374, 2019. [Online]. Available: `https://api.semanticscholar.org/CorpusID:198402250`.

[5] R. Perez-Llorens, T. Albero-Albero, and J. Silvestre-Blanes, "A dataset of raw fabric grayscale images for defect detection," *Preprints*, Apr. 2026. DOI: `10.20944/preprints202604.1335.v1`. [Online]. Available: `https://doi.org/10.20944/preprints202604.1335.v1`.

Development of a data-driven PINN to solve the neutron diffusion equation in nuclear reactors

E. Navarro-Gamón[♭],[1] A. Vidal-Ferràndiz[♮], D. Ginestar[♮], A. Carreño[♭] and G. Verdú[♭]

(♭) I.U. de Seguridad Industrial, Radiofísica y Medioambiental, Universitat Politècnica de València, Camí de Vera s/n, València, Spain.

(♮) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València, Camí de Vera s/n, València, Spain.

1 Introduction

The analysis and simulation of neutron behavior within a nuclear reactor core are fundamental pillars to guarantee safety and optimize the design of these facilities [1], [2]. To guarantee the safety, the continuous development and updating of numerical calculation codes that can predict these phenomena accurately and adapt to complex geometries is essential [2]. Classic deterministic methods, such as finite differences or nodal methods, have reached a high degree of maturity; however, they usually require an intense computational cost and significant efforts in mesh discretization when the dimensionality of the problem increases [2], [3].

In recent years, the rise of deep learning techniques has revolutionized the field of scientific computing [2]. Machine learning offers a promising alternative approach due to its ability to act as a universal function approximator without rigidly depending on meshes [2], [3]. Nevertheless, purely data-driven methods suffer from a major shortcoming in nuclear engineering: the cost of acquiring high-fidelity experimental or simulation data is usually prohibitive [3].

To mitigate this data dependency, the paradigm of Physics-Informed Neural Networks (PINNs) has emerged [4]. These networks not only learn from the scarce available data but directly integrate physical laws, in the form of partial differential equations, into the loss function of the network, evaluating the residuals through automatic differentiation [2], [3]. Since the initial formulation of neural network methodologies to solve differential equations by Lagaris et al. [5] and the subsequent generalization of PINNs by Raissi et al. [4], multiple variants have been developed to solve eigenvalue problems, quantum mechanics, and fluid dynamics [6], [7].

This document is organized as follows. Section 2 details the methods used, including the mathematical foundations of the diffusion equation, the generation of reference solutions via centered finite differences, and the neural network training design. Section 3 discusses the numerical results. Finally, section 4 summarizes the main conclusions of this study.

2 The neutron diffusion eigenvalue problem

The neutron diffusion approximation is one of the most common approximations to the Boltzmann neutron transport equations. This represents a valid mathematical description of the behavior of

[1] emnaga@upvnet.upv.es

the neutrons inside a nuclear reactor core [1]. This approximation describes a balance equation between the neutrons that enter the control volume and the neutrons that are removed. Such steady state balance equation results in an eigenvalue problem in which the eigenvector, associated with the dominant eigenvalue, is the neutron flux solution of the problem that depends on the position vector $\vec{r}$ and the group of energy g.

For a steady-state problem, the multigroup diffusion equation can be written as in equation (1),

$$-\vec{\nabla}\cdot\left(D_g\vec{\nabla}\Phi_g\left(\vec{r}\right)\right)+\Sigma_{rg}\left(\vec{r}\right)\Phi_g\left(\vec{r}\right)=\sum_{\substack{g'=1\\g'\neq g}}^{G}\Sigma_{sg'\to g}\left(\vec{r}\right)\Phi_{g'}\left(\vec{r}\right)+\frac{\chi_g\left(\vec{r}\right)}{\lambda}\sum_{g'=1}^{G}\nu_{g'}\Sigma_{fg'}\left(\vec{r}\right)\Phi_{g'}\left(\vec{r}\right), \quad (1)$$

where D_g, Σ_{rg}, $\Sigma_{sg'\to g}$ and $\Sigma_{fg'}$ are space-dependent parameters: the diffusion coefficient and the removal cross-section for the g^{th} group of energy, the scattering cross-section from the group g' to the g energy group and the fission cross-section, respectively. The space-dependent fission spectrum is χ_g and the average number of neutrons produced per fission is $\nu_{g'}$ [1]. Then, the diffusion equation can be written in a matrix form as in (2), being $\mathcal{L}$ and $\mathcal{M}$ the loss operator and the production operator, respectively, Φ is the eigenvector and λ is the eigenvalue.

$$\mathcal{L}\Phi=\frac{1}{\lambda}\mathcal{M}\Phi. \quad (2)$$

2.1 Physically Informed Neural Networks

2.2 Neural network design

In order to obtain accurate results, PINNs require a sufficient density of points to properly represent the domain geometries and minimize the physical residuals but without the intervention of a structured grid. Consequently, the design of the network must be carefully optimized.

As illustrated in Figure 1, the network takes the spatial position vector $\vec{r}_i$ as the input layer. Therefore, it natively supports multi-dimensional geometries by adapting the dimensions of the input tensor. This information is processed through h fully-connected hidden layers, each containing n neurons. Finally, the output layer delivers the continuous mapping of the scalar neutron fluxes for all the G energy groups involved in the diffusion equation. The neural network is developed and trained using the PyTorch open-source library [8]. The activation function applied to the hidden layers is the hyperbolic tangent, tanh.

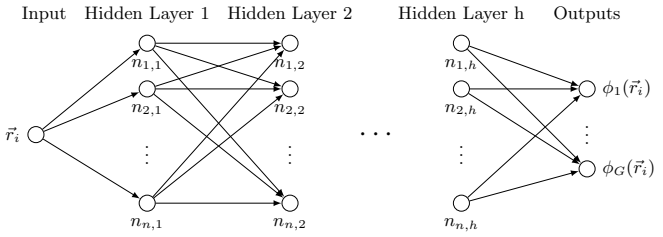


Figure 1: Generic fully-connected neural network architecture designed for multi-dimensional and multi-group problems. The network receives the spatial position vector $\vec{r}_i$ as input and maps it through h hidden layers to predict the neutron flux ϕ_g for all G energy groups.

2.3 Training design

Here is a comprehensive overview of the proposed methodology. The overall execution pipeline of the data-driven PINN is structured into five sequential stages, as illustrated in Figure 2. This modular architecture ensures a clear separation between data generation, the physics-informed training, and a traditional numerical validation.

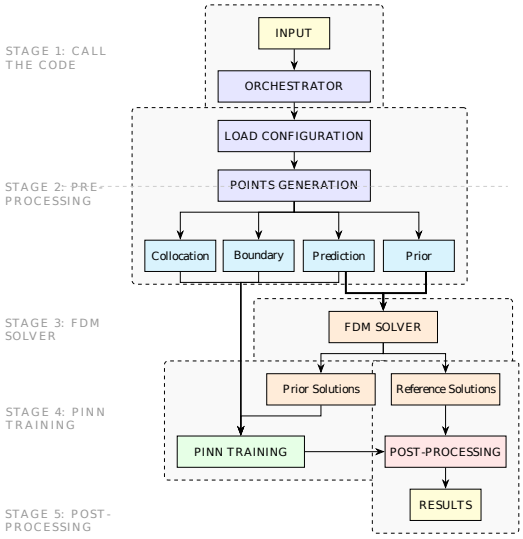
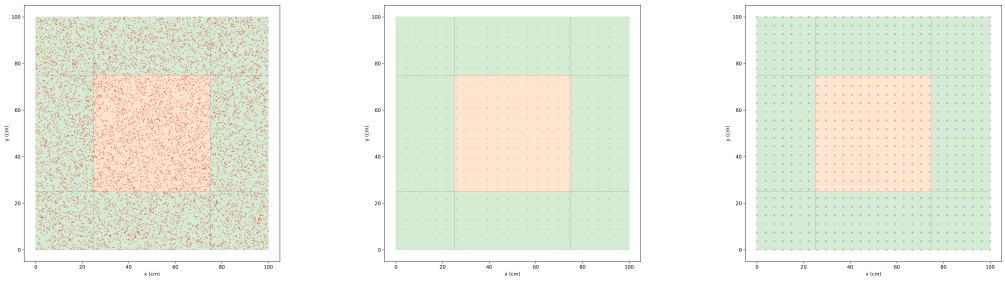


Figure 2: Modular block diagram representing the 5-stage pipeline for the data-driven PINN execution, from raw input data to final error assessment against CFDM reference solutions.

The diffusion solver begins with the master orchestrator script handling the entire workflow in-memory, managing data flow between the physics-based configurations and the deep learning training loop. The configuration file reads the input information, which contains the problem geometry, the multi-group material cross-sections and the boundary conditions. Subsequently, the spatial domain is discretized into different sets of points strategically sampled for the PINN, the *collocation points*, *boundary points*, *prior points* and *prediction points*, as seen in Figure 3.



(a) 10000 collocation points. (b) 484 prior points. (c) 809 prediction points.

Figure 3: Distribution of the training points through the 2D Heterogeneous problem.

To provide the neural network known flux solutions, a traditional Centered Finite Differences Method (CFDM) solver is executed. This stage acts as the ground-truth reference, solving the eigenvalue problem to obtain accurate scalar neutron fluxes, Φ and the eigenvalue, λ.

The PINN is trained by minimizing a multi-objective loss function. The total loss function

represents the error made by a given state of the neural network and is defined as a weighted sum of three independent terms as in equation (3). Where α, β, and γ are experimentally selected penalty weights independent for each optimizer.

$$Loss = \alpha L_r + \beta L_b + \gamma L_p, \tag{3}$$

The *collocation points* are generated using a Latin Hypercube Sampling, LHS, across the continuous domain to evaluate the residual, L_r. It evaluates the error that the model commits when satisfying the equations at the N_c generated collocation points. To construct this loss, the residual R_g for each energy group $g \in \{1, \ldots, G\}$ is evaluated, as in equation (4), where $\hat{\Phi}$ represents the continuous flux mapped by the network and λ is treated as a trainable scalar variable. Thus, the residual loss is computed as the mean of the squared residuals over all collocation points and energy groups as in equation (5).

$$R_g(\vec{r}_i) = -\vec{\nabla}\cdot\Big(D_g(\vec{r}_i)\vec{\nabla}\hat{\Phi}_g\left(\vec{r}_i\right)\Big) + \Sigma_{rg}\left(\vec{r}_i\right)\hat{\Phi}_g\left(\vec{r}_i\right) - \sum_{\substack{g'=1 \\ g'\neq g}}^{G} \Sigma_{sg'\to g}\left(\vec{r}_i\right)\hat{\Phi}_{g'}\left(\vec{r}_i\right) - \frac{\chi_g}{\lambda}\sum_{g'=1}^{G}\nu_{g'}\Sigma_{fg'}\left(\vec{r}_i\right)\hat{\Phi}_{g'}\left(\vec{r}_i\right), \tag{4}$$

$$L_r = \frac{1}{G\cdot N_c}\sum_{g=1}^{G}\sum_{i=1}^{N_c}\left|R_g(\vec{r}_i)\right|^2. \tag{5}$$

The *boundary points* are sampled on the external faces to enforce Dirichlet or Neumann conditions. The boundary loss L_b is defined as the mean square error between the prediction of the network and the physical boundary values on the N_b boundary points. For a system with Dirichlet boundary conditions it would follow equation (6).

$$L_b = \frac{1}{G\cdot N_b}\sum_{g=1}^{G}\sum_{i=1}^{N_b}\left|\hat{\Phi}_g(\vec{r}_i^{\,b})\right|^2. \tag{6}$$

To accelerate the convergence and prevent the PINN from falling into non-physical local minima or the trivial zero solution, the network is supplied with N_p *Prior points* extracted from the CFDM solver. The prior loss L_p drives the network to match these high-fidelity spatial distributions on a reduced grid as in equation (7).

$$L_p = \frac{1}{G\cdot N_p}\sum_{g=1}^{G}\sum_{i=1}^{N_p}\left|\hat{\Phi}_g(\vec{r}_i^{\,p}) - \Phi_{g,CFDM}(\vec{r}_i^{\,p})\right|^2. \tag{7}$$

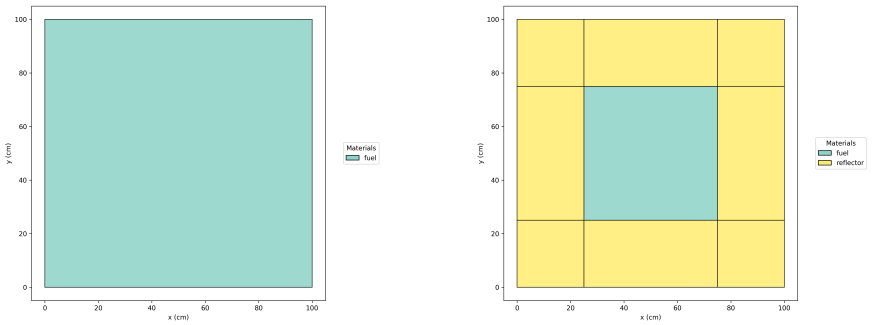
Finally, the *prediction points* are mapped to obtain the spatial distributions from de CFDM solver to evaluate the network final inferred flux calculated in this points. Once the training is completed via combined Adam and L-BFGS optimization, the execution moves to the post-process where the predicted PINN solutions in the N_{pred} points are compared against the reference solutions from the solver for the N_{pred} points to compute absolute errors and validate the accuracy of the neural network.

3 Results

3.1 Homogeneous 2-Dimensional Reactor

To assess the performance of the proposed data-driven PINN, a homogeneous 2D reactor core is initially analyzed. This problem serves as a verification case to check the ability of the network

to capture the fundamental profiles of the neutron flux in both fast ($g = 1$) and thermal ($g = 2$) energy groups. To better understand the challenges the neural network faces in each scenario, the geometry for the homogeneous core is displayed in Figure 4a. It represents a simplified continuous medium with uniform cross-sections, ideal for testing the base mapping capabilities.



(a) Homogeneous 2D reactor. (b) Heterogeneous 2D reactor.

Figure 4: Material layout, geometry and boundary conditions of the evaluated 2D reactor cores.

The PINN is provided with 300 collocation points, 200 boundary points and 25 prior solutions, with 809 prediction points. Figure 5 displays the behavior of the network during the training optimization for the homogeneous case, using 4 hidden layers and 80 neurons per layer. The left plot represents the decay of the total loss function, where a smooth, stable reduction is observed during the first 4×10^4 epochs using the Adam optimizer. Subsequently, a sharp drop in the loss is achieved when switching to the L-BFGS optimizer, reaching values below 10^{-5}, which indicates a highly accurate solution to the system of equations. In parallel, the right plot shows the convergence of the trainable parameter λ. The PINN successfully predicts a final effective multiplication factor of $\lambda = 0.99891$, which compared against the reference solution represents an error of 308.49 pcm.

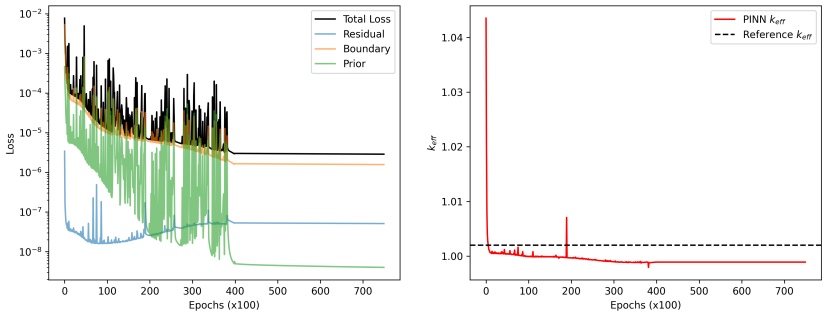


Figure 5: Training loss convergence and the eigenvalue ($\lambda = k_{eff}$) evolution for the 2D homogeneous reactor case. On the left plot, the vertical dashed line indicates the transition from the Adam optimizer to the L-BFGS fine-tuning.

The predicted physical distributions are illustrated in Figure 6. As expected for the homogeneous core with zero flux boundary conditions, both energy groups exhibit a smooth, symmetric sine-like distribution peaking at the center of the core. The fast flux presents higher magnitude values compared to the thermal flux, honoring the physical scale of the neutron population. A detailed inspection of the local absolute error distributions depicted in the bottom row of Figure 6 reveals that the discrepancies are not uniformly distributed. Instead, the highest residuals are con-

centrated at the boundaries of the core. This behavior is typical in PINN training, as the network tends to produce larger absolute deviations in regions with higher discontinuities. Quantitatively, the maximum point-wise absolute error remains below 0.03. A summary of the loss convergence and eigenvalue and eigenvector error errors is collected in Table 1.

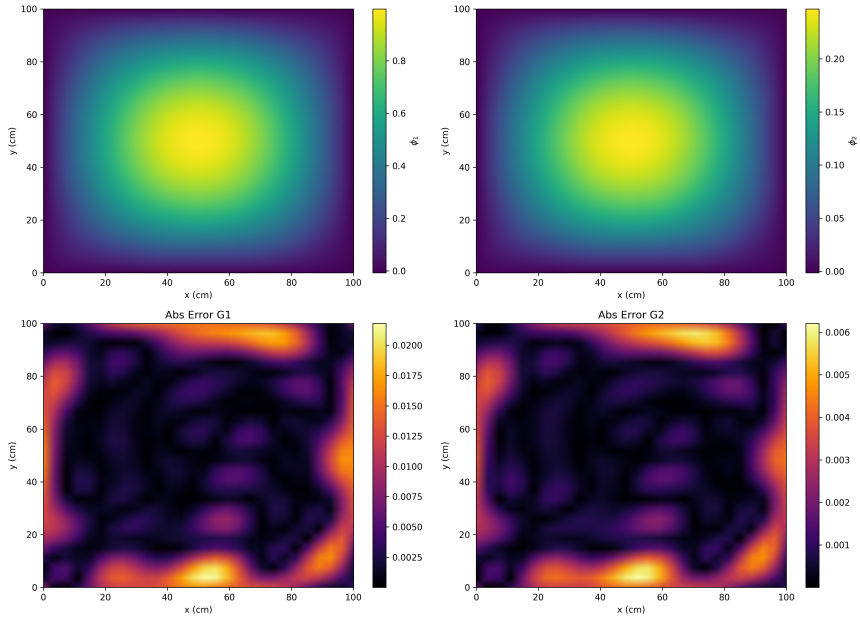


Figure 6: Spatial distribution of the neutron flux for Fast (left) and Thermal (right) groups. Below, a comparison between the CFDM reference solutions and the PINN predictions.

To further quantify the local accuracy, the global statistical errors between the CFDM and PINN mappings are evaluated using the Maximum Absolute Error (MAPE) and the Root Mean Square Error (RMSE) defined in equation below. Where ϕ_{ref} is the reference value of the neutron flux and ϕ_{pred} is the predicted neutron flux. Resulting for the homogeneous problem in a MAPE of 0.0218 and a RMSE of 0.0070 for the fast group, and a MAPE of 0.0062 with an RMSE of 0.0019 for the thermal group. Also the eigenvalue error is evaluated with the absolute difference between the reference eigenvalue, λ_{ref}, obtained using CFDM and the predicted eigenvalue, λ_{pred}.

$$\Delta\lambda = 10^5 \left|\lambda_{\text{ref}} - \lambda_{\text{pred}}\right| \quad \text{(pcm)}$$

$$\text{MAPE} = \max \left|\phi_{ref} - \phi_{pred}\right|$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{Reference Value}_i - \text{Predicted Value}_i)^2}$$

Table 1: Loss convergence, λ predictions, and statistical errors.

Reactor Case	**Final Loss**	**Pred. λ**	**$\Delta\lambda$ (pcm)**	**Group**	**MAPE**	**RMSE**
Homogeneous 2D	$< 10^{-5}$	0.99891	308.49	Fast	0.0218	0.0070
				Thermal	0.0062	0.0019
Heterogeneous 2D	$< 10^{-5}$	0.94172	710.57	Fast	0.0050	0.0013
				Thermal	0.1835	0.0198

3.2 Heterogeneous 2-Dimensional Reactor

To evaluate the PINN under more challenging conditions, a heterogeneous 2D reactor with material discontinuities is analyzed. Figure 4b illustrates the core layout for a heterogeneous core composed of fuel and reflector. This configuration generates sharp transitions in the cross-sections which introduce important discontinuities that the PINN must learn to map.

A summary of the loss convergence, λ performance, and statistical errors for both configurations of 2D reactors is collected in Table 1. In both scenarios, the behavior of the training optimization exhibited a stable reduction during the first 4×10^4 epochs using the Adam optimizer. Subsequently, a sharp drop in the total loss was achieved when switching to the L-BFGS optimizer, reaching final values below 10^{-5}. Regarding the global neutron balance, the PINN successfully predicts a λ of 0.94172 for the heterogeneous core. When compared against their respective reference solutions, these predictions yield errors of 710.57 pcm for the eigenvalue. Finally, the values recorded for both the MAPE and RMSE confirm that adding coarse prior solutions efficiently guides the model toward the physical global minimum, maintaining high local precision and avoiding trivial zero solutions even in complex multi-material domains.

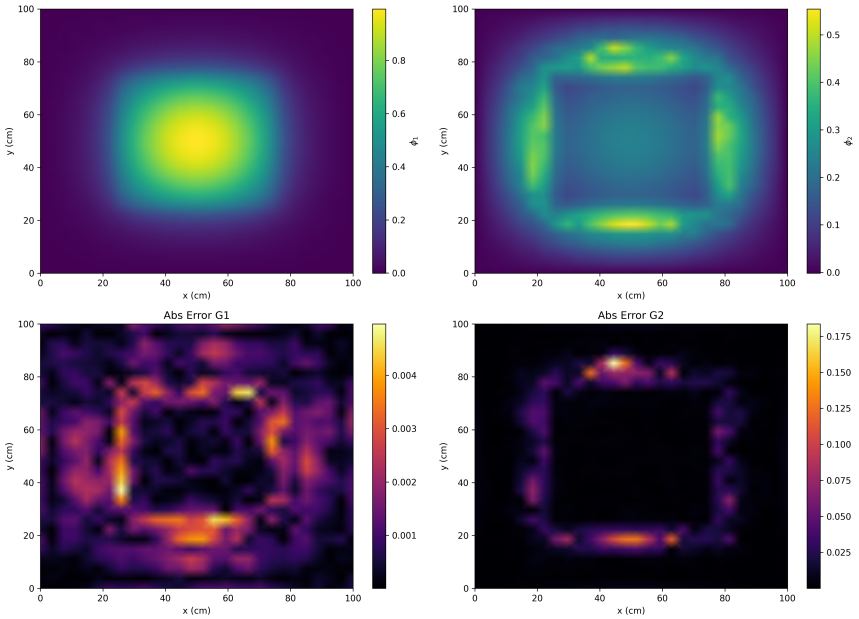


Figure 7: Spatial distribution of the neutron flux for Fast (left) and Thermal (right) groups in the heterogeneous core.

Figure 7 presents the spatial distribution of the neutron fluxes. The heterogeneous core induces local flux depressions and peaks at the material interfaces, especially visible in the thermal group due to the moderation effect. An examination of the local absolute error distributions located in the bottom row of Figure 7, reveals a distinct behavior compared to the homogeneous problem. Instead of concentrating at the center, the highest residuals exhibit a grid-like pattern aligned with the material interfaces and assembly corners. This is expected, as the sharp discontinuities in the cross-sections induce steep gradients in the continuous neural network mapping that are harder to replicate than smooth profiles. Quantitatively, absolute error for the fast group peaks at around 0.005. In contrast, the thermal group, shows maximum local errors of 0.175. Despite these localized peaks at the boundaries, the overall precision remains high across the domain.

4 Conclusions

In this work, a data-driven Physics-Informed Neural Network (PINN) framework has been successfully developed and applied to solve the multi-dimensional multi-group neutron diffusion eigenvalue problem. The modular execution pipeline reliably handles the continuous mapping of neutron fluxes and the prediction of the dominant eigenvalue (λ). The methodology was thoroughly verified through two simple scenarios: a homogeneous reactor and a multi-material heterogeneous core. In all cases, the combination of the Adam and L-BFGS optimizers, supported by the integration of coarse prior solutions, effectively guided the network to avoid non-physical local minima. The numerical results demonstrate to be accurate. The errors for the eigenvalue remained reasonable, reaching 710.57 pcm for the 2D heterogeneous problem. A local inspection of the absolute residuals revealed that while continuous networks tend to concentrate errors at steep gradient regions, such as material corners and reflector boundaries, the global behavior of the reactor is captured with high fidelity. These results confirm that the proposed data-driven PINN scheme is a possibility to integrate experimental data with physical laws of a differential equation.

Acknowledgments

This project has been partially funded with "Ayuda a Primeros Proyectos de Investigación (PAID-06-24)", Vicerrectorado de Investigación of the Universitat Politècnica de València (UPV), project CIGE/2024/72 funded by the Conselleria d'Educació, Cultura, Universitats i Ocupació of Generalitat Valenciana (GE 2025) and project PID2023-152604OB-I00 by MICIU/AEI /10.13039/501100011033 and FEDER, EU.

References

[1] W. M. Stacey, *Nuclear Reactor Physics*, 2nd. Wiley-VCH, 2007.

[2] Y. Yang et al., "A data-enabled physics-informed neural network with comprehensive numerical study on solving neutron diffusion eigenvalue problems," *Annals of Nuclear Energy*, vol. 183, p. 109 656, 2023.

[3] M. H. Elhareef and Z. Wu, "Physics-informed neural network method and application to nuclear reactor calculations: A pilot study," *Nuclear Science and Engineering*, vol. 197, no. 4, pp. 601–622, 2023.

[4] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[5] I. E. Lagaris, A. Likas, and D. I. Fotiadis, "Artificial neural networks for solving ordinary and partial differential equations," *IEEE Transactions on Neural Networks*, vol. 9, no. 5, pp. 987–1000, 1998.

[6] Q. Yang, Y. Deng, Y. Yang, Q. He, and S. Zhang, "Neural networks based on power method and inverse power method for solving linear eigenvalue problems," *Computers and Mathematics with Applications*, vol. 147, pp. 14–24, 2023.

[7] H. Baty, "A hands-on introduction to physics-informed neural networks for solving partial differential equations with benchmark tests taken from astrophysics and plasma physics," *arXiv preprint arXiv:2403.00599*, 2024.

[8] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019.

Proper orthogonal decomposition for neutron noise simulation an its application to small modular reactors.

Antoni Vidal-Ferràndiz[♭],[1] Amanda Carreño[♮], Damián Ginestar[♭], Gumersindo Verdú[♮]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

(♮) I.U. de Seguridad Industrial, Radiofísica y Medioambiental, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

The numerical simulation of neutron flux fluctuations, commonly referred to as neutron noise, constitutes a fundamental problem in the field of nuclear reactor analysis. This phenomenon can be modelled by the time-dependent neutron diffusion equation, which is computationally expensive to solve since the spatial discretization of full-core geometries leads to large systems of ordinary differential equations. In addition, the presence of both prompt and delayed neutrons introduces different time scales, making the system stiff and requiring the use of implicit time integration schemes. Each time step then involves solving a large linear system, which significantly has a large computational cost, and consequently a large computational time [1]. Different computations such as the placement of the neutron nooise source [2] imply to solve this problem many times. This motivates the use of reduced-order models that can reproduce the main dynamics of the system with a much lower computational cost.

To address this, we propose to use a Reduced Order Modeling (ROM) strategy based on Proper Orthogonal Decomposition (POD) [3], [4]. The main idea behind building reduce order models using POD is to find a low-dimensional subspace from the high-dimensional solution space that captures the most important dynamics of the system. This is done through of a set of snapshots that are a set of high-fidelity solutions. They are computed for a representative range of parameters or times, and an orthonormal basis is then extracted using a Singular Value Decomposition (SVD). These basis functions, often called 'POD modes', are chosen to minimize the reconstruction error. The main advantage is that the projection of the original operator onto the reduced POD basis greatly reduces the size of the system. However, the quality of the approximation depends strongly on the number of snapshots used to build the basis. Different sets of snapshots are computed to test the reduced order models POD for two reactor benchmarks, related with the neutron noise problem.

[1] anvifer2@upv.es

2 Neutron diffusion equation

The behavior of the neutron distribution can be approximated by the multi-group neutron diffusion equation, which for G energy groups and N_p families of delayed neutron precursors. This model is expressed as [5]

$$\frac{1}{v_g}\frac{\partial \Phi_g(\vec{r},t)}{\partial t} - \nabla\cdot(D_g(\vec{r},t)\nabla\Phi_g(\vec{r},t)) + \Sigma_{t,g}(\vec{r},t)\Phi_g(\vec{r},t) = S_{f,g}(\vec{r},t) + S_{s,g}(\vec{r},t) + S_{d,g}(\vec{r},t) \quad (1)$$

where the source terms for fission ($S_{f,g}$), scattering ($S_{s,g}$), and delayed neutrons ($S_{d,g}$) are defined as

$$\begin{aligned} S_{f,g}(\vec{r},t) &= (1-\beta)\chi_{p,g}\sum_{g'=1}^{G}\nu\Sigma_{f,g'}(\vec{r},t)\Phi_{g'}(\vec{r},t), \\ S_{s,g}(\vec{r},t) &= \sum_{g'\neq g}\Sigma_{s,g'\to g}(\vec{r},t)\Phi_{g'}(\vec{r},t), \\ S_{d,g}(\vec{r},t) &= \sum_{p=1}^{N_p}\chi_{d,p,g}\lambda_p\mathcal{C}_p(\vec{r},t). \end{aligned} \quad (2)$$

The concentration of the p-th precursor group follows the balance equation

$$\frac{\partial \mathcal{C}_p(\vec{r},t)}{\partial t} = \beta_p\sum_{g'=1}^{G}\nu\Sigma_{f,g'}(\vec{r},t)\Phi_{g'}(\vec{r},t) - \lambda_p\mathcal{C}_p(\vec{r},t), \quad p=1,\ldots,N_p. \quad (3)$$

In the previous expressions, Φ_g denotes the neutron flux for the g-th energy group, and $\mathcal{C}_p$ represents the delayed neutron precursor concentration for the p-th family. The remaining magnitudes are known as the macroscopic cross-sections anddepend on the materials in the reactor core configuration [5].

In this work, we are interested on studying the evolution in time of neutron noise defined as

$$\delta\Phi_g(\vec{r},t) = \Phi_g(\vec{r},t) - \Phi_{g,0}, \quad (4)$$

where $\Phi_{g,0}$ is the neutron flux obtained by forcing the criticality of the reactor at steady-state.

For the spatial discretization of model (1), a finite element method is applied (see [6] for more details). The resulting time-dependent algebraic system is known as the full order model (FOM) [7]

$$V\frac{d\mathbf{\Phi}}{dt} + L\,\mathbf{\Phi} = F\,\mathbf{\Phi} + \sum_{p=1}^{N_p} M_p\,C_p, \quad (5)$$

$$P\,\frac{dC_p}{dt} = -\lambda_p\,P\,\mathbf{C}_p + \beta_p\,R_p\,\mathbf{\Phi}, \qquad p=1,\ldots,N_p, \quad (6)$$

where $\mathbf{\Phi}$ and C_p are the discretized vectors of the neutron flux and the p-th precursor concentration, respectively. The matrices V, L, F, M_p, R_p and P arise from the spatial discretization of the velocity, loss, fission, precursor production, precursor coupling and mass operators, respectively. This system is solved by using a semi-implicit method [7].

As the time scales of the fast and thermal neutrons and neutron precursors differ by several orders of magnitude, this system is stiff, and implicit or semi-implicit methods are necessary to perform the time integration [7]. Therefore, a large linear system must be solved at each time step, which is very expensive when the simulation is repeated for different parameters or configurations. This motivates the use of a reduced-order model (ROM).

3 Proper Orthogonal Decomposition

In this section, we describe the Proper Orthogonal Decomposition (POD) strategy as a ROM to reduce the computational cost of the FOM. To define the POD, the state of the dynamical system is represented by $\mathbf{X}(t) = [\Phi_1, \ldots, \Phi_G, C_1, \ldots, C_{N_p}]^T$, where the vectors $\Phi_g \in \mathbb{R}^n$ and $C_k \in \mathbb{R}^n$ are the discretized neutron flux and neutron precursor concentration by suing the finite element method, and n is the number of degrees of freedom. The POD approach seeks to approximate this high-dimensional solution by projecting it onto a low-rank subspace spanned by a set of basis functions $\{\psi_m\}_{m=1}^M$, known as POD modes, such that

$$X(\vec{r}, t) \approx \sum_{m=1}^{M} a_m(t)\psi_m(\vec{r}). \tag{7}$$

This substitution, followed by a Galerkin projection, transforms the high-dimensional system into a reduced system where the unknowns are the coefficients $a_m(t)$ and whose temporal integration can be done with low computational cost.

3.1 Reduced basis generation from snapshots

The POD methodology looks for a low-dimensional subspace that describes the main behavior of the high-fidelity solution $X(\vec{r}, t)$. To construct this subspace, we use the solution of the FOM obtained by using an implicit method and setting 'large' time-steps.

Let the time interval $[0, T]$ be discretized into a set of K discrete time steps $\{t_1, t_2, \ldots, t_K\}$. The high-fidelity solution (or 'snapshot') at each time t_k is represented by a vector $x_k \in \mathbb{R}^N$, where $N = n \cdot (G + N_p)$. We define the snapshot matrix $\mathcal{S} \in \mathbb{R}^{N \times K}$ as

$$\mathcal{S} = [x_1, x_2, \ldots, x_K] \tag{8}$$

The aim is to find an orthonormal basis $\{\psi_m\}_{m=1}^M$(with $M \leq K \ll N$) that maximizes the energy captured from $\mathcal{S}$. This solution is usually obtained by computing the Singular Value Decomposition (SVD) of the snapshot matrix

$$\mathcal{S} = U\Sigma V^T, \tag{9}$$

where the matrix of the left singular vectors $U = [\psi_1, \psi_2, \ldots, \psi_K] \in \mathbb{R}^{N \times K}$ contains the POD modes. The singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_K \geq 0$ in the diagonal matrix Σ represent the relevance or energetic contribution of each mode to the solution.

In practice, the dimension M of the reduced basis is selected according to the singular value decay criterion

$$I(M) = \frac{\sum_{k=1}^{M} \sigma_k^2}{\sum_{k=1}^{K} \sigma_k^2} > 1 - \varepsilon, \tag{10}$$

where ε is a tolerance. By retaining only the first M modes, we define the projection matrix $W = [\psi_1, \ldots, \psi_M]$, which is used to project the original time-dependent operators into the reduced problem.

3.2 Reduced order model

Once the POD basis $\{\psi_m\}_{m=1}^M$ has been obtained from snapshots, we approximate the time-dependent state vector $X(\vec{r}, t)$ as a linear combination of these spatial modes

$$X(\vec{r}, t) \approx \sum_{m=1}^{M} a_m(t)\psi_m(\vec{r}) = W\, a(t) \tag{11}$$

where $W \in \mathbb{R}^{N\times M}$ is the matrix containing M POD modes and $a(t) \in \mathbb{R}^M$ is the vector of reduced coefficients.

The full order semi-discretized system of the neutron diffusion equation once the FEM is applied (5) can be written in a general form as

$$\frac{d}{dt}X(t) = T(t)X(t), \tag{12}$$

where T represents the inverse of the velocity operator and the net production and loss operators (diffusion, scattering, and fission). To derive the Reduced Order Model (ROM), we substitute the expansion (11) into (12) and we pre-multiply the system by W^T (Galerkin method) taking into account that W is an orthonormal basis. This yields the reduced system

$$\frac{da(t)}{dt} = \underbrace{\left\langle W^T, T(t)W \right\rangle}_{\tilde{T}(t)} a(t), \tag{13}$$

which is an ordinary differential system of $M \ll N$ equations. This system is solved by using an implicit backward differentiation formula and adaptive time-step control from the PETSc library [8].

Once the $a(t)$ is computed the neutron flux is reconstructed using the equation (11) .

4 Numerical results

To compare the time-domain results obtained with the ROM against the FOM results, the following errors have been defined

$$\text{RMSE}(\delta\phi_g) = 100 \times \frac{\|\delta\phi_{g,\,ROM} - \delta\phi_{g,\,FOM}\|_2}{\|\delta\phi_{g,\,FOM}\|_2} \qquad (\%), \qquad derror = \delta\phi_{1,\text{ROM}} - \delta\phi_{1,\text{FOM}}, \tag{14}$$

and tow neutorn noise problems have been sutidied, one related with a 1D reactor and another problem associated with a 3D small modular reactor.

4.1 1D UOX reactor

Figure 1: One-dimensional problem.

The first reactor consists of a one-dimensional reactor composed of 17 homogeneous square fuel pins of length 1.26 cm (Figure 1) [9]. Each fuel pin is divided into 8 cells: 4 cells of length 0.13215 cm to model the moderator (blue color) and 4 cells of length 0.18285 cm to model the fuel (red color). Two extra cells of moderator of 0.08 cm are added at the edges of the problem. The cross sections used are defined in [9]. A perturbation located in the seventh fuel pin (marked in green) is introduced as a sinusoidal variation of the cross sections defined as

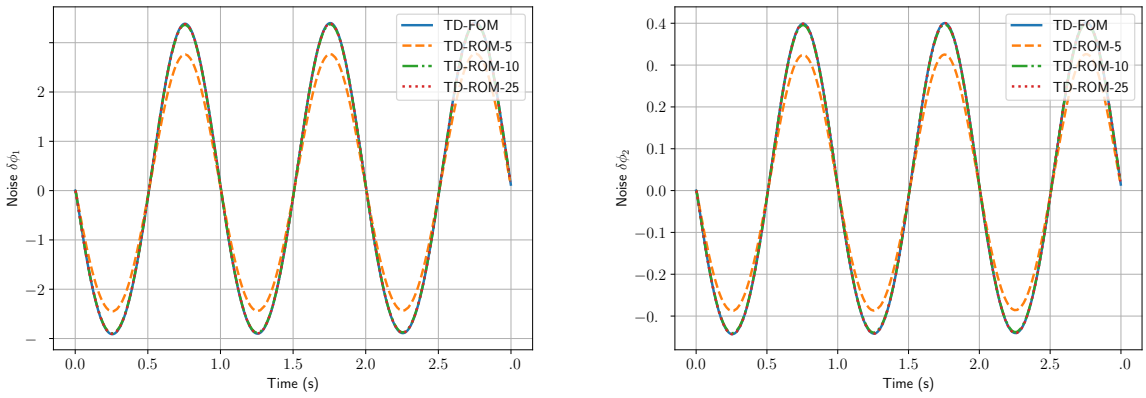
$$\begin{aligned}
\delta\Sigma_t(t) &= 0.004\,\Sigma_t^0 \sin(2\pi t)\,,\\
\delta\Sigma_s(t) &= 0.0034\,\Sigma_s^0 \sin(2\pi t)\,,\\
\delta\Sigma_f(t) &= 0.002\,\Sigma_f^0 \sin(2\pi t)\,.
\end{aligned}$$

We test the POD method for different numbers of snapshots ($k = n_{snaps}$). The set of snapshots is computed by solving the full time-dependent problem by using a large and constant time-step defined by $(t_{end} - t_0)/n_{snaps}$. Figure 2 shows the time-domain neutron noise $\delta\phi_g(t) = \phi_g(t) - \phi_g(0)$ obtained for 5, 10 and 25 snapshots in comparison with the noise obtained from the full order model. Also, the *derror* of the ROM is shown in Figure 3. To quantify the errors, Table 1 shows the RMS error obtained for the different configurations of the POD in comparison with the results of the full order model.

In particular, Table 1 presents the RMSE of the neutron flux noise $\delta\Phi_1$ and $\delta\Phi_2$ for different numbers of snapshots applied to the 1D UOX reactor. With only 5 snapshots the errors exceed 17% for both groups, indicating that the reduced basis is insufficient to represent the solution accurately. However, an important improvement is observed from 10 snapshots onwards, with errors dropping below 1% and continuing to decrease as more snapshots are added, reaching minimum values of 0.02% and 0.10% with 50 snapshots.

Table 1: RMSE results for different number of snapshots for the 1D UOX reactor.

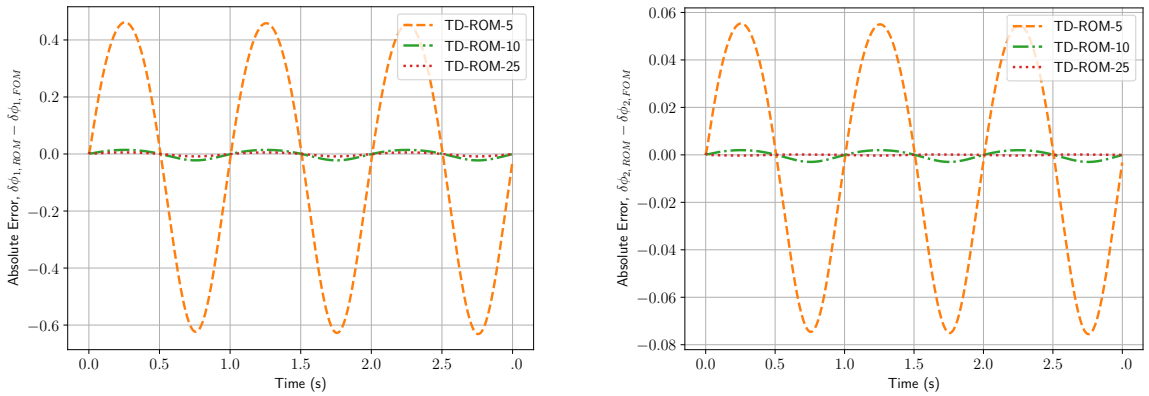
Number of Snapshots	RMSE($\delta\phi_1$) (%)	RMSE($\delta\phi_2$) (%)
5	17.19	17.27
10	0.55	0.73
25	0.23	0.28
50	0.02	0.10
100	0.06	0.15



(a) Fast neutron noise

(b) Thermal neutron noise

Figure 2: Neutron noise results for the 1D UOX problem.



(a) Fast neutron noise

(b) Thermal neutron noise

Figure 3: *derror* using different numbers of snapshots and taking the full order time domain computation as the reference.

4.2 3D ACP-100 SMR reactor

To assess the methodology proposed, a neutron noise problem associated with the small modular reactor (ACP-100) [10] is studied. This pressurized water reactor is developed by China National Nuclear Corporation (CNNC), featuring a short construction cycle, economic competitiveness and flexibility in application. Two-group nuclear cross sections for this reactor along its geometry can be found in [11]. Figure 4 shows a scheme of its geometry.

To induce a transient state suitable for neutron noise analysis, a localized sinusoidal cross-section perturbation given by $\Sigma_{a2}(t) = 0.04\Sigma_{a2}(0)\sin(2\pi t)$ is applied to the macroscopic cross sections in the core's central plane, simulating the effect of a moderator density change. The snapshots were extracted uniformly over the transient duration.

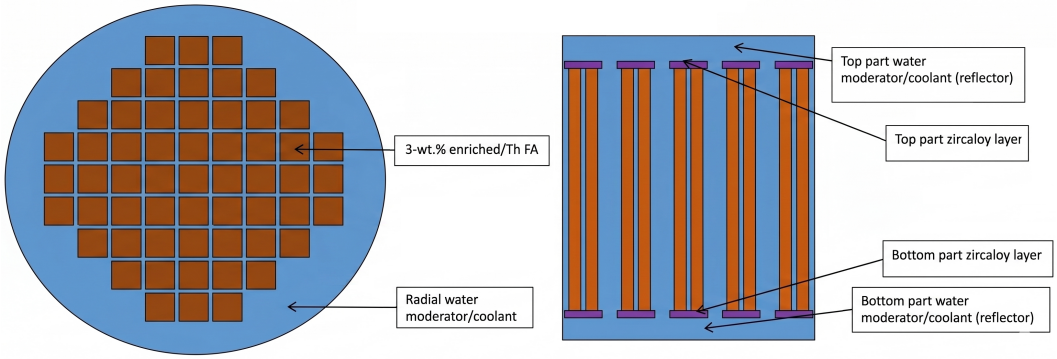


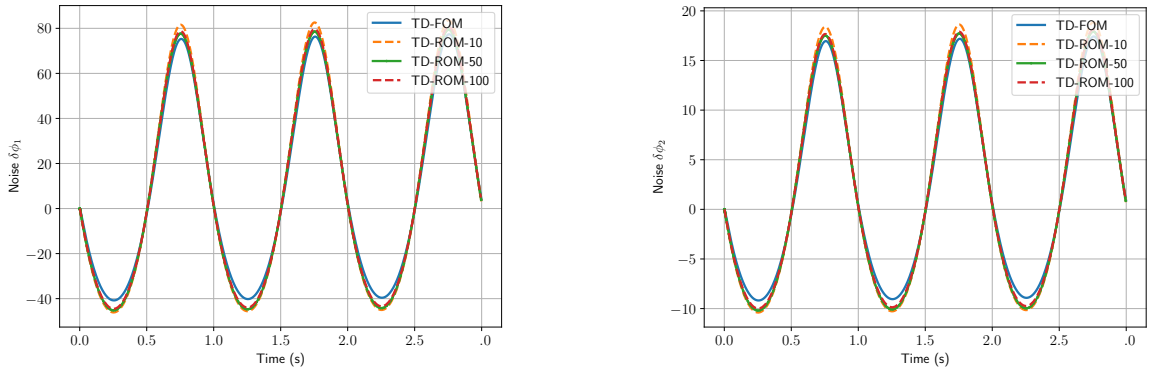
Figure 4: Radial and axial view of the 3D FA mesh. Adapted from [11].

Table 2 presents the RMSE of the two neutron noise flux, $\delta\Phi_1$ and $\delta\Phi_2$, obtained for different numbers of snapshots. The time evolution of the neutron noise for both energy groups is shown in Figure 5. Graphically, one can observe that ROM configurations reproduce the oscillatory behaviour of the full-order model with good agreement. Results show that all cases achieve errors below 2.1%, which demonstrates the overall accuracy of the proposed method. Regarding $\delta\Phi_1$, the lowest error (1.05%) is obtained with only 10 snapshots, while for $\delta\Phi_2$, the best performance is

achieved with 25 snapshots (1.63%). Then, increasing the number of snapshots does not lead to a improvement in accuracy in this case. The configuration with 25 snapshots offers a good balance between accuracy and computational cost.

Table 2: RMSE results for different number of snapshots for the ACP 100 SMR reactor.

Number of Snapshots	RMSE($\delta\phi_1$) (%)	RMSE($\delta\phi_2$) (%)
10	1.05	2.09
25	1.27	1.63
50	1.18	1.78



(a) Fast neutron noise

(b) Thermal neutron noise

Figure 5: Neutron noise results for the 3D ACP-100 SMR reactor

5 Conclusions

In this work, a Proper Orthogonal Decomposition (POD) approach was successfully implemented to reduce the computational burden of simulating neutron noise problems in the time-domain in nuclear reactors. The methodology was verified against a 1D UOX benchmark and also applied to a realistic 3D ACP-100 SMR model.

For the one-dimensional case, the results show that the accuracy of the ROM depends heavily on the number of snapshots. With only 5 snapshots, the RMSE is around 17%, which drops significantly to under 1% when using 10 snapshots, and achieves excellent agreement (0.02% error) with 50 snapshots. In the three-dimensional ACP-100 case, the POD approach maintained an RMSE between 1% and 2% for both fast and thermal energy groups using 10 to 50 snapshots. These results confirm that projecting the stiff, high-dimensional neutron diffusion equations onto a low-rank POD basis preserves the critical system dynamics while vastly reducing the size of the algebraic system to be solved at each time step. Future work will explore non-uniform snapshot sampling techniques to further optimize the basis generation for highly localized noise sources.

Acknowledgments

This work has been partially funded with "Ayuda a Primeros Proyectos de Investigación (PAID-06-24)", Vicerrectorado de Investigación of the Universitat Politècnica de València (UPV), project

CIGE/2024/72 funded by the Conselleria d'Educació, Cultura, Universitats i Ocupació of Generalitat Valenciana (GE 2025) and project PID2023-152604OB-I00 by MICIU/AEI /10.13039/501100011033 and FEDER, EU.

References

[1] A. Vidal-Ferràndiz, A. Carreño, D. Ginestar, C. Demazière, and G. Verdú, "A time and frequency domain analysis of the effect of vibrating fuel assemblies on the neutron noise," *Annals of Nuclear Energy*, vol. 137, p. 107 076, Mar. 2020, ISSN: 0306-4549. DOI: 10.1016/j.anucene.2019.107076.

[2] M. Chillarón, A. Vidal-Ferràndiz, V. Vidal, and G. Verdú, "Classification and location of neutron noise perturbations using convolutional neural networks," *Nuclear Science and Engineering*, vol. 199, no. sup1, S606–S616, 2025.

[3] S. Volkwein, "Proper orthogonal decomposition: Theory and reduced-order modelling," *Lecture Notes, University of Konstanz*, vol. 4, no. 4, pp. 1–29, 2013.

[4] A. Buchan et al., "A pod reduced order model for resolving angular direction in neutron/photon transport problems," *Journal of Computational Physics*, vol. 296, pp. 138–157, Sep. 2015, ISSN: 0021-9991. DOI: 10.1016/j.jcp.2015.04.043.

[5] W. M. Stacey, *Nuclear reactor physics*. John Wiley & Sons, 2018.

[6] A. Vidal-Ferrandiz, R. Fayez, D. Ginestar, and G. Verdú, "Solution of the lambda modes problem of a nuclear power reactor using an h–p finite element method," *Annals of Nuclear Energy*, vol. 72, pp. 338–349, 2014.

[7] Y. Fontenla, A. Vidal-Ferràndiz, A. Carreño, D. Ginestar, and G. Verdú, "Femffusion and its verification using the c5g7 benchmark," *Annals of Nuclear Energy*, vol. 196, p. 110 239, Feb. 2024, ISSN: 0306-4549. DOI: 10.1016/j.anucene.2023.110239.

[8] S. and Balay et al., *PETSc/TAO Users Manual Revision 3.24*. 2025. DOI: 10.2172/2998643. [Online]. Available: http://dx.doi.org/10.2172/2998643.

[9] A. Carreño, A. Vidal-Ferràndiz, D. Ginestar, and G. Verdú, "Frequency-domain models in the SP_N approximation for neutron noise calculations," *Progress in Nuclear Energy*, vol. 148, p. 104 233, 2022, ISSN: 0149-1970. DOI: 10.1016/j.pnucene.2022.104233.

[10] B. Xu, *CNNC's ACP100 SMR: Technique features and progress in china*, Presentation at the 13th INPRO Dialogue Forum on Legal and Institutional Issues in the Global Deployment of Small Modular Reactors, IAEA Headquarters, Vienna, Austria, Oct. 2016. [Online]. Available: https://nucleus.iaea.org/sites/INPRO/df13/Presentations/011_CNNC%27s%20ACP100%20SMR-Technique%20Features%20and%20Progress%20in%20China.pdf.

[11] M. A. R. Ishraq and A. E. Kruglikov, "Assessing the suitability of two-group cross-sections and diffusion coefficients derived from serpent-2 for small modular reactor acp-100," *Nuclear Energy and Technology*, vol. 11, no. 1, pp. 1–12, Feb. 2025, ISSN: 2452-3038. DOI: 10.3897/nucet.11.130422.

A Neural Network Approach for Inverse Kinematics in Arm Pose Reconstruction

Gloria Campos[1]

I.U. de Investigación en Tecnología Centrada en el Ser Humano, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

Over the last decade, virtual reality technologies as head-mounted displays (HMDs) have transitioned from entertainment-focused tools into valuable platforms for a wide range of applications, including engineering and medicine [1]. Among their most notable capabilities is integrated hand tracking, directly embedded within the devices themselves, eliminating the need for external sensors or markers. Advances in computer vision have substantially improved the accuracy and robustness of these systems, enabling reliable real-time estimation of hand position and orientation across a wide range of conditions [2].

In the field of neurorehabilitation, where assessing the patient's motor status is essential for both diagnosis and the evaluation of recovery progress, these devices can be used for detailed tracking of multiple hand and finger landmarks, enabling fine-grained motion capture and a precise representation of hand kinematics. Importantly, recent studies have shown that the resulting kinematic estimations achieve a relatively high level of accuracy and are sufficiently reliable for several quantitative analysis and application scenarios [3, 4].

However, despite these advances, a key limitation of HMD-based hand-tracking is their restricted access to proximal upper-limb kinematics. While HMD-based systems provide highly accurate tracking of the distal upper limb, they offer limited or no direct information about proximal segments [3]. In many implementations, the forearm and upper arm are simplified or not explicitly modelled. In neurorehabilitation, this limitation is critical, as hand tracking alone is insufficient for a complete assessment of arm motor function. Reconstructing full upper-limb pose, particularly at the shoulder and elbow, is therefore essential to evaluate postural behavior, joint coordination, and the mechanical load during task execution.

Existing techniques are primarily based on classical IK formulations, where the upper limb is modeled as an articulated kinematic chain with multiple degrees of freedom (DoF). In these approaches, the hand pose is used as an end-effector constraint, and the joint angles of proximal segments (wrist, elbow and shoulder) are estimated by solving the IK problem. However, the problem is mathematically challenging due to redundancy: the human upper limb has more DoF than constraints provided by the hand pose, leading to an underdetermined system with infinitely many valid solutions. Moreover, the forward kinematics mapping from joint space to Cartesian space is nonlinear, making the inverse problem difficult to solve analytically. As a result, most

[1] glocamhe@htech.upv.es

methods rely on numerical optimization with additional constraints such as joint limits, smoothness or energy minimization to obtain a unique and anatomically plausible solution, although these approaches can be computationally demanding and sensitive to initialization [5, 6].

More recently, artificial intelligence approaches have been proposed to overcome these limitations by neural networks that learn the mapping between end-effector pose and full upper-limb configuration, predicting the location of the segments in the kinematic chain. In robotics, such approaches have already demonstrated promising results, often achieving higher accuracy and improved computational efficiency compared to traditional methods [7, 8].

Therefore, the main hypothesis of this study is that the IK problem can be effectively addressed through a neural network that learns the nonlinear mapping from end-effector information to the full spatial configuration of the upper limb, without requiring prior knowledge of the shoulder or elbow positions, nor of the orientations of the arm joints. Accordingly, the objectives of this work are: first, to formulate the IK problem as a regression task and solve it using both a classical numerical optimization method and a neural network-based model; and second, to compare the performance of both approaches in terms of accuracy and computational efficiency.

2 Methods

2.1 Experimental dataset

The dataset used in this study was obtained from a previously conducted experimental study approved by the Ethics Committee of the Universitat Politècnica de València (Approval No: P16280225). The original study involved the acquisition of upper limb joint positions during simple reaching tasks, recorded by a laboratory-grade motion capture system (Optitrack, NaturalPoint Inc., USA) at a frequency of 60 Hz. The anatomical points captured by the motion capture system in the three-dimensional space were shoulder, elbow, wrist and forehead.

Data from 39 healthy adults (20 men (51.3%) and 19 women (48.7%), mean age 41.6 $\pm$ 13.9 years) were included in the analysis. Each participant performed 12 repetitions of the task, resulting in a total of 468 trials. Each trial lasted approximately 6 seconds, yielding between 300 and 400 samples per trial. Thus, the resulting dataset contained a total of 145,413 samples.

2.2 Data preprocessing

A preprocessing pipeline was implemented to prepare the data for both the numerical and neural-based IK models. First, the upper arm and forearm segment lengths were estimated by computing the Euclidean norm of the vectors defined between consecutive joint centers (shoulder-to-elbow and elbow-to-wrist). Then, spatial normalization was applied to all samples by translating all joint coordinates such that the wrist was mapped to the fixed reference point [1,1,1]. This transformation referenced all trajectories in a relative coordinate system removing global position variability. A non-zero-centered reference point was chosen to avoid null-centered inputs.

For the neural network approach, the dataset was divided using an intra-subject partitioning strategy, reserving 4 of the 12 trials from each individual (33.3% of the data) for testing to evaluate the model's generalization performance on unseen movement sequences.

2.3 Kinematic model of human arm

The human arm is modeled as a kinematic chain with 7 DoF distributed across the shoulder, elbow, and wrist joints [9]. In this simplified representation, the arm is composed of two rigid links, neglecting soft tissue deformation and joint translation effects: the upper arm and the forearm. As shown in Figure 1, the shoulder is modeled as a 3-DoF spherical joint accounting for abduction/adduction,

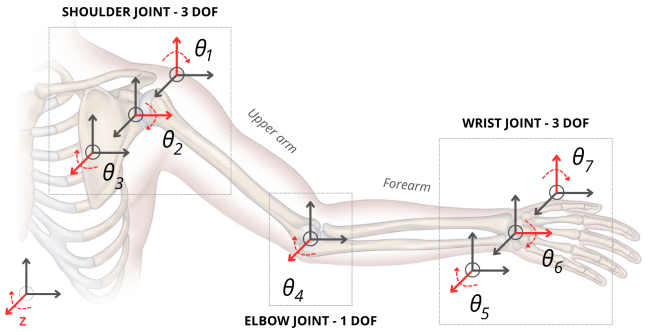


Figure 1: Human arm model with 7-DoF.

flexion/extension, and internal/external rotation. The elbow is modeled as a 1-DoF hinge joint representing flexion/extension. The wrist is represented as a 3-DoF joint capturing the functional range of motion of the distal upper limb, including flexion/extension, radial/ulnar deviation, and forearm pronation/supination.

The Denavit–Hartenberg (DH) convention [10] is employed to describe the transformation from the shoulder to the end-effector (hand) (Table 1). Here, the z_{i-1} axis defines the axis of rotation for each joint, while the x_i axis is the common normal between consecutive joint axes. The parameters θ_i, d_i, a_i, and α_i represent the joint angle, link offset, link length, and link twist, respectively.

Table 1: DH parameters of the 7-DoF human arm model.

Joint (i)	θ_i	d_i	a_i	α_i
1	θ_1	0	0	$\pi/2$
2	θ_2	0	0	$\pi/2$
3	θ_3	l_{upper}	0	$\pi/2$
4	θ_4	0	0	$-\pi/2$
5	θ_5	$l_{forearm}$	0	$\pi/2$
6	θ_6	0	0	$\pi/2$
7	θ_7	0	0	0

2.4 Numerical inverse kinematics

Given the 7-DoF redundant nature of the human arm model, the objective is to estimate the joint configuration vector $\mathbf{q} \in \mathbb{R}^7$ that minimizes the spatial error between the predicted end-effector position and the experimental target. The problem is formulated in a shoulder-fixed coordinate frame, where the shoulder position $\mathbf{P}_{shoulder}$ is used as the origin of the kinematic chain for each frame.

Forward kinematics formulation

The mapping from joint space to Cartesian space is defined through forward kinematics. Based on the DH parameters in Table 1, the homogeneous transformation matrix $\mathbf{A}_i$ for each joint is given by:

$$\mathbf{A}_i(\theta_i) = \begin{bmatrix} \cos\theta_i & -\sin\theta_i\cos\alpha_i & \sin\theta_i\sin\alpha_i & a_i\cos\theta_i \\ \sin\theta_i & \cos\theta_i\cos\alpha_i & -\cos\theta_i\sin\alpha_i & a_i\sin\theta_i \\ 0 & \sin\alpha_i & \cos\alpha_i & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

The full transformation from the shoulder to any joint is obtained by sequential multiplication of the corresponding DH matrices:

$$\mathbf{T}^i_{shoulder}(\mathbf{q}) = \prod_{j=1}^{i} \mathbf{A}_j(\theta_j) \tag{2}$$

In particular, the wrist position is extracted from the last transformation matrix as:

$$\hat{\mathbf{x}}_{wrist} = \left[\mathbf{T}^7_{shoulder}(\mathbf{q})\right]_{1:3,4} \tag{3}$$

Similarly, the elbow position is reconstructed as:

$$\hat{\mathbf{x}}_{elbow} = \left[\mathbf{T}^4_{shoulder}(\mathbf{q})\right]_{1:3,4} \tag{4}$$

Non-linear optimization with constraints

The IK problem is formulated as a non-linear least-squares optimization problem, where the objective is to minimize the Euclidean distance between the predicted wrist position and the target wrist position expressed in the shoulder-centered frame:

$$\min_{\mathbf{q}} J(\mathbf{q}) = \frac{1}{2} \left\| \hat{\mathbf{x}}_{wrist}(\mathbf{q}) - (\mathbf{x}_{target} - \mathbf{P}_{shoulder}) \right\|_2^2 \tag{5}$$

The optimization is solved using the Trust Region Reflective (TRF) algorithm, which is particularly suitable for constrained nonlinear least-squares problems. This method allows the incorporation of box constraints, enabling the enforcement of anatomical joint limits ($\theta_i \in [-\pi, \pi]$) while maintaining numerical stability in the presence of kinematic redundancies [11]. The solver does not require an explicit analytical Jacobian, as it is internally approximated through finite differences.

To ensure temporal consistency and improve convergence stability, a warm-start strategy is employed. For each frame $t > 0$, the initial solution is set to the previous estimate:

$$\mathbf{q}^0_t = \mathbf{q}^*_{t-1} \tag{6}$$

This strategy promotes smoothness in joint space trajectories and improves convergence behavior in sequential motion reconstruction.

2.5 Neural-based inverse kinematics

In this study, the IK problem is reformulated as a supervised regression task aimed at enabling full upper-limb reconstruction from signals that are available in HMDs. Unlike conventional analytical IK, the considered problem is redundant, since multiple shoulder and elbow configurations may produce the same wrist position. Therefore, rather than computing a unique closed-form solution, the proposed neural model aims to learn a plausible mapping from observable variables to joint configurations directly from experimental motion data, implicitly capturing biomechanical patterns between the shoulder and elbow. In particular, the input space is defined to match the type of information that can be reliably obtained from current HMD-based hand tracking systems, namely the Cartesian coordinates of the wrist and the device pose, used as an approximation of

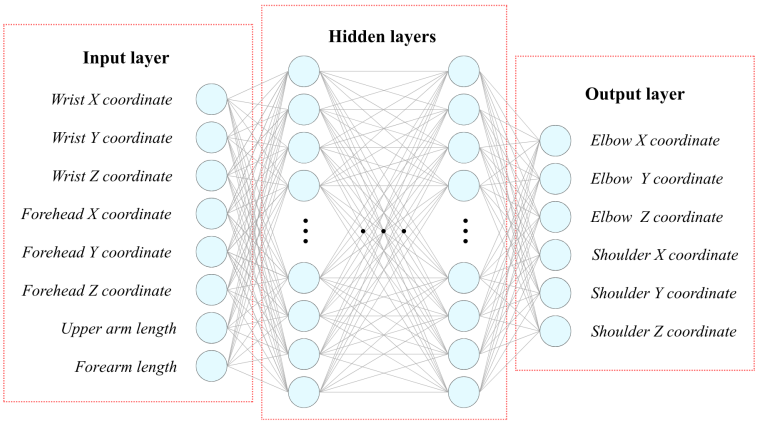


Figure 2: Architecture of the proposed MLP for neural-based upper-limb reconstruction.

the forehead position, together with subject-specific anthropometric parameters (upper arm and forearm lengths). This choice ensures that the trained model can later perform inference using hand-tracking data extracted from an HMD.

The objective is to learn a continuous mapping $\mathcal{F} : \mathcal{X} \subset \mathbb{R}^8 \rightarrow \mathcal{Y} \subset \mathbb{R}^6$, where the output space $\mathcal{Y}$ corresponds to the Cartesian positions of the shoulder and elbow joints. The mapping $\mathcal{F}$ is approximated using a Multilayer Perceptron (MLP), defined as a composition of L layers:

$$\mathbf{h}^{(l)} = \sigma^{(l)} \left(\mathbf{W}^{(l)} \mathbf{h}^{(l-1)} + \mathbf{b}^{(l)} \right), \tag{7}$$

with $\mathbf{h}^{(0)} = \mathbf{x}$ and $\hat{\mathbf{y}} = \mathbf{h}^{(L)}$. The final layer is linear and produces the predicted joint coordinates in $\mathbb{R}^6$. The network architecture, shown on Figure 2, consists of an input layer matching the dimensionality of $\mathcal{X}$, followed by h hidden layers with nonlinear activation functions. Dropout regularization with rate δ was considered on hyperparameter selection to improve generalization. The output layer is a fully connected linear layer.

Training and optimization protocol

The network parameters Ω are optimized by minimizing the Mean Squared Error (MSE):

$$J(\Omega) = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2^2 \tag{8}$$

Optimization is performed using the Adam optimizer with learning rate η and batch size β. Gradients are computed via backpropagation. Early stopping with a patience of 12 epochs and best-weight restoration is used to prevent overfitting and improve generalization.

Hyperparameter selection

A grid search was conducted over a structured hyperparameter space Θ, where each model configuration is defined as:

$$\theta = \{\mathcal{A}, \sigma, \eta, \beta, \delta\}$$

where $\mathcal{A}$ denotes the network architecture, defined by the number of hidden layers (3 to 4) and the corresponding number of neurons in each layer. The explored space is defined as:

- Network architecture $\mathcal{A} \in \{[64, 64, 64], [128, 64, 32], [64, 128, 64], [64, 64, 64, 64], [128, 64, 32, 16], [128, 128, 64, 64]\}$
- Activation function $\sigma \in \{\text{ReLU, Leaky ReLU, Tanh}\}$ [12]
- Learning rate $\eta \in \{0.01, 0.001\}$
- Batch size $\beta \in \{16, 32, 64, 128\}$
- Dropout rate $\delta \in \{0.0, 0.1, 0.2\}$

The optimal configuration was selected as the model achieving the lowest validation Mean Absolute Error (MAE):

$$\theta^* = \arg\min_{\theta \in \Theta} \frac{1}{|\mathcal{D}_{val}|} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_{val}} \|\mathbf{y}_i - \mathcal{M}(\mathbf{x}_i; \theta)\|_2 \tag{9}$$

2.6 Evaluation metrics

The performance of both IK methods was assessed in terms of spatial reconstruction accuracy and computational efficiency. Spatial accuracy was quantified using the MAE computed in 3D space. For each frame, the Euclidean distance between the ground-truth and predicted joint positions was first calculated, and the final error was obtained by averaging across all samples:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2 \tag{10}$$

where $\mathbf{x}_i \in \mathbb{R}^3$ and $\hat{\mathbf{x}}_i \in \mathbb{R}^3$ denote the ground-truth and reconstructed Cartesian coordinates of a given joint at frame i, respectively.

For the numerical IK method, the MAE was computed for the elbow joint only. For the neural-based model, MAE was computed independently for both elbow and shoulder joint reconstructions.

Computational efficiency was evaluated by measuring the average inference time per frame for each method.

3 Results

3.1 Optimal configuration of the neural-based method

The grid search procedure identified the optimal neural configuration θ^* as a MLP with four hidden layers (128, 128, 64, 64 neurons per layer), Leaky ReLU activation, learning rate of 0.001, batch size of 32, and no dropout regularization. This architecture follows a progressive width reduction strategy, enabling high-dimensional feature extraction in the initial layers and progressively refined regression mappings in deeper layers.

This configuration achieved a minimum validation MAE of 4.8 mm, representing a substantial improvement over the 66.5 mm obtained by the worst-performing architectural configuration. Best configuration total training time was 4 minutes and 5 seconds.

3.2 Comparison between numerical and neural-based IK

Both approaches were evaluated on the same samples corresponding to the neural-network test set. The comparison considered three criteria: spatial reconstruction accuracy, computational efficiency, and required input information.

As reported in Table 2, the numerical IK method obtained an average elbow reconstruction error of 13.07 ± 9.23 cm, whereas the neural-based method reduced this error to 1.99 ± 1.35 cm. In addition, the numerical solver required 1.97 s per sample, while the trained neural model performed inference in 0.06 s per sample.

Regarding shoulder reconstruction, the neural-based model achieved an average error of 1.86 ± 1.33 cm, indicating consistent reconstruction performance across both proximal and distal joints. In contrast, the numerical formulation does not explicitly estimate shoulder position, as it operates in a shoulder-centered coordinate frame.

The two formulations also differ in their input requirements. In the present setting, where only Cartesian positions are available and no rotational information is considered, the numerical IK approach requires explicit shoulder coordinates to define the shoulder-centered kinematic chain and solve the optimization problem. In contrast, the neural-based method only uses wrist and forehead coordinates, measurements typically available in HMDs, together with subject-specific segment lengths. However, the numerical IK formulation provides an additional advantage, as it yields the full joint configuration of the 7-DoF kinematic chain. Specifically, the solution of the optimization problem returns the complete set of joint angles, enabling reconstruction of the entire upper-limb kinematics in joint space. This provides access to the underlying joint-level representation of the movement, which is not explicitly modeled in the neural-based approach.

Table 2: Performance of IK methods.

Method	Elbow prediction MAE (cm) [mean ± sd]	Shoulder prediction MAE (cm) [mean ± sd]	Computation time per sample (s)
Numerical IK	13.07 ± 9.23	–	1.97
Neural-based IK	1.99 ± 1.35	1.86 ± 1.33	0.06

4 Conclusions

This study compared a classical numerical optimization method and a neural-based regression model for solving the IK problem in upper-limb reconstruction using only Cartesian positions and segment lengths. Results showed that the proposed neural model consistently outperformed the numerical IK approach in both accuracy and computational efficiency on the same test data. In particular, the MLP achieved substantially lower reconstruction error while reducing inference time by approximately 97%, supporting its suitability for real-time applications.

From a practical perspective, the neural-based approach relies on inputs compatible with HMD systems (wrist and forehead positions and anthropometric parameters), whereas the numerical method requires explicit shoulder localization and operates in a shoulder-centered coordinate frame. Therefore, when applying the numerical IK formulation in an HMD scenario, the shoulder position must be estimated or inferred prior to optimization.

A key limitation of this work is that rotational information was not incorporated in either formulation due to its limited availability and consistency in the dataset. Consequently, both methods rely exclusively on positional data, which simplifies the problem but may reduce the accuracy of the reconstructed kinematics, particularly for joints where orientation is relevant. Additionally, evaluation was restricted to controlled motion capture data with similar movement patterns, and validation on real HMD data remains necessary.

Future work will explore the integration of rotational information to improve kinematic fidelity. In addition, the effect of adding new movements to the dataset will be investigated, along with alternative input representations and network architectures to enhance model performance and generalization.

Acknowledgments

This work was supported by the Ministerio de Ciencia e Innovación of Spain (grant PID2022-141498OA-I00) and by the Conselleria de Innovación, Industria, Comercio y Turismo of the Generalitat Valenciana (grant INREIA/2024/73 and CIACIF/2024/252).

References

[1] E. A. Keshner, P. T. Weiss, D. Geifman, and D. Raban, "Tracking the evolution of virtual reality applications to rehabilitation as a field of study," *Journal of NeuroEngineering and Rehabilitation*, vol. 16, p. 76, 12 2019.

[2] D. Reimer, I. Podkosova, D. Scherzer, and H. Kaufmann, "Evaluation and improvement of hmd-based and rgb-based hand tracking solutions in vr," *Frontiers in Virtual Reality*, vol. 4, 7 2023.

[3] A. Casile, G. Fregna, V. Boarini, C. Paoluzzi, F. Manfredini, N. Lamberti, A. Baroni, and S. Straudi, "Quantitative comparison of hand kinematics measured with a markerless commercial head-mounted display and a marker-based motion capture system in stroke survivors," *Sensors*, vol. 23, 9 2023.

[4] D. Borzelli, V. Boarini, and A. Casile, "A quantitative assessment of the hand kinematic features estimated by the oculus quest 2," *Scientific Reports*, vol. 15, 12 2025.

[5] J. S. Toquica, P. S. Oliveira, W. S. Souza, J. M. S. Motta, and D. L. Borges, "An analytical and a deep learning model for solving the inverse kinematic problem of an industrial parallel robot," *Computers and Industrial Engineering*, vol. 151, 1 2021.

[6] I. Dulba and M. Opałka, "A comparison of jacobian-based methods of inverse kinematics for serial robot manipulators," *International Journal of Applied Mathematics and Computer Science*, vol. 23, pp. 373–382, 6 2013.

[7] F. L. Tagliani, N. Pellegrini, and F. Aggogeri, "Machine learning sequential methodology for robot inverse kinematic modelling," *Applied Sciences (Switzerland)*, vol. 12, 10 2022.

[8] V. Kramar, O. Kramar, and A. Kabanov, "An artificial neural network approach for solving inverse kinematics problem for an anthropomorphic manipulator of robot sar-401," *Machines*, vol. 10, 4 2022.

[9] A. Bertomeu-Motos, A. Blanco, F. J. Badesa, J. A. Barios, L. Zollo, and N. Garcia-Aracil, "Human arm joints reconstruction algorithm in rehabilitation therapies assisted by end-effector robotic devices," *Journal of NeuroEngineering and Rehabilitation*, vol. 15, 2 2018.

[10] J. Denavit and R. S. Hartenberg, "A kinematic notation for lower-pair mechanisms based on matrices," *Journal of Applied Mechanics*, vol. 22, pp. 215–221, 6 1955.

[11] M. A. Branch, T. F. Coleman, Y. Li, S. Siam, and J. C. Sci, "A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems," *SIAM Journal on Scientific Computing*, 1999.

[12] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation functions in deep learning: A comprehensive survey and benchmark," *Neurocomputing*, vol. 503, pp. 92–108, 9 2022.

Machine learning model for strength of high-performance concrete forecast by using the hybrid WOA optimized MARS technique

P.J. García–Nieto[1,*], E. García–Gonzalo[1], J.P. Paredes–Sánchez[2] and L.A. Menéndez-García[1]

(1) Department of Mathematics, Faculty of Sciences, University of Oviedo, 33007 Oviedo, Spain.
(2) Department of Energy, Polytechnic School of Engineering of Gijón, University of Oviedo, C/Luis Ortiz Berrocal, Campus de Gijón, 33203 Gijón, Spain.

1 Introduction

In the concrete construction industry, high-performance concrete (HPC) is a new word. HPC must be made with additional cementitious materials like fly ash and blast furnace slag, as well as chemical admixtures like superplasticizer, in addition to the three fundamental components of conventional concrete: Portland cement, fine and coarse aggregates, and water [1,2]. Because high-performance concrete is such a complicated material, it is challenging to simulate its behavior. Numerous independent investigations have demonstrated that the development of concrete strength is controlled by the concentration of other concrete elements in addition to the water-to-cement ratio [1]. Because high-performance concrete is such a complicated material, it is particularly challenging to simulate its behavior [2–4]. The aim of this study was to obtain some predictive models able to perform the accurate calculation of the compressive strength of high-performance concrete (CSHPC) (output variable) as a function of eight specific input variables (cement, fly ash, blast furnace slag, water, superplasticizer, coarse aggregate, fine aggregate and age of testing) [3–5]. This study presents a robust algorithm, relied on multivariate adaptive regression splines (MARS) approach [6,7] and whale optimization algorithm (WOA) [8], for predicting the CSHPC from eight physical input parameters. Additionally, a multilayer perceptron network (MLP) and M5 model tree were fitted to the data with comparison purposes [9–14]. The results of the present study are two-fold. In the first place, the importance (or relevance) of each input variables on the CSHPC (output variable) is presented through the MARS model. Secondly, a MARS model for forecasting compressive strength of high-performance concrete is obtained with success. Indeed, a MARS regression was performed and a coefficient of determination equal to 0.8808 for the CSHPC estimation was obtained when this MARS model was applied to 1030 diverse samples. Furthermore, the results obtained with the MLP approach [10–12] and M5 model tree [13,14] are worse than those obtained with the MARS model. The agreement between experimental data and the model confirmed the good performance of the latter. The results of this investigation support the feasibility of using machine learning tools to estimate

* pjgarcia@uniovi.es

high-performance concrete parameters as an appropriate and accurate approximation. Finally, conclusions of this innovative research work are exposed.

2 Methods

2.1 Experimental dataset

The strength model's reliability was tested using experimental data from 17 different sources [1–5]. Test results were compiled for concrete that contained fly ash, blast furnace slag, and superplasticizer in addition to cement. In order to offer the full information needed for such a study, it was decided that these mixtures were a fairly representative set governing all of the important parameters that determine the strength of HPC. A total of 1030 concrete samples from the aforementioned studies were assessed. Due to larger particles (more than 20 mm), unique curing circumstances, etc., some of the concrete samples were removed from the data during the evaluation. Approximately 700 concrete samples that were prepared using regular Portland cement and allowed to cure normally were assessed. Specimens of various sizes and forms were employed in various investigations. Following established protocols, all of these specimen types were transformed into 15-cm cylinders.

Even while only one name is used to describe each component, these terms actually refer to a number of different forms. Cement, for instance, can have a variety of chemical compositions and be pulverized to varied levels of fineness. In addition to component kinds, mixing quantities and mixing preparation methods have an impact on the characteristics of concrete. Despite the fact that millions of various mixes are described by experimental data found in technical references, no composite of this data has yet been created. Furthermore, it is very difficult to forecast strength from the existing data because a blend is rarely defined with all of the crucial details mentioned. Consequently, the following eight input variables determine the concrete's compressive strength in this method: (1) cement; (2) fly ash; (3) blast furnace slag; (4) water; (5) superplasticizer; (6) coarse aggregate; (7) fine aggregate; and (8) age of testing. Table 1 shows the physico-chemical input and output parameters used in this investigation as well as their means, standard deviations and minima and maxima.

Table 1: The physico-chemical input and output parameters employed in this study jointly with their means and standard deviations.

Input variables	Name of the parameter	Mean	Standard deviation	Minimum value	Maximum value
Cement (kg/m^3)	C	281.2	104.5	102.0	540.0
Fly ash (kg/m^3)	FA	54.2	63.9	0.0	200.1
Blast furnace slag (kg/m^3)	BFS	73.9	86.3	0.0	359.4

Water (kg/m^3)	W	181.6	21.4	121.8	247.0
Superplasticizer (kg/m^3)	S	6.2	5.9	0.0	32.2
Coarse aggregate (kg/m^3)	CA	972.9	77.6	801.0	1145.0
Fine aggregate (kg/m^3)	FA	773.6	80.2	549.0	992.6
Age of testing (days)	AT	45.7	63.2	1.0	63.2
Output variable					
CSHPC (MPa)	CSHPC	35.8	16.7	2.3	82.6

2.2 Multivariate adaptive regression splines (MARS) method

A multivariate nonparametric classification and regression method is called multivariate adaptive regression splines (MARS) [6,7,9]. MARS uses a set of coefficients and piecewise-defined polynomials of degree q (basis functions), also known as splines, that are fully driven by the regression data to establish the functional relationship between the dependent and independent variables [6]. Fitting basis functions to various intervals of the independent variables creates the MARS regression model. Splines often depict the behavior of the dependent variable by having sections that are smoothly related to one another. Typically, the degree q of the splines is chosen by striking a balance between model complexity and performance.

Generally speaking, the following model is used by all MARS models: $\sum_{i=1}^{k} c_i B_i(x)$, where $B_i(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$ represent the model's basis functions, and c_i stands for constant coefficients. A two-step procedure is used to prune and produce the final MARS model. First, an extremely complicated and overfitted model results from a gradual selection of basis functions. Despite being able to fit the data, this model is not very good at predicting new things. Using a regression approach, duplicate basis functions are eliminated one at a time to enhance prediction. The generalized cross validation (*GCV*) approach is used to decide which basis functions will be part of the model. This methodology divides the root mean squared residual error by a penalty parameter that is dependent on the model's complexity. The *GCV* equation is $GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f_M(x_i)\right)^2}{\left(1 - \frac{M + 1 + d \cdot M}{n}\right)^2}$, where d is a penalty for each base function in the model, and M is the number of basis functions [6]. The explanatory variables' contribution to the model's goodness of fit can be used to determine their relevance once the MARS model has been built. MARS determines the variable important scores by calculating the degree to which each variable's goodness of fit is diminished.

2.3 Whale optimization algorithm (WOA)

The Whale Optimization Algorithm (WOA) is a metaheuristic optimizer proposed by Seyedali Mirjalili in 2016, inspired by the hunting strategy of humpback whales, particularly their bubble-net feeding behavior [8]. The algorithm models three main mechanisms: (1) prey encirclement, where solutions are updated based on the current best individual; (2) spiral attack, which simulates helical movement toward the prey by combining contraction and rotation; and (3) global search, which introduces exploration through random movements relative to selected individuals. The transition between exploration and exploitation is controlled by parameters that decrease adaptively over the iterations. WOA is valued for its simplicity, low parameter count, and ability to avoid local optima, making it applicable to continuous optimization problems, feature selection, and model fitting in machine learning.

2.4 Multilayer perceptron (MLP) model

The MLP makes use of the function $f: X \subset \Re^d \to Y \subset \Re^c$, which is explained by [11,12] $f(x) = \phi(\psi(x)) = (\phi \circ \psi)(x)$, $\phi: X \subset \Re^d \to U \subset \Re^h$, $\psi: U \subset \Re^h \to Y \subset \Re^c$, where U is the space of hidden variables, also known as the characteristics space. Considering the MLP architecture, it can be determined that: $\psi_j(x) = \psi(w_j^T + w_{j0})$, where $\psi_j \in \Re^d$ is the parameter vector of the separate neurons, $w_{j0} \in \Re$ is a limit value known as the threshold value, and ψ is the activation function of the neurons in the hidden layer. The hyperbolic tangent and logistic function are the two typical activation functions ψ, both of which are sigmoid functions. $\phi_j(u) = \phi(c_j^T u + c_{j0})$, where ϕ is the neurons' activation function in the output layer, $c_j \in \Re^h$ is their weight vector, and $c_{j0} \in \Re$ is a limit value called the threshold value. In real life, ϕ is typically a dichotomy function, Heaviside step function, or identity transformation. Lastly, $f(x) = \sum_{j=1}^{h} c_j \psi(w_j^T x + w_{j0}) + c_0$ is the formulation of the function that the MLP uses.

2.5 M5 model tree

The M5 model tree is a machine learning algorithm for regression that combines the interpretability of decision trees with the predictive power of linear models. It was proposed by J. Ross Quinlan as an extension of traditional regression trees [13,14]. The algorithm constructs a tree by recursively partitioning the attribute space, selecting at each node the partition that maximizes the error reduction, typically measured by the standard deviation of the target variable. At each leaf, instead of a constant value, a linear regression model of the following type is fitted: $\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$, where $\hat{y}$ is the prediction, x_i are the input variables and β_i are the locally estimated coefficients. After the tree is constructed, pruning processes are applied to reduce overfitting, and smoothing processes are applied, combining the predictions of the leaf nodes with those of their ancestor nodes, typically through weighted interpolation. This mechanism can be generally expressed as: $\hat{y}_{smooth} = \frac{n \cdot \hat{y}_{leaf} + k \cdot \hat{y}_{parent}}{n + k}$, where n is the number of instances in the leaf and k is a smoothing parameter (the standard value is $k = 15$). Thanks to this hybrid structure, the M5 model tree captures non-linear relationships at a

global level through the tree, while modeling local linear dependencies, achieving a balance between interpretability and accuracy in continuous regression tasks.

3 Results

Table 2 gives the determination and correlation coefficients for the WOA/MARS–relied approximation, MLP-type model and M5 model tree for the CSHPC output factor applied to the testing data, each in order.

Table 2: Coefficients of determination (R2) and correlation coefficients (r) for the WOA/MARS–based approximation, MLP approach and M5 model tree fitted in this investigation to the CSHPC variable applied to the testing data.

Model	WOA/MARS	MLP	M5 model tree
R^2	0.8808	0.7149	0.8379
r	0.9385	0.8455	0.9154

Using the WOA/MARS–based model, this study was able to accurately and successfully forecast the compressive strength of high-performance concrete (CSHPC) (output variable) in accordance with the actual experimental values obtained. In fact, Figure 1 compares the CSHPC values observed and predicted using WOA/MARS–based model. The WOA/MARS–based model has been the best fitting, since these results clearly match the outcome criterion of "quality of fit" ($R^2 = 0.8808$).

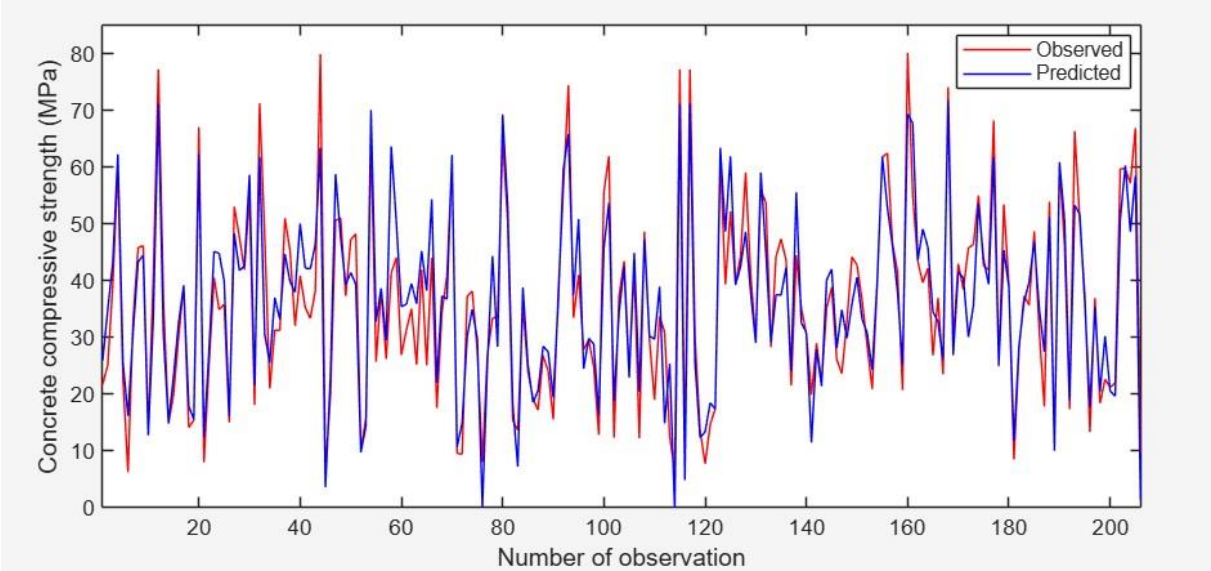


Figure 1: Comparison between the CSHPC values observed and predicted by using the best WOA/MARS–based model ($R^2 = 0.8808$).

4 Conclusions

Because high-performance concrete is such a complicated material, simulating its behavior is challenging. The purpose of this study was to show how WOA/MARS–relied approach

could be modified to forecast concrete's compressive strength. The laboratory created a series of trial batches of HPC that demonstrated positive experimental outcomes. The following conclusions were drawn from this study:

1. The WOA/MARS–relied approximation used in this study is a suitable option to evaluate the CSHPC output variable in various types of concretes with enough precision. Indeed, it was confirmed that the compressive strength of high-performance concrete can be exactly calculated in the construction industry using a WOA/MARS–relied approximation using testing data (20% experimental data not used for training), since a coefficient of determination with a value of 0.8808 is obtained.
2. Due to the fact that MARS approximations produce an explicit mathematical expression of the CSHPC from the input variables as a summation of basis functions (hinge functions and the product of two or more functions of this kind), the MARS approach can be configured into a low-cost microcontroller-based device to achieve a reliable forecasting performance of CSHPC (concrete automation applications).
3. It is possible to arrange the input variables used in the foretelling of the CSHPC in order of their significance. One of the main findings of this investigation is this. In this way, the age of testing (AT) could be considered the most significant predictor of the CSHPC output variable, followed by the contents of cement and water, blast furnace slag, fly ashes, fine aggregate, superplasticizer, and coarse aggregate, each in order.
4. Finally, in future research, for the definition of hybrid models based on concretes' analyses, the ideas presented here are very general and could, in theory, be broadened to include more independent variables. To sum up, this WOA/MARS approach could be successfully applied to other types of concretes with similar or distinct origins, but it is essential to always consider the distinctive qualities of each of its constituents.

The models created using this methodology can be used to calculate the compressive strength with success. These models are practical and simple to use for numerical experiments to examine how each variable affects the mix proportions. The strength model, for instance, can be used to investigate how age or the water-to-binder ratio affect strength.

Acknowledgements

The University of Oviedo's Department of Mathematics provided computational support, for which the authors are grateful. Additionally, we would like to thank Anthony Ashworth for correcting the investigation's English language and spelling.

References

[1] Chang, T.-P., Chung, F.-C., Lin, H.-C., A mix proportioning methodology for high performance concrete. *Journal of the Chinese Institute of Engineers*, 19: 645--655, 1996. http://dx.doi.org/10.1080/02533839.1996.9677830.

[2] Domone, P. L. J., Soutsos, M., An approach to the proportioning of high strength concrete mixes. *Concrete International*, 6 (10): 26--31, 1994.

[3] Arnold, R., Working with Concrete (For Pros By Pros). New York (USA), Taunton Press, 2003.

[4] Nolan, K., Concrete Construction. Carlsbad, California (USA), Craftsman Book Company, 2010.

[5] Gillette, H.P., Shattuck, H.C., Concrete Construction: Methods and Costs. Hampshire (UK), Science & Technology Publishing, 2025.

[6] Friedman, J.H., Multivariate adaptive regression splines. *The Annals of Statistics*, 42: 1--67, 1991.

[7] García-Nieto, P.J., García-Gonzalo, E., Sánchez Lasheras, F., Paredes-Sánchez, J.P., Riesgo Fernández, P., Forecast of the higher heating value in biomass torrefaction by means of machine learning techniques. *Journal of Computational and Applied Mathematics*, 357: 284--301, 2019.

[8] Mirjalili, S., Lewis, A., The Whale Optimization Algorithm. *Advances in Engineering Software*, 95: 51--67, 2016.

[9] García-Nieto, P.J., García-Gonzalo, E., Bové, J., Arbat, G., Duran-Ros, M., Puig-Bargués, J., Modeling pressure drop produced by different filtering media in microirrigation sand filters using the hybrid ABC-MARS-based approach, MLP neural network and M5 model tree. *Computers and Electronics in Agriculture*, 139: 65--74, 2017.

[10] García-Nieto, P.J., Sánchez Lasheras, F., García-Gonzalo, E., de Cos Juez, F.J., PM_{10} concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: A case study, MLP neural network and M5 model tree. *Science of The Total Environment*, 621: 753--761, 2018.

[11] Bishop, C.M., Neural Networks for Pattern Recognition. Oxford (UK), Oxford University Press, 1995.

[12] Ripley, B.D., Pattern Recognition and Neural Networks. Cambridge (UK), Cambridge University Press, 2014.

[13] Quinlan, J.R., Learning with continuous classes. In: *Proceedings of Australian Joint Conference on Artificial Intelligence*, World Scientific Press, Singapore, pp. 343--348, 1992.

[14] Kisi, O., Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *Journal of Hydrology*, 528: 312--320, 2015.

Mathematical model for the thermoelectric response of a bipolar radiofrequency-based hemostatic device

Antonio Romero[♭, 1], Gabriel Ignacio Maureira[♮] and Macarena Trujillo [¥]

(♭) Departamento de Matemática Aplicada, Universitat Politècnica de València
Camino de Vera s/n 46022. Valencia. Spain
(♮) Universitat Politècnica de València,
Camino de Vera s/n 46022. Valencia. Spain.
(¥) BioMIT. Departamento de Ingeniería Electrónica, Universitat Politècnica de València
Camino de Vera s/n. 46022. Valencia. Spain.

1 Introduction

Radiofrequency (RF)-based hemostatic devices have become indispensable tools liver surgery for minimizing intraoperative hemorrhage. These devices operate by delivering high-frequency electrical currents directly to the tissue. This energy induces molecular oscillation (ionic agitation) within the tissue, converting electrical energy into thermal energy that coagulates blood vessels and surrounding tissue. [1]. An important advancement in the hemostatic devices is the integration of saline-linked bipolar sealers using Transcollation technology, such as the Aquamantys system (Medtronic, Portsmouth, NH).

The Aquamantys system is a bipolar, surface-contact device characterized by continuous external irrigation with an isotonic saline solution. It comprises two parallel metallic electrodes, one acting as anode and the other as cathode for RF energy delivery, converting electrical energy into thermal energy by the Jouel effect. Each electrode is integrated with a dedicated internal channel for continuous saline irrigation. The synergy between the saline solution and RF energy improves electrode-tissue electrical coupling by reducing surface impedance, while simultaneously providing a cooling effect that prevents overheating and tissue carbonization under standard clinical operating conditions [2].

Developing a mathematical model that reproduces the behavior of Aquamantis has a great interest in clinical practice, because is an interesting way to evaluate and assess the efficiency of the device. Some previous models have been done [4], but without considering the thermal problem. In this study, a full coupled electro-thermal mathematical model is developed to simulate a more realistic response of the Aquamantys device and evaluate the efficacy of saline irrigation.

2 Mathematical model

2.1 Model domains

The geometry involves as domains the tissue and two metallic tips that corresponds to the electrodes of the Aquamantys device through which the RF currents are applied. The saline which flows outside the electrode tips is modeled as a layer, and it is also a domain (see Figure 1). The tissue is modeled as a block with dimensions 50 mm length, 50 mm width and 50 mm height. Overlying it, the saline layer of constant thickness of 5 mm is modeled and the tips of the Aquamantys device have a rounded form with

[1] aromtar@mat.upv.es

3.5 mm of diameter and lateral orifice with a diameter of 1.44 mm, the center of which is positioned 2.5 mm from the electrode tip and oriented opposite to the second electrode. The two tips are 3 mm separate. A three-dimensional model was developed, but due to the symmetry plane only half of the geometry is represented as shown in Figure 1.

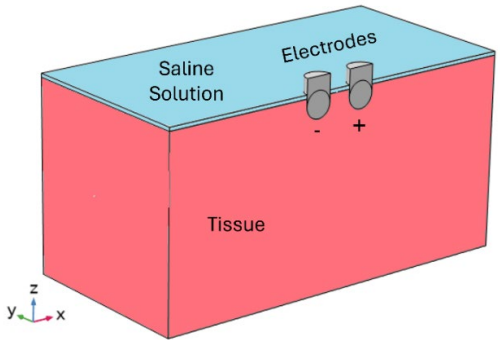


Figure 1: Scheme of the geometry used in the model. Scheme out of scale.

2.2 Governing equations and conditions

The governing equation of the thermal problem is the Bioheat Equation modified by the Enthalpy method to include the intratissue water vaporization [3]:

$$\frac{\partial(\rho h)}{\partial t} = \nabla \cdot (k \nabla T) + q + Q_p \tag{1}$$

where ρ (kg/m^3) is tissue density, h $(J/kg \cdot K)$ the enthalpy, k$(W/m \cdot k)$ the thermal conductivity, T(°C) the temperature and t(s) the time. Q_p represents heat loss by blood perfusion

$$Q_p = \rho_b \cdot c_b \cdot \omega_b (T_b - T) \tag{2}$$

the subindex b, indicates that all these properties are from blood, where ρ_b is the blood density $(1000\ kg/m^3)$, c_b is the specific heat of blood $(4180\ J/kg \cdot K)$, ω_b blood perfusion rate $(6.4 \times 10^{-3}\ 1/s)$ and T_b the blood temperature (37 ℃).

q represents the heat source generated by the electrical problem, serving as the coupling term between the electrical and thermal physics. $q = \sigma(T)|E|^2$, where E is the electric field, $\sigma(S/m)$ electrical conductivity that depends on temperature. The next equation is the governing equation of the electric field:

$$\nabla \cdot \sigma(T) V = 0 \tag{3}$$

where V is the voltage. The electric field is calculated from $E = -\nabla V$. The electrical conductivity dependence of temperature on tissue and saline layer was modeled using the equation:

$$\sigma_i(T) = \begin{cases} \sigma_i \cdot e^{0.015 \cdot (T-37)} & 0 \leq T \leq 99\ ℃ \\ \sigma_i \cdot 2.535 & 99\ ℃ < T \leq 100\ ℃ \\ \sigma_i \cdot 2.535 - \sigma_i \cdot 0.502(T-100) & 100\ ℃ < T \leq 105\ ℃ \\ \sigma_i \cdot 2.535 \cdot 10^{-2} & T > 105\ ℃ \end{cases} \tag{4}$$

where de subindex i, indicates the tissue or the saline solution electrical conductivity measured at 37 ºC. Tissue enthalpy is coupled with temperature using the enthalpy-based method [3].

$$\left.\frac{\partial(\rho h)}{\partial t}\right]_{Tissue} = \frac{\partial T}{\partial t} \cdot \begin{cases} \rho_l c_l & 0 < T \leq 99\ ℃ \\ h_{fg} \cdot \rho_m & 99\ ℃ < T \leq 100\ ℃ \\ \rho_g c_g & T \geq 100\ ℃ \end{cases} \quad (5)$$

where ρ_i and c_i are the density and the specific heat for the tissue when the intra-tissue water is in liquid phase (i=l) and in gas phase (i=g), h_{fg} is the product of the latent heat of vaporization and the water density at $100\ ℃$, and C is the water content in liver (68%). The saline solution (0.9%) is described by:

$$\left.\frac{\partial(\rho h)}{\partial t}\right]_{Saline} = \frac{\partial T}{\partial t} \cdot \begin{cases} \rho_{saline} c_{saline} & 0 < T \leq 99℃ \\ h_{fg} \cdot \rho_m & T \geq 99℃ \end{cases} \quad (6)$$

where ρ_{saline} and c_{saline} are saline density and specific heat of the saline solution, and ρ_m is the water density at 99 ºC ($958\ kg/m^3$).

To perform the calculations, the used material properties are found in Table 1 [4].

Table 1: Material properties (Tissue properties assessed at 37 ºC)

	$\rho\ (\mathrm{kg/m^3})$	$\sigma(S/m)$	$c\ (J/kg \cdot K)$	$\mathrm{k(W/m \cdot k)}$	$\omega\ (1/\mathrm{s})$
Hepatic tissue	1080 (1)	0.31	3455	0.502	
	370.44 (2)		2155		
Blood	1000		4180	0.543	0.0064
Saline Solution	1004	0.774	3200	0.58	
Electrode	6450		840	70	

(1) Tissue in liquid phase. (2) Tissue in gas phase.
ρ: density, σ: electrical conductivity, *c*: specific heat, k: thermal conductivity, ω: blood perfusion.

To evaluate the model, a zero-flux Neumann boundary condition is applied at the symmetry plane for thermal and electrical problems. For the thermal problem, a Dirichlet condition of constant T_0 temperature is applied at the outers of the geometry. T_0 is also equal to the blood temperature and equal to 37ºC. For the electrical problem, a constant volage of +52 V is applied on one electrode and on the other electrode a -52 V is applied. Given the nonlinearity of the governing equations and the complexity of the geometries involved, the model is solved computationally using COMSOL Multiphysics, a software based on the Finite Element Method.

3 Results

Our interest is to check if including saline layer produces a better efficiency of the application or RF currents in terms of the volume of coagulated zone. Figures 2A and 2B show the temperature distribution at the symmetry plane within the tissue without and with including the irrigation saline layer, respectively. The white line is the isoline of T = 80 ºC, temperature in which coagulation is produced. It is observed that the temperature within the tissue is higher when there is saline irrigation. The saline layer, with higher electrical conductivity, provokes that the electrical currents were applied with better efficiency and consequently a higher increase in temperature is produced. In points of 1, 2 and 3 mm depth from the tissue surface in the half plane between the electrodes, the temperature is 97.83 ºC, 94.11 ºC, 79.81 ºC vs 95.27 ºC, 93.80 ºC, 78.09 ºC, for the cases of including or not including the saline layer, respectively.

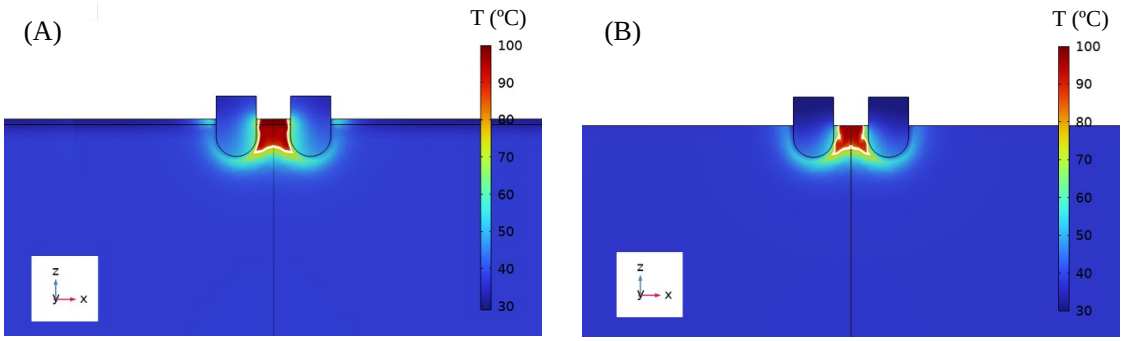


Figure 2: Temperature distribution measured at t = 1 s in the symmetry plane. Without the irrigation saline layer (A) and with irrigation saline layer (B).

4 Conclusions

The model serves as a robust starting point for studying the thermoelectric response of tissue when using a bipolar radiofrequency-based hemostatic device, as it effectively replicates the influence of the saline layer.

Acknowledgements

This research was funded by Spanish Ministerio de Ciencia e Innovation, Agencia Estatal de Investigacion, Fondo Europeo de Desarrollo Regional (Grant PID2022-136273OBC31 funded by MCIN/AEI/10.13039/501100011033/FEDER, UE).

References

[1] S. Ran, X. Chen, Y. Zhang, “Heat conduction in live tissue during radiofrequency electrosurgery”. Journal of the Royal Society Interface, Vol. 21, pp. 1--15, 2024.

[2] M. Trujillo, E. Berjano, “Review of the mathematical functions used to model the temperature dependence of electrical and thermal conductivities of biological tissue in radiofrequency ablation”. International Journal of Hyperthermia, Vol. 29, pp. 590--597, 2013.

[3] Abraham, J. P., Sparrow, E. M., A thermal-ablation bioheat model including liquid-to-vaporphase change, pressure-and necrosis-dependent perfusion, and moisture-dependent properties. International Journal of Heat and Mass Transfer, 50(13-14),2537-2544, 2007.

[4] J. Arenas, J. J. Perez, M. Trujillo, E. Berjano, “Computer modeling and ex vivo experiments with a (saline-linked) irrigated electrode for RF-assisted heating”. BioMedical Engineering OnLine, Vol. 16, pp. 1--15, 2017.

Aerodynamic Shape Optimization using Graph Neural Networks and Neural ODEs

Guillem Pérez[♭,1], Victoria Muerza[♮], Fernando Pascual[♭,♯] and Emilio Larrodé[♭]

(♭) Department of Mechanical Engineering, University of Zaragoza
c/ María de Luna 3, 50018, Zaragoza (Spain)

(♮) Department of Applied Economics,
Gran Vía 2, 50005, Zaragoza (Spain)

(♯) Calypso Instruments EMEA, S.L.,
c/ de Alfonso Solans 12, 50014, Zaragoza (Spain)

1 Introduction

Aerodynamic Shape Optimization (ASO) is a subfield of aerodynamics that focuses on improving the performance of outdoors objects in interaction with air by means of their shape. Traditionally, parameterizing the geometry in ASO has been necessary through algorithms like Free-Form Deformation [1]. Recently, the development of novel neural network architectures has allowed studying the geometry explicitly through meshes [2]. According to this tendency, this paper proposes a new method for ASO.

The method is composed of two concatenated neural networks. The first one is a Graph Neural Network (GNNs) which is used as a surrogate model to assess the geometry performance via the objective function of the problem. Once the GNN is trained, it is employed by the second neural network, a Neural Ordinary Differential Equation (Neural ODE), to search through the design space. The neural ODE works as a diffeomorphic flow, deforming an initial geometry to perform the search, but keeping its topological properties.

To validate the approach, the method is tested afterwards in the two dimensional case of airfoils.

2 Methods

2.1 Graph Neural Networks

GNNs are a type of artificial neural networks proposed by Scarselli [3]. Conceptually, they are identical to the widely known Convolutional Neural Networks (CNNs) used in image processing. However, instead of being applied to an image where each pixel shares information to its neighbours, its input data is a graph and the convolution operation or the message passing is performed between nodes that are connected through the edges of the graph.

In this context, GNNs have already been used to predict airfoil performance [4]. The method presented in this work employs a GNN model to predict the aerodynamic performance given a geometry.

[1] guillem.perez@unizar.es

2.2 Neural Ordinary Differential Equations (Neural ODEs)

Neural ODEs were proposed by Chen [5]. They are a family of deep neural networks, which parameterize the derivative of a hidden state. In this case, this hidden state will be a velocity field for the deformation of the space. Then, this deformation field will be applied to the initial geometry, and the GNN will evaluate the deformed geometry performance. The deformation field is given by the following differential equation

$$\frac{d\mathbf{X}}{dt} = \mathbf{V}_\theta(\mathbf{X}, t) \tag{1}$$

$\mathbf{X}(t)$ is a general point in the space, $t \in [0, 1]$ an artificial time, where $\mathbf{X}(t = 0) = \mathbf{X}_0$ are the initial points and $\mathbf{X}(t = 1) = \mathbf{X}_\theta$ the deformation. And $\mathbf{V}_\theta(\mathbf{X}, t)$ is the neural network with parameters $\{\theta\}$.

Integrating the equation we obtain the final deformation,

$$\mathbf{X}(t) = \mathbf{\Phi}(\mathbf{X}_0; t) \tag{2}$$

where $\mathbf{\Phi}$ is the deformation field. If $\mathbf{V}_\theta$ is smooth and Lipschitz continuous in the interval $t \in [0,\ 1]$, then the flow $\mathbf{\Phi}$ is diffeomorphic [6], and consequently it is injective and preserves topological properties. This characteristic makes the method avoid unwanted geometries like selfintersecting surfaces. Therefore, the neural ODE, given a set of parameters $\{\theta\}$, proposes a geometry that is introduced into the previously trained GNN to evaluate its performance.

Neural ODEs are a new family of models with increasing number of applications. Examples of these applications fields are dynamical systems modeling [5] and image treatment [7], [8]. The contribution of this work is the implementation of neural ODEs as an algorithm for the search in a design space for ASO.

3 Results

This work presents a new method in the ASO subfield. The obtained algorithm has the advantage of not parameterizing the geometry, which would limit the design space explored. The only limitations to the design space are imposed through the design of the neural ODE to guarantee some basic desired properties for the geometry.

The method developed has been implemented and tested in Python. The reason for this is the library PyTorch and the framework and community around it for the development of neural networks software.

4 Conclusions

This work introduces a novel framework for ASO utilizing recent neural network architectures with new characteristics respect to previous frameworks. GNNs and Neural ODEs are used since their designs offer desired properties for the model. Future work will apply this method to current ASO problems with more complexity.

Acknowledgments

This study was supported by the Ministry of Science and Innovation project "NUEVOS ALGORITMOS PARA EL DESARROLLO DE UN INSTRUMENTO DE MEDICIÓN DE ELEVADA PRECISIÓN PARA EL SECTOR EÓLICO WMETERPLUS" (ref. CPP2023-010582).

References

[1] T. W. Sederberg and S. R. Parry, "Free-form deformation of solid geometric models," *SIGGRAPH Comput. Graph.*, vol. 20, no. 4, 151–160, Aug. 1986, ISSN: 0097-8930. DOI: 10.1145/15886.15903 [Online]. Available: https://doi.org/10.1145/15886.15903

[2] J. Jabón, S. Corbera, R. Álvarez, and R. Barea, "Aerodynamic shape optimization using graph variational autoencoders and genetic algorithms," *Structural and Multidisciplinary Optimization*, vol. 67, no. 3, p. 32, 2024, ISSN: 1615-1488. DOI: 10.1007/s00158-024-03771-5 [Online]. Available: https://doi.org/10.1007/s00158-024-03771-5

[3] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The Graph Neural Network Model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009. DOI: 10.1109/TNN.2008.2005605

[4] S. Jena, G. D. Weymouth, A. K. Lidtke, and A. Coraddu, *Predicting airfoil pressure distribution using boundary graph neural networks*, 2025. arXiv: 2503.18638 [physics.flu-dyn]. [Online]. Available: https://arxiv.org/abs/2503.18638

[5] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural Ordinary Differential Equations," 2019. arXiv: 1806.07366 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1806.07366

[6] E. Dupont, A. Doucet, and Y. W. Teh, *Augmented Neural ODEs*, 2019. arXiv: 1904.01681 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1904.01681

[7] H. Niu et al., "On the applications of neural ordinary differential equations in medical image analysis," *Artificial Intelligence Review*, vol. 57, no. 9, p. 236, 2024, ISSN: 1573-7462. DOI: 10.1007/s10462-024-10894-0 [Online]. Available: https://doi.org/10.1007/s10462-024-10894-0

[8] S. Sun, K. Han, D. Kong, H. Tang, X. Yan, and X. Xie, *Topology-Preserving Shape Reconstruction and Registration via Neural Diffeomorphic Flow*, 2022. arXiv: 2203.08652 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2203.08652

Explaining Couette Flows using AI

Clara Trillo-Yagüe[♭1], Andrés Cremades[♭], Ricardo Vinuesa[♮] and Sergio Hoyas[♭]

(♭) Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

(♮) University of Michigan

Abstract

Couette flows, defined as the fluid motion between two infinite parallel plates moving at different relative velocities, represent one of the canonical configurations in wall-bounded turbulence. These flows have been extensively studied due to their unique physical characteristics. A distinctive feature of turbulent Couette flows is the emergence of persistent, extreme-scale coherent structures in the form of streamwise rolls that span the entire core region. Despite decades of research using high-fidelity Direct Numerical Simulations (DNS) and experimental analysis, the fundamental mechanisms governing the genesis, organization, and self-sustenance of these large-scale secondary flow structures remain only partially understood. Traditional structural identification methods often rely on predefined thresholds, leaving the objective causal importance of these regions in the overall flow evolution as an open question.

This work proposes a novel framework to identify which regions and structures are most influential in the formation and sustenance of the aforementioned secondary flow patterns. Building upon recent advances in explainable artificial intelligence (XAI) for fluid mechanics [1], [2], the present approach employs SHAP (Shapley Additive Explanations) values to evaluate the objective contribution of each grid point within the computational domain. To achieve this, a deep-learning surrogate model based on a U-Net architecture is developed and trained to predict the short-term temporal evolution of the flow field from DNS data. Subsequently, the Gradient-SHAP algorithm is implemented to compute the sensitivity of the prediction error with respect to the input fluctuations. These importance maps are then used to define and characterize SHAP-based structures, providing a new perspective on which physical features are truly determinant in the self-sustaining cycle of turbulent Couette flows.

References

[1] A. Cremades, S. Hoyas, and R. Vinuesa, "Classically studied coherent structures only paint a partial picture of wall-bounded turbulence," *Nature Communications*, vol. 16, no. 1, p. 10 189, 2025.

[2] A. Cremades et al., "Identifying regions of importance in wall-bounded turbulence through explainable deep learning," *Nature Communications*, vol. 15, no. 1, p. 3864, 2024.

[1] ctriyag@upv.es

Entropy and Explainability: Exploring Information and Causality in Turbulent Flows

Belén Reverte Badillo[♭1], Andrés Cremades Botella[♭] and Sergio Hoyas Calvo[♭]

(♭) Universitat Politècnica de València (UPV),
Instituto Universitario de Matemática Pura y Aplicada (IUMPA),
Camí de Vera s/n, 46022 Valencia, Spain.

Abstract

Understanding causality in turbulent flows remains one of the central challenges in fluid mechanics, with direct implications for modeling, prediction, and control. Recent advances in data-driven methods have introduced new tools to identify dynamically influential regions directly from data. In particular, explainable artificial intelligence (XAI) techniques based on additive feature attribution, such as SHAP (SHapley Additive exPlanations), have demonstrated the ability to identify flow regions that play a dominant role in the evolution of turbulence. These methods provide a quantitative measure of the contribution of each spatial location or flow variable to model predictions, offering an interpretable framework for analyzing complex turbulent dynamics [1].

Recent studies have shown that SHAP-based analyses can reveal influential regions that do not necessarily coincide with classically defined coherent structures, suggesting that traditional turbulence diagnostics may capture only a partial picture of the mechanisms governing flow evolution [2]. Furthermore, combining explainable deep learning with deep reinforcement learning (DRL) has been shown to significantly improve turbulence control strategies, enabling controllers to outperform conventional reward formulations while requiring substantially less actuation energy [3].

Parallel developments in turbulence research have explored causality from the perspective of information theory. In this framework, entropy-based measures, such as conditional Shannon entropy, provide a quantitative description of the information content of the system, while causal relationships can be interpreted in terms of information flow [4].

In this work, we investigate the relationship between feature attribution and information-theoretic measures in turbulence by combining SHAP-based analyses with entropy-driven diagnostics. Using a turbulent channel flow dataset, we estimate probability density functions of the streamwise velocity at different wall-normal locations and construct a three-dimensional, time-dependent entropy field from consecutive flow snapshots. This enables a direct comparison between the spatial distributions of entropy and SHAP attribution fields at matching time instants. The analysis is further performed across wall-normal planes, providing a scale- and location-dependent characterization of the relationship between regions of varying information content and model attribution.

This approach aims to explore whether there is a systematic relationship between SHAP attribution and entropy-based measures, without presupposing its nature or direction. In particular,

[1] brevbad@upv.es

the analysis is designed to assess whether data-driven attribution methods capture aspects of the information content of the flow or reveal complementary structures not directly reflected by entropy. More broadly, this work contributes to ongoing efforts to connect machine learning interpretability with physically grounded descriptions of turbulent dynamics.

References

[1] A. Cremades, S. Hoyas, and R. Vinuesa, "Additive-feature-attribution methods: A review on explainable artificial intelligence for fluid dynamics and heat transfer," *International Journal of Heat and Fluid Flow*, vol. 112, p. 109 662, 2025.

[2] A. Cremades, S. Hoyas, and R. Vinuesa, "Classically studied coherent structures only paint a partial picture of wall-bounded turbulence," *Nature Communications*, vol. 16, no. 1, p. 10 189, 2025.

[3] M. Beneitez, A. Cremades, L. Guastoni, and R. Vinuesa, "Improving turbulence control through explainable deep learning," *arXiv preprint arXiv:2504.02354*, 2025.

[4] A. Lozano-Durán and G. Arranz, "Information-theoretic formulation of dynamical systems: causality, modeling, and control," *Physical Review Research*, vol. 4, no. 2, p. 023 195, 2022.

Physics-Based Mathematical Decomposition of Electric Vehicle Energy Consumption: Model Formulation, Python Implementation and Real-World Validation Against OBD Measurements

Paula Reguera Carrera[♭], Emilio Larrodé Pellicer[♭] and Victoria Muerza Marín[♮]
(♭) Universidad de Zaragoza | Escuela de Ingeniería y Arquitectura,
C/ María de Luna, 50018 Zaragoza (España).
(♮) Universidad de Zaragoza | Facultad de Economía y Empresa,
Gran Vía de Don Santiago Ramón y Cajal, 18, 50005 Zaragoza (España).

1 Introduction

The accurate prediction of energy consumption in electric vehicles remains an open mathematical problem. While laboratory test cycles provide standardised benchmarks, real-world consumption differs substantially due to terrain, traffic and, critically, aerodynamic conditions that depend on the apparent wind flow, the vector combination of ambient wind and vehicle velocity. The energy consumption of electric vehicles depends on multiple factors that affect vehicle range, making accurate prediction a key challenge.

Existing approaches either rely on purely data-driven models that lack physical interpretability, or on simplified physics-based models that assume flat terrain, constant velocity and zero ambient wind, and their prediction accuracy remains around 5–8% in real-world conditions. A third limitation, less addressed in the literature, is the treatment of motor efficiency: most models assign a constant value rather than a function of operating point.

This work presents a physically-grounded mathematical model implemented in Python that decomposes the battery energy demand of an instrumented electric van into five analytical contributions, aerodynamic, rolling resistance, kinetic, potential and auxiliary, and validates it instantaneously against OBD-II measurements in real urban routes. The model incorporates: (i) measured apparent wind flow as direct input to the aerodynamic term; (ii) motor efficiency maps as functions of torque and RPM; and (iii) a multi-stage signal processing pipeline to correct GPS altitude noise. The central question is whether a closed-form physical model, without free calibration parameters, can reproduce the energy curve measured by the battery management system with engineering precision.

[1] preguera@unizar.es

2 Methods

2.1 Instrumentation and data acquisition

A Citroën ë-Jumpy electric van (100 kW, 1949 kg) was instrumented with three synchronised data sources: an ultrasonic anemometer (apparent wind speed and angle), an external GPS device (speed, position and altitude) and an OBD-II logger (battery power, motor torque and RPM). Three urban routes were recorded in Zaragoza, Spain (24.6 km total). The OBD high-voltage battery power signal provides the ground-truth energy reference against which the model is validated.

2.2 Signal processing pipeline

Sensor heterogeneity introduces two mathematical challenges before the energy model can be applied. First, temporal alignment: the three data streams have different timestamps and sampling rates. Cross-correlation between speed-related signals (GPS velocity and OBD vehicle speed) is used to estimate the clock offset between each pair, after which all signals are resampled to a common 1 Hz grid. Second, GPS altitude denoising: raw barometric-GPS altitude contains high-frequency spikes that introduce large errors in the gravitational energy term. A three-stage filter is applied: (a) Hampel identifier with a sliding window to replace statistical outliers by the local median; (b) Savitzky–Golay polynomial filter, which smooths data by fitting successive subsets of adjacent data points with a low-degree polynomial by least squares, preserving the signal tendency without distorting gradients; (c) a causal slope-limiting pass that prevents the corrected altitude from implying gradients exceeding ±5%, which is the physical maximum on these urban routes.

2.3 Mathematical energy model

The total mechanical energy demand at the wheels per time step $\Delta t$is:

$$E_T = W_{Aero} + W_{Rod} + \Delta E_{cin} + \Delta E_{pot}$$

where each term is defined as:

$$W_{Aero} = \frac{1}{2}\,\rho\, C_x\, S_{eff}\, v_a\ |\, v_a\,|\ \Delta d$$
$$W_{Rod} = M\, g\, (c_0 + c_2\, v^2)\, \Delta d$$
$$\Delta E_{cin} = \frac{1}{2}\, M\, (v_k^2 - v_{k-1}^2)$$
$$\Delta E_{pot} = M\, g\, \Delta h$$

Here $v_a$is the apparent wind speed projected onto the vehicle axis (measured by the anemometer), $v$is vehicle speed, $\Delta d$is distance travelled, $\Delta h$is altitude change, $M$is vehicle mass, $\rho$is air density, $C_x$is the drag coefficient and $S_{eff}$is the effective frontal area. The rolling resistance coefficient follows the empirical law $f_r = c_0 + c_2 v^2$.

Battery energy is obtained by applying the drivetrain efficiency chain. For traction ($E_T > 0$):

$$E_{bat} = \frac{E_T}{\eta_{motor}(RPM, P) \cdot \eta_{trans}}$$

For regenerative braking ($E_T < 0$):

$$E_{regen} = \mid E_T \mid \cdot \eta_{trans} \cdot \eta_{motor,regen}(RPM, P) \cdot \eta_{bat}$$

Motor efficiency η_{motor} is not a scalar constant but a two-dimensional function of shaft power and rotational speed, obtained from published efficiency maps. RPM and torque from OBD are used as real-time inputs to this function at each time step, via polygon containment queries on efficiency contours. A model based on the PMSM efficiency map allows estimation of the global efficiency of the electric vehicle and optimisation of its energy consumption. Auxiliary consumption ($E_{aux} \approx 400$ W, constant) is detected from OBD power measurements during stationary periods. Net battery energy per route is:

$$E_{consumida} = \sum E_{bat} - \sum E_{regen} + \sum E_{aux}$$

2.4 Python implementation

The complete pipeline — data loading, synchronisation, altitude correction, energy model and visualisation — is implemented as a standalone desktop application in Python (tkinter · pandas · scipy · matplotlib). The mechanical power at the wheels is calculated from the physical equations and mapped to electric power using efficiency functions, with OBD data used as both input and validation reference. The application generates cumulative energy curves, instantaneous power time series, efficiency plots and a quantitative summary report comparing calculated and OBD-measured energy by route.

3 Results

The model is validated on three real urban routes in Zaragoza, Spain (T1: 4.65 km, T2: 10.69 km, T3: 9.29 km; 24.6 km and 62 minutes total). For each route, the cumulative battery energy calculated by the model is compared against the energy integrated from the OBD high-voltage battery power signal:

Route	E_calc (kWh)	E_OBD (kWh)	Error (%)
T1	0.591	0.671	−12.0 %
T2	1.647	1.634	+0.7 %
T3	1.640	1.612	+1.7 %
Global	**3.877**	**3.918**	**−1.1 %**

On routes T2 and T3 the model achieves errors of 0.7% and 1.7% respectively — within the measurement uncertainty of the OBD sensor itself. The global error across all three routes is **−1.1%**, obtained without any free calibration parameter: all physical coefficients (C_x, c_0, η_{trans}, η_{bat}) are taken from technical documentation and literature. The cumulative energy curves from the model track the OBD reference throughout the full duration of each route, confirming that the mathematical decomposition captures the dominant physical phenomena at urban driving speeds.

Route T1 shows a larger discrepancy (−12%), attributed to two model limitations: the motor efficiency maps used are scaled from a different motor platform than the one installed in the test vehicle, and the transmission efficiency is assumed constant rather than speed- and torque-dependent. T1 has a higher proportion of regenerative braking events relative to traction, which amplifies both sources of error.

4 Conclusions

A physics-based mathematical model implemented in Python achieves a global energy prediction error of **−1.1%** against OBD measurements on 24.6 km of real urban driving, without any free calibration parameter. The main conclusions are:

- The five-term energy decomposition reproduces the battery energy curve measured by OBD with engineering precision on routes with reliable altitude data
- Motor efficiency as a two-dimensional function of RPM and shaft power, rather than a scalar constant, is essential for accurate traction and regeneration modelling in urban stop-and-go driving, where the operating point varies continuously
- A three-stage GPS altitude correction pipeline (Hampel filter + Savitzky–Golay smoothing + slope limiting) is a necessary preprocessing step: uncorrected altitude noise introduces errors in the gravitational term that propagate directly into the energy balance
- The global error of −1.1% across three routes of different profiles demonstrates that the mathematical model captures the dominant energy phenomena without route-specific fitting
- The remaining discrepancy in T1 (−12%) points to two open problems: the need for vehicle-specific motor efficiency maps and a torque-dependent transmission efficiency model, rather than a constant value
- The Python implementation provides a reproducible, modular framework extensible to other electric vehicle platforms and to highway conditions, where aerodynamic and gravitational contributions become dominant

Future work will validate the methodology on electric heavy-duty trucks operating on interurban routes, where the higher vehicle mass and speed make the aerodynamic and gravitational terms mathematically more significant and the model assumptions more demanding.

References

[1] M. Varnhagen et al., "Effect of Wind on Electric Vehicle Energy Consumption: Sensitivity Analyses and Implications for Range Estimation and Optimal Routing", *ACM Journal on Autonomous Transportation Systems*, Vol. 1, No. 1, 2024. https://doi.org/10.1145/3633460
[2] E. Chowdhury et al., "Estimation of the Impact of Wind on Vehicle Energy Consumption in Battery Electric Vehicles", *International Journal of Automotive Technology*, 2025. https://doi.org/10.1007/s12239-025-00344-3
[3] B. Vlachos et al., "Data Driven Methods for Finding Coefficients of Aerodynamic Drag and Rolling Resistance of Electric Vehicles", *World Electric Vehicle Journal*, Vol. 14, No. 6, pp. 134, 2023. https://doi.org/10.3390/wevj14060134
[4] R. De Cauwer et al., "Energy Consumption Prediction for Electric Vehicles Based on Real-World Data", *Energies*, Vol. 8, No. 8, pp. 8573–8593, 2015.
[5] A. Masmoudi et al., "Permanent Magnet Motor Efficiency Map Calculation and Small Electric Vehicle Consumption Optimization", *Journal of Electrical Systems*, Vol. 14, No. 1, 2018.
[6] K. Layne, M. Booysen, "Electric Vehicle Drivetrain Efficiency and the Multi-Speed Transmission Question", *World Electric Vehicle Journal*, Vol. 14, No. 12, pp. 342, 2023. https://doi.org/10.3390/wevj14120342
[7] R. Schafer, "What Is a Savitzky–Golay Filter?", *IEEE Signal Processing Magazine*, Vol. 28, No. 4, pp. 111–117, 2011. https://doi.org/10.1109/MSP.2011.941097
[8] H. Chaoui, P. Sicard, "Hybrid Energy Consumption Prediction for Battery Electric Vehicles", *Energies*, Vol. 15, No. 17, pp. 6490, 2022. https://doi.org/10.3390/en15176490

Part V

Infectious Diseases and Mathematical Modeling in Public Health

Predictive analysis of hospital mortality in patients admitted from the emergency department

R.de Ramón Frías[b, 1], I. Barrachina Martínez[h] and S. González-de Julián[h]
(b) Universitat Politècnica de València- PhD in Business Management and Administration
(h) INECO, Universitat Politècnica de València

1 Introduction

It is a verifiable fact that the saturation suffered by emergency services has been increasing for some years now- In Spain, the number of visits to emergency services in increased by approximately 27% between 2013 and 2023. [1-3]
As a result, patients are waiting longer than necessary to be visited. Many studies link this increase in emergency room waiting time with an increase in mortality among patients who end up being admitted [4-10].
The objective of this study is to build a predictive model of mortality risk for patients admitted from the emergency department, considering ED waiting times and the sociodemographic and clinical characteristics of these patients, to identify and prioritize actions to improve their care.

2 Methods

This retrospective, observational study is based on a cohort of 13.314 patients seen in the ED and admitted to hospital at a secondary hospital in Valencia in 2023. Data related to ED visits were retrieved from the database of the healthcare district, while hospitalization data were extracted from the Minimum Basic Data Set (MBDS).

The study was approved by the Ethics Committee for Research with Medication (CEIm) on 4 February 2025 (CElm Registration No.: PÎ 5 2025), ensuring protected data confidentiality and anonymization.

2.1 Variables included

a) **Sociodemographic variables available at ED admission:** age, sex, transfer from nursing home, weekend admission, ED readmission (if <72 hours since previous admission)

b) **Clinical variables regarding the patient's prior health status:** level of chronicity according to Clinical Risk Groups (CRG), hospitalization readmission (considered if <30 days since previous admission)

c) **Variables related to initial care in the ED:**
-Length of stay (LOS) until discharge: defined as the time spent in the emergency department (ED) from patient admission until discharge papers for hospitalization are signed.

[1] RMDERAM@upv.edu.es

-Time until bed occupancy: defined as the time between patient admission to the ED and arrival at the assigned bed.
-Prolonged wait (time in ED > 6h)
- ICD-10 code of the main emergency diagnoses

d) Diagnostic variables related to the hospital episode: diagnoses at discharge**,** Diagnostic-Related Group (APR-DRG), severity of illness (severity of the disease and the consumption of resources), risk of mortality (probability of death)

e) In-hospital mortality (outcome variable): identified when the reason for hospital discharge was coded as "exitus"

2.2 Model construction

The predictive model for hospital outcomes is based on data available at the time of emergency department admission.

Preprocessing
The total data was divided into a training group (80%) and a test group (20%) using a train-test split with stratification by the target variable, ensuring that the proportion of positive cases remained consistent in both subsets.

- Train: 10,648 episodes → 93.48% alive / 6.52% deceased
- Test: 2,663 episodes → 93.47% alive / 6.53% deceased

The ED main diagnostics were coded by grouping the 10 most frequent and incorporating the rest as "other" category, generating dummy variables using one-hot encoding. The final dataset consisted of 9 original variables plus the 10 most frequent diagnoses. Given the marked class imbalance, the parameter `class_weight="balanced"` was used in models that allowed it, which is equivalent to assigning a greater penalty to errors in the minority class (deaths).

Final Model
Although three classification models were compared (Logistic Regression, Random Forest, and Gradient Boosting), Logistic Regression with `class_weight="balanced"` and a threshold of 0.3 was ultimately selected because it offered the greatest sensitivity and best clinical interpretability of its coefficients.
The logistic regression model quantified the joint influence of the explanatory variables (independent variables, regressors, or covariates), which were considered predictive, on the probability (1) of death (dependent variable or regressor).

The mathematical expression of the binomial logistic regression for the 11 independent variables included, is:

$$Ln\left\{\frac{P\left(y=\frac{1}{x_1,x_2,\dots\dots,x_n}\right)}{P\left(y=\frac{0}{x_1,x_2,\dots\dots,x_n}\right)}\right\}=\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n$$

where Y is now a binary variable. It takes the value of 1 when the patient dies and 0 when they survive.

Multiple binomial logistic regression allows the direct calculation of the likelihood of the binomial process for different values of the independent variable, which is a logistic distribution function, using the following expression:

$$P\left(y=\frac{1}{x_1,x_2,\dots\dots,x_n}\right)=\frac{e^{\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n}}{1+e^{\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n}}=\frac{1}{1+e^{-(\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n)}}$$

The likelihood ratio indicates how many times more likely a patient is to die for each of the independent variables. This ratio is defined as follows:

$$Odds=\frac{P\left(y=\frac{1}{x_1,x_2,\dots\dots,x_n}\right)}{P\left(y=\frac{0}{x_1,x_2,\dots\dots,x_n}\right)}=\frac{P\left(y=\frac{1}{x_1,x_2,\dots\dots,x_n}\right)}{1+P\left(y=\frac{0}{x_1,x_2,\dots\dots,x_n}\right)}=e^{\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n}$$

The odds ratio (OR) provides a measure of the association between mortality and different levels of each of the independent variables. This is the ratio between the probability with the presence of the independent variables and the probability with the absence of the independent variables.

$$OR=\frac{\frac{1}{1+e^{-(\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n)}}}{1-\frac{1}{1+e^{-(\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n)}}}}{\frac{\frac{1}{1+e^{-\alpha}}}{1-\frac{1}{1+e^{-\alpha}}}}$$

The OR indicates how much the risk of death increases with different levels of the independent variables.

When all independent variables are Xi = 0, or are absent, the odds ratio is equal to:
$Odds=e^{\alpha}$ $\alpha=Ln\,Odd$

When the independent variable Xi is present, the OR is equal to:

$Odds=e^{\beta_i}$ $\beta_i=Ln\,OR$

It quantifies the magnitude of the association between the outcome (exitus) and the factor of interest (age, sex, etc.). This is the odds ratio (OR) used to increase the variable Xi by one unit, while the others remain constant.
The estimators are calculated using maximum likelihood, and the Wald tests and associated p-values are obtained for each.
The test on the general model is performed by calculating the logarithm of the likelihood ratio (log. Likelihood Ratio) and the Hosmer-Lemeshow tests.

3 Results

Descriptive analysis:
First, a descriptive summary of the variables included in the analysis is presented for the 13.214 patients included after cleaning the database.

The mean age of patients admitted through the ED was 70 years (48% over 75 years); the mean time spent in the ED until discharge was 390 minutes, and time until bed occupancy was 617
minutes, a gap suggesting a time management issue. As shown in Table 1, chronicity stage 0 was the least frequent (10%), while distribution across the other three stages was similar (27% and 31%). In total, 11% were referred from a nursing home, 10% had been previously hospitalized within the last 30 days, and 23% were admitted on the weekend. The mortality rate was 6.8%.
Stratifying patients according to the cutoff of 6 hours until bed occupancy revealed that almost 70% waited longer than 6 hours and 49% were over 75 years old Chronicity levels were distributed proportionally across both groups. Weekend admissions were lower in the >6 h group. Also, mortality was higher in this group.

	All cases	*Time until bed occupancy*		*p value*
		<6h	***>6h***	
Chronicity level				*0.01*
none	497 (3.8%)	187(4.5%)	310(7.5%)	
0	1289(9.8%)	422(10.3%)	867(9.5%)	
1	3753(28.4%)	1218(29.6%)	2535(27.8%)	
2	3558(26.9%)	1089(26.5%)	2469(27.1%)	
3	4117(31.1%)	1194(29.1%)	2923(32.1%)	
From nursing home				*0.00*
Yes	1479 (11.1%)	467 (11.4%)	1012 (11.1%)	
Hospital readmission within 30 days of discharge				*0.00*
Yes	1372 (10.4%)	428 (10.4%)	944 (10.4%)	
Weekend admission				*0.00*
Yes	3050 (23.1%)	1063 (25.86%)	1981 (21.75%)	
Discharge destination				
Death	854 (6.5%)	238 (5.8%)	616 (6.8%)	*0.00*

Table 1.- Descriptive analysis of patients

Bivariate Descriptive Analysis
Before proceeding with the modeling, a bivariate descriptive analysis was performed to explore the relationship between each predictor variable and the target variable, *Exitus*. This analysis identifies which variables exhibit a differential distribution between deceased patients (*Exitus* = 1) and survivors (*Exitus* = 0), providing initial evidence of their predictive potential.
For continuous numerical variables, the distributions between both groups were compared using histograms and boxplots, visually assessing potential differences in central tendency and dispersion. For categorical and binary variables, relative frequencies were calculated for each group and presented using

bar charts, allowing for the identification of categories with a higher proportion of deaths. (Results shown in the **appendix**)

Logistic model

The results obtained from the logistic regression analysis for each level of the independent variables are presented in **Table 2** and **Figure 1** where the main clinical conclusions derived from the model are:

-**Previous hospital readmission** (OR =2.40): This is one of the most relevant risk factors. Patients with recent hospitalizations are more than twice as likely to die, reflecting a state of accumulated clinical frailty.
-**Residence in a nursing home** (OR = 1.78): Institutionalized patients have a significantly higher risk, consistent with their profile of greater dependency, comorbidity, and lower functional reserve.
-**Level of chronicity** (OR = 1.31): The higher the level of chronicity, the greater the risk of mortality, validating this variable as a marker of clinical severity.
-**Weekend admission** (OR = 1.22): This is associated with a modest but significant increase in risk, possibly related to differences in resource availability or healthcare organization.
-**Age** (OR = 1.04): Each additional year increases the probability of death by 4%. Although the individual effect is small, its cumulative impact is clinically relevant.
-**ED Diagnosis**: Appendicitis (ICD_K35.80, OR = 0.22) is significantly associated with lower mortality, consistent with its good prognosis and standardized treatment.

Variable	Coefficient	OR	Lower CI 95%	Upper CI 95%
Previous hospital readmission *(reing_HOSP)*	0.8749	2.3986	2.0734	2.7749
Residence in a nursing home *(residencia)*	0.5774	1.7814	1.5494	2.0481
Bronchopneumonia_J18*(ED diagnosis)*	0.2749	1.3164	0.8766	1.9768
Level of chronicity *(nivel_cronicidad)*	0.2681	1.3074	1.2450	1.3730
Weekend admission *(finde)*	0.1995	1.2208	1.1026	1.3517
Age *(edad)*	0.0433	1.0442	1.0415	1.0469
Sex *(sexo)*	0.0254	1.0257	0.9412	1.1179
LOS until discharge *(tpo en urg)*	0.0003	1.0003	1.0002	1.0005
LOS until bed occupancy *(tpo llegar a cama)*	0.0002	1.0002	1.0000	1.0003
ED readmission *(reing_URG)*	-0.1977	0.8206	0.3812	1.7664
Unknown_ED_Diagnosis	-0.2457	0.7822	0.6253	0.9784
Heart failure_I50.9 *(ED diagnosis)*	-0.2509	0.7781	0.5884	1.0290
Pneumonia_J18.9 *(ED diagnosis)*	-0.3521	0.7032	0.5412	0.9137
Others_ED_Diagnosis	-0.4354	0.6470	0.5377	0.7785
Fever_R50.9 *(ED diagnosis)*	-0.9175	0.3995	0.2616	0.6100
Urinary Tract Infection_N39.0 *(ED diagnosis)*	-0.9474	0.3878	0.2936	0.5121
Appendicitis_K35.80 *(ED diagnosis)*	-1.5295	0.2166	0.1083	0.4335
Acute bronchitis_I20.0 *(ED diagnosis)*	-2.0400	0.1300	0.0753	0.2245
COPD_J44.1 *(ED diagnosis)*	-2.5847	0.0754	0.0437	0.1303

Table 2: Logistic regression results with 95% confidence intervals

The model achieved a mean AUC-ROC of 0.7397 ± 0.0047, indicating moderate-to-good discriminatory ability and, above all, high consistency between folds. The low variability between interactions rules out overfitting and confirms that the model reliably generalizes unseen data.

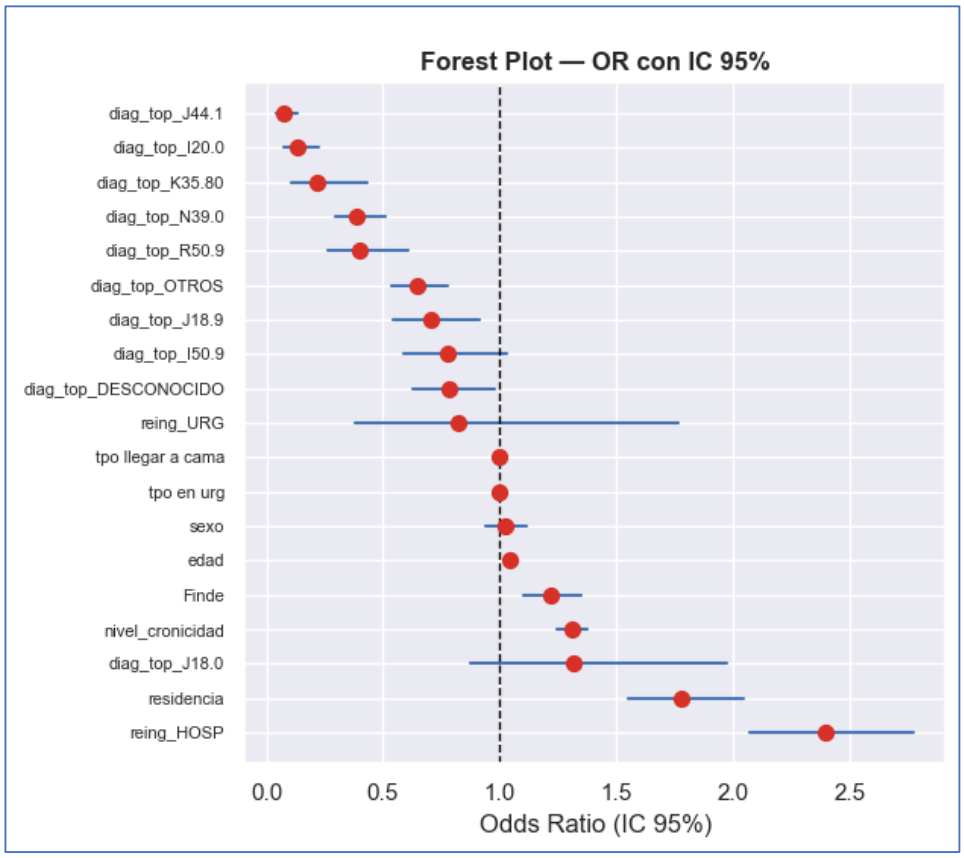


Figure 1.- *OR results for each variable*

Conclusions

Patient care in EDs is a complex process that must be analysed from multiple perspectives. In this study we examined specific characteristics of ED admitted patients and their impact on inpatient mortality. Our analysis combines these characteristics, resulting in a more comprehensive patient profile, composed primarily of elderly individuals with a high degree of frailty.

It is possible to construct a predictive model of in-hospital mortality with acceptable discriminatory capacity (AUC 0.74) using exclusively information available at admission, the main risk factors being previous hospital readmission, residence in a social care center and the level of chronicity.

These findings corroborate the need to implement efficient interventions in the ED for managing patients admitted with risk factors, as for example adding a geriatric consultant to the ED

References

[1] Informe barómetro sanitaria 2024. Secretaría General de Salud Digital, Información e Innovación del SNS Subdirección General de Información Sanitaria https://www.sanidad.gob.es/estadEstudios/estadisticas/BarometroSanitario/Barom_Sanit_2024/BS_2024_Totales/Informe_BS_2024.pdf

[2] https://revistasanitariadeinvestigacion.com/saturacion-en-los-servicios-de-urgencias-causas-y-consecuencias/ (Acceso 29 de julio 2025)

[3] https://www.kingsfund.org.uk/insight-and-analysis/long-reads/whats-going-on-with-ae-waiting-times (Acceso 29 de julio 2025)

[4] Jones, S., Moulton, C., Swift, S., Molyneux, P., Black, S., Mason, N., Oakley, R., & Mann, C. (2022). Association between delays to patient admission from the emergency department and all-cause 30-day mortality. Emergency Medicine Journal. https://doi.org/10.1136/emermed-2021-211572

[5] Singer, A. J., Thode, H. C., Viccellio, P., & Pines, J. M. (2011). The association between length of emergency department boarding and mortality. Academic Emergency Medicine, 18(12), 1324–1329. https://doi.org/10.1111/j.1553-2712.2011.01236.x

[6] Roussel, M., Teissandier, D., Yordanov, Y., Balen, F., Noizet, M., Tazarourte, K., Bloom, B., Catoire, P., Berard, L., Cachanado, M., Simon, T., Laribi, S., & Freund, Y. (2023). Overnight Stay in the Emergency Department and Mortality in Older Patients. JAMA Internal Medicine, 183(12), 1378–1385. https://doi.org/10.1001/jamainternmed.2023.5961

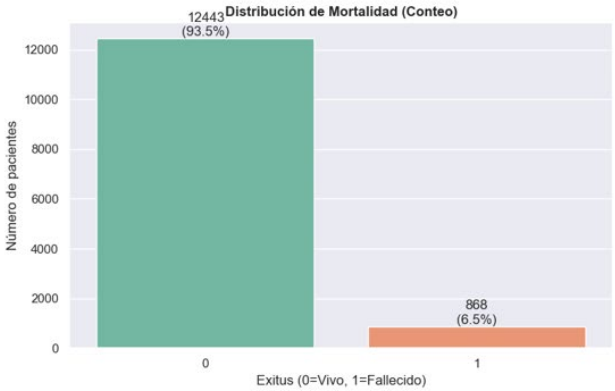
[7] Lauque, D.; Khalemsky, A.; Boudi, Z.; Östlundh, L.; Xu, C.; Alsabri, M.; Onyeji, C.; Cellini, J.; Intas, G.; Soni, K.D.; et al. Length-of-Stay in the Emergency Department and In-Hospital Mortality: A Systematic Review and Meta-Analysis. J. Clin. Med. 2023, 12, 32. https://doi.org/10.3390/ jcm12010032

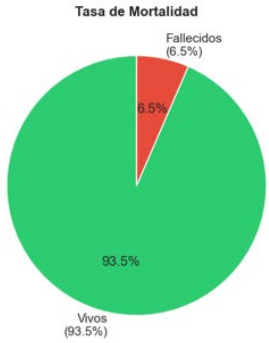
[8] Otto, R., Blaschke, S., Schirrmeister, W., Drynda, S., Walcher, F., & Greiner, F. (2022). Length of stay as quality indicator in emergency departments: analysis of determinants in the German Emergency Department Data Registry (AKTIN registry). Internal and Emergency Medicine, 17(4), 1199–1209. https://doi.org/10.1007/s11739-021-02919-1 6

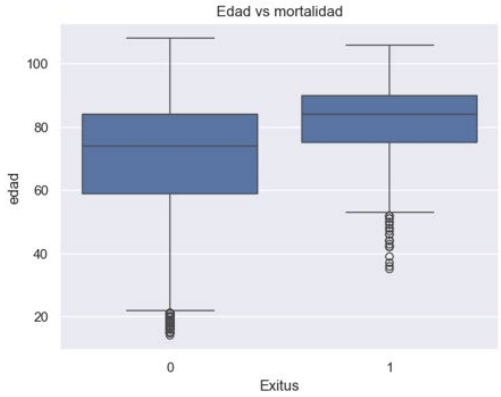
[9] Kim, Y. E., & Lee, H. Y. (2022). The effects of an emergency department length-of-stay management system on severely ill patients' treatment outcomes. BMC Emergency Medicine, 22(1). https://doi.org/10.1186/s12873-022-00760-z 10

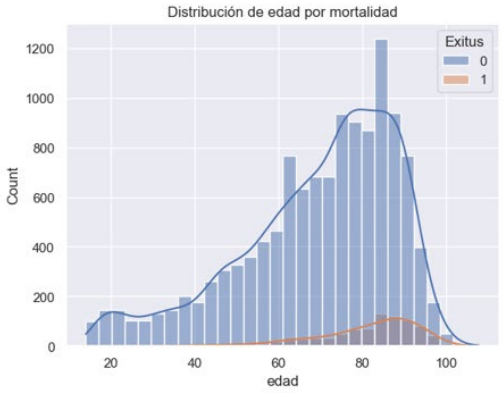
[10] Díaz-Gegúndez, M., Paluzie, G., Sanz-Ballester, C., Boada-Mejorana, M., Terré-Ohme, S., & Ruiz-Poza, D. (2011). Evaluación de un programa de intervención en residencias geriátricas para reducir la frecuentación hospitalaria. *Revista Espanola de Geriatria y Gerontologia*, *46*(5), 261–264. https://doi.org/10.1016/j.regg.2011.03.001

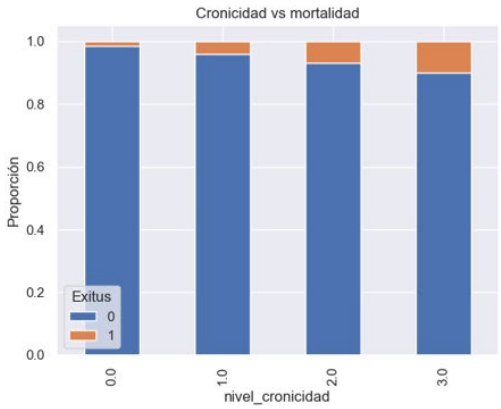
APPENDIX

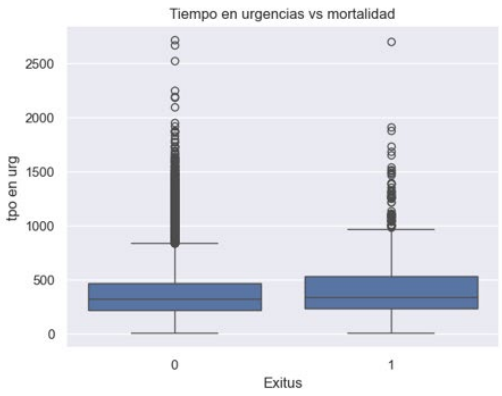

Distribución de Mortalidad (Conteo)
12443
(93.5%)
868
(6.5%)
Número de pacientes
Exitus (0=Vivo, 1=Fallecido)


Tasa de Mortalidad
Fallecidos
(6.5%)
6.5%
93.5%
Vivos
(93.5%)


Edad vs mortalidad
edad
Exitus


Distribución de edad por mortalidad
Exitus
Count
edad


Cronicidad vs mortalidad
Exitus
Proporción
nivel_cronicidad


Tiempo en urgencias vs mortalidad
tpo en urg
Exitus

A data-driven approach to cognitive performance profiling

Pedro Fernández de Córdoba[♭], Álvaro González Cortés[♭1], Esperanza Navarro Pardo[#]

(♭) I.U. de Matemática Pura y Aplicada, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.
(#) Departamento de Psicología Evolutiva y de la Educación, Universitat de València
Avenida Blasco Ibáñez 21, València, Spain.

1 Abstract

The assessment of cognitive performance often relies on aggregate descriptors (typically the mean reaction time and the overall proportion of correct responses) as the primary indicators of individual ability. Although such summary statistics are convenient and readily interpretable, they reduce a rich behavioural signal to a handful of central tendency measures, discarding the distributional structure that often carries the most informative content about how individuals differ. The consequences of this reduction are particularly serious in heterogeneous populations, where meaningful variability operates at scales that group means cannot resolve. A related concern arises from the widespread use of normative scales and diagnostic thresholds as the principal reference for individual evaluation. While indispensable in clinical practice, these categorical frameworks tend to impose discrete boundaries on what is, in essence, a continuous space of cognitive functioning, and may conceal the substantial overlap and internal heterogeneity that characterise both clinical and non-clinical groups.

A complementary perspective consists in describing each individual relative to the empirical distribution of performance observed in the sample under study, rather than against externally defined cut-offs. Within this perspective, latent subgroups are not posited in advance but recovered from the data themselves through unsupervised statistical methods, which simultaneously yield a probabilistic assignment of each participant to the resulting profiles. This data-driven orientation is naturally suited to continuous performance tasks, in which participants respond to long sequences of simple visual stimuli and produce dense trial-by-trial records of response latencies and errors. Tasks of this kind possess two practical features that make them especially valuable for large-scale cognitive characterisation. First, they are inexpensive to deploy and require only minimal computing resources, which makes their administration feasible in a wide range of educational and research settings. Second, because the stimuli are typically non-verbal and context-free, the tasks are largely insensitive to cultural and linguistic background, and can therefore be applied across diverse populations without introducing systematic bias. These considerations have motivated their use as scalable proxies for attentional control and processing efficiency, particularly in studies involving children with attentional difficulties.

Considerable methodological effort has been devoted to the analysis of the latency component of these tasks. Empirical reaction time distributions are typically positively skewed, and the

[1] agoncor@upvnet.upv.es

ex-Gaussian distribution — defined by three parameters (μ, σ, τ) — has become the standard descriptive model in the field, with each parameter associated with a relatively well-defined cognitive interpretation in terms of routine processing speed, trial-to-trial variability, and the prevalence of unusually slow responses. The dataset analysed in the present study consists in reaction time data from a sample of school-aged children in a visual flanker task. Building on that experimental basis, the objective here is methodological rather than descriptive: the aim is to develop a statistical framework that goes beyond the temporal dimension of performance to incorporate response accuracy as an equally informative source of behavioural variability.

The integration of accuracy with latency information has been addressed in the literature through several theoretical traditions, including hierarchical psychometric models, diffusion decision models, and response mixture frameworks. Each of these approaches presupposes a specific structural assumption — be it an item-level psychometric architecture, a mechanistic decision process, or a finite set of latent response strategies — that is fixed before the data are observed. None is intended for the purely data-driven discovery of behavioural phenotypes in a multidimensional performance space free of such prior structure.

The present work develops a framework designed to fill this gap. The proposal rests on two complementary components. The first is the construction of an integrated performance vector that summarises individual cognitive behaviour through four coordinates: the three ex-Gaussian parameters fitted to the distribution of correct response latencies, and an additional dimension representing individual error propensity. This fourth coordinate is obtained from a population-level Beta-Binomial model of error counts, which is then used as an informative prior to derive participant-specific estimates within a coherent Bayesian construction, thereby ensuring statistical consistency between population-level heterogeneity and individual-level inference. The second component is the application of unsupervised Gaussian Mixture Modelling on this four-dimensional space, which enables the identification of latent cognitive profiles without recourse to predefined categories. The framework is illustrated on the dataset described, and yields a fully data-driven characterisation of behavioural heterogeneity anchored to the empirical distribution of the population under study.

Solving Inverse Problems in X-ray CT with Implicit Neural Representations

Hyoung Suk Park[♭],[1] Kiwan Jeon[♭]

(♭) National Institute for Mathematical Sciences, Daejeon, 34047, Republic of Korea

1 Introduction

In X-ray computed tomography (CT), the presence of metal violates the assumptions of the standard CT forward model, resulting in a highly ill-posed inverse problem. In this study, we first review classical approaches for metal artifact reduction (MAR) and then introduce our proposed implicit neural representation (INR)-based method for MAR [1].

2 Methods

The proposed method first generates a high-quality prior image by training a conditioned INR using metal-unaffected projections, and then incorporates this prior into a normalized metal artifact reduction (NMAR) framework for projection completion. To improve robustness under severe metal corruption, we pretrain the encoder-conditioned INR on paired metal-corrupted and artifact-free CT images, thereby embedding data-driven prior knowledge into the INR parameter space. This prior-embedded initialization reduces sensitivity to random initialization and accelerates convergence during measurement-specific refinement. After projection completion with the INR prior, we further suppress residual artifacts using a metal-conditioned correction network, in which the metal mask modulates intermediate features to target metal-induced secondary artifacts while preserving anatomical structures.

3 Results

The image quality was evaluated using CT number (CTN), noise, sharpness, streak artifact, SSIM, metal integrity, bone integrity, and perceptual bias ratio (PBR). The proposed method achieved 0.72 in CTN, 0.01 in noise, 0.73 in sharpness, 1.33 in streak artifact, 0.72 in SSIM, 1.13 in metal integrity, 0.93 in bone integrity, and 1.55 in PBR, resulting in the lowest final score of 0.89.

4 Conclusions

The proposed INR-based MAR framework effectively integrates data-driven prior learning with projection-domain correction, enabling robust reconstruction under severe metal artifacts. By leveraging prior-embedded initialization and metal-conditioned refinement, the method achieves

[1] hspark@nims.re.kr

superior overall performance and consistent improvements across multiple image quality metrics compared to existing approaches.

Acknowledgments

H. S. Park and K. Jeon were supported by the National Institute for Mathematical Sciences (NIMS) grant funded by the Korean government (No. NIMS-B26A10000). H. S. Park and K. Jeon were partially supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00338419).

References

[1] H. S. Park and K. Jeon, "Mgmar: Metal-guided metal artifact reduction for x-ray computed tomography," *arXiv preprint arXiv:2603.13447*, 2026.

Post-Pandemic Social Contact Patters in South Korea: A Nationwide Survey Dataset and Analysis

Min-Kyung Chae[1], Woo-Sik Son[1], Dong-Uk Hwang[1], Kyeongah Nah[2], Minsoo Kim[3], Jong-Hoon Kim[4] and Jonggul Lee[1,*]

(1) Department of Fundamental Research on the Public Agenda, National Institute for Mathematical Sciences,
70 Yuseong-daero 1689 beon-gil, Yuseong-gu, 34047 Daejeon, South Korea.
(2) Busan Center for Medical Mathematics , National Institute for Mathematical Sciences,
179 Gudeok-ro, Seo-gu, 49241 Busan, South Korea.
(3) Department of Electronics Communications Engineering, National Korea Maritime & Ocean University,
727 Taejong-ro, Yeongdo-gu, Busan, South Korea.
(4) Epidemiology, Public Health, Impact, International Vaccine Institute,
1 Gwanak-ro, Gwanak-gu, Seoul, South Korea.

1 Introduction

Age-structured contact matrices are essential inputs for mathematical models of infectious disease transmission, underpinning the estimation of the basic reproduction number (R_0) and the evaluation of age-targeted public health interventions. In South Korea, however, context-specific empirical contact data have not been publicly available, compelling researchers to rely on synthetic contact matrices derived from international studies—most notably POLYMOD [3] and Prem et al. [4]. This reliance is particularly concerning given that the COVID-19 pandemic profoundly altered social behaviour: non-pharmaceutical interventions, widespread adoption of teleworking, and the shift to online education substantially reshaped how and with whom people interact. Evidence from multiple countries suggests that post-pandemic contact patterns have not fully returned to pre-pandemic baselines [5-6], raising the possibility of systematic bias when pre-pandemic or internationally derived matrices are applied to model current transmission dynamics in South Korea. The present study addresses this gap by (i) reporting the first na tionwide, empirically derived social contact dataset for South Korea collected during the post-pandemic period, and (ii) quantifying the error introduced in R_0 estimation when pre-pandemic synthetic contact matrices are used in place of the empirical post-pandemic matrices.

* jlee@nims.re.kr

2 Methods

A national-level social contact survey was conducted over two one-week periods: December 6–12, 2023 (weekdays during the school semester) and February 7–13, 2024 (covering weekdays during school vacation, weekends, and Lunar New Year holidays). A total of 2,415 participants were recruited via quota sampling stratified by age group (10-year intervals, 0–79 years) and ge ographic region (five major provinces), with 1,987 participants retained after rigorous data cleaning. Participants recorded daily close contacts—defined as either skin-to-skin physical contact or conversational exchange of at least three words—including the contact person's age, sex, relationship, location, frequency, and duration of interaction. Daily contact rates were mod eled using a Bayesian generalized linear mixed model with a negative binomial distribution, ac counting for age, sex, household size, and survey period as covariates. Age-structured contact matrices were derived using a bootstrap procedure (2,000 iterations) incorporating reciprocity correction. To quantify changes relative to pre-pandemic contact levels, the next-generation ap proach was applied to compare the dominant eigenvalues of the empirical contact matrices with those from pre-pandemic estimates.

3 Results

A total of 133,776 contacts were recorded across 1,987 participants over 14 days, yielding a mean daily contact rate of 4.81 contacts per person per day. Contact rates varied markedly by period: weekdays during the semester were highest (5.49/day), followed by weekdays during vacation (4.90/day), Lunar New Year holidays (4.46/day), and weekends (3.67/day). Contacts with extended family members increased by more than 1,000% during the Lunar New Year hol idays, accompanied by a pronounced enhancement of inter-regional social mixing. The age-spe cific contact profile exhibited a distinctive post-pandemic pattern that departs substantially from pre-pandemic estimates: contact rates dropped sharply among adults aged 20–29 years—likely reflecting the lasting adoption of teleworking and online education—while the 70+ age group showed the highest contact rates among all adults aged 20 and above, possibly driven by ex- panded grandparental childcare and increased elderly community participation. Applying the next-generation matrix framework, the ratios of R_0 derived from the empirical post-pandemic matrices relative to those from the Prem et al. (2017) pre-pandemic matrix ranged from 0.52 to 0.70 across the four periods, confirming that social contact levels remain substantially below pre-pandemic norms. Critically, the pre-pandemic synthetic matrix systematically overesti- mated the transmission contribution of young adults (20–29 years) and underestimated the in- fection risk among older adults (70+ years), yielding biased age-specific reproduction numbers that could misdirect age-targeted intervention strategies.

4 Conclusions

Post-pandemic contact patterns in South Korea remain 30–48% below pre-pandemic levels and show a distinctive age profile—reduced mixing among young adults and unexpectedly elevated contact rates among the elderly—that diverges markedly from pre-pandemic synthetic matrices. Applying these outdated matrices in epidemiological models therefore introduces systematic bias, overestimating the transmission role of young adults while underestimating infection risk in older age groups. With South Korea now a super-aged society, correcting this bias is not merely a methodological refinement but a prerequisite for designing reliable age-targeted interventions. The publicly available dataset (Figshare: https://doi.org/10.6084/m9.figshare.29312222) provides a reproducible foundation for updated infectious disease modelling in South Korea.

Acknowledgements

This work was supported by the National Institute for Mathematical Sciences (NIMS) grant funded by the Korean government (No. NIMS-B25730000 and NIMS-B26730000).

References

[1] W.-S. Son, M.-K. Chae, D.-U. Hwang, K. Nah, M. Kim, J.-H. Kim and J. Lee, “Social contact patterns in South Korea: an analysis of a survey conducted in 2023–2024”. *BMC Infectious Diseases*, Vol. 25, pp. 295, 2025.

[2] M.-K. Chae, W.-S. Son, K. Nah, M. Kim, J.-H. Kim and J. Lee, “A Nationwide Social Contact Survey Dataset for Public Health and Social Sciences Research in South Korea”. *Scientific Data*, Vol. 13, pp. 603, 2026.

[3] J. Mossong et al., “Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases”. *PLoS Medicine*, Vol. 5, pp. e74, 2008.

[4] K. Prem, A.R. Cook and M. Jit, “Projecting social contact matrices in 152 countries using contact surveys and demographic data”. *PLoS Computational Biology*, Vol. 13, pp. e1005697, 2017.

[5] C.I. Jarvis et al., “Social contact patterns following the COVID-19 pandemic: a snapshot of post-pandemic behaviour from the CoMix study”. *Epidemics*, Vol. 48, pp. 100778, 2024.

[6] J.A. Backer, E.R. Vos, G. den Hartog, C.C. van Hagen, H.E. de Melker, F.R. van der Klis and J. Wallinga, “Contact behaviour before, during and after the COVID-19 pandemic in the Netherlands: evidence from contact surveys, 2016 to 2017 and 2020 to 2023”. *Eurosurveillance*, Vol. 29, pp. 2400143, 2024.

Multi-scale RSV dynamics: SLIRS with ILI-informed calibration and time-series correction

Saebom Jeon[♭],[1] Almudena Sánchez-Sánchez[♮] and Rafael Jacinto Villanueva Micó[♮]

(♭) School of Mathematics, Statistics and Data Science, Sungshin Women's University
2, Bomun-ro 34da-gil, Seongbuk-gu, Seoul, Republic of Korea.
(♮) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

Respiratory syncytial virus (RSV) is a leading cause of acute lower respiratory tract infections in infants, the elderly, and immunocompromised individuals, generating a substantial healthcare burden each winter season [1], [2]. Reliable forecasts of RSV epidemic timing and intensity are essential for resource allocation, hospital surge planning, and the timing of immunoprophylaxis campaigns.

Mechanistic compartmental models based on susceptible–infectious–recovered frameworks have been widely used to capture RSV transmission dynamics and seasonal patterns [3], [4]. These models reproduce the overall seasonal waveform observed in surveillance data, but often exhibit limited interannual variability, with relatively stable peak timing and amplitude compared to the substantial fluctuations observed empirically.

Recent studies have highlighted the role of environmental and epidemiological drivers, including interactions with other respiratory pathogens such as influenza, in shaping RSV dynamics [5], [6]. In particular, statistical associations between respiratory viruses suggest that information from influenza-like illness (ILI) surveillance may provide predictive value for RSV dynamics.

In this work we propose a two-step post-hoc correction framework that combines (1) an ILI-informed calibration of peak amplitude and timing, and (2) an autoregressive (AR) residual model to capture short-term temporal dependence not explained by the mechanistic model. This approach aims to improve the representation of both interannual variability and short-term fluctuations in RSV incidence without modifying the underlying transmission model structure.

2 Methods

2.1 Data and Baseline Mechanistic Model

Weekly laboratory-confirmed RSV hospitalizations and ILI consultation rates were obtained from a national surveillance system covering five complete seasons (2015–2019). The baseline model is a previously validated SLIRS compartmental model that incorporates seasonal forcing of the

[1] alwaysns@sungshin.ac.kr

transmission rate and generates a weekly predicted incidence curve $\hat{R}_s(w)$ for each season s and epidemiological week w.

For each season s, the observed RSV peak amplitude Y_s and peak timing t_s^{RSV} are identified as

$$Y_s = \max_w R_s^{\mathrm{obs}}(w), \qquad t_s^{\mathrm{RSV}} = \arg\max_w R_s^{\mathrm{obs}}(w),$$

where $R_s^{\mathrm{obs}}(w)$ denotes the observed incidence. Analogous quantities are derived from the SLIRS predictions: $\hat{Y}_s^{\mathrm{SLIRS}}$ and $\hat{t}_s^{\mathrm{SLIRS}}$. The previous-season ILI peak I_s^{prev} is matched using the observed 44–51 week lag between ILI and RSV peaks.

2.2 ILI-Informed Peak Calibration

Exploratory analysis confirmed a strong interannual co-movement between Y_s and I_s^{prev}, as well as a systematic shift in t_s^{RSV} relative to $\hat{t}_s^{\mathrm{SLIRS}}$ that correlates with ILI activity. We therefore model the observed peak amplitude as a linear function of the SLIRS-predicted peak and the previous-season ILI peak:

$$Y_s = \beta_0 + \beta_1 \hat{Y}_s^{\mathrm{SLIRS}} + \beta_2 I_s^{\mathrm{prev}} + \epsilon_s, \tag{1}$$

with $\epsilon_s \sim \mathcal{N}(0, \sigma^2)$. The calibrated peak amplitude is then

$$\hat{Y}_s^{\mathrm{cal}} = \hat{\beta}_0 + \hat{\beta}_1 \hat{Y}_s^{\mathrm{SLIRS}} + \hat{\beta}_2 I_s^{\mathrm{prev}}.$$

A season-specific scaling factor $C_s = \hat{Y}_s^{\mathrm{cal}} / \hat{Y}_s^{\mathrm{SLIRS}}$ is applied to the entire SLIRS trajectory to preserve the within-season epidemic shape.

A similar linear model is used to correct peak timing. The observed timing error $E_s = t_s^{\mathrm{RSV}} - \hat{t}_s^{\mathrm{SLIRS}}$ is regressed against I_s^{prev}:

$$E_s = \alpha_0 + \alpha_1 I_s^{\mathrm{prev}} + \xi_s, \quad \xi_s \sim \mathcal{N}(0, \tau^2),$$

and the predicted shift $\hat{E}_s = \hat{\alpha}_0 + \hat{\alpha}_1 I_s^{\mathrm{prev}}$ is rounded to the nearest integer week.

The final calibrated trajectory is obtained by first shifting the SLIRS weekly prediction and then scaling it:

$$\hat{R}_s^{\mathrm{final}}(w) = C_s \cdot \hat{R}_s(w - \hat{E}_s),$$

with missing weeks at the boundaries omitted. This two-stage correction maintains the mechanistic epidemic shape while aligning the peak in both magnitude and timing with the ILI-driven interannual signal.

2.3 Autoregressive Residual Correction

The ILI-based calibration addresses interannual peak variability but does not explicitly model short-term temporal dependence in the residuals $e_s(w) = R_s^{\mathrm{obs}}(w) - \hat{R}_s^{\mathrm{final}}(w)$. Autocorrelation and partial autocorrelation functions of the SLIRS-only residuals reveal a strong autoregressive structure of order 1 (AR(1)). We therefore fit an AR(1) model directly to the residuals,

$$e_t = \phi\, e_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2),$$

and add the predicted $\hat{e}_t$ to the mechanistic output, producing a corrected prediction $\tilde{R}_t = \hat{R}_t^{\mathrm{final}} + \hat{e}_t$. This step captures the within-season persistence of prediction errors and further reduces short-term discrepancies, especially around the epidemic peak.

3 Results

3.1 Improvement in Peak Amplitude and Timing

Table 1 summarises the observed and predicted peak amplitudes across the five study seasons. The SLIRS model alone produces nearly flat peaks (range 1022–1054) that fail to track the observed interannual variation (range 1078–1460). After ILI-based calibration, the predicted peaks adjust markedly, reducing the mean absolute error (MAE) in peak amplitude from 200 to 95.2 cases. The addition of the timing correction further reduces the mean peak timing error to 0.8 weeks.

Table 1: Observed, SLIRS-predicted and calibrated RSV peak amplitudes. Calibration substantially reduces amplitude error and captures interannual variability.

Season	Observed Peak	SLIRS Peak	Calibrated	SLIRS Error	Calibrated Error
2015	1172	1022	1201	150	-29
2016	1108	1042	1068	66	40
2017	1460	1050	1384	410	76
2018	1401	1053	1279	348	122
2019	1078	1054	1287	24	-209

Figure 1 displays the full weekly trajectories. The SLIRS model (black dashed) reproduces the overall seasonal shape but misses the height and exact placement of the peak. The amplitude-only correction (orange) adjusts the magnitude, while the combined amplitude and timing correction (blue) aligns the peak week as well. The largest improvements occur in the high-incidence seasons 2017 and 2018, where the raw mechanistic model under-predicts the peak by 28–29%; the calibrated version reduces this to 5–9%.

3.2 Overall and Peak-Period Accuracy

To assess the practical value of the corrections, we computed the root-mean-square error (RMSE) over the whole epidemic curve and over the top 5% of observed values (the peak period). As shown in Table 2, the SLIRS model achieves the lowest overall RMSE (124.07) because its smooth oscillatory pattern fits the bulk of the non-peak weeks well. However, during the epidemiologically critical peak period, the calibrated models dramatically outperform the baseline. The combined amplitude+timing correction reduces the peak-period RMSE from 267.07 to 96.05, a 64% reduction. This trade-off—slightly higher overall error but much lower peak error—is acceptable for public health applications where anticipating the peak is the primary goal.

Table 2: Overall and peak-period RMSE for the different prediction strategies.

Model	RMSE (Overall)	RMSE (Peak Period)
SLIRS (baseline)	124.07	267.07
Amplitude calibration	145.99	115.43
Amplitude + timing calibration	172.87	96.05

3.3 Residual Autoregressive Structure

Autocorrelation analysis of the SLIRS residuals confirms strong, slowly decaying positive correlation and a dominant partial autocorrelation at lag 1. The Ljung-Box test rejects white noise (p ¡

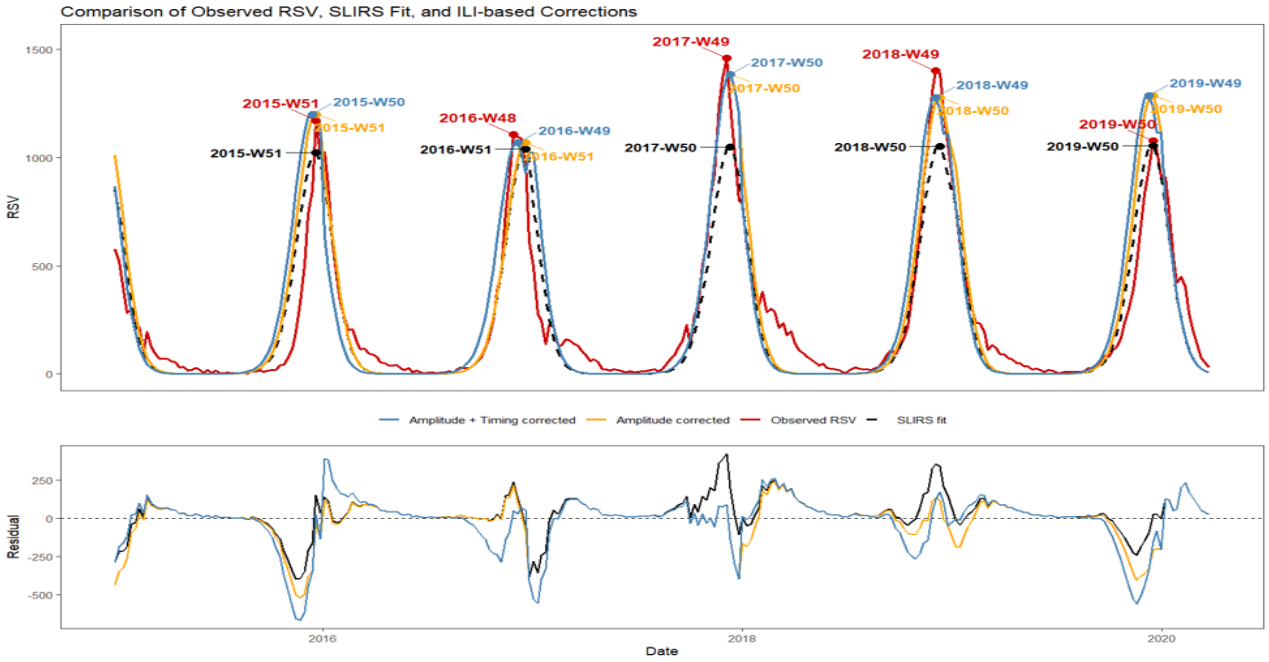


Figure 1: Observed RSV incidence (red), SLIRS baseline (black dashed), amplitude-calibrated (orange) and amplitude+timing calibrated (blue) predictions. The combined correction provides the best alignment of peak timing and magnitude.

0.001). An AR(1) model fitted to the residuals captures this dependence and, when added to the mechanistic output, further halves the overall RMSE (57.41 vs. 121.98) and reduces peak-period RMSE to 112.23. In contrast, expanding the AR model with ILI covariates (ARX) or time-varying volatility (GARCH) brings no additional improvement, indicating that the short-term errors are essentially autoregressive and not directly driven by ILI.

3.4 Multi-Scale Interpretation

The clear dichotomy between the ILI-based calibration (which corrects interannual peak variability) and the AR correction (which handles week-to-week persistence) reveals that RSV dynamics operate on at least two separable time scales. ILI activity, integrated over a 44–51 week lag, explains about 50–60% of the variance in next-season RSV peak magnitude, whereas short-term prediction errors are dominated by intrinsic autoregressive inertia. Combining the two approaches—first aligning the seasonal peak with ILI, then adjusting the weekly path with AR residuals—offers a simple yet powerful way to enhance mechanistic epidemic models without modifying their internal structure.

4 Conclusions

We have presented a lightweight, post-hoc correction framework that substantially improves the prediction of RSV epidemic peaks by exploiting two complementary sources of information: previous - season ILI activity and the autoregressive structure of mechanistic model residuals. The ILI-informed calibration successfully introduces interannual variability into otherwise rigid SLIRS predictions, reducing peak amplitude error by more than 50% and bringing peak timing to within one week of the observed value. The addition of an AR(1) residual model further reduces short-term

errors, particularly during the critical peak weeks.

Importantly, the two corrections operate on distinct temporal scales, revealing a multi-scale architecture of RSV transmission: seasonal variation is linked to broad ILI-related forcing, while week-to-week deviations reflect intrinsic persistence. This insight not only improves forecasting accuracy but also contributes to a better understanding of the drivers of RSV epidemic variability.

From a practical standpoint, the framework is easy to implement: it requires only the output of an existing mechanistic model and routinely available ILI surveillance data. As such, it can be readily integrated into real-time public health monitoring systems to support more responsive resource planning and intervention timing. Future work will extend the approach to other respiratory pathogens and evaluate its performance in prospective forecasting settings.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) RS-2025-23525071 to Jeon and partially supported by the Vicerrectorado de Investigación of the Universitat Politècnica de València PAID-11-25 grant.

References

[1] H. Nair et al., "Global burden of acute lower respiratory infections due to respiratory syncytial virus in young children: A systematic review and meta-analysis," *The Lancet*, vol. 375, no. 9725, pp. 1545–1555, 2010. DOI: 10.1016/S0140-6736(10)60206-1.

[2] T. Shi et al., "Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: A systematic analysis," *The Lancet*, vol. 390, no. 10098, pp. 946–958, 2017. DOI: 10.1016/S0140-6736(17)30938-8.

[3] V. E. Pitzer et al., "Environmental drivers of the spatiotemporal dynamics of respiratory syncytial virus in the united states," *PLoS Pathogens*, vol. 11, no. 1, e1004591, 2015. DOI: 10.1371/journal.ppat.1004591.

[4] N. C. Grassly and C. Fraser, "Mathematical models of infectious disease transmission," *Nature Reviews Microbiology*, vol. 6, no. 6, pp. 477–487, 2008. DOI: 10.1038/nrmicro1845.

[5] Y. Li, R. M. Reeves, X. Wang, et al., "Global patterns in monthly activity of influenza virus, respiratory syncytial virus, parainfluenza virus, and metapneumovirus: A systematic analysis," *The Lancet Global Health*, vol. 7, no. 8, e1031–e1045, 2019. DOI: 10.1016/S2214-109X(19)30264-5.

[6] S. Nickbakhsh et al., "Virus–virus interactions impact the population dynamics of influenza and the common cold," *Proceedings of the National Academy of Sciences*, vol. 116, no. 52, pp. 27 142–27 150, 2019. DOI: 10.1073/pnas.1911083116.

Predicting Patient and Graft Survival in Kidney Transplantation Using Machine Learning and Cox Regression

Laura Shalabayeva[1], Marc Romeu Ferras[1], Javier Díaz-Carnicero[1], María Ramos Cebrián[2], Alberto Budía Alba[2] and David Vivas-Consuelo[1]

(1) Centro de Investigación en Economía y Gestión de la Salud, CIEGS.
Univesitat Politècnica de València, UPV.
Camí de Vera, s/n, Edificio 7J, 3ª planta, Algirós, 46022 València, Valencia.
(2) Unidad de Litotricia y Endourología,
La Fe University and Polytechnic Hospital.
Av. Fernando Abril Martorell, nº 106, 46026 València, Valencia.

1 Introduction

Kidney transplantation is the most suitable treatment for the last stage of chronic kidney disease [1], which affects about 10% of the general population all over the world. This equals approximately 800 million individuals [2]. Also, there is a mismatch between the number of people who need a transplantation and the availability of transplants. For example, it was estimated in the UK that about 3,000 kidney transplantation surgeries are performed per year, while the number of patients who require this procedure is 5,000 [1]. This marks kidney transplants as a valuable resource, whose wastage cannot be permitted [3]. On the other hand, achieving long-term graft survival is still a challenge [4]. Thus, it is essential to develop a tool that would be able to predict the survival of kidney transplants.

Survival analysis is a subfield of statistics, capable of handling censored data, where the event of interest is not always observed due to time limitations or loss of follow-up. This approach allows estimating the time to the occurrence of the event of interest for new instances using their predictor variables [5]. One of the most widely used techniques in survival analysis is Cox regression. However, it is primarily suited to low-dimensional datasets and does not scale well to high-dimensional feature spaces. To address this limitation, machine learning (ML) methods can be employed [6].

ML is one of the branches of artificial intelligence (AI), which develops predictive models based on available data by automatically identifying patterns between clinical features and post-procedural outcomes [4, 7]. Currently, several studies have focused on developing such models for organ transplantation, showing promising results [8]. Thus, the objective of this study is to develop a predictive model for kidney transplant survival using available data.

[1] lshalab@upv.edu.es

2 Methods

2.1 Database

The database includes 525 kidney transplant cases treated at La Fe University and Polytechnic Hospital (Valencia, Spain), with a follow-up period of up to 15 years. Among them there are 112 cases of graft loss and 110 cases of exitus. These are 2 outcomes to predict in the study. There are several blocks of predictive features: donor-related characteristics, recipient-related characteristics, transplant variables, immunosuppressive therapy, intraoperative and post-transplant complications and events.

2.2 Data preprocessing

First, a data cleaning procedure was conducted. The following exclusion criteria were applied: predictive features with more than 60% missing data were removed; cases with more than 20% missingness across numerical features and non-nominal cathegorical were excluded; and cases with missing outcome labels were discarded.

According to Scheffner et al., the remaining missing values were imputed using the median for continuous numerical variables and the mode for discrete numerical and binary variables. Nominal categorical variables were encoded using one-hot encoding, and additional binary indicators were created to flag missing values.

2.3 Exploratory Data Analysis

Once the data have been cleaned and preprocessed, exploratory data analysis (EDA) is performed. Kaplan–Meier estimators are used to estimate graft and patient survival functions. Categorical variables are summarized using frequency distributions, while numerical variables are described using descriptive statistics, including mean, standard deviation, median and interquartile range.

2.4 Data splitting

The dataset was split into training and test subsets using an 80/20 ratio. Stratification based on the outcome event was applied.

2.5 Cox regression model development and evaluation

Cox regression is a semi-parametric method for survival analysis whose performance is commonly evaluated using the concordance index (C-index). This metric represents discrimination ability by estimating the probability that, for a randomly selected pair of individuals, the patient who experiences the event earlier is assigned a higher predicted risk by the model. The C-index ranges from 0 to 1, where 0.5 corresponds to a random guess and 1 indicates perfect discrimination [6].

Two Cox models were trained for each outcome, applying different strategies to improve feature selection. One strategy was based on sequentially discarding variables with the highest p-value, while the second selected the best set of variables by minimizing the Akaike Information Criterion (AIC). The effect of variables was assessed via the Hazard Ratio (HR) and p-value.

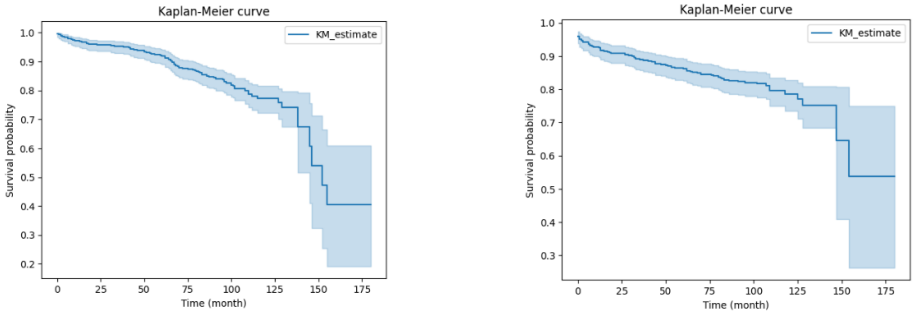
2.6 ML model development and evaluation

For ML model development, the following architectures were selected: Random Survival Forest (RSF) and Gradient Boosting Survival Analysis (GBSA). The models were evaluated with C-index, Integrated Brier Score (IBS), and area under the time-dependent receiver operating characteristic curve (AUC(t)).
Brier Score measures the error between the ground truth survival status at a specific time point and the survival probability predicted by the model, while IBS calculates this error along the observational time by applying integration. This metric can take values between 0 and 1, where lower values indicate better performance. AUC(t), similarly to the C-index, evaluates the discrimination power of the model at specific time points [6]. Normally, AUC values in the range between 0.7 and 0.8 are considered good performance, while AUC higher than 0.8 is considered excellent [3].
In order to select the hyperparameters RandomizedSearchCV python function was applied. To obtain the variables' importance permutation_importance function was used.

3 Results

3.1 Demographics

The database consists of 61.7% male patients and 38.3% female patients, with a median donor and recipient age of 54 years. The probability of patient survival at 5 and 10 years is 91.94% and 77.31%, respectively, while the corresponding probabilities for graft survival are 86.18% and 78.51%. Kaplan–Meier survival curves are shown in Figure 1, and a more detailed description of the database is provided in Table 1.



(**a**) (**b**)

Figure 1. Kaplan-Meier survival curves: (**a**) Exitus; (**b**) Graft loss.

Table 1. Database description.

Variable	Possible values with frequency and descriptive statistics for cathegorical and numerical variables, respectively
Donor age (years)	Mean: 52.4, std:12.63, min: 1.5, 25%: 50, 50%: 54, 75%: 56, max: 79
Donor gender	Male: 74.41%, Female: 25.59%
Donor hypertension	Yes: 12.11%, No: 87.89%
Donor diabetes	Yes: 4.3%, No: 95.7%
Donor obesity	Yes: 6.64%, No: 93.36%
Recipient age (years)	Mean: 51.84, std:13.81, min: 15, 25%: 41, 50%: 54, 75%: 63, max: 78
Recipient gender	Male: 67,1%, Female: 38,3%
Diabetes	At 1 year: 15.82%, Pre-transplant: 21.09%, No: 36.91%
Diabetes treated with insulin	Yes: 5.66%, No: 94.34%
Hepatitis B	Yes: 3.52%, No: 96.48%
Hepatitis C	Yes: 2.93%, No: 97.07%
HIV	Yes: 0.2%, No: 99.8%
Hypertension	Yes: 12.9%, No: 87.1%
Obesity	Yes: 20.31%, No: 79.69%
Dyslipidemia	Yes: 60.35%, No:39,65%
Transplant number	Mean: 1.13, std: 0.36, min: 1, 25%: 1, 50%: 1, 75%: 1, max: 4
Steroids	Yes: 99.61%, No: 0.39%
Initial calcineurin inhibitor (tacrolimus)	Yes: 100%
Initial antimet	Yes: 97.85%, No: 2.15%
Initial ImTOR	Yes: 3.32%, No: 96.67%
Delayed graft function	Yes: 65.04%, No: 34.96%
Need for dialysis	Yes: 26.95%, No: 73.05%
Acute rejection	Yes: 17.97%, No: 82.03%
Chronic rejection	Yes: 4.49%, No: 95.51%
Intraoperative complication	Yes: 9.57%, No: 90.43%
Early surgical complication (<30 days)	Yes: 25.78%, No: 74.22%
Late surgical complication (>30 days)	Yes: 26.37%, No: 73.63%
Immunological complication	Yes: 16.02%, No: 83.98%
Opportunistic infection	Yes: 27.73%, No: 72.27%
Cardiovascular event	Yes: 16.21%, No: 83.79%
Tumors	Yes: 18.36%, No: 81.64%
Cause of donor death	Stroke: 0.2%, Hemorrhagic stroke: 26.95%, Ischemic stroke: 7.81%, Living donor: 0.59%, Voluntary donor: 2.34%, Encephalopathy: 2.73%, Myocardial infarction: 0.98%, CNS infection / meningitis: 0.2%, Cardiorespiratory arrest: 0.2%, Traumatic brain injury: 11.52%, Pulmonary thromboembolism: 0.2%, Cerebral anoxia: 0.2%, Hydrocephalus – Arnold-Chiari malformation: 0.2%, Pneumococcal meningitis: 0.2%, Polytrauma: 0.2%, Other: 2.54%, NaN: 42.97%
Blood group	O: 41.41%, A: 46.29%, AB: 2.93%, B: 8.79%, NaN: 0.59%
Combined	Bilateral: 0.98%, Hepatorenal: 2.54%, Pancreas-kidney: 8.01%, No: 88.48%
Transplant donor type	Deceased donor: 89.84%, DCD type 3: 0.2%, Living donor: 8.59%, NaN: 1.37%
Nephropathy	Renal dysplasia: 0.2%, Glomerular disease: 23.44%, Hereditary: 3.32%, Infectious: 0.2%, Diabetic: 13.28%, Diabetic nephropathy and unilateral nephrectomy:

	0.2%, Interstitial: 4.3%, End-stage: 6.45%, Nephrectomies: 0.2%, Unknown etiology : 25.39%, ADPKD: 12.5%, Chronic pyelonephritis: 0.39%, Vesicoureteral reflux: 1.37%, Systemic: 5.86%, Tubulointerstitial: 0.2%, Calcineurin inhibitor toxicity: 0.2%, Other: 1.37%, NaN: 1.17%
Dialysis	PD: 23.05%, HD: 60.35%, HD+PD: 11.33%, Predialysis: 3.12%, NaN: 2.15%
Induction	Basiliximab: 72.27%, Daclizumab: 0.39%, Thymoglobulin: 8.59%, No: 0.2%, NaN: 5.86%
Tacrolimus	Advagraf: 75.78%, Prograf: 18.36%, NaN: 5.86%
Type of initial antimet	Cellcept: 65.43%, Myfortic: 26.37%, NaN: 8.2%

3.2 Cox regression models

The discriminative ability of the developed Cox models is summarized in Table 2. For the Exitus outcome, both models showed good and stable performance, with slightly better discrimination for the model selected by AIC. In contrast, for graft loss, the model based on p-value selection outperformed the AIC selection model.

Table 2. C-index of Cox models.

C-index	Exitus		Graft loss	
	Variable selection by p-value	**Variable selection by AIC**	**Variable selection by p-value**	**Variable selection by AIC**
Train	0.76	0.78	0.77	0.81
Test	0.79	0.80	0.65	0.53

The variable importance of the best-performing Cox models is illustrated in the forest plots (Figure 2 and 3).

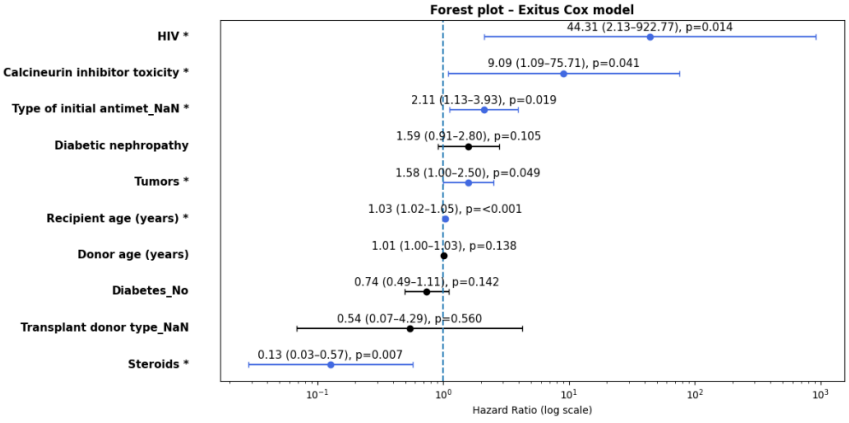


Figure 2. Forest plot of the best Exitus Cox model. Hazard ratios are shown with 95% confidence intervals. Variables with p < 0.05 are highlighted in blue and marked with an asterisk (*).

For exitus, increased risk was associated with HIV, calcineurin inhibitor toxicity, diabetic nephropathy, tumors, and older donor and recipient age (HR > 1). Additionally, missing values

for the type of initial antimetabolite were associated with higher risk. Protective factors (HR < 1) included absence of diabetes, missing donor type, and steroid use.

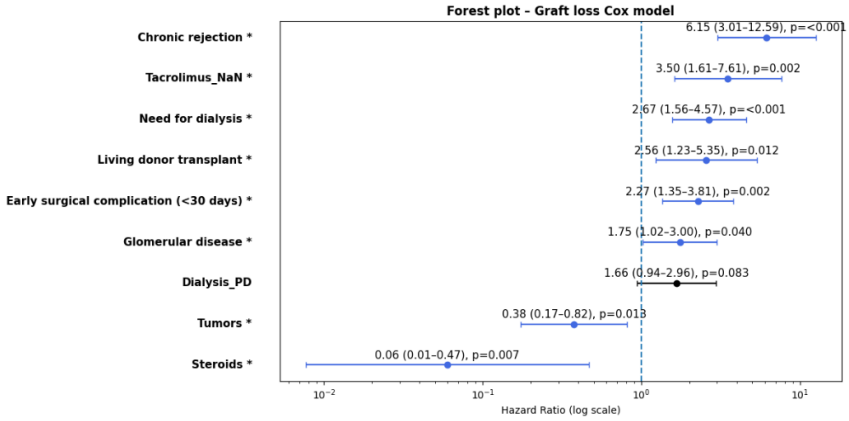


Figure 3. Forest plot of the best Graft loss Cox model. Hazard ratios are shown with 95% confidence intervals. Variables with $p < 0.05$ are highlighted in blue and marked with an asterisk (*).

For graft loss, the main risk factors were chronic rejection, missing tacrolimus values, need for dialysis, living donor transplantation, early surgical complications (<30 days), glomerular disease, and peritoneal dialysis. Tumors and steroid use were associated with a reduced risk. Although the apparent protective effect of tumors may seem contradictory, Beviá-Romero et al. described the same effect and attributed it to survivorship bias, as tumors tend to be diagnosed in patients with longer follow-up, likely due to prolonged exposure to immunosuppressive therapy, rather than representing a true protective factor.

3.3 ML models

ML models performance metrics are summarized at Table 3.

Table 3. ML models performance metrics.

Metrics	Exitus		Graft loss	
	RSF	**GBSA**	**RSF**	**GBSA**
Train C-index	0.796	0.808	0.800	0.815
Test set				
C-index	0.787	0.756	0.665	0.628
IBS	0.088	0.085	0.121	0.122
AUC at 1 year	0.587	0.483	0.816	0.711
AUC at 2 years	0.602	0.535	0.824	0.732
AUC at 3 years	0.604	0.539	0.703	0.569
AUC at 5 years	0.760	0.714	0.687	0.594
AUC at 7 years	0.875	0.837	0.682	0.597
AUC at 10 years	0.835	0.830	0.748	0.767

Exitus models showed good discrimination, with C-index values around 0.80 in training and test sets (RSF: 0.796/0.787; GBSA: 0.808/0.756), and low prediction error (IBS: 0.088 and 0.085, respectively). Time-dependent AUC analysis revealed limited predictive performance in the first years after transplantation (AUC ≈ 0.48–0.60), but substantially improved discrimination from year 5 onwards, reaching values above 0.7. RSF showed more stable performance across datasets, while GBSA exhibited near-random discrimination in the first three years. Overall, both models demonstrated stronger predictive capacity in the mid- to long-term follow-up period. Figure 4 displayed top 20 predictors of Exitus RSF model, being recipient and donor age and diabetes-related variables main predictors. These main predictors also were labeled as important by Cox model.

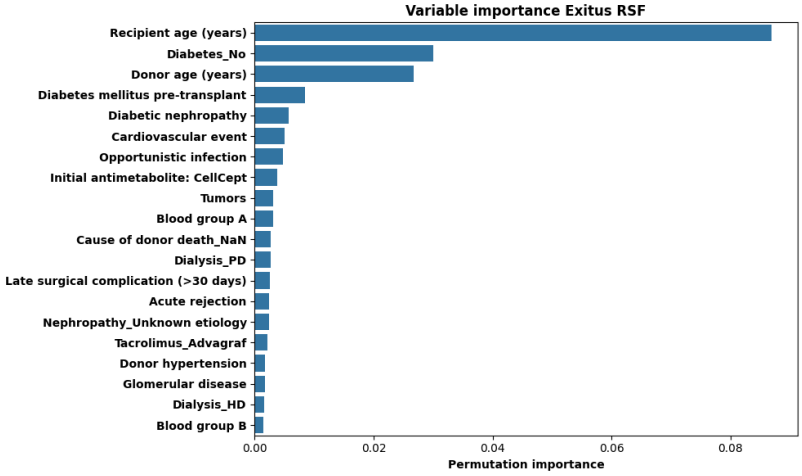


Figure 4. Permutation feature importance for the top 20 predictors identified by Exitus RSF model.

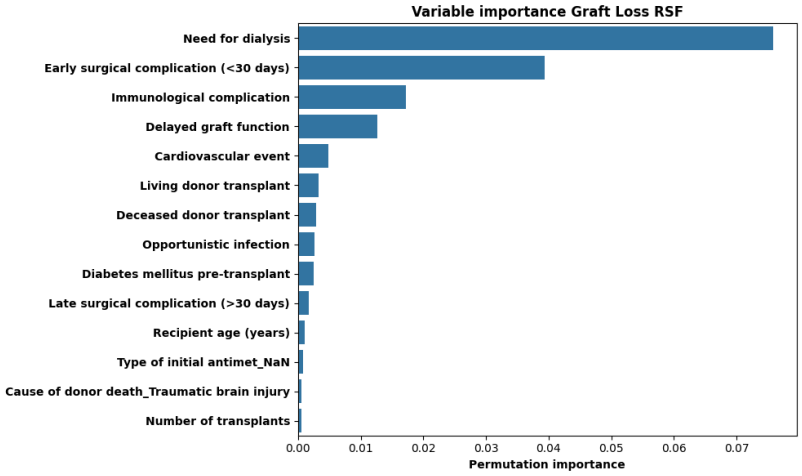


Figure 5. Graft loss RSF model: Variables with positive permutation importance score.

For graft loss, both RSF and GBSA models showed signs of overfitting, with better performance in the training set than in the test set, suggesting limited generalizability. The integrated Brier score remained acceptable for both models. Time-dependent AUC indicated stronger discrimination in early and late follow-up, with a decline in intermediate periods. Overall, RSF outperformed GBSA in AUC and C-index. According to Figure 5 postoperative need for dialysis, early surgical complications, immunological complications and delayed graft function are main predictors, with the first two also selected for the final Cox model.

4 Conclusions

In kidney transplant patients, ML models showed comparable performance to the Cox model, which is conventionally used for survival analysis. In addition, ML-based models demonstrated utility in identifying key variables and exhibited time-dependent performance, supporting their use for early prediction of graft loss and long-term mortality.

References

[1] S. Riley, Q. Zhang, W.-Y. Tse, A. Connor, and Y. Wei, "Using Information Available at the Time of Donor Offer to Predict Kidney Transplant Survival Outcomes: A Systematic Review of Prediction Models". *Transplant International*, Vol. 35, Jun. 2022. doi: 10.3389/ti.2022.10397.

[2] C. P. Kovesdy, "Epidemiology of chronic kidney disease: an update 2022". Apr. 01, 2022. Elsevier B.V. doi: 10.1016/j.kisu.2021.11.003.

[3] H. Ali et al., "Improved survival prediction for kidney transplant outcomes using artificial intelligence-based models: development of the UK Deceased Donor Kidney Transplant Outcome Prediction (UK-DTOP) Tool". *Ren Fail*, Vol. 46, no. 2, 2024. doi: 10.1080/0886022X.2024.2373273.

[4] G. K. Aksoy et al., "Predicting graft survival in paediatric kidney transplant recipients using machine learning". *Pediatric Nephrology*, Vol. 40, no. 1, pp. 203–211, Jan. 2025. doi: 10.1007/s00467-024-06484-5.

[5] P. Wang, Y. Li, and C. K. Reddy, "Machine Learning for Survival Analysis". *ACM Computing Surveys*, Vol. 51, no. 6, pp. 1–36, Feb. 2019. doi:10.1145/3214306

[6] A. Sarica *et al.*, "Explainability of random survival forests in predicting conversion risk from mild cognitive impairment to alzheimer's disease". *Brain Informatics*, Vol. 10, no. 1, Nov. 2023. doi:10.1186/s40708-023-00211-w

[7] S. A. Naqvi, K. Tennankore, A. Vinson, P. C. Roy, and S. S. Abidi, "Predicting kidney graft survival using machine learning methods: Prediction model development and feature Significance Analysis Study". *Journal of Medical Internet Research*, Vol. 23, no. 8, Aug. 2021. doi:10.2196/26843

[8] D. B. Olawade, S. Marinze, N. Qureshi, K. Weerasinghe, and J. Teke, "The impact of Artificial Intelligence and machine learning in organ retrieval and transplantation: A comprehensive review". *Current Research in Translational Medicine*, Vol. 73, no. 2, p. 103493, Apr. 2025. doi:10.1016/j.retram.2025.103493

[9] I. Scheffner, M. Gietzelt, T. Abeling, M. Marschollek, and W. Gwinner, "Patient survival after kidney transplantation: Important role of graft-sustaining factors as determined by predictive modeling using Random Survival Forest Analysis". *Transplantation*, Vol. 104, no. 5, pp. 1095–1107, May 2020. doi:10.1097/tp.0000000000002922

[10] Á. Beviá-Romero et al., "Kidney Transplant: Survival Analysis and Prognostic Factors after 10 Years of Follow-Up". *Mathematics*, Vol. 11, no. 7, 2023. doi: 10.3390/math11071640.

Describing the Transmission Dynamics of Hepatitis A in Children Under One Year of Age: An Age-Structured Model with Uncertainty

Clara Burgos Simón [♭,1] Cristina Valcárcel Nazco[♮] Carmen Guirado Fuentes[♮] and Rafael Jacinto Villanueva Micó [♭]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València Camí de Vera s/n, València, Spain.

(♮) Fundación Canaria Instituto de Investigación Sanitaria de Canarias (FIISC), Barranco de la Ballena, s/n, Edificio Anexo al Hospital Universitario de Gran Canaria Dr. Negrín, 35019 Las Palmas de Gran Canaria.

1 Introduction

Hepatitis A is an acute infectious liver disease caused by the Hepatitis A virus (HAV), an RNA virus of the *Picornaviridae* family, mainly transmitted via the fecal–oral route through direct contact or contaminated food and water [1], [2]. Although HAV infection does not lead to chronic disease, its clinical severity strongly depends on age: it is often asymptomatic in early childhood, while in adolescents and adults it is more likely to cause symptomatic illness and complications [3].

Epidemiological patterns of Hepatitis A vary worldwide depending on socioeconomic conditions and vaccination strategies [4]. In low-endemicity countries such as Spain, a shift of infection toward older age groups has been observed, increasing the clinical burden and favoring the occurrence of outbreaks [2], [5]. This change has motivated the study of HAV transmission dynamics and the evaluation of control measures.

In this context, the age structure of the population plays a key role. In this work, two age groups are considered: children under 1 year of age and individuals older than 1 year. This classification is motivated by the age-dependent clinical expression of the disease and the fact that infections in young children are frequently asymptomatic, which may lead to underreporting [2]. Moreover, routine vaccination is typically administered from 12 months of age, making the analysis of transmission in children under 1 year particularly relevant.

The aim of this work is to develop a mathematical model describing the transmission dynamics of Hepatitis A in Spain, with a special focus on children under 1 year of age. To this end, a random modeling framework is proposed to account for the inherent uncertainty in both transmission mechanisms and epidemiological data.

The data used in this study have been provided by the Centro Nacional de Epidemiología and correspond to annual reported cases of Hepatitis A in Spain, disaggregated by autonomous communities and age groups. These data have been regrouped into the two age categories considered in the model and aggregated at the national level. Figure 1 shows the annual reported cases for each group, highlighting both the overall trend and the presence of epidemic outbreaks.

[1] clabursi@upv.es

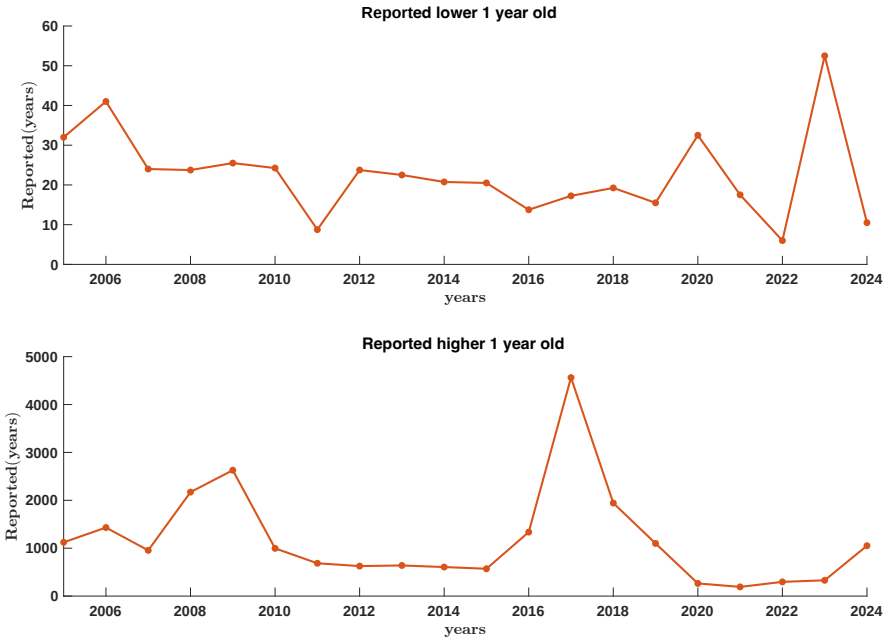


Figure 1: Annual reported cases of Hepatitis A in Spain for the two age groups considered: under 1 year of age and over 1 year of age.

2 Mathematical model

In this work, the transmission dynamics of Hepatitis A is described by means of a two-age-group SEIR compartmental model. The population is divided into two age classes: children under 1 year of age (G_1) and individuals older than 1 year (G_2). Since the available data are given on an annual basis, the time variable t is measured in years.

For each age group G_i, $i = 1, 2$, the total population is divided into susceptible (S_i), latent (L_i), infectious (I_i), and recovered (R_i) individuals.

As indicated in [1], [2], Hepatitis A is transmitted through direct contact with infected individuals. At time $t > 0$, a susceptible individual belonging to $S_1(t)$ or $S_2(t)$ becomes latent, entering $L_1(t)$ or $L_2(t)$, respectively, after contact with an infectious individual from either of the two infectious subpopulations $I_1(t)$ or $I_2(t)$.

We denote by β_1 and β_2 the transmission rates associated with the first age group, describing the infection of susceptible individuals in $S_1(t)$ due to contact with infectious individuals in $I_1(t)$ and $I_2(t)$, respectively. Thus, the incidence term governing the transition from $S_1(t)$ to $L_1(t)$ is given by

$$S_1(t)\left(\beta_1 I_1(t) + \beta_2 I_2(t)\right).$$

Similarly, β_3 and β_4 denote the transmission rates associated with the second age group, accounting for the infection of individuals in $S_2(t)$ due to contact with infectious individuals from both age groups. In this case, the corresponding transition term from $S_2(t)$ to $L_2(t)$ is

$$S_2(t)\left(\beta_3 I_1(t) + \beta_4 I_2(t)\right).$$

Next, we focus on the dynamics of the latent compartment. We denote by σ the rate at which latent individuals L_i progress to the infectious class I_i, which is assumed to be the same for both age groups ($i = 1, 2$). Therefore, the term modeling this transition is $\sigma L_i(t)$. As reported in [6], an

individual remains in the latent state for approximately two weeks. Since the model is formulated on an annual time scale, this value is translated into $\sigma = 52/2 = 26$ year^{-1}.

Once an individual enters the infectious state, the average duration of this stage is about 38 days, independently of the age group [6]. We denote by γ the recovery rate from the infectious compartment I_i to the recovered compartment R_i, $i = 1, 2$, and therefore $\gamma = 365/38$. The corresponding transition term is given by $\gamma I_i(t)$.

Taking all these mechanisms into account, the two-age-group SEIR model describing the transmission dynamics of Hepatitis A is given by

$$\begin{aligned}
\frac{dS_1(t)}{dt} &= -\left(\frac{\beta_1 I_1(t)}{N_1} + \frac{\beta_2 I_2(t)}{N_2}\right) S_1(t) + \mu - (d_1 + c)\, S_1(t),\\
\frac{dL_1(t)}{dt} &= \left(\frac{\beta_1 I_1(t)}{N_1} + \frac{\beta_2 I_2(t)}{N_2}\right) S_1(t) - \sigma L_1(t) - (d_1 + c)\, L_1(t),\\
\frac{dI_1(t)}{dt} &= \sigma L_1(t) - \gamma I_1(t) - (d_1 + c)\, I_1(t),\\
\frac{dR_1(t)}{dt} &= \gamma I_1(t) - (d_1 + c)\, R_1(t),\\
\frac{dS_2(t)}{dt} &= -\left(\frac{\beta_3 I_1(t)}{N_1} + \frac{\beta_4 I_2(t)}{N_2}\right) S_2(t) - d_2\, S_2(t) + c\, S_1(t),\\
\frac{dL_2(t)}{dt} &= \left(\frac{\beta_3 I_1(t)}{N_1} + \frac{\beta_4 I_2(t)}{N_2}\right) S_2(t) - \sigma L_2(t) - d_2\, L_2(t) + c\, L_1(t),\\
\frac{dI_2(t)}{dt} &= \sigma L_2(t) - \gamma I_2(t) - d_2\, I_2(t) + c\, I_1(t),\\
\frac{dR_2(t)}{dt} &= \gamma I_2(t) - d_2\, R_2(t) + c\, R_1(t).
\end{aligned} \tag{1}$$

3 Random fitting

The aim of this section is to determine the parameters of model (1) that best fit the data shown in Figure 1. To this end, we perform a random fitting procedure. More precisely, the model is calibrated in a stochastic manner following the procedure described below.

First, using Particle Swarm Optimization (PSO), we perform a deterministic fitting of the model. During this process, we store all the evaluations of the objective function, together with their corresponding parameter sets, generated by the PSO algorithm until convergence to the minimum. In this way, instead of obtaining a single optimal solution, we identify the region in the parameter space where the global minimum is attained. This information is essential for constructing a reliable random fitting.

Among all the stored evaluations (and their associated parameter sets), many of them may only contribute noise to the fitting process. Therefore, the next step consists of selecting those parameter sets that provide relevant information. To this end, we sort the parameter sets according to their error (i.e., the value of the objective function to be minimized), and progressively select them—starting from the lowest error—until the available data are fully covered.

However, this procedure may still involve a large number of parameter sets, leading to a high computational cost. Moreover, many of these parameter sets may lie within the same region of the parameter space and thus provide redundant information. To address this issue, we cluster the parameter sets and, by means of a Latin Hypercube Sampling strategy, we select those that are more dispersed within the parameter space.

A known drawback of this type of approach is the potentially high variance, which may lead to wide prediction intervals. For this reason, we further construct subsets of parameter sets and

select the one that, on the one hand, exhibits lower variance and, on the other hand, maximizes the number of observed data points contained within its confidence interval.

4 Results

In this section, we present the results obtained using the random fitting procedure described above. Figure 2 shows the resulting fit. The red dots represent the reported data, the blue line corresponds to the mean of the selected outputs, and the shaded blue region denotes the 95% confidence interval associated with these outputs.

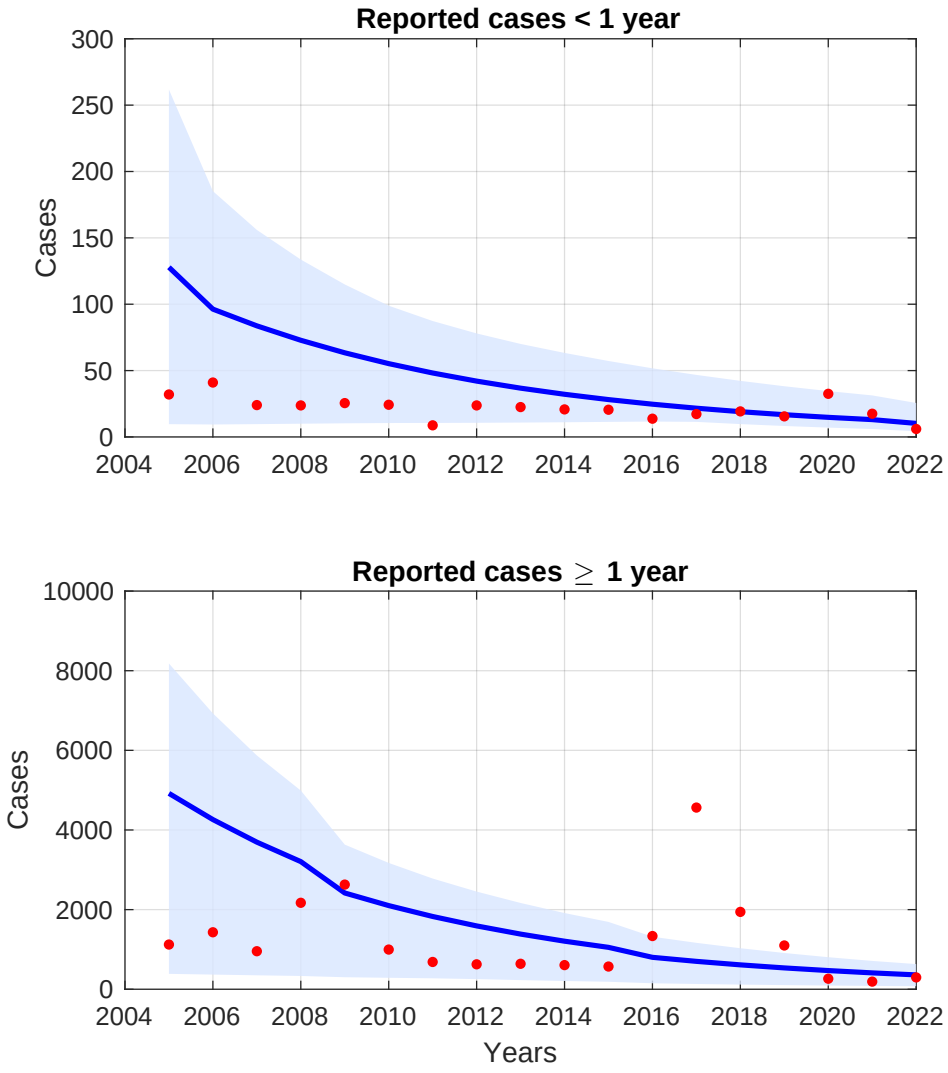


Figure 2: Random fitting results for the Hepatitis A model. The red dots represent the reported data, the blue line corresponds to the mean of the selected outputs, and the shaded region denotes the 95% confidence interval.

5 Conclusions

In this work, a two-age-group SEIR model has been proposed to describe the transmission dynamics of Hepatitis A in Spain, with special emphasis on the population under 1 year of age. The model incorporates the main epidemiological mechanisms of the disease and accounts for the interaction between different age groups, which is essential to capture the observed transmission patterns.

A random fitting procedure has been developed to calibrate the model parameters using epidemiological data provided by the Centro Nacional de Epidemiología. This approach allows not only the identification of optimal parameter values, but also the characterization of the region in the parameter space that provides a good fit to the data.

The results show that the proposed methodology provides a satisfactory fit to the reported data, while also quantifying the uncertainty associated with the model outputs through confidence intervals. This is particularly relevant in the context of Hepatitis A, where underreporting and variability in transmission play a significant role.

Acknowledgments

This work has been funded by: Vicerrectorado de Investigación of the Universitat Politècnica de València PAID-11-25 and Grant CIGE/2023/79, Generalitat Valenciana (Spain).

References

[1] E. Franco, C. Meleleo, L. Serino, D. Sorbara, and L. Zaratti, "Hepatitis a: Epidemiology and prevention in developing and developed countries," *World Journal of Hepatology*, vol. 4, no. 3, pp. 68–73, 2012. DOI: 10.4254/wjh.v4.i3.68.

[2] World Health Organization, *Hepatitis a*, https://www.who.int/news-room/fact-sheets/detail/hepatitis-a, Accessed 2026, 2019.

[3] S. M. Lemon, J. J. Ott, P. Van Damme, and D. Shouval, "Type a viral hepatitis: A summary and update on the molecular virology, epidemiology, pathogenesis and prevention," *Journal of Hepatology*, vol. 68, no. 1, pp. 167–184, 2018. DOI: 10.1016/j.jhep.2017.08.034.

[4] World Health Organization, "Who position paper on hepatitis a vaccines – june 2012," *Weekly Epidemiological Record*, vol. 87, no. 28, pp. 261–276, 2012.

[5] K. H. Jacobsen and J. S. Koopman, "The effects of socioeconomic development on worldwide hepatitis a virus seroprevalence patterns," *International Journal of Epidemiology*, vol. 34, no. 3, pp. 600–609, 2005. DOI: 10.1093/ije/dyi062.

[6] C. Valcárcel-Nazco, E. Sanromá-Ramos, L. García-Pérez, R. J. Villanueva-Micó, C. Burgos-Simón, and J. Mar, "Coste-efectividad de la vacunación universal infantil frente a la hepatitis a en españa: Un enfoque dinámico," *Gaceta Sanitaria*, vol. 37, p. 102 292, 2023. DOI: 10.1016/j.gaceta.2023.102292. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0213911123000067.

Neural ODE-Based Longitudinal Forecasting of Pulmonary Disease Progression in Chest Radiographs

Gonzalo Atienza[♭1], Rocío Del Amor[♭♮] and Fernando Lopez[♭♮]

(♭) I.U. de Investigación en Bioingeniería y Tecnología Orientada al Ser Humano, Universitat Politècnica de València, Camí de Vera s/n, València, Spain.
(♮) Artikode Intelligence S.L., Valencia, Spain

1 Introduction

The proposed framework aims to model the temporal evolution of chest radiographs as a continuous trajectory in a high-dimensional latent space. Given an initial image observation at time zero, the goal is to predict the state of the image at any future time point by solving an initial value problem. This approach treats disease progression not as a sequence of static states but as a dynamic process governed by underlying biological laws.

Traditionally, deep learning models for time series analysis in medicine have relied on discrete architectures such as Recurrent Neural Networks or Long Short-Term Memory networks that assume fixed time intervals or require prior discretization of the underlying biological process. This approach is often insufficient to capture the continuous nature of anatomical and pathological changes. In this context, Neural Ordinary Differential Equations, also known as Neural ODEs [1], have emerged as a powerful paradigm, allowing the derivative of the system state to be parameterized by a neural network. This framework enables the modeling of disease dynamics as a continuous flow in a representation space, facilitating inference at any temporal point regardless of the original sampling frequency.

This work builds upon the ImageFlowNet framework [2], which extends Neural ODEs to the medical image domain through a multiscale architecture. While the original framework was validated in domains such as ophthalmology and neuroimaging, this research explores its capacity to model pulmonary dynamics in chest radiography. To this end, a massive database has been consolidated and normalized from the public repositories PADCHEST [3] and MIMIC-CXR [4], employing Large Language Models LLMs [5] for the semantic standardization of the associated medical reports.

The main objective is to predict the future visual appearance of a radiograph from an initial image, capturing the structural transformations of the lungs and the cardiac area. In this article, we present preliminary results demonstrating the convergence of the model and its ability to generate visually coherent and structurally accurate trajectories, as measured by image fidelity metrics

[1] gonatsel@cam.upv.es

such as Peak Signal to Noise Ratio PSNR and Structural Similarity Index SSIM [6]. Finally, the implications of these results are discussed as a basis for future multimodal systems that integrate clinical language into the conditioning of predictions, allowing for a more personalized and explainable prognosis simulation.

2 Related Work

The field of medical imaging has experienced a paradigm shift with the advent of generative neural networks which allows models to learn the underlying probability distributions of complex clinical data. Early research focused on generative adversarial networks (GANs) and variational autoencoders (VAEs) to synthesize realistic anatomical structures and mitigate the challenges of data scarcity in specialized domains. While these architectures established the feasibility of synthetic medical imaging they often suffered from training instabilities such as mode collapse where the generator produces a limited variety of outputs. More recently denoising diffusion probabilistic models (DDPMs) have emerged as the state of the art for high fidelity synthesis by employing a gradual denoising process to transform Gaussian noise into structured images [7]. However applying these models to longitudinal clinical data remains difficult because standard diffusion frameworks do not inherently preserve patient identity or anatomical consistency over long temporal sequences.

Predicting disease progression requires moving beyond static image synthesis toward longitudinal modeling that accounts for temporal dynamics. Traditional sequence modeling often relies on recurrent architectures that assume regular time intervals between observations which is rarely the case in real world clinical practice. To handle irregularly sampled data researchers have turned to deformation based models that learn spatial transformations to map anatomical changes over time [8]. By treating these transformations as a continuous process neural ordinary differential equations ODEs allow for the parameterization of velocity fields that define the trajectory of a patient within a latent manifold [1]. This approach ensures that the predicted evolution is smooth and mathematically consistent with the underlying biological processes. Frameworks like ImageFlowNet address this by integrating multiscale anatomical transformations with temporal and semantic conditioning to provide a complete picture of disease evolution [2]. Furthermore self supervised strategies like SimSiam are increasingly utilized to organize the representation space ensuring that images from the same longitudinal series remain proximal and coherent in the latent manifold [9].

The ambition of this project is to move beyond static image synthesis by exploring a framework for continuous and personalized forecasting. By integrating multiscale latent dynamics with semantic report standardization, we aspire to develop a methodology capable of modeling complex pathological trajectories that maintain anatomical consistency, even across irregular clinical time intervals.

3 Mathematical Framework

3.1 Multiscale Latent Dynamics

The model describes the progression of image features as a continuous-time process within a set of joint representation spaces. For each resolution scale k, the evolution of the latent state $\mathbf{z}^k$ is defined by a position-parameterized neural ordinary differential equation [1]. The rate of change in the manifold is determined by a velocity field f_θ^k, as shown in the following expression:

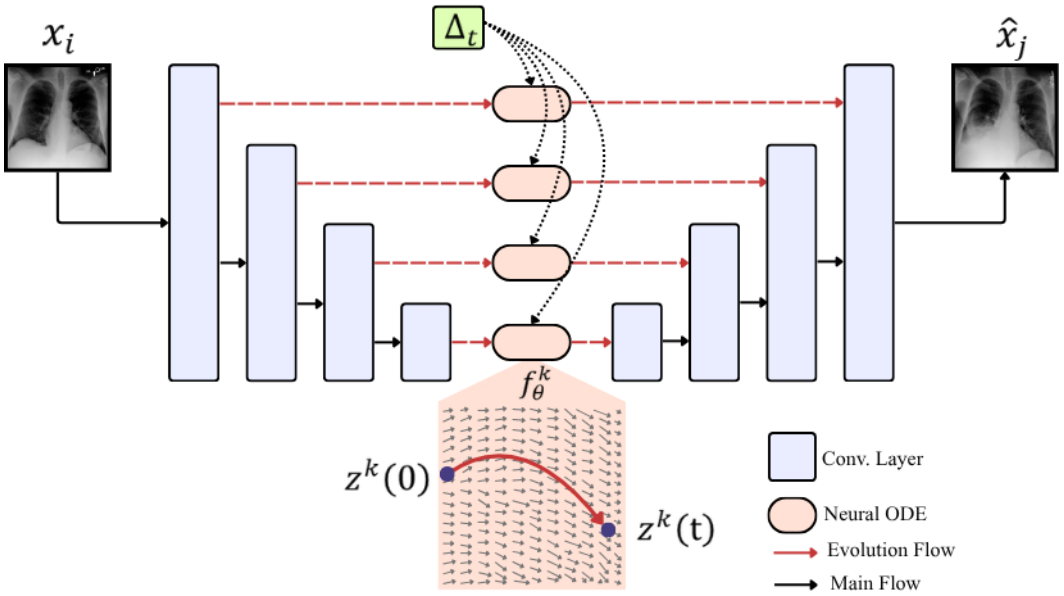


Figure 1: The architectural design follows a multiscale convolutional encoder-decoder structure where longitudinal dynamics are captured by convolutional layers parameterizing the velocity field f_θ^k within the skip connections. These layers define the rate of change for the latent state $z^k(0)$ according to the specific time interval Δ_t. An ODE solver then utilizes this learned field to compute the evolved representation $z^k(t)$ as defined in the mathematical framework. This structural modification allows the decoder to reconstruct the future anatomy by integrating these time conditioned features across all hierarchical scales.

$$\frac{d\mathbf{z}^k(t)}{dt} = f_\theta^k(\mathbf{z}^k(t)) \tag{1}$$

To compute the state of the latent representation at a target time point t, the framework integrates the learned velocity field starting from the initial state at time 0. This integration effectively accumulates the infinitesimal transformations learned by the network along the trajectory. The resulting latent state is formally expressed through the following integral formulation:

$$\mathbf{z}^k(t) = \mathbf{z}^k(0) + \int_0^t f_\theta^k(\mathbf{z}^k(\tau))\, d\tau \tag{2}$$

3.2 Two-Stage Training and Trajectory Integration

The training of the framework is conducted in two distinct phases to ensure stability and representational fidelity [2]. In the first stage, the multiscale autoencoder is trained for reconstruction. The encoder and decoder are optimized to minimize the discrepancy between the input and the reconstructed image, effectively defining a fixed joint representation space that captures the essential morphology of the thorax. This phase ensures that the latent manifold possesses the necessary geometric properties to represent chest radiographs accurately across different scales, leveraging architectures such as U-Net [10].

In the second stage, the autoencoder weights can either be frozen or fine-tuned with a reduced learning rate to preserve the integrity of the latent space, while the optimization focuses on the

velocity fields f_θ^k. During this phase, the model learns to map transitions between latent states over time. The velocity field at each scale is implemented as a sequence of convolutional layers that parameterize the gradient space of these latent transitions. The resulting latent states are then passed through the decoder to synthesize the predicted future radiograph. This separation allows the velocity field to focus on modeling progression dynamics without altering the underlying anatomical representation.

3.3 Comprehensive Optimization Objective

The training process is governed by a multi-task objective function designed to ensure representational fidelity, semantic consistency, and temporal smoothness. The total optimization objective is defined as:

$$\mathcal{L} = \mathcal{L}_r + \lambda_v \mathcal{L}_v + \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s \tag{3}$$

The reconstruction loss $\mathcal{L}_r$ employs Mean Squared Error to minimize pixel-wise differences. The visual feature term $\mathcal{L}_v$ aligns high-level semantic representations using an encoder pre-trained on ImageNet such as ConvNeXT [11]. The contrastive term $\mathcal{L}_c$ follows a SimSiam formulation [9] to enforce consistency within longitudinal trajectories. Finally, the smoothness term $\mathcal{L}_s$ regularizes the velocity field to ensure stable and smooth temporal evolution. The full loss is defined as:

$$\begin{aligned}
\mathcal{L} = & \frac{1}{HWC} \sum_{h \in H} \sum_{w \in W} \sum_{c \in C} \|\hat{x}_j[h,w,c] - x_j[h,w,c]\|_2^2 \\
& + \lambda_v \left(- \frac{e(\hat{x}_j)^\top e(x_j)}{\|e(\hat{x}_j)\|_2 \, \|e(x_j)\|_2} \right) \\
& + \lambda_c \left(- \frac{p_d(p_j(z_{t_i}))^\top \, p_j(z_{t_j})}{2 \, \|p_d(p_j(z_{t_i}))\|_2 \, \|p_j(z_{t_j})\|_2} - \frac{p_d(p_j(z_{t_j}))^\top \, p_j(z_{t_i})}{2 \, \|p_d(p_j(z_{t_j}))\|_2 \, \|p_j(z_{t_i})\|_2} \right) \\
& + \lambda_s \, \|f_\theta\|_2^2
\end{aligned} \tag{4}$$

4 Methodology and Results

4.1 Longitudinal CXR registration

To establish a unified longitudinal framework, the MIMIC-CXR [4] and PadChest [3] datasets were homogenized by organizing patient studies chronologically and retaining only individuals with at least two valid examinations. Images were processed in grayscale, resized to the target resolution, and normalized to ensure representational consistency across both data sources. From these series, temporal pairs were generated by defining the earlier study as the fixed image and the subsequent study as the moving image.

The registration pipeline utilized the Advanced Normalization Tools ANTs framework to perform a hierarchical non-rigid registration, beginning with an affine alignment followed by a Symmetric Normalization SyN refinement [12], [13]. To preserve clinically relevant structures and prevent excessive warping of pathological regions, a coarse anatomical guidance mask was constructed. This mask integrated lung regions, a pathology aware exclusion mask, and a central anchor region to define a healthy tissue mask used during all registration stages. Optimization was driven by masked similarity, ensuring that the transformation was estimated primarily from stable anatomical features while minimizing the influence of abnormal regions. This strategy resulted in thoracic alignments that were visually consistent and limited unwanted local deformations around the lungs

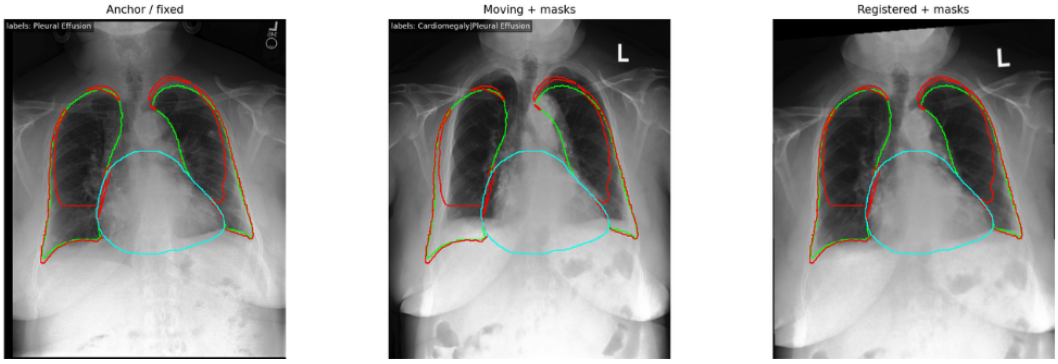


Figure 2: Longitudinal chest X-ray registration with anatomically guided masks. The anatomical masks shown are derived from the fixed (anchor) image and are displayed consistently across all three panels for reference. *Left panel*: fixed image with overlaid masks. *Center panel*: moving image from a later time point, with the same fixed-image masks overlaid for visual alignment comparison. *Right panel*: registered moving image after hierarchical ANTs alignment. Mask interpretation: *green contours* = lung segmentation (tissue of interest), *red contours* = pathology-aware exclusion region (to prevent the deformable registration from being driven by abnormal findings), *cyan contours* = central mediastinal anchor region (stable anatomical landmark used to constrain excessive local deformation).

and mediastinum. A representative example of the fixed image, moving image, registered output, and estimated displacement field is shown in Fig. 2.

From these series, temporal pairs were generated by defining the earlier study as the fixed image and the subsequent study as the moving image. For the experiments conducted in this study, the dataset was specifically filtered to include pairs representing the progression of pleural effusion. In this configuration, the initial observation x_i represents either a healthy state or an existing condition, while the target state x_j is characterized by the definitive presence of pleural effusion.

4.2 Experimental results

The training evolution of the framework over 120 epochs reveals the balance between pixel-level fidelity and anatomical consistency. As illustrated in Fig. 3, the Peak Signal to Noise Ratio PSNR (left panel) exhibits an initial increase followed by a gradual descent as the neural ordinary differential equation ODE solvers begin to dominate the latent transformations. This trend is attributed to the model transitioning from generating images that closely resemble the high detail initial state x_0 to producing temporal forecasts that prioritize structural alignment with the target image x_t. While pixel-level metrics such as PSNR (21.57 ± 2.66 dB) reflect the characteristic smoothing introduced by MSE-based optimization, they do not fully capture the clinical relevance of the synthesized images. In longitudinal forecasting, high pixel-level fidelity is not necessarily indicative of clinical utility, as the model is expected to generate structurally different future states rather than replicate the initial observation.

This trade-off is further reflected in the evolution of the Hausdorff distance HD (right panel). In this visualization, the y-axis is inverted such that an upward trajectory represents a reduction in distance and, consequently, an improvement in boundary alignment. While the PSNR decreases, the HD improves and stabilizes above the baseline, indicating that the model successfully learns

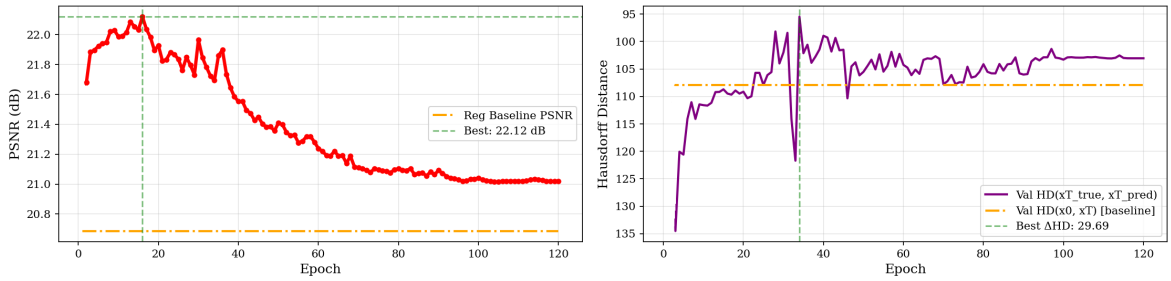


Figure 3: Evolution of validation metrics during training. The Peak Signal to Noise Ratio PSNR shows the shift from initial detail preservation to temporal modeling. The Hausdorff distance HD, plotted on an inverted y-axis, demonstrates improved anatomical alignment relative to the orange baseline representing the initial and target image comparison.

to adjust the anatomical contours toward the ground truth despite the loss of fine textural detail. The orange dashed line in both panels represents the baseline comparison, defined as the direct evaluation of the initial state x_0 against the target ground truth x_t without any temporal evolution.

The qualitative performance of the system is shown in Fig. 4, which displays an example of a sequence from the initial observation to a 14 day prediction. The synthesized output captures the progression of pathological features, such as the expansion of a pleural effusion, while maintaining the global thoracic structure. Although the predicted image exhibits less high frequency detail than the reconstruction at $t = 0$, the difference map confirms that the error remains spatially bounded and aligned with the areas of clinical change. The quantitative summary of these preliminary results is provided in Table 1. The low Dice coefficient (0.329 ± 0.220) indicates that the model struggles to capture precise patient-specific morphology, as the optimization focuses on modeling temporal intensity dynamics rather than explicit boundary segmentation. Consequently, the framework prioritizes the visual progression of the pathology over fine structural precision.

Metric	Value (Mean ± Standard Deviation)
Peak Signal to Noise Ratio PSNR (dB)	21.57 ± 2.66
Structural Similarity Index SSIM	0.677 ± 0.083
Mean Absolute Error MAE	0.120 ± 0.040
Mean Squared Error MSE	0.034 ± 0.025
Dice Coefficient	0.329 ± 0.220
Hausdorff Distance HD (pixels)	99.73 ± 60.47

Table 1: Quantitative summary of preliminary test results.

A more meaningful assessment of the model's performance is therefore provided by pathology-driven validation. To this end, an independent validation was performed to ensure that the framework effectively synthesizes the emerging pathology rather than simply replicating the initial state. A binary classifier was trained to distinguish between healthy and pathological radiographs, achieving an Area Under the Curve AUC of 93.44% and a precision of 91.30%. The overall accuracy and balanced accuracy reached 81.67%, with a recall of 70.00% and an F1 score of 79.25%.

When applied to the synthetic images $\hat{x}_j$ generated by the framework, the independent classifier correctly identified 88.61% of the predictions as containing pleural effusion, while only 11.39% were classified as healthy. This high classification rate confirms that the Neural Ordinary Differential

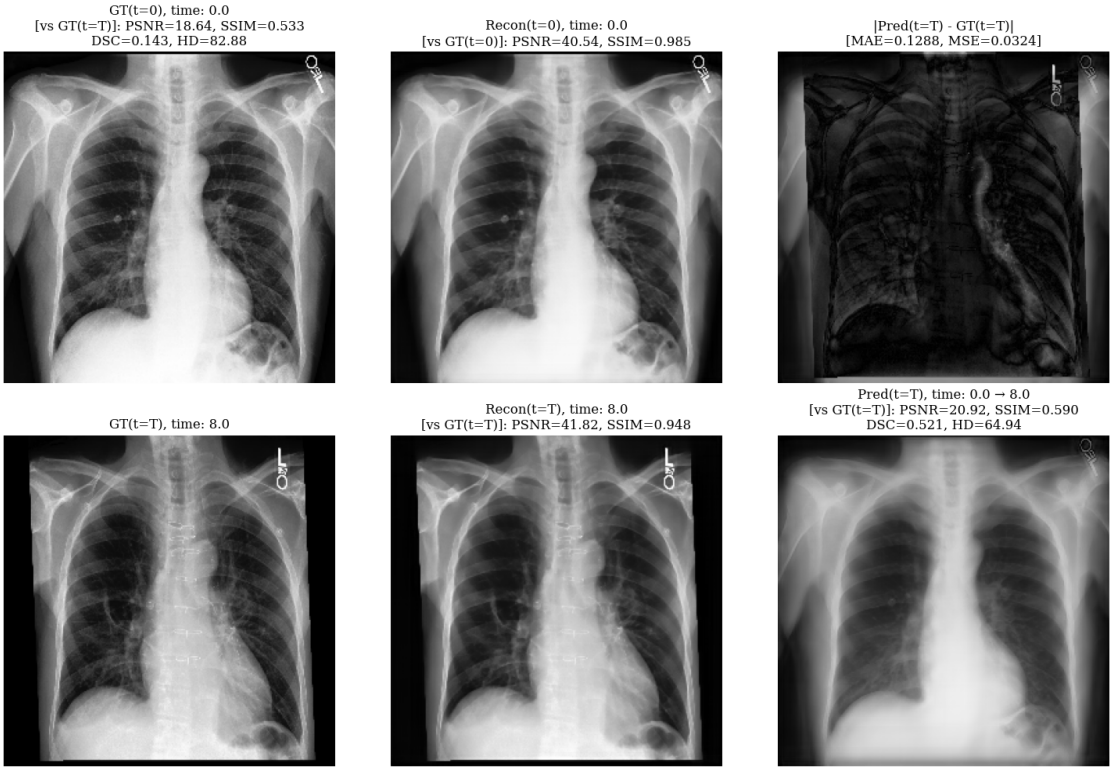


Figure 4: Qualitative example of pleural effusion progression over time. The top row shows the initial study together with its high-fidelity reconstruction, while the bottom row shows the target ground truth and the corresponding model prediction. The prediction captures the global anatomical evolution toward the target state, although fine details are visibly smoothed, which is consistent with an MSE-based optimization objective.

Equation ODE solvers successfully modify the latent representations to encode the visual characteristics of the target disease, ensuring that the generated images are clinically relevant to the modeled progression.

5 Conclusion

This work presents an initial framework for modeling the progression of pulmonary diseases as a continuous flow within a hierarchical latent manifold. The preliminary results confirm that the proposed architecture effectively generates visually coherent and anatomically accurate trajectories, capturing the evolution of clinical findings such as pleural effusion. Specifically, the pathology-driven validation using an independent classifier demonstrates that the model successfully synthesizes the target disease in 88.61% of cases, confirming its ability to forecast pathological transitions rather than merely replicating initial states.

The stability observed in the validation metrics and the successful alignment of lung boundaries demonstrate the model's capacity to generalize from complex longitudinal data with irregular temporal sampling. To improve the sharpness of the predictions and mitigate the characteristic smoothing caused by MSE optimization, future research will focus on incorporating additional

semantic information. The implementation of medical reports will be explored to better organize the representation space, facilitating the generation of radiographs with higher levels of detail and providing a more personalized, explainable prognosis simulation framework.

Acknowledgments

This work has been funded by GVA through PROGRAMA PROMETEO 2023-CIPROM/2022/20 and project PRACOMIA AP2025-09 funded by the Polytechnic University of Valencia.

References

[1] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *Advances in neural information processing systems*, 2018.

[2] C. Liu, K. Xu, et al., "Imageflownet: Forecasting multiscale image-level trajectories of disease progression with irregularly-sampled longitudinal medical images," *arXiv preprint arXiv:2406.14794*, 2024.

[3] A. Bustos, A. Pertusa, J.-M. Carmona, and S. Iglesias, "Padchest: A large chest x-ray dataset with multi-label annotated reports," *Medical image analysis*, vol. 66, 2020.

[4] A. E. Johnson, T. J. Pollard, et al., "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, 2019.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE transactions on image processing*, 2004.

[7] W. Wang et al., "Diffusion model for medical image denoising, reconstruction and translation," *Computerized Medical Imaging and Graphics*, 2025.

[8] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: A learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[9] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, 2015.

[11] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[12] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[13] B. B. Avants, N. J. Tustison, and G. Song, "A reproducible evaluation of ants similarity metric performance in brain image registration," *NeuroImage*, vol. 54, no. 3, pp. 2033–2044, 2011.

Mathematical Modelling of the Atypical Epidemic Behaviour of Seasonal Influenza in the Valencian Community During the 2015–2016 Season

Rocío Bagán Navarro[♮], Cristina-Luisovna Pérez[♭] and Rafael J. Villanueva Micó[♭]

(♮) Universidad Europea de Valencia, ,
Valencia, Spain.
(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Abstract

Seasonal influenza remains one of the most persistent recurrent challenges in public health, exhibiting marked interannual variability in epidemic timing, transmissibility, age distribution, and clinical burden [1], [2]. Whilst epidemic waves in temperate regions generally follow a recognisable winter seasonal pattern, certain seasons depart substantially from historical expectations. Such anomalous outbreaks provide valuable natural experiments through which the mechanisms governing respiratory epidemic dynamics may be investigated.

The influenza season of 2015–2016 in the Valencian Community (Spain) constituted one such case. Surveillance records revealed a substantially delayed epidemic peak, a more moderate cumulative incidence, and an epidemiological profile distinct from neighbouring seasons. In contrast, the 2016–2017 season displayed the more conventional pattern typically observed in Mediterranean settings. The present work examines whether this atypical behaviour may be explained within a parsimonious mechanistic framework, without invoking structural changes in transmission processes.

To this end, a susceptible–latent–infectious–removed (SLIR) compartmental model was implemented and calibrated against weekly surveillance data using Particle Swarm Optimisation. Estimated parameters indicate that the 2015–2016 season was characterised by lower effective transmissibility, reduced epidemic growth, and delayed peak formation relative to 2016–2017. Practical identifiability analysis further demonstrated robust parameter recovery in the vicinity of the optimum solution.

These findings suggest that unusual epidemic behaviour may emerge from moderate shifts in a small set of epidemiologically interpretable parameters rather than from qualitatively distinct mechanisms. More broadly, the study illustrates the strategic value of mathematical modelling for outbreak interpretation, prospective surveillance, and preparedness planning in infectious disease systems.

References

[1] World Health Organization, "Up to 650 000 people die of respiratory diseases linked to seasonal flu each year," *WHO News Release*, 2018. [Online]. Available: `https://www.who.int`.

[2] J. Oliva, C. Delgado-Sanz, A. Larrauri, et al., "Burden of seasonal influenza in spain: Epidemiology and economic impact," *Vaccines*, vol. 10, no. 4, pp. 1–15, 2022.

Commol: A Rust-Powered Python Library for Compartment Dynamic Models

David Martínez-Rodríguez[♭,1]
Carlos Andreu-Vilarroig,[♭],
and Rafael Jacinto Villanueva Micó[♭]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

Compartment models are widely utilized across various mathematical modeling domains, such as epidemiology, population dynamics, pharmacokinetics, or chemical reaction systems [1]. These models describe complex biological or physical systems as a finite set of homogeneous, well-mixed subsystems—termed compartments—that exchange materials, individuals, or energy according to conservation laws and rate equations [2]. When developing these models, researchers frequently encounter a technical compromise between the high-level usability of scientific programming languages and the raw computational efficiency required for extensive simulations, sensitivity analyses, or complex parameter calibration tasks. Historically, mathematical models are manually recoded for each specific study or optimized in lower-level languages like C++ or Fortran once Python or R prototypes reach their performance limits. This manual implementation often results in rigid software structures that lack modularity, making real-world problems involving multiple compartments, non-linear transitions, and stratified populations difficult to reproduce, scale, and reuse across different research groups.

To address these limitations, Commol [3] was developed as an open-source compartment modeling library designed to eliminate the "two-language problem" in compartment modelling coding. Commol aims to enhance compartment modeling by coupling a fluent, user-friendly Python [4] Application Programming Interface (API) with a highly optimized, memory-safe Rust [5] core. This framework allows to define complex dynamic systems natively in Python while strictly delegating execution to a Rust compiled backend. By doing so, Commol standardizes the inclusion of advanced calibration methodologies and uncertainty quantification, reducing the barrier to entry for developing highly performant, production-ready mathematical models.

2 Software Architecture and Design

The software architecture of the Commol library is structured around three distinct operational layers, establishing a separation of concerns between the user-facing declarative interface and the high-performance numerical engine. The Python API layer serves as the primary point of user

[1] damarro3@upv.es

interaction, exposing the central classes and builder patterns required for model construction, validation, and analytical post-processing. Sitting beneath this interface, the PyO3 binding [6] layer acts as a critical intermediary, facilitating seamless interoperability and zero-copy memory sharing between the Python interpreter and the compiled backend. This binding layer is specifically engineered to release the Python Global Interpreter Lock (GIL) during computationally intensive tasks, allowing true multithreading. Finally, the Rust core implementation assumes responsibility for the computational workloads. This core is logically partitioned into specialized crates, primarily including components dedicated to memory-safe core data structures, difference equation solvers, graphical network representations, and parameter calibration routines. The integration and compilation of these cross-language components are managed by the maturin [7] build system and the Cargo [8] workspace manager, ensuring cross-platform compatibility and distribution.

Achieving high performance in dynamic systems modeling needs specific design principles, particularly when executing iterative algorithms over prolonged time horizons. A computational bottleneck in dynamic modeling is the continuous evaluation of user-defined mathematical expressions during the inner simulation loop. Traditionally, these expressions are evaluated using interpreted runtimes, which introduces overhead. Commol solves this fundamental limitation through the implementation of Just-In-Time (JIT) compilation technologies. Mathematical rate expressions, defined by the user as standard algebraic strings, are parsed exactly once into Abstract Syntax Trees (ASTs) during the initialization phase. Utilizing the Cranelift code generation framework [9], these parsed syntax trees are subsequently compiled directly into native machine code. This JIT compilation pipeline allows arbitrary, user-defined, non-linear equations to execute at hardware speeds that are strictly comparable to hardcoded Rust implementations, entirely bypassing the Python runtime during simulation.

The execution phase is optimized for memory efficiency. The Rust simulation engine relies on pre-allocated, reusable memory buffers. By avoiding new dynamic memory allocations during the inner simulation loop, the framework avoids the latency introduced by system calls, achieving low execution times. Additionally, for computationally intensive tasks that require exploring expansive parameter spaces—such as multi-run optimization, sensitivity analysis, or ensemble calibration—Commol leverages the Rayon library [10]. Rayon provides a data-parallelism framework utilizing a highly efficient work-stealing scheduler, effortlessly parallelizing execution across all available CPU cores without requiring explicit thread management from the user.

On the Python frontend, data integrity and mathematical soundness are guaranteed prior to any simulation execution. The Pydantic [11] library is utilized for programmatic data validation, ensuring that all model parameters, topological bounds, and initial conditions conform exactly to expected mathematical types and ranges. Furthermore, the library incorporates the pint package [12] to provide optional, physical unit handling and dimensional validation. This integration prevents unit errors, ensuring that rate constants and state variables interact seamlessly and produce dimensionally consistent derivatives. The analytical ecosystem is rounded out by NumPy [13] for vectorized data extraction and tensor manipulation, alongside Matplotlib [14] and Seaborn [15] for generating publication-ready visualizations of simulated trajectories and empirical data fits.

3 Model Definition and Stratification

The construction of a mathematical model within Commol utilizes a builder programming pattern, allowing users to programmatically and incrementally construct the dynamic system with minimal boilerplate code. Users instantiate the elements of the system through the API, logically categorizing them into distinct operational definitions that mirror the formal mathematical description of the system.

Compartments are defined as discrete state variables intended to represent specific states of the

entities being modeled, such as susceptible individuals in an epidemic. Parameters are explicitly declared to define the numerical rate constants governing the velocity of the system. These parameters can be configured with specific initial scalar values for forward simulation, or with defined boundary constraints (minimum and maximum limits) for inverse problem-solving and parameter calibration.

Transitions act as the directional edges of the model graph, delineating the physical or state-based flows between designated source and target compartments. These transitions are not limited to linear constants; they execute arbitrary mathematical rate expressions written in standard algebraic notation. To maximize flexibility, these expressions can utilize a dedicated variable placeholder to dynamically reference the current state value of the source compartment, destination compartment, or global population totals during each step of numerical integration.

Beyond basic topological definition, Commol supports complex stratification. Real-world modeling frequently demands dividing homogeneous compartments into heterogeneous sub-populations based on demographic attributes—such as age groups, spatial locations, or vulnerability profiles. Instead of requiring the user to manually declare hundreds of individual compartments and transition permutations, the architecture allows for procedural stratification. A base model can be defined once and subsequently expanded through stratification matrices, automatically generating the corresponding interconnected sub-graphs. This reduces the errors in the model formulation and accelerates the prototyping phase for complex, multi-layered physiological and epidemiological systems.

4 Simulation and Parameter Calibration

Simulation execution within the current architecture leverages difference equation solvers to project discrete time trajectories. The technique is an explicit fixed-step recurrence, analogous to the forward Euler method [16]: at each step, every transition rate is evaluated at the current compartment state and the resulting flows are applied simultaneously to obtain the next state, moving quantities from source to target compartments over a constant time increment. Since the next state depends only on the present one, no iterative solve is required, making each step a single inexpensive pass and entire trajectories cheap to generate. This approach outputs time series data in multiple accessible formats supported by integrated visualization utilities for immediate inspection. However, while forward simulation is computationally trivial for modern architectures, the primary computational challenge in applied compartment modeling lies in fitting unknown parameter values to noisy, real-world empirical data. Commol addresses this bottleneck through a parallelized optimization module.

The deterministic parameter calibration framework fits simulation outputs to observed data points utilizing several available optimization algorithms tailored for non-linear systems. The Rust core supports both the Nelder-Mead simplex method [17] and Particle Swarm Optimization (PSO) [18], a stochastic, population-based metaheuristic. These optimization routines iteratively minimize deviations between the simulated trajectories and empirical data based on configurable statistical loss functions, specifically Sum of Squared Errors (SSE), Root Mean Square Error (RMSE), or Mean Absolute Error (MAE), depending on the underlying noise distribution of the dataset. The calibration process is highly customizable; it allows users to enforce minimum and maximum bounds on specific parameters, apply domain-knowledge inequality constraints, and introduce scaling parameters. These scaling parameters are particularly crucial in epidemiology, as they explicitly account for under-reporting, testing constraints, or reporting delays in empirical surveillance data, mapping the true simulated incidence to the observed metric.

To support robust Uncertainty Quantification (UQ), Commol implements a probabilistic calibration workflow based on ensemble methods. Biological and physical systems frequently exhibit

parameter identifiability issues, where multiple distinct sets of parameters yield nearly identical fits to the empirical data. To capture this multimodality, the system executes a predefined number of independent, randomized optimization runs across the parameter hyperspace. The resulting parameter solutions are subsequently grouped utilizing K-means clustering algorithm [19]. This clustering identifies similar sets, allowing the system to extract diverse, representative parameter configurations from distinct local minima.

Following the clustering phase, an ensemble optimization is performed utilizing the NSGA-II (Non-dominated Sorting Genetic Algorithm II) [20]. The algorithm operates on two primary, competing objectives: first, to select parameter sets that maximize the coverage of empirical data points (ensuring the model reliably captures the observed system dynamics), and second, to minimize the width of the resulting temporal confidence intervals (ensuring high predictive precision). By iteratively evolving a population of candidate parameter sets, the algorithm identifies the Pareto front, a set of non-dominated solutions where improvements in data coverage necessarily necessitate a trade-off in interval width. By simulating from this ensemble, Commol automatically derives temporal confidence intervals, statistical coverage metrics, and descriptive parameter statistics, such as means, standard deviations, and 95% confidence intervals. This provides decision-makers with a quantitative view of model uncertainty, which is indispensable for identifying high-risk scenarios in public health interventions or ensuring safety thresholds in clinical drug dosing.

5 Conclusion

Commol is distributed as open-source software under a MIT license and is currently in its version 0.1.0 alpha development phase. The foundational architecture, core discrete-time difference-equation features, and calibration modules are fully implemented and optimized. The packaged library is actively available on the Python Package Index (PyPI) and can be installed via standard package managers for Linux, Windows, and macOS operating systems.

The development roadmap focuses on expanding the mathematical foundations and application scopes of the library to transition toward a stable 1.0 release. Future releases will implement continuous-time Ordinary Differential Equation (ODE) solvers. The inclusion of adaptive step-size algorithms and implicit solvers for stiff systems will complement the existing discrete-time framework.

Additionally, the software architecture is being extended to support advanced network-based topological models, including agent-based and stochastic network representations. By leveraging the same builder syntax, this unified design allows to express different modeling paradigms for the exact same system, such as transitioning from a mean-field ODE approach to an individual-based agent model or a stochastic network simulation with minimal modifications. This modularity will enable to directly compare how different structural assumptions influence model behavior, facilitating the evaluation of how individual heterogeneity or local network topology affects emergent phenomena.

Acknowledgments

This work has been partially supported by the grant PID2025-172735NB-I00 funded by MICIU/AEI/10.13039/501100011033 and by the grant CIPROM/2026/22 funded by the Generalitat Valenciana (Conselleria de Educación, Cultura, Universidades y Empleo).

References

[1] F. Brauer, "Compartmental models in epidemiology," in *Mathematical Epidemiology*, vol. 1945, 2008, pp. 19–79. DOI: 10.1007/978-3-540-78911-6_2.

[2] K. Godfrey, *Compartmental Models and Their Application*. London, UK: Academic Press, 1983.

[3] R. J. Villanueva Micó, C. Andreu Vilarroig, and D. Martínez Rodríguez, *Commol: A high-performance compartment modelling library*, 2026. [Online]. Available: https://github.com/MUNQU/commol.

[4] G. Van Rossum and F. L. Drake, *Python3*, 2026. [Online]. Available: https://www.python.org.

[5] N. D. Matsakis and F. S. Klock, "The rust language," *Ada Lett.*, vol. 34, no. 3, pp. 103–104, Oct. 2014, ISSN: 1094-3641. DOI: 10.1145/2692956.2663188.

[6] PyO3 Project and Contributors, *PyO3*, 2026. [Online]. Available: https://github.com/PyO3/pyo3.

[7] PyO3 Project and Contributors, *Maturin*, 2026. [Online]. Available: https://github.com/pyo3/maturin.

[8] The Rust Project Developers, *The Cargo book: The Rust package manager*, https://doc.rust-lang.org/cargo/, 2025. Accessed: Jul. 2026.

[9] Bytecode Alliance, *Cranelift*, 2026. [Online]. Available: https://github.com/bytecodealliance/wasmtime/tree/main/cranelift.

[10] Stone, Josh and Matsakis, Nicholas D., *Rayon: A data parallelism library for Rust*, 2026. [Online]. Available: https://github.com/rayon-rs/rayon.

[11] S. Colvin et al., *Pydantic Validation*, version v2.14.0a1, May 2026. [Online]. Available: https://github.com/pydantic/pydantic.

[12] H. E. Grecco, J. Chéron, A. Savage, and contributors, *Pint: Operate and manipulate physical quantities in Python*, 2026. [Online]. Available: https://github.com/hgrecco/pint.

[13] C. R. Harris et al., "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020. DOI: 10.1038/s41586-020-2649-2.

[14] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. DOI: 10.1109/MCSE.2007.55.

[15] M. L. Waskom, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021. DOI: 10.21105/joss.03021. [Online]. Available: https://doi.org/10.21105/joss.03021.

[16] S. Elaydi, *An Introduction to Difference Equations* (Undergraduate Texts in Mathematics), 3rd. Springer, 2005.

[17] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, Jan. 1965, ISSN: 0010-4620. DOI: 10.1093/comjnl/7.4.308. eprint: https://academic.oup.com/comjnl/article-pdf/7/4/308/1013182/7-4-308.pdf. [Online]. Available: https://doi.org/10.1093/comjnl/7.4.308.

[18] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4, 1995, 1942–1948 vol.4. DOI: 10.1109/ICNN.1995.488968.

[19] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, Berkeley, CA, USA: University of California Press, 1967, pp. 281–297.

[20] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002. DOI: 10.1109/4235.996017.

Inverse Problems in Dental CBCT for Digital Dentistry

Hyoung Suk Park and Kiwan Jeon[1]

National Institute for Mathematical Sciences
Yuseong-gu, Daejeon, Republic of Korea, 34047

1 Introduction

Digital dentistry relies fundamentally on volumetric imaging, where cone-beam computed tomography (CBCT) plays a central role in diagnosis and treatment planning. However, due to compact system design and low-dose requirements, dental CBCT inherently suffers from ill-posed inverse problems, primarily caused by photon scattering and truncated projection data [1], [2]. These factors lead to severe shading artifacts and structural distortions, which degrade both visual quality and downstream tasks such as segmentation. From a mathematical perspective, CBCT reconstruction can be formulated as an inverse problem that requires a careful balance between data fidelity and prior information under incomplete measurements.

2 Methods

We consider inverse problem formulations for CBCT reconstruction under incomplete measurements arising from scattering and truncated projections. In the image domain, shading artifacts are addressed using generative adversarial networks (GANs), where learned mappings act as data-driven regularizations applied after reconstruction [3]. In the projection domain, implicit neural representations (INRs) are employed to model attenuation beyond the reconstruction region, enabling compensation for missing measurements within the forward model [4]. In addition, practical reconstruction is achieved by combining INR-based projection completion with analytic reconstruction methods such as FDK, allowing efficient computation under limited data.

3 Results

Across these approaches, different formulations address distinct sources of degradation in CBCT imaging. Image-domain restoration is effective in reducing shading artifacts caused by scattering, while projection-domain approaches focus on compensating for missing data due to field-of-view truncation. These approaches provide complementary improvements in image quality and structural consistency. In addition, practical reconstruction schemes offer a balance between image quality and computational efficiency.

[1]jeonkiwan@nims.re.kr

4 Conclusions

This work highlights different inverse problem formulations in digital dentistry, including both data-driven image restoration and physics-based reconstruction. These approaches address distinct sources of degradation in CBCT imaging and can provide complementary improvements in image quality and structural consistency. Practical reconstruction strategies further allow a balance between image quality and computational efficiency.

Acknowledgments

This work was supported by the National Institute for Mathematical Sciences (NIMS) grant funded by the Korean government (No. NIMS-B26A10000) partially supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00338419)

References

[1] C. M. Hyun, K. Jeon, S. H. Kang, S.-H. Lee, J. K. Seo, and H. S. Park, "Clinical applications of low-dose dental cone-beam computed tomography," in *Emerging Radiation Detection: Technology and Applications*, K. Iniewski and H. Gadey, Eds. Springer Nature Switzerland, 2024, pp. 147–163. DOI: 10.1007/978-3-031-63897-8_9.

[2] H. S. Park, K. Jeon, and J. K. Seo, "Deep learning-based artefact reduction in low-dose dental cone beam computed tomography with high-attenuation materials," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 383, no. 2305, p. 20 240 045, Sep. 2025. DOI: 10.1098/rsta.2024.0045.

[3] H. S. Park, K. Jeon, S.-H. Lee, and J. K. Seo, "Unpaired-paired learning for shading correction in cone-beam computed tomography," *IEEE Access*, vol. 10, pp. 26 140–26 148, 2022. DOI: 10.1109/ACCESS.2022.3155203.

[4] H. S. Park and K. Jeon, "An iterative reconstruction method for dental cone-beam computed tomography with a truncated field of view," 2025. arXiv: 2508.07618. [Online]. Available: https://arxiv.org/abs/2508.07618.

Mathematical Modelling of Cognitive Load During Mental Arithmetic Tasks Using EEG Signals

Israr Ul Haq and Luis Acedo[1]

Department of Mathematics, Centro Universitario de Plasencia,
Avda. Virgen del Puerto, 2,
University of Extremadura, 10600, Plasencia, Spain.

1 Introduction: Electroencephalography: Technique, Variants, and Clinical Applications

Electroencephalography (EEG) is a non-invasive neurophysiological technique that records the brain's electrical activity through surface electrodes placed on the scalp. The method, first introduced by Hans Berger in 1929 [1], captures oscillatory patterns generated primarily by postsynaptic potentials of cortical pyramidal neurons. EEG has since become a cornerstone of clinical neurophysiology and computational neuroscience due to its high temporal resolution and ability to monitor dynamic brain states [2].

1.1 Technical Foundations

EEG acquisition typically follows the international 10–20 electrode placement system, ensuring spatial standardization across recordings. Electrodes detect voltage fluctuations that are amplified, filtered, and digitized for further analysis. Recordings may be performed during wakefulness, rest, sleep, or under activation procedures such as hyperventilation or photic stimulation. Depending on the clinical context, EEG studies may range from short routine sessions to prolonged or ambulatory monitoring. The technique is safe, painless, and suitable for repeated evaluations [3].

1.2 Types of EEG Recordings

Several EEG modalities are used in clinical and research settings:

- **Routine EEG**: Short baseline recording used as an initial diagnostic tool.
- **Long-term EEG**: Extended monitoring to capture infrequent paroxysmal events, particularly in epilepsy.
- **Ambulatory EEG**: Continuous recording during normal daily activities.
- **Video-EEG**: Simultaneous EEG and video recording to correlate electrical activity with clinical manifestations.

[1] acedo@unex.es

- **Intraoperative EEG**: Monitoring during neurosurgical or high-risk procedures to detect ischemia or functional compromise.
- **Polysomnographic EEG**: Integrated into sleep studies to characterize sleep stages and diagnose sleep disorders.

1.3 Clinical Applications

EEG plays a central role in the diagnosis and management of neurological disorders. Its primary indication is the evaluation of epilepsy, enabling the detection of epileptiform discharges, classification of seizure types, and guidance of therapeutic decisions. EEG is also valuable in assessing metabolic or toxic encephalopathies, infectious diseases of the central nervous system, cerebrovascular events, and disorders of consciousness. Additional applications include confirmation of brain death, evaluation of sleep disorders, and monitoring depth of anesthesia or medically induced coma [4].

1.4 Mathematical and Computational Models Based on EEG

EEG signals provide a rich substrate for mathematical modeling due to their temporal precision and complex dynamical structure. Several modeling approaches are widely used:

- **Spectral and time–frequency models**: Fourier analysis, wavelet transforms, and multitaper methods quantify oscillatory components and transient events.
- **Neural mass and neural field models**: Mesoscopic models such as the Jansen–Rit [5] and Wendling models [6] simulate population-level cortical dynamics and reproduce EEG rhythms and epileptic discharges.
- **Network and graph-theoretical models**: Functional connectivity networks derived from coherence, phase-locking value, or Granger causality allow characterization of large-scale brain organization [7].
- **Nonlinear dynamical systems**: Techniques such as Lyapunov exponents, entropy measures, and recurrence analysis capture the nonlinear structure of EEG activity.
- **Machine learning and deep learning models**: Convolutional neural networks, recurrent architectures, and modern transformer-based models enable automated classification, seizure prediction, and biomarker discovery [8].

These mathematical frameworks support the development of predictive tools, enhance understanding of pathological brain states, and contribute to intelligent monitoring systems in clinical environments [9].

2 Techniques Used to Study EEG Dynamics in Acedo & Aranda (2012)

In our collaborative work [10], we investigated the temporal structure of electroencephalographic (EEG) signals using analytical, computational, and statistical tools. The central objective was to determine whether the *damped oscillatory relaxation* observed in real EEG recordings could emerge from simple stochastic models of collective neural activity.

2.1 SIS Stochastic Cellular Automaton Model

We modeled the brain as a stochastic cellular automaton based on the susceptible–infected–susceptible (SIS) epidemic process on a complete graph. Firing neurons correspond to "infected" sites, while quiescent neurons correspond to "susceptible" sites. The mean-field evolution of the fraction of active neurons is governed by the nonlinear recurrence

$$x(t+1) = x(t) + (Na - \beta)x(t) - Na\,x(t)^2, \tag{1}$$

quoted in the article as "$x(t+1) = x(t) + (Na - B)x(t) - Nax^2(t)$".

To solve this system, we employed a *modal series expansion*

$$x(t) = \sum_{k=0}^{\infty} A_k\, a_k^t, \tag{2}$$

which yields exact recurrence relations for the coefficients A_k. This method allows us to determine whether the fundamental mode a is positive (pure exponential decay) or negative (oscillatory decay). The latter case produces the damped oscillations that are also observed in EEG autocorrelations.

Analytical Computation of the Autocorrelation Function

Because EEG does not directly measure the number of firing neurons, we focused on the temporal self-correlation function

$$\phi(t) = \langle x(t|x_0)\, x_0 \rangle - \langle x(t) \rangle^2, \tag{3}$$

defined in the paper as "$\phi(t) = (x(t|x_0)x_0) - (x(t))^2$".

We assumed that fluctuations around the stationary state A_0 follow a Gaussian distribution,

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, \exp\left[-\frac{(x - A_0)^2}{2\sigma^2}\right],$$

quoted as "$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-A_0)^2/(2\sigma^2)}$". Integrating over all possible initial conditions yields a self-correlation function with *damped oscillatory relaxation*, matching the behavior seen in real EEG data.

Comparison with Elementary Cellular Automata

To test the generality of this phenomenon, we simulated several elementary cellular automata (ECA). We found periodic oscillations in rules 26, 28, 38, 50, and 52, and damped oscillations in rule 30. As stated in the article, "Damped oscillations are found for ECA rule 30". This suggests that oscillatory relaxation is a robust feature of complex automata, not specific to the SIS model.

Autocorrelation Analysis of Real EEG Signals

We analyzed EEG recordings (delta and alpha rhythms) by dividing long signals into independent 10-second segments. For each segment we computed

$$\phi(t) = \langle V(t)V(0) \rangle - \langle V(t) \rangle^2, \tag{4}$$

quoted as "$\phi(t) = \{V(t)V(t=0)\} - \langle V(t) \rangle^2$".

Both delta and alpha rhythms exhibit clear damped oscillations, with faster decay in deep sleep than in wakefulness. The similarity between model predictions and empirical EEG supports the interpretation of EEG as a *statistical epiphenomenon* emerging from collective neural dynamics.

Understanding cognitive activity through quantitative analysis of brain signals is an important interdisciplinary challenge spanning applied mathematics, physics, and neuroscience. Electroencephalography (EEG) provides high-temporal-resolution recordings that can be modeled as multivariate, nonstationary time series, making them particularly suitable for spectral, coherence, and scaling-based mathematical methods. The dataset by Zyma et al. [11] was curated specifically to support such analyses by providing well-controlled EEG recordings during mental arithmetic, a standardized paradigm for inducing cognitive load.

3 Dataset Description

The dataset consists of EEG recordings from 36 healthy young adults performing a mental arithmetic task based on serial subtraction, along with baseline resting-state recordings. EEG signals were acquired using a 23-channel system following the International 10–20 montage, sampled at 500 Hz. For each subject, artifact-free segments of 180 s (rest, eyes closed) and 60 s (mental arithmetic) are provided.

Participants were divided into two groups according to task performance: "good counters" and "bad counters", based on the number and accuracy of arithmetic operations completed within a fixed time window. This grouping allows the investigation of inter-individual differences in task difficulty and cognitive workload using mathematical descriptors derived from EEG signals [11].

All recordings are distributed in European Data Format (EDF) and organized according to the Brain Imaging Data Structure (BIDS) standard [12], facilitating reproducible computational and mathematical analysis.

4 Relevance to Applied Mathematics

From an applied mathematics perspective, the dataset enables systematic comparison between traditional frequency-domain approaches, such as Fourier power spectral density and coherence, and nonlinear time-series methods. In particular, detrended fluctuation analysis (DFA) provides access to scaling exponents that quantify long-range temporal correlations in nonstationary EEG signals [13], [14].

Such analyses support the development and validation of mathematical models of brain dynamics, including multiscale processes, coupling between brain regions, and changes in temporal correlation structure induced by cognitive load. The availability of both rest and task conditions, as well as performance-based subject grouping, makes this dataset a valuable benchmark for applied mathematical research on complex biological signals.

The EEG dataset summarized in this paper constitutes a robust and well-documented resource for applied mathematicians and signal-processing researchers. Its design explicitly supports spectral, coherence, and scaling analyses, enabling quantitative investigations of multiscale brain dynamics during cognitive workload. The open-access nature and standardized formatting further enhance its value for reproducible and methodologically rigorous research.

5 Results

Preliminary results for the mathematical modelling of the signal are given below.

5.1 Time-Domain EEG Signal

Figure 1 shows a representative segment of the EEG signal recorded from the frontal channel Fp1 during resting-state conditions. The signal exhibits oscillatory activity with amplitudes on the

order of tens of microvolts and no dominant high-frequency components, which is consistent with baseline neural activity in a conscious and relaxed state.

This time-domain representation provides a qualitative reference for subsequent frequency-domain analyses and illustrates the inherent variability of frontal EEG dynamics prior to task engagement.

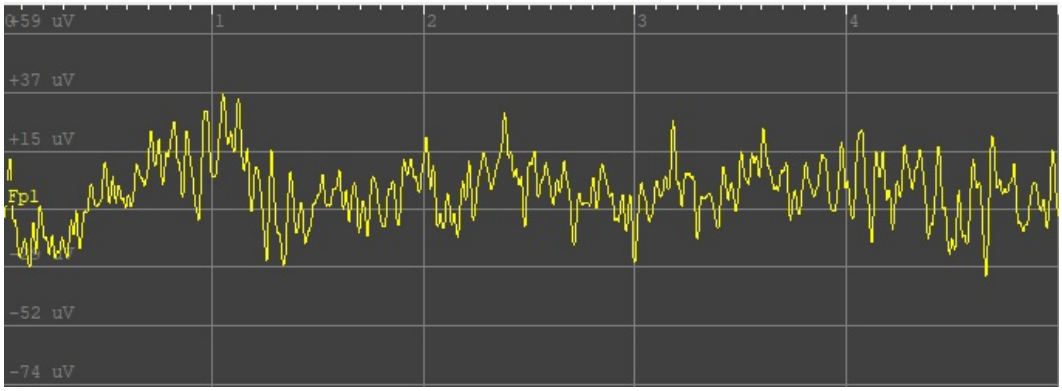


Figure 1: Representative EEG signal segment recorded from channel Fp1 during resting-state conditions, illustrating baseline frontal neural activity. The EEG recorded corresponds to 5 sec. The amplitude of the wave is measured in microVolts (μV).

5.2 Power Spectrum Analysis: Resting State

The power spectral density (PSD) of the EEG signal during resting conditions is shown in Figure 2. A clearly defined spectral peak is observed in the 8–12 Hz frequency range, corresponding to the alpha band. This alpha-band dominance is a well-known neurophysiological signature of conscious but relaxed wakefulness.

The presence of this peak indicates minimal cognitive load and serves as a reference state for comparison with task-related neural activity.

5.3 Power Spectrum Analysis: Mental Arithmetic

Figure 3 shows the power spectral density obtained during mental arithmetic task performance. In contrast with the resting condition, the alpha-band peak is markedly attenuated, and spectral power is redistributed toward other frequency components.

This reduction of alpha activity reflects increased cognitive load and enhanced cortical activation associated with active mental processing. The observed spectral reorganisation constitutes a clear neurophysiological marker of the transition from resting to task-engaged cognitive states.

5.4 Comparative Interpretation

The comparison between resting-state and task-related spectra highlights the sensitivity of frequency-domain EEG observables to changes in cognitive state. The prominent alpha-band peak observed during rest and its suppression during mental arithmetic are consistent across trials, despite inter-subject variability.

From a modelling perspective, the mental arithmetic task can be interpreted as a perturbation of the baseline neural dynamical system, leading to a departure from alpha-dominated oscillatory behaviour toward a more broadband spectral configuration. These results provide a clear empirical basis for incorporating spectral features as observables in mathematical models of cognitive load.

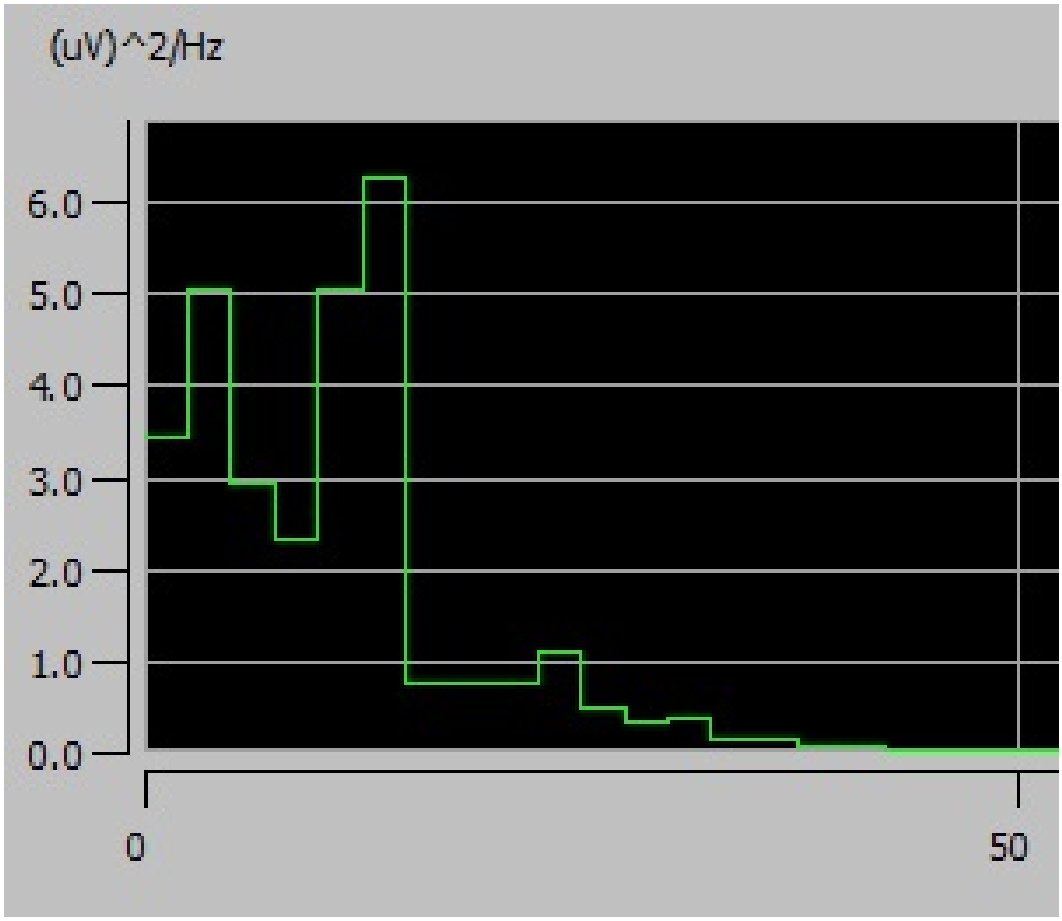


Figure 2: Power spectral density of the EEG signal during resting conditions. A pronounced peak in the 8–12 Hz range is observed, corresponding to alpha-band activity characteristic of a relaxed conscious state.

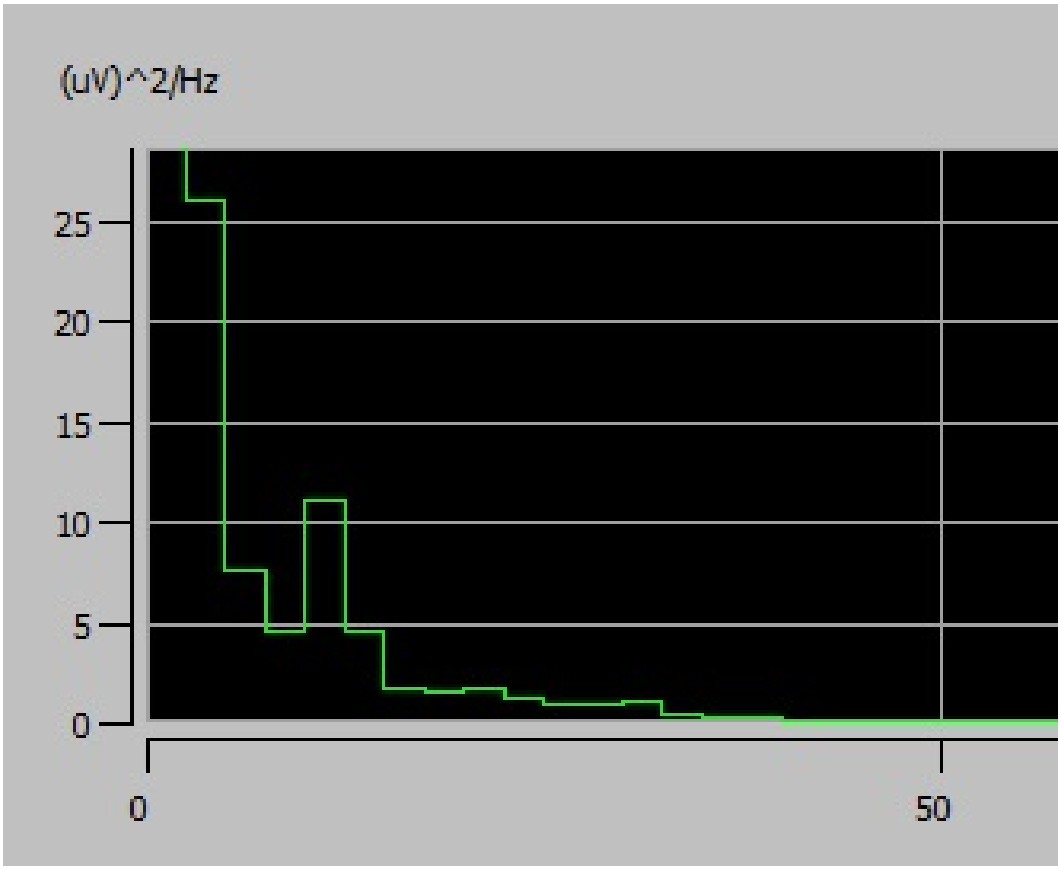


Figure 3: Power spectral density of the EEG signal during mental arithmetic task performance. Compared to the resting state, alpha-band power is reduced and spectral energy is redistributed, indicating increased cognitive load.

6 Conclusions

In this work, a mathematical modelling approach for the analysis of human cognitive behaviour during mental arithmetic tasks has been presented, based on the study of electroencephalographic (EEG) signals. By combining time-domain representations with frequency-domain analysis, it was possible to identify clear neurophysiological markers associated with different cognitive states.

The resting-state condition is characterised by a pronounced alpha-band peak in the 8–12 Hz range, which is a well-established signature of conscious yet relaxed wakefulness. During mental arithmetic performance, this alpha activity is significantly attenuated, accompanied by a redistribution of spectral power, reflecting increased cognitive load and enhanced cortical engagement. The contrast between these two conditions provides empirical evidence of task-dependent modulation of neural dynamics.

From a modelling perspective, these observations support the interpretation of cognitive activity as a dynamical system whose observable outputs are modulated by task demands. Spectral features derived from EEG signals emerge as suitable observables for capturing transitions between cognitive states, despite inherent inter-subject variability.

The results presented here demonstrate the potential of mathematical and statistical frameworks to bridge engineering methodologies and cognitive neuroscience in the study of human behaviour. Future work will focus on introducing quantitative metrics, such as spectral ratios and dynamical order parameters, as well as extending the analysis to larger datasets and more complex modelling approaches. The present extended abstract thus provides a solid foundation for the development of a full-length journal article, with *Physica A* as a natural target venue.

References

[1] H. Berger, "Über das elektrenkephalogramm des menschen," *Archiv für Psychiatrie und Nervenkrankheiten*, vol. 87, pp. 527–570, 1929.

[2] E. Niedermeyer and F. L. da Silva, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Lippincott Williams & Wilkins, 2005.

[3] F. L. da Silva, "Eeg and meg: Relevance to neuroscience," *Neuron*, vol. 80, no. 5, pp. 1112–1128, 2013.

[4] G. Buzsáki, C. A. Anastassiou, and C. Koch, "The origin of extracellular fields and currents—eeg, ecog, lfp and spikes," *Nature Reviews Neuroscience*, vol. 13, no. 6, pp. 407–420, 2012.

[5] B. H. Jansen and V. G. Rit, "Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns," *Biological Cybernetics*, vol. 73, no. 4, pp. 357–366, 1995.

[6] F. Wendling, F. Bartolomei, J.-J. Bellanger, and P. Chauvel, "Epileptic fast activity can be explained by a model of impaired gabaergic dendritic inhibition," *European Journal of Neuroscience*, vol. 15, no. 9, pp. 1499–1508, 2002.

[7] C. J. Stam, "Functional connectivity patterns of human magnetoencephalographic recordings: A small-world network approach," *Neuroscience Letters*, vol. 355, no. 1–2, pp. 25–28, 2004.

[8] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[9] M. Breakspear, "Dynamic models of large-scale brain activity," *Nature Neuroscience*, vol. 20, no. 3, pp. 340–352, 2017.

[10] L. Acedo and D. F. Aranda, "Self-correlations of electroencephalograms," *Complex Systems*, vol. 20, no. 4, pp. 289–302, 2012.

[11] I. Zyma et al., "Electroencephalograms during mental arithmetic task performance," *Data*, vol. 4, no. 1, p. 14, 2019. DOI: 10.3390/data4010014. [Online]. Available: https://www.mdpi.com/2306-5729/4/1/14.

[12] K. J. Gorgolewski et al., "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments," *Scientific Data*, vol. 3, p. 160 044, 2016. DOI: 10.1038/sdata.2016.44.

[13] C.-K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger, "Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series," *Physical Review E*, vol. 49, no. 2, pp. 1685–1689, 1994. DOI: 10.1103/PhysRevE.49.1685.

[14] K. Kiyono, "Estimating long-range temporal correlations in non-stationary time series," *Progress of Theoretical and Experimental Physics*, vol. 2015, no. 12, 12A102, 2015. DOI: 10.1093/ptep/ptv158.

A proposal to select the optimal immunization for Respiratory Syncytial Virus using a calibrated Agent-Based Model

J. Villanueva-Oller[1,*], R. J. Villanueva[2], J. C. Cortés[2], C. Pérez-Diukina[2]

(1) Departamento de Informática y Estadística, Universidad Rey Juan Carlos, Av. del Alcalde de Móstoles, s/n, 28933 Móstoles, Spain.
(2) Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Camí de Vera, s/n, 46022 València, Spain.

Abstract

Respiratory Syncytial Virus (RSV) is a common, highly contagious virus that primarily affects the respiratory tract. With symptoms similar to the common cold, it is a leading cause of lower respiratory infections in infants and young children. It is the primary cause of bronchiolitis and pneumonia in children under one year of age, accounting for 1.6% to 2.1% of RSV-associated hospitalizations in the overall population cohort.

In the Spanish Valencian Community (CVA), Nirsevimab, a long-acting monoclonal antibody, was introduced during the 2023–2024 RSV season as a universal pre-exposure prophylaxis for high-risk infants and all children under 6 months of age, with a positive impact.

The aim of prophylaxis is to protect children < 1 year in the months of highest incidence (November - February). Immunization campaigns immunize newborns just after birth, and children born outside the campaign period are immunized in the first month of the campaign, ensuring maximum protection in the months of highest incidence.

Under these premises, and using an Agent-Based Model (ABM) calibrated for RSV dynamics, which incorporates a representation of the CVA population including actual demographic data, annual birth distributions, and age-specific mortality rates, our proposal is to

a) incorporate into the model the protection provided by the immunoprophylaxis to the population of children up to 1 year old.
b) study how close the prophylaxis strategy implemented in the CVA is to the best possible, according to the model, in terms of hospitalizations and infections, with special attention to children under 6 months old.

*Javier.villanueva@urjc.es

A mathematical model with bistability dynamics explaining malaria prevalence

Gilberto González-Parra[a,1] Sharmin Sultana[b]
Luis Fernando Chaves[b]
and Daniela Lezcano[a]
(Gilberto González-Parra)
(Sharmin Sultana),sultans@iu.edu.
(Luis Fernando Chaves),lfchaves@iu.edu.
(Daniela Lezcano),daniela.lezcanocuello@student.nmt.edu.

(a) Department of Mathematics, New Mexico Tech, Socorro, New Mexico, USA.
(b) Department of Environmental and Occupational Health, School of Public Health, Indiana University, Bloomington, Indiana, USA

1 Introduction

There are several mathematical models that have been proposed to investigate the dynamics of malaria Oftentimes these models are based on systems of differential equations where the state variables represent humans grouped by the disease progression stage. In addition, other variables represent the mosquito population classified by their capacity to infect humans. These models have a variety of forms depending on how the virus is transmitted. However, there are other mathematical models that include other state variables that represent environmental factors such as climate or land.

In this work we constructed a mathematical model to describe the dynamics of Malaria by including humans, mosquitoes and land. In particular, the model considers three types of land. The first one refers to land that is used for urbanization. The second one is land devoted to agriculture and the last type is the forest land. The aim of this work is to explore from a mathematical point of view the effect of the type of land on the dynamics of malaria and more particularly on the steady states of the system. Another main aim of this work is to explain real-world data regarding malaria prevalence in different regions. This data presents a very interesting feature which is bistability. This qualitative behavior is not commonly observed in previous studies that investigated malaria with mathematical models. Moreover, we proposed a particular and plausible mathematical model that allows us to obtain the bistability behavior.

2 Mathematical model for land use

In this section we present the mathematical models for the dynamics of land and for malaria. These models are coupled through the biting rate a and the amount of cover land in agriculture land available A. A crucial part of the mathematical modeling approach is the assumption that the

[1] Gilberto.GonzalezParra@nmt.edu

biting rate a increases with the amount of cover land in agriculture. The constructed mathematical model is composed by a system of nonlinear differential equations that describes the dynamics of three land types and the malaria transmission process. The three land types considered in this study are agriculture (A), empty (E) and forest (F). The constructed mathematical model includes the human and mosquito populations in order to describe the dynamics of malaria. In particular, we have susceptible, infected non-infectious and infectious mosquitoes.

2.1 The human population model

This submodel categorizes the human population into distinct classes.vS stands for the susceptible or vulnerable individuals, while E stands for the exposed class. Those who are infected are further separated into those who do not undergo clinical therapy (I) and those who do (C). Some people recover from malaria naturally, and after that, they go through a phase known as the *refractory period*, during which they are impervious (immune) to malaria. R stands for the class that is part of the refractory phase. They must return to the susceptible class(S) in order to regain the infection. In contrast to the Chitnis's concept [1], this immunological class does not aid in transmission. Exposure to infected mosquito bites (the strength of transmission) determines the infectious period for superinfection [2], [3], [4]. This human submodel was presented in [3]. The ODE system is:

$$\begin{aligned} \dot{S} &= B - \beta S + \sigma R - \delta_H S + \rho C, \\ \dot{E} &= \beta S - \delta_H E - \gamma_H E, \\ \dot{I} &= (1-\chi)\gamma_H E - \eta\beta I + \nu C - rI - \delta_H I, \\ \dot{R} &= -\sigma R + rI - \delta_H I, \\ \dot{C} &= \chi\gamma_H E + \eta\beta I - (\nu + \rho + \delta_H)C. \end{aligned} \tag{1}$$

Variables	Description
S	Susceptible class of human who can acquire malaria
E	Exposed class of human in the latent phase
I	Infected individuals not receiving clinical treatment
R	Infected patient recovered from malaria naturally
C	Clinically treated infected patients
N	Total human population size
W	Total number of infectious mosquitoes
Parameters	**Description**
B	Re-population of susceptible by migration or birth
δ_H	Per capita human mortality rate
γ_H	Incubation rate for exposed individuals
β	Force of infection
σ	Rate of immunity loss in recovered individuals
ρ	Recovery rate from clinically treated patient to susceptible
ν	rate of transition from clinically treated patient C to infected class I
r	Recovery rate for untreated infected individuals
χ	Probability of clinical detection and given medical attention
η	Reduction factor for reinfection in the infected class

Table 1: Description of the variables and parameters of the mathematical model (1).

In malaria-endemic regions, higher transmission intensity—driven by an abundance of infectious mosquitoes and frequent mosquito bites—leads to slower recovery from infection. This phenomenon

is explained by the transmission intensity (Λ), which quantifies how often humans are exposed to the parasite through mosquito bites. The transmission intensity is given by the equation:

$$\Lambda = a\frac{W}{N} \tag{2}$$

where, a is the biting rate of mosquitoes, W represents the number of infectious mosquitoes, and N is the total human population. In [3] the effective recovery rate is expressed by

$$r(\Lambda) = \frac{\Lambda}{exp(\Lambda/r_0) - 1} \tag{3}$$

Then in [3], rate of immunity loss ($\sigma(\Lambda)$) in human is expressed as the following

$$\sigma(\Lambda) = \frac{\Lambda}{exp(\Lambda/\sigma_0) - 1} \tag{4}$$

The transmission rate (β) of malaria is expressed as the following equation

$$\beta = b\frac{W}{N}a \tag{5}$$

where, b is the probability that a mosquito bite will successfully transmit malaria from the mosquito to the human, W is the total number of infectious mosquitoes in the area, which reflects the density of malaria-transmitting mosquitoes. The following equation expands on the transmission dynamics of malaria by incorporating both local transmission (via mosquitoes) and external sources of infection. The equation is

$$\beta = b\frac{W}{N}a + \beta_e \tag{6}$$

β_e is the external force of infection, which represents malaria transmission from external sources, such as infected individuals entering the area from other regions.

2.2 The mosquito population model

In [3] Larvae and adults are regarded as the most critical phases in a mosquito's life cycle. Thus, the simplified two-stage sub-model for mosquitoes can be expressed as follows:

$$\dot{L} = f\,M\,\frac{(K_0 - L)}{K_0} - (\delta_L + d_L)\,L, \tag{7}$$

$$\dot{M} = d_L\,L - \delta_M\,M \tag{8}$$

Fecundity of the female mosquito depends on blood meal i.e. biting rate a. So, $f = N_L\,a$, N_L is the number of larval recruits per oviposition (Egg-laying) event. d_L and δ_L are the development rate and the death rate of Larva. K_0 is the carrying capacity of the larval habitat, which limits larval growth and is influenced by environmental factors like rainfall. δ_M is the death rate of the mosquito.

In the study of mosquito populations and vector-borne disease transmission, adult mosquitoes(M) are typically categorized into three classes based on their infection status and ability to transmit pathogens, such as those responsible for diseases like malaria, dengue, and Zika. These three classes are (i) Susceptible Non-Infected Mosquitoes (X), (ii) Infected Non-Infectious Mosquitoes (V) and (iii) Infectious Mosquitoes (W) [3]. In other words, $M = X + V + W$. These classes are crucial for understanding the dynamics of disease spread and the effectiveness of control strategies. The sub-model based on the adult mosquitoes is given below

Variables	Description
L	Larval stage
M	Mosquuito stage
Parameters	**Description**
f	Per capita rate of facundity
a	Biting rate of mosquito
d_L	Development rate at which larvae transition to adults
δ_L	Density-independent larval mortality rate
δ_M	Death rate of Mosquito
N_L	No of larval recruits per oviposition (Egg-laying) event
K_0	Carrying capacity of the larval growth
y	Fraction of infectious humans in the total human population
$G(t)$	Transition rate from exposed to infectious mosquitoes.

Table 2: Description of the variables and parameters of the mathematical model (7)-(8)

$$\begin{aligned} \dot{X} &= -c\,a\,y\,X + d_L\,L - \delta_M\,X, \\ \dot{V} &= c\,a\,y\,X - G(t) - \delta_M\,V \\ \dot{W} &= G(t) - \delta_M\,W \end{aligned} \tag{9}$$

The sum of the equations for X, V, and W from model (9) leads to a simplified equation (8) for the total mosquito population M.

$$\frac{d(X+V+W)}{dt} = d_L\,L - (X+V+W)\delta_M.$$

2.3 The Land use change model

Land use transitions between forests, agricultural land, and urban areas are complex processes influenced by various economic, social, and environmental factors. Forests are often cleared to accommodate agriculture or urban development due to increasing demand for food, grazing land, and urban expansion driven by population growth and economic development [5]. The proposed Land use change model is

$$\begin{aligned} \dot{F} &= c_1\,U + r_2\,A - d_1\,F, \\ \dot{A} &= d_1\,F - r_2\,A - d_2\,A, \\ \dot{U} &= d_2\,A - c_1\,U. \end{aligned} \tag{10}$$

Thus, the state variable F represents the amount of free vegetation land, A represents the amount of land devoted to agriculture and the variable E represents the amount of urbanized land.

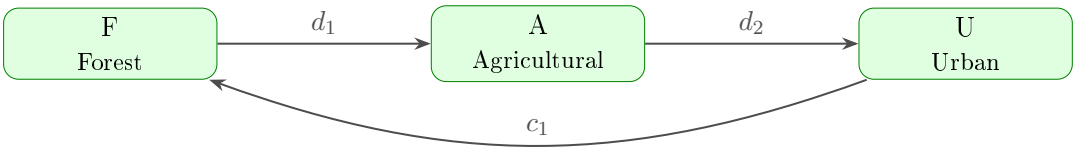


Figure 1: Land use dynamics

2.4 Relation between mosquito biting rate and the land use change

The linear relation $a = x_1 + x_2\,A$ links land use (agriculture) with mosquito dynamics by making the biting rate a dependent on agricultural land (A). This integration allows the land-use model to influence the mosquito-human model and highlights the role of land management in malaria control. Here x_1 represents the baseline mosquito biting rate in the absence of agricultural land ($A = 0$) and x_2 indicates how much the biting rate increases per unit increase in agricultural land.

The previous equation provides a mechanism to couple the land-use model and the mosquito-human model. Land-Use Model determines A (agricultural land fraction) based on forest and urban land dynamics. Mosquito-Human Model uses A to compute a, which affects malaria transmission rates. This coupling allows us to study how land-use changes (e.g., agricultural expansion) affect malaria prevalence and vice versa.

Thus the coupled mosquito-human-land use model becomes

$$\begin{aligned}
\dot{L} &= f\,M\,\tfrac{(K_0-L)}{K_0} - (\delta_L + d_L)\,L,\\
\dot{X} &= -c\,a\,y\,X + d_L\,L - \delta_M\,X,\\
\dot{V} &= cayX - n_P\,\gamma_P\,V - \delta_M\,V\\
\dot{W} &= n_P\,\gamma_P\,V - \delta_M\,W\\
\dot{F} &= c_1 - c_1\,A + r_2\,A - c_1\,F - d_1\,F,\\
\dot{A} &= d_1\,F - r_2\,A - d_2\,A,\\
a &= x_1 + x_2\,A
\end{aligned} \tag{11}$$

Here, y is the fraction of human population transmitting the parasite to mosquito vector. Note that $y = \frac{I+C}{N}$. c is per bite human to mosquito infectivity probability and plasmodium related parameters n_P, γ_P are the exposed number and the incubation rate.

3 Bifurcation diagram and bistability

In this section, we present a preliminary bifurcation analysis that allows us to explain different behaviors and which are supported by data regarding malaria and land use. We have found the steady states of the three models of the system separately. It can be proven that the basic reproduction number $\mathcal{R}_0$ is given by

$$\mathcal{R}_0 = a\sqrt{bc}\sqrt{\frac{K d_L}{N\delta_M\,f}\left(f - \delta_M\left(1 + \frac{\delta_L}{d_L}\right)\right)}\sqrt{\frac{1}{(\delta_H + r_0)\,\delta_M}}\left(\frac{n_p\,\gamma_p}{n_p\,\gamma_p + \delta_M}\right)^{\frac{n_p}{2}}\left(\frac{n_H\,\gamma_H}{n_H\,\gamma_H + r_0 + \delta_H}\right)^{\frac{n_H}{2}}.$$

Figure 2 shows a bifurcation diagram that illustrates the relationship between F (forest land fraction) and y (the infectious human fraction). It shows how changes in forest land F influence the equilibrium states of y, revealing stability and transitions in the system.

4 Conclusions

In this preliminary work, we proposed a mathematical model that presents bistability and which agrees with real-word data from Malaria in Panama. We found the conditions such that the proposed mathematical model displays a backward bifurcation, leading to bistable dynamics. Future, works will include a full stability and bifurcation analysis using different mathematical techniques including the Center Manifold theory.

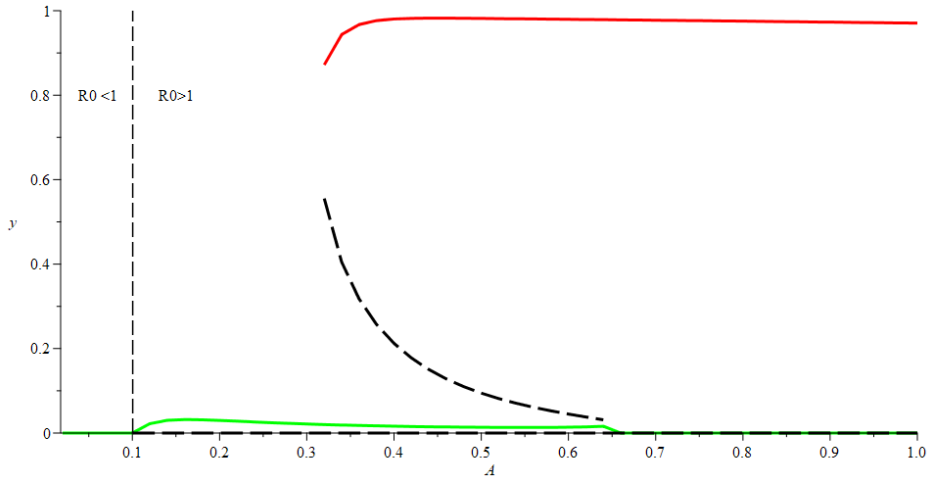


Figure 2: Relationship between saddle-node bifurcation and the agricultural land using the mathematical model (11).

References

[1] N. Chitnis, J. M. Cushing, and J. Hyman, "Bifurcation analysis of a mathematical model for malaria transmission," *SIAM Journal on Applied Mathematics*, vol. 67, no. 1, pp. 24–45, 2006.

[2] J. L. Aron and R. M. May, "The population dynamics of malaria," in *The population dynamics of infectious diseases: theory and applications*, Springer, 1982, pp. 139–179.

[3] D. Alonso, A. Dobson, and M. Pascual, "Critical transitions in malaria transmission models are consistently generated by superinfection," *Philosophical Transactions of the Royal Society B*, vol. 374, no. 1775, p. 20 180 275, 2019.

[4] K. Dietz, L. Molineaux, and A. Thomas, "A malaria model tested in the african savannah," *Bulletin of the World Health Organization*, vol. 50, no. 3-4, p. 347, 1974.

[5] D. Pimentel, "Environmental and economic costs of the application of pesticides primarily in the united states," *Environment, development and sustainability*, vol. 7, pp. 229–252, 2005.

Impact of temperature on the life cycle and vectorial capacity of Aedes aegypti, Puerto Rico, 2022-2025

Juan Ángel Aledo[1], Juan C. Orengo[2],[1] Mario Alberto Espinoza[3] and Rafael-J. Villanueva[4]

(1) Departamento de Matemáticas, Universidad de Castilla-La Mancha, Albacete, Spain.

(2) Public Health Program, Ponce Health Sciences University, Ponce, Puerto Rico, USA.

(3) Department of Computer Science, Universidad Internacional del Ecuador, Quito, Ecuador.

(4) Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Valencia, Spain.

Introduction

Dengue is an arboviral disease caused by the dengue virus, transmitted to humans by arthropods, with *Aedes aegypti* being the main vector in Puerto Rico [1]. Dengue disease is endemic in Puerto Rico, with five epidemics declared in the last 30 years, totaling approximately 100,000 cases [2]. The purpose of this study was to assess, using real-world data, the period of the year with the highest transmission capacity of the virus, based on vectorial capacity (r_{VC}) and daily development rate (r_T) that are temperature-dependent [3], [4], [5].

Methods

A retrospective, ecological, observational, time-series study was implemented for the period from January 1, 2022, to December 31, 2025. The climatic data sources were the National Weather Service, National Oceanic and Atmospheric Administration (NOAA) and Weather Underground [6], [7]. The r_{VC}, which is the vector's capacity to transmit the virus within a population, was calculated using the Ross-McDonald equation [3]. We used the Sharpe-DeMichele equation [8] to calculate the r_T from egg hatching until the adult mosquito emerges; all parameter values, with the exception of temperature, were taken from Rueda et al. (1990) [4], [8].

Results

During 2022–2025, the temperature remained optimal for virus transmission (21.39°C–30.22°C). The r_T between egg hatching and adult mosquito emergence ranged from 6.26 to 14.72 days. The r_{VC} was > 0.2 (days^{-1}) throughout the study period, which falls within the minimum threshold range associated with epidemic potential. Monthly analysis revealed that temperature, r_T, and

[1]jorengo@psm.edu

r_{VC} for June, July, August, and September were not significantly different from one another, but showed a statistically significant difference ($p < 0.01$) compared to the other months of the year.

Conclusions

The optimal months for virus transmission and mosquito development were June, July, August, and September, which coincide with the highest expected incidence of dengue fever in Puerto Rico. Integrated dengue control and prevention strategies should be maintained year-round given the persistent risk of epidemic transmission. Incorporating local climatological data into mathematical models would improve the accuracy of vector-borne transmission risk assessments.

Acknowledgments

Juan A. Aledo was supported by PID2022-139293NB-C32 (MICIU/AEI/10.13039/501100011033 and ERDF, EU) and 2025-GRIN-38476 (Universidad de Castilla-La Mancha and ERDF A way of making Europe). Rafael Villanueva was supported by the Vicerrectorado de Investigación de la Universitat Politècnica de València PAID-11-25.

Conflict of interest

The authors declare no conflict of interest.

References

[1] D. J. Gubler, R. J. Novak, E. Vergne, N. A. Colon, M. Velez, and J. Fowler, "Aedes (Gymnometopa) mediovittatus (Diptera: Culicidae), a potential maintenance vector of dengue viruses in Puerto Rico," *Journal of Medical Entomology*, vol. 22, no. 5, pp. 469–475, Sep. 1985. DOI: 10.1093/jmedent/22.5.469.

[2] Pan American Health Organization, *Dengue data and analysis*, https://www.paho.org/en/arbo-portal/dengue-data-and-analysis/dengue-analysis-country, Accessed: April 16, 2026, Apr. 2026.

[3] J. Liu-Helmersson, H. Stenlund, A. Wilder-Smith, and J. Rocklöv, "Vectorial capacity of Aedes aegypti: Effects of temperature and implications for global dengue epidemic potential," *PLOS ONE*, vol. 9, no. 3, e89783, Mar. 2014. DOI: 10.1371/journal.pone.0089783.

[4] L. M. Rueda, K. J. Patel, R. C. Axtell, and R. E. Stinner, "Temperature-dependent development and survival rates of Culex quinquefasciatus and Aedes aegypti (Diptera: Culicidae)," *Journal of Medical Entomology*, vol. 27, no. 5, pp. 892–898, Sep. 1990. DOI: 10.1093/jmedent/27.5.892.

[5] E. Abbasi, "Aedes aegypti and dengue: Insights into transmission dynamics and viral lifecycle," *Epidemiology and Infection*, vol. 153, e88, Aug. 2025. DOI: 10.1017/S0950268825100320.

[6] National Weather Service and National Oceanic and Atmospheric Administration (NOAA), *National weather service*, https://www.weather.gov/, Accessed: January 8, 2026, Jan. 2026.

[7] Weather Underground, *Weather underground*, https://www.wunderground.com/, Accessed: January 8, 2026, Jan. 2026.

[8] P. J. Sharpe and D. W. DeMichele, "Reaction kinetics of poikilotherm development," *Journal of Theoretical Biology*, vol. 64, no. 4, pp. 649–670, Feb. 1977. DOI: 10.1016/0022-5193(77)90265-x.

Memory Effects or Modeling Errors? (In)Consistency of Fractional-Order Systems in Epidemiology

Carla M. A. Pinto[♭],[1] Julia Calatayud[♮] and Marc Jornet

(♭) School of Engineering, Polytechnic of Porto
Porto, Portugal.

(♮) Departament de Matemàtiques, Universitat Jaume I,
12071 Castellón de la Plana, Castellón, Spain,
() .

1 Introduction

While compartmental systems effectively capture transitions in epidemiological and sociological phenomena, the role of Caputo fractional calculus remains under scrutiny. This work provides an expository analysis of fractional modeling, specifically addressing how the non-Markovian property relates to variable latency and transition risks. By comparing these formulations with traditional delay differential equations, we highlight several mechanistic inconsistencies inherent in the fractional approach. Ultimately, we argue that the practical contribution of fractional calculus to these models remains mathematically and biologically unclear [1], [2], [3].

References

[1] J. Calatayud, M. Jornet, and C. M. Pinto, "On the interpretation of caputo fractional compartmental models," *Chaos, Solitons & Fractals*, vol. 186, p. 115 263, 2024, ISSN: 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2024.115263.

[2] J. Calatayud, M. Jornet, and C. Pinto, "On partial caputo fractional models," *Computational and Applied Mathematics*, vol. 44, p. 241, 2025.

[3] J. M. Pinto Carla M. A. Calatayud J., "To fractionalize or not to fractionalize epidemiological compartmental models, that is the question," in *Proceedings of NODYCON2026*, W. Lacarbonara, Ed., New Jersey, Portugal: Springer, 2026.

[1] cap@isep.ipp.pt

A framework for time-varying risk assessment of *P. vivax* transmission in Republic of Korea

Woo-Sik Son[♭], Ah-Young Lim[♮], Dong-Uk Hwang[♭] and Kyeongah Nah[♭,1]

(♭) National Institute for Mathematical Sciences,
70, Yuseong-daero 1689 beon-gil, Yuseong-gu, Daejeon, 34047, Republic of Korea.
(♮) London School of Hygiene & Tropical Medicine,
Keppel Street, London, WC1E 7HT, United Kingdom.

1 Introduction

Eliminating *Plasmodium vivax* (*P. vivax*) malaria remains a challenge in temperate regions like Republic of Korea due to its complex biology, including prolonged dormant liver stages and highly seasonal transmission. Traditional epidemiological surveillance often struggles to capture the dynamic transmission potential in these low-incidence settings. This study develops a mechanistic modeling framework to estimate the time-varying reproduction number by incorporating biologically explainable, seasonality dependent generation intervals.

2 Methods

2.1 A subsection

We utilize weekly data on symptom onset, mosquito abundance, and temperature from main endemic regions in South Korea between 2013 and 2024. The generation interval was reconstructed by dividing the transmission cycle into four components: human-to-mosquito transmission, temperature-dependent extrinsic incubation in mosquitoes, mosquito-to-human transmission, and a bimodal intrinsic incubation period in humans. Using these seasonally constrained intervals, we generalized renewal equation-based methods to estimate the case reproduction number (R_t), the average number of secondary cases generated by an individual who becomes infected at calendar time t.

3 Results

Our analysis reveals that while the absolute number of secondary cases peaks during the summer, the average case reproduction number is highest in early spring. This indicates that

[1] knah@nims.re.kr

spring-onset cases, though fewer in number, possess a significantly higher per-capita potential to propagate transmission compared to summer cases. Furthermore, the case reproduction number frequently exceeded the epidemic threshold ($R_t > 1$), indicating that elimination has not yet been achieved.

4 Conclusions

We developed a mechanistic modeling framework to estimate the time-varying reproduction number by incorporating biologically explainable, seasonality dependent generation intervals. The findings emphasize a critical window for malaria intervention in early spring. Prioritizing timely diagnosis and early treatment of spring cases could interrupt transmission chains before the seasonal peak, optimizing resource allocation for malaria elimination. To support the exploration of these dynamics, we provide a web-based interactive tool that visualizes estimated transmission timelines and reproduction numbers.

Acknowledgements

This work was supported by the National Institute for Mathematical Sciences (NIMS) grant funded by the Korean government (No. NIMS-B25730000). AL was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R1A6A3A03061207).

References

[1] W.S. Son, A.Y. Lim, D.U. Hwang, K. Nah, "Estimating the effective reproduction number of *Plasmodium vivax* malaria transmission in South Korea," submitted

[2] D.U. Hwang, "Web-based interactive plot visualizing the symptom onset timing of infectees based on their infector's symptom onset", https://observablehq.com/@nims-epid/interactive-plot-for-p-vivax-malaria

Numerical validation of the quasi-stationary distribution in a stochastic SID coinfection Markov model

Carlos Andreu-Vilarroig[♭1], Sten Madec[♮] and Erida Gjini[♯]

(♭) Instituto de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

(♮) Institut Denis Poisson, University of Tours, Tours, France.

(♯) Center for Computational and Stochastic Mathematics, University of Lisbon, Lisbon, Portugal.

1 Introduction

Co-infection, the simultaneous infection of a host by multiple pathogens, is a widespread phenomenon with significant implications for disease dynamics [1]. While deterministic compartmental models have been the standard tool for its mathematical study, they fail to capture the stochastic variability inherent to transmission in finite populations. Continuous-time Markov chains (CTMCs) offer a principled alternative, allowing one to characterise the full probability distribution of the system state and study extinction events [2]. Despite a rich body of theory for classical Markov epidemic models [3], the stochastic analysis of co-infection remains largely unexplored. Building on a stochastic Susceptible-Infected-Double-infected (SID) model, we focus here on the Ornstein–Uhlenbeck (OU) approximation of the quasi-stationary distribution (QSD), a systematic study of how model parameters shape it, and its numerical validation via Gillespie simulations.

2 Methods

2.1 The stochastic SID model and its quasi-stationary distribution

We consider a closed population of constant size N_p divided into susceptible $S(t)$, infected $I(t)$, and double infected (or coinfected) $D(t)$ individuals. The model dynamics are as follows:

- a susceptible becomes infected at rate $\beta(I + D)/N_p$, where $\beta > 0$ is the infection rate.
- an infected individual acquires a second infection at the attenuated rate $k\beta(I+D)/N_p$, where $k \in (0, 1)$ is the attenuation factor of the double infection.
- and both classes I and D clear at rate m, where $m > 0$ is clearance (death + recovery) rate.

[1] caranvi1@upv.es

The bivariate CTMC $\mathbf{X}(t) = (I(t), D(t))^T$ evolves on $\Omega = \{(i,d) : i,d \geq 0,\ i+d \leq N_p\}$ and its probability distribution $\mathbf{p}(t)$ satisfies the Kolmogorov forward equations

$$\frac{d\mathbf{p}}{dt} = Q\mathbf{p},$$

with Q a tridiagonal block generator matrix [4]. The absorbing state $(0,0)$ is accessible, so the infection goes extinct with probability 1. The behaviour of practical interest is described by the *quasi-stationary distribution* (QSD) $\mathbf{q}^*$, the limiting conditional distribution given non-extinction. Since the chain restricted to $\Omega \setminus \{(0,0)\}$ is irreducible, $\mathbf{q}^*$ is unique and coincides with the positive eigenvector of the reduced generator $\tilde{Q}$ associated with its dominant eigenvalue ρ_1 [3].

2.2 Ornstein–Uhlenbeck approximation and parameter effects on the QSD

For N_p large and $R_0 = \beta/m > 1$, the normalized fluctuations $\delta\mathbf{x} = \mathbf{X}(t)/N_p - \mathbf{x}^*$ of the random vector $\mathbf{X}(t)$ around the normalized endemic equilibrium point

$$\mathbf{x}^* = \mathbf{X}^*/N_p = \left(\frac{1 - R_0^{-1}}{1 + k(R_0 - 1)},\ \frac{k(R_0-1)(1-R_0^{-1})}{1+k(R_0-1)} \right)^T$$

can be linearized and approximated to an Ornstein–Uhlenbeck process [5] with the local drift matrix $\mathcal{B}$ and diffusion matrix $\mathcal{S}$, which converges to the QSD when $t \to \infty$:

$$\delta\mathbf{x}(t) \ \sim\ \mathrm{OU}(\mathcal{B}, \mathcal{S}) \ \xrightarrow[t\to\infty]{}\ \mathcal{N}_2(\mathbf{0},\ \Sigma)\,, \tag{1}$$

$$\mathbf{X}(t) \ \xrightarrow[t\to\infty]{}\ \mathbf{q}^* \ \approx\ \mathcal{N}_2(N_p\mathbf{x}^*,\ N_p\Sigma)\,, \tag{2}$$

where Σ solves the Lyapunov equation $\mathcal{B}(\mathbf{x}^*)\Sigma + \Sigma\mathcal{B}^T(\mathbf{x}^*) = -\mathcal{S}(\mathbf{x}^*)$. Since m can be cancelled in the Lyapunov equation, the approximation (1) is fully characterized by the reduced set (R_0, k, N_p).

A theoretical analysis three parameters shape the QSD in distinct and interpretable ways:

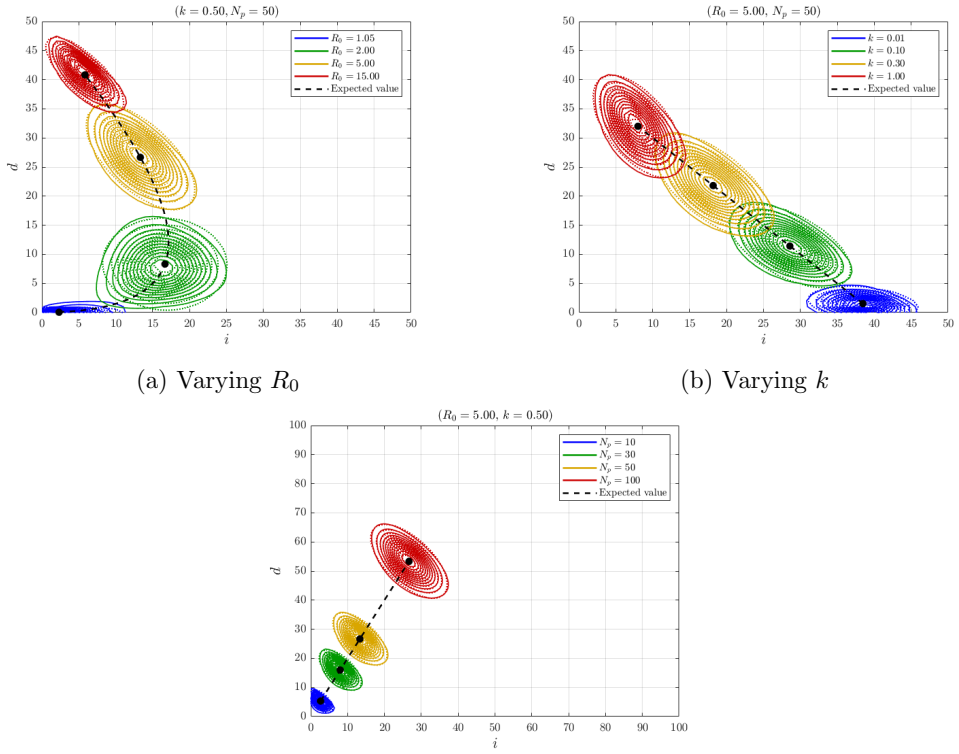
- The reproduction number R_0 controls both location and concentration. Near the threshold ($R_0 \to 1^+$) the QSD collapses towards the origin with large relative fluctuations, while for large R_0 it concentrates on the double-infected class with $\Sigma \sim \Sigma_\infty / R_0 \to \mathbf{0}$.
- The attenuation factor k governs the balance between infected classes. For $k \to 0^+$ the QSD reduces to that of a single-infection model with fluctuations confined to the I direction, while for $k \to 1^-$ both compartments coexist with increasing dominance of D and a negative correlation between their fluctuations.
- Finally, N_p rescales mean and variance uniformly, Moreover, absolute fluctuations grow as $\mathcal{O}(\sqrt{N_p})$ while relative fluctuations shrink as $1/\sqrt{N_p}$, recovering the deterministic endemic equilibrium in the limit $N_p \to \infty$.

3 Results

We validate the OU approximation (1) against the exact QSD, computed numerically as the dominant eigenvector of $\tilde{Q}$. Two complementary experiments are carried out.

Experiment 1: parameter effect on the QSD (Figure 1). We compute both the exact QSD and the OU approximation while varying each parameter (R_0, k, N_p) marginally, keeping the other two fixed. For each parameter configuration, we overlay the exact QSD and the OU

bivariate normal in the (I, D) plane, allowing a direct visual assessment of how well the Gaussian approximation tracks the shift in location and the change in spread and orientation of the true distribution as the parameters vary. This experiment serves as a numerical confirmation of the analytical parameter effects described in Section 2.2.



(a) Varying R_0

(b) Varying k

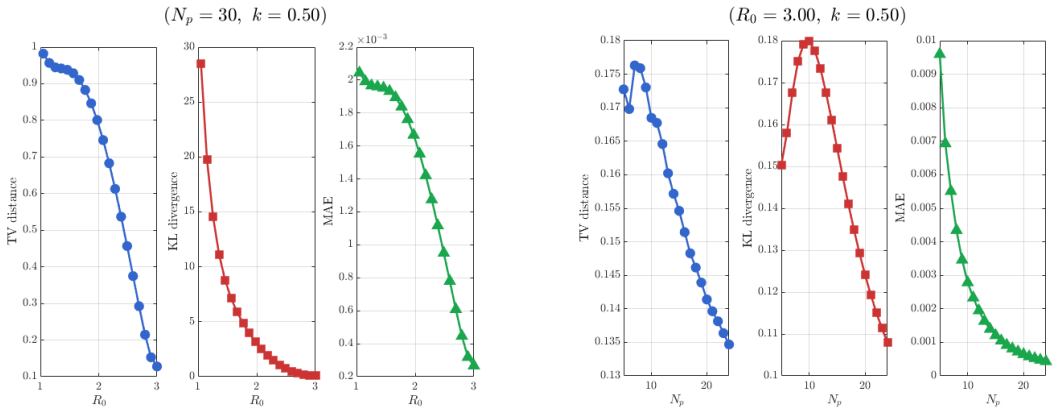
(c) Varying N_p

Figure 1: Exact QSD (solid line) and OU approximation (dashed line) in the (I, D) plane under marginal variations of R_0 (a), k (b), and N_p (c), with the remaining parameters fixed.

Experiment 2: approximation error near the endemic boundary (Figure 2). The OU approximation relies on a linearization around $\mathbf{x}^*$ and is expected to deteriorate as $R_0 \to 1^+$ (where the endemic equilibrium vanishes and the distribution is no longer well approximated by a Gaussian) or for small N_p (where demographic noise is large relative to the mean). To quantify this, we compute three error metrics between the exact QSD and the OU approximation: the total variation distance (TVD), the mean absolute error (MAE) and the Kullback–Leibler (KL) divergence. These are reported as functions of R_0 and N_p separately. As predicted by the theory, all three metrics decay (in this case, exponentially) as R_0 increases above 1 and as N_p grows, confirming that the approximation improves in the strongly supercritical and large-population regimes. The dependence on k is comparatively mild, consistent with the fact that k does not alter the population rescaling.

4 Conclusions

In this work, we have studied the quasi-stationary distribution of a stochastic SID coinfection Markov model and its Ornstein–Uhlenbeck Gaussian approximation. The approximation is fully characterized by the reduced parameter set (R_0, k, N_p) and provides closed-form expressions for



(a) Error metrics as a function of R_0.

(b) Error metrics as a function of N_p.

Figure 2: TVD, MAE, and KL divergence between the exact QSD and the OU approximation as a function of R_0 (a) and N_p (b), for selected values of k.

both the location and the covariance structure of the endemic fluctuation regime.

The numerical experiments confirm the theoretical predictions on two fronts. First, systematic comparison of the exact QSD and the OU approximation under marginal parameter variations shows that the Gaussian approximation accurately tracks the qualitative and quantitative changes in the distribution: the shift toward the double-infected class as R_0 increases, the rebalancing between infected compartments as k varies, and the concentration around the uniform reescalation as N_p grows. Second, error analysis through TVD, MAE and KL divergence quantifies the validity regime of the approximation: the error decays consistently as R_0 moves away from the epidemic threshold and as N_p increases, confirming that the OU approximation is reliable in the highly supercritical and large-population regimes.

Beyond validation, the numerical results shed light on the limits of the approximation. Near the epidemic threshold ($R_0 \to 1^+$) and for small population sizes, the linearization underlying the OU approach breaks down and the error metrics grow significantly, signalling that in these regimes a more refined stochastic analysis would be required. These findings provide a practical guide for the applicability of the OU approximation in realistic co-infection scenarios.

References

[1] E. C. Griffiths, A. B. Pedersen, A. Fenton, and O. L. Petchey, "The nature and consequences of coinfection in humans," *Journal of Infection*, vol. 63, no. 3, pp. 200–206, 2011.

[2] T. Britton and H. Andersson, *Stochastic Epidemic Models and Their Statistical Analysis* (Lecture Notes in Statistics). New York: Springer, 2000, vol. 151.

[3] I. Nåsell, "Stochastic models of some endemic infections," *Mathematical Biosciences*, vol. 179, no. 1, pp. 1–19, 2002.

[4] L. J. S. Allen, *An Introduction to Stochastic Epidemic Models* (Lecture Notes in Mathematics), F. Brauer, P. van den Driessche, and J. Wu, Eds. Berlin: Springer, 2008, vol. 1945, pp. 81–130.

[5] E. J. Allen, *Modeling with Itô Stochastic Differential Equations* (Mathematical Modelling: Theory and Applications). Dordrecht: Springer, 2007, vol. 22.

Part VI

Linear Algebra, Matrix Analysis and Applications

Optimal order-24 matrix exponential approximation with applications to generative AI flows

Jorge Sastre[♭],[1] José Miguel Alonso[♮], Javier Ibáñez[†], Julio Andrés[♭], Roger B. Dannenberg[♯], Daniel Faronbi[‡], and Nuria Lloret[*]

(♭) I.U. de Telecomunicaciones y Aplicaciones Multimedia, (♮) Instituto de Instrumentación para Imagen Molecular, (†) I.U. de Matemática Multidisciplinar, (∗) Instituto de Diseño para la Fabricación y Producción Automatizada, Universitat Politècnica de València
Camino de Vera s/n, Valencia, Spain.

(♯) Carnegie Mellon University, Pittsburgh, USA.

(‡) Music and Audio Research Lab, New York University, New York, USA.

1 Introduction

The computation of the matrix exponential is a core operation in numerical linear algebra [1], [2], [3]. Scaling and squaring methods combined with polynomial approximations are arising for high-performance computing [4], [5], [6], [7]. Recent research has demonstrated that specialized evaluation schemes [5], [7], [8], [9] can surpass the Paterson–Stockmeyer limit [10], being more efficient than the algorithms based on Padé approximants [3].

This work presents a degree–32 optimized polynomial that matches the matrix exponential Taylor expansion, denoted by

$$T_m(A) = \sum_{k=0}^{m} \frac{A^k}{k!}, \tag{1}$$

up to degree 24, while requiring only five matrix multiplications. This optimized polynomial approximation can be written as different evaluation formulas with the same cost which are equivalent in exact arithmetic, however they are not equivalent in finite arithmetic. Tests will be done to select the most accurate evaluation formula. Then, an error analysis taking into account the absolute and relative forward error and the absolute backward error is provided to enable a simple algorithm for variable precision tolerance ϵ higher or equal to the unit roundoff error in a target precision.

The proposed method is particularly suitable for generative AI flows, where efficiency and invertibility are paramount [11], representing an improvement over the results in [12]. While previous work [12] relied on Taylor-based approximations up to degree 15 and required two Taylor remainder terms to determine the scaling parameter, this work demonstrates that a single remainder term is sufficient for practical applications when utilizing Taylor-based approximations, including the proposed order-24 approximation.

[1] jsastrem@upv.es

2 Evaluation of the 24+ Approximation

Following the nested search methodology for matrix function approximations from [13, Sec. 2], see [4], [9], a new evaluation formula for the matrix exponential has been derived. This approximation matches the Taylor series up to $n = 24$ and reaches degree $m = 32$, and following the notation from [4] this kind of Taylor based approximation will be denoted as a $m = 24+$ approximation. The coefficients of the corresponding evaluation formulas were obtained through a large-scale search using MATLAB's `vpasolve` with random initialization and 32 variable precision decimal digits. The search space involved thousands of degree-32 Taylor-based evaluation schemes with an approximation order $m = 23+$. Only schemes with real coefficients were selected as initial guesses to ensure convergence toward the final order-24 approximation, which matches the Taylor series expansion up to degree 24. The original version of the scheme, denoted by V0, is defined as:

$$y_0 = A^2(c_1A^2 + c_2A), \tag{2a}$$

$$y_1 = (y_0 + c_3A^2 + c_4A)(y_0 + c_5A^2) + c_6y_0, \tag{2b}$$

$$y_2 = (y_1 + c_7A^2 + c_8A)(y_1 + c_9y_0 + c_{10}A^2 + c_{11}A) + c_{12}y_1 + c_{13}y_0 + c_{14}A^2, \tag{2c}$$

$$y_3 = (y_2 + c_{15}y_0 + c_{16}A^2 + c_{17}A)(y_2 + c_{18}y_1 + c_{19}A^2 + c_{20}A) + c_{21}y_2 + c_{22}y_1 + c_{23}y_0 + c_{24}A^2 + A + I, \tag{2d}$$

where the coefficients c_i with 32 decimal digit precision are given in Table 1. Evaluating (2) requires only five matrix products: A^2 and four subsequent multiplications (y_0, y_1, y_2, y_3).

Table 1: Coefficients c_i from (2) with 32 decimal digit precision

i	c_i
1	0.00003066988351753155671655530404559 7
2	0.00017693205574968558387865926764 94
3	0.00156301553635919022242809459567 9
4	0.19228195836715997126659516746105
5	0.00717765845832223226922741637827 4
6	1.96056906627053561366943164384 62
7	0.02350698208745188231018302106 19
8	0.14247044623861512093553455514624
9	5.78928996266281992422739750209 6
10	0.00512695081041867546638117765568 37
11	-0.49007616198291019496884135528642
12	2.96277447272722296578903504773 31
13	-224.25107960412600375437856015944
14	0.06982136948861304251433252230107 5
15	249.33865472444327886061273617255
16	-0.09308055894614124457539636023497
17	-1.56409427861893129776813737009 37
18	83.46036957811685067300865672746 4
19	0.84555981452154771700210428212474
20	-1.84087731333089224156626234287 61
21	203.10675518177674007955246073858
22	779.15245428143528269003748435 96
23	51902.24568688138729129607400922 9
24	-2.37930567342023826024019587170 65

Table 2: Relative errors of the computed polynomial coefficients b_k with respect to the exact Taylor series coefficients $1/k!$ for degrees $k = 25$ to $k = 32$.

Degree k	Relative Error
25	6.3464×10^{-2}
26	1.0328×10^{-1}
27	1.7358×10^{-1}
28	2.7075×10^{-1}
29	4.3333×10^{-1}
30	5.6976×10^{-1}
31	7.0290×10^{-1}
32	7.9400×10^{-1}

Writing y_3 from (2d) as a polynomial

$$y_3 = P_m(A) = \sum_{k=0}^{32} b_k A^k, \tag{3}$$

the relative errors of the resulting coefficients b_k with respect to the corresponding Taylor coefficients $1/k!$, i.e. $|1/k! - b_k|k!$, for degrees $k = 25$ to 32 are given in Table 2. Note that for those terms, the corresponding 24+ approximation for the matrix logarithm found in [13, Sec. 2] had higher relative errors, ranging from 0.66 to 1.11. In contrast, this new approximation ranges from 0.06 to 0.79, thereby increasing the accuracy of the approximation in exact arithmetic.

To improve accuracy in finite precision, different operation orderings have been proposed. Among them, scheme (4), denoted as V2, is defined as:

$$y_0 = A^2(c'_1 A^2 + c'_2 A), \tag{4a}$$
$$y_1 = (y_0 + c'_3 A^2 + c'_4 A)(y_0 + c'_5 A^2), \tag{4b}$$
$$\begin{aligned} y_2 = {} & (y_1 + c'_6 y_0 + c'_7 A^2 + c'_8 A)(y_1 + c'_9 y_0 + c'_{10} A^2 + c'_{11} A) \\ & + c'_{12} y_1 + c'_{13} y_0 + c'_{14} A^2, \end{aligned} \tag{4c}$$
$$\begin{aligned} y_3 = {} & (y_2 + c'_{15} y_0 + c'_{16} A^2 + c'_{17} A)(y_2 + c'_{18} y_1 + c'_{19} y_0 + c'_{20} A^2 + c'_{21} A) \\ & + c'_{22} y_2 + c'_{23} y_1 + c'_{24} y_0 + c'_{25} A^2 + A + I. \end{aligned} \tag{4d}$$

Similarly, scheme (5), denoted as V5a, is given by:

$$y_0 = A^2(A^2 + c''_1 A), \tag{5a}$$
$$y_1 = (y_0 + c''_2 A^2 + c''_3 A)(y_0 + c''_4 A^2), \tag{5b}$$
$$y_2 = (y_1 + c''_5 y_0 + c''_6 A^2 + c''_7 A)(y_1 + c''_8 y_0 + c''_9 A^2 + c''_{10} A), \tag{5c}$$
$$\begin{aligned} y_3 = {} & (c''_{11} y_2 + c''_{12} y_1 + c''_{13} y_0 + c''_{14} A^2 + c''_{15} A) \\ & \times (c''_{16} y_2 + c''_{17} y_1 + c''_{18} y_0 + c''_{19} A^2 + c''_{20} A) \\ & + c''_{21} y_2 + c''_{22} y_1 + c''_{23} y_0 + c''_{24} A^2 + A + I. \end{aligned} \tag{5d}$$

The coefficients c'_i and c''_i are determined using the values from Table 1, ensuring that expressions (4) and (5) remain algebraically equivalent to (2) in exact arithmetic.

Note that (5) reduces the number of summing terms with respect to the original (2). This operation count reduction may also reduce the rounding error. Since the matrix powers in expression (5) are presented in descending order, we also evaluated a variant using an ascending

summation order. In this version, the lowest-degree terms in each sub-expression (y_0, y_1, y_2, and y_3) are summed first to determine whether this mitigates rounding errors and enhances numerical stability in finite-precision arithmetic. One of the objectives of this work will be the selection of the most accurate evaluation formulas among all the proposed variations.

3 Stability Analysis

The stability of the scheme was verified using the criterion established in [8, Ex. 3.1]. Let $\tilde{c}_i$ denote the value of the coefficients c_i from the evaluation formulas from (5) rounded to the target precision. Let $\bar{b}_k$ denote values of the reconstructed polynomial coefficients b_k from (3) by using the rounded coefficients $\tilde{c}_i$. Therefore, the stability criterion from [8, Ex. 3.1] consists of checking that the maximum relative error of the reconstructed coefficients $\bar{b}_k(\tilde{c}_i)$ with respect to the exact Taylor series coefficients $1/k!$, given by

$$e_k = |1/k! - \bar{b}_k(\tilde{c}_i)|k!, \tag{6}$$

for $k = 0$ to $k = 24$, remains on the order of the unit roundoff u in the target precision (e.g., $u = 2^{-53} \approx 1.11 \times 10^{-16}$ for double precision).

From all the real 24+ solutions found in the large-scale MATLAB search, we selected the most stable and accurate one, given by (2) with coefficients from Table 1. In this approximation, the maximum relative reconstruction error (6) for the coefficients $\bar{b}_k$ using the $\tilde{c}_i$ values in IEEE double-precision arithmetic was less than $10u$, thereby satisfying the stability criterion. Moreover, the terms for the degrees higher than 24 presented a relative error with respect to the corresponding Taylor coefficients lower than 0.79, see Table 2. Note that the reorderings from (4) and (5) also meet the stability check, and give the same relative error for the terms of degree higher than 24.

4 Error Analysis

To analyze the accuracy of the proposed method, we summarize the framework for the scaling and squaring Taylor approximation of the matrix exponential described in [14], [15]. Let

$$R_m(A) = \sum_{k=m+1}^{\infty} \frac{A^k}{k!} \tag{7}$$

denote the remainder of the degree-m Taylor approximation to the matrix exponential. For the scaled matrix $2^{-s}A$, with $s \in \mathbb{N} \cup \{0\}$, the relationship between the computed approximation and the exact exponential is given by

$$\left(T_m(2^{-s}A)\right)^{2^s} = e^A \left(I + g_{m+1}(2^{-s}A)\right)^{2^s} = e^{A+2^s h_{m+1}(2^{-s}A)}, \tag{8}$$

where:

$$g_{m+1}(2^{-s}A) = -e^{-2^{-s}A} R_m(2^{-s}A), \quad h_{m+1}(2^{-s}A) = \log\left(I + g_{m+1}(2^{-s}A)\right). \tag{9}$$

Here, log denotes the principal logarithm. Both $g_{m+1}(2^{-s}A)$ and $h_{m+1}(2^{-s}A)$ are holomorphic functions of A in the set $\Omega_m = \{X \in \mathbb{C}^{n\times n} : \rho(e^{-X}T_m(X) - I) < 1\}$, and thus commute with A. As shown in [14], h_{m+1} and g_{m+1} relate to the backward and forward errors in exact arithmetic, respectively. [15, eq. (22)] gives the explicit series expansion of the forward relative error

$$g_{m+1}(x) = -\sum_{k=0}^{\infty} \frac{(-1)^k x^{m+1+k}}{k!m!(m+1+k)}. \tag{10}$$

Using Taylor expansion in (8) one gets

$$g_{m+1}(x) = e^{h_{m+1}(x)} - 1 = \sum_{k\geq 1} \frac{(h_{m+1}(x))^k}{k!}, \tag{11}$$

$$h_{m+1}(x) = \sum_{k\geq 1} \frac{(-1)^{k+1}(g_{m+1}(x))^k}{k} = \sum_{k\geq m+1} c_k^{(m)} x^k, \tag{12}$$

and using (10) it follows that

$$c_k^{(m)} = \frac{(-1)^k x^{m+1+k}}{k!m!(m+1+k)} \text{ for } k = m+1, \ldots, 2m+1. \tag{13}$$

Furthermore, if $\|g_{m+1}\| \ll 1$ or $\|h_{m+1}\| \ll 1$ then $h_{m+1}(2^{-s}A) \approx g_{m+1}(2^{-s}A)$ [15, eq. (16) and (17)].

In [14], [15] the scaling parameter s is chosen such that:

$$\|h_{m+1}(2^{-s}A)\| \leq \max\{1, \|2^{-s}A\|\}u, \tag{14}$$

where $u = 2^{-53}$ is the unit roundoff error in IEEE double precision. If $2^{-s}\|A\| \geq 1$, the backward error satisfies $\Delta A \leq \|A\|u$, leading to $(T_m(2^{-s}A))^{2^s} = e^{A+\Delta A} \approx e^A$. Conversely, if $2^{-s}\|A\| < 1$, we obtain $\|R_m(2^{-s}A)\| \approx \|T_m(2^{-s}A)\|u$. Since our algorithm explicitly evaluates $T_m(2^{-s}A)$, it follows that $e^{2^{-s}A} \approx T_m(2^{-s}A)$, rendering further increases in m or s unnecessary for higher accuracy.

In a similar way to [4, Sec. 4] using the MATLAB Symbolic Math Toolbox, high-precision arithmetic, 200 terms of the series, and a zero-finding algorithm, we determined the maximal value θ_{24+} of $\theta = \|2^{-s}A\|$ for the 24+ approximation (2) such that:

$$\|h_{m+1}(2^{-s}A)\| \leq \sum_{k\geq m+1} |c_k^{(m)}|\theta^k \leq \max\{1, \theta\}u, \tag{15}$$

where the Taylor based matrix exponential polynomial approximation 24+ is given by (3). Consequently, if $\|2^{-s}A\| \leq \theta_{24+}$, then condition (14) is satisfied.

Table 3 gives the corresponding θ values for the proposed 24+ approximation alongside other state-of-the-art approximations with a comparable computational cost. Let M denote the cost of a single matrix multiplication. The compared methods include the degree 12 Taylor approximation via the Paterson–Stockmeyer method [14, Tab. 1], the degree 18 Taylor approximation [5, Tab. 2], the degree 20 Taylor approximations [16, Sec. 5.2] and [17, Sec. 4], the 21+ approximation [4], and the 32-optimal degree approximation [7, Tab. 5], all of which require $5M$. We also include the Padé rational approximant [3, Tab. 3.1], which has a cost of $4M + 1D$, where D denotes the cost of solving a multiple right-hand side linear system. Note that, as per [18, Appendix 10], solving such a system generally satisfies $D \approx \frac{4}{3}M > M$.

Note that it was demonstrated in [19] that the maximum of $\|A^k\|^{1/k}$ for $k \geq m+1$ occurs in the range $k = m+1, m+2, \ldots, 2m+1$. In that reference, this maximum value is approximated by $\max\{\|A^{m+1}\|^{1/(m+1)}, \|A^{m+2}\|^{1/(m+2)}\}$ using the matrix power norm estimation algorithm from [20], with no adverse effect on the accuracy of the matrix exponential computation algorithm proposed there. Note also that by using (7), (11) and (13) it follows that the first term of the Taylor remainder $R_m(2^{-s}A)$, the relative forward error $g_{m+1}(2^{-s}A)$ and the absolute backward error $h_{m+1}(2^{-s}A)$ is $(2^{-s}A)^{m+1}/(m+1)!$.

Motivated by these results, in the following subsection we analyze if setting the scaling parameter such that $(2^{-s}A)^{m+1}/(m+1)! \leq \max\{\epsilon, \|2^{-s}A\|\epsilon\}$ is practical for building a scaling and squaring algorithm for computing the matrix exponential with Taylor approximations and a prescribed tolerance ϵ higher than or equal to the unit roundoff error in the working machine precision.

Table 3: Maximal values of θ for different matrix exponential approximations with a similar computational cost.

Approximation	Source	Cost	θ
Paterson–Stockmeyer Taylor degree 12	[14, Tab. 1]	$5M$	$\theta_{12} = 3.28 \times 10^{-1}$
Padé approximant	[3, Tab. 3.1]	$4M + 1D$	$\theta_{14} = 9.50 \times 10^{-1}$
Taylor degree 18	[5, Tab. 2]	$5M$	$\theta_{18} = 1.1$
Taylor degree 20	[16], [17]	$5M$	$\theta_{20} = 1.44$
Approximation 21+	[4, Tab. 5]	$5M$	$\theta_{21+} = 1.68$
Optimal degree approximation of degree 32	[7, Tab. 5]	$5M$	$\theta_{degopt\,32} = 1.9$
Proposed 24+	-	5M	$\theta_{24+} = 2.47$

*M: Matrix multiplication cost, D: multiple right-hand side linear system solving cost.

4.1 Assessment of first-term-based scaling

If only the first term of the backward error for a Taylor approximation of degree $m = 24$ in (15) is used, with $\epsilon = u = 2^{-53}$ in IEEE double-precision arithmetic, one obtains

$$\frac{(2^{-s}A)^{m+1}}{(m+1)!} \leq \frac{\theta^{m+1}}{(m+1)!} \leq \max\{u, \theta u\} \rightarrow \theta_{24} = (25!u)^{1/24} \approx 2.43, \tag{16}$$

which is slightly lower than the value θ_{24+} given in Table 3 for the new 24+ approximation. Therefore, taking into account that for normal matrices it follows that $\|A^k\|_2 = \|A\|_2^k$, for $k = 2, 3, \ldots$, if we select the two-norm, using only the first term of the Taylor series remainder of the Taylor approximation of degree 24 for calculating the scaling parameter is a more conservative selection than needed, giving a higher or equal scaling parameter s.

In order to analyze the case of non-normal matrices and using the one-norm, note that from (7), the remainder of the Taylor approximation for the scaled matrix $B = 2^{-s}A$ can be written as

$$\|R_m(B)\| = \frac{\|B^{m+1}\|}{(m+1)!}\left(1 + \frac{\|B^{m+2}\|}{\|B^{m+1}\|(m+2)} + \frac{\|B^{m+3}\|}{\|B^{m+1}\|(m+2)(m+3)} + \cdots\right).$$

Hence, the truncation error is determined by the first neglected term multiplied by a correction factor that collects the contribution of higher-order terms.

To assess the practical relevance of scaling strategies based on the condition

$$\|(2^{-s}A)^{m+1}\|/(m+1)! \leq \max\{\epsilon, \|2^{-s}A\|\epsilon\}, \tag{17}$$

we examine the behavior of the correction factor

$$\mathcal{R}_{m,p}(B) = \frac{\displaystyle\sum_{j=2}^{p} \frac{\|B^{m+j}\|}{(m+j)!}}{\dfrac{\|B^{m+1}\|}{(m+1)!}}, \tag{18}$$

using representative 16×16 matrices test matrices, similarly to [3, Fig. 1.1]. In this case, the set of test matrices was augmented with all the matrices from the Matrix Computation Toolbox (MCT) [21] and the Eigtool MATLAB Package (EMP) [22]. The matrices where `expm` function from MATLAB gave `NaN` or `inf` values were excluded from all of them, using finally 59 matrices. Figures 1 and 2 report the correction factor $\mathcal{R}_{m,p}(B)$ with $B = 2^{-s}A$ for $p = 5$ and $p = 50$, respectively, using the scaling criterion from (17), i.e.

$$s = \max\left\{0, \min\left\{\left\lceil \frac{1}{m+1}\log_2\left(\frac{\|A^{m+1}\|}{(m+1)!\,\epsilon}\right)\right\rceil, \left\lceil \frac{1}{m}\log_2\left(\frac{\|A^{m+1}\|}{(m+1)!\,\|A\|\,\epsilon}\right)\right\rceil\right\}\right\}, \tag{19}$$

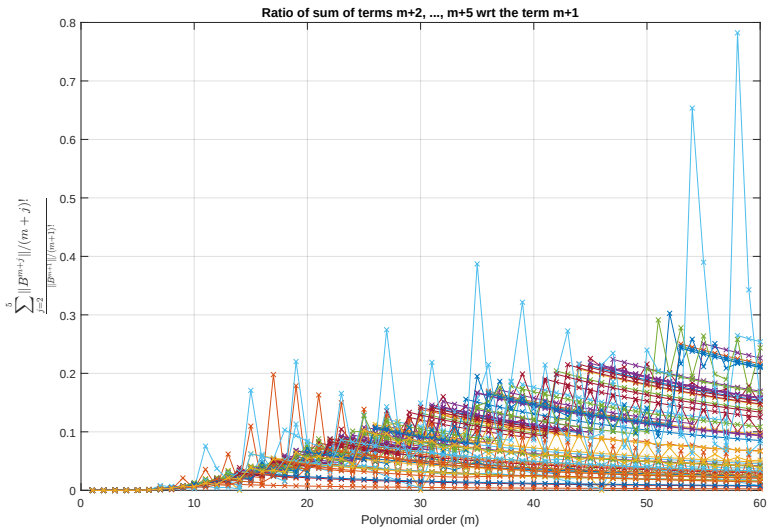


Figure 1: Ratio of the partial sum of the Taylor remainder $\sum_{j=2}^{5} \|B^{m+j}\|/(m+j)!$ to the first neglected term $\|B^{m+1}\|/(m+1)!$, for the scaled matrix $B = 2^{-s}A$. The scaling parameter s is selected such that $\|B\| \leq \max\{1, \|B\|\}u$, where u denotes the unit roundoff in IEEE double precision arithmetic.

where s depends on m.

The numerical results show that, for all matrices considered, the curves corresponding to $p = 5$ and $p = 50$ are indistinguishable. Moreover, except for two test matrices, the value of $\mathcal{R}_{m,p}(B)$ remains below 0.8. This indicates that, once the first neglected term $\|B^{m+1}\|/(m+1)!$ is sufficiently small, the contribution of higher-order terms to the Taylor remainder is negligible.

The two exceptions are the `godunov_demo` matrix from the EMP and the `chebspec` matrix from the MCT, both of which deserve special mention.For the `godunov_demo` matrix, the maximum value of the correction factor (18) for $p = 50$ over the range $m = 1, 2, \ldots, 60$ was 8.84×10^1. This yields an effective error tolerance $\epsilon \approx (1 + 8.84 \times 10^1)u$ for the truncation, which does not represent a significant increase in the overall error bound. However, while MATLAB R2023a function `expm` produced a finite result, its relative error was approximately 5.34×10^{35} when compared to a reference solution computed in high-precision arithmetic (using Taylor and Padé approximations with a mutual relative difference below the unit roundoff u). Although no `NaN` or `Inf` values are returned, the result is numerically unreliable, and this matrix likely lies outside the regime where either first-term-based or $\|A^k\|^{1/k}$-based error indicators remain meaningful.

For the `chebspec` matrix, the relative one-norm error of the MATLAB R2023a function `expm` with respect to the reference solution is 3.81×10^{-11}, indicating a reasonably accurate result. The maximum value of the correction ratio (18) for $p = 50$ over the range $m = 1, 2, \ldots, 60$ was 5.49, which is close to unity. This implies that the effective tolerance ϵ used in the truncation is approximately $6.49u$, i.e., of order $\mathcal{O}(u)$. Consequently, the scaling prescribed by (19) should lead to a small theoretical truncation error.

Using (19) with the bound $\|A^{m+1}\|_1 \leq \|A^2\|_1^{m/2}\|A\|_1$ (for $m = 24$) gives $s = 6$. Evaluating the matrix exponential via formula V5a (5) with this scaling, and with its coefficients rounded to IEEE double precision, results in a relative error of 5.21×10^{-12}, which is lower than that of `expm`. Furthermore, estimating $\|A^{m+1}\|_1$ using the matrix power norm estimation algorithm from [20] and (19) yields $s = 1$. The same scaling is obtained using the earlier algorithm from [23, Alg. 4.1]. Applying formula V5a (5) with the new scaling $s = 1$ produces an error norm of 2.62×10^{-11}, which

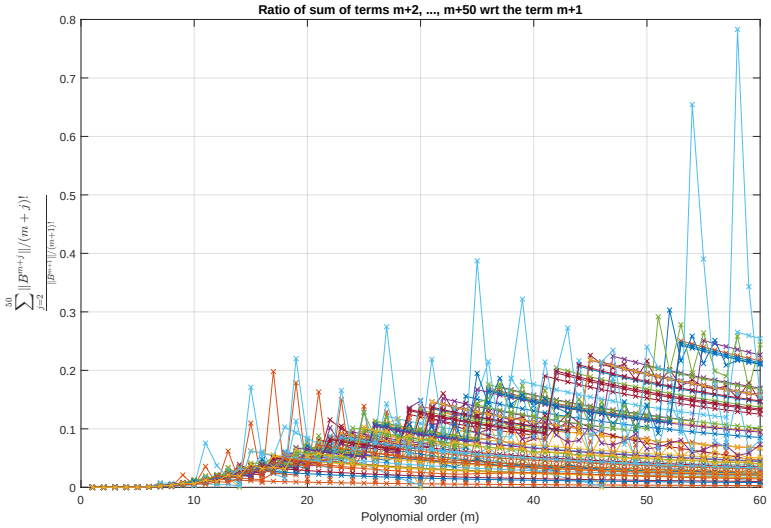


Figure 2: Ratio of the partial sum of the Taylor remainder $\sum_{j=2}^{50} \|B^{m+j}\|/(m+j)!$ to the first neglected term $\|B^{m+1}\|/(m+1)!$, for the scaled matrix $B = 2^{-s}A$. The scaling parameter s is selected such that $\|B\| \leq \max\{1, \|B\|\}u$, where u denotes the unit roundoff in IEEE double precision arithmetic.

also outperforms `expm`. Note that the cost of matrix power norm estimations is not negligible for small to medium-sized matrices, as reported in [4], [5]. Therefore, its use should be limited in such cases.

Remark 1 (Comparison with $\|A^k\|^{1/k}$-based scaling criteria) *Scaling criteria based on $\|A^k\|^{1/k}$ characterize the asymptotic growth of matrix powers but do not explicitly account for the factorial damping present in the Taylor remainder. The numerical experiments show that, for normal matrices and for non-normal matrices outside regimes of extreme non-normality, controlling the first neglected term $\|B^{m+1}\|/(m+1)!$ provides a sharp and practically effective estimate of the truncation error.*

The case of `chebspec` *illustrates that moderate non-normality may lead to correction ratios larger than one, even when the reference solution computed by* `expm` *is accurate. Nevertheless, the observed correction factors remain small enough to keep the effective tolerance at the level of $\mathcal{O}(u)$, and the first-term-based scaling still yields competitive or improved accuracy. Only in regimes of extreme non-normality, such as those exemplified by* `godunov_demo`*, does this approach break down, a situation in which even standard reference implementations fail to deliver reliable results.*

Finally, note that in [15, Sec. 2.1] a classical rounding error analysis was performed showing that, since the number of floating point operations increase with both the size matrix n and the polynomial approximation degree m, then the expected rounding error increases with both parameters n and m.

In the one-norm, the evaluation of $T_m(2^{-s}A)$ in floating-point arithmetic satisfies the worst-case forward error bound [15, eq. (5]

$$\frac{\|\mathrm{fl}(T_m(2^{-s}A)) - T_m(2^{-s}A)\|_1}{\|T_m(2^{-s}A)\|_1} \leq \phi(m,n)\, u,$$

where $\phi(m,n)$ depends on both the truncation degree m and the matrix dimension n. Such bounds are known to be pessimistic. In practice, roundoff errors range between u and mnu, and a commonly

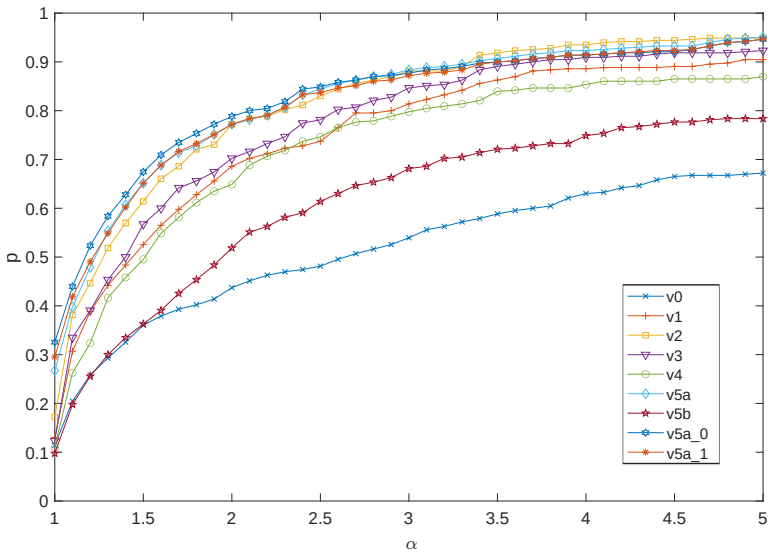


Figure 3: Performance profile comparing different evaluation orderings of the 24+ expression (2), including (4), (5), and the inverse ordering of (5). The matrix exponential is computed without norm estimations using 59 test matrices from the MCT and EMP collections. All variants use a forward error criterion based on u, except for V5a_0, which employs a backward error criterion $\|2^{-s}A\|u$, and V5a_1, which uses the more aggressive scaling criterion $\|2^{-s}A\|\sqrt{mn}\,u$.

used rule of thumb (see, e.g., [24, p. 52]) suggests that when a bound of the form $f(n)u$ is available, then the observed error is typically of order $\sqrt{f(n)}\,u$. This result is supported by assuming that the rounding errors are independent random variables and applying the central limit theorem. Applying this heuristic yields $\phi(m,n) \approx \sqrt{mn}$.

Although rounding errors do not necessarily behave as independent random variables, numerical experiments in [15, Sec. 3] showed that replacing the worst-case estimate by $\sqrt{mn}\,u$ lead to a reduction in computational cost, with no noticeable degradation in accuracy in the majority of cases.

This observation justifies the use of a more aggressive scaling strategy, in which the tolerance ϵ is increased to $\sqrt{mn}\,\epsilon$, leading to an accelerated variant of the algorithm. Figure 3 presents the performance profile [25] for the reordered 24+ evaluation formulas V0 (2), V2 (4), V5a (5), the inverse ordering of (5), i.e. V5b, and other variants, applied to the MCT and EMP test set. In this plot, the parameter $\alpha \in$ (sampled in steps of 0.1) represents the error ratio, while p denotes the probability that a specific method achieves a relative error within a factor α of the minimum error recorded across all tested methods. All variants compute the matrix exponential vith no norm estimation, using the 59 test matrices from the MCT and EMP collections.

All implementations employ a forward error criterion based on u, except for version V5a_0, which applies the expressions of V5a with a backward error criterion $\|2^{-s}A\|u$, and version V5a_1, which uses the more aggressive scaling criterion $\|2^{-s}A\|\sqrt{mn}\,u$. Among all tested variants, V5a_0 exhibits the best overall performance. Nevertheless, further numerical experiments are required to confirm the robustness of this observation.

5 Application to Musical Generative Flows

In [12], Taylor based approximations of the type of [8, Ex. 5.1] up to approximation order 15+ were used. This is a Taylor based polynomial approximation of degree 16, matching the corresponding

matrix exponential Taylor approximation up to degree 15. In this approximation the corresponding coefficient for degree 16, denoted by b_{16}, has a low absolute relative error with respect to the corresponding Taylor coefficient equal to $|b_{16} - 1/16!|16! = 0.454$. Following the approach for generative flows in [11], the first Taylor remainder term of the matrix exponential was adopted as an error measure to select the matrix scaling parameter. Since for the 15+ approximation this first term is $|b_{16} - 1/16!| = 0.454/16! < 1/16!$, the remainder second term $1/17!$ was also used to increase the scaling parameter if necessary. Then, the algorithm proposed there was applied to Generative flows for generation of images, accelerating the training and inference processes with respect to the original matrix exponential algorithm from [11].

In this work, we plan to apply a new matrix exponential computation algorithm based on the new 24+ approximation from Section 2, and using the new proposed scaling parameter selection from Section 4 for Generative flows in music generation. Generative flows are an approach to probabilistic modeling that allow for invertible transformation of a random variable. Flows are modeled using the change of variable formulation. A function is used to transform the variable and it is normalized by the absolute determinant of the transformed Jacobian matrix to preserve volume. Calculating the determinant of the Jacobian matrix can be computationally expensive if the Jacobian does not have specific properties. Because of this, two formulations of flows are generally used to ensure a traceable Jacobian matrix: autoregressive flows, which give the Jacobian matrix a triangular shape [26], or coupling layers, which divide the Jacobian into four regions [27]. Each of these formulations provide a traceable way to calculate invertible transforms. Recently, matrix exponentials have been added to flow models to increase the speed of training flows [11], [12]. Specifically, matrix exponentials were added to coupling layers to improve training stability.

In a live performance, a live musician improvises musical phrases and we use generative models to form phrase variations that can be played back in response in real time. Performed phrases are transcribed using a simple strategy of frequency estimation and quantization. To create variations of phrases, we encode phrases as a matrix containing 32 one-hot vectors, each representing a 16th note of the phrase. Using a flow model trained on a corpus of symbolic melodic phrases, we use the inverse of the model to find latent representations of the phrase. A latent representation can then be perturbed by adding noise. Running the model forward produces an output phrase that bears similarity to the original phrase, but, depending on the amount of noise added and which latent space(s) are perturbed, the output will be altered. The intent is to form musical variations, and the output is strongly biased by the distribution learned from the musical training corpus. The variations can consist of minor ornamentation when the latent space close to the output layer is perturbed, but variation can be more structural when earlier latent spaces are perturbed.

More details: We use the YIN [28] algorithm for fundamental frequency estimation. The estimates are quantized to an equal-temperament scale. A median filter is then used to reject outliers such as octave errors and to fill gaps that sometimes appear in YIN output. Pitches are quantized in time at regular intervals considered to be 16th notes, forming a phrase. In this encoding, we consider repeated pitches to be continuations of the previous 16th note, so notes can only begin after a rest or a pitch change.

To encode the phrases, we consider them to be sequences of classes, with a class for each pitch over the playable pitch range plus one class representing "rest" (silence, no pitch). Each class for a 16th note is encoded as a one-hot vector, so 32 one-hot vectors constitute a phrase.

To decode phrases, we consider repeated classes to be continuations, so a class that repeats in N consecutive time slots represents a pitch with a duration of N 16th notes.

In this work, we will apply the most accurate version of different reorderings from evaluation formulas (2) using the scaling parameter given by the new proposed error analysis to the approximation to the generation of musical variations. In this context, the output is strongly biased by the distribution learned from the musical training corpus. The use of the 24+ approximation

allows for a degree-32 Taylor-like behavior with the computational cost of only five multiplications, maintaining the structural fidelity of the learned musical features while optimizing inference time in large-scale models.

6 Conclusions

In this work, we have introduced a 24+ polynomial approximation for computing the matrix exponential. By reaching degree 32 while matching the Taylor expansion up to degree 24 using only five matrix multiplications, this scheme improves computational efficiency compared to traditional Padé approximants, previous Paterson-Stockmeyer implementations and recent state-of-the-art efficient polynomial approximations.

An error analysis based on the first neglected Taylor remainder term, $\|A^{m+1}\|/(m+1)!$, has been presented and evaluated. Our empirical assessment shows that, for both normal and moderately non-normal matrices, this first-term-based criterion provides a practical estimate of the theoretical truncation error. Furthermore, through stability analyzes over the MCT and EMP test matrix collections, we evaluated several operation reorderings of the polynomial to mitigate finite precision arithmetic errors. Preliminary numerical experiments indicate that the V5a_0 formulation—, employing a backward error criterion, yields the lowest error among the tested variants.

Finally, we plan to apply this scheme in the domain of generative AI, specifically within invertible flow models for real-time musical improvisation. The proposed 24+ approximation reduces the computational cost associated with the evaluation of matrix exponentials in coupling layers. By avoiding unnecessary overscaling while maintaining fidelity to the learned mathematical features, the method provides an efficient alternative for large-scale generative architectures.

Acknowledgments

This work has been supported by the Generalitat Valenciana Grant CIAICO/2023/275.

References

[1] C. B. Moler and C. F. Van Loan, "Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later," *SIAM Review*, vol. 45, no. 1, pp. 3–49, 2003.

[2] N. J. Higham, *Functions of Matrices: Theory and Computation*. Philadelphia, PA: SIAM, 2008.

[3] A. H. Al-Mohy and N. J. Higham, "A new scaling and squaring algorithm for the matrix exponential," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 970–989, 2009.

[4] J. Sastre, J. Ibáñez, and E. Defez, "Boosting the computation of the matrix exponential," *Applied Mathematics and Computation*, vol. 340, pp. 206–220, 2019.

[5] P. Bader, S. Blanes, and F. Casas, "Computing the matrix exponential with an optimized Taylor polynomial approximation," *Mathematics*, vol. 7, no. 12, p. 1174, 2019.

[6] P. Bader, S. Blanes, F. Casas, and M. Seydaoğlu, "An efficient algorithm to compute the exponential of skew-hermitian matrices for the time integration of the Schrödinger equation," *Mathematics and Computers in Simulation*, vol. 194, pp. 383–400, 2022.

[7] E. Jarlebring, M. Fasi, and E. Ringh, "Computational graphs for matrix functions," *ACM Transactions on Mathematical Software*, vol. 48, no. 4, pp. 1–35, 2023.

[8] J. Sastre, "Efficient evaluation of matrix polynomials," *Linear Algebra and its Applications*, vol. 539, pp. 229–250, 2018.

[9] J. Sastre and J. Ibáñez, "Evaluation of matrix polynomials beyond the Paterson–Stockmeyer method," *Mathematics*, vol. 9, no. 14, p. 1600, 2021.

[10] M. S. Paterson and L. J. Stockmeyer, "On the number of nonscalar multiplications necessary to evaluate polynomials," *SIAM Journal on Computing*, vol. 2, no. 1, pp. 60–66, 1973.

[11] C. Xiao and L. Liu, "Generative flows with matrix exponential," in *Proceedings of the 37th International Conference on Machine Learning*, ser. PMLR, vol. 119, 2020, pp. 10 452–10 461.

[12] J. Sastre, D. Faronbi, J. M. Alonso, P. Traver, J. Ibáñez, and N. Lloret, "Improving matrix exponential for generative ai flows: A taylor-based approach beyond paterson–stockmeyer," *arXiv preprint arXiv:2512.20777*, 2025. DOI: 10.48550/arXiv.2512.20777. arXiv: 2512.20777 [cs.LG].

[13] J. Sastre, "Advances on the evaluation of matrix polynomials beyond the paterson–stockmeyer method," in *Proceedings of the 25th Conference on Modelling for Engineering & Human Behaviour 2023*, J. Lucas, S. Izquierdo, J. Macián, and Y. Villacampa, Eds., Valencia, Spain: Instituto Universitario de Matemática Multidisciplinar, Jul. 2023, pp. 442–448.

[14] J. Sastre, J. Ibáñez, E. Defez, and P. Ruiz, "Accurate matrix exponential computation to solve coupled differential models in engineering," *Mathematical and Computer Modelling*, vol. 54, pp. 1835–1840, 2011.

[15] J. Sastre, J. Ibáñez, P. Ruiz, and E. Defez, "Accurate and efficient matrix exponential computation," *International Journal of Computer Mathematics*, vol. 91, no. 1, pp. 97–112, 2014.

[16] E. Jarlebring and G. Lorentzon, "The polynomial set associated with a fixed number of matrix-matrix multiplications," *SIAM Journal on Applied Algebra and Geometry*, 2026, Accepted.

[17] J. Sastre, J. Ibáñez, J. M. Alonso, and E. Defez, "Beyond Paterson–Stockmeyer: Advancing matrix polynomial computation," *WSEAS Transactions on Mathematics*, vol. 24, pp. 684–693, 2025, Presented at the 5th International Conference on Applied Mathematics, Computational Science and Systems Engineering (AMCSE), Paris, France, April 14-16, 2025. DOI: 10.37394/23206.2025.24.68. [Online]. Available: https://doi.org/10.37394/23206.2025.24.68.

[18] L. S. Blackford and J. Dongarra, "Installation guide for LAPACK," Department of Computer Science, University of Tennessee, LAPACK Working Note 41, 1999.

[19] P. Ruiz, J. Sastre, J. Ibáñez, and E. Defez, "High performance computing of the matrix exponential," *Journal of Computational and Applied Mathematics*, vol. 291, pp. 370–379, 2016.

[20] N. J. Higham and F. Tisseur, "A block algorithm for matrix 1-norm estimation, with an application to 1-norm pseudospectra," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1185–1201, 2000. DOI: 10.1137/S089547989834164X.

[21] N. J. Higham, *The matrix computation toolbox*, Available at https://github.com/higham/mctoolbox, 2025.

[22] T. G. Wright, *Eigtool, version 2.1*, Available at https://www.cs.ox.ac.uk/pseudospectra/eigtool/, 2009.

[23] N. J. Higham, "Fortran codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation," *ACM Transactions on Mathematical Software*, vol. 14, no. 4, pp. 381–396, 1988. DOI: 10.1145/50087.50088.

[24] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Second. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2002, ISBN: 0-89871-521-0. DOI: 10.1137/1.9780898718027.

[25] E. D. Dolan and J. J. Moré, "Benchmarking optimization software with performance profiles," *Mathematical Programming*, vol. 91, no. 2, pp. 201–213, 2002. DOI: 10.1007/s101070100263.

[26] G. Papamakarios, T. Pavlakou, and I. Murray, "Masked autoregressive flow for density estimation," in *Advances in Neural Information Processing Systems 30*, 2017.

[27] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in Neural Information Processing Systems 31*, 2018.

[28] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002. DOI: 10.1121/1.1458024. [Online]. Available: https://doi.org/10.1121/1.1458024.

A new Taylor method for the matrix sign function

Emilio Defez[♭],[1] José Miguel Alonso[♮], Javier Ibáñez[♭], Jorge Sastre[†]

(♭) Instituto de Matemática Multidisciplinar
(♮) Instituto de Instrumentación para Imagen Molecular
(†) Instituto de Telecomunicación y Aplicaciones Multimedia
Universitat Politècnica de València. Camino de Vera s/n, 46022, Valencia, Spain.

1 Introduction

The study of matrix functions constitutes a fundamental area within linear algebra and numerical analysis, with relevant applications in differential equations, control theory, and computational methods (see [1], [2]). Among the most important matrix functions are the exponential, the logarithm, and the sign function. The latter plays a particularly significant role from both theoretical and computational perspectives. In particular, it has notable applications in solving Riccati equations [3], computing invariant subspaces [4] and matrix decompositions [2]. Recently, the matrix sign function has found prominent applications in Artificial Intelligence and Machine Learning [5], [6], serving both as a target for deep learning approximations and as a core algorithmic component in advanced neural network optimizers [7], [8], [9].

The last year, in Reference [10], the following result (Theorem 2.8) has been established:

Theorem 1 *Let $A \in \mathbb{C}^{n\times n}$ have no purely imaginary eigenvalue and $k \in \mathbb{R}^+$. It holds for a large number k true that*

$$\operatorname{sign}(A) \approx \tanh(kA). \tag{1}$$

In this work, we first provide a new proof of this result, and subsequently use it to develop a new method for computing the matrix sign function. Finally, we present some examples illustrating the application of this method.

We now prove Theorem 1. Let $z = a + bi, a \neq 0$ be a complex number. Then

$$\tanh(z) = \frac{\operatorname{senh}(z)}{\cosh(z)} = \frac{e^z - e^{-z}}{e^z + e^{-z}}. \tag{2}$$

Using Euler's identity, it follows that

[1] edefez@imm.upv.es

$$\begin{aligned}\tanh(z) &= \frac{e^z - e^{-z}}{e^z + e^{-z}} \\ &= \frac{e^{a+bi} - e^{-a-bi}}{e^{a+bi} + e^{-a-bi}} \\ &= \frac{e^a\, e^{bi} - e^{-a}\, e^{-bi}}{e^a\, e^{bi} + e^{-a}\, e^{-bi}} \\ &= \frac{e^a(\cos(b) + i\,\mathrm{sen}(b)) - e^{-a}(\cos(b) - i\,\mathrm{sen}(b))}{e^a(\cos(b) + i\,\mathrm{sen}(b)) + e^{-a}(\cos(b) - i\,\mathrm{sen}(b))}. \end{aligned} \tag{3}$$

1) Suppose that $a > 0$ and let $a \to \infty$. Then $\lim_{a\to\infty} e^{-a} = 0$ while functions $\cos(b), \mathrm{sen}(b)$ remain bounded. Hence

$$\lim_{\Re(z)\to\infty} \tanh(z) = 1. \tag{4}$$

2) Suppose that $a < 0$ and let $a \to \infty$. Then $\lim_{a\to\infty} e^{a} = 0$ while functions $\cos(b), \mathrm{sen}(b)$ remain bounded. Hence

$$\lim_{\Re(z)\to\infty} \tanh(z) = -1. \tag{5}$$

Recall that the sign function $\mathrm{sign}(z)$ for complex numbers with $\Re(z) \neq 0$ is defined as

$$\mathrm{sign}(z) = \begin{cases} 1 & \text{if} \quad \Re(z) > 0 \\ -1 & \text{if} \quad \Re(z) < 0 \end{cases} \tag{6}$$

Therefore, from (4)-(5) we obtain

$$\lim_{\Re(z)\to\infty} \tanh(z) = \mathrm{sign}(z). \tag{7}$$

We need the following lemma:

Lema 1 *Let $y(z) = \tanh(z)$. Then the $n-$th derivative of $y(z)$ can be written as*

$$y^{(n)}(z) = Q_{n-1}(y(z))\left(1 - (y(z))^2\right)\ ,\ n \geq 1, \tag{8}$$

where $Q_n(x)$ is a polynomial of degree n with integer coefficients.

Proof: We proceed by induction. For $y(z) = \tanh(z)$ one gets

$$y'(z) = 1 - (\tanh(z))^2 = 1 - (y(z))^2$$

so the result holds for $Q_0(x) = 1$. Differentiating again

$$y''(z) = -2y(z)\left(1 - (y(z))^2\right),$$

so $Q_1(x) = -2x$ the result holds. Differentiating again and simplifying, we can write

$$y^{(3)}(z) = \left(6(y(z))^2 - 2\right)\left(1 - (y(z))^2\right),$$

thus, taking $Q_2(x) = 6x^2 - 2$ identity (8) holds, and $Q_2(x)$ has degree 2 with integers coefficients. Let us now suppose that the identity (8) holds for $n \geq 1$, ie:

$$y^{(n)}(z) = Q_{n-1}\left(y(z)\right)\left(1 - \left(y(z)\right)^2\right), \tag{9}$$

where $Q_{n-1}(x) = \sum_{k=0}^{n-1} \beta_k x^k, \beta_k \in \mathbb{Z}$. Let's see that the identity (8) is true for $n+1$. To do this, we calculate $y^{(n+1)}(z)$, which, applying the induction hypothesis:

$$\begin{aligned} y^{(n+1)}(z) &= \frac{d}{dz}\left[Q_{n-1}\left(y(z)\right)\left(1 - \left(y(z)\right)^2\right)\right] \\ &= \frac{d}{dz}\left[Q_{n-1}\left(y(z)\right)\right]\left(1 - \left(y(z)\right)^2\right) + \frac{d}{dz}\left[\left(1 - \left(y(z)\right)^2\right)\right] Q_{n-1}\left(y(z)\right). \end{aligned} \tag{10}$$

But $Q_{n-1}\left(y(z)\right) = \sum_{k=0}^{n-1} \beta_k \left(y(z)\right)^k$, by the chain rule

$$\begin{aligned} \frac{d}{dz}\left[Q_{n-1}\left(y(z)\right)\right] &= \left(\sum_{k=1}^{n-1} k\beta_k \left(y(z)\right)^{k-1}\right) y'(z) \\ &= \left(\sum_{k=1}^{n-1} k\beta_k \left(y(z)\right)^{k-1}\right)\left(1 - \left(y(z)\right)^2\right). \end{aligned} \tag{11}$$

Therefore, substituting (11) into (10) we have

$$\begin{aligned} y^{(n+1)}(z) &= \frac{d}{dz}\left[Q_{n-1}\left(y(z)\right)\right]\left(1 - \left(y(z)\right)^2\right) + \frac{d}{dz}\left[\left(1 - \left(y(z)\right)^2\right)\right] Q_{n-1}\left(y(z)\right) \\ &= \left(\sum_{k=1}^{n-1} k\beta_k \left(y(z)\right)^{k-1}\right)\left(1 - \left(y(z)\right)^2\right)^2 + Q_{n-1}\left(y(z)\right)\left(-2y(z)y'(z)\right) \\ &= \left(\sum_{k=1}^{n-1} k\beta_k \left(y(z)\right)^{k-1}\right)\left(1 - \left(y(z)\right)^2\right)^2 - 2Q_{n-1}\left(y(z)\right) y(z)\left(1 - \left(y(z)\right)^2\right) \\ &= \underbrace{\left[\left(\sum_{k=1}^{n-1} k\beta_k \left(y(z)\right)^{k-1}\right)\left(1 - \left(y(z)\right)^2\right) - 2Q_{n-1}\left(y(z)\right) y(z)\right]}_{=Q_n(y(z))}\left(1 - \left(y(z)\right)^2\right) \\ &= Q_n\left(y(z)\right)\left(1 - \left(y(z)\right)^2\right), \end{aligned} \tag{12}$$

where

$$Q_n(x) = \left(\sum_{k=1}^{n-1} k\beta_k x^{k-1}\right)\left(1 - x^2\right) - 2xQ_{n-1}(x)$$

is a polynomial of degree n with integer coefficients. Thus, the proposed formula is true. □

Remark 1 *This Lemma 1 provides a simple way to compute higher-order derivatives of* $\tanh(z)$. *The polynomials* $Q_n(x)$ *are defined recursively by*

$$Q_0(x)=1\ ,\ Q_n(x)=-2xQ_{n-1}(x)+(1-x^2)Q'_{n-1}(x)=\frac{d}{dx}\left((1-x^2)Q_{n-1}(x)\right)\ ,\ n\geq 1. \quad (13)$$

Once calculated $Q_{n-1}(x)$*, the variable* x *is changed to* $Q_{n-1}\left(\tanh\left(z\right)\right)$ *and it follows that the* $n-$*th derivative of* $\tanh\left(z\right)$ *can be written as*

$$\frac{d^n}{dz^n}\tanh\left(z\right)=Q_{n-1}\left(\tanh\left(z\right)\right)\left(1-\left(\tanh\left(z\right)\right)^2\right)\ ,\ n\geq 1. \quad (14)$$

For example, we have

$$\begin{aligned}Q_{14}(x) &= 1307674368000x^{14}-5666588928000x^{12}+9909265766400x^{10}-8888034355200x^8\\ &+ 4275610859520x^6-1039052851200x^4+103028914176x^2-1903757312,\end{aligned}$$

then

$$\begin{aligned}Q_{14}\left(\tanh\left(z\right)\right) &= 1307674368000\left(\tanh\left(z\right)\right)^{14}-5666588928000\left(\tanh\left(z\right)\right)^{12}\\ &+ 9909265766400\left(\tanh\left(z\right)\right)^{10}-8888034355200\left(\tanh\left(z\right)\right)^{8}+4275610859520\left(\tanh\left(z\right)\right)^{6}\\ &- 1039052851200\left(\tanh\left(z\right)\right)^{4}+103028914176\left(\tanh\left(z\right)\right)^{2}-1903757312,\end{aligned}$$

and finally one gets

$$\begin{aligned}\tanh^{(15)}\left(z\right) &= Q_{14}\left(\tanh\left(z\right)\right)\left(1-\left(\tanh\left(z\right)\right)^2\right)\\ &= -1307674368000\tanh^{16}(z)+6974263296000\tanh^{14}(z)-15575854694400\tanh^{12}(z)\\ &+ 18797300121600\tanh^{10}(z)-13163645214720\tanh^{8}(z)+5314663710720\tanh^{6}(z)\\ &- 1142081765376\tanh^{4}(z)+104932671488\tanh^{2}(z)-1903757312.\square\end{aligned}$$

Now suppose that $A\in\mathbb{C}^{r\times r}$ has no purely imaginary eigenvalues and admits the Jordan decomposition $A=TJ_AT^{-1}$, where

$$J_A=diag\left\{J_{\lambda_1},J_{\lambda_2},\ldots,J_{\lambda_s}\right\}$$

where J_{λ_i} are the different Jordan blocks associated with the matrix A. Let us consider an arbitrary block associated with the eigenvalue $\lambda\in\sigma(A)$ given by the block $n\times n$:

$$J_\lambda=\begin{pmatrix}\lambda & 1 & 0 & \cdots & 0\\ 0 & \lambda & 1 & \vdots & \vdots\\ 0 & 0 & \ddots & \ddots & \vdots\\ \vdots & \cdots & \ddots & \lambda & 1\\ 0 & \cdots & \cdots & 0 & \lambda\end{pmatrix}$$

We then have, see for example [11], if $\Re(\lambda)>0$ then

$$\operatorname{sign}\left(J_\lambda\right)=\begin{pmatrix}1 & 0 & 0 & \cdots & 0\\ 0 & 1 & 0 & \vdots & \vdots\\ 0 & 0 & \ddots & \ddots & \vdots\\ \vdots & \cdots & \ddots & 1 & 0\\ 0 & \cdots & \cdots & 0 & 1\end{pmatrix},$$

and if $\Re(\lambda) < 0$ then

$$\operatorname{sign}(J_\lambda) = \begin{pmatrix} -1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \vdots & \vdots \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \cdots & \ddots & -1 & 0 \\ 0 & \cdots & \cdots & 0 & -1 \end{pmatrix}.$$

We have that $\operatorname{sign}(A) = Tdiag\left\{\operatorname{sign}(J_{\lambda_1}), \operatorname{sign}(J_{\lambda_2}), \ldots, \operatorname{sign}(J_{\lambda_s})\right\} T^{-1}$.

Since $\tanh(z)$ is an analytic function, it follows that by the definition of a matrix function from Jordan blocks, we have that

$$\tanh(J_\lambda) = \begin{pmatrix} \tanh(\lambda) & \frac{\tanh'(\lambda)}{1!} & \frac{\tanh''(\lambda)}{2!} & \cdots & \frac{\tanh^{(n-1)}(\lambda)}{(n-1)!} \\ 0 & \tanh(\lambda) & \frac{\tanh'(\lambda)}{1!} & \vdots & \frac{\tanh^{(n-2)}(\lambda)}{(n-2)!} \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \cdots & \vdots & \tanh(\lambda) & \frac{\tanh'(\lambda)}{1!} \\ 0 & \cdots & \cdots & 0 & \tanh(\lambda) \end{pmatrix}$$

Applying lemma 1 we can write

$$\tanh(J_\lambda)$$

$$= \begin{pmatrix} \tanh(\lambda) & \frac{Q_0(\tanh(\lambda))\left(1-(\tanh(\lambda))^2\right)}{1!} & \frac{Q_1(\tanh(\lambda))\left(1-(\tanh(\lambda))^2\right)}{2!} & \cdots & \frac{Q_{n-2}(\tanh(\lambda))\left(1-(\tanh(\lambda))^2\right)}{(n-1)!} \\ 0 & \tanh(\lambda) & \frac{Q_0(\tanh(\lambda))\left(1-(\tanh(\lambda))^2\right)}{1!} & \vdots & \frac{Q_{n-3}(\tanh(\lambda))\left(1-(\tanh(\lambda))^2\right)}{(n-2)!} \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \cdots & \vdots & \tanh(\lambda) & \frac{Q_0(\tanh(\lambda))\left(1-(\tanh(\lambda))^2\right)}{1!} \\ 0 & \cdots & \cdots & 0 & \tanh(\lambda) \end{pmatrix} \tag{15}$$

If we assume that $\lambda = a + ib$ with $a > 0$, taking limits in (15) when $\Re(\lambda) \to \infty$, we have that $\tanh(\lambda) \to 1$, and since the Q_i are (continuous) polynomials, we have that $Q_i(\tanh(\lambda))\left(1 - (\tanh(\lambda))^2\right) \to 0$, so

$$\lim_{\Re(\lambda)\to\infty} \tanh(J_\lambda)$$

$$= \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \vdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \cdots & \vdots & 1 & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix} = \operatorname{sign}(J_\lambda) \tag{16}$$

If we assume that $\lambda = a + ib$ with $a < 0$, taking limits in (15) when $\Re(\lambda) \to \infty$, we have that $\tanh(\lambda) \to -1$, and since the Q_i are (continuous) polynomials, we have that $Q_i(\tanh(\lambda))\left(1 - (\tanh(\lambda))^2\right) \to 0$, so

$$\lim_{\Re(\lambda)\to\infty} \tanh(J_\lambda)$$
$$= \begin{pmatrix} -1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \vdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \cdots & \vdots & -1 & 0 \\ 0 & \cdots & \cdots & 0 & -1 \end{pmatrix} = \operatorname{sign}(J_\lambda). \tag{17}$$

Therefore, we finally conclude that $\operatorname{sign}(A) = \lim_{k\to\infty} \tanh(kA)$, and the theorem 1 has been proven.

2 Proposed Method

Let $A \in \mathbb{C}^{n\times n}$ have no purely imaginary eigenvalues. Choose $s > 0$ such that $\|2^{-s}A\| < \frac{\pi}{2} \approx 1.5708$. Next, we will use the reference [12] to calculate $\tanh(2^{-s}A)$ using Taylor expansion

$$\tanh(A) = \sum_{n\geq 1} \frac{2^{2n}\left(2^{2n}-1\right)\mathfrak{B}_{2n}}{(2n)!} A^{2n-1}\ ,\ \|A\| < \frac{\pi}{2} \tag{18}$$

where $\mathfrak{B}_{2n}$ are Bernoulli numbers. Once calculated $\tanh(2^{-s}A)$ we apply the recurrence (5) from [12]:

$$\left.\begin{array}{rcl} T_0 & = & \tanh(2^{-s}A) \\ T_i & = & 2\left(I + T_{i-1}^2\right)^{-1} T_{i-1} \end{array}\right\} \tag{19}$$

For $i = 1, \ldots, s$, the recurrence (19) constructs $\tanh(A)$, and for $i > s$, recurrence (19) computes $\tanh(2^{i-s}A)$. As $i \to \infty$, this acts as the $k \to \infty$ multiplier, thus T_i converge to the matrix sign function $\operatorname{sign}(A)$.

The idea is to take a tolerance $\varepsilon > 0$ and continue the iteration (19) until

$$\|T_i - T_{i-1}\| < \varepsilon,$$

to finally take $\operatorname{sign}(A) = T_i$. Let's look at an example.

Ejemplo 1 *Given the matrix* $A = \begin{pmatrix} 1 & 9 & 9 & 0 & 2 & 8 \\ 4 & 3 & 7 & 7 & 0 & 2 \\ 2 & 9 & 8 & 6 & 7 & 9 \\ 5 & 6 & 8 & 3 & 2 & 5 \\ 3 & 1 & 5 & 3 & 1 & 1 \\ 4 & 0 & 3 & 8 & 3 & 10 \end{pmatrix}$ *with* $\|A\| = 29.8732$*, we take* $s = 5$*. We take the tolerance* $\varepsilon = 10^{-16}$*. It is easy to calculate that*

$$\operatorname{sign}(A) \approx \begin{pmatrix} -0.748891 & 0.503848 & 0.374005 & 0.453926 & -0.166385 & 0.486995 \\ 0.105645 & 0.314174 & 0.57375 & 0.475182 & -0.850379 & -0.346366 \\ 0.20234 & 0.219606 & -0.0594646 & 0.212872 & 1.12669 & 0.229067 \\ 0.183091 & 0.803413 & 0.573947 & -0.576762 & -0.212876 & 0.0876392 \\ -0.00452786 & -0.184043 & 0.654739 & -0.160791 & 0.284293 & -0.18435 \\ 0.46488 & -0.530631 & -0.284078 & 0.525589 & 0.0542177 & 0.78665 \end{pmatrix}$$

Following the proposed algorithm, we have the following table shows the different results:

i	$\|T_i - T_{i-1}\|$
1	0.327626
2	0.569644
3	0.932561
4	0.893653
5	0.339008
6	0.0733985
7	0.00278619
8	3.53463×10^{-6}
9	5.66125×10^{-12}
10	2.67693×10^{-16}
11	2.97293×10^{-16}
12	7.63655×10^{-17}

we will therefore take

$$\operatorname{sign}(A) \approx T_{12} = \begin{pmatrix} -0.748891 & 0.503848 & 0.374005 & 0.453926 & -0.166385 & 0.486995 \\ 0.105645 & 0.314174 & 0.57375 & 0.475182 & -0.850379 & -0.346366 \\ 0.20234 & 0.219606 & -0.0594646 & 0.212872 & 1.12669 & 0.229067 \\ 0.183091 & 0.803413 & 0.573947 & -0.576762 & -0.212876 & 0.0876392 \\ -0.00452786 & -0.184043 & 0.654739 & -0.160791 & 0.284293 & -0.18435 \\ 0.46488 & -0.530631 & -0.284078 & 0.525589 & 0.0542177 & 0.78665 \end{pmatrix}$$

with absolute error $\|\operatorname{sign}(A) - T_{12}\| = 4.37256346637199 \times 10^{-15}$.

Remark 2 *Although it is not the objective of this work, using Lemma 1, we can obtain a similar result for the n-th derivative of function* $\tan(z)$. *We define polynomials* $S_n(x)$ *by*

$$S_0(x) = 1 \ , \ S_n(x) = 2xS_{n-1}(x) + (1+x^2)S'_{n-1}(x) = \frac{d}{dx}\left((1+x^2)S_{n-1}(x)\right) \ , \ n \geq 1. \tag{20}$$

Once calculated $S_{n-1}(x)$, *the variable* x *is changed to* $S_{n-1}(\tan(z))$ *and it follows that the n-th derivative of* $\tan(z)$ *can be written as*

$$\frac{d^n}{dz^n}\tan(z) = S_{n-1}(\tan(z))\left(1 + (\tan(z))^2\right) \ , \ n \geq 1. \tag{21}$$

For example,

$$\begin{aligned} S_{14}(x) &= 1307674368000x^{14} + 5666588928000x^{12} + 9909265766400x^{10} + 8888034355200x^8 \\ &+ 4275610859520x^6 + 1039052851200x^4 + 103028914176x^2 + 1903757312 \end{aligned}$$

then

$$\begin{aligned} S_{14}(\tan(z)) &= 1903757312 + 103028914176\tan(z)^2 + 1039052851200\tan(z)^4 \\ &+ 4275610859520\tan(z)^6 + 8888034355200\tan(z)^8 + 9909265766400\tan(z)^{10} \\ &+ 5666588928000\tan(z)^{12} + 1307674368000\tan(z)^{14} \end{aligned}$$

and one gets that

$$\begin{aligned}\tan^{(15)}(z) &= S_{14}(\tan(z))\left(1+(\tan(z))^2\right)\\ &= 1307674368000\tan^{16}(z)+6974263296000\tan^{14}(z)+15575854694400\tan^{12}(z)\\ &+ 18797300121600\tan^{10}(z)+13163645214720\tan^{8}(z)+5314663710720\tan^{6}(z)\\ &+ 1142081765376\tan^{4}(z)+104932671488\tan^{2}(z)+1903757312.\square\end{aligned}$$

Acknowledgments

This work has been supported by the Generalitat Valenciana Grant CIAICO/2023/275.

References

[1] E. D. Denman and A. N. Beavers Jr, "The matrix sign function and computations in systems," *Applied mathematics and Computation*, vol. 2, no. 1, pp. 63–94, 1976.

[2] K.-L. Du, M. Swamy, Z.-Q. Wang, and W. H. Mow, "Matrix factorization techniques in machine learning, signal processing, and statistics," *Mathematics*, vol. 11, no. 12, p. 2674, 2023.

[3] R. Byers, "Solving the algebraic Riccati equation with the matrix sign function," *Linear Algebra and its Applications*, vol. 85, pp. 267–279, 1987.

[4] Z. Bai and J. Demmel, "Using the matrix sign function to compute invariant subspaces," *SIAM Journal on Matrix Analysis and Applications*, vol. 19, no. 1, pp. 205–225, 1998.

[5] R. Padmanabhan and S. Brugiapaglia, "Approximating matrix functions with deep neural networks and transformers," *arXiv preprint arXiv:2602.07800*, 2026.

[6] M. Kansal, V. Sharma, P. Sharma, and L. Jäntschi, "A globally convergent iterative method for matrix sign function and its application for determining the eigenvalues of a matrix pencil," *Symmetry*, vol. 16, no. 4, p. 481, 2024.

[7] N. Amsel, D. Persson, C. Musco, and R. M. Gower, "The polar express: Optimal matrix sign methods and their application to the muon algorithm," in *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

[8] Anonymous, "MUON optimizes under spectral norm constraints," in *Forty-second International Conference on Machine Learning (ICML)*, 2025.

[9] N. Amsel, T. Chen, A. Greenbaum, C. Musco, and C. Musco, "Nearly optimal approximation of matrix functions by the lanczos method," in *Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[10] H. Fatoorehchi, H. G. Bohr, and P. L. Fosbøl, "Accelerated BIBO stability criterion for dynamical systems based on matrix hyperbolic tangent function: H. fatoorehchi et al.," *International Journal of Dynamics and Control*, vol. 13, no. 6, p. 239, 2025.

[11] C. S. Kenney and A. J. Laub, "The matrix sign function," *IEEE transactions on automatic control*, vol. 40, no. 8, pp. 1330–1348, 2002.

[12] J. Ibáñez, J. M. Alonso, J. Sastre, E. Defez, and P. Alonso-Jordá, "Advances in the approximation of the matrix hyperbolic tangent," *Mathematics*, vol. 9, no. 11, p. 1219, 2021.

The PageRank Vector with Memory: From Discrete Iterations to Continuous Dynamical Systems (Part I)

J. Flores[♭1], E. Primo[♭ ♮], D. Rodríguez [♭ ♮], and M. Romance[♭ ♮]

(♭) Department of Applied Mathematics, Science and Engineering of Materials, and Electronic Engineering, ESCET, Universidad Rey Juan Carlos, Móstoles (Madrid, Spain).

(♮) Laboratory of Computational Mathematics on Complex Networks and their Applications, Universidad Rey Juan Carlos, Móstoles (Madrid, Spain).

Abstract

PageRank is one of the most influential centrality measures in network science, originally introduced for ranking web pages and later generalized to a wide range of applications in social, biological, economic, and technological networks [1]. While classical PageRank theory focuses on the existence and uniqueness of a stationary distribution for a fixed personalization vector, the dynamics of PageRank under time-varying personalization vectors remain less explored [2]. In this work, we present an extension of this model by incorporating a"memory function" into the personalization vector, exploring the transition from discrete-time iterations to continuous-time dynamical systems.

In the first part, we analyze a discrete iterative scheme where the personalization vector at each step is updated as a weighted average of its preceding states. The mathematical foundations and convergence analysis of this discrete model—specifically the proof that it converges to the left-hand Perron vector of the row-stochastic adjacency matrix—have been recently established in [3]. We discuss how this mechanism acts as an "influence amplifier" for the underlying graph structure.

The second part extends these results to a continuous-time framework. We model the evolution of the centrality vector via a system of Volterra integro-differential equations. In this continuous setting, we establish rigorous results concerning the existence and uniqueness of solutions and prove that the system preserves the required stochastic properties (positivity and unit sum). Finally, we investigate the asymptotic behavior under various memory functions, showing that while for some traditional functions the solution converges (asymptotically) to the Perron vector of the row-stochastic adjacency matrix, certain oscillatory memories can give rise to asymptotically periodic solutions, revealing complex dynamics in memory-dependent networks.

References

[1] S. Brin and L. Page, Computer networks and ISDN systems 30 (1998), 107–117.

[2] P. Boldi, M. Santini and S. Vigna, ACM Transactions on Information Systems 27(4) (2009), 1–23.

[3] D. Aleja, J. Flores, E. Primo, D. Rodriguez, and M. Romance, Lin. Alg. Appl. 733 (2026), 233–272.

[1] julio.flores@urjc.es

The PageRank Vector with Memory: From Discrete Iterations to Continuous Dynamical Systems (Part II)

J. Flores[♭],E. Primo[♭ ♮ 1], D. Rodríguez [♭ ♮], and M. Romance[♭ ♮]

(♭) Department of Applied Mathematics, Science and Engineering of Materials, and Electronic Engineering, ESCET, Universidad Rey Juan Carlos, Móstoles (Madrid, Spain).

(♮) Laboratory of Computational Mathematics on Complex Networks and their Applications, Universidad Rey Juan Carlos, Móstoles (Madrid, Spain).

Abstract

PageRank is one of the most influential centrality measures in network science, originally introduced for ranking web pages and later generalized to a wide range of applications in social, biological, economic, and technological networks [1]. While classical PageRank theory focuses on the existence and uniqueness of a stationary distribution for a fixed personalization vector, the dynamics of PageRank under time-varying personalization vectors remain less explored [2]. In this work, we present an extension of this model by incorporating a"memory function" into the personalization vector, exploring the transition from discrete-time iterations to continuous-time dynamical systems.

In the first part, we analyze a discrete iterative scheme where the personalization vector at each step is updated as a weighted average of its preceding states. The mathematical foundations and convergence analysis of this discrete model—specifically the proof that it converges to the left-hand Perron vector of the row-stochastic adjacency matrix—have been recently established in [3]. We discuss how this mechanism acts as an "influence amplifier" for the underlying graph structure.

The second part extends these results to a continuous-time framework. We model the evolution of the centrality vector via a system of Volterra integro-differential equations. In this continuous setting, we establish rigorous results concerning the existence and uniqueness of solutions and prove that the system preserves the required stochastic properties (positivity and unit sum). Finally, we investigate the asymptotic behavior under various memory functions, showing that while for some traditional functions the solution converges (asymptotically) to the Perron vector of the row-stochastic adjacency matrix, certain oscillatory memories can give rise to asymptotically periodic solutions, revealing complex dynamics in memory-dependent networks.

References

[1] S. Brin and L. Page, Computer networks and ISDN systems 30 (1998), 107–117.

[2] P. Boldi, M. Santini and S. Vigna, ACM Transactions on Information Systems 27(4) (2009), 1–23.

[3] D. Aleja, J. Flores, E. Primo, D. Rodriguez, and M. Romance, Lin. Alg. Appl. 733 (2026), 233–272.

[1] eva.primo@urjc.es

Row completion of polynomial matrices with partial prescription of the singular structure

Agurtzane Amparan[♭] Itziar Baragaña[♯] Silvia Marcaida[♭] and Alicia Roca[♯],[1]

(♭) Departamento de Matemáticas, Universidad del País Vasco UPV/EHU,
Barrio Sarriena s/n, Leioa, Spain.

(♮) Departamento de Ciencia de la Computacioón e I.A., Universidad del País Vasco UPV/EHU,
Paseo M. de Lardizabal 1, Donostia-San Sebastián, Spain.

(♯) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

A frequently studied problem in Linear Algebra consists in characterizing the existence of a matrix having certain properties. Within this framework, an important problem is the *row (column) completion of matrices*. It consists in characterizing the existence of a matrix with some of its rows prescribed together with other structural properties. We are interested in the row (column) completion problem of polynomial matrices with some rows, the degree, and the *eigenstructure* (or part of it) prescribed. The eigenstructure of a polynomial matrix is formed by the invariant factors, the partial multiplicities at infinity, and the minimal indices of the left and right null spaces.

Polynomial matrices often arise in many different scientific areas such as engineering, physics, economics, and biology, in the study of the dynamical behavior of systems of differential or difference equations [1], [2], [3], [4]. On the other hand, matrix pencil completion problems play an important role in the study of a variety of problems in control theory. A matrix pencil is a polynomial matrix of degree one. An enormous progress has been made in this area in the last decades ([5], [6], [7]).

The row (column) completion problem of polynomial matrices with prescribed eigenstructure and given degree has been studied in [8]. As the eigenstructure is formed by four different types of elements, the study of the partial prescription of the eigenstructure leads to 15 different problems. In [8], it was also solved the partial prescription problem in some of those cases. More precisely, it was solved the row completion problem when the whole eigenstructure but the row (column) minimal indices are prescribed, and when the finite and/or the partial multiplicities at infinity are prescribed.

Here we give necessary and sufficient conditions for the existence of a polynomial matrix of given degree with prescribed singular structure and some of its rows (the results were obtained in [9]). Precisely written the problem solved is the following,

Problem 1.1 *Let $P(s) \in \mathbb{F}[s]^{m\times n}$ be a polynomial matrix of* $\deg(P(s)) = d$. *Find necessary and sufficient conditions for the existence of a polynomial matrix $W(s) \in \mathbb{F}[s]^{z\times n}$ of* $\deg(W(s)) \leq d$ *such that* $\begin{bmatrix} P(s) \\ W(s) \end{bmatrix}$ *has the singular structure prescribed.*

[1] correspondingauthor@mail.com

2 Methods

2.1 Theoretical results

To achieve our results we turn a row completion problem of a polynomial matrix into a matrix pencil completion problem, hence being able to take advantage of the results in ([6], [7]). In turn, to do it, we need to choose an appropriate linearization of the polynomial matrix. The linearization of choice is the *First Frobenius companion form.*

Key results for solving Problem 1.1 are the *Index sum theorem for matrix polynomials* ([10, Theorem 6.5]) and the result of the prescription of the eigenstructure to a polynomial matrix [8, Theorem 4.2]

Lemma 2.1 ([10, Theorem 6.5]. Index Sum Theorem for Matrix Polynomials) *Let $P(s) \in \mathbb{F}[s]^{m\times n}$ be a polynomial matrix,* $\deg(P(s)) = d$, $\operatorname{rank}(P(s)) = r$. *Let $\phi_1(s,t) \mid \cdots \mid \phi_r(s,t)$, $c_1 \geq \cdots \geq c_{n-r}$ and $u_1 \geq \cdots \geq u_{m-r}$ be the homogeneous invariant factors, column minimal indices and row minimal indices of $P(s)$, respectively. Then,*

$$\sum_{i=1}^{r} \deg(\phi_i) + \sum_{i=1}^{n-r} c_i + \sum_{i=1}^{m-r} u_i = rd. \tag{1}$$

Theorem 2.2 ([8, Theorem 4.2]. Prescription of the eigenstructure) *Let $P(s) \in \mathbb{F}[s]^{m\times n}$,* $\deg(P(s)) = d$, $\operatorname{rank}(P(s)) = r$. *Let $\phi_1(s,t) \mid \cdots \mid \phi_r(s,t)$ be the homogeneous invariant factors, $\mathbf{c} = (c_1, \ldots, c_{n-r})$ the column minimal indices, and $\mathbf{u} = (u_1, \ldots, u_{m-r})$ the row minimal indices of $P(s)$, where $u_1 \geq \cdots \geq u_\eta > u_{\eta+1} = \cdots = u_{m-r} = 0$.*

Let z and x be integers such that $0 \leq x \leq \min\{z, n-r\}$. Let $\gamma_1(s,t) \mid \cdots \mid \gamma_{r+x}(s,t)$ be monic homogeneous polynomials, and $\mathbf{d} = (d_1, \ldots, d_{n-r-x})$ and $\mathbf{v} = (v_1, \ldots, v_{m+z-r-x})$ two partitions, where $v_1 \geq \cdots \geq v_{\bar{\eta}} > v_{\bar{\eta}+1} = \cdots = v_{m+z-r-x} = 0$. There exists a polynomial matrix $W(s) \in \mathbb{F}[s]^{z\times n}$ such that $\deg(W(s)) \leq d$, $\operatorname{rank}\left(\begin{bmatrix} P(s) \\ W(s) \end{bmatrix}\right) = r+x$, *and $\begin{bmatrix} P(s) \\ W(s) \end{bmatrix}$ has $\gamma_1(s,t) \mid \cdots \mid \gamma_{r+x}(s,t)$ as homogeneous invariant factors, $d_1, \ldots, d_{n-r-x}$ as column minimal indices and $v_1, \ldots, v_{m+z-r-x}$ as row minimal indices if and only if*

$$\gamma_i(s,t) \mid \phi_i(s,t) \mid \gamma_{i+z}(s,t), \quad 1 \leq i \leq r, \tag{2}$$

$$\bar{\eta} \geq \eta, \tag{3}$$

$$\mathbf{c} \prec' (\mathbf{d}, \mathbf{a}), \tag{4}$$

$$\mathbf{v} \prec' (\mathbf{u}, \mathbf{b}), \tag{5}$$

$$\sum_{i=1}^{r+x} \deg(\operatorname{lcm}(\phi_{i-x}, \gamma_i)) \leq \sum_{i=1}^{m+z-r-x} v_i - \sum_{i=1}^{m-r} u_i + \sum_{i=1}^{r+x} \deg(\gamma_i), \quad \text{with equality when } x = 0, \tag{6}$$

where $\mathbf{a} = (a_1, \ldots, a_x)$ and $\mathbf{b} = (b_1, \ldots, b_{z-x})$ are

$$\begin{aligned} a_1 = {} & \textstyle\sum_{i=1}^{m+z-r-x} v_i - \sum_{i=1}^{m-r} u_i \\ & + \textstyle\sum_{i=1}^{r+x} \deg(\gamma_i) - \sum_{i=1}^{r+x-1} \deg(\operatorname{lcm}(\phi_{i-x+1}, \gamma_i)) - d, \\ a_j = {} & \textstyle\sum_{i=1}^{r+x-j+1} \deg(\operatorname{lcm}(\phi_{i-x+j-1}, \gamma_i)) - \sum_{i=1}^{r+x-j} \deg(\operatorname{lcm}(\phi_{i-x+j}, \gamma_i)) - d, \\ & \qquad 2 \leq j \leq x, \end{aligned} \tag{7}$$

$$\begin{aligned} b_1 = {} & \textstyle\sum_{i=1}^{m+z-r-x} v_i - \sum_{i=1}^{m-r} u_i \\ & + \textstyle\sum_{i=1}^{r+x} \deg(\gamma_i) - \sum_{i=1}^{r+x} \deg(\operatorname{lcm}(\phi_{i-x-1}, \gamma_i)), \\ b_j = {} & \textstyle\sum_{i=1}^{r+x} \deg(\operatorname{lcm}(\phi_{i-x-j+1}, \gamma_i) - \sum_{i=1}^{r+x} \deg(\operatorname{lcm}(\phi_{i-x-j}, \gamma_i)), \\ & \qquad 2 \leq j \leq z-x. \end{aligned} \tag{8}$$

3 Results

We solve Problem 1.1, i.e., we find necessary and sufficient conditions for the existence of a polynomial matrix when some rows, the degree, and the singular structure are prescribed. The result is the next one.

Theorem 3.1 ([8, Theorem 3.12]. Prescription of singular structure) *Let $\mathbb{F}$ be an algebraically closed field. Let $P(s) \in \mathbb{F}[s]^{m\times n}$, $\deg(P(s)) = d$, $\operatorname{rank}(P(s)) = r$. Let $\phi_1(s,t) \mid \cdots \mid \phi_r(s,t)$ be the homogeneous invariant factors, $\mathbf{c} = (c_1, \ldots, c_{n-r})$ the column minimal indices, and $\mathbf{u} = (u_1, \ldots, u_{m-r})$ the row minimal indices of $P(s)$, where $u_1 \geq \cdots \geq u_\eta > u_{\eta+1} = \cdots = u_{m-r} = 0$.*

Let z and x be integers such that $0 \leq x \leq \min\{z, n-r\}$ and let $d_1 \geq \cdots \geq d_{n-r-x} \geq 0$ and $v_1 \geq \cdots \geq v_{\bar{\eta}} > v_{\bar{\eta}+1} = \cdots = v_{m+z-r-x} = 0$ be non negative integers. Let

$$E = \sum_{i=1}^{n-r-x} d_i - \sum_{i=1}^{n-r} c_i + \sum_{i=1}^{m+z-r-x} v_i - \sum_{i=1}^{m-r} u_i - xd. \tag{9}$$

There exists $W(s) \in \mathbb{F}[s]^{z\times n}$ such that $\deg(W(s)) \leq d$, $\operatorname{rank}\left(\begin{bmatrix} P(s) \\ W(s) \end{bmatrix}\right) = r + x$, $\begin{bmatrix} P(s) \\ W(s) \end{bmatrix}$ has $\mathbf{v} = (v_1, \ldots, v_{m+z-r-x})$ as row minimal indices and $\mathbf{d} = (d_1, \ldots, d_{n-r-x})$ as column minimal indices if and only if (3),

$$E \leq \sum_{i=r+x-z+1}^{r} \deg(\phi_i), \tag{10}$$

$$\sum_{i=1}^{m+z-r-x} v_i - \sum_{i=1}^{m-r} u_i \geq \max\{0, E\}, \tag{11}$$

$$\mathbf{c} \prec' (\mathbf{d}, \tilde{\mathbf{a}}), \tag{12}$$

$$\mathbf{v} \prec' (\mathbf{u}, \tilde{\mathbf{b}}), \tag{13}$$

where $\tilde{\mathbf{a}} = (\tilde{a}_1, \ldots, \tilde{a}_x)$ and $\tilde{\mathbf{b}} = (\tilde{b}_1, \ldots, \tilde{b}_{z-x})$ are

$$\begin{array}{rl} \tilde{a}_1 = & \sum_{i=1}^{m+z-r-x} v_i - \sum_{i=1}^{m-r} u_i - E - d, \\ \tilde{a}_j = & -d, \quad 2 \leq j \leq x, \end{array} \tag{14}$$

$$\begin{array}{rl} \tilde{b}_1 = & \sum_{i=1}^{m+z-r-x} v_i - \sum_{i=1}^{m-r} u_i + \min\{0, \deg(\phi_r) - E\}, \\ \tilde{b}_j = & \min\{0, \sum_{i=r-j+1}^{r} \deg(\phi_i) - E\} - \min\{0, \sum_{i=r-j+2}^{r} \deg(\phi_i) - E\}, \\ & \qquad\qquad 2 \leq j \leq z - x. \end{array} \tag{15}$$

The necessity of the results obtained holds for arbitrary fields, whilst to prove the sufficiency we need that the underlying field is algebraically closed.

As corollaries of this result, we solve the same problem when only the row (column) minimal indices are prescribed.

We will emphasize some properties of the results obtained presenting some examples.

4 Conclusions

The row (column) completion problem of polynomial matrices is studied when the degree and the singular structure of the completed matrix is prescribed. The results are part of the 15 problems project of characterizing the existence of a polynomial matrix when some of its rows (columns), the degree, and part of its eigenstructure are prescribed. The whole project has been solved in [8] and [9]. The necessity of the characterization analyzed in this work holds for arbitrary fields, whilst the proof of the sufficiency requires that the underlying field is algebraically closed.

Acknowledgments

This work was supported by grants PID2021-124827NB-I00 and RED2022-134176-T funded by MCIN/AEI/ 10.13039/501100011033 and by "ERDF A way of making Europe" by the "European Union". The first and third authors were also supported by grant GIU21/020 funded by UPV/EHU..

References

[1] T. Betcke, N. J. Higham, V. Mehrmann, C. Schröder, and F. Tisseur, "NLEVP: A collection of nonlinear eigenvalue problems," *ACM Transactions on Mathematical Software*, vol. 39, no. 2, 1â28, 2013.

[2] G. D. Forney, "Minimal bases of rational vector spaces with applications to multivariable linear systems," *SIAM Journal Control*, vol. 13, no. 3, 143â520, 1975.

[3] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*. SIAM, 1985.

[4] T. Kailath, *Linear Systems*. Prentice Hall, 1980.

[5] J. Loiseau, S. Mondié, I. Zaballa, and P. Zagalak, "Assigning the Kronecker invariants of a matrix pencil by row or column completion," *Linear Algebra Appl.*, vol. 278, 327â336, 1998.

[6] M. Dodig, "Completion up to a matrix pencil with column minimal indices as the only nontrivial kronecker invariants," *Linear Algebra Appl.*, vol. 438, 3155â3173, 2013.

[7] M. Dodig and M. Stošić, "The general matrix completion problem: A minimal case," *SIAM J. Matrix Anal. Appl.*, vol. 40, no. 1, 347â369, 2019.

[8] A. Amparan, I. Baragaña, S. Marcaida, and A. Roca, "Row or column completion of polynomial matrices of given degree," *SIAM J. Matrix Anal. Appl.*, vol. 45, no. 1, 478â503, 2024.

[9] A. Amparan, I. Baragaña, S. Marcaida, and A. Roca, "Row or column completion of polynomial matrices of given degree ii," *SIAM J. Matrix Anal. Appl.*, vol. 708, 252â279, 2025.

[10] F. De Terán, F. M. Dopico, and D. S. Mackey, "Spectral equivalence of matrix polynomials and the index sum theorem," *Linear Algebra Appl.*, vol. 459, 264â333, 2014.

High relative accurate algorithms for Riordan arrays associated with Schröder matrices

Jorge Ballarín[♭], Jorge Delgado[♮1] and Juan Manuel Peña[♭]

(♭) Edificio de Matemáticas, Universidad de Zaragoza
Calle Pedro Cerbuna 12, 50009 - Zaragoza, Spain.

(♮) Edificio Torres Quevedo, Universidad de Zaragoza
Calle María de Luna 3, 50018 - Zaragoza, Spain.

1 Introduction

Riordan matrices arise in combinatorics within the framework of Riordan arrays, which play a central role in studying recurrence relations, lattice path problems, combinatorial identities, and related areas (see [1]). The large and little Schröder numbers arise in the enumeration of specific classes of lattice paths, giving rise to the corresponding Riordan arrays and associated triangular matrices. The inverses of the large and little Schröder matrices also admit a natural combinatorial interpretation. In this talk, we will show the total positivity of the large and little Schröder matrices, and we will demonstrate that a wide range of linear algebra computations can be carried out to high relative accuracy (HRA) with them and their inverses.

2 Methods

2.1 Bidiagonal decomposition

We will present the bidiagonal decomposition of a family of recursive matrices. This bidiagonal decomposition will help us to show the total positivity of large and little Schröder matrices. In addition, this bidiagonal decomposition will also play a key role to obtain HRA algorithms for linear algebra problems with large and little Schröder matrices.

3 Results

In this talk we will present a bidiagonal factorization of large and little Schröder matrices, $\mathcal{P}_n$ and $\mathcal{Q}_n$, as well as of their inverses. From these two bidiagonal decompositions we derive the total positivity of both $\mathcal{P}_n$ and $\mathcal{Q}_n$. In addition, we will show that both bidiagonal decompositions can be computed with HRA. These HRA bidiagonal decompositions, together with the algorithms devised by Koev in [2], [3], will allow us to solve certain linear algebra problems for $\mathcal{P}_n$ and $\mathcal{Q}_n$ also to high relative accuracy. The theoretical results are illustrated numerically. For more details on the results see [4].

[1] jorgedel@unizar.es

Acknowledgments

This research was partially supported through the Spanish research grants PID2022-138569NB-I00 and RED2022-134176-T (MCIU/AEI), and by Gobierno de Aragón (E41_23R).

References

[1] R. Sprugnoli, "Riordan arrays and combinatorial sums," *Discrete Mathematics*, vol. 132, pp. 267–290, 1994.

[2] P. Koev, "Accurate eigenvalues and svds of totally nonnegative matrices," *SIAM J. Matrix Anal. Appl.*, vol. 27, pp. 1–23, 2005.

[3] P. Koev, "Accurate computations with totally nonnegative matrices," *SIAM J. Matrix Anal. Appl.*, vol. 29, pp. 731–751, 2007.

[4] J. M. P. Jorge Ballarín Jorge Delgado, "Accurate computations with riordan arrays associated with schröder matrices," *Calcolo*, 2026, Accepted for publication.

Preserving observability and structural decodability with output feedback transformations on ISO representations of convolutional codes

Noemí DeCastro-García[♭♮], Lucía Mallo-Fernández[♭♮1], and Miguel V. Carriegos[♭♮]

(♭) Departamento de Matemáticas, Universidad de León,
Campus de Vegazana s/n, 24007, León, Spain.

(♮) Research Institute of Applied Sciences in Cybersecurity, Universidad de León,
Campus de Vegazana s/n, 24007, León, Spain.

Abstract

In this work, we study the effect of output feedback transformations on Input/State/Output (ISO) representations of convolutional codes [1].

Convolutional codes are error-correcting codes that can be formally described as submodules of $\mathbb{F}[z]^n$ of rank k, where n denotes the number of output symbols, k the number of input symbols, and δ the complexity of the code. A convolutional code $\mathcal{C}$ can be represented by an encoder $G(z)$, which maps an input sequence $u(z)$ to an output sequence $v(z)$ via $v(z) = G(z)u(z)$.

An alternative description is provided by linear dynamical systems through ISO realizations. An ISO representation of $\mathcal{C}$ is a tuple $\Sigma(\mathcal{C}) = (A, B, C, D) \in \mathbb{F}^{\delta\times\delta} \times \mathbb{F}^{\delta\times k} \times \mathbb{F}^{(n-k)\times\delta} \times \mathbb{F}^{(n-k)\times k}$ such that

$$\mathcal{C} = \left\{ v(z) = \begin{pmatrix} y(z) \\ u(z) \end{pmatrix} \in \mathbb{F}[z]^n : \exists\, x(z) \in \mathbb{F}[z]^\delta \text{ such that } \begin{cases} x_{t+1} = Ax_t + Bu_t, \\ y_t = Cx_t + Du_t, \\ v_t = \begin{pmatrix} y_t \\ u_t \end{pmatrix}, \\ x_0 = 0. \end{cases} \right\}.$$

Here, $x_t \in \mathbb{F}^\delta$ is the state vector, $u_t \in \mathbb{F}^k$ the input (information) vector, and $y_t \in \mathbb{F}^{n-k}$ the output vector, while v_t is the transmitted codeword.

Output feedback is a classical tool in control theory, but its role in coding theory requires further investigation. Let Σ be a linear dynamical system over $\mathbb{F}$ given by

$$\Sigma : \begin{cases} x_{t+1} = Ax_t + Bu_t, \\ y_t = Cx_t + Du_t. \end{cases}$$

Given a feedback matrix $K \in \mathbb{F}^{k\times(n-k)}$, we consider the static output feedback law

$$u_t = Ky_t + v_t,$$

[1] lmalf@unileon.es

where $v_t \in \mathbb{F}^k$ is a new external input. Substituting $y_t = Cx_t + Du_t$ yields

$$(I_k - KD)u_t = KCx_t + v_t.$$

We say that the system is *well-posed* if $I_k - KD \in GL_k(\mathbb{F})$. In that case, setting

$$\Theta := (I_k - KD)^{-1},$$

the transformed system induced by K, denoted by Σ_K, is

$$\Sigma_K := \left(A_K := A + B\Theta KC,\, B_K := B\Theta,\, C_K := (I_{n-k} - DK)^{-1}C,\, D_K := (I_{n-k} - DK)^{-1}D\right). \tag{1}$$

This transformation provides a constructive mechanism for generating new convolutional codes [2]. While the equivalence of convolutional codes and their state-space realizations has been widely studied, our approach focuses on using output feedback transformations to obtain non-equivalent codes with desirable structural properties [1]. For instance, although minimality is preserved under well-posed feedback, other properties, such as invertibility of the state matrix, may be lost [3], [4].

We derive explicit algebraic conditions on the feedback matrix K that guarantee the preservation of key structural properties of the ISO representation, enabling the construction of observable and well-decodable (GDP) convolutional codes. We also provide constructive criteria for selecting such feedback matrices and establish quantitative lower bounds on their density within the parameter space.

Our results show that admissible feedback matrices form a Zariski-open and dense subset of the parameter space. In particular, structural decodability is generically preserved when the field size is sufficiently large. Overall, this yields a deterministic and quantitative framework for generating new observable convolutional codes via output feedback while maintaining desirable system-theoretic and decoding properties, typically producing non-equivalent encoders.

References

[1] N. DeCastro-García, L. Mallo-Fernández, and M. V. Carriegos, "Output feedback transformations for preserving observability and structural decodability in convolutional codes," *accepted at International Workshop on the Arithmetic of Finite Fields (WAIFI 2026)*, 2026.

[2] N. DeCastro-García, M. V. Carriegos, and A. L. M. Castañeda, "Construction of observable and MDP convolutional codes with good decodable properties by I/S/O representations," *International Journal of Applied Mathematics and Computer Science*, vol. 36(1), pp. 141–154, 2026.

[3] T. Kailath, *Linear Systems*. Prentice-Hall, Englewood Cliffs, NJ, 1980.

[4] H. Gluesing-Luerssen and G. Schneider, "State space realizations and monomial equivalence for convolutional codes," *Linear Algebra and Its Applications*, vol. 425(2-3), pp. 518–533, 2007.

Part VII

Maths & Physics

Entangled photons detection

Guillem Gómez i Blanch[♭],[1] Màrius Josep Fullana i Alfonso[♭] and Roberto Dale Valdivia[♮]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

(♮) Centro de Investigación Operativa, Universidad Miguel Hernández
Edificio Torretamarit - Avda. de la Universidad s/n 03202 Elche, Spain.

1 Introduction

The subject of this work is the detection of entangled photons in a beam of photons of given direction and frequency. The interest of this topic lies both in the investigation of photon beams in the laboratory and in that of incident radiation from the outside, specifically in relation to the background radiation (CBR). In our approach we treat the incident photon beam as a convex function of entangled and unentangled photon pairs, investigating their proportion by measuring the correlation of results at various angles between the detector polarimeters.

The EPR paradox [4] is at the center of our approach. The EPR refers to a system of two particles in a singlet state of interaction, which move away in opposite directions along the same direction. Bell [1] studied the correlation between the spins of both particles as a function of the angle between the detectors. Our interest is different from Bell's, in comparing two different views of entanglement, that of quantum and local realism; it is limited only to the distinction between independent and entangled particle pairs. We extend this approach in this work to the case in which the particles arrive in parallel directions and the same sense to the detectors, after having experienced reflections that do not alter their quantum state, as occurs with photons coming from the atmosphere.

In the EPR paradox, Bohm and Ahanorov [3] replaced the electrons of the initial hypothesis by photons, changing the role played by the spin of the electrons by the polarization of the photons. This equivalence and the fact that non-relativistic quantum mechanics could be applied to particles that travel at the speed of light, as is the case with photons, deserve a first consideration in the present work, showing that wave functions can be assigned to photons, where the linear momentum vector and polarization intervene, so that non-relativistic quantum mechanics can be applied to them.

In our approach we present a theoretical-experimental scheme on which to establish the detection of entangled photons, based on a beam of incident photons subjected to an initial preparation consisting of fixing their direction by collimation and energy by spectral analysis, so that photons of determined momentum vector and therefore energy act perpendicularly on a pair of polarimeters, one of which at least is capable of rotating in the plane perpendicular to the incidence of the radiation.

Since photons are characterized not only by the linear momentum vector and the corresponding energy but also by the polarization, in this initial preparation the orientation of their polarization

[1] guigobla@upv.es

is not defined and the incident photons, both those entangled and those that are not, reach the detectors with an orientation completely independent of the orientation that the experimenter has imposed on his polarimeters. The photons, impinging on the polarimeters necessarily make a transition to the polarization imposed by each of the instruments, giving rise to the specific observable magnitude which consists of a value associated with the verticality and another with the horizontality of the polarity.

Therefore, we will consider the incident photon flux on the polarimeters as a combination of entangled and unentangled pairs in the form of a convex function, in proportion characterized by a fraction of entangled in value f and of unentangled in value $1 - f$.

We will consider as entangled photons exclusively those in the singlet state with orthogonal polarizations, assimilating all other possible entanglement as non-entangled. The theoretical treatment of the fraction of entangled photons shows a certain angular dependence with respect to the rotation of the polarimeter. The same treatment on the non-entangled fraction of arbitrary proportionality shows a similar but not equal angular dependence that allow us to establish the calculation of the entangled fraction, using the correlation of the experimental results.

Below we show an abbreviated description of the proposed approach.

2 Methods

2.1 Quantum characterization of the photon and of the photon systems

Maxwell's equations are a solid starting point in a fully relativistic representation of electromagnetic radiation and, therefore, of its corpuscular version, the photons. Starting from this, in effect, we arrive at a wave equation that represents the photon and that can be used semi-classically in non-relativistic quantum mechanics. The series expansion of the electromagnetic vector potential $\mathbf{A}$ as a function of the wave quadrivector and a canonical change of coordinates leads to a representation of the field by elementary harmonic oscillators. For each value of the momentum vector $\mathbf{k}$ we have the position vector $\mathbf{Q}_{\mathbf{k}n}$ and the linear momentum vector, $\mathbf{P}_{\mathbf{k}n}$ where n=1,2 represents the two values that each vector takes in the plane perpendicular to $\mathbf{k}$ and which represents the polarization direction. The Hamilton function then takes the form:

$$H = \frac{1}{2} \sum_{\mathbf{k}n} \left(\mathbf{P}_{\mathbf{k}n}^2 + \omega^2 \mathbf{Q}_{\mathbf{k}n}^2 \right) \tag{1}$$

From here we proceed to the canonical quantification, replacing the variables by operators in the previous expression. Following Berestetskii [2] , we use the linear combination operator $\omega \mathbf{Q}_{\mathbf{k}n} \pm \mathbf{P}_{\mathbf{k}n}$ to obtain wave functions that can make some non-relativistic quantum approaches valid, which, in the representation of moments has the form:

$$\mathbf{A}_{\mathbf{k}\alpha}(\mathbf{k}'\alpha') = \sqrt{2\pi} \frac{\mathbf{u}^{(\alpha)}}{\sqrt{\omega}} \, \delta_{\mathbf{k}'\mathbf{k}} \, \delta_{\alpha'\alpha} \tag{2}$$

with $\mathbf{k}$ linear momentum, ω energy and $\mathbf{u}^{\alpha}$ unit polarization vector. The wave function and the vector observe the transverse condition $\mathbf{k} \cdot A(\mathbf{k} = 0)$. The total angular momentum of the photon j can take only integer values. Of course, in addition to the total angular momentum of the photon we must consider the parity, corresponding to the transformation of the wave function into an inversion of the coordinate system.

Photons can form binary states, i.e. entangled pairs, as Landau [6], showed. The total angular momentum of the entangled pair of photons takes only integer values and is determined by the behavior towards parity; for the singlet state, with zero total momentum. Formally, the symmetric and antisymmetric parts of the system are separated.

Bohm and Aharonov [3] proposed the use of entangled photons, with orthogonal polarizations. Among the states that systems of two photon pairs can adopt, we are interested in the singlet state with the parities of the two perpendicular photons.

Therefore, our present approach starts from the premises that it makes physical sense to assign to each individual photon a wave function characterized by a polarization vector and a momentum vector - and, therefore, an energy and a spatial orientation - determined, in representation of moments and that this characterization extends to photons with the same energy, polarization and propagation vector representing their quantum state, in the usual identification of individual particles that have the same characteristics with the same quantum state.

2.2 Framework of the proposed experiment

The starting point of the proposed experiment is a photon beam where it is presumed that there are both entangled and non-entangled photon pairs. In this situation, a preliminary preparation consisting of a collimation is made in order to filter only photons with momentum of a defined direction and sense and a frequency filtering, which will condition that the photons that pass the preparation all have the same directional vector $\mathbf{u}$ and the same frequency ω therefore the same momentum vector $\mathbf{k}$ as symbolized in the figure 1 and subsequent collimation to conveniently direct the photons towards the polarimeters.

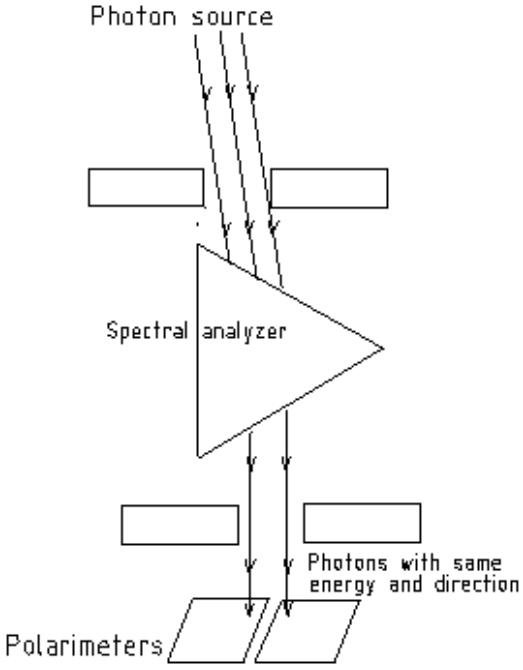


Figure 1: *Collimation and photon energy selection.*

Thus, entangled or non-entangled photon pairs are defined with respect to vector momentum, energy and direction; but not with respect to polarization; we will limit ourselves to those that have polarization perpendicular to each other. These non-separable entangled pairs will therefore be considered as a limiting case of mixed systems, although they turn out to be pure states, as can be demonstrated. Likewise, non-entangled photon pairs will be considered as mixed and separable systems, since it is always possible to express them as a product of two independent functional factors.

To experimentally investigate a beam of photons with the same energy ω and propagation vector $\mathbf{k}$ we have a pair of polarimeters placed perpendicular to the beam. We have foreseen a second collimation before the polarimeters. Each polarimeter is able to detect whether the polarization of the incident photon is horizontal or vertical, constituting an observable with dichotomous eigenvalues: " vertical" or " horizontal". The consideration of both results for a pair of photons that impinge simultaneously -it is, within a very short delay of time- allow us to qualify as possibly correlated or not correlated at all. Both polarimeters are initially aligned and at least one of them can rotate perpendicular to the incidence of the photons.

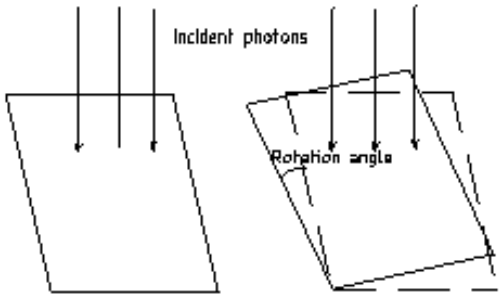


Figure 2: *One of the two polarimeters can rotate respect to the other.*

Starting from the data provided by the polarimeters on pairs of photons, each incident on a different polarimeter, we establish the correlation between these values. The behavior of entangled photons will be seen to be distinguishable from non-entangled ones if the detection plane is rotated, which makes it possible to incorporate correlation conditions dependent on the rotation angle and define a fraction of entangled photons.

The photons are made to impinge orthogonally on the polarimeters. Each polarimeter will be characterized by two vectors: one, horizontal and the other vertical. To each polarimeter we correspond a Hilbert space, $\mathcal{H}il_1$ with vectors $|H_1\rangle$ and $|V_1\rangle$ and $\mathcal{H}il_2$ with vectors $|H_2\rangle$ and $|V_2\rangle$. These two spaces will have the geometric vectors that characterize them, h_1, v_1 and h_2, v_2 parallel and in correspondence with the aforementioned kets. Together, both Hilbert spaces will configure a tensor product space of both, with vectors $|H_1\rangle|H_2\rangle$, $|H_1\rangle|V_2\rangle$, $|V_1\rangle|H_2\rangle$ and $|V_1\rangle|V_2\rangle$.

As mentioned, at least one of the two polarimeters will be able to rotate an angle α in a plane normal to the direction of incidence of the photon flux $\vec{\mathbf{n}}$ configuring a Hilbert space $\mathcal{H}il'_1$ with vectors $|H'_1\rangle$ and $|V'_1\rangle$, which, with the other polarimeter, will configure a product space of vectors $|H'_1\rangle|H_2\rangle$, $|H'_1\rangle|V_2\rangle$, $|V'_1\rangle|H_2\rangle$ and $|V'_1\rangle|V_2\rangle$

To establish an unified reference system, it will be necessary to project the vectors from $\mathcal{H}il'_1$ to $\mathcal{H}il_1$ by means of the corresponding plane rotation of angle α.

2.3 General approach to the treatment of photon fluxes as mixed systems

Mixture systems are described by the functions of the pure reference states multiplied by the corresponding probability. To characterize them by the density operator, which when adopting a reference system becomes a density matrix, and responds to the expression:

$$\rho = \sum_i |\Psi_i\rangle p_i \langle\Psi_i| \tag{3}$$

with Ψ_i the reference Hilbert vector and p_i its probability of presentation, therefore, it is necessary to know first the probability of each basis state. In our case, it is a question of calculating the

probability of presentation of each pair of basis vectors as a function of the angle between the polarimeters and forming the correlation function.

In the incident photon beam we have two systems, corresponding to entangled systems Ψ_E and non-entangled systems, Ψ_N; therefore, it is necessary to know first the probability of each basis state. As we have said, the reference vectors are two orthogonal in each polarimeter, which give a tensor product of four vectors.

To calculate the relative probability for a two-photon entangled system we use an approach similar to Kaiser's [5]. For example we evaluate the amplitude corresponding to the situation $|H_1\rangle|V_2\rangle$:

$$\mathcal{A}_{\mathcal{E}}(H_1, V_2) = \{\langle H_1|\langle V_2|\}|\Psi_E\rangle \tag{4}$$

which is actually identifiable with a pure system, and from here we can calculate the corresponding probabilities as the square of the amplitude. The product of the probability by the value assigned to each experimental situation allows us to calculate the correlation and thus be able to compare the correlation between pairs of correlated and uncorrelated photons.

For non-entangled systems the calculation becomes a little longer, being convenient to resort to the density operator:

$$P_N = \sum_{k,n} Tr(\rho_k |u_n\rangle\langle u_n|) \tag{5}$$

,with $|u_n\rangle$ the corresponding vector of the tensor product of spaces, for example $|H_1\rangle|H_2\rangle$.

To be able to distinguish between the two situations it is necessary to apply the solution used by Bell in the analysis of the EPR paradox: move at least one of the two polarimeters. Then the elementary vectors must be replaced by:

$$\langle H_1^{\alpha}| = cos\,\alpha\,\langle H_1| + sin\,\alpha\,\langle V_1| \tag{6}$$

$$\langle V_1^{\alpha}| = -sin\,\alpha\,\langle H_1| + cos\,\alpha\,\langle V_1| \tag{7}$$

With this substitution we obtain a difference between the correlations of entangled and non-entangled systems. If the fraction of entangled is f the fraction of non-entangled (including any fraction of entangled of some other type than the one considered) will be $1 - f$:

$$\langle \Psi_T|\Psi_T\rangle = f \cdot \langle \Psi_E|\Psi_E\rangle + (1 - f)\, P_N \tag{8}$$

Then, by giving different values for the angle α we form equations with the experimental correlation found and thus we can calculate the various unknowns and particularly the fractions of photon pairs correlated by the indicated form.

3 Results

We establish a convex model for the photon beams consisting of entangled and non-entangled pairs after a preparation in which their energy and propagation vector are fixed. We consider the Hilbert space tensor product of the two corresponding to the detection devices and investigate the probability of polarization detections as a function of the angle between the polarimeters. From these data, we elaborate the correlation of the data and study the fraction of entangled pairs.

4 Conclusions

According to our investigation, it seems possible to distinguish and determine the entanglement fraction f in a photon's beam. Therefore, it is could open a possibility of experimental investigation.

Acknowledgments

This work is partially supported by PAID-11-25 of the Universitat Politècnica de València.

References

[1] Bell, J. On the Einstein-Podolsky-Rosen paradox. *Physics 1*, 3 (1964), 195–200.

[2] Berestetskii, V., Lifschitz, E., and Pitaevskii, L. *Teoria Cuantica Relativista*. Editorial Reverte S.A., Barcelona, 1975.

[3] Bohm, D., and Ahanorov, Y. Discussion of experimental proof for the paradox of einstein, rosen and podolsky. *Physical Review 108*, 4 (1957), 1070–1076.

[4] Einstein, A., Podolski, B., and Rosen, N. Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, 47 (1935), 777–780.

[5] Kaiser, D. Bell's inequality and quantum entanglement. *Lecture Notes for 8.225/STS.02 Physics in the 20th Century* (2020).

[6] Landau, L. On the angular momentum of a system of two photons. *Dokl.Akad.Nauk SSSR.*, 60 (1948), 471–473.

More about harmonic oscillators on the lattice

Emilio Defez Candel,[1] Michael M. Tung and Màrius Josep Fullana i Alfonso

I.U. de Matemàtica Multidisciplinària, Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain.

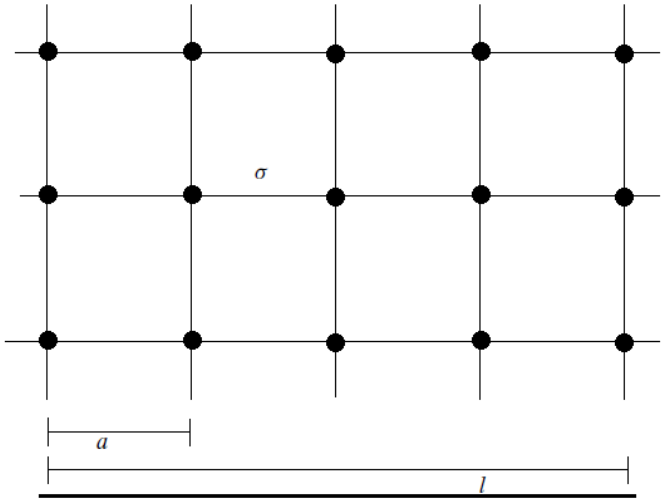
1 Introduction

This work extends the simulation of a two-dimensional ideal lattice having simple tetragonal geometry to a more general case (see [1]). Now, we consider an alternance of masses. The masses alternate between m and $2m$ both in the X axis and in the Y axis. This is a previous step to consider even more complicated structures. The harmonic character of the oscillators give rise to a system of second-order linear differential equations, which can be recast into matrix form. The explicit solutions which govern the dynamics of this system can be expressed in terms of matrix trigonometric functions. For the derivation we employ the Lagrangian formalism to determine the correct solutions, which extremize the underlying action of the system. In this paper, only the equations of movement are presented. We are working on the analytical and numerical solutions and they will be presented further on.

2 Previous calculations

In order to clarify the new computations explained in the introduction, we briefly present here the previous ones (see [1]). Let be a square network of length l, composed of n^2 equidistant nodes ij, where $i, j \in \{1, ..., n\}, n \in \mathbb{N}$. Its spacing is, then, $a = l/(n+1)$. Each element ij has components x_{ij} and y_{ij}, respectively, which will be time-dependent $t \geq 0$.

[1] edefez@imm.upv.es



If one of the positional indices is either 0 or $n+1$, the element will fall on the wall of the network and will be neglected. We note that the geometry of the network imposes the symmetry condition that all physical formulas are invariant under the simultaneous changes $x \Leftrightarrow y$, $i \Leftrightarrow j$ and the transposition (exchange of the position of the two indices). We will assume that the different elements are two-dimensional harmonic oscillators with the same stiffness constant σ and we will neglect friction. We choose the equilibrium length as the natural spacing $a > 0$ and we assume that the elements at the edges are also fixed with this distance to a constant suspension. The interior elements have four neighbors at distance a in equilibrium. We assume, first that the elements have equal mass $m > 0$.

If ξ_{ij} and η_{ij} denote, respectively, the displacements in the directions of the X and Y axes relative to the equilibrium position, then the kinetic energy on each node ij is given by:

$$T_{ij} = \frac{1}{2}\, m \left(\dot{\xi_{ij}}^{\,2} + \dot{\eta_{ij}}^{\,2} \right) \tag{1}$$

where the dot means derivative with respect to time.

Then, the total kinetic energy is:

$$T = \sum_{i,j=0}^{n} T_{ij} = \sum_{i,j=0}^{n} \frac{1}{2}\, m \left(\dot{\xi_{ij}}^{\,2} + \dot{\eta_{ij}}^{\,2} \right) \tag{2}$$

The potential energy V_{ij} for each node, ij, is obtained by applying Hooke's law, considering elastic interactions with neighbors:

$$V_{ij} = \frac{1}{2}\, \sigma \left[(\xi_{i+1,j} - \xi_{i,j})^2 + (\eta_{i+1,j} - \eta_{i,j})^2 \right] \tag{3}$$

where $\sigma > 0$ is the same elasticity constant for all nodes.

Then, the total potential energy is:

$$V = \sum_{i,j=0}^{n} V_{ij} = \frac{1}{2}\, \sigma \sum_{i,j=0}^{n} \left[(\xi_{i+1,j} - \xi_{i,j})^2 - (\eta_{i+1,j} - \eta_{i,j})^2 \right] \tag{4}$$

Therefore, the Lagrangian, L, has the form:

$$L = T - V = \frac{1}{2}\sum_{i,j=0}^{n}\left[m\left(\dot{\xi_{ij}}^2 + \dot{\eta_{ij}}^2\right) - \sigma\left(\xi_{i+1,j} - \xi_{i,j}\right)^2 - \sigma\left(\eta_{i+1,j} - \eta_{i,j}\right)^2\right] \tag{5}$$

Let us now apply the Euler-Lagrange equations to obtain the equations of motion. We will apply the Euler-Lagrange equations for each coordinate ξ_{kl} and η_{kl}, with $k, l \in \{1, \ldots, n\}$, taking into account that the edges $i = 0$, $i = n + 1$ or $j = 0$, $j = n + 1$ are fixed (the displacements are zero) and are not dynamic degrees of freedom.

Equation for ξ_{kl}:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\xi}_{kl}}\right) - \frac{\partial L}{\partial \xi_{kl}} = 0. \tag{6}$$

1. **Kinetic term:**

$$\frac{\partial L}{\partial \dot{\xi}_{kl}} = m\dot{\xi}_{kl},$$

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\xi}_{kl}}\right) = m\ddot{\xi}_{kl}.$$

2. **Potential term:** appears in any term of L that contains ξ_{kl}. - When $(i, j) = (k, l)$:

$$-\frac{\sigma}{2} \cdot 2(\xi_{k+1,l} - \xi_{k,l}) \cdot (-1) = +\sigma(\xi_{k+1,l} - \xi_{k,l}).$$

(Derivative with respect to ξ_{kl} of $-\frac{\sigma}{2}(\xi_{k+1,l} - \xi_{kl})^2$ gives $+\sigma(\xi_{k+1,l} - \xi_{kl})$).

- When $(i, j) = (k - 1, l)$:

$$-\frac{\sigma}{2} \cdot 2(\xi_{k,l} - \xi_{k-1,l}) \cdot (1) = -\sigma(\xi_{k,l} - \xi_{k-1,l}).$$

(Derivative with respect to ξ_{kl} of $-\frac{\sigma}{2}(\xi_{k,l} - \xi_{k-1,l})^2$ gives $-\sigma(\xi_{k,l} - \xi_{k-1,l})$).

- There are no contributions of $(i, j) = (k, l - 1)$ or $(k, l + 1)$ for ξ because ξ only appears with index $i + 1, i$ in the first term of the difference $(\xi_{i+1,j} - \xi_{ij})$; it does not appear in $(\eta_{i+1,j} - \eta_{ij})$.

So,

$$\frac{\partial L}{\partial \xi_{kl}} = \sigma(\xi_{k+1,l} - \xi_{k,l}) - \sigma(\xi_{k,l} - \xi_{k-1,l}).$$

Euler-Lagrange equations:

$$m\ddot{\xi}_{kl} - [\sigma(\xi_{k+1,l} - \xi_{k,l}) - \sigma(\xi_{k,l} - \xi_{k-1,l})] = 0, \tag{7}$$

$$m\ddot{\xi}_{kl} = \sigma(\xi_{k+1,l} - 2\xi_{k,l} + \xi_{k-1,l}). \tag{8}$$

The equations in the displacements in Y axis are identical to (7) and (8), but considering variables η. The boundary conditions are $\xi_{0,j} = \xi_{n+1,j} = \xi_{i,0} = \xi_{i,n+1} = 0$ and similarly for η.

If we now express the equations of motion in matrix form. First for the case of all equal masses. If we compare the equations of motion in ξ and η, we observe that the exchange $\xi \leftrightarrow \eta$ implies simultaneously the transposition of the matrix elements such that $(ij) \leftrightarrow (ji)$. This symmetry operation is related to the reflection of the lattice about the diagonal in the first quadrant of the coordinates.

If we identify the square matrices $X = (\xi_{kl}), Y = (\eta_{kl}) \in \mathbb{R}^{n\times n}$, then we can express the equations of motion in matrix form compactly like this:

$$m\ddot{X} + \sigma X A_0 = 0 \tag{9}$$

$$m\ddot{Y} + \sigma Y A_0 = 0 \tag{10}$$

where $A_0 \in \mathbb{R}^{n\times n}$ is the invertible tridiagonal matrix:

$$A_0 = \begin{pmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & 0 & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & 0 & & & -1 & 2 & -1 \\ & & & & & -1 & 2 \end{pmatrix} \tag{11}$$

If the values of the matrices X and Y and their derivatives are known at a time $t = 0$, the dynamic evolution of the physical system is fully determined by the following initial value problems:

$$m\ddot{X}(t) + \sigma X(t) A_0 = 0, \quad t \in [0, T] \tag{12}$$
$$X(0) = X_0, \quad \dot{X}(0) = \dot{X}_0$$

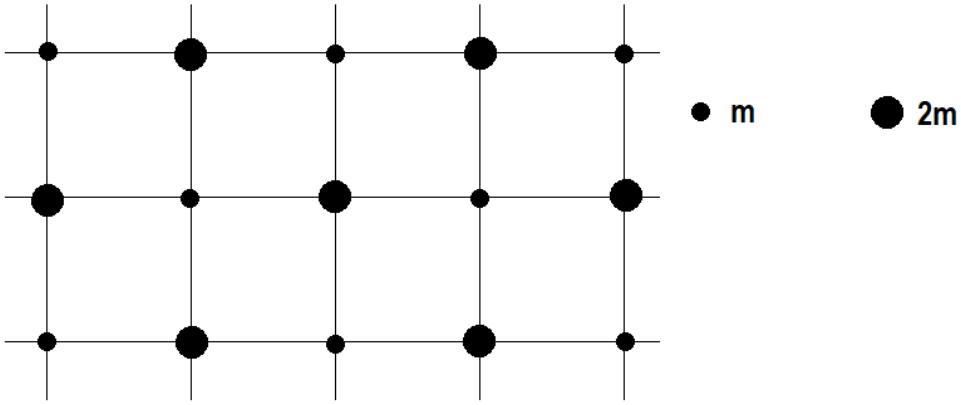
$$m\ddot{Y}(t) + \sigma Y(t) A_0 = 0, \quad t \in [0, T] \tag{13}$$
$$Y(0) = Y_0, \quad \dot{Y}(0) = \dot{Y}_0$$

Let us now solve analytically the equations of motion for the case of equal masses.
Taking $A = \dfrac{\sigma}{m} A_0$ we can write equations (12)-(13) in the form

$$\ddot{X}(t) + X(t)A = 0, \ddot{Y}(t) + Y(t)A = 0, \quad t \in [0, T]. \tag{14}$$

3 Results in the new context

Let us now modify the problem so that the masses alternate with m and $2m$ on both the X and Y axes.



Let us write the new kinetic energy and the new potential energy, the new Lagrangian and the equations of motion (see [2]).

We will modify the problem so that the masses alternate in the mesh: each node (i, j) will have a mass m_{ij} that is worth m or $2m$ alternatively in both directions i and j.

$$m_{ij} = m\,(1 + \varepsilon_{ij})\,, \tag{15}$$

where $\varepsilon_{ij} = 1$ if $i + j$ is odd (or even, depending on the convention), and 0 otherwise.

But specifically: alternation means that if we define $m_{ij} = m$ when $i+j$ is even, and $m_{ij} = 2m$ when $i + j$ is odd, then there is a mass m and another $2m$ alternating like a chessboard.

Kinetic energy:

$$T = \frac{1}{2}\sum_{i,j=1}^{n} m_{ij}\left(\dot{\xi}_{ij}^2 + \dot{\eta}_{ij}^2\right), \quad m_{ij} = \begin{cases} m & \text{if } i+j \text{ even}, \\ 2m & \text{if } i+j \text{ odd}. \end{cases} \tag{16}$$

Potential energy:

Assume that the elastic couplings are the same for all neighboring interactions in the i direction (the same constant σ for all springs), and there is no coupling in the j direction. This is given by the original statement and we will keep it:

$$V = \frac{1}{2}\,\sigma \sum_{i,j=0}^{n} \left[(\xi_{i+1,j} - \xi_{i,j})^2 + (\eta_{i+1,j} - \eta_{i,j})^2\right]. \tag{17}$$

As before, $\xi_{0,j} = \xi_{n+1,j} = 0$, etc.

Lagrangian:

$$L = T - V = \frac{1}{2}\sum_{i,j=1}^{n} m_{ij}(\dot{\xi}_{ij}^2 + \dot{\eta}_{ij}^2) - \frac{1}{2}\sigma \sum_{i,j=0}^{n} \left[(\xi_{i+1,j} - \xi_{i,j})^2 + (\eta_{i+1,j} - \eta_{i,j})^2\right]. \tag{18}$$

Equations of motion for ξ_{kl}:

We apply Euler-Lagrange:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\xi}_{kl}}\right) - \frac{\partial L}{\partial \xi_{kl}} = 0.$$

1. **Kinetic term:**

$$\frac{\partial L}{\partial \dot{\xi}_{kl}} = m_{kl}\dot{\xi}_{kl} \quad \Rightarrow \quad \frac{d}{dt}(\cdots) = m_{kl}\ddot{\xi}_{kl}.$$

2. **Potential term:**

Contributions only from the links $(\xi_{i+1,j} - \xi_{i,j})^2$ that contain ξ_{kl}. - From the link between (k, l) and $(k + 1, l)$: term $-\frac{\sigma}{2}(\xi_{k+1,l} - \xi_{kl})^2$ gives derivative with respect to ξ_{kl}: $+\sigma(\xi_{k+1,l} - \xi_{kl})$. - From the link between $(k - 1, l)$ and (k, l): term $-\frac{\sigma}{2}(\xi_{kl} - \xi_{k-1,l})^2$ gives derivative with respect to ξ_{kl}: $-\sigma(\xi_{kl} - \xi_{k-1,l})$.

So,

$$\frac{\partial L}{\partial \xi_{kl}} = \sigma(\xi_{k+1,l} - \xi_{kl}) - \sigma(\xi_{kl} - \xi_{k-1,l}) = \sigma(\xi_{k+1,l} - 2\xi_{kl} + \xi_{k-1,l}).$$

Equations of movement:

$$m_{kl}\ddot{\xi}_{kl} - \sigma(\xi_{k+1,l} - 2\xi_{kl} + \xi_{k-1,l}) = 0, \tag{19}$$

$$m_{kl}\ddot{\xi}_{kl} = \sigma(\xi_{k+1,l} - 2\xi_{kl} + \xi_{k-1,l}). \tag{20}$$

Equations for η_{kl}: similarly:

$$m_{kl}\ddot{\eta}_{kl} = \sigma(\eta_{k+1,l} - 2\eta_{kl} + \eta_{k-1,l}). \tag{21}$$

Now we will find a similar way to obtain matricial equations of movement like (9) and (10), with another different matrix A_0, but that captures the alternation of masses and has a formulation like the one that comes out when the masses are equal.

Indeed, the alternation of masses can be absorbed in a different matrix A that is not the same for all components, so that the equations remain in the form:

$$\ddot{X} = -\frac{\sigma}{m} XA \quad \text{(or similar)}$$

The idea is the following: the original equations with alternating masses are:

$$m_{kl}\ddot{\xi}_{kl} = \sigma\big(\xi_{k+1,l} - 2\xi_{k,l} + \xi_{k-1,l}\big).$$

If we define $\mu_{kl} = \frac{m}{m_{kl}}$, that is,

$$\mu_{kl} = \begin{cases} 1, & k+l \text{ even}, \\ \frac{1}{2}, & k+l \text{ odd}, \end{cases} \tag{22}$$

then the equation can be rewritten as:

$$\ddot{\xi}_{kl} = \frac{\sigma}{m}\mu_{kl}\big(\xi_{k+1,l} - 2\xi_{k,l} + \xi_{k-1,l}\big).$$

This suggests defining a matrix $B \in \mathbb{R}^{n\times n}$ with elements:

$$B_{kl} = \mu_{kl} \quad \text{(coefficients alternating 1 and 1/2)}.$$

Then the operator on the right is: at each node (k,l), we multiply $(\xi_{k+1,l} - 2\xi_{kl} + \xi_{k-1,l})$ by B_{kl} and then by σ/m.

We can write this as:

$$\ddot{X} = -\frac{\sigma}{m}\,(B \odot (A_0 X))$$

where $\odot$ is the Hadamard product, and A_0X is the usual product. In order to numerically solve the equations of movement, it will be better to avoid the Hadamard product.

To avoid this, we note that B_{kl} depends only on $k+l$. Can we define a block-diagonal matrix? Not exactly, because B is not diagonal, but has a checkerboard pattern. However, we can introduce a matrix A such that XA produces finite differences with different weights depending on the position. Would we achieve this if instead of using the same matrix A_0 for all rows, we use a tridiagonal matrix **with row-dependent elements**? Let's look at it:

The operator $(\xi_{k+1,l} - 2\xi_{kl} + \xi_{k-1,l})$ is the element (k,l) of XA_0^{T} if A_0^{T} acts from the right, but with changed sign as we saw. Specifically,

$$\xi_{k+1,l} - 2\xi_{kl} + \xi_{k-1,l} = -(A_0X)_{kl}.$$

Then the equation becomes:

$$\ddot{\xi}_{kl} = -\frac{\sigma}{m} B_{kl}(A_0X)_{kl}.$$

This is not a matrix product because B_{kl} multiplies component by component.

To make it matrix-free without Hadamard product, the matrix B can be introduced as a **block-diagonal matrix** if we rearrange the unknowns into a vector. Let $\mathbf{x} = \mathrm{vec}(X) \in \mathbb{R}^{n^2}$ (stacked

columns). Then the operator (A_0X) in vectorized form is $(I_n \otimes A_0)\mathbf{x}$ (what if A_0 acts from the left? Be careful).

Actually:

$$\text{vec}(A_0X) = (I_n \otimes A_0)\text{vec}(X) = (I_n \otimes A_0)\mathbf{x}.$$

The term $B_{kl}(A_0X)_{kl}$ is the Hadamard product of $\text{vec}(B)$ and $\text{vec}(A_0X)$, that is, $\text{diag}(\text{vec}(B)) \cdot (I_n \otimes A_0)\mathbf{x}$.

So the vector equation is

$$\ddot{\mathbf{x}} = -\frac{\sigma}{m}\text{diag}(\mathbf{b})(I_n \otimes A_0)\mathbf{x},$$

where $\mathbf{b} = \text{vec}(B) \in \mathbb{R}^{n^2}$.

Now we define a matrix $A_{\text{total}} \in \mathbb{R}^{n^2 \times n^2}$ as:

$$A_{\text{total}} = \text{diag}(\mathbf{b})(I_n \otimes A_0). \tag{23}$$

Then the equation is:

$$\ddot{\mathbf{x}} = -\frac{\sigma}{m}A_{\text{total}}\mathbf{x}, \tag{24}$$

which has exactly the same form (but now is a vector equation) as the case of equal masses but with a different matrix A_{total}, which incorporates the alternation of masses.

Analogously for $\mathbf{y} = \text{vec}(Y)$.

4 Conclusions

In this communication, we present the first steps for the generalization of the study of harmonic oscillators on the lattice (see [1]). This first work extends the study to the alternance of the masses in the nodes of the lattice. We have deduced the equations of movement for this new case using the Euler-Lagrange formalism. Then we have prepared these equations of movement for their analytical and numerical resolution. Notice that the new equations have a similar form as the one for constant masses (see [1]), but now matrix A has a more general form. This new general form requires a specific numerical study. This study is in progress and should be presented in a near future. Moreover, this work will allow us to apply our results to ideal metallic and crystal molecular structures.

Acknowledgments

This work has been partially supported by PAID-11-25 of the Universitat Politècnica de València and the Generalitat Valenciana Grant CIAICO/2023/275

References

[1] M. M. Tung, J. Ibáñez, E. Defez, and J. Sastre, "Simulation of harmonic oscillators on the lattice," *Mathematical Methods in the Applied Sciences*, vol. 43, no. 14, pp. 8237–8252, 2020.

[2] L. D. Landau and E. M. Lifshitz, *Mechanics. Volume 1 of* Course of Theoretical Physics. Butterwoth-Heinenann, Elsevier Ltd; Oxford, England, 1976.

Quantum Trajectories of Logistic Dynamics

Joan C. Micó[b, 1]

(b) Institut Universitari de Matemàtica Multidisciplinar. Universitat Politècnica de València
Camí de Vera s/n, ciutat de València, València (Spain).

1 Introduction

The translation of the quantum formalism from physical systems to Abstract Dynamical Systems (ADS) has been able to be done in [1]. In fact, the work [1] presents a first application to the quantum ADS dynamics given by the simplest case, the one-dimensional stationary case, provided by the quantum logistic dynamics. The objective of this application case is to obtain its quantum trajectories. However, this application fails in some parts of its theoretical development, due to, the work [2] tries to correct those failing parts. However, once the corrections are carried out in [2], the conclusion is amazing: the quantum terms, i.e., the quantum energies and the quantum wave functions vanish in the quantum trajectories. In fact, this problem also happens in the one-dimensional stationary cases of physical systems, and a great amount of scientific literature have been devoted to solve this problem. Two highlighting works [3, 4] present widely the advances done in physics to solve that problem by focusing the quantum dynamics on the Madelung hydrodynamic formulation of the wave function [5], rather in the solution of the corresponding time-independent Schrödinger equation, also followed by Bohm and Hiley [6]. As a consequence, the Ermakov equation plays a fundamental role to obtain, at least theoretically, the quantum trajectories.

2 The ADS one-dimensional stationary case

The classical ADS approach for this case is provided by the one-dimensional autonomous differential equation [1, 2]:

$$\dot{q}(t) = f(q)\,;\; q(t_0) = q_0 \tag{1}$$

Eq. 1 provides the time-independent Schrödinger equation [1, 2]:

$$\psi''(q) + \left(\frac{f^2(\mathrm{q})}{\sigma^2} + \frac{2E}{\sigma^2}\right)\psi(q) = 0 \tag{2}$$

[1] jmico@mat.upv.es

In Eq. 2, E is the system energy and σ the system Planck constant. The corresponding ADS quantum formulation of Eq. 1 is provided by the following equations [1, 2]:

$$\left.\begin{array}{c} \dot{q}(t) = f(q) - \frac{x}{2}\ ;\ q(t_0) = q_0 \\ \dot{x}(t) = Q'(q)\ ;\ x(t_0) = 0 \end{array}\right\} \tag{3}$$

where $Q(q)$ is the quantum potential:

$$Q(q) = -\frac{\sigma^2}{2R(q)} R''(q) \tag{4}$$

In Eq. 4, $R(q) = |\psi(q)|$, which corresponds to the polar split of the wave function, Eq. 2, provided by Madelung [5] for the stationary case:

$$\psi(q) = |\psi(q)|\, e^{i\frac{W(q)}{\sigma}} = R(q)\, e^{i\frac{W(q)}{\sigma}} \tag{5}$$

being $W(q)$ the quantum principal Hamilton function, related with the quantum action, such as it happens in the classical case, as: $S(t,q) = -E \cdot t + W(q)$. Note that the corresponding time-dependent wave function is then written as:

$$\Psi(t,q) = R(q)\, e^{i\frac{S(t,q)}{\sigma}} = R(q)\, e^{i\frac{W(q)}{\sigma}} e^{-i\frac{E}{\sigma}t} \tag{6}$$

Besides, in [2], the solution of the time-independent Schrödinger equation, Eq. 2, for the case of the logistic dynamics ($f(q) = a \cdot q - b \cdot q^2$; $a, b > 0$; and $c = a/b$ is the attractor of the logistic dynamics) is provided for the following boundary problem:

$$\left.\begin{array}{c} \theta_n''(q) + \left(\theta_n'(q)\right)^2 + \frac{f^2(q)}{\sigma^2} = \frac{2E_n}{\sigma^2}\left(e^{-4\theta_n(q)} - 1\right)\ ;\ \theta_n(0) = \theta_n(c) = 0 \\ \varphi_n'(q) = \frac{\sqrt{2E_n}}{\sigma} e^{-2\theta_n(q)}\ ;\ \varphi_n(0) = 0\ ;\ \varphi_n(c) = \pi \cdot n\ ;\ n \in \mathbb{Z}_{>0} \end{array}\right\} \tag{7}$$

$$\psi_n(q) = A_n \cdot e^{\theta_n(q)} \cdot sin\left(\frac{\sqrt{2E_n}}{\sigma} \int_0^q e^{-2\theta_n(y)}\, dy\right) \tag{8}$$

where $\theta(q) \to -\infty$ as $q \to \pm\infty$, then $\psi_n(q) \to 0$ as $q \to \pm\infty$, i.e., the eigenfunctions $\psi_n(q)$ are bounded states. In Eqs. 7 and 8 is assumed that $\psi_n(0) = \psi_n(c) = 0$ to obtain the quantum energies $E_n > 0$. However, the change $\omega_n(q) = e^{\theta_n(q)}$ (being $\omega_n(q)$ non-dimensional) provides the version of the solution here used for the same boundary problem:

$$\left.\begin{array}{c} \omega_n''(q) + \left(\frac{2E_n}{\sigma^2} + \frac{f^2(q)}{\sigma^2}\right)\omega_n(q) + \frac{f^2(q)}{\sigma^2} = \frac{2E_n}{\sigma^2}\frac{1}{\omega_n^3(q)}\ ;\ \omega_n(0) = \omega_n(c) = 1 \\ \varphi_n'(q) = \frac{\sqrt{2E_n}}{\sigma}\frac{1}{\omega_n^2(q)}\ ;\ \varphi_n(0) = 0\ ;\ \varphi_n(c) = \pi \cdot n\ ;\ n \in \mathbb{Z}_{>0} \end{array}\right\} \tag{9}$$

Then, the quantized stationary wave functions are written as:

$$\psi_n(q) = A_n \cdot \omega_n(q) \cdot sin\left(\frac{\sqrt{2E_n}}{\sigma}\int_0^q \frac{dy}{\omega_n^2(y)}\right) \quad ; \quad n \in \mathbb{Z}_{>0} \tag{10}$$

The quantization of the wave function, Eq. 10, implies the quantization of the quantum potential, $Q_n(q)$, Eq. 4. Therefore, Eq. 3 becomes an infinite set of quantum trajectories:

$$\left.\begin{array}{c} \dot{q}_n(t) = f(q) - \frac{x_n}{2}\ ;\ q_n(t_0) = q_0 \\ \dot{x}_n(t) = Q_n'(q)\ ;\ x_n(t_0) = 0 \end{array}\right\} \quad ; \quad n \in \mathbb{Z}_{>0} \tag{11}$$

However, note that the derivative of the absolute value of a real function in Eq. 11, $Q_n'(q)$, such that $R_n(q_n) = |\psi_n(q_n)|$, is computed as $R_n'(q) = \frac{\psi_n(q)}{|\psi_n(q)|}\psi_n'(q)$. From this last result and that from $|\psi_n(q)|^2 = \psi_n^2(q)$:

$$Q_n(q) = \frac{f^2(q_n)}{2} + E_n\ ;\ Q_n'(q) = f(q) \cdot f'(q) \tag{12}$$

In other words, both the quantum energies, E_n, and the quantum wave functions, $\psi_n(q)$, vanish to compute the quantum trajectories of Eq. 11. Is there any solution for this problem? The following sections are devoted to find this solution.

3 The hydrodynamic approach to the quantum logistic dynamics

The hydrodynamic approach was provided by Madelung [5] by the polar split of the wave function. Following the conclusions by Holland [3] and Wyatt [4], this approach provides more information for the quantum trajectories than that provided only by the wave function. The hydrodynamic approach for the one-dimensional stationary case arises with the substitution of the polar split, Eq. 5, in the time-independent Schrödinger equation, Eq. 2, and subsequently equalling the real and imaginary parts to zero:

$$-E + \frac{1}{2}\big(W'(\mathrm{q})\big)^2 - \frac{1}{2}f^2(\mathrm{q}) - \frac{\sigma^2}{2R(q)}R''(q) = 0 \tag{13}$$

$$\frac{d}{dq}\big(W'(\mathrm{q}) \cdot R^2(q)\big) = 0 \tag{14}$$

Eq. 14 provides $W'(\mathrm{q}) = K/R^2(q)$, being K an indetermined constant. The substitution of this result in Eq. 13, the multiplication by $2R(q)$ and subsequent reorganization, provides the following Ermakov differential equation for $R(q)$:

$$R''(q) + \left(\frac{f^2(\mathrm{q})}{\sigma^2} + \frac{2E}{\sigma^2}\right)R(q) = \frac{K^2}{\sigma^2}\frac{1}{R^3(q)} \tag{15}$$

Note that the dimensions $[R(q)] = [q]^{-1/2}$, due to $R^2(q)$ has the dimensions of a probability density, therefore in Eq. 15: $[K] = [\sigma][q]^{-1/2}$. However, comparing Eq. 15 with Eq. 2, via Eq. 5, $R(q) \neq |\psi(q)|$. Only if $K = 0$ in Eq. 15, $R(q) = |\psi(q)|$. This subtle difference makes that

Eq. 15 provides more quantum information than just Eq. 2 does it. In addition, the analytical solution of the Ermakov equation can be found in [7] (in Section 2.9.1-1., item 2, and referred there as *Yermakov equation*):

$$C \cdot R^2(q) = \eta^2(q)\left[\frac{K^2}{\sigma^2} + \left(D + C\int_0^q \frac{dy}{\eta^2(y)}\right)^2\right] \tag{16}$$

In Eq. 16, $C \neq 0$ and D are two arbitrary constants, and $\eta(q)$ is a non-trivial solution of the associated homogeneous equation of Eq. 15:

$$\eta''(q) + \left(\frac{f^2(\mathrm{q})}{\sigma^2} + \frac{2E}{\sigma^2}\right)\eta(q) = 0 \tag{17}$$

Note in Eq. 17 that the identification $\eta(q) = |\psi(q)|$ can be held. In fact, in Eq. 16, $R(q) = 0$ if $\eta(q) = 0$. Therefore, a non-trivial solution of $\eta(q)$ can be taken as in Eq. 2: $\eta(0) = \eta(c) = 0$, which provides the same quantized solutions of Eq. 9 for $\eta(q)$ (in fact, the same symbols are used) and its substitution in Eq. 16:

$$C_n \cdot R_n^2(q) = A_n^2 \cdot \omega_n^2(q) \cdot sin^2\big(\varphi_n(q)\big)\left[\frac{K_n^2}{\sigma^2} + \left(D_n + \frac{C_n}{A_n^2}\int_0^q \frac{dy}{\omega_n^2(y)\cdot sin^2(\varphi_n(y))}\right)^2\right] \tag{18}$$

In Eq.18, both $\omega_n(q)$ and $\varphi_n(q)$ obey Eq. 7. Note in addition that Eq. 18 just depends on three independent family of constants, $F_n^2 = A_n^2/C_n$, K_n^2 and D_n, that is:

$$R_n^2(q) = F_n^2 \cdot \omega_n^2(q) \cdot sin^2\big(\varphi_n(q)\big)\left[\frac{K_n^2}{\sigma^2} + \left(D_n + \frac{1}{F_n^2}\int_0^q \frac{dy}{\omega_n^2(y)\cdot sin^2(\varphi_n(y))}\right)^2\right] \tag{19}$$

In addition, as Holland [3] and Wyatt [4] also emphasize for physical systems, the solution $R_n(q)$ of Eq. 19 can be substituted in the polar split of Eq. 5 to obtain a generalized form of the wave function, by also computing $W_n(q)$ by integrating $W_n'(\mathrm{q}) = K_n/R_n^2(q)$ from Eq. 14. However, the objective of this article is to find the quantum trajectories by Eq. 11. The study of the corresponding generalized eigenfunctions $\psi_n(q) = R_n(q)\, e^{i\frac{W_n(q)}{\sigma}}$ as a possible orthonormal base of a Hilbert space must be studied in subsequent investigations.

In addition, note that once the solution of the Ermakov equation (Eq. 15) has been found out by Eq. 19, the quantized quantum potential $Q_n(q)$ (Eq. 4) can be computed. However, the use of Eq. 15 is an easier way to compute it instead the direct derivation on $R_n(q)$ in Eq. 19. Dividing Eq. 15 by $R_n(q)$ and multiplying it by $-\sigma^2/2$:

$$Q_n(q) = -\frac{\sigma^2}{2R_n(q)}R_n''(q) = \frac{f^2(q)}{2} + E_n - \frac{K_n^2/2}{\left(R_n^2(q)\right)^2} \tag{20}$$

Therefore, the derivative of the quantum potential, arising in the quantum trajectories of Eq. 11, becomes:

$$Q_n'(q) = f(q)\cdot f'(q) + K_n^2 \frac{\left(R_n^2(q)\right)'}{\left(R_n^2(q)\right)^3} \tag{21}$$

being the numerator multiplied by K_n^2:

$$\left(R_n^2(q)\right)' = 2F_n^2\left[\omega_n(q)\cdot\omega_n'(q)\cdot sin^2\left(\varphi_n(q)\right) + \frac{\sqrt{2E_n}}{\sigma} sin\left(\varphi_n(q)\right)\cdot cos\left(\varphi_n(q)\right)\right]\left[\frac{K_n^2}{\sigma^2} + \left(D_n + \frac{1}{F_n^2}\int_0^q \frac{dy}{\omega_n^2(y)\cdot sin^2\left(\varphi_n(y)\right)}\right)^2\right] + 2\left(D_n + \frac{1}{F_n^2}\int_0^q \frac{dy}{\omega_n^2(y)\cdot sin^2\left(\varphi_n(y)\right)}\right) \tag{22}$$

Note that the dependence of the quantum potential derivative $Q_n'(q)$ on three family of constants, F_n^2, K_n^2 and D_n, makes difficult their values' choice. The objective of this values' choice is to prove that it provides new information in the quantum trajectories by Eq. 11 that the classical trajectories by Eq. 1 do not provide. Therefore, the first work's hypotheses done in this research about this values' choice have been: $F_n = 1$, $D_n = 0$ and $K_n = \mathrm{X}_n\cdot\sigma$. The two first values' choice are just simplifying hypotheses, and the third value's choice is based in the dimensional analysis, deduced above from Eq. 15: $[K_n] = [\sigma][q]^{-1/2}$. Note then that $[\mathrm{X}_n] = [q]^{-1/2}$. The substitution of these constants' values in Eqs. 19 and 22 (after inserting the result of Eq. 21 in Eq. 22) provide for these equations:

$$R_n^2(q) = \omega_n^2(q)\cdot sin^2\left(\varphi_n(q)\right)[\mathrm{X}_n^2 + \phi_n^2(q)] \tag{23}$$

$$Q_n'(q) = f(q)\cdot f'(q) + \frac{2\cdot\sigma^2\cdot\mathrm{X}_n^2}{\left(R_n^2(q)\right)^3}\left\{\left[\omega_n(q)\cdot\omega_n'(q)\cdot sin^2\left(\varphi_n(q)\right) + \frac{\sqrt{2E_n}}{\sigma} sin\left(\varphi_n(q)\right)\cdot cos\left(\varphi_n(q)\right)\right][\mathrm{X}_n^2 + \phi_n^2(q)] + \phi_n(q)\right\} \tag{24}$$

such that in both equations:

$$\phi_n(q) = \int_0^q \frac{dy}{\omega_n^2(y)\cdot sin^2\left(\varphi_n(y)\right)} \tag{25}$$

Therefore, as theoretical conclusion, the computation of the quantum trajectories need to follow the steps: (1) find the solution of the boundary problem given by Eq. 9, which provides the $\omega_n(q)$ and $\varphi_n(q)$ functions; (2) compute $Q_n'(q)$ by Eq. 24; (3) substitute the result of Eq. 24 in Eq. 11 for any $n \in \mathbb{Z}_{>0}$.

However, providing $R_n^2(q)$ of Eq. 23 normalized to unit is possible. Unlike Eq. 10, where the normalization consists in doing $\int_0^{+\infty}\psi_n(q)\,dq = 1$, from which $A_n = \int_0^{+\infty}\omega_n(q)\cdot sin\left(\frac{\sqrt{2E_n}}{\sigma}\int_0^q \frac{dy}{\omega_n^2(y)}\right)^{-1}$, Eq. 23 is the solution of the Ermakov equation, Eq. 15, which is a nonlinear differential equation. Then, if in Eq. 15 the substitution $\bar{R}_n(q) = \bar{K}_n\cdot R_n(q)$, $\bar{K}_n \neq 0$, is done (quantization is assumed to be done already), the result is:

$$\bar{R}_n''(q) + \left(\frac{f^2(q)}{\sigma^2} + \frac{2E}{\sigma^2}\right)\bar{R}_n(q) = \frac{K_n^2}{\bar{K}_n^4 \cdot \sigma^2}\frac{1}{\bar{R}_n^3(q)} \tag{26}$$

whose general solution is (conveniently with $D_n = 0$), as in Eq. 19:

$$\bar{R}_n^2(q) = \bar{F}_n^2 \cdot \omega_n^2(q) \cdot sin^2\big(\varphi_n(q)\big)\left[\frac{K_n^2}{\bar{K}_n^4 \cdot \sigma^2} + \frac{1}{\bar{F}_n^4}\phi_n^2(q)\right] \tag{27}$$

being $\bar{F}_n \neq 0$. Taking in Eq.27, again $K_n = \mathrm{X}_n \cdot \sigma$, and from Eq. 23, $\bar{K}_n^2 = \bar{F}_n^2 = \int_0^{+\infty} R_n^2(q)\, dq = \mathrm{X}_n^2 \cdot I_n + J_n = 1$, with $I_n = \int_0^{+\infty} \omega_n^2(q) \cdot sin^2\big(\varphi_n(q)\big)\, dq$ and $J_n = \int_0^{+\infty} \omega_n^2(q) \cdot sin^2\big(\varphi_n(q)\big) \cdot \phi_n^2(q)\, dq$, the normalized version of Eq. 23 arises by dividing Eq. 23 by $\bar{K}_n^2 = \bar{F}_n^2 = \mathrm{X}_n^2 \cdot I_n + J_n$:

$$\bar{R}_n^2(q) = \frac{1}{\mathrm{X}_n^2 \cdot I_n + J_n}\omega_n^2(q) \cdot sin^2\big(\varphi_n(q)\big)[\mathrm{X}_n^2 + \phi_n^2(q)] \tag{28}$$

In addition, proving that $Q_n'(q)$ in Eq. 24 becomes invariant by this change is trivial.

4 Application case

All computations and figures here presented have been done with the software MATHEMATICA 14.3. In addition, a first empirical case that provides a possible quantum effect for the logistic dynamics is identified for the world population evolution in the period 1955-2025. The empirical data have been taken, in population billions, from the web page https://www.worldometers, cited in Wikipedia. Although, at short term, the world population dynamics is considered as exponential, at middle or long terms its dynamics tends asymptotically to a saturation value, which nature is explained by the limits of resources needed for the population growth [8]. The logistic dynamics is capable to explain this pattern. If in Eq. 1 the logistics function is written as $f(q) = a \cdot q - b \cdot q^2$, with parameters $a, b > 0$, the analytical solution of Eq. 1 is:

$$q(t) = \frac{c}{1 + \left(\frac{c}{q_0} - 1\right)\mathrm{e}^{-a(t-t_0)}} \tag{29}$$

In Eq. 29, $c = a/b$ is the dynamical attractor or saturation level to which the dynamics tends asymptotically as $t \to +\infty$. If Eq. 29 is fitted with the *NonlinearModelFit* function of MATHEMATICA 14.3, the outcomes provided are $a = 0.026$ and $b = 0.002$, then $c = 13.0$, with a determination coefficient of $r^2 = 0.9997$, being $t_0 = 1950$ and $q_0 = 2.4931$ the corresponding population (in billions) for this year. In fact, Figure 1 (right) shows, at first sight, that the predicted curve aligns with the empirical data. However, the centre of Figure 1 shows that the residuals versus time have a double oscillating dynamical pattern between -0.06 and 0.05 amplitudes, which have not been captured by the logistic dynamics fitting. In addition, the Anderson-Darling normality test for the residuals provides a *pvalue* of 0.003, and for the Chi-Square test a *pvalue* of 0.04, i.e., significantly the residuals are not noise, and its dynamics provides an oscillating pattern that has not been captured by the logistic dynamics. As

hypothesis, this oscillating pattern is assumed to be caused by the quantum dynamics provided by Eqs. 11 and 24.

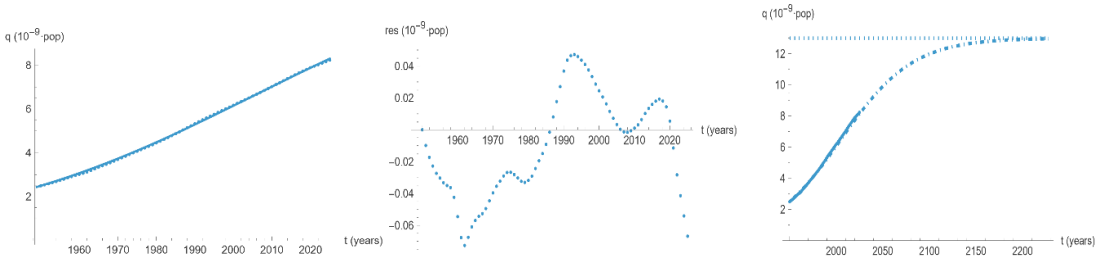


Figure 1. Left, the empirical world population dynamics in the period 1955-2025 (dots) together the logistic dynamics fitted versus time in the same period (continuous curve). Centre, the residuals versus time. Right, the fitted extrapolation logistic dynamics to the 2227 year (dashed curve) together the empirical data (dots) and the explicit saturation value *c*=13.0 population billions (dotted straight line).

In addition, Figure 1 (right) also shows the extrapolation of the logistic dynamics to the 2227 year for which the dynamics is significantly close to the saturation level. This datum could provide a clue to identify the quantum level n for the boundary problem given by Eq. 9: if 2 cycles are identified visually in the period 1955-2025, there would be 7 cycles in the period 1955-2227 and, then, $n = 8$; besides, if 4 cycles are identified visually in the period 1955-2025, there would be 13 cycles in the period 1955-2227 and, then, $n = 14$. This is like this because $n - 1$ cycles in $q \in]0, c[$ correspond to n zeros of Eq. 24, which would be translated to Eq. 24.

Therefore, to solve the boundary problem given by Eq. 9 for the integers referred, the *NDSolve* function of MATHEMATICA 14.3 has been used. However, the system Planck constant's value, σ, must be chosen. Due to time is for ADS the only common and constant variable, and that a change in the temporal scale can produce a change in the laws of the ADS dynamics, the hypothesis here stated is that σ has dimensions of time and that it represents the limit of the dynamics' knowledge. Then, in the application case, this temporal limit is provided by 1 year, due to the value chosen is $\sigma = 1.0$. Using then the *NDSolve* function, the quantum energies of the quantum levels $n = 1, 2, 3$, 8, 14 have been computed by Eq. 9: $E_1 = 0.874$, $E_2 = 0.115$, $E_3 = 0.261$, $E_8 = 1.867$ and $E_{14} = 5.658$. Note that E_1 should be lesser than the other energy values. This is one of the limitations of the software, i.e., the loss of precision in some cases.

Besides, the *NDSolve* function computes the corresponding $\omega_n(q)$ and $\varphi_n(q)$ of Eq. 9. Through them, the corresponding $\phi_n(q)$ functions of Eq. 25 can be also computed by the equivalent differential problem $\phi_n'(q) = 1/\left(\omega_n^2(q) \cdot sin^2(\varphi_n(q))\right)$ with $\phi_n(0) = 0$, from which the corresponding Ermakov normalized functions of Eq. 28, $\bar{R}_n^2(q)$, and the corresponding q-derivatives of the quantum potential in Eq. 24, $Q_n'(q)$ (which is invariant under normalization). For questions of space, only $\bar{R}_8^2(q)$ is presented in Figure 2 jointly the corresponding quantum dynamics of Eq. 11 for n =8, with $X_8 = 0.76$. This value permits to compute the normalization constant, $X_8^2 \cdot I_8 + J_8 = 4.16 \cdot 10^{23}$ in $\bar{R}_8^2(q)$ of Eq. 28, and the quantum dynamics of Eq. 11 for n =8. Note that the quantum dynamics do not align with the empirical data. In fact, a slight

change in the X_8 value provides a quantum evolution below the empirical data instead above them, as showed in Figure 2. This result provides another limitation of this study.

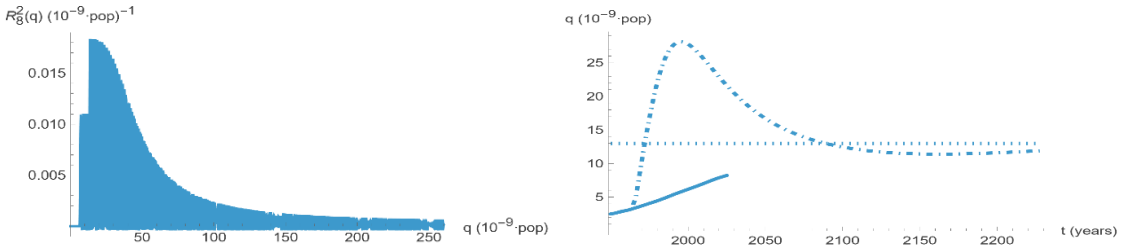


Figure 2. Left, the Ermakov normalized function $\bar{R}_8^2(q)$. Right, the empirical world population in the period 1955-2025 (dots) together the quantum logistic dynamics for n =8 extrapolated to the 2227 year (dashed curve) together the explicit saturation value *c*=13.0 population billions (dotted straight line).

5 Conclusions

(1) Program the shooting method with a Dev C++ program for Eq. 9. (2) Use subsequently a genetic algorithm, also with a Dev C++ program, to find the X_n value that can fit the quantum dynamics of Eq. 9 to the empirical data.

References

[1] J. C. Micó, "A Quantum Formalism for Abstract Dynamical Systems". *Mathematics*, 12: 1076, 2024.

[2] J. C. Micó, "Quantum Logistics Dynamics". *Modelling for Engineering & Human Behaviour 2025* (Universitat Politècnica de València, ciutat de València), 316-323, 2025.

[3] P. R. Holland, *The quantum Theory of Motion*. Cambridge, Springer, 1993.

[4] R. E. Wyatt, *Quantum Dynamics with Trajectories*. New York, Springer, 2005.

[5] E. Madelung, "Quantum Theory in Hydrodynamical Form". *Z. Für Phys. (J. Phys.)*, 40: 1–4, 1927.

[6] D. Bohm, B J. Hiley, *The Undivided Universe: An Ontological Interpretation of Quantum Theory*. London-New York, Routledge, 1995.

[7] A. D. Polyanin, V. F. Zaitsev, *Handbook of exact solutions for ordinary differential equations*. London-New York-Washington, D.C., CRC Press Company, 2003.

[8] D. H. Meadows, D. Meadows, J. Randers, *Beyond the Limits*. Vermont, Chelsea Green Publishing, 1992.

Data-Driven Scatter Correction in PET via Weighted Maximum Likelihood Expectation Maximization

Joan Prats-Climent[♭],[1] Alejandro Sanz-Sánchez[♭] and María José Rodríguez-Álvarez[♭]

(♭) Instituto de Instrumentación para Imagen Molecular (I3M)
Universitat Politècnica de València - Consejo Superior de Investigaciones Científicas
Camí de Vera s/n, 46022, València, Spain.

1 Introduction

Scatter correction remains a critical challenge in positron emission tomography (PET), as scattered photons degrade contrast and quantitative accuracy. Conventional approaches such as energy window (EW) methods and single scatter simulation (SSS) provide effective corrections but rely on model-based assumptions and may be sensitive to noise, energy resolution, or inaccuracies in attenuation maps.

In this work, we propose a data-driven scatter correction framework based on a weighted maximum likelihood expectation maximization (W-MLEM) reconstruction [1]. Event-by-event statistical weights are introduced to modulate the contribution of each coincidence during reconstruction, enabling both hard classification (binary rejection of scatter events, equivalent to filtering out the data) and soft weighting (probabilistic suppression). These weights are assumed to be derived from prior information coming from machine learning classifiers trained to distinguish between true and scattered events [2].

The proposed approach is evaluated using Monte Carlo simulated list-mode PET data. For benchmarking purposes, reconstructions obtained with the W-MLEM framework (including both hard and soft thresholding) are quantitatively compared against two widely used scatter correction strategies: the Triple Energy Window [3] method and the openSSS algorithm [4], both of which were implemented within our reconstruction pipeline.

2 Methods

2.1 List-mode MLEM formulation

Emission and detection of photons in PET imaging follow Poisson statistics. In list-mode PET reconstruction, the measured data consist of a set of detected coincidence events $\{y_i\}_{i=1}^{N}$, where each event i is associated with a specific line of response (LOR). The expected number of counts for event i is given by:

[1] joapracl@i3m.upv.es

$$\hat{y}_i = \sum_j p_{ij}\lambda_j + r_i \tag{1}$$

where λ_j is the activity in voxel j, p_{ij} is the system matrix element representing the detection probability, and r_i accounts for additive contributions such as scatter and random coincidences. The log-likelihood of the list-mode data can be written as:

$$\mathcal{L}(\lambda) = \sum_{i=1}^{N} [\log \hat{y}_i - \hat{y}_i] \tag{2}$$

Maximization using the expectation-maximization (EM) algorithm leads to the standard list-mode MLEM update equation [5]:

$$\lambda_j^{(n+1)} = \frac{\lambda_j^{(n)}}{\sum_i p_{ij}} \sum_i p_{ij} \frac{y_i}{\hat{y}_i^{(n)}} \tag{3}$$

2.2 Weighted likelihood formulation (W-MLEM)

To incorporate prior information regarding the reliability of each detected event—such as the probability of an event being a true coincidence versus a scattered one—we introduce event-wise statistical weights $w_i \in [0, 1]$.The weighted log-likelihood is then defined as:

$$\mathcal{L}_w(\lambda) = \sum_{i=1}^{N} w_i [\log \hat{y}_i - \hat{y}_i] \tag{4}$$

Maximizing this weighted likelihood within the EM framework results in the following update equation for the W-MLEM algorithm:

$$\lambda_j^{(n+1)} = \frac{\lambda_j^{(n)}}{\sum_i w_i p_{ij}} \sum_i w_i p_{ij} \frac{y_i}{\hat{y}_i^{(n)}} \tag{5}$$

This framework allows for the implementation of different scatter suppression strategies, ranging from *hard thresholding* (where $w_i \in \{0, 1\}$) to *soft weighting* (where w_i takes continuous or step-wise values). This approach enables precise modulation of each event contribution to the final reconstructed image based on its estimated classification, instead of adding the correction in the additive term. Two weighting schemes are considered based on the output of a neural network classifier. The performance of this classifier is characterized by a false positive rate (FPR) of 25% (misclassified scatter) and a false negative rate (FNR) of 6% (misclassified trues):

Hard thresholding

$$w_i = \begin{cases} 1 & \text{if event classified as true} \\ 0 & \text{if event classified as scatter} \end{cases} \tag{6}$$

Soft weighting

$$w_i \in [0, 1] \tag{7}$$

where w_i represents the probability that event i corresponds to a true coincidence. Three different values for w_i were explored in this scenario, w = 0.10, 0.25 and 0.50.

2.3 Simulation setup

Monte Carlo simulations were performed using the *Geant4 Application for Tomographic Emission* (GATE) framework [6], modeling a clinical PET system corresponding to the Siemens Biograph Vision PET scanner. The simulated object consisted of a cylindrical phantom with 6 cm radius and 15 cm axial length, representative of a simplified Utah-like phantom. Two cylindrical inserts were embedded within the background: one hot and one cold lesion, both with a radius of 1.25 cm and same 15 cm axial length. The activity concentration of the hot insert was set to a ratio of 4:1 with respect to the uniform background, while the cold insert contained no activity.

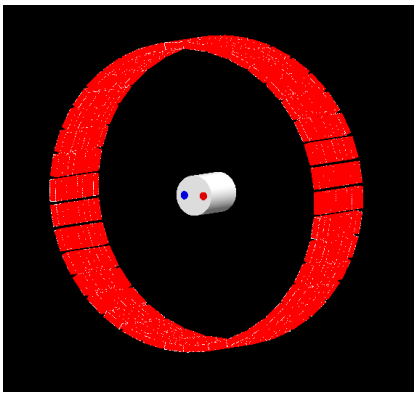


Figure 1: PET system simulation: Siemens Biograph Vision with Utah-like phantom

This phantom configuration was selected to provide a controlled yet challenging scenario for scatter correction evaluation. The combination of a uniform background with both hot and cold regions enables simultaneous assessment of contrast recovery and spill-in effects caused by scattered events. In particular, the cold insert is highly sensitive to scatter contamination, as any residual scatter leads to artificially elevated activity, while the hot insert allows evaluation of contrast loss and quantitative bias.

The simulated list-mode data were subsequently reconstructed using the CASToR software for PET image reconstruction [7]. For the proposed weighted reconstruction, the CASToR source code was modified to enable the ingestion of event-wise statistical weights obtained a priori from a neural network classifier. These weights were read from an extended list-mode format and incorporated during the reconstruction process to modulate the contribution of each event, allowing both hard and soft weighting strategies to be evaluated. This implementation enables a direct comparison between conventional reconstruction pipelines and the proposed data-driven W-MLEM approach.

2.4 Reconstruction performance

Reconstruction performance was evaluated through both qualitative and quantitative analysis. Line profiles across the reconstructed images were analyzed to assess spatial resolution and contrast behavior. In addition, several image quality metrics were computed, including hot source contrast (HC), cold source contrast (CC), contrast recovery coefficient (CRC), and spill-over ratio (SOR), which are standard figures of merit in PET reconstruction. Let $\bar{C}_{\text{hot}}$, $\bar{C}_{\text{cold}}$, and $\bar{C}_{\text{bg}}$ denote the mean reconstructed activity concentration within the hot insert, cold insert, and background regions, respectively. Let $C_{\text{bg,true}}$ and $C_{\text{hot,true}}$ denote the true activity concentrations.

Hot contrast (HC) The hot source contrast is defined as:

$$\mathrm{HC} = \frac{\bar{C}_{\mathrm{hot}} - \bar{C}_{\mathrm{bg}}}{\bar{C}_{\mathrm{bg}}} \tag{8}$$

Cold contrast (CC) The cold source contrast is defined as:

$$\mathrm{CC} = \frac{\bar{C}_{\mathrm{bg}} - \bar{C}_{\mathrm{cold}}}{\bar{C}_{\mathrm{bg}}} \tag{9}$$

Contrast Recovery Coefficient (CRC) The contrast recovery coefficient for the hot insert is defined as:

$$\mathrm{CRC} = \frac{\frac{\bar{C}_{\mathrm{hot}}}{\bar{C}_{\mathrm{bg}}} - 1}{\frac{C_{\mathrm{hot,true}}}{C_{\mathrm{bg,true}}} - 1} \tag{10}$$

Spill-Over Ratio (SOR) The spill-over ratio is defined as:

$$\mathrm{SOR} = \frac{\bar{C}_{\mathrm{cold}}}{\bar{C}_{\mathrm{bg}}} \tag{11}$$

3 Results

The quantitative evaluation of the different scatter correction methods is summarized in Table 1. The proposed data-driven W-MLEM framework showed high efficacy in recovering image contrast. The NN_hard strategy achieved a Contrast Recovery Coefficient (CRC) of 0.808 and the lowest SOR among all evaluated methods (0.267), surpassing both the conventional Energy Window (EW) and the model-based openSSS in terms of background suppression in the cold region. While openSSS provided the highest hot contrast, the NN_hard approach remained highly competitive while offering a more favorable spill-over profile.

Table 1: Quantitative comparison of image contrast metrics.

Method	**HC**	**CC**	**CRC**	**SOR**
Raw	2.141	0.6151	0.714	0.385
EW	2.299	0.6612	0.767	0.339
openSSS	**2.490**	0.830	**0.830**	0.275
NN_hard	2.424	**0.733**	0.808	**0.267**
NN_soft010	2.414	0.731	0.804	0.269
NN_soft025	2.398	0.725	0.799	0.275
NN_soft050	2.342	0.702	0.780	0.298
True	**2.600**	**0.779**	**0.867**	**0.220**

Regarding the different weighting schemes, the results indicate that hard thresholding consistently provided better performance than the soft weighting variants. As the penalty weight increased from 0.10 to 0.50, a progressive decrease in CRC and an increase in SOR were observed. This trend suggests that, given the high specificity and sensitivity of the neural network classifier, the sub-optimal events rejected by the hard threshold contain minimal true signal. Consequently,

applying a soft weight reintroduces residual scatter without significantly improving the signal-to-noise ratio, making the binary rejection (hard threshold) the optimal operation point for this specific scenario.

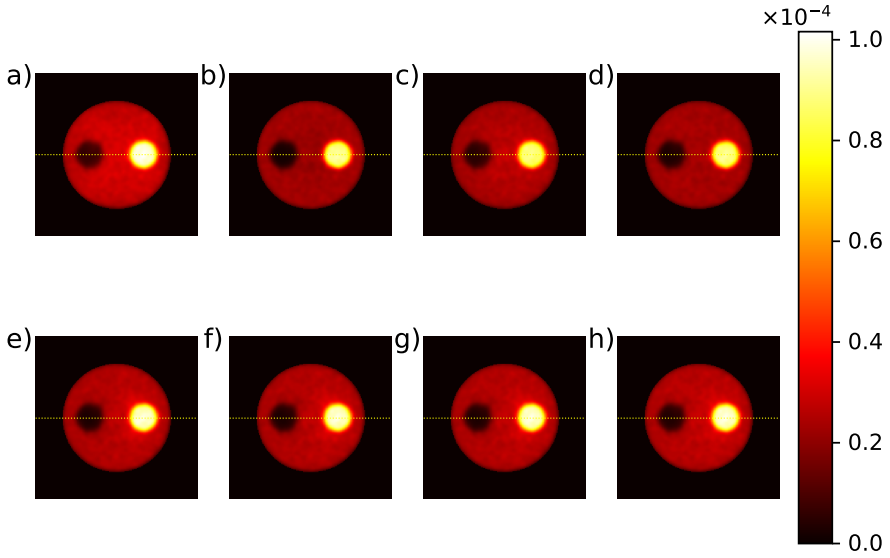


Figure 2: Reconstructed images for the proposed algorithms. (a) Raw image (b) Ideal (True) image (c) EW (d) openSSS (e) NN_hard (f) NN_soft010 (g) NN_soft025 (h) NN_soft050

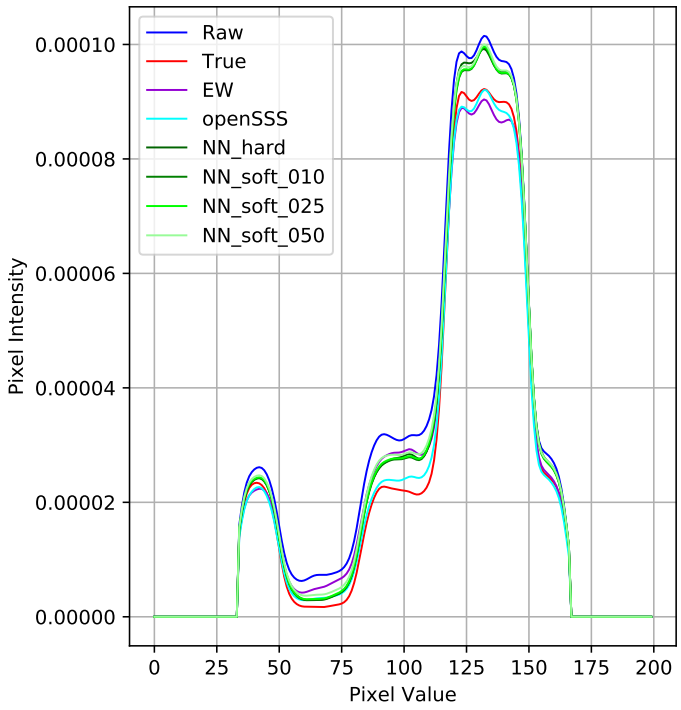


Figure 3: Central slice profile of reconstructed images for the proposed algorithms.

These quantitative findings are visually confirmed in the reconstructed images (Figure 2) and the line profiles (Figure 3). The Raw reconstruction shows elevated activity in the cold insert (Figure 2a). In contrast, the NN reconstruction (Figures 2e - 2h) exhibits sharp boundaries and a baseline level in the cold region that closely tracks the True ground truth (Figure 2b). The line profiles in Figure 3 further illustrate this behavior.

4 Conclusions

This work presented a data-driven scatter correction framework for PET based on a weighted list-mode MLEM (W-MLEM) reconstruction. Results demonstrate that the W-MLEM framework effectively suppresses scatter contributions while maintaining image quality and statistical stability. By introducing event-wise statistical weights derived from a neural network classifier, the proposed method constitutes a flexible and promising alternative to traditional scatter correction techniques, particularly in the context of data-driven PET reconstruction.

A key finding of this study is that hard thresholding emerged as the most effective weighting strategy, consistently outperforming the soft weighting variants in terms of contrast recovery and background suppression. While soft weighting is theoretically intended to mitigate sensitivity loss by recovering misclassified true events, our results indicate that the high precision achieved by the neural network makes binary rejection optimal. The robust performance of the NN_hard method is a direct consequence of the classifier's high sensitivity. With only 6% of true events being incorrectly flagged as scatter (FNR), the statistical loss incurred by hard rejection is minimal. Conversely, the 25% false positive rate means that some scatter is still accepted with w=1; however, our results show that increasing w for the rejected events (soft weighting) only exacerbates background contamination without significantly recovering true signal, as the rejected pool is predominantly composed of correctly identified scatter. Nevertheless, we hypothesize that soft weighting could regain its advantage in low-statistics scenarios, such as short acquisition times, or in cases where the classifier exhibits higher false negative rates. These results highlight the potential of event-wise statistical weighting to provide a personalized and efficient solution for advanced PET image reconstruction.

Acknowledgments

All authors declare that they have no known conflicts of interest in terms of competing financial interests or personal relationships that could have an influence or are relevant to the work reported. This work was supported by PTA2022-022171-I funded by MCIU/AEI/10.13039/501100011033 and 'ESF Investing in your future'.

References

[1] M. Oehler and T. M. Buzug, *The λ-MLEM Algorithm: An Iterative Reconstruction Technique for Metal Artifact Reduction in CT Images*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 42–47.

[2] J. Prats-Climent, F. Sánchez, H. Espinos-Morato, A. González, and M. Rodríguez-Álvarez, "Scatter and random correction of pet list-mode data using machine learning approaches," in *Proceedings of the Conference on Modelling for Engineering and Human Behaviour 2023 (MME&HB 2023)*, València, Spain, 2023, pp. 72–77.

[3] L. Shao, R. Freidelfer, and J. Karp, "Triple energy window scatter correction technique in pet," *IEEE Transactions On Medical Imaging*, vol. 13, no. 4, pp. 641–648, 1994.

[4] R. José Santo, A. Salomon, H. W. A. M. de Jong, S. Stute, T. Merlin, and C. Beijst, "openSSS: An open-source implementation of scatter estimation for 3D TOF-PET," *EJNMMI Physics*, vol. 12, no. 17, pp. 1–15, 2025.

[5] L. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE Transactions On Medical Imaging*, vol. 1, no. 2, pp. 113–122, 1982.

[6] S. Jan et al., "Gate: A simulation toolkit for pet and spect.," *Phys Med Biol*, vol. 49, no. 19, pp. 4543–4561, 2004.

[7] T. Merlin et al., "Castor: A generic data organization and processing code framework for multi-modal and multi-dimensional tomographic reconstruction," *Phys Med Biol*, vol. 63, no. 185005, 2018.

Parameter Estimation of Sub-Solar-Mass Objects with Mini Extreme Mass-Ratio Inspirals

Qiming Song[a,b], Huaike Guo[a,b], and Ju Chen[a,b]

[a] International Centre for Theoretical Physics Asia-Pacific (ICTP-AP), University of Chinese Academy of Sciences (UCAS), Beijing, China.

[b] Taiji Laboratory for Gravitational Wave Universe (Beijing/Hangzhou), University of Chinese Academy of Sciences (UCAS), Beijing, China.

1 Introduction

The direct observation of gravitational waves (GWs) has opened a new way to probe compact objects whose masses, formation channels, or internal structure may be difficult to access electromagnetically. The LIGO–Virgo–KAGRA (LVK) collaboration has reported more than one hundred compact-binary candidates, including binary black holes, binary neutron stars, and neutron-star–black-hole systems [1, 2, 3]. Most of these sources are compatible with conventional stellar evolution. However, compact objects outside the usual stellar-remnant mass ranges remain especially interesting, since they could point to exotic compact objects (ECOs), including primordial black holes (PBHs), boson stars, or other nonstandard remnants.

Sub-solar-mass compact objects are particularly clean targets for new physics. Ordinary stellar evolution does not produce black holes with masses below the solar scale, and white dwarfs are not compact enough to form ultracompact binaries radiating in the LIGO band. Therefore, a robust GW detection of a sub-solar-mass ultracompact object would provide strong evidence for a PBH-like or ECO origin. Searches for sub-solar-mass binaries have been carried out with data from the first, second, and third observing runs of Advanced LIGO and Advanced Virgo [4, 5, 6, 7]. These searches have covered progressively broader mass ratios, spin assumptions, and phenomenological models, but no statistically significant detection has been reported so far. Complementary probes have also been proposed, including continuous-wave searches for planetary-mass PBHs [8] and microlensing surveys, but the accessible parameter space remains far from exhausted.

Extreme mass-ratio inspirals (EMRIs) are another important class of GW sources. In the standard sense, an EMRI consists of a stellar-mass compact object orbiting a massive black hole with mass $M \sim 10^4$–$10^7\,M_\odot$, and is a primary target for space-borne detectors such as LISA [9, 10, 11]. The broader EMRI family also includes intermediate mass-ratio inspirals, early-stage EMRIs, extremely large mass-ratio inspirals, and binary EMRIs, each highlighting a different unequal-mass regime. A mini extreme mass-ratio inspiral (mini-EMRI) can be viewed as a scaled-down analogue: a stellar-mass black hole orbited by a sub-solar-mass ECO [12, 13]. Because the total mass is much smaller than that of a standard EMRI, the late inspiral can enter the frequency band of ground-based detectors.

This work studies whether mini-EMRIs can be used to identify sub-solar-mass compact objects through parameter estimation. We focus on nonspinning, nearly circular systems and ask how accurately the secondary mass can be measured before the innermost stable circular orbit (ISCO). We

compare a relativistically corrected circular-orbit waveform with a waveform generated by FastEMRIWaveforms, and use the Fisher information matrix (FIM) to forecast the uncertainty of the secondary mass. Since the sources considered here are at Galactic-to-nearby-galaxy distances, cosmological redshift is neglected and source-frame masses are treated as detector-frame masses.

2 Waveform Model and Detector Response

We consider a mini-EMRI with primary mass m_1, secondary mass m_2, luminosity distance D_L, inclination angle ι, and coalescence phase Φ_0. Three waveform descriptions are useful for understanding the problem.

The Newtonian quadrupole waveform gives a simple leading-order description of a circular inspiral. Its two polarizations may be written as [14]

$$h_+(t) = \frac{1}{D_L}\left(\frac{G\mathcal{M}_c}{c^2}\right)^{5/4}\left(\frac{5}{c\tau}\right)^{1/4}\left(\frac{1+\cos^2\iota}{2}\right)\cos\Phi(\tau), \tag{1}$$

$$h_\times(t) = \frac{1}{D_L}\left(\frac{G\mathcal{M}_c}{c^2}\right)^{5/4}\left(\frac{5}{c\tau}\right)^{1/4}\cos\iota\,\sin\Phi(\tau), \tag{2}$$

where $\mathcal{M}_c = (m_1 m_2)^{3/5}/(m_1+m_2)^{1/5}$ is the chirp mass, $\tau = t_{\rm coal} - t$, and

$$\Phi(\tau) = -2\left(\frac{5G\mathcal{M}_c}{c^3}\right)^{-5/8}\tau^{5/8} + \Phi_0. \tag{3}$$

This model is valuable as a reference, but it is not suitable for the main Fisher analysis: at leading Newtonian order the waveform depends on the component masses only through $\mathcal{M}_c$, so m_1 and m_2 are strongly degenerate.

To break this degeneracy, we adopt a relativistically corrected waveform for circular orbits. Following the Teukolsky–Sasaki–Nakamura based treatment of Ref. [15], the dominant quadrupolar harmonic can be written schematically as

$$h_{2,+} = \frac{4}{D_L}\left(\frac{G\mathcal{M}_c}{c^2}\right)^{5/3}\left(\frac{\pi f}{c}\right)^{2/3}\left(\frac{1+\cos^2\iota}{2}\right)\mathcal{H}_{o,2}\cos\Phi, \tag{4}$$

$$h_{2,\times} = \frac{4}{D_L}\left(\frac{G\mathcal{M}_c}{c^2}\right)^{5/3}\left(\frac{\pi f}{c}\right)^{2/3}\cos\iota\,\mathcal{H}_{o,2}\sin\Phi, \tag{5}$$

where $\mathcal{H}_{o,2}$ encodes relativistic corrections to the amplitude and depends on the orbital configuration. In the weak-field limit $\mathcal{H}_{o,2} \to 1$, the expression reduces to the quadrupole result. In practice, the relativistic correction and the associated inspiral evolution introduce information beyond the chirp mass alone.

As a second and more complete relativistic model, we use FastEMRIWaveforms [16, 17, 18]. For Schwarzschild eccentric inspirals, the complex strain can be expressed as a harmonic sum

$$h = h_+ - ih_\times = \frac{\mu}{D_L}\sum_{lmn} A_{lmn}(t)\,{}_{-2}Y_{lm}(\theta,\phi)e^{-i\Phi_{mn}(t)}, \tag{6}$$

where $\mu = m_1 m_2/(m_1+m_2)$, A_{lmn} are Teukolsky-based mode amplitudes, and Φ_{mn} are accumulated orbital phases. We use the `FastSchwarzschildEccentricFlux` model in the nearly circular limit. The inspiral trajectory is obtained by evolving the orbital parameters using precomputed fluxes, while the waveform amplitudes are generated efficiently by reduced-order and neural-network interpolation. This model is therefore well suited for large parameter scans and Fisher-matrix forecasts.

The strain measured by a ground-based detector is the projection of the two polarizations onto the detector response,

$$h_D(t) = F_+(t;\alpha,\delta,\psi)h_+(t) + F_\times(t;\alpha,\delta,\psi)h_\times(t), \tag{7}$$

where the antenna-pattern functions depend on sky position, polarization angle, detector orientation, and Earth's rotation [19]. The numerical results below use a single-detector response; extension to the H1–L1 network is a natural next step.

3 Signal-to-Noise Ratio and Fisher Parameter Estimation

For stationary Gaussian detector noise, the noise-weighted inner product between two real waveforms $A(t)$ and $B(t)$ is

$$(A|B) = 4\,\mathrm{Re}\int_0^\infty \frac{\tilde{A}^*(f)\tilde{B}(f)}{S_n(f)}\,df, \tag{8}$$

where $S_n(f)$ is the one-sided noise power spectral density. The signal-to-noise ratio is then

$$\mathrm{SNR}^2 = (h|h). \tag{9}$$

The SNR indicates whether the signal is likely to be detectable and also controls, at leading order, the reliability of the Fisher approximation.

The detector data are modeled as

$$d(t) = h(t;\boldsymbol{\theta}) + n(t), \tag{10}$$

where $\boldsymbol{\theta}$ denotes the source parameters and $n(t)$ is noise. Under the same Gaussian-noise assumption, the likelihood is

$$\mathcal{L}(d|\boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2}(d-h(\boldsymbol{\theta})|d-h(\boldsymbol{\theta}))\right]. \tag{11}$$

In a full Bayesian analysis this likelihood is multiplied by a prior to obtain the posterior distribution. In this preliminary study, we instead use the Fisher information matrix as a fast forecasting tool and assume that priors are broad over the local region constrained by the likelihood.

The Fisher matrix is defined by [14]

$$\Gamma_{ij} = \left(\frac{\partial h}{\partial\theta_i}\middle|\frac{\partial h}{\partial\theta_j}\right). \tag{12}$$

When the SNR is high and the waveform is approximately linear over the posterior support, the covariance matrix is approximated by

$$\Sigma_{ij} \simeq (\Gamma^{-1})_{ij}, \qquad \Delta\theta_i \simeq \sqrt{\Sigma_{ii}}. \tag{13}$$

The waveform derivatives are computed numerically with a second-order central finite difference,

$$\frac{\partial h}{\partial\theta_i} \simeq \frac{h(\boldsymbol{\theta}+\Delta_i\mathbf{e}_i) - h(\boldsymbol{\theta}-\Delta_i\mathbf{e}_i)}{2\Delta_i}. \tag{14}$$

In the scans below we include at least the correlated parameters

$$\boldsymbol{\theta} = \{m_1, m_2, D_L, \Phi_0\}, \tag{15}$$

even though the main quantity of interest is the relative uncertainty $\Delta m_2/m_2$.

The Fisher matrix also diagnoses degeneracies. If two parameters change the waveform in nearly the same direction, Γ_{ij} develops a nearly singular direction and the inferred uncertainty grows. This is precisely why the Newtonian waveform is inadequate for the present goal: it is sensitive mainly to the chirp mass, not to m_1 and m_2 separately. Finally, the Cramer–Rao inequality implies that Γ^{-1} gives a lower bound on the covariance of any unbiased estimator. The Fisher results should therefore be interpreted as optimistic forecasts rather than full parameter-estimation uncertainties.

4 Results and Discussion

We scan mini-EMRI systems with

$$m_1 \in [5, 50]\, M_\odot, \qquad m_2 \in [10^{-5}, 10^{-3}]\, M_\odot, \tag{16}$$

using the final 1000 s of inspiral before the ISCO. The primary-mass interval is chosen to keep the late inspiral broadly inside the sensitive band of ground-based detectors and to avoid the conventional lower and upper mass-gap regions [20, 21, 22, 23, 24, 25]. In this setup, the characteristic GW frequencies span roughly 64–879 Hz across the mass range, overlapping well with the LIGO sensitivity band.

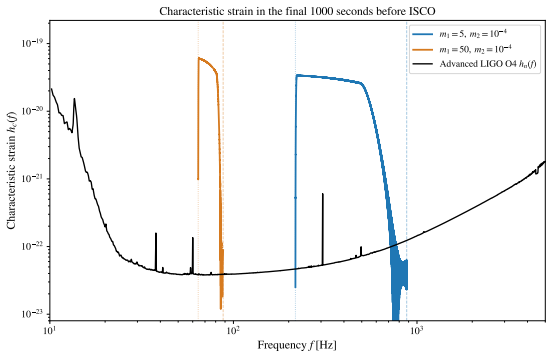
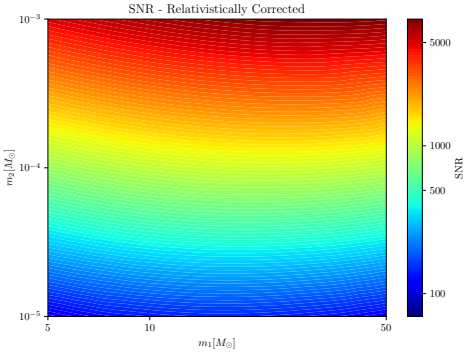
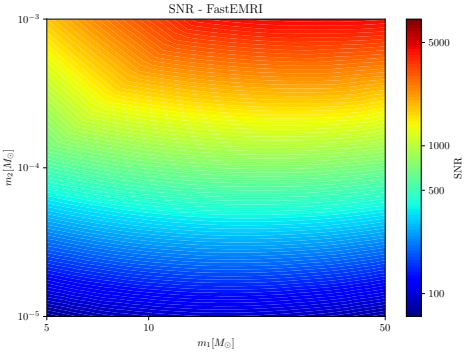


Figure 1: The characteristic strain of representative mini-EMRI signals during the final 1000 s before ISCO, compared with the LIGO design sensitivity curve.

The first result is conceptual but important: the Newtonian quadrupole waveform cannot be used to estimate the two masses independently. Since the leading waveform is parameterized by $\mathcal{M}_c$, the Fisher matrix contains an approximate mass degeneracy. The relativistically corrected waveform and FastEMRIWaveforms both separate the component-mass dependence more effectively, and are therefore adopted for the main forecasts. In the remainder of this section we emphasize the FastEMRIWaveforms results, while using the relativistically corrected waveform as a comparison model.



(a) Relativistically corrected waveform.

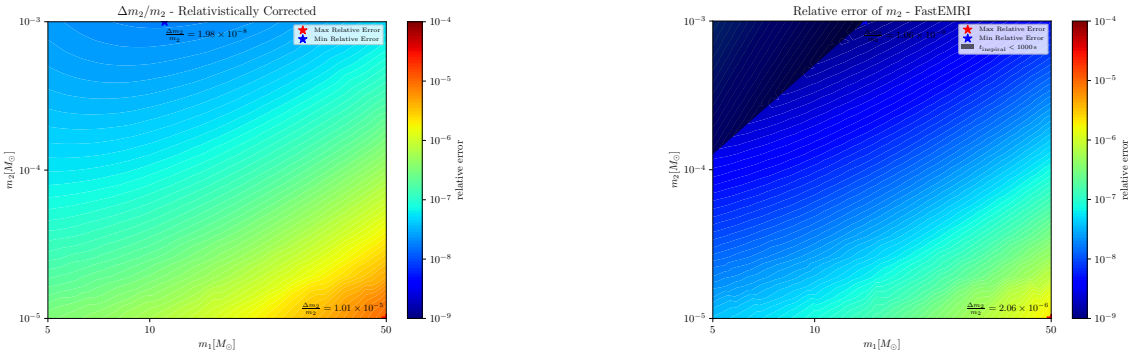


(b) FastEMRIWaveforms waveform.

Figure 2: SNR scans over the mini-EMRI mass parameter space at the reference distance.

For a reference distance of $D_L = 1$ Mpc, the SNR is high over a large part of the scanned parameter space, supporting the use of the Fisher approximation. At fixed m_2, the SNR depends only weakly on m_1: increasing the primary mass raises the amplitude scale but shifts the signal to lower frequencies

and reduces the effective bandwidth. These effects partly compensate each other. At fixed m_1, the SNR increases with m_2, as expected from the stronger radiation sourced by the larger secondary mass.



(a) Relativistically corrected waveform.

(b) FastEMRIWaveforms waveform.

Figure 3: The relative Fisher error $\Delta m_2/m_2$ over the mass parameter space.

The corresponding Fisher uncertainties show that the secondary mass can be measured very accurately in the optimistic high-SNR regime. For most of the scanned parameter space, $\Delta m_2/m_2$ is below 10^{-5}. The uncertainty grows for heavier primaries because the final 1000 s contain fewer useful inspiral cycles and a narrower bandwidth. It decreases for larger secondary masses, following the SNR trend. The FastEMRIWaveforms waveform generally gives smaller relative errors than the relativistically corrected waveform, consistent with its richer relativistic harmonic content and more accurate inspiral dynamics. Some small-m_1, large-m_2 systems are excluded from the FastEMRIWaveforms scan because the allowed semi-latus-rectum range of the current implementation does not cover a full 1000 s segment before ISCO.

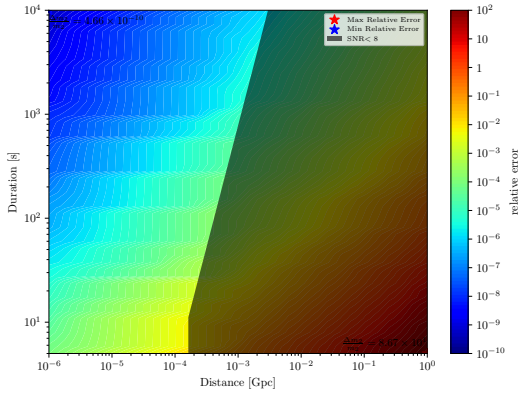


Figure 4: The dependence of $\Delta m_2/m_2$ on luminosity distance and observation duration for a representative system with $m_1 = 20\,M_\odot$ and $m_2 = 10^{-4}\,M_\odot$. The dark region can be used for systems below a chosen detectability threshold, e.g. SNR < 8.

We also examine how the secondary-mass uncertainty varies with distance and observation duration for the representative system $(m_1, m_2) = (20\,M_\odot, 10^{-4}\,M_\odot)$. The trend is simple: $\Delta m_2/m_2$ increases with distance because the signal becomes weaker, and decreases with observation duration because more inspiral cycles are accumulated. In practice, the useful region must also satisfy a detection threshold. We use SNR > 8 as a reference criterion and assume, optimistically, that the Fisher approximation is applicable throughout the detectable region.

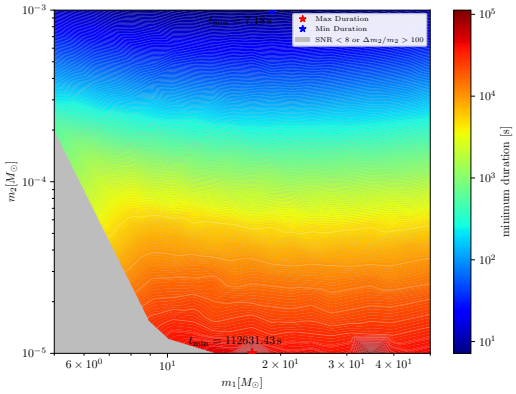


Figure 5: The minimum observation time required to identify the secondary as sub-solar-mass under the criteria SNR > 8 and $\Delta m_2/m_2 \leq 100$.

Finally, we estimate the minimum duration needed to classify the secondary as a sub-solar-mass object. Since the goal is classification rather than precision astrophysics, we adopt the loose requirement $\Delta m_2/m_2 \leq 100$ together with SNR > 8. At $D_L = 1\,\mathrm{Mpc}$, the scan indicates that detectability is typically the limiting condition: once the signal is detectable, the secondary mass is constrained far below the classification threshold. For example, systems with $m_2 = 3 \times 10^{-4}\,M_\odot$ and $m_1 = 5\,M_\odot$ or $50\,M_\odot$ reach SNR $\simeq 8$ with observation times of order a few hundred seconds, while the Fisher errors remain much smaller than the adopted threshold. This suggests that mini-EMRIs could provide a viable route to identifying sub-solar-mass compact objects if such systems occur within the accessible volume.

5 Conclusions

We have presented a Fisher-matrix forecast for parameter estimation of mini extreme mass-ratio inspirals, motivated by the search for sub-solar-mass compact objects. Such objects are difficult to explain through ordinary stellar evolution and would be natural candidates for PBHs or other ECOs. Mini-EMRIs provide a complementary probe because a stellar-mass black hole with a sub-solar-mass companion can radiate in the frequency band of ground-based detectors during the final inspiral.

The waveform model is central to this problem. The Newtonian quadrupole waveform is useful for intuition, but its dependence on the chirp mass causes a degeneracy between m_1 and m_2 in the Fisher matrix. Relativistically corrected waveforms and FastEMRIWaveforms waveforms break this degeneracy more effectively and allow the secondary mass to be estimated separately. In the scanned parameter space, $m_1 \in [5, 50]\,M_\odot$ and $m_2 \in [10^{-5}, 10^{-3}]\,M_\odot$, the final 1000 s before ISCO can yield high SNRs at a reference distance of 1 Mpc, and the inferred Fisher errors for m_2 are small over most of the detectable region.

These results should be interpreted as optimistic forecasts. The inverse Fisher matrix gives a Cramer–Rao lower bound, and the analysis assumes high SNR, local waveform linearity, broad priors, a single-detector response, and simplified nonspinning circular dynamics. Future work should extend the calculation to the H1–L1 network, include more realistic orbital eccentricity and spin effects, and validate the Fisher estimates with full Bayesian parameter inference. Even with these caveats, the present study indicates that mini-EMRIs are a promising channel for testing the existence of sub-solar-mass compact objects with ground-based GW detectors.

6 Acknowledgments

We gratefully acknowledge the helpful discussions with Zixuan Wang. The support provided by China Scholarship Council (CSC) during a visit to Valencia is acknowledged.

References

[1] B. P. Abbott et al. GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs. *Phys. Rev. X*, 9:031040, 2019. doi: 10.1103/PhysRevX.9.031040.

[2] R. Abbott et al. GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo during the First Half of the Third Observing Run. *Phys. Rev. X*, 11:021053, 2021. doi: 10.1103/PhysRevX.11.021053.

[3] R. Abbott et al. GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo during the Second Part of the Third Observing Run. *Phys. Rev. X*, 13:041039, 2023. doi: 10.1103/PhysRevX.13.041039.

[4] B. P. Abbott et al. Search for Subsolar Mass Ultracompact Binaries in Advanced LIGO's First Observing Run. *Phys. Rev. Lett.*, 121:231103, 2018. doi: 10.1103/PhysRevLett.121.231103.

[5] B. P. Abbott et al. Search for Subsolar-Mass Ultracompact Binaries in Advanced LIGO's Second Observing Run. *Phys. Rev. Lett.*, 123:161102, 2019. doi: 10.1103/PhysRevLett.123.161102.

[6] R. Abbott et al. Search for Subsolar-Mass Binaries in the First Half of Advanced LIGO's and Advanced Virgo's Third Observing Run. *Phys. Rev. Lett.*, 129:061104, 2022. doi: 10.1103/PhysRevLett.129.061104.

[7] R. Abbott et al. Search for Subsolar-Mass Black Hole Binaries in the Second Part of Advanced LIGO's and Advanced Virgo's Third Observing Run. *Mon. Not. Roy. Astron. Soc.*, 524:5984–5992, 2023. doi: 10.1093/mnras/stad1887.

[8] Andrew L. Miller, Nancy Aggarwal, Sebastien Clesse, and Federico De Lillo. Probing Planetary-Mass Primordial Black Holes with Continuous Gravitational Waves. *Phys. Rev. D*, 102:123009, 2020. doi: 10.1103/PhysRevD.102.123009.

[9] Pau Amaro-Seoane et al. Intermediate and Extreme Mass-Ratio Inspirals: Astrophysics, Science Applications and Detection Using LISA. *Class. Quant. Grav.*, 24:R113–R169, 2007. doi: 10.1088/0264-9381/24/17/R01.

[10] Stanislav Babak et al. Science with the Space-Based Interferometer LISA. V. Extreme Mass-Ratio Inspirals. *Phys. Rev. D*, 95:103012, 2017. doi: 10.1103/PhysRevD.95.103012.

[11] Pau Amaro-Seoane et al. Laser Interferometer Space Antenna. *arXiv preprint arXiv:1702.00786*, 2017.

[12] Hai-Tian Guo et al. Searching for Subsolar-Mass Compact Objects with Mini Extreme Mass-Ratio Inspirals. *arXiv preprint*, 2022. Bibliographic details to be verified.

[13] Andrew L. Miller et al. Mini Extreme Mass-Ratio Inspirals as Probes of Subsolar-Mass Compact Objects. *arXiv preprint*, 2024. Bibliographic details to be verified.

[14] Michele Maggiore. *Gravitational Waves. Volume 1: Theory and Experiments*. Oxford University Press, 2007. ISBN 9780198570745.

[15] Lee Samuel Finn and Kip S. Thorne. Detection, Measurement, and Gravitational Radiation from Inspiraling Compact Binaries. *Phys. Rev. D*, 62:124021, 2000. doi: 10.1103/PhysRevD.62.124021.

[16] Alvin J. K. Chua, Michael L. Katz, Niels Warburton, and Scott A. Hughes. Rapid Generation of Fully Relativistic Extreme-Mass-Ratio-Inspiral Waveform Templates for LISA Data Analysis. *Phys. Rev. Lett.*, 126:051102, 2021. doi: 10.1103/PhysRevLett.126.051102.

[17] Michael L. Katz, Alvin J. K. Chua, Lorenzo Speri, Niels Warburton, and Scott A. Hughes. Fast Extreme-Mass-Ratio-Inspiral Waveforms: New Tools for Millihertz Gravitational-Wave Data Analysis. *Phys. Rev. D*, 104:064047, 2021. doi: 10.1103/PhysRevD.104.064047.

[18] Charlotte Chapman-Bird et al. FastEMRIWaveforms. `https://doi.org/10.5281/zenodo.15630565`, 2025.

[19] Piotr Jaranowski, Andrzej Krolak, and Bernard F. Schutz. Data Analysis of Gravitational-Wave Signals from Spinning Neutron Stars. *Phys. Rev. D*, 58:063001, 1998. doi: 10.1103/PhysRevD.58.063001.

[20] Feryal Ozel, Dimitrios Psaltis, Ramesh Narayan, and Jeffrey E. McClintock. The Black Hole Mass Distribution in the Galaxy. *Astrophys. J.*, 725:1918–1927, 2010. doi: 10.1088/0004-637X/725/2/1918.

[21] Will M. Farr, Niharika Sravan, Andrew Cantrell, Laura Kreidberg, Charles D. Bailyn, Ilya Mandel, and Vicky Kalogera. The Mass Distribution of Stellar-Mass Black Holes. *Astrophys. J.*, 741:103, 2011. doi: 10.1088/0004-637X/741/2/103.

[22] Laura Kreidberg, Charles D. Bailyn, Will M. Farr, and Vicky Kalogera. Mass Measurements of Black Holes in X-Ray Transients: Is There a Mass Gap? *Astrophys. J.*, 757:36, 2012. doi: 10.1088/0004-637X/757/1/36.

[23] Maya Fishbach and Daniel E. Holz. Where Are LIGO's Big Black Holes? *Astrophys. J. Lett.*, 851:L25, 2017. doi: 10.3847/2041-8213/aa9bf6.

[24] R. Abbott et al. GW190814: Gravitational Waves from the Coalescence of a 23 Solar Mass Black Hole with a 2.6 Solar Mass Compact Object. *Astrophys. J. Lett.*, 896:L44, 2020. doi: 10.3847/2041-8213/ab960f.

[25] R. Abbott et al. GW190521: A Binary Black Hole Merger with a Total Mass of 150 Solar Masses. *Phys. Rev. Lett.*, 125:101102, 2020. doi: 10.1103/PhysRevLett.125.101102.

Unit Killing Initial Data

Salvador Mengual Sendra[♭1] and Igor Khavkine[♮]

(♭) Departament d'Astronomia i Astrofísica, Universitat de València
Av. Vicent Andrés Estellés, 19 46100 Burjassot, València, Spain.

(♮) Institute of Mathematics of the Czech Academy of Sciences,
Žitná 25, 110 00 Praha 1, Czech Republic.

1 Introduction

Killing initial data (KID) equations consist of a set of partial differential equations (PDEs) intrisic to a spacelike hypersurface S in a Einstein vacuum Lorentzian manifold (M, g), whose special property is that their solutions are in bijection with the Killing vector fields in the domain of dependence of S [1]. In particular, if S is a Cauchy surface for (M, g), solving the KID equations on S allows one to check whether the metric initial data will evolve into a spacetime with Killing vectors, without first solving the Einstein equations. Moreover, due to their geometric character, the KID equations can be formulated on an arbitrary choice of Cauchy surface and coordinates thereon.

One application of the KID equations is to provide a geometric characterization of initial data of the rotating Kerr black hole intrinsic to an arbitrary Cauchy surface [2]. That characterization combines a so-called *IDEAL characterization* of the Kerr geometry [3] with the KID equations and a way of geometrically expressing Kerr's stationary Killing vector directly from the curvature. It would be interesting to extend similar intrinsic initial data characterizations to other physically interesting solutions of Einstein equations with fluid matter. While these solutions do possess Killing vectors, they can only be obtained from the curvature in unit normalized form. Already back in the well-known reference [4, §2.2(vi)], Ehlers noticed that a timelike unit vector that is a rescaling of a Killing vector can be identified by a related set of equations and referred to it as *isometric*.[2] For reference, let us call these the *unit Killing equations*. A natural question is whether the initial data characterization strategy from [2] can be generalized to the situation where only a unit normalized Killing vector is available. Part of a positive answer would be identifying the unit Killing initial data (uKID) equations.

In this work, we precisely derive the uKID equations for vacuum with cosmological constant, following the *propagation identity* framework outlined in [5, 6]. Thus, we first need to recall what we mean by a *propagation identity* for a differential equation $E[\phi] = 0$, with the Killing equation and its unit variant being examples of our primary concern. However, for our current purposes, we must generalize the key lemma [5, Lem.2.1] that allows us to connect such a propagation identity with E-initial data conditions. The reason is that we will have to rely on a more complicated notion

[1] salvador.mengual@uv.es

[2] Strictly speaking, Ehlers referred to the flow of such a vector field as an *isometric flow*. We might as well refer to the vector field itself as isometric.

of hyperbolic PDE system and corresponding well-posed initial value problem, where different equations in the system may have different or unexpected differential orders (*mixed-order systems*).

2 Propagation identity

The *principal symbol* of a linear differential operator $L[w]$ on $\mathbb{R}^n$ of order k consists of all the terms containing derivatives of order k and, replacing each occurrence of a partial derivative ∂_i by the indeterminate p_i, is denoted by $L^0(p)$ [7]. Replacing $\mathbb{R}^n$ by a manifold M, and a matrix valued differential operator by a smooth linear differential operator $L\colon \Gamma(W) \to \Gamma(v)$ that maps between sections of vector bundles $V \to M$ and $W \to M$, the indeterminates p_i are interpreted as coordinates on the fibers of the cotangent bundle T^*M, and then principal symbol $L^0(p)$ becomes a fiber-wise homogeneous polynomial function on T^*M valued in the linear vector bundle morphisms $\mathrm{Hom}(W, V)$. Below, we will discuss block decompositions of L and $L^0(p)$, which on a manifold must be interpreted with respect to some direct decompositions $V = \bigoplus_i V_i$, $W = \bigoplus_j W_j$ of the vector bundles.

Definition 1 ([7, §§II.4.1, II.7.2.1]) *Fix a time coordinate t and a corresponding derivative ∂_t.*

(a) A linear differential operator $y = L[w]$ is called (uniformly) Agmon-Douglis-Nirenberg (ADN) strictly hyperbolic *if the the polynomial $\det L^0(p)$ is (i) monic (has unit coefficient) with respect to the highest power of p_t, (ii) all of its roots with respect to p_t (all other variables are fixed) are real, and (iii) these roots separated from each other uniformly by at least $\delta\|p_\perp\|$, for some $\delta > 0$, over all spatial momenta $p_\perp$ and spacetime points.*

(b) Consider a linear differential operator $y = L[w]$, with the following block structure under the decomposition $y = (y_1, \dots, y_m)$, $w = (w_1, \dots, w_m)$:

$$L[w] = \begin{bmatrix} L_{11} & \cdots & L_{1m} \\ \vdots & & \vdots \\ L_{m1} & \cdots & L_{mm} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix}. \tag{1}$$

We say that L is Agmon-Douglis-Nirenberg (ADN) strictly hyperbolic *when there exist integers* weights $t_j, s_j \geq 0$, $j = 1, \dots, m$, *such that L_{ij} is a differential operator of order $(t_j - s_i)$ and the* weighted principal symbol *has the block diagonal structure*

$$L^0(p) := \begin{bmatrix} L^0_{11}(p) & \cdots & L^0_{1m}(p) \\ \vdots & & \vdots \\ L^0_{m1}(p) & \cdots & L^0_{mm}(p) \end{bmatrix} = \begin{bmatrix} L^0_{11}(p) & & \\ & \ddots & \\ & & L^0_{mm}(p) \end{bmatrix}, \tag{2}$$

with $L^0_{ij}(p)$ being the principal symbol of L_{ij} as an operator of order $(t_j - s_i)$, where each $L^0_{ii}(p)$ is strictly hyperbolic.

The following well-posedness result can be stated for the initial value problem of ADN strictly hyperbolic systems.

Definition 2 ([7, §II.7.2.4]) *Consider an inhomogeneous PDE system $L[w] = y$ with block structure $L_{ij}[w]$ and ADN weights t_j, s_i, as well as a time coordinate t. The* mixed order initial data *(at $t = 0$) consists of smooth functions $(w_j^{(l)})_{j=1,\dots,m}^{l=0,\dots,t_j-1}$ on the surface $t = 0$. The corresponding* mixed-order initial value problem *consists of finding a smooth w that satisfies $L[w] = 0$ for all time and (at $t = 0$) the conditions*

$$\begin{gathered} \partial_t^l w_j|_{t=0} = w_j^{(l)} \ (l = 0, \dots, t_j - 1) \\ \text{and} \quad \partial_t^l \sum_{j=1}^m L_{ij}[w_j]|_{t=0} = \partial_t^l y_i|_{t=0} \ (l = 0, \dots, s_i - 1), \end{gathered} \tag{3}$$

where the second equality condition is empty whenever $s_i = 0$, and otherwise its left-hand side is expressible, by formal differentiation, purely in terms of the $(w_j^{(l)})$ data. The above initial value problem is well-posed *when there exists for all time a unique smooth w satisfying $L[w] = 0$ and the conditions* (3)*.*

Note that, when $y = 0$, due to the linearity of the equation $L[w] = 0$, if the initial data $(w_j^{(l)})_{j=1,\dots,m}^{l=0,\dots,t_j-1} = 0$, the unique solution is $w = 0$.

Proposition 3 ([7, §II.7.2.4]) *Consider a globally uniformly ADN strictly hyperbolic system $L[w] = 0$ with smooth coefficients, and a temporal function t, such that $t = 0$ is a Cauchy surface. Then its initial value problem (as in Definition 2) is well-posed.*

The following generalizes Lemma 2.1 of [5]:

Lemma 4 *Let $E[\phi] = 0$ be a PDE (system) on a spacetime (M, g), defined on some (possibly multicomponent) field ϕ. Suppose that there exist ADN mixed-order equations $P[\psi] = 0$, $Q[\phi] = 0$ (of respective weights $t_j(P)$, $s_i(P)$ and $t_j(Q)$, $s_i(Q)$) with a well-posed initial value problem (as in Definition 2), where the differential operators P and Q satisfy the identities*

$$P[E[\phi]] = \sigma[Q[\phi]] \quad \textit{and} \quad Q[\phi] = \rho[E[\phi]], \tag{4}$$

for some linear differential operator σ and ρ. Then, given a Cauchy surface $\Sigma \subset M$ with unit timelike normal n^a, the unique solution $\phi = (\phi_j)$ of $Q[\phi] = 0$ with initial data $\phi_j|_\Sigma = \phi_j^{(0)}, \dots, \nabla_n^{t_j(Q)-1}\phi|_\Sigma = \phi_j^{t_j(Q)-1}$ satisfies the equation $E[\phi] = 0$ provided the initial data $\psi_j|_\Sigma = 0, \dots, \nabla_n^{t_j(Q)-1}\psi_j|_\Sigma = 0$, for $\psi = E[\phi]$, vanish.

In addition, there exists a purely spatial linear PDE on Σ, $E^\Sigma[\phi_j^{(0)}, \dots, \phi_j^{t_j(Q)-1}] = 0$ such that the conditions $Q[\phi] = 0$ and $E^\Sigma[\phi_j|_\Sigma, \dots, \nabla_n^{t_j(Q)-1}\phi_j|_\Sigma] = 0$ imply the vanishing of the initial data $\psi_j|_\Sigma = 0, \dots, \nabla_n^{t_j(P)-1}\psi_j|_\Sigma = 0$ for $\psi = E[\phi]$.

3 Killing and unit isometric vectors

Consider the Killing equation on a Lorentzian manifold:

$$K_{ab}[v] := \nabla_a v_b + \nabla_b v_a = 0 \quad (E[\phi] = 0). \tag{5}$$

The corresponding propagation equations are

$$\Box v_a + R_a{}^b v_b = 0 \quad (Q[\phi] = 0), \tag{6}$$

$$\Box h_{ab} - 2R^c{}_{ab}{}^d = 0 \quad (P[\phi] = 0), \tag{7}$$

where h_{ab} is considered to be symmetric, while the propagation identities (4) take the form

$$\Box K_{ab}[v] - 2R^c{}_{ab}{}^d K_{cd}[v] = K_{ab}[Q[v]] + 2R_{(a}{}^c K_{b)c}[v] - 2\mathcal{L}_v R_{ab}, \tag{8}$$
$$(P[E[\phi]] = \sigma[Q[\phi]] + \tau[G])$$

$$Q[v] = \Box v_a + R_a{}^b v_b = \nabla^b K_{ab}[v] - \frac{1}{2}\nabla_a K^b{}_b[v], \tag{9}$$
$$(Q[\phi] = \rho[E[\phi]])$$

where we denoted $(R \cdot v)_a = R_a{}^b v_b$ and $\mathcal{L}_v R_{ab} = v^c \nabla_c R_{ab} + 2R_{c(a}\nabla_{b)}v^c$ is the Lie derivative of R_{ab} with respect to the vector field v.

Suppose now that $v_b := \lambda u_b$, where $u_b u^b = -1$, hence $v_b v^b = \lambda^2$. Also define the orthogonal spatial projection $\gamma_{ab} = g_{ab} + u_a u_b$. Substituting in the Killing operator, we get:

$$K_{ab}[\lambda u] = \lambda(K_{ab}[u] + 2u_{(a}\nabla_{b)} \ln \lambda); \tag{10}$$

Contracting with u (using $\mathring{(-)} = u^e \nabla_e(-)$) gives:

$$\begin{aligned} \lambda^{-1} u^a K_{ab}[\lambda u] &= u^a \nabla_a u_b + u^a \nabla_b u^a + u^a \nabla_a \ln \lambda - \nabla_b \ln \lambda \\ &= \mathring{u}_b + \tfrac{1}{2}\nabla_b (u^a u_a) - \nabla_b \ln \lambda + \mathring{\ln} \lambda u_b, \end{aligned} \tag{11}$$

$$\lambda^{-1} u^a u^b K_{ab}[\lambda u] = u^a \mathring{u}_a - 2\mathring{\ln} \lambda; \tag{12}$$

Now, if we define:

$$\bar{K}_{ab}[u] := \gamma_a^c \gamma_b^d K_{cd}[u], \tag{13}$$

$$A_a[\lambda, u] := \mathring{u}_a - \nabla_a \ln \lambda, \tag{14}$$

$$B_{ab}[u] := (\mathrm{d}A[\lambda, u])_{ab} = 2\nabla_{[a}\mathring{u}_{b]}; \tag{15}$$

using the contractions, the definitions and $u^a u_a = -1$, the Killing operator can be rewritten as:

$$K_{ab}[\lambda u] = \lambda(\gamma_a^c \gamma_b^d K_{cd}[u] - 2u_{(a}(\mathring{u} - \nabla \ln \lambda)_{b)}) = \lambda(\bar{K}_{ab}[u] - 2u_{(a}A_{b)}[\lambda, u]). \tag{16}$$

The system $\bar{K}_{ab}[u] = 0, B_{ab}[u] = 0$ on u_a is equivalent to $\bar{K}_{ab}[u] = 0, A_a[\lambda, u] = 0$ on (λ, u_a), which is equivalent to the original $K_{ab}[v] = 0$ on $v_a = \lambda u_a$.

Let us introduce a new affine connection:

$$\mathbb{D}_a S_b = \nabla_a S_b + u^c(\nabla_a u_b) S_c - u_b(\nabla_a u^c) S_c, \tag{17}$$

$$\mathbb{D}_a T^b = \nabla_a T^b - u^b(\nabla_a u_c) T^c + u_c(\nabla_a u^b) T^c, \tag{18}$$

which also acts as $\mathbb{D}_a U = \nabla_a U$ on scalars and extends to higher rank mixed tensors by the Leibniz rule. It has the compatibility properties

$$\nabla_a(S_b T^b) = (\mathbb{D}_a S_b) T^b + S_b(\mathbb{D}_a T^b), \quad \mathbb{D}_a g_{ab} = 0, \quad \mathbb{D}_a u_b = 0, \quad \mathbb{D}_a u^b = 0, \tag{19}$$

but it is not symmetric (it has torsion).

We substitute the above parametrization into (9) and hide all explicit appearances of derivatives of λ through the identity $\nabla_a \ln \lambda = \mathring{u}_a - A_a$. The u-orthogonal and u-parallel components of equality (9) then implies the identities

$$\hat{\Box} u_a := (\mathbb{D}^c + 2\mathring{u}^c)\nabla_c u_a - (\nabla^c u_c)\mathring{u}_a = (\mathbb{D}^c + \mathring{u}^c)[\bar{K}_{ca} - \tfrac{1}{2}\gamma_{ca}\bar{K}_b^b] + \mathring{u}^d \bar{K}_{da} - u^c B_{ca}, \tag{20}$$

$$\mathbb{D}^c \mathring{u}_c + 2\mathring{u}^c \mathring{u}_c + (\nabla^c u^b)(\nabla_c u_b) + \frac{2\Lambda}{d-2} = \tfrac{1}{2}\mathring{\bar{K}}_b^b + (\nabla^c u^d)\bar{K}_{cd}. \tag{21}$$

which makes $\hat{\Box} u_a$ a propagation operator for the u-data of the $\bar{K}_{ab}[u] = 0$, $B_{ab}[u] = 0$ Killing system.

Next, we expand the identity (8). The u-orthogonal and u-parallel components of the resulting expression imply the identities

$$
\begin{aligned}
&(\mathbb{D}^c + \mathring{u}^c)(\mathbb{D}_c + \mathring{u}_c)\bar{K}_{ab} + (\mathring{u}^c\mathring{u}_c + \mathring{u}^c\mathbb{D}_c)\bar{K}_{ab} \\
&\quad + 2\bar{K}^{cd}\rho_{ac:bd} + 4\bar{B}_{(a|}{}^c\nabla_c u_{|b)} - 4\mathring{u}_{(a}C_{b)} \\
&\quad + 2\bar{K}_{(a|d}(\nabla_c u^d)(\nabla^c u_{|b)}) \\
&\quad - u^c[2(\mathbb{D}_d\bar{K}_{(a|c})(\nabla^d u_{|b)}) + (\nabla^d u_d)(\mathbb{D}_c\bar{K}_{ab})] \\
&\quad - 2(\nabla + uu^e\nabla_e)_{(a}u_{b)}[\tfrac{1}{2}\mathring{\bar{K}}{}^c_c + (\nabla^c u^d)\bar{K}_{cd} + \frac{2\Lambda}{d-2}] \qquad (*) \\
&= 2(\mathbb{D} + \mathring{u} + uu^c\mathbb{D}_c)_{(a}\hat{\Box}u_{b)}, \\
\\
&-\Box_{dR}B_{ab} + 4(\nabla_{[a}\mathring{u}^c)B_{b]c} + 4\nabla_{[a|}C^c\nabla_c u_{|b]} - 4\nabla_{[a}\bar{K}^{cd}\sigma_{b]:cd} \\
&\quad + 2\nabla_{[a|}[(\mathbb{D}^c + 3\mathring{u}^c - (\nabla^e u_e)u^c)(\nabla_c u^d) + 2(\nabla_c u^d)\mathbb{D}^c]\bar{K}_{|b]d} \\
&\quad + 2\nabla_{[a}u_{b]}[(\nabla^d u^c)(\nabla_d u^e)\bar{K}_{ce} + \bar{K}^{cd}\tau_{cd}] \\
&= -2\nabla_{[a|}[-u^c\mathbb{D}_c\hat{\Box}u + (\nabla u^c)\hat{\Box}u_c]_{|b]} \qquad (**)
\end{aligned}
$$

where we have decomposed B_a as $B_{ab} = \bar{B}_{ab} + (u \wedge C)_{ab} = \bar{B}_{ab} + u_aC_b - u_bC_a$, where $u^a\bar{B}_{ab} = 0$ and $u^aC_a = 0$.

If we take the identities $(*)$ and $(**)$, together with (20) we get a system of propagation identities for the $\bar{K}_{ab}[u] = 0, B_{ab}[u] = 0$ system of unitary Killing equations:

$$
L\begin{bmatrix}\bar{K}[u]\\ B[u]\end{bmatrix} = \begin{bmatrix}* & *\\ * & *\end{bmatrix}\begin{bmatrix}\bar{K}[u]\\ B[u]\end{bmatrix} = \begin{bmatrix}*\\ *\end{bmatrix}\hat{\Box}u.
$$

To show that this system is ADN strictly hyperbolic, we need to find the weights, with respect to which it satisfies 1. Using the above block decomposition, let us assign the weights $(t_1, t_2) = (3, 2)$ and $(s_1, s_2) = (1, 0)$. Then the weighted principal symbol is determined by the orders:

$$
\begin{bmatrix}t_1 - s_1 & t_2 - s_1\\ t_1 - s_1 & t_2 - s_2\end{bmatrix} = \begin{bmatrix}2 & 1\\ 3 & 2\end{bmatrix}: \quad L^0(p) = \begin{bmatrix}p^2 & 0\\ 0 & p^2\end{bmatrix},
$$

where the first p^2 is the principal symbol of $\Box$ multiplies the identity operator on $\bar{K}[u]$ the same for the the second p^2 but on $B[u]$, which clearly satisfies the desired definition.

4 Initial data

Now that we have the propagation equations, we find the corresponding initial data conditions. For n-dimensional Λ-vacua, the KID equations are [6, Eq.(20)]

$$
\begin{aligned}
&\mathrm{KID}^{(1)}_{AB}[v] := D_Av_B + D_Bv_A - 2\pi_{AB}v_0 = 0, \\
&\mathrm{KID}^{(2)}_{AB}[v] := D_AD_Bv_0 + (2(\pi\cdot\pi)_{AB} - \pi\pi_{AB} - r_{AB})v_0 \\
&\qquad\qquad - 2\pi_{(B}{}^CD_{A)}v_C - (D^C\pi_{AB})v_C + \frac{4\Lambda}{n-2}g_{AB}v_0 = 0.
\end{aligned} \qquad (22)
$$

where D_A is the Levi-Civita connection of the induced metric g_{AB} on a spacelike (Cauchy) surface, and π_{AB} is its extrinsic curvature. The Riemann and Ricci curvatures of the induced metric are denoted by r_{ABCD} and r_{AB}. If t is the time function in Gaussian normal coordinates of the surface and h^A_a is the inclusion of spatial vectors into spacetime vectors, then $v_a = v_0(\nabla_a t) + h^A_a v_A$ is the restriction of the putative Killing vector to the surface.

The rescaling $v_a \mapsto \lambda v_a$ is equivalent to $(v_0, v_A) \mapsto (\lambda v_0, \lambda v_A)$ and the normalization $u_a u^a = -1$ is equivalent to $u_A u^A = u_0^2 - 1$. Now, choose λ so that $v_a = \lambda u_a$, with u_a unit normalized. Under rescaling:

$$\begin{aligned}
\lambda^{-1}\,\mathrm{KID}^{(1)}_{AB}[\lambda u] &= \mathrm{KID}^{(1)}_{AB}[u] + u_A D_B \ln\lambda + u_B D_A \ln\lambda,\\
\lambda^{-1}\,\mathrm{KID}^{(2)}_{AB}[\lambda u] &= \mathrm{KID}^{(2)}_{AB}[u] + u_0\lambda^{-1} D_A D_B \lambda + 2u_0 (D_{(A}u_0)(D_{B)}\ln\lambda) - 2u_C \pi^C{}_{(B} D_{A)} \ln\lambda\\
&= \mathrm{KID}^{(2)}_{AB}[u] + u_0 [D_A D_B \ln\lambda + (D_A\ln\lambda)(D_B\ln\lambda)]\\
&\quad + 2(D_{(A}u_0)(D_{B)}\ln\lambda) - 2u_C \pi^C{}_{(B} D_{A)}\ln\lambda.
\end{aligned}$$

Contracting with u_A gives:

$$\begin{aligned}
u^A u^B \lambda^{-1}\,\mathrm{KID}^{(1)}_{AB}[\lambda u] &= u^A u^B\,\mathrm{KID}^{(1)}_{AB}[u] + 2(u_0^2-1) u^B D_B \ln\lambda,\\
u^B \lambda^{-1}\,\mathrm{KID}^{(1)}_{AB}[\lambda u] &= u^B\,\mathrm{KID}^{(1)}_{AB}[u] + u_A (u^B D_B \ln\lambda) + (u_0^2-1) D_A \ln\lambda.
\end{aligned}$$

Hence, defining

$$L_A[v] := \frac{v^B\,\mathrm{KID}^{(1)}_{AB}[v]}{(v_C v^C)} - \frac{v_A (v^B v^C\,\mathrm{KID}^{(1)}_{BC}[v])}{2(v_C v^C)^2}, \tag{23}$$

we get the identity (recalling that $u_A u^A = u_0^2 - 1$)

$$L_A[\lambda u] = L_A[u] + D_A \ln\lambda.$$

Next, let us define the following operators acting on a unit vector (u_0, u_A):

$$\begin{aligned}
\mathrm{uKID}^{(0)}_{AB}[u] &:= 2D_{[A}L_{B]}[\lambda u],\\
\mathrm{uKID}^{(1)}_{AB}[u] &:= \lambda^{-1}\,\mathrm{KID}^{(1)}_{AB}[\lambda u] - 2u_{(A}L_{B)}[\lambda u],\\
\mathrm{uKID}^{(2)}_{AB}[u] &:= \lambda^{-1}\,\mathrm{KID}^{(2)}_{AB}[\lambda u] - u_0\left(D_{(A}L_{B)}[\lambda u] + L_A[\lambda u]L_B[\lambda u] - 2L_{(A}[u]L_{B)}[\lambda u]\right)\\
&\quad - 2L_{(A}[\lambda u]D_{B)}u_0 + 2u_C\pi^C{}_{(A}L_{B)}[\lambda u],
\end{aligned} \tag{24}$$

where the the λ-dependence on the rhs completely cancels. Setting the three uKID operators to 0 gives the unit KID conditions on (u_0, u_A). Namely, if the KID conditions (22) hold on $(v_0, v_A) = (\lambda u_0, \lambda u_A)$ the definitions (23) and (24) imply that also the uKID conditions hold. Conversely, note first that

$$\mathrm{uKID}^{(0)}_{AB}[u] = D_{[A}L_{A]}[u] = 0 \quad\Longleftrightarrow\quad L_A[u] = -D_A\ln\lambda \quad\Longleftrightarrow\quad L_A[\lambda u] = 0.$$

Then, once the last equality holds for some λ, the KID conditions on $(\lambda u_0, \lambda u_A)$ are reconstructed as

$$\begin{aligned}
\lambda^{-1}\,\mathrm{KID}^{(1)}_{AB}[\lambda u] &= \mathrm{uKID}^{(1)}_{AB}[u] + 2u_{(A}L_{B)}[\lambda u],\\
\lambda^{-1}\,\mathrm{KID}^{(2)}_{AB}[\lambda u] &= \mathrm{uKID}^{(2)}_{AB}[u] + u_0\left(D_{(A}L_{B)}[\lambda u] + L_A[\lambda u]L_B[\lambda u] - 2L_{(A}[u]L_{B)}[\lambda u]\right)\\
&\quad + 2L_{(A}[\lambda u]D_{B)}u_0 - 2u_C\pi^C{}_{(A}L_{B)}[\lambda u].
\end{aligned}$$

In other words, the uKID conditions are satisfied by a unit (u_0, u_A) if, and only if, there exists a λ such that the KID conditions are satisfied by $(\lambda u_0, \lambda u_A)$.

Acknowledgments

S.M. acknowledges financial support from the Generalitat Valenciana Project AICO/2020/125, the Universitat Politècnica de València Project PAID-11-25 and the Generalitat Valenciana (grant CIACIF/2021/028). I.K. was partially supported by GAČR project GA25-15544S and RVO: 67985840.

References

[1] R. Beig and P. T. Chrusciel, "Killing initial data," *Classical and Quantum Gravity*, vol. 14, pp. A83–A92, 1997.

[2] A. García-Parrado Gómez-Lobo, "Local non-negative initial data scalar characterization of the Kerr solution," *Physical Review D*, vol. 92, p. 124053, 2015.

[3] M. Mars, "A spacetime characterization of the kerr metric," *Classical and Quantum Gravity*, vol. 16, p. 2507, 1999.

[4] J. Ehlers, "Contributions to the relativistic mechanics of continuous media," *General Relativity and Gravitation*, vol. 25, pp. 1225–1266, 1993.

[5] A. García-Parrado and I. Khavkine, "Conformal Killing initial data," *Journal of Mathematical Physics*, vol. 60, p. 122502, 2019.

[6] A. García-Parrado and I. Khavkine, "Closed conformal Killing-Yano initial data," *Classical and Quantum Gravity*, vol. 39, p. 105002, 2022.

[7] S. Gindikin and L. Volevich, *Mixed Problem for Partial Differential Equations with Quasihomogeneous Principal Part*, vol. 147 of *Translations of Mathematical Monographs*. American Mathematical Society, 1995.

On the Einstein-Proca Equations in Spherical Symmetry

Alicia Herrero[♭1], Roberto Dale[♮] and Jesús Martínez[♮]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València, Camí de Vera s/n, València, Spain.

(♮) Departamento de Estadística, Matemática e Informática and Centre of Operations Research (CIO), Universidad Miguel Hernández de Elche, Avda. de la Universidad de Elche, Elche, Spain.

1 Introduction

Vector fields play a fundamental role in modern physics. They appear as gauge fields in the Standard Model, as generators of geometrical symmetries in relativistic theories, as vector potentials in electrodynamics, and as additional dynamical degrees of freedom in modified gravity, cosmology and vector-tensor theories of gravitation.

The dynamics and energetic content of a broad but mathematically controlled class of vector-field theories can be analyzed considering the space-time metric as a prescribed geometrical background during the derivation and without assuming a specific gravitational field equation. This makes the results independent of any particular metric theory of gravity and applicable to arbitrary curved space-times.

On the other hand, in Field Theory, massive bosons with unit spin are described using a relativistic vector field that satisfies the Proca Equation, a generalization of Maxwell's Equation with a mass term. The differential of the 1-form associated with the Proca field by the space-time metric is the 2-form of Proca, which is regular if it has two independent lightlike eigendirections, or singular if it has only one. In parallel to how the Maxwell energy tensor is defined, the Proca energy tensor is introduced, which can be considered as a source of the Einstein Equations of General Relativity. The study of these equations with spherical symmetry has been widely studied in Numerical Relativity by Spanish and Portuguese groups, concentrating mainly on the regular case applied to astrophysical situations (Proca stars [1], [2], [3]).

In the present communication we derive the Euler–Lagrange field equations associated with the most general quadratic Lagrangian density built from the vector field, the metric, the curvature scalar and the first covariant derivatives of the field. In addition, the Hilbert energy-momentum tensor associated with the same Lagrangian, separating the contribution of each invariant term, is also obtained. Moreover, we also analyze the essential equations of Einstein-Proca in spherical symmetry for both classes of fields, regular and singular. We show that the equations naturally become the Einstein-Maxwell ones by cancelling the mass of the particle, and how the Reissner-Nordström solution is recovered (to the regular case) [4] and the Mineur-Vaidya solution (to the singular case) [5] (see too the reedition,[6]). The presented results open the door to the (numerical) study of the

[1] aherrero@mat.upv.es

Einstein-Proca solutions for a singular Proca 2-form, which to our knowledge remains unexplored in Relativistic Astrophysics.

2 Vector field dynamics

Consider a space-time with a metric g and its Levi-Civita connection ∇. From the covariant derivative of a vector field ξ, two second-order tensors are defined, a symmetric part S and an antisymmetric part F:

$$S_{\mu\nu} = \nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu. \tag{1}$$

The tensor S is also the Lie derivative od the metric along ξ, $S = \mathcal{L}_\xi g$, and F can also be interpreted as the exterior derivative of the one-form metrically associated with ξ, $F = d\xi$. This structure is analogous to electromagnetism, where the electromagnetic field tensor is obtained as the exterior derivative of the vector potential.

The quadratic Lagrangian density in the vector field and its first-order derivatives is written using scalar invariants built from S and F and the square norm of the vector field as:

$$L = a\,\mathrm{tr}S^2 + b\,\mathrm{tr}F^2 + c\,(\mathrm{tr}S)^2 + (\mu + \nu R)\xi^2, \tag{2}$$

where a, b, c, μ and ν are real parameters, R is the scalar curvature of the metric, and $\xi^2 = g(\xi, \xi)$.

The restriction to quadratic terms and first order derivatives is physically and mathematically motivated since higher-order invariants, for instance terms cubic in S, would generally introduce nonlinear products of first and second order derivatives in the field equations. Also, a term of the form

$$i^2(\xi)\mathrm{Ric} = R_{\mu\nu}\xi^\mu\xi^\nu, \tag{3}$$

can be rewritten, up to a total divergence, in terms of the invariants already present in the Lagrangian [7].

Applying the variational principle with respect to the vector field ξ, assuming variations that vanish on the boundary of the integration domain, the Euler–Lagrange equations are

$$\nabla_\alpha \left(\frac{\partial L}{\partial \nabla_\alpha \xi_\beta} \right) - \frac{\partial L}{\partial \xi_\beta} = 0 \tag{4}$$

which results in the form

$$a\nabla \cdot S - b\nabla \cdot F + 2c\, d(\nabla \cdot \xi) - \frac{1}{2}(\mu + \nu R)\xi = 0. \tag{5}$$

or equivalently as

$$\alpha\, \Delta\xi + \beta\, i(\xi)\mathrm{Ric} + \gamma\, d(\nabla \cdot \xi) + \rho\, \xi = 0 \tag{6}$$

where $\alpha = a - b$, $\beta = a + b$, $\gamma = a + b + 2c$, and $\rho = -\dfrac{1}{2}(\mu + \nu R)$.

As we can see, this equation has a wave-like term $\Delta\xi$ contribution, an algebraic coupling with the Ricci tensor, a term associated with the divergence of the vector field, and a mass-like or scalar-curvature coupling term. Moreover, it includes several well-known equations as particular cases.

For instance, the Proca field, which describes a massive vector field, has a Lagrangian of the form

$$L_P = \frac{1}{4}\mathrm{tr}F^2 - \frac{\mu_p^2}{2}\,\xi^2, \tag{7}$$

where μ_p is the Proca mass. Then, the corresponding field equation is

$$\nabla \cdot F - \mu_p^2 \xi = 0 \tag{8}$$

or, equivalently,

$$\Delta\xi - i(\xi)\mathrm{Ric} - \mu_p^2\,\xi = 0. \tag{9}$$

which is a particular case of equation (5).

Other known situations can also be recovered from the general equation (5):

- The Almost-Killing equation generalizes the condition of a Killing vector field. It is written as

$$\Delta\xi + i(\xi)\mathrm{Ric} + (1-\lambda)d(\nabla\cdot\xi) = 0. \tag{10}$$

  which corresponds to $\alpha = 1$, $\beta = 1$, $\gamma = 1-\lambda$, and $\rho = 0$. This equation appears in general relativity, for example as a possible gauge condition in 3+1 formulations of Einstein's equations.

- The Will equation, introduced in vector-tensor theories of gravitation, can also be easily recovered.

- The Dale–Sáez equation, used in certain cosmological models of vector-tensor gravity, also appears as a particular case. Its characteristic form is

$$\nabla\cdot F - \frac{\eta}{2\omega-\eta}d(\nabla\cdot\xi) = 0, \tag{11}$$

  or, equivalently,

$$\Delta\xi - i(\xi)\mathrm{Ric} - \frac{2\omega}{2\omega-\eta}d(\nabla\cdot\xi) = 0. \tag{12}$$

The systematic comparison among these models shows that apparently different vector-field theories can be understood as different choices of the free parameters in the general Lagrangian.

3 Energy tensor of a vector field

In this section we will focus on the Proca field to get its energy-momentum tensor. For a Lagrangian L, the energy-momentum tensor is defined by

$$T^{\mu\nu} = \frac{2}{\sqrt{-g}}\frac{\delta\mathcal{L}}{\delta g_{\mu\nu}}, \tag{13}$$

where $\mathcal{L} = L\sqrt{-g}$.

Since the general Lagrangian contains several independent terms, the total tensor is obtained as the sum of the contributions associated with each piece. This strategy makes it possible to identify which part of the energy arises from the symmetric component, the antisymmetric component, the divergence term and the mass term.

The term $\nabla\cdot\xi$ generates a null energy tensor, that is, a total divergence contributes neither to the field equations nor to the energy-momentum tensor.

Now, the term $(\nabla\cdot\xi)^2$ will generate a energy-momentum tensor of the form

$$T = -\left[(\nabla\cdot\xi)^2 + 2i(\xi)d(\nabla\cdot\xi)\right]g + 2\,\xi\widetilde{\otimes}d(\nabla\cdot\xi). \tag{14}$$

where $\widetilde{\otimes}$ denotes the symmetrized tensor product, that is, $x\widetilde{\otimes}y = x\otimes y + y\otimes x$, being x and y vector fields.

The symmetric part of the lagrangian with the form $\frac{1}{4}\mathrm{tr}S^2$ will give a energy-momentum tensor of the form

$$T = -S^2 + \frac{1}{4}(\mathrm{tr}S^2)g - (\nabla\cdot\xi)S + \xi\widetilde{\otimes}(\nabla\cdot S) - \mathcal{L}_\xi S^*, \tag{15}$$

where S^* represents the contravariant tensor metrically equivalent to $S = \mathcal{L}_\xi\, g$. That is, $S^* = (\mathcal{L}_\xi g)^* = -\mathcal{L}_\xi g^*$, where g^* is the contravariant metric tensor.

Respect to the antisymmetric part, a term of the form $\frac{1}{4}\mathrm{tr}F^2$ has associated a Hilbert tensor given by

$$T = -F^2 + \frac{1}{4}(\mathrm{tr}F^2)g, \tag{16}$$

which is the electromagnetic energy-momentum tensor form.

Finally, if a Proca mass term is also included, the energy-momentum tensor of the Proca field will be

$$T_P = -F^2 + \frac{1}{4}(\mathrm{tr}F^2)g + \mu_p^2\left[\xi\otimes\xi - \frac{1}{2}\xi^2 g\right]. \tag{17}$$

So, the total energy-momentum tensor associated with the Lagrangian will be the addition of the corresponding contribution given by every term considered.

4 Proca field in Spherical Symmetry

The line element of a spherically symmetric space-time (V_4, g), is usually expressed in coordinates adapted to the symmety (t, r, θ, φ), in which the metric has a diagonal form:

$$ds^2 = -e^{2\Phi}dt^2 + e^{2\Lambda}dr^2 + r^2d\theta^2 + r^2\sin^2\theta d\varphi^2$$

with $\Phi(t,r)$ and $\Lambda(t,r)$ as arbitrary functions. However, this metric can have other expressions when we consider different coordinate systems. In fact, working in an orthonormal basis $\{e_\alpha\}_{\alpha=0}^3$, associated to the coordinate system by

$$e_0 = e^{-\Phi}\frac{\partial}{\partial t}, \quad e_1 = e^{-\Lambda}\frac{\partial}{\partial r}, \quad e_2 = \frac{1}{r}\frac{\partial}{\partial\theta}, \quad e_3 = \frac{1}{r\sin\theta}\frac{\partial}{\partial\varphi},$$

the metric expression is

$$g = ds^2 = -(\omega^0)^2 + (\omega^1)^2 + (\omega^2)^2 + (\omega^3)^2$$

where $\{\omega_\alpha\}_{\alpha=0}^3$ is the algebraic dual basis: $\omega^0 = e^\Phi dt, \quad \omega^1 = e^\Lambda dr, \quad \omega^2 = r\ d\theta, \quad \omega^3 = r\ \sin\theta\ d\varphi$. This coordinate system allows to write the Einstein tensor as $G = G^\beta_{\ \alpha} e_\beta \otimes \omega^\alpha$ with

$$G^0_{\ 0} = -\frac{1}{r^2} + e^{-2\Lambda}\left(\frac{1}{r^2} - \frac{2}{r}\Lambda'\right) \tag{18}$$

$$G^1_{\ 0} = -G^0_{\ 1} = \frac{2}{r}\dot{\Lambda}\ e^{-\Lambda-\Phi} \tag{19}$$

$$G^1_{\ 1} = -\frac{1}{r^2} + e^{-2\Lambda}\left(\frac{1}{r^2} + \frac{2}{r}\Phi'\right) \tag{20}$$

$$G^2_{\ 2} = G^3_{\ 3} = e^{-2\Lambda}\left(\Phi'' + \Phi'^2 - \Lambda'\Phi' + \frac{\Phi'}{r} - \frac{\Lambda'}{r}\right) - e^{-2\Phi}\left(\ddot{\Lambda} + \dot{\Lambda}^2 - \dot{\Lambda}\dot{\Phi}\right) \tag{21}$$

and the other components equal to zero. In this expressions the dot and the prime represent the derivatives with respect to the t or the r variables, respectively.

On the other hand, a vector field ξ can be written in both basis having that

$$\xi = A^\alpha e_\alpha \equiv A^0 e^{-\Phi}\frac{\partial}{\partial t} + A^1 e^{-\Lambda}\frac{\partial}{\partial r} + \frac{A^2}{r}\frac{\partial}{\partial\theta} + \frac{A^3}{r\sin\theta}\frac{\partial}{\partial\varphi}$$

and its corresponding associated $1-$form by the metric will result:

$$\xi = -A^0\omega^0 + A^1\omega^1 + A^2\omega^2 + A^3\omega^3 = -A^0e^\Phi dt + A^1e^\Lambda dr + A^2 r d\theta + A^3 r\sin\theta d\varphi.$$

As we have seen, the energy-momentum tensor associated with a Proca vector field is given by expression (17), which has an electromagentic-form part $M = -F^2 + \frac{1}{4}(\mathrm{tr}F^2)g$ and the Proca mass term $N = \mu_p^2\left[\xi\otimes\xi - \frac{1}{2}\xi^2 g\right]$. These two parts can be analyzed separately having that the study of M will depend on the character of the $2-$form F.

- If F is a regular 2-form, $M = M^\beta_\alpha e_\alpha\otimes\omega^\beta = \mathrm{diag}(-f^2,-f^2,f^2,f^2)$, that is, $M^0_0 = M^1_1 = -M^2_2 = -M^3_3 = -f^2$ and the other components are equal to zero. In this case

$$2f^2 = e^{-2(\Phi+\Lambda)}F^2_{tr} + \frac{1}{r^4\sin^2\theta}F^2_{\theta\varphi} \tag{22}$$

  with

$$F_{tr} = (A^0\ e^\Phi)' + (A^1\ e^\Lambda)^\cdot \qquad \text{and} \qquad F_{\theta\varphi} = -(A^2 r)_{,\varphi} + r\sin\theta(A^3)_{,\theta} + r\cos\theta A^3.$$

- If F is singular $M = \rho\, l\otimes l$, where l is the fundamental direction of F, which can be written as

$$l = -e^\Phi dt + \varepsilon\ e^\Lambda dr = -\omega^0 + \varepsilon\ \omega^1 \qquad \text{or} \qquad l = e^{-\Phi}\frac{\partial}{\partial t} + \varepsilon\ e^{-\Lambda}\frac{\partial}{\partial r} = e_0 + \varepsilon\ e_1,$$

  with $\varepsilon = \pm 1$. Then, $M^0_0 = -M^1_1 = -\rho$, $M^0{}_1 = -M^1{}_0 = \varepsilon\rho$ and the other components equal zero.

For the Proca mass term, in components, we have:

$$N^0_0 = -\frac{\mu_p}{4}\Big((A^0)^2 + (A^1)^2 + (A^2)^2 + (A^3)^2\Big) \tag{23}$$

$$N^i_0 = -N^0_i = \frac{\mu_p}{2}A^0A^i, \qquad i = 1,2,3 \tag{24}$$

$$N^1_1 = \frac{\mu_p}{4}\Big((A^0)^2 + (A^1)^2 - (A^2)^2 - (A^3)^2\Big) \tag{25}$$

$$N^i_1 = N^1_i = \frac{\mu_p}{2}A^1A^i, \qquad i = 2,3 \tag{26}$$

$$N^2_2 = \frac{\mu_p}{4}\Big((A^0)^2 - (A^1)^2 + (A^2)^2 - (A^3)^2\Big) \tag{27}$$

$$N^3_2 = N^2_3 = \frac{\mu_p}{2}A^2A^3 \tag{28}$$

$$N^3_3 = \frac{\mu_p}{4}\Big((A^0)^2 - (A^1)^2 - (A^2)^2 + (A^3)^2\Big) \tag{29}$$

Now, the Einstein equations $G^\alpha_\beta = \kappa T^\alpha_\beta$ will consider the addition of both terms, using the corresponding electromagnetic term depending on the F character. On one hand, we can observe that if the Proca mass vanishes, $\mu_p = 0$, the Einstein-Maxwell equations are obtained, which can be solved leading to the Reissner-Nordström solution when F is regular (see [4]) and to the Mineur-Vaidya solution for a singular F (see [5] and [6]).

On the other hand, in the generic case $\mu_p \neq 0$, for any F, the components $G^0_i = \kappa T^0_i$ and $G^1_i = \kappa T^1_i$, for $i = 2,3$, and $G^3_2 = \kappa T^3_2$ result

$$0 = \frac{\mu_p}{2}A^0A^i, \qquad 0 = \frac{\mu_p}{2}A^1A^i, \qquad \text{and} \qquad 0 = \frac{\mu_p}{2}A^2A^3$$

for $i = 2,3$. This equations imply two situations: the nullity of the components 2 and 3 of the vector field; or the nullity of the components A^0, A^1 and A^2 or A^3. The case of a radial field in the $\{t,r\}-$plane, that is, $\xi = A^0\ e_t + A^1\ e_r$ is usually considered in the literature. So, in the following, we will focus in this generic case.

Next, we consider the Proca field equation (8) for this vector field. This equation can be developed and written in this coordinate system. We begin with the factor $\nabla \cdot F$, which can be calculated using the relation:

$$(\nabla \cdot F)^{\beta} = \nabla_{\alpha} F^{\alpha\beta} = \frac{1}{\sqrt{-g}} \partial_{\alpha} (\sqrt{-g}\ F^{\alpha\beta})$$

Consequently, the Proca field equations are:

$$\frac{e^{-\Lambda}}{r^2}(r^2 e^{-(\Phi+\Lambda)} F_{tr})' + \frac{e^{-\Phi}}{r^2 \sin\theta}(\sin\theta F_{t\theta})_{,\theta} + \frac{e^{-\Phi}}{r^2 \sin^2\theta}(F_{t\varphi})_{,\varphi} = \frac{\mu}{2} A^0 \tag{30}$$

$$-e^{-\Phi}(e^{-(\Phi+\Lambda)} F_{tr})^{\cdot} - \frac{e^{-\Lambda}}{r^2 \sin\theta}(\sin\theta F_{r\theta})_{,\theta} - \frac{e^{-\Lambda}}{r^2 \sin^2\theta}(F_{r\varphi})_{,\varphi} = \frac{\mu}{2} A^1 \tag{31}$$

$$(e^{-\Phi+\Lambda} F_{t\theta})^{\cdot} - (e^{\Phi-\Lambda} F_{r\theta})' = 0 \tag{32}$$

$$(e^{-\Phi+\Lambda} F_{t\varphi})^{\cdot} - (e^{\Phi-\Lambda} F_{r\varphi})' = 0 \tag{33}$$

with

$$F_{tr} = (A^0\ e^{\Phi})' + (A^1\ e^{\Lambda})^{\cdot} \qquad F_{t\theta} = (A^0\ e^{\Phi})_{,\theta} = e^{\Phi} A^0_{,\theta} \tag{34}$$

$$F_{t\varphi} = (A^0\ e^{\Phi})_{,\varphi} = e^{\Phi} A^0_{,\varphi} \qquad F_{r\theta} = -(A^1\ e^{\Lambda})_{,\theta} = -\ e^{\Lambda} A^1_{,\theta} \tag{35}$$

$$F_{r\varphi} = -(A^1\ e^{\Lambda})_{,\varphi} = -\ e^{\Lambda} A^1_{,\varphi} \tag{36}$$

Moreover, taking into account that th Proca field also satisfies $\nabla \cdot \xi = 0$, we have to add this equation to the previous ones to have completely defined the field. This last equation can be written as:

$$\nabla \cdot \xi = \frac{1}{\sqrt{-g}} \partial_{\alpha}\big(\sqrt{-g}\ \xi^{\alpha}\big) = 0 \quad \Longrightarrow \quad e^{\Lambda}(\dot{A}^0 + A^0 \dot{\Lambda}) + e^{\Phi}\left((A^1)' + A^1\Phi' + \frac{2}{r} A^1\right) = 0. \tag{37}$$

These equations relate the Proca field form and the metric components. In the next section the Einstein equations will join these Proca equations to complete the system.

5 Einstein-Proca equations in the regular case

In this section we will focus on the Einstein-Proca equations for a radial Proca field in spherical symmetry considering that the 2-form $F = d\xi$ is regular. In the previous section we have seen that the Einstein equations are constructed from the Einstein tensor and the energy-momentum tensor as $G^{\alpha}_{\ \beta} = \kappa T^{\alpha}_{\ \beta}$. So, the field has the form $\xi = A\ e_t + B\ e_r$, where A and B are functions of all the coordinates, and the Einstein equations are written as:

$$-\kappa\ \frac{\mu}{4}\Big(A^2 + B^2\Big) - \kappa f^2 = -\frac{1}{r^2} + e^{-2\Lambda}\left(\frac{1}{r^2} - \frac{2}{r}\Lambda'\right) \tag{38}$$

$$\kappa\ \frac{\mu}{2} A\ B = \frac{2}{r}\dot{\Lambda}\ e^{-\Lambda-\Phi} \tag{39}$$

$$\kappa\ \frac{\mu}{4}\Big(A^2 + B^2\Big) - \kappa f^2 = -\frac{1}{r^2} + e^{-2\Lambda}\left(\frac{1}{r^2} + \frac{2}{r}\Phi'\right) \tag{40}$$

$$\kappa\ \frac{\mu}{4}\Big(A^2 - B^2\Big) + \kappa f^2 = e^{-2\Lambda}\left(\Phi'' + \Phi'^2 - \Lambda'\Phi' + \frac{\Phi'}{r} - \frac{\Lambda'}{r}\right) - \tag{41}$$

$$- \ e^{-2\Phi}\Big(\ddot{\Lambda} + \dot{\Lambda}^2 - \dot{\Lambda}\dot{\Phi}\Big)$$

with f given by (22). In addition, ξ has to satisfy the Proca field equations (30)-(33) and (37) too. Combining these equations, naming $P = Ae^{\Phi}$ and eliminating redundat equations, we arrive at the following result:

Statement. Consider a general spherically symmetric space-time

$$ds^2 = -e^{2\Phi(t,r)}dt^2 + e^{2\Lambda(t,r)}d\rho^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2,$$

a Proca 1-form $\xi = -P(t,r)dt$ and $F = d\xi$. If F is a regular 2-form, the Einstein Proca equations for ξ in g imply that Λ, Φ and P only depend on the r coordinate and, besides, they satisfy the Obukhov-Vlachynsky equations:

$$\begin{aligned}
\Lambda' + \Phi' &= \kappa\, \frac{\mu_p}{4}\, r\, e^{2(\Lambda-\Phi)}\, P^2 \\
\Phi' - \Lambda' &= \frac{1}{r}(e^{2\Lambda} - 1) - \frac{\kappa}{2}\, r\, e^{-2\Phi} P'^2 \\
0 &= P'' + \left(\frac{2}{r} - \kappa\, \frac{\mu}{4} e^{2(\Lambda-\Phi)} r\, P^2\right) P' - \frac{\mu}{2} e^{2\Lambda} P.
\end{aligned}$$

Note that in [8], the authors obtained and analyzed the above system of ordinary differential equations for Λ, Φ and P from a static spherically symmetric space-time. It is worthwhile to remark that in this work we obtain the staticity of the metric from the Einstein-Proca equations in the regular case, a result that extends the idea of the Jensen-Birkoff theorem for a Proca field.

6 Conclusions

This work provides a general and unified formulation for studying relativistic vector fields in curved space-times. A general quadratic Lagrangian for a vector field in curved space-time has been constructed, depending at most on first-order derivatives of the field. Its main value lies in allowing several known theories, including Proca, Almost-Killing, Will and Dale–Sáez models, to be compared within the same framework.

The Hilbert energy-momentum tensor has been obtained systematically by separating the contributions of each term in the Lagrangian.

A study of the Proca field equations in spherical symmetry has also been done and the Einstein-Proca system has been analyzed for a radial vector field that generates a regular 2-form. The particular case of a time-like vector field directly leads to a static solution.

In conclusion, this article provides a solid mathematical basis for analysing the dynamics and energy content of relativistic vector fields in curved space-times. Its formulation is general enough to encompass known models while remaining structured enough to support further developments.

Acknowledgments

This work has been supported by the Universitat Politècnica de València by project PAID-11-25.

References

[1] N. Sanchis-Gual, C. Herdeiro, E. Radu, J. C. Degollado, and J. A. Font, "Numerical evolutions of spherical Proca stars," *Physical Review D*, vol. 95, p. 104 028, 2017. DOI: 10.1103/PhysRevD.95.104028.

[2] R. Brito, V. Cardoso, C. A. R. Herdeiro, and E. Radu, "Proca stars: Gravitating Bose-Einstein condensates of massive 1 particles," *Physics Letters B*, vol. 752, p. 291, 2016. DOI: 10.1016/j.physletb.2015.11.051.

[3] R. Luna, M. Llorens-Monteagudo, A. Lorenzo-Medina, J. C. Bustillo, and et al., "Numerical relativity surrogate waveform models for exotic compact objects: The case of head-on mergers of equal-mass Proca stars," *Physical Review D*, vol. 110, p. 024 004, 2024. DOI: 10.1103/PhysRevD.110.024004.

[4] R. Adler, M. Bazin, and M. Schiffer, *Introduction to General Relativity*. New York: McGraw-Hill Book Company, 1965.

[5] P. C. Vaidya, "The external field of a radiating star in general relativity," *Current Science*, vol. 12, 1943. DOI: https://ui.adsabs.harvard.edu/abs/1943CSci...12..183C.

[6] P. C. Vaidya, "The Gravitational Field of a Radiating Star," *General Relativity and Gravitation*, vol. 31, 1999.

[7] R. Dale, A. Herrero, and J. A. Morales-Lladosa, "Vector field dynamics: field equations and energy tensor," *Eur. Phys. J. Plus*, vol. 139, p. 1013, 2024. DOI: 10.1140/epjp/s13360-024-05782-z.

[8] Y. N. Obukhov and E. J. Vlachynsky, "Einstein-Proca model: spherically symmetric solutions," *Ann. Phys.*, vol. 8 (6), pp. 497–510, 1999. DOI: 10.1002/(SICI)1521-3889(199909)8:6<497::AID-ANDP497>3.0.CO;2-5.

A Preliminary Study of CMB Polarisation Amplitude Statistics and Map Downgrading Techniques

R. Dale[♭1] and J. M. Gandía[♭]

(♭) Departament d'Estadística, Matemàtiques i Informàtica, Universitat Miguel Hernández d'Elx
Avinguda de la Universitat d'Elx, s/n, 03202 Elx, Alacant, Spain.

1 Introduction

Recent work has investigated the possibility of testing Bell-type inequalities in a cosmological context using observations of the cosmic microwave background (CMB). In Ref. [1], a cosmic analogue of the Clauser–Horne–Shimony–Holt (CHSH) inequality was formulated under the assumption of local realism and applied to COBE temperature data. Although no violation was found in that first analysis, the same framework was subsequently extended to higher-resolution WMAP and Planck temperature maps in Ref. [2], where significant violations of the corresponding cosmic CHSH inequalities were reported.

These studies provide the conceptual and computational starting point for the next stage of the programme: the extension of the formalism from CMB temperature anisotropies to CMB polarisation. This step is not merely a matter of replacing one observable by another. The linear polarisation field is described by the Stokes parameters Q and U, which are not scalar fields on the sphere, and therefore require particular care when constructing scalar observables and when degrading maps to lower angular resolution.

In the present contribution we focus on the simplest scalar quantity associated with the linear polarisation field, namely the polarisation amplitude

$$P = \sqrt{Q^2 + U^2}.$$

Before constructing angular correlation functions or CHSH-type combinations from P, it is necessary to understand its empirical statistical behaviour. In particular, P is positive by construction, non-Gaussian, and generally asymmetric. This makes the choice of a dichotomisation threshold non-trivial and motivates the use of the empirical median, rather than the mean, as the natural reference threshold.

A further point requiring attention is the distinction between inpainted and non-inpainted Planck products. The inpainted Commander maps provide a full-sky reconstruction in which masked or low-confidence regions are filled by a statistical procedure, whereas the non-inpainted maps retain the corresponding masked regions explicitly. Since the Planck Collaboration stresses that inpainted CMB maps are primarily intended for visualisation and presentation purposes, rather than for a full scientific analysis, we regard the non-inpainted maps with an explicit polarisation

[1] rdale@umh.es

mask as the more conservative reference for the statistical interpretation [3]. The inpainted maps are nevertheless useful as a robustness check, allowing us to assess the sensitivity of the results to the reconstruction procedure.

Map downgrading constitutes another essential methodological aspect of the present analysis. As already illustrated in the temperature-anisotropy study of Ref. [2], reducing the angular resolution of CMB maps is a practical requirement when dealing with large data sets. In the polarisation case, however, the procedure is more delicate, since constructing P before or after degrading the Stokes maps leads to genuinely different low-resolution observables. The aim of this work is therefore to compare several degradation strategies for Planck Commander polarisation maps, considering both inpainted and non-inpainted products, and to analyse their impact on the basic statistics and empirical distribution of P.

2 Polarisation amplitude and dichotomisation threshold

The first step towards extending the temperature-based cosmic CHSH formalism to CMB polarisation is the identification of a suitable scalar observable. The linear polarisation field is described by the Stokes parameters Q and U, which depend on the choice of the local basis on the tangent plane to the sphere. Under a rotation of this basis, Q and U mix with one another. Therefore, they cannot be treated, individually, as scalar fields in the same sense as the CMB temperature anisotropy.

A natural scalar quantity is instead provided by the polarisation amplitude

$$P(\hat{n}) = \sqrt{Q^2(\hat{n}) + U^2(\hat{n})}, \tag{1}$$

which is invariant under local rotations of the polarisation basis. It should be stressed that, in this preliminary analysis, P is used as an empirical scalar observable constructed from the Commander Stokes maps, rather than as a noise-debiased estimator of the true polarisation amplitude. In this respect, P is the simplest candidate for constructing a dichotomous variable analogous to that used in the temperature analysis. However, the analogy is not complete. In the temperature case, the field is approximately centred and close to Gaussian on the relevant angular scales, so that a threshold based on the mean value is a natural choice. By contrast, P is positive by construction and its empirical distribution is neither centred nor Gaussian.

This difference has a direct consequence for the construction of a dichotomous observable. If the mean value of P were used as a threshold, the resulting binary field would in general be unbalanced, since the positive skewness of the distribution implies that the mean and the median do not coincide. It is therefore necessary to examine the empirical distribution of P before defining the threshold entering any subsequent correlation analysis.

We have performed this preliminary statistical analysis using the original Planck Commander polarisation maps, considering both inpainted and non-inpainted products. In each case, the polarisation amplitude was constructed from the Stokes parameters according to Eq. (1), and the corresponding distribution was fitted to several standard families supported on the positive real line. The Rayleigh distribution is the natural reference model, since it is obtained when Q and U are independent centred Gaussian variables with equal variance. Nevertheless, departures from this ideal case may arise from residual foregrounds, noise, masking, and map-processing effects.

Table 1 summarises the most relevant distributional fits for the original Commander maps. For the inpainted map, Weibull, Nakagami and Rayleigh provide very similar values of the Akaike information criterion (AIC), indicating a distribution close to the Rayleigh limit. For the non-inpainted map, Weibull and Nakagami remain the preferred descriptions, but the departure from the Rayleigh reference is more visible. In this case, beta and gamma distributions also provide competitive fits, suggesting a more complex empirical structure.

Table 1: Relevant distributional fits for the polarisation amplitude $P = \sqrt{Q^2 + U^2}$ obtained from the original Planck Commander maps. The inpainted and non-inpainted cases are shown separately. The Rayleigh distribution is included as the reference model corresponding to two independent centred Gaussian Stokes components with equal variance.

Map	Distribution	Par1	Par2	Par3	LogL	AIC
Inpainted	Weibull	2.299×10^{-3}	1.953	–	2.378×10^{8}	-4.757×10^{8}
Inpainted	Nakagami	9.699×10^{-1}	5.339×10^{-6}	–	2.378×10^{8}	-4.756×10^{8}
Inpainted	Rayleigh	1.634×10^{-3}	–	–	2.378×10^{8}	-4.756×10^{8}
Non-inpainted	Weibull	3.931×10^{-2}	1.833	–	1.124×10^{8}	-2.249×10^{8}
Non-inpainted	Nakagami	8.844×10^{-1}	1.612×10^{-3}	–	1.124×10^{8}	-2.248×10^{8}
Non-inpainted	Rayleigh	2.839×10^{-2}	–	–	1.122×10^{8}	-2.243×10^{8}
Non-inpainted	Beta	2.692	7.449×10^{1}	–	1.122×10^{8}	-2.245×10^{8}
Non-inpainted	Gamma	2.770	1.260×10^{-2}	–	1.122×10^{8}	-2.243×10^{8}

The parameters in Table 1 follow the usual parametrisation of each distribution family.

The Rayleigh reference is also reflected in the fitted parameters: the Weibull shape parameter is close to 2, and the Nakagami shape parameter is close to 1, especially in the inpainted case. The non-inpainted map remains broadly Rayleigh-like, although with a more appreciable deviation from this limiting behaviour.

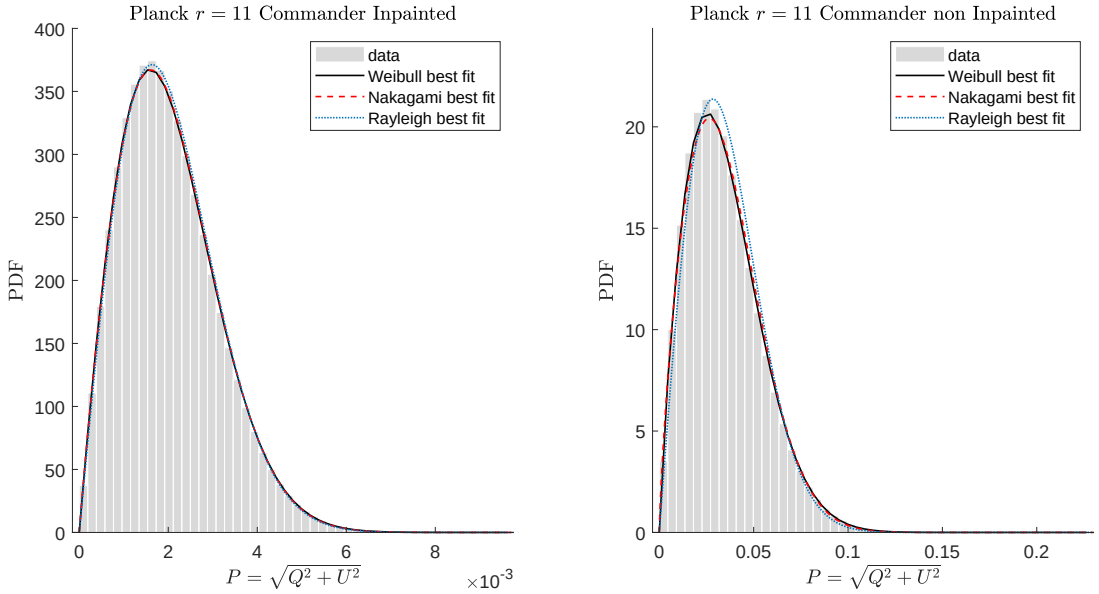


Figure 1: Normalised histograms of $P = \sqrt{Q^2 + U^2}$ for the original Planck Commander maps: inpainted (left) and non-inpainted (right). The empirical histograms are compared with Weibull, Nakagami and Rayleigh fits.

For a like-for-like comparison, the same polarisation mask is applied to both the inpainted and non-inpainted maps when computing the empirical statistics.

The same conclusion is reflected in the basic empirical statistics. Let μ_P, σ_P and P_{50} denote, respectively, the mean, standard deviation and median of the polarisation amplitude over the valid pixels. The values obtained for the original Commander maps are shown in Table 2. In both cases,

$\mu_P > P_{50}$, which is the expected signature of a positively skewed distribution. This confirms that the mean and the median define genuinely different thresholds for P.

Table 2: Basic empirical statistics of the polarisation amplitude $P = \sqrt{Q^2 + U^2}$ for the original Planck Commander maps, after removing masked pixels. All quantities are expressed in $\mathrm{mK_{CMB}}$.

Map	μ_P	σ_P	P_{50}	N_{pix}
Inpainted	2.0376×10^{-3}	1.0898×10^{-3}	1.8986×10^{-3}	43516974
Non-inpainted	3.4898×10^{-2}	1.9841×10^{-2}	3.1914×10^{-2}	43516974

For this reason, we adopt the empirical median P_{50}, computed after applying the relevant mask, as the reference threshold for the main dichotomisation. The corresponding binary field is defined as

$$F_P(\hat{n}) = \mathrm{sign}\left[P(\hat{n}) - P_{50}\right], \tag{2}$$

where

$$P_{50} = \mathrm{median}\{P(\hat{n}) : \hat{n} \in \mathcal{M}\}, \tag{3}$$

and $\mathcal{M}$ denotes the set of valid pixels after masking.

This choice has a simple statistical advantage: except for possible numerical ties exactly at the threshold, it divides the valid pixels into two subsets of approximately equal cardinality. Hence,

$$f_+ = \frac{N(P > P_{50})}{N_{\mathrm{pix}}} \simeq \frac{1}{2}, \qquad f_- = \frac{N(P < P_{50})}{N_{\mathrm{pix}}} \simeq \frac{1}{2}, \tag{4}$$

so that the one-point mean of the dichotomised field is approximately zero,

$$\langle F_P \rangle \simeq 0. \tag{5}$$

This avoids introducing, at the thresholding stage, a spurious global imbalance between the two possible outcomes.

The mean-based threshold,

$$F_P^{(\mu)}(\hat{n}) = \mathrm{sign}\left[P(\hat{n}) - \mu_P\right], \tag{6}$$

may still be retained as a robustness check and as a formal comparison with the temperature-based construction. However, it is not the primary choice for the polarisation amplitude, precisely because P is positive and skewed. The subsequent analysis will therefore use P_{50} as the natural dichotomisation threshold, recomputed separately for each map and for each low-resolution construction of the polarisation amplitude.

3 Low-resolution constructions of the polarisation amplitude

The preceding section shows that the statistical properties of P must be examined before using it to define a dichotomous field. A further point arises when the analysis is performed at degraded angular resolution. Since the definition of P is nonlinear in the Stokes parameters, constructing the polarisation amplitude and degrading the map are not interchangeable operations. Consequently, different downgrading procedures may lead to different empirical distributions, and hence to different values of the median threshold P_{50}.

We consider three different low-resolution constructions of the polarisation amplitude. The first one consists of degrading the Stokes maps Q and U to the target resolution and only then constructing the amplitude,

$$P(Q_{\mathrm{deg}}, U_{\mathrm{deg}})(\hat{n}) = \sqrt{Q^2_{\mathrm{deg}}(\hat{n}) + U^2_{\mathrm{deg}}(\hat{n})}. \tag{7}$$

This construction measures the amplitude of the smoothed Stokes components. Since the smoothing acts on Q and U before the nonlinear operation defining P, cancellations between positive and negative values of the Stokes parameters may occur before the modulus is taken.

The second construction follows the opposite order. The amplitude P is first computed at the original resolution and the resulting scalar map is then degraded using the standard pixel-based procedure. We denote by $\mathcal{D}_{\rm ud}$ the corresponding pixel-domain degradation operator. The low-resolution field obtained in this way is

$$P_{\rm ud}(\hat{n}) = \mathcal{D}_{\rm ud}\left[\sqrt{Q^2+U^2}\right](\hat{n}). \tag{8}$$

In this case, the quantity being averaged is already positive.

The third construction is analogous, but the scalar field P, constructed at the original resolution, is degraded by means of a harmonic-space filtering procedure. We denote by $\mathcal{D}_{\rm alm}$ the corresponding harmonic-space degradation operator. The resulting map is

$$P_{\rm alm}(\hat{n}) = \mathcal{D}_{\rm alm}\left[\sqrt{Q^2+U^2}\right](\hat{n}). \tag{9}$$

Here, the degradation is performed through the spherical-harmonic representation of the scalar field.

These three constructions should not be regarded as interchangeable numerical implementations of a single operation. The reason is the nonlinearity of the definition of P. In general,

$$\sqrt{(\mathcal{D}Q)^2+(\mathcal{D}U)^2} \neq \mathcal{D}\left(\sqrt{Q^2+U^2}\right), \tag{10}$$

where $\mathcal{D}$ denotes a generic degradation operator. Thus, the first construction defines the amplitude of the degraded Stokes components, whereas the latter two constructions define degraded versions of an amplitude already formed at the original resolution.

The impact of this distinction is already visible at the level of the basic statistics. Table 3 shows the empirical mean, standard deviation and median for the original maps and for the three low-resolution constructions, considering both inpainted and non-inpainted Commander products. The original maps have resolution $r = 11$, whereas all low-resolution maps have resolution $r = 6$.

Table 3: Basic empirical statistics of the polarisation amplitude for different low-resolution constructions, comparing inpainted and non-inpainted Planck Commander maps. The original maps correspond to resolution $r = 11$, whereas the remaining rows correspond to resolution $r = 6$. All quantities are expressed in $\mathrm{mK_{CMB}}$.

Map	Construction	μ_P	σ_P	P_{50}
Inpainted	Original $r = 11$	2.0376×10^{-3}	1.0898×10^{-3}	1.8986×10^{-3}
Inpainted	$P(Q_{\rm deg}, U_{\rm deg})$	7.9601×10^{-4}	4.2839×10^{-4}	7.3980×10^{-4}
Inpainted	$P_{\rm ud}$	2.0382×10^{-3}	5.1759×10^{-4}	1.9810×10^{-3}
Inpainted	$P_{\rm alm}$	2.0380×10^{-3}	3.3287×10^{-4}	2.0180×10^{-3}
Non-inpainted	Original $r = 11$	3.4898×10^{-2}	1.9841×10^{-2}	3.1914×10^{-2}
Non-inpainted	$P(Q_{\rm deg}, U_{\rm deg})$	8.7743×10^{-4}	4.7803×10^{-4}	8.0947×10^{-4}
Non-inpainted	$P_{\rm ud}$	3.4884×10^{-2}	6.8715×10^{-3}	3.5335×10^{-2}
Non-inpainted	$P_{\rm alm}$	3.4883×10^{-2}	6.7977×10^{-3}	3.5368×10^{-2}

Several features follow from Table 3. First, the construction $P(Q_{\rm deg}, U_{\rm deg})$ leads to a substantial reduction in both the mean and the median. This is especially clear in the non-inpainted case, where the original map has a much larger characteristic amplitude, whereas the amplitude reconstructed from the degraded Stokes components is reduced by more than one order of magnitude. This

behaviour is expected, since the Stokes components are smoothed before forming the positive amplitude.

By contrast, the constructions $P_{\rm ud}$ and $P_{\rm alm}$ preserve the mean value of P to high accuracy. This occurs because the positive scalar amplitude has already been constructed at the original resolution before the degradation is applied. However, these two procedures markedly reduce the standard deviation, showing that the resulting maps are considerably narrower around their characteristic positive scale.

The median threshold is affected accordingly. For $P(Q_{\rm deg}, U_{\rm deg})$, the median is shifted to much smaller values, consistently with the reduction in the whole amplitude scale. For $P_{\rm ud}$ and $P_{\rm alm}$, the median remains close to the original scale, but the distribution becomes more concentrated. This reinforces the need to recompute P_{50} independently for each construction, rather than transferring a threshold obtained from another map or from another degradation procedure.

3.1 Distributional behaviour after map downgrading

The differences described above are not restricted to the first empirical moments. They also affect the full distributional shape of the polarisation amplitude. To quantify this effect, we fitted the degraded maps to the same families of positive distributions considered in the original-resolution analysis. The Rayleigh distribution is again included as a physical reference, since it is expected when the amplitude is formed from two independent centred Gaussian components with equal variance.

Table 4 summarises the most relevant fits for the three low-resolution constructions. The construction $P(Q_{\rm deg}, U_{\rm deg})$ remains close to the Rayleigh reference in both the inpainted and non-inpainted cases. The fitted Weibull and Nakagami parameters remain close to their Rayleigh limiting values, and the AIC values of the Rayleigh fits are very close to those of the best-fitting distributions.

By contrast, the distributions of $P_{\rm ud}$ and $P_{\rm alm}$ are substantially different. In these cases, Rayleigh fits are no longer competitive: the empirical distributions are narrower, more concentrated around a positive scale, and have much less probability mass near $P = 0$. This behaviour is a direct consequence of degrading a positive amplitude already formed at the original resolution.

Table 4: Summary of the main distributional fits at resolution $r = 6$. For each construction, the best-fitting family is shown together with the Rayleigh reference.

Map	Construction	Fit	AIC	Comment
Inpainted	$P(Q_{\rm deg}, U_{\rm deg})$	Weibull / Rayleigh	-5.270×10^5 / -5.269×10^5	Rayleigh-like
Inpainted	$P_{\rm ud}$	GEV / Rayleigh	-5.092×10^5 / -4.771×10^5	non-Rayleigh
Inpainted	$P_{\rm alm}$	Gamma / Rayleigh	-5.434×10^5 / -4.816×10^5	non-Rayleigh
Non-inpainted	$P(Q_{\rm deg}, U_{\rm deg})$	Weibull / Rayleigh	-5.184×10^5 / -5.182×10^5	Rayleigh-like
Non-inpainted	$P_{\rm ud}$	Logistic / Rayleigh	-2.946×10^5 / -2.460×10^5	non-Rayleigh
Non-inpainted	$P_{\rm alm}$	Logistic / Rayleigh	-2.956×10^5 / -2.461×10^5	non-Rayleigh

This behaviour is illustrated in Figs. 2 and 3. The construction $P(Q_{\rm deg}, U_{\rm deg})$ remains close to the Rayleigh-like reference, whereas $P_{\rm ud}$ and $P_{\rm alm}$ are clearly narrower and shifted away from $P = 0$.

Overall, the three procedures define different low-resolution observables. The construction $P(Q_{\rm deg}, U_{\rm deg})$ preserves a Rayleigh-like character, whereas $P_{\rm ud}$ and $P_{\rm alm}$ preserve the mean amplitude more closely but suppress the low-P region and reduce the relative dispersion. Hence the degradation strategy affects not only μ_P, σ_P and P_{50}, but also the full empirical distribution of P. The median threshold must therefore be recomputed for each map and each degradation procedure before constructing angular correlations or CHSH-type combinations.

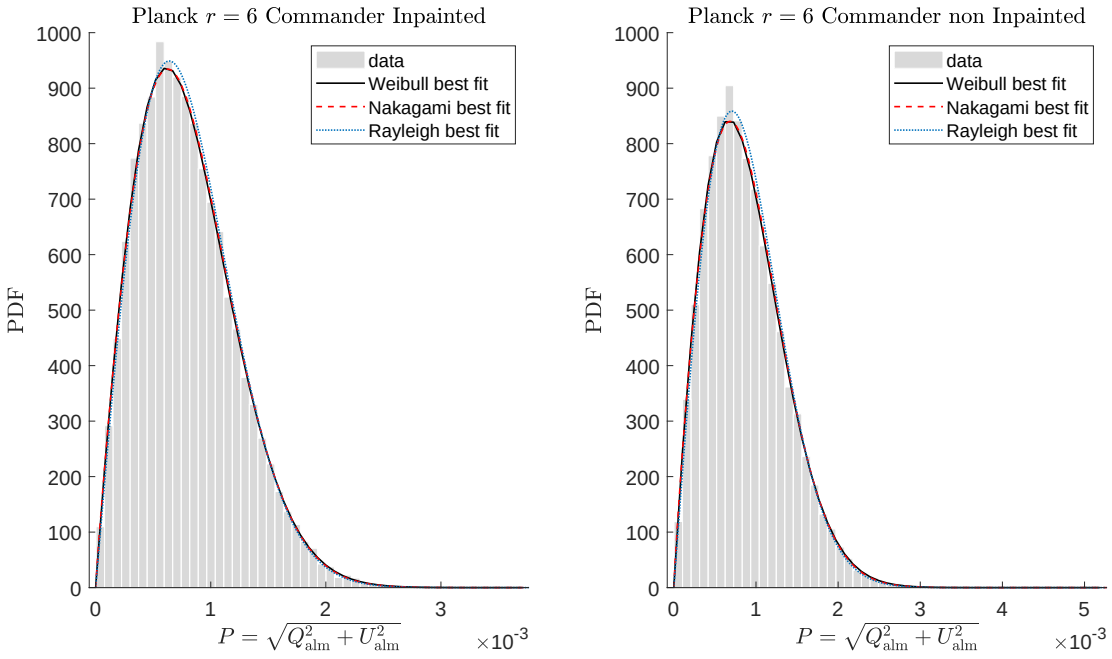


Figure 2: Normalised histograms of $P(Q_{\rm deg}, U_{\rm deg})$ at resolution $r = 6$, for inpainted (left) and non-inpainted (right) Commander maps. Weibull, Nakagami and Rayleigh fits are shown for comparison.

4 Conclusions

We have presented a preliminary statistical study of the CMB polarisation amplitude

$$P = \sqrt{Q^2 + U^2}$$

as a necessary step towards extending cosmic CHSH-type analyses from temperature anisotropies to polarisation. Since P is positive by construction and has a non-Gaussian, asymmetric empirical distribution, the thresholding strategy used for temperature cannot be transferred without modification. In particular, the mean value is not the most natural reference threshold. We have therefore adopted the empirical median P_{50}, computed after masking, as the primary dichotomisation threshold.

The main result of this work is that map downgrading and the nonlinear construction of P do not commute. Consequently, the three low-resolution constructions considered here,

$$P(Q_{\rm deg}, U_{\rm deg}), \qquad P_{\rm ud}, \qquad P_{\rm alm},$$

should not be interpreted as equivalent implementations of the same observable. The first construction, in which the Stokes maps are degraded before forming the amplitude, preserves a Rayleigh-like distributional behaviour, consistent with an amplitude built from two smoothed components. By contrast, the direct degradation of P, either through the pixel-domain operator or through harmonic-space filtering, preserves the mean amplitude much more closely but produces narrower, positively shifted and clearly non-Rayleigh-like distributions.

This distinction affects not only the empirical mean and standard deviation, but also the median threshold P_{50} and the full distributional shape of the polarisation amplitude. Therefore, P_{50} must

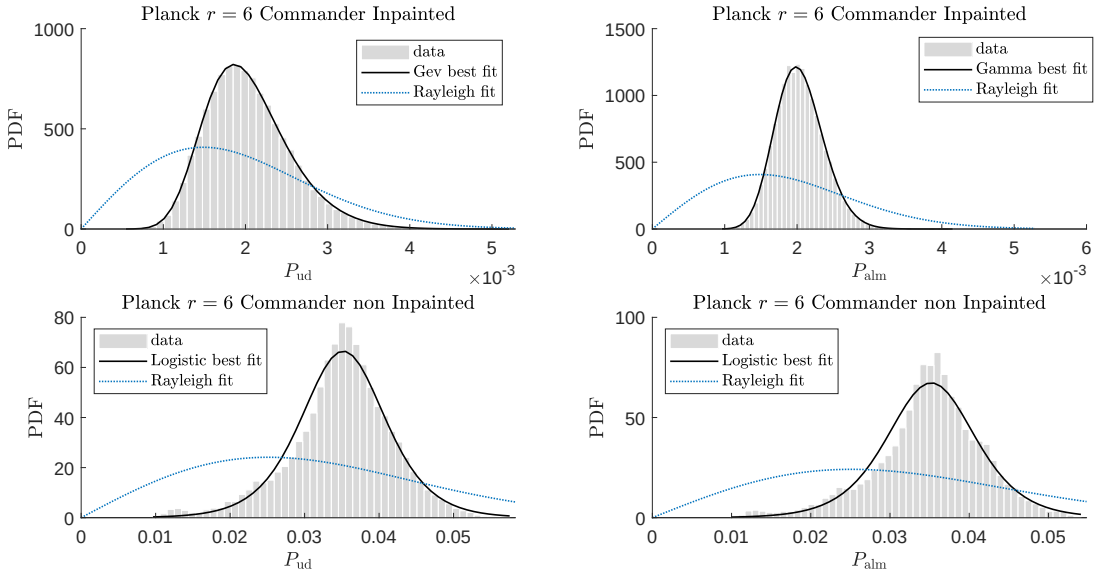


Figure 3: Normalised histograms of the directly degraded amplitudes $P_{\rm ud}$ and $P_{\rm alm}$ at resolution $r = 6$. Top row: inpainted maps; bottom row: non-inpainted maps. The Rayleigh reference clearly fails for these constructions.

be recomputed independently for each map and for each degradation strategy. The non-inpainted Commander maps with an explicit polarisation mask provide the most conservative reference for this statistical analysis, while the inpainted maps offer a useful robustness check.

These results establish the methodological basis for the next stage of the programme, namely the construction of angular correlation functions and CHSH-type combinations from CMB polarisation data. In that subsequent analysis, the low-resolution polarisation observable, the mask, and the associated dichotomisation threshold must be fixed a priori before any interpretation of possible Bell-type tests is attempted.

Acknowledgments

This work was funded by the Conselleria d'Educació, Universitats i Ocupació of the Generalitat Valenciana through project CIAICO/2024/284.

References

[1] R. Dale, R. Lapiedra, and J. A. Morales-Lladosa, "Cosmic primordial density fluctuations and bell inequalities," *Physical Review D*, vol. 107, p. 023 506, 2023. DOI: 10.1103/PhysRevD.107.023506

[2] R. Dale, J. M. Gandía, J. A. Morales-Lladosa, and R. Lapiedra, "Violation of bell inequalities from cosmic microwave background data," *Journal of Cosmology and Astroparticle Physics*, vol. 2025, no. 07, p. 044, 2025. DOI: 10.1088/1475-7516/2025/07/044 arXiv: 2502.13846.

[3] Planck Collaboration, "Planck 2018 results. IV. Diffuse component separation," *Astronomy & Astrophysics*, vol. 641, A4, 2020. DOI: 10.1051/0004-6361/201833881 arXiv: 1807.06208.

A geometric study on methods of temporary reparametrization in a two-body problem

José Antonio López Ortí[♭] ,[1] María José Martínez Usó[♮] , Francisco José Marco Castillo[♭]

(♭) Departamento de Matemáticas. Universitat Jaume I de Castellón
Av. Sos Baynat s/n, 12071 Castellón de la Plana, Spain.

(♮) Departamento de Matemática Aplicada. Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

The two-body problem is a well-known classical problem in celestial mechanics [1], [2], [3]. The equation of relative motion of the secondary body with respect to the primary body can be described by the equation:

$$\ddot{\vec{r}} = -\frac{mu}{r^3}\vec{r} = -\vec{\nabla}V, \quad V = \frac{mu}{r}.$$

This problem is completely integrable and includes, among others, constants such as the integral of areas, the integral of energy, and the Laplace vector [4], [5]. One of the problems that arises in the equation of motion is the irregularity of the origin, which leads to time reparameterizations [6], [7], [8] so that, by introducing a new time variable τ, it is possible to eliminate the singularity. As a first example of these transformations, we have the Sundman transformation, given by $dt = r\,d\tau$. The power of the transformation becomes clearer using a Hamiltonian formulation, which in our case is embodied in the Hamiltonian:

$$H = \frac{1}{2}(p_x^2 + p_y^2 + p_z^2) - \frac{\mu}{r},$$

where $p_X = \dot{x}$, $p_y = \dot{y}$, $p_z = \dot{x}$, $r = \sqrt{x^2 + y^2 + z^2}$ The motion is described by Hamilton's equations [9].

$$\dot{p_i} = -\frac{\partial H}{\partial q_i}, \qquad \dot{q} = \frac{\partial H}{\partial p}, \quad i = 1, 2, 3.$$

where $(x, y, z) = (q_1, q_2, q_3)$, $(p_x, p_y, p_z) = (p_1, p_2, p_3)$. Since time does not appear explicitly in the Hamiltonian, the energy is constant h. Hamilton's equations are invariant under the addition of a constant to the Hamiltonian, so we can replace H with $\hat{H} = H - h$, thus, the trajectory in phase space is contained in the manifold $\hat{H} = 0$. This consideration is important because a transformation in the time variable given by $dt = Q(r)d\tau$ reduces to transforming the Hamiltonian into $K = Q(r)\hat{H}$, thus:

$$\frac{dp_i}{d\tau} = -\frac{\partial H}{\partial q_i}, \quad \frac{dq_i}{d\tau} = \frac{\partial H}{\partial p_i}, \quad i = 1, 2, 3.$$

[1] lopez@mat.uji.es

Thus, for the Sundman transformation, we obtain

$$K = \frac{r}{2}(p_x^2 + p_y^2 + p_z^2) - \mu - h\, r,$$

, a Hamiltonian that is regular at the origin. It is also important to note that, for the Sundman transformation, when combined with a coordinate transformation, the problem can be reduced to a harmonic oscillator. If we study the problem in the planar case [10], and also eliminate the Hamiltonian constant, we obtain the following:

$$K = \frac{r}{2}(p_x^2 + p_y^2) - h\, r.$$

Introducing a Levi-Civita transformation $z = w^2$ where $z = x + i\ , y$, $w = \xi + i\,\eta$, we have $x = \xi^2 - \eta^2$, $y = 2\xi\,\eta$ which is a point transformation that in turn induces a symplectic diffeomorphism by simply equating the Liouville 1-form in both systems. With this nomenclature, we have $r = |z| = |w|^2$

$$p_x dx + p_y dy = p_\xi d\xi + p_\eta d\eta,$$

from which, by differentiating and operating, we obtain that

$$p_x = \frac{\eta p_\xi - \xi p_\eta}{2r}, \quad p_y = \frac{\eta p_\xi + 2p_{xi} p_\eta}{2r}$$

and therefore

$$\hat{K} = \frac{1}{8}(p_\xi^2 + p_\eta^2) - h(\xi^2 + \eta^2),$$

which corresponds to the Hamiltonian of a harmonic oscillator of mass $m = 2$ ($h < 0$). The three-dimensional case can also be reduced to the harmonic oscillator form by combining the Sundman time transformation with the Kustaanheimo–Stiefel transformation [11].

Furthermore, aside from the theoretical interest of these transformations, numerical integration of the problem using physical time as the integration variable presents difficulties in the elliptical case when the eccentricities are very large. This is because equal time intervals produce point distributions on the orbit with very high density in the apoapsis region and very low density in the periapsis region. In contrast, the problem dynamics require an inverse point distribution, since the velocity is much higher in the periapsis region than in the apoapsis region. The methods must also strive to include sufficient points in the regions of greatest curvature, which correspond to the apoapsis and periapsis regions.

To obtain a better distribution of the integration base points, an analytical regularization of the integration step can be used by means of a change of variable, of the form $dt = Q(r)d\tau$. In the case of the transformation, the introduction of a constant factor $\frac{1}{na}$, $ndt = \frac{r}{a}d\tau$ is equivalent to the relation $dM = \frac{r}{a}dg$, where a is the semi-axis of the ellipse, n the mean motion, M the mean anomaly, and g the eccentric anomaly. With this transformation, a distribution of points symmetric with respect to the axes of the ellipse is achieved.

The structure of this work is as follows. Section 1 introduces the problem and its general background. Section 2 presents a series of transformation families corresponding to elliptic motion, which, where possible, are extended to hyperbolic and parabolic motion. Section 3 uses certain geometric properties of curvature to extend the two-parameter family to all types of orbits in the two-body problem. Section 4 presents the main conclusions of the work.

2 Some families of temporary transformations in the two-body problem

In this section, we will describe a set of transformations commonly used in mechanics, initially designed for elliptical motion.

Elliptical motion is a periodic motion in which the mean anomaly M described earlier can be used as a time variable. This anomaly is obtained by simply setting the periapsis value to zero and rescaling its time variation using the mean motion n as a factor. In elliptical motion, the time variables are resolved so that they vary by $[0.2\pi]$ in one revolution, allowing us to develop the magnitudes of the two-body problem in a Fourier series as a function of these variables. This fact, in principle natural, introduces problems, as we will see, when transforming from elliptical to parabolic and hyperbolic motion, which can occur, for example, in the perturbed motion of a highly eccentric comet.

Among the most commonly used time variables is the true anomaly f, which is related to the mean anomaly by the relation

$$dM = \frac{r^2}{a^2\sqrt{1-e^2}}df,$$

and the Nacozy anomaly τ [7], defined as $dt = Cr^{3/2}d\tau$, where C is a constant that can be introduced to satisfy the periodicity condition.

All these anomalies are of the form $dM = C\,r^{\alpha}d\Psi$. Janin and Bond introduce a one-parameter family of equivalent time transformations in the elliptic case [12], $dM = C_{\alpha}\,r^{\alpha}d\Psi$, which they call the generalized Sundman transformation, where C_{α} is the normalization constant.

Furthermore, by denoting r' as the distance from the secondary lens to the secondary focus of the conic section, and by f' as the angle measured from the secondary focus between the direction of the secondary lens and the periapsis, a two-parameter family of variables $\Psi_{\alpha,\beta}$ appears, defined by the authors [13], including f' itself, called the antifocal anomaly by Fukushima [14], all the variables of the generalized Sundman family ($\beta = 0$), the reduced arc length s^* introduced by Brumberg [15], the elliptical anomaly [16] w introduced by Brumberg and Fukushima [16], and the semifocal anomaly. $\Psi = \frac{f+f'}{2}$ introduced by the authors [17]. Table 1 shows the parameters α and β that define these anomalies in the biparametric family.

Table 1: Representation of anomalies in the biparameric family

	M	g	f	τ^*	Ψ_{α}	s^*	w	f'	Ψ
α	0	1	2	1.5	α	0.5	1.5	1	2
β	0	0	0	0	0	-0.5	0.5	1	1

3 Extension to the parabolic and hyperbolic two-body problem

The anomalies in the previous section depend on a normalization factor $C_{\alpha,\beta}$, which generally depends on the semi-major axis and the eccentricity. In the limiting case of parabolic motion, this translates to $e = 1$ and the semi-major axis disappearing ($a = +\infty$), as does r'. The definition of the constants $C_{\alpha,\beta}$ also poses problems, even for the generalized Sundman family.

To avoid this problem, we will first define the transformations from a projective point of view, taking $dt = r^{\alpha}r'\beta$ as our representative. Secondly, we will relate the product $r\,r'$ to curvature, thereby extending the family to any case.

The radius vector with respect to a conic section in a conic section is given as a function of the mean anomaly and the parameter, as

$$r = \frac{1}{1+e\cos f}, \qquad r' = \frac{1}{1-e\cos f}, \quad \forall e \neq 1,$$

defined expression, except in the case of the parabola. The curvature of a conic section for $e \neq 1$, as a function of r, r', can be obtained through [18]

$$\kappa(\theta) = \frac{\left|r^2 + 2\dot{r}^2 - r\ddot{r}\right|}{\left(r^2 + (\dot{r}^2\right)^{3/2}},$$

where $\dot{r}$ denotes the derivative with respect to the angular variable. From this expression and after certain calculations, we obtain:

$$\kappa = \frac{(r\,r')^{3/2}}{p^2},$$

which relates the product of the radius vectors to the curvature. Thus, $r\,r' = p^{8/3}\kappa^{2/3}$, and substituting into the family biparametric.

$$r^{\alpha} r'^{\beta} = r^{\alpha-\beta}(r\,r')^{\beta} = p^{\frac{8\beta}{3}} r^{\alpha-\beta} \kappa^{\frac{2\beta}{3}},$$

thus eliminating constants, the variable $\Psi_{\alpha,\beta}$ is equivalent from a projective point of view to the variable $\Phi_{\gamma,\delta}$ defined as

$$dt = r^{\gamma}\kappa^{\delta} d\Phi_{\gamma,\delta}, \qquad \gamma = \alpha - \beta, \quad \delta = \frac{2\beta}{3},$$

a non-singular expression for the parabolic case, since it no longer depends on r'.

4 Concluding Remarks

The two-parameter family of time transformations defined for the elliptic case can be naturally extended to hyperbolic motion.

In the parabolic case, the secondary focus disappears, so, in principle, the family of transformations is meaningless.

The product $r\,r'$ is directly related to the curvature, making it possible to define an equivalent family for parabolic motion.

The new family must be understood in a projective sense, so constants are eliminated, which, in many cases, are infinite or zero in the parabolic case.

Acknowledgments

This work has been partially supported by grant GACUJIMB/2025/12 of University Jaume I of Castellón.

References

[1] D. Brouwer and G. Clemence, *Methods of Celestial Mechanics*. Academic Press: New York, NY, USA, 1961.

[2] Y. Hagihara, *Celestial Mechanics*. MIT Press: Cambridge, MA, USA, 1970.

[3] F. Tisserand, *Traité de Mécanique Céleste*. Gauthier-Vilars, Paris, 1894.

[4] D. Boccaletti and G. Puccaco, *Theory of Orbits: Perturbative and Geometrical Methods*. Springer, Berlin, 1999.

[5] G. Beutler, *Methods of Celestial Mechanics*. Springer, Berlin, 2010.

[6] K. Sundman, "Memoire sur le probleme des trois corps," *Acta Math.*, vol. 36, pp. 105–179, 1912.

[7] P. Nacozy, "The intermediate anomaly," *Celest. Mech.*, vol. 16, pp. 309–313, 1997.

[8] J. Cariñena, E. Martínez, and M. Muñoz, "Infinitesimal time reparametrisation and its applications," *J. Nonlinear Math. Phys.*, vol. 29, pp. 523–555, 2022.

[9] V. Arnold, *Mathematical Methods of Calsiccal Mechanics*. Springer, Berlin, 1978.

[10] J. Waldvogel, "A new regularization of the planar problem of three bodies," *Celest. Mech.*, vol. 6, pp. 221–231, 1972.

[11] P. Saha, "Interpreting the kustaanheimo-stiefel transform in gravitational dynamics," *Mon. Not. R. Astron. Soc.*, vol. 400, pp. 228–231, 2009.

[12] G. Janin and V. Bond, *The Elliptic Anomaly*. NASA Technical Memorandum 58228. Greenbelt, MD, USA, 1980.

[13] J. López, V. Agost, and M. Barreda, "The elliptic anomalya new bi-parametric family of temporal transformations to improve the integration algorithms in the study of the orbital motion," *J. Comput. Appl. Math.*, vol. 309, pp. 482–492, 2017.

[14] T. Fukushima, "Efficient orbit integration by the orbital longitude method uusing antifocal anomaly," *Astron. J.*, vol. 128, pp. 1455–1464, 2004.

[15] E. Brumberg, "Length of arc as independent argument for highly eccentric orbits," *Celest. Mech.*, vol. 53, pp. 323–328, 1992.

[16] E. Brumberg and T. Fukushima, "Expansions of elliptic motion based on elliptic functions theory," *Celest. Mech. Dyn. Astron.*, vol. 60, pp. 69–89, 1994.

[17] J. López, V. Agost, and M. Barreda, "Cmmse: Study of a new symmetric anomaly in the elliptic, hyperbolic, and parabolic keplerian motion," *Math. Methods Appl. Sci.*, vol. 46, pp. 1429–1442, 2023.

[18] P. do Carmo, *Differential Geometry of Curves and Surfaces. 2 Ed.* Dover Publications Inc, Mineola, NY, USA, 2017.

Part VIII

Uncertainty Quantification and Modelling

Nonuniform Polynomial Behavior for Random Discrete-Time Systems

Oana Albescu[b, 1], Larisa Elena Biris[γ], Traian Ceausu[γ]
Nicolae Marian Seimeanu[γ]

(b) Faculty of Mathematics and Computer Science, Transilvania University of Brasov, Romania
(γ)West University of Timisoara, Department of Mathematics,
Timisoara, Romania

1 Introduction

The study investigates nonuniform polynomial stability for linear random one-sided discrete-time systems within the framework of random dynamical systems. While exponential behavior has been extensively studied, polynomial growth and decay remain less explored, particularly in the random discrete setting, despite their relevance when exponential properties fail.

2 Methods

2.1 Lyapunov norms

We establish necessary and sufficient conditions for these properties and provide Lyapunov-type characterizations. In addition, we analyze the corresponding behavior of the associated adjoint and inverse systems. The results extend the existing theory and contribute to a better understanding of nonuniform dynamics in random discrete systems.

3 Results

The obtained results emphasize the relevance of polynomial asymptotic phenomena in situations where exponential behavior is not present and provide a framework for their systematic analysis.

4 Conclusions

We studied nonuniform polynomial stability and instability for linear random one-sided discrete-time systemsWe obtained necessary conditions describing the asymptotic behavior of the system. We also investigated the corresponding properties for the adjoint and inverse systems. These results contribute to the development of the theory

[1] oana.albescu@unitbv.ro

of polynomial asymptotic behavior in random discrete dynamical systems and provide a basis for further research in this area.

References

[1] L. E. Biriş, T. Ceauşu, I.-L. Popa, N. Seimeanu, On uniform polynomial splitting of variational nonautonomous difference equations in Banach spaces, An. Univ. Vest Timi¸s. Ser. Mat.-Inform. 58 (2) (2022), 22-37.

[2] R. Datko, Uniform asymptotic stability of evolutionary processes in a Banach space, {\it SIAM J. Math. Anal.} \textbf{3} (1972), 428–445.

[3] I.-L. Popa, T. Ceausu, M. Megan, Nonuniform power instability and Lyapunov sequences, Appl. Math. Comput., 247 (2014), 969--975.

[4] I.-L. Popa, T. Ceausu, M. Megan, On exponential stability for linear discrete-time systems in Banach spaces, Comput. Math. Appl., 63 (2012), 1497--1503.

[5] I.-L. Popa, Lyapunov functions for random semidynamical systems in terms of tempered exponential splitting, Math. Meth. Appl. Sci., 44 (15) (2021), 11923--11932.

Probabilistic Analysis of Linear Delay Difference Equations with Uncertain Parameters

A. Navarro-Quiles[♭1], J.-V. Romero[♭] and M.-D. Roselló[♭]

(♭) Instituto Universitario de Matemática Multidisciplinar
Universitat Politècnica de València, Spain

1 Introduction

Delay difference equations constitute an important class of mathematical models for describing discrete-time dynamical systems with memory. In contrast to standard difference equations, the future evolution depends not only on the current state but also on previous observations, allowing delayed responses to be naturally incorporated into the model. Such equations arise in numerous applications including population dynamics, epidemiology, economics, supply chain management and digital control.

In many real-world applications, however, the parameters governing the system are not known exactly. Experimental measurements, environmental variability and modelling assumptions introduce uncertainty that may significantly affect the predicted behaviour of the solution. For this reason, representing uncertain parameters by random variables provides a more realistic framework than purely deterministic models.

This work studies a first-order linear delay difference equation with an uncertain feedback coefficient. Rather than analysing individual sample trajectories, our objective is to quantify how uncertainty propagates through the system by deriving analytical expressions for the first two statistical moments of the solution. Since the fundamental solution can be expressed as a polynomial in the random parameter, both the expectation and the variance are obtained directly from its moments, avoiding repeated stochastic simulations.

Finally, the proposed methodology is illustrated through a numerical example in which the analytical results are compared with Monte Carlo simulations, showing that the moment-based approach provides an accurate and computationally efficient description of the stochastic dynamics.

2 A Linear Delay Difference Model

Consider the first-order delay difference equation

$$x_{n+1} = x_n + \alpha x_{n-\tau} + u_n, \qquad n \geq 0, \tag{1}$$

where $\tau \in \mathbb{N}$ denotes a fixed delay, $\alpha \in \mathbb{R}$ is a constant coefficient and $\{u_n\}_{n\geq 0}$ is a prescribed forcing sequence. The initial history is given by

$$x_n = h_n, \qquad n = -\tau, \ldots, 0. \tag{2}$$

[1] annaqui@upv.es

Equation (1) describes systems in which the current state depends simultaneously on the present configuration and on information recorded τ time steps earlier. Such delayed interactions appear in many applications, including biological systems with maturation periods, economic models with deferred responses and engineering processes affected by communication or processing lags. To analyse (1), we first consider the corresponding homogeneous problem,

$$x_{n+1} = x_n + \alpha x_{n-\tau}, \qquad n \geq 0. \tag{3}$$

Its dynamics are governed by the fundamental sequence $\{G_n\}_{n\geq -\tau}$ defined recursively as

$$\begin{cases} G_n = 1, & n = -\tau, \ldots, 0, \\ G_{n+1} = G_n + \alpha G_{n-\tau}, & n \geq 0. \end{cases} \tag{4}$$

Using this sequence, the solution of the homogeneous equation can be written as

$$x_n^{(h)} = G_n h_{-\tau} + \sum_{j=-\tau+1}^{0} G_{n-\tau-j}\,(h_j - h_{j-1}), \qquad n \geq -\tau. \tag{5}$$

The complete solution of the non-homogeneous problem is then obtained through the discrete variation-of-constants formula,

$$x_n = x_n^{(h)} + \sum_{j=0}^{n-1} G_{n-\tau-j-1}\, u_j. \tag{6}$$

An important feature of this representation is that each element of the fundamental sequence is a polynomial in the parameter α. Consequently, for every fixed time instant, the solution is itself a polynomial function of α. This observation provides the basis for the probabilistic analysis developed in the next section, where the parameter is assumed to be random.

3 Probabilistic Analysis

Suppose now that both the feedback coefficient and the initial history are uncertain. More precisely, let $\alpha : \Omega \to \mathbb{R}$ be a random variable and assume that $H_j : \Omega \to \mathbb{R}$, $j = -\tau, \ldots, 0$, represent the random initial history. Consequently, the solution of the model (1) becomes a stochastic process $\{X_n\}_{n\geq -\tau}$. From (5) and (6),

$$X_n = G_n(\alpha) H_{-\tau} + \sum_{j=-\tau+1}^{0} G_{n-\tau-j}(\alpha)(H_j - H_{j-1}) + \sum_{j=0}^{n-1} G_{n-\tau-j-1}(\alpha) u_j.$$

Since every function G_n is polynomial in α, the solution is a nonlinear transformation of the random vector

$$(\alpha, H_{-\tau}, \ldots, H_0).$$

Although the complete probability distribution can be obtained through standard transformation techniques, such an approach rapidly becomes computationally demanding as the dimension of the random vector increases. For this reason, we focus on the first two statistical moments of the solution.

Assuming that α is independent of the initial history,

$$\mathbb{E}[X_n] = \mathbb{E}[G_n(\alpha)]\mathbb{E}[H_{-\tau}] + \sum_{j=-\tau+1}^{0} \mathbb{E}[G_{n-\tau-j}(\alpha)]\mathbb{E}[H_j - H_{j-1}] + \sum_{j=0}^{n-1} u_j \mathbb{E}[G_{n-\tau-j-1}(\alpha)].$$

Moreover, since every G_n is a polynomial in α, its expectation can be computed directly from the moments of the random parameter,

$$\mathbb{E}[G_n(\alpha)] = \sum_{k=0}^{d_n} a_{k,n}\, \mathbb{E}[\alpha^k].$$

Similarly, the variance is obtained from

$$\mathrm{Var}(X_n) = \mathbb{E}[X_n^2] - \mathbb{E}[X_n]^2,$$

where the second-order moment depends only on mixed moments of the initial conditions together with the moments of α. These expressions provide an efficient alternative to Monte Carlo simulation, since they avoid repeatedly solving the delayed difference equation while still capturing the main statistical features of the solution process.

4 Numerical Example

To illustrate the proposed methodology, we consider the delay difference equation

$$x_{k+1} = x_k + b(\omega)\, x_{k-4}, \qquad k \geq 0,$$

where the delay is fixed to $m = 4$. The uncertainty is introduced through the model coefficient and the initial history. More precisely, the random coefficient is assumed to follow a uniform distribution,

$$b(\omega) \sim \mathcal{U}(-0.35, -0.15),$$

whereas the initial history is taken as

$$x_k(\omega) = z(\omega), \qquad k = -4, \ldots, 0,$$

with

$$z(\omega) = 0.6 + 0.8y(\omega), \qquad y(\omega) \sim \mathrm{Beta}(2, 5),$$

and $b(\omega)$ and $z(\omega)$ are assumed to be mutually independent.

Figure 1 illustrates the delayed exponential $e_m^{b,k}$ for several values of the delay parameter with $b = -0.25$ deterministic. As the delay increases, the oscillatory behaviour becomes more pronounced and the transient dynamics persist for longer times, highlighting the strong influence of the memory length on the solution.

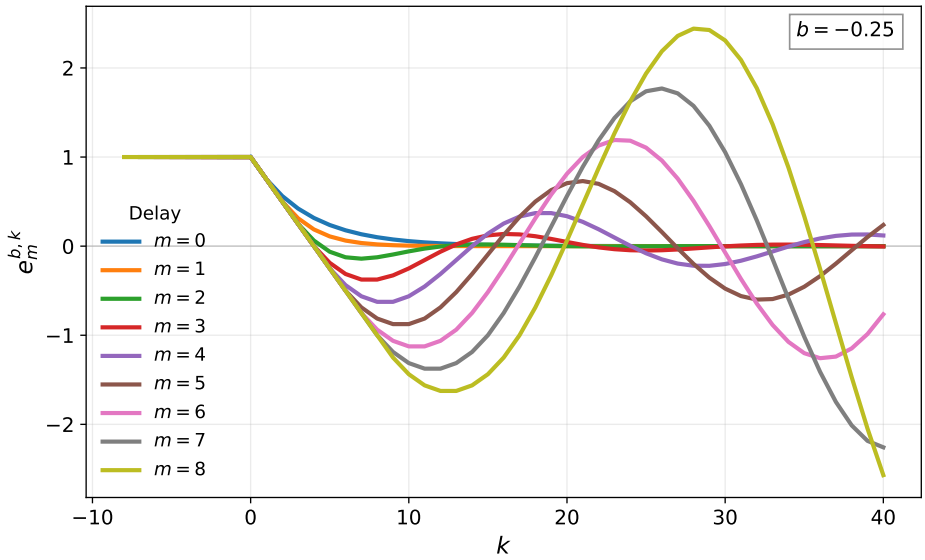
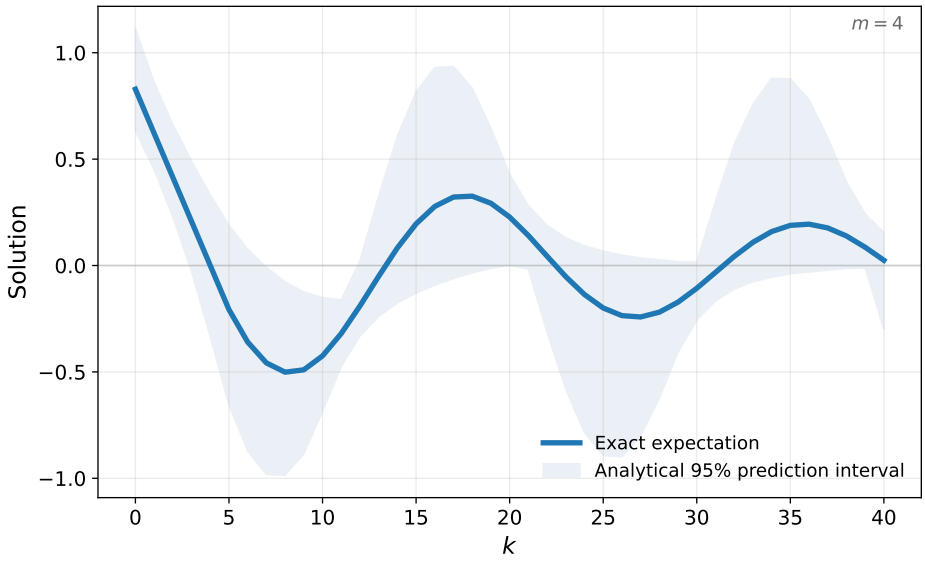
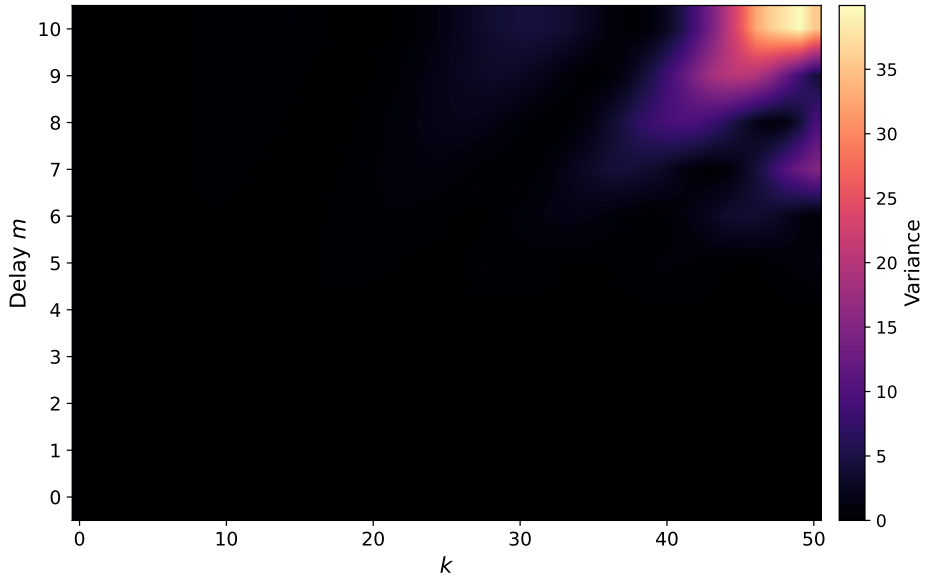


Figure 2 presents the exact expectation of the solution together with the analytical 95% prediction interval. The uncertainty propagates over time while preserving the oscillatory behaviour induced by the delay.



Finally, Figure 3 shows the variance of the solution for different values of the delay. Larger delays produce substantially higher levels of variability, particularly at later time instants, revealing that the interaction between randomness and memory effects amplifies uncertainty during the transient dynamics.



5 Conclusions

This work has developed a probabilistic analysis of linear delay difference equations with uncertain parameters. The proposed approach takes advantage of the polynomial representation of the delayed fundamental solution to derive analytical expressions for the expectation and variance of the stochastic solution without resorting to sampling-based techniques.

The numerical results confirm the accuracy of the proposed framework and illustrate the influence of the delay on uncertainty propagation. Whereas the mean solution preserves the oscillatory dynamics of the deterministic model, increasing the delay substantially amplifies the variance, highlighting the impact of memory effects on the stochastic response.

Future research will focus on extending the methodology to systems with multiple delays, random forcing terms and nonlinear delay difference equations, as well as on deriving analytical approximations for the full probability distribution of the solution.

Funding Acknowledgments

This work has been partially supported by the grant PID2025-172735NB-I00 funded by MICIU/AEI/10.13039/501100011033 and by the grant CIPROM/2026/22 funded by the Generalitat Valenciana (Conselleria de Educación, Cultura, Universidades y Empleo).

Bibliography

[1] J. Diblík and D. Ya. Khusainov. Representation of solutions of linear discrete systems with constant coefficients and pure delay. *Advances in Difference Equations*, 2006:80825, 2006.

[2] R. L. Graham, D. E. Knuth and O. Patashnik. *Concrete Mathematics*. Addison–Wesley, 2nd edition, 1994.

[3] L. M. Milne-Thomson. *The Calculus of Finite Differences*. Macmillan, London, 1933.

A Probabilistic Framework for Heat Transfer Models with Implicit Solutions

E. Morellà, R.-J. Villanueva and J.-C. Cortés

I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

Cooling and heating equations are a classical example of how the dynamics of physical systems can be modeled using differential equations. The exact solutions to these differential equations depend on a variety of parameters related to the material (e.g., heat capacity, density) and the environment (e.g., ambient temperature), and their explicit form is often unknown. In cases where the form is known, the standard method for determining the parameters is fitting based on experimental measurements. This approach is statistical in nature and is necessary because measurements exhibit inconsistencies, arising both from limitations of the experimental setup and from inherent imperfections in the measurement process. However, these sources of error are an intrinsic part of the experiment itself. The uncertainty in the parameter values can naturally be represented through probability distributions. In this context, a theoretical framework is proposed that incorporates parameter uncertainty into the model and allows for obtaining solutions for the probability distribution of temperature at any time, as a function of the distribution of the parameters. Moreover, the procedure only requires knowledge of an implicit form of the solution. To demonstrate the capabilities of the proposed method, computational experiments are conducted using synthetic data. These experiments show how to predict the temporal evolution of the temperature distribution, via the so called one-point probability distribution, based on a prior parametric distribution of the parameters and, alternatively, how to estimate the parameter distribution from samples (experimental measurements) of temperature at specific time instants.

2 Problem Statement

Let $x(t) \in \mathbb{R}$ be a time-dependent real-valued scalar quantity governed by a deterministic dynamical system, that can be measured by conducting an experiment. Assume that an implicit representation of the solution is known

$$h(x(t), \vec{\eta}, t) = 0, \tag{1}$$

where $\vec{\eta} = (\eta_1, \ldots, \eta_N) \in \mathbb{R}^N$ is a parameter vector, $h : \mathbb{R} \times \mathbb{R}^N \times \mathbb{R}^+ \to \mathbb{R}$ a differentiable map, that is injective in its first argument, as well as in at least one parameter argument, say the N-th, η_N. All previous conditions are satisfied by a general heat transfer model, growth models [1], and many autonomous differential equations.

While the dynamics of the deterministic dynamical system are completely characterized by the parameters, and the true evolution of $x(t)$ is deterministic, the system's state remains uncertain in practice. This uncertainty arises mainly because the exact values of the parameters $\vec{\eta}$ are unknown, and knowing them accurately and precisely is often infeasible. Consequently, the parameter vector $\vec{\eta}$ is best described as a random vector $\vec{H}$ with joint probability density function $p_{\vec{H}}(\vec{\eta})$. Accordingly, $X(t)$ becomes a *stochastic process* and $h(X(t), \vec{H}, t)$ an identically null deterministic quantity, or equivalently a random variable with a Dirac delta distribution centered at zero.

The main theoretical problem is to derive the distribution of the solution $X(t)$ at any given time for a known parameter distribution $p_{\vec{H}}(\vec{\eta})$, which is the probabilistic analogue of solving the governing differential equation for a specific set of parameters.

Conversely, the main practical problem is to estimate the unknown parameter distribution from experimental measurements, conceptually similar to a standard curve fitting, but with an enriched perspective, due to the consideration of the whole probability space. Our proposed method's core idea is to first parametrize the joint probability distribution of the uncertain parameters using *shape* parameters. In many applications, such as our heat transfer model, these parameters are mutually independent if defined wisely, which greatly simplifies the parameterization. We then find the shape parameters that best fit the data according to a distribution-aware objective function. In this work the choice for the parametrization is the product of beta distributions, an $\mathcal{L}_1$-norm based Cramér-von Mises-like metric for the objective function, and a Particle Swarm Optimization (PSO) for the search algorithm.

3 Theoretical framework

In this Section an expression for the probability density function, $X(t)$ of a stochastic process satisfying

$$h(X(t), \vec{H}, t) = 0, \quad \forall t > 0 \tag{2}$$

is obtained, where $\vec{H}$ is a vector of parameter-like random variables, $h : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ is injective with respect to its first argument and at least another parameter argument, say H_N. The expression is useful for computation when H_N is independent of $(H_1, \ldots, H_{N-1})$.

Let $Y(t)$ be the stochastic process given by $Y(t) = h(X(t), \vec{H}, t)$. The joint probability density function of $X, \vec{H}$ can be expressed in terms of $Y, \vec{H}$'s by the RVT (Random Variable Transformation) technique, which holds due to injectivity of h with respect to the first argument. For each $t \geq 0$

$$p_{X(t),\vec{H}}(x, \vec{\eta}) = \left| \frac{\partial h}{\partial x}(x, \vec{\eta}, t) \right| p_{Y(t),\vec{H}}(h(x, \vec{\eta}, t), \vec{\eta}). \tag{3}$$

Since $Y(t)$ is deterministically zero its distribution is a Dirac delta and it is independent of any

other random variable. Therefore,

$$p_{X(t),\vec{H}}(x,\vec{\eta}) = \left|\frac{\partial h}{\partial x}(x,\vec{\eta},t)\right| p_{\vec{H}}(\vec{\eta})\delta(h(x,\vec{\eta},t)), \tag{4}$$

Integrating with respect to $\vec{\eta}$

$$p_{X(t)}(x) = \int_{\mathbb{R}^N} \left|\frac{\partial h}{\partial x}(x,\vec{\eta},t)\right| p_{\vec{H}}(\vec{\eta})\,\delta(h(x,\vec{\eta},t))\,\mathrm{d}\vec{\eta}. \tag{5}$$

Choosing a parameter component η_N with respect to which h is monotonic, defining $u = h(x,\eta_1,\ldots,\eta_{N-1},\eta_N,t)$ and solving for $\eta_N = h^{-1}(x,\eta_1,\ldots,\eta_{N-1},u,t)$

$$\begin{aligned} p_{X(t)}(x) = \int_{\mathbb{R}}\cdots\int_{\mathbb{R}}\int_{\mathbb{R}} & \frac{\left|\frac{\partial h}{\partial x}(x,\eta_1,\ldots,\eta_{N-1},h^{-1}(x,\eta_1,\ldots,\eta_{N-1},u,t),t)\right|}{\left|\frac{\partial h}{\partial \eta_N}(x,\eta_1,\ldots,\eta_{N-1},h^{-1}(x,\eta_1,\ldots,\eta_{N-1},u,t),t)\right|} \\ & p_{\vec{H}}(\eta_1,\ldots,\eta_{N-1},h^{-1}(x,\eta_1,\ldots,\eta_{N-1},u,t))\delta(u)\,\mathrm{d}u\,\mathrm{d}\eta_{N-1}\,\ldots\,\mathrm{d}\eta_1. \end{aligned} \tag{6}$$

Define $\eta_N^\star = h^{-1}(x,\eta_1,\ldots,\eta_{N-1},0,t)$ so that

$$p_{X(t)}(x) = \int_{\mathbb{R}}\cdots\int_{\mathbb{R}} \frac{\left|\frac{\partial h}{\partial x}(x,\eta_1,\ldots,\eta_{N-1},\eta_N^\star,t)\right|}{\left|\frac{\partial h}{\partial \eta_N}(x,\eta_1,\ldots,\eta_{N-1},\eta_N^\star,t)\right|}\cdot p_{\vec{H}}(\eta_1,\ldots,\eta_{N-1},\eta_N^\star)\,\mathrm{d}\eta_{N-1}\,\ldots\,\mathrm{d}\eta_1. \tag{7}$$

Since H_N is assumed to be independent of $H_1,\ldots,H_{N-1}$, then the previous equation can be simplified to

$$p_{X(t)}(x) = \mathbb{E}_{H_1,\ldots,H_{N-1}}\left[\frac{\left|\frac{\partial h}{\partial x}(x,\eta_1,\ldots,\eta_{N-1},\eta_N^\star,t)\right|}{\left|\frac{\partial h}{\partial \eta_N}(x,\eta_1,\ldots,\eta_{N-1},\eta_N^\star,t)\right|}\cdot p_{H_N}(\eta_N^\star)\right], \tag{8}$$

where the expectation is taken with respect to the joint probability density.

4 Heat transfer models

In order to exhibit the capabilities of the proposed framework we present a detailed application to a physical model for heat transfer.

The evolution of the temperature of an object interacting with its environment has been long studied. In that regard, Newton's well-known Law of cooling states that an object's rate of cooling is directly proportional to the temperature difference with its medium. It is broadly utilized as an approximation under the condition that the temperature difference is small. However, in the case of forced convection —where the heat is transferred from an externally produced current of matter, e.g. a breeze, as in Newton's original experiment— the Law works particularly well.

In contrast, radiative cooling, which consists of heat loss through the emission of light, is determined by Stefan's law (derived by Boltzmann). According to which the energy radiated per unit area and unit time is proportional to the fourth power of the temperature. Additionally, there may be other sources of heat in the system either internal steaming for example from an electrical

current and described by Joule's Law or external such as the incidence of sunlight. The total heat the system exchanges with its environment per unit area and time is [2]

$$q_{\text{tot}} = -\underbrace{\epsilon\sigma(T^4 - T_e^4)}_{q_{\text{rad}}} - \underbrace{h(T - T_v)}_{q_{\text{conv}}} + \phi, \tag{9}$$

, where $\sigma = 5.67 \times 10^{-8} \frac{\text{W}}{\text{m}^2 \cdot \text{K}^4}$ is the Stefan-Boltzmann constant, ϵ is the emissivity, T_e is the ambient temperature, h the heat transfer coefficient, T_v is the temperature of the current, and ϕ is the aggregation of all other heat sources.

The effect of the transferred heat on the temperature is determined by the specific heat capacity c, which depends on the concrete material the object is made of, and its mass m

$$mc\frac{\mathrm{d}T}{\mathrm{d}t} = Aq_{\text{tot}}, \tag{10}$$

where A is the surface area.

Finally, the system is governed by a first-order autonomous differential equation

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{A}{mc}\left(-\epsilon\sigma(T^4 - T_e^4) - h(T - T_v) + \phi\right), \quad T(0) = T_0. \tag{11}$$

The parameters are assumed to be temperature independent over the experiment's range of temperatures. The values of the thermodynamic parameters (emissivity, specific heat capacity, heat transfer coefficient, etc.) are tabulated. However, different conditions result in radically different values, leading to a high degree of uncertainty.

For instance, h is highly dependent on whether the current flow is turbulent or laminar. In reality, the flow may not be completely laminar nor turbulent. Similarly, different materials have different emissivities, making its determination particularly hard for heterogeneous materials. Even a single material has vastly different emissivities depending on the level of oxidation in the surface. Since it is impossible to tabulate every single combination of conditions, most real systems remain uncertain.

Other sources of uncertainty may stem from the heat sources and the initial conditions —or the initial time, however they can be unified in a single uncertain term.

Plausible distributions for the parameters ϵ, h, ϕ, T_0 are modeled on a case to case basis.

5 Proposed Probabilistic Framework

5.1 Forward Uncertainty Propagation

There exist two fundamental quantities of interest in the heat transfer model, both described by implicit equations: the transient, i.e the solution to the differential equation, which at any given time is a random variable; and the steady-state, which is a global equilibrium of the model and is also a random variable.

For the steady-state note that it can be written as $g(T_{\text{eq}}, \epsilon, h, \phi) = 0$, where g is the function

given by

$$g(T, \epsilon, h, \phi) = -\epsilon\sigma(T^4 - T_e^4) - h(T - T_v) + \phi. \tag{12}$$

Note that the previous expression is independent of the initial condition, since the equilibrium is global. The probability distribution of the solution is obtained using (8) with ϕ as the *special* parameter η_N

$$p_{T_{\text{eq}}}(x) = \mathbb{E}_{E,H}\left[(-4\epsilon\sigma x^3 - h) \cdot p_\Phi(\phi^\star)\right], \quad \phi^\star = \epsilon\sigma(x^4 - T_e^4) + h(x - T_v) \tag{13}$$

Similarly, for the transient define the function f so that $f(T(t), \epsilon, h, \phi, T_0, t) = 0$ is the solution to the differential equation (11). Then f is given by

$$f(T, \epsilon, h, \phi, T_0, t) = \int_{T_0}^{T} \frac{\mathrm{d}u}{g(u, \epsilon, h, \phi)} - \frac{A}{mc}t. \tag{14}$$

The first probability density function of the temperature can be obtained using (8) with T_0 as the *special* parameter η_N

$$p_{T(t)}(x) = \mathbb{E}_{E,H,\Phi}\left[\frac{|g(T_0^\star, \epsilon, h, \phi)|}{|g(x, \epsilon, h, \phi)|} \cdot p_{\mathcal{T}_0}(T_0^\star)\right], \tag{15}$$

where $T_0^\star$ is given implicitly by solving

$$\int_{T_0^\star}^{T} \frac{\mathrm{d}u}{g(u, \epsilon, h, \phi)} = t. \tag{16}$$

5.2 Backward Parameter Calibration

In order to infer distributions for the uncertain parameters from experimental data, we propose to use an optimization approach. in order for it to be computationally plausible the search space needs to be drastically reduced. Therefore, the distributions of the uncertain parameters are parametrized using appropriate families of distributions. For instance, since for heat transfer models ranges for the parameters can be given, we employ beta distributions. In other models, lognormal distributions can be used for strictly positive random variables, mixtures for multi-modal distributions, etc.

Moreover, the objective function should take into account the whole shape of the distributions, for that reason we employ a Cramér-von Mises-based metric [3]. Suppose the available data are sets of N_x measurements taken at N_t different times

$$\{(t_i, x_i^{(j)})\}_{j=1}^{N_x}, \quad i = 1, \dots N_t \tag{17}$$

then the metric is constructed as the average of the different $\mathcal{L}_1$-Cramér-von Mises statistics

$$\frac{1}{N_t}\sum_{i=1}^{N_t}\int_{-\infty}^{\infty} |F_i(x) - F_{T(t_i)}(x)|\, p_{T(t_i)}(x)\mathrm{d}x, \tag{18}$$

where $p_{T(t_i)}(x), F_{T(t_i)}(x)$ are the predicted (using a forward uncertainty propagation step) proba-

bility and cumulative distribution functions, respectively; and $F_i(x)$ is the empirical distribution of the data at t_i.

Funding Acknowledgments

This work has been partially supported by the grant PID2025-172735NB-I00 funded by MICIU/AEI/10.13039/501100011033 and by the grant CIPROM/2026/22 funded by the Generalitat Valenciana (Conselleria de Educación, Cultura, Universidades y Empleo).

References

[1] F. Rodríguez, F. Z. Lachachi, I. Medjahdi, M. Á. Castro, and J. C. Cortés, "A note on the application of the rvt method to general classes of single-species population models formulated by random differential equations," Computational and Applied Mathematics, vol. 43, no. 5, p. 286, 2024, ISSN: 1807-0302. DOI: 10.1007/s40314-024-02809-0.

[2] J. HOLMAN, Heat Transfer, 10-th Edition. New York: McGraw-Hill, 2010.

[3] F. Schmid and M. Trede, "An L1-variant of the Cramer-von Mises test," Statistics & Probability Letters, vol. 26, no. 1, pp. 91–96, 1996, ISSN: 0167-7152. DOI: https://doi.org/10.1016/0167-7152(95)00256-1.

Impact of Light-Cycle duration on Algal Biomass Production: A Random PSF Model

J-C. Cortés[♭], A. Navarro-Quiles[♭], Š. Papáček[♮] and C.-L. Pérez[♭]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

(♮) Institute of Information Theory and Automation, Czech Academy of Sciences,
Prague, Czech Republic.

1 Introduction

Optimizing the cultivation of microalgae is crucial because of their wide array of uses, which span the pharmaceutical sector, the production of high-quality animal feed, and their role as biological agents for capturing urban CO_2 [1]. Nevertheless, large-scale microalgal cultivation in photobioreactors remains challenging due to elevated operational costs. Consequently, the optimization of microalgal growth rates is very important for achieving economically viable and financially sustainable production systems with photobioreactors.

Central to this optimization question is understanding how light availability, which is the primary driver of photosynthesis, affects the microalgae. While traditional approaches utilize the deterministic Photosynthetic Factory (PSF) model, experimental evidence suggests that light/dark cycles are not constant [2], [3]. Instead, the inherent randomness of hydrodynamic mixing and other physical processes introduces a level of uncertainty in the duration of these light cycles that is unavoidable in industrial settings.

In this research, we are extending the deterministic PSF model into a random framework in order to better model the effects of variability of the light cycles within a photobioreactor. By modelling the light/dark cycle duration as a random variable we will move from a dynamic model governed by a system of ordinary differential equations (ODEs) to a system of random differential equations (RDEs). We derive the first order probability density function (1-PDF) of the solution for any given time point and from it the PDF of some indicators of growth within the photobioreactor. Through supporting numerical examples for *Porphyridium* sp., we show that the efficiency of algal biomass production will be affected even by small variation in light-dark cycle length.

2 Model

2.1 Deterministic Model: The Photosynthetic Factory

This work extends the Photosynthetic Factory (PSF) framework of Eilers and Peeters [4] which we briefly explain in this section.
The model describes the state of a photosynthetic unit (PSU) using three dimensionless fractions:

resting (x_{R}), active (x_{A}), and inhibited (x_{B}). The transition dynamics are governed by the following mechanisms:

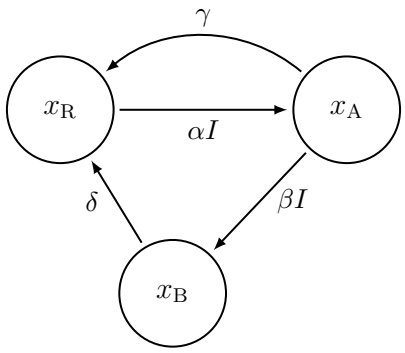


Figure 1: Compartmental diagram of the three-state PSF model. Transitions driven by irradiance I are photon-dependent, while γ and δ represent intrinsic kinetic rates.

2.2 Deterministic Model: System Dynamics and Re-parameterization

The specific growth rate, $\mu = \dot{c}_x/c_x$, is linked to the internal PSU states by the flux from the active to the resting state [4]:

$$\mu = \kappa\gamma x_{\mathrm{A}}, \tag{1}$$

where κ is a dimensionless proportionality constant. Utilizing the constraint $x_{\mathrm{R}} + x_{\mathrm{A}} + x_{\mathrm{B}} = 1$, the system is reduced to two states $(x_{\mathrm{A}}, x_{\mathrm{B}})$ and re-parameterized following [5] to isolate steady-state effects from kinetic dynamics allows us to re-write the PSF model as

$$\frac{1}{q_4}\begin{bmatrix}\dot{x}_{\mathrm{A}}\\ \dot{x}_{\mathrm{B}}\end{bmatrix} = \begin{bmatrix}-q_2(1+q_5) & 0\\ 0 & -\frac{q_5}{q_2(1+q_5)}\end{bmatrix}\begin{bmatrix}x_{\mathrm{A}}\\ x_{\mathrm{B}}\end{bmatrix} + u(t)\begin{bmatrix}-(1+q_5) & -1\\ q_5 & 0\end{bmatrix}\begin{bmatrix}x_{\mathrm{A}}\\ x_{\mathrm{B}}\end{bmatrix}, \tag{2}$$

where $u(t) = I(t)/q_1$ is the normalized irradiance. Parameters q_1, q_2, q_3 define the steady-state production curve, while q_4 and q_5 represent the timescales of light reactions and photoinhibition, respectively (Table 1).

Param.	Unit	Exp. Value [6]
q_1	$\mu\mathrm{E\,m^{-2}\,s^{-1}}$	2.50×10^{2}
q_2	–	3.02×10^{-1}
q_3	$\mathrm{s^{-1}}$	1.77×10^{-4}
q_4	$\mathrm{s^{-1}}$	4.83×10^{-1}
q_5	–	3.00×10^{-4}

Table 1: Re-parameterized PSF model values for *Porphyridium* sp.

2.3 Reduced Model under Intermittent Light

In industrial bioreactors, mechanical mixing and the Lambert-Beer law render light exposure intermittent. This is typically modeled as a square wave $u(t)$ with period h:

$$u(t) = \begin{cases} u_a, & t \in [0, rh) \\ u_b, & t \in [rh, h) \end{cases} \tag{3}$$

where $r \in (0,1)$ is the duty cycle. To bypass the stiffness of the full ODE system, a fast-timescale reduction as outlined in [5] approximates the dynamics of the active state x_{A} for sufficiently fast cycling ($h < e^4$):

$$\dot{x}_{\mathrm{A}}(t) = -q_4 \left[u(t) + q_2\right] x_{\mathrm{A}}(t) + q_4 u(t) \left[1 - x_{\mathrm{B}}^{\mathrm{ss}}(u_{\mathrm{av}})\right] \tag{4}$$

where $x_{\mathrm{B}}^{\mathrm{ss}}$ is the steady-state inhibited fraction at average irradiance u_{av}.

The solution $x_{\mathrm{A}}(t)$ converges to a unique h-periodic orbit as proved in [7], [8]. For optimization purposes, we focus on the cycle-averaged growth rate $\mu(h)$ and the relative productivity $p(h)$:

$$\mu(h) = \frac{\kappa\gamma}{h} \int_{Kh}^{(K+1)h} x_{\mathrm{A}}(t)\, dt, \quad p(h) = \frac{\mu(h)}{\mu_o(u_{\mathrm{av}})} \tag{5}$$

where $\mu_o(u_{\mathrm{av}})$ is the growth rate under a constant irradiance equivalent to the cycle average. This formulation allows for the efficient evaluation of growth performance under the stochastic light regimes typical of large-scale production. A productivity of 1 menas that the system is performing under intermittent light just as effectively as under the industrially unachievable condition of continuous light. A value greater than 1 means that it outperforms the continuous light regime which has been shown to be possible in flashing-light experiments [9].

2.4 Random Reduced PSF Model

Experimental evidence suggests that light/dark cycle frequencies f are variable due to hydrodynamic mixing which is complex to model. We define the cycle length as a random variable $H = 1/f$ on a probability space $(\Omega, \mathcal{F}_\Omega, \mathbb{P})$.

The randomized irradiance $u(t;H)$ becomes a time-modulated dichotomous input. The resulting RDE for the active state X_{A} for a given time t is:

$$\dot{X}_{\mathrm{A}}(t;H) = -q_4[u(t;H) + q_2]X_{\mathrm{A}}(t;H) + q_4 u(t;H)[1 - x_{\mathrm{B}}^{\mathrm{ss}}(u_{\mathrm{av}})] \tag{6}$$

where the average irradiance u_{av} remains deterministic.

Using the Random Variable Transformation (RVT) method [10], we derive the 1-PDF of $X_{\mathrm{A}}(t)$. Assuming independence between the initial condition X_0 and H, the 1-PDF is given by the expectation:

$$f_{X_A}(x,t) = \mathbb{E}_H \left[f_{X_0} \left(x\nu(t;H) - \mathcal{I}(t;H) \right) \nu(t;H) \right] \tag{7}$$

where $\nu(t;H)$ is the integrating factor and $\mathcal{I}(t;H)$ represents the accumulated light-driven transitions. To ensure computational efficiency, $\mathcal{I}(t;H)$ is evaluated by partitioning the integral into $J = \lfloor t/H \rfloor$ complete cycles and a remaining fractional period, exploiting the piecewise-constant structure of the irradiance signal.

2.5 Randomized Average Growth Rate

Long-term system performance is characterized by the random average growth rate $M(r;H)$. In the quasi-steady regime ($t \geq s$), the influence of the initial condition X_0 vanishes, and variability in M is driven by the random cycle length H. Analogous to the deterministic case, M is defined as:

$$M(r;H) := \frac{\kappa\gamma}{H} \int_{KH}^{(K+1)H} X_A(t,r;H)\, \mathrm{d}t, \tag{8}$$

where K is an integer ensuring the interval lies within the quasi-steady regime.

To derive the PDF of M, we apply the RVT method. Given the inverse mapping of the growth rate relation, the PDF $f_M(\mu)$ can be expressed in expectation form:

$$f_M(\mu) = \mathbb{E}_H \left[f_{X_0}\left(\frac{1}{I_2(H)} \left(\frac{H\mu}{\kappa\gamma} - I_1(H) \right) \right) \left| \frac{H}{\kappa\gamma I_2(H)} \right| \right], \tag{9}$$

where $I_1(H)$ and $I_2(H)$ are integrals of the model solution over one cycle. This formulation reveals that while intracellular dynamics act as a fast stabilizing mechanism, the long-term dispersion in biomass productivity is primarily controlled by the statistics of the external hydrodynamic environment.

To illustrate the framework, we perform numerical simulations in *Mathematica* using 5,000 samples for X_0 and H. We assume an initial state near darkness, $X_0 \sim U(0.001, 0.01)$, and evaluate the normalized random productivity:

$$P(H, r) := \frac{M(H, r)}{\mu_o(u_{\text{av}})}, \tag{10}$$

with a fixed light-fraction $r = 0.5$. The cycle length H is modeled using three distinct distribution families, all centered at the experimental mean $h_e = 45\,\text{s}$ with equivalent variance to isolate the effect of the distribution shape on growth performance:

$$H \sim \mathcal{D}, \quad \text{where } \mathcal{D} \in \begin{cases} \mathcal{U}(h_e - \delta, h_e + \delta), & \delta = \sqrt{0.15 h_e} \\ \mathcal{N}(h_e, \sigma), & \sigma = 0.05 h_e \\ \frac{1}{2}\sum \mathcal{N}(h_e \pm \sigma_b, 1) & \end{cases} \quad \text{with } h_e = 45. \tag{11}$$

3 Results

The distribution of normalized productivity P is sensitive to the functional form of H outlines in (11), even when the first two moments are identical. For all test cases, we have:

$$\mathbb{E}[P] \approx 0.6738, \quad \mathbb{V}[P] \approx 1.223 \times 10^{-5}$$

However, the non-equivalence of the resulting distributions is confirmed by the extremely low p-values in Table 2. The Cramér–von Mises tests provide strong evidence to reject the null hypothesis that the productivity samples originate from the same distribution.

Comparison	**Cramér–von Mises (p)**
Uniform vs Normal	3.88×10^{-13}
Uniform vs Bimodal	1.56×10^{-3}
Normal vs Bimodal	4.87×10^{-6}

Table 2: Pairwise statistical tests for the distributions of P. In all cases, the null hypothesis (identical distributions) is rejected.

These findings underscore that higher-order characteristics of the light-cycle duration, such as modality and tail behavior, are important for system performance analysis. Consequently, accurate productivity optimization in industrial photobioreactors requires precise experimental characterization of the light-regime PDF, as mean values alone are insufficient to capture the dynamics of the system.

4 Conclusions

In this work, we have investigated the growth dynamics of microalgae under light/dark cycles whose frequency is modeled as a continuous random variable. We did so, by extending the deterministic Photosynthetic Factory (PSF) model in a Random Differential Equation (RDE) framework. This approach allowed us to derive the probability density functions (PDFs) for key system states, such as the proportion of active photosynthetic units at any given time, and also for the resulting average growth rate and relative productivity. These results suggest that future experiments with a detailed study of hydrodynamic mixing and perceived light regimes are important for the optimization of microalgal cultures in photobioreactors. A realistic construction of the PDF of H is essential for transitioning from idealized models to predictive tools for real-world photobioreactors.

Acknowledgments

Cristina Luisovna Pérez has been awarded an FPU21/02343 grant as well as EST/00567 from the Ministry of Universities.
This work has been supported by the grant PID2025-172735NB–I00 granted by the research project MCIN/AEI/10.13039/501100011033, and by the research project CIGE/2023/79 (Generalitat Valenciana).

References

[1] S. N. Chanquia, G. Vernet, and S. Kara, "Photobioreactors for cultivation and synthesis: Specifications, challenges, and perspectives," *Engineering in Life Sciences*, vol. 22, no. 12, pp. 712–724, 2022. DOI: https://doi.org/10.1002/elsc.202100070. eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/elsc.202100070. [Online]. Available: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/elsc.202100070.

[2] S.-K. Wang, A. R. Stiles, C. Guo, and C.-Z. Liu, "Microalgae cultivation in photobioreactors: An overview of light characteristics," *Engineering in Life Sciences*, vol. 14, no. 6, pp. 550–559, 2014. DOI: 10.1002/elsc.201300170.

[3] A. Chiarini and M. Quadrio, "The light/dark cycle of microalgae in a thin-layer photobioreactor," *Journal of Applied Phycology*, vol. 33, pp. 183–195, 2021. DOI: 10.1007/s10811-020-02310-1.

[4] P. Eilers and J. Peeters, "A model for the relationship between light intensity and the rate of photosynthesis in phytoplankton," *Ecological Modelling*, vol. 42, pp. 199–215, 1988.

[5] S. Papáček, S. Čelikovský, B. Rehák, and D. Štys, "Experimental design for parameter estimation of two time-scale model of photosynthesis and photoinhibition in microalgae," *Mathematics and Computers in Simulation*, vol. 80, pp. 1302–1309, 2010.

[6] X. Wu and J. Merchuk, "A model integrating fluid dynamics in photosynthesis and photoinhibition processes," *Chemical Engineering Science*, vol. 56, pp. 3527–3538, 2001.

[7] S. Papáček, S. Čelikovský, D. Štys, and J. J. Ruiz-León, "Bilinear system as a modelling framework for analysis of microalgal growth," English, *Kybernetika*, vol. 43, no. 1, pp. 1–20, 2007.

[8] S. Celikovský, S. Papáček, A. Cervantes-Herrera, and J. Ruiz-Leon, "Singular perturbation based solution to optimal microalgal growth problem and its infinite time horizon analysis," *IEEE Transactions on Automatic Control*, vol. 55, no. 3, pp. 767–772, 2010. DOI: 10.1109/TAC.2010.2040498.

[9] S. Abu-Ghosh, D. Fixler, Z. Dubinsky, and D. Iluz, "Flashing light in microalgae biotechnology," *Bioresource technology*, vol. 203, pp. 357–363, 2016.

[10] T. Soong, *Random Differential Equations in Science and Engineering*. USA: Academic Press, 1973.

Characterization of the Unstationnary Stochastic Response of Mechanical Systems Using a Hilbertian–Bayesian Approach

Carol Santos Almonte[♭], Oscar Sanchez Jimenez[♭], Eduardo Souza de Cursi[♭1] and Emmanuel Pagnacco[♭]

(♭) INSA Rouen Normandie, Normandie Univ,
LMN UR 3828, F-76000, Rouen, France.

1 Abstract

This work presents a methodology for characterizing the unstationnary stochastic response of mechanical systems subject to uncertainty, arising from external loads or variability in their components. The approach covers both Gaussian and non-Gaussian scenarios, as well as stationary and non-stationary cases. The main objective is to construct the temporal evolution of the marginal probability density function (PDF). To achieve this, a functional expansion in a Hilbert space is employed, where the PDF is represented as a truncated series of orthogonal polynomials. To determine the coefficients of this expansion, two approaches are compared: a least-squares method (deterministic collocation) and a Bayesian inference approach. In the latter, the posterior distribution is computed from a likelihood function (defined through a Lagrangian enforcing the mean and variance) and two variants of prior distributions (either via Bootstrap resampling or via optimal moment projection).

The accuracy of the methodology is validated through benchmarks with exact analytical solutions, including modulated Gaussian processes, mixtures of processes, the non-central Chi-square distribution [1], and the harmonic vibration of an Euler–Bernoulli beam [2]. The quality of the reconstruction is quantified using the Mean Squared Error (MSE). The results confirm that these metamodels capture complex dynamics with high fidelity while requiring significantly fewer simulations than the Monte Carlo method.

References

[1] Y.-Y. Weng, T. Liu, X.-Y. Zhang, and Y.-G. Zhao, "Probability density estimation of polynomial chaos and its application in structural reliability analysis," *Reliability Engineering & System Safety*, vol. 253, p. 110 537, 2025, ISSN: 0951-8320.

[2] Y. Zhou and J. Li, "Analytical solution of the generalized density evolution equation for stochastic systems: Euler-bernoulli beam under noisy excitations and nonlinear vibration of kirchhoff plate," *Structural Safety*, vol. 117, p. 102 619, 2025, ISSN: 0167-4730.

[1] eduardo.souza@insa-rouen.fr

Probabilistic Analysis of an RC Circuit for a Stochastic Delta-Correlated Process

Juan-Carlos Cortés[♭], José-Vicente Romero[♭], María-Dolores Roselló[♭] and Joaquin F. Valencia Sullca[♭],[1]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

It is well known that many important physical systems in engineering are modeled using first-order linear differential equations. As the parameters of such equations need to be set from measurements or other procedures that often involves errors, it is natural to consider them as random variables or stochastic processes rather than numbers or functions, respectively. This approach leads to the formulation of random differential equations (RDEs) [1]-[2]. In this contribution, we study the following first-order linear differential equation, that describes the voltage of an RC circuit,

$$\begin{aligned} \frac{\mathrm{d}V(t)}{\mathrm{d}t} &= -\frac{V(t)}{RC} + \frac{V_s}{RC} + \frac{\alpha C(t)}{RC}, \quad t > 0, \\ V(0) &= V_0, \end{aligned} \tag{1}$$

i.e.

$$\begin{aligned} V^{'}(t) &= rV(t) + \beta C(t) + b, \quad t > 0, \\ V(0) &= V_0, \end{aligned} \tag{2}$$

where $V(t)$ denotes the voltage of the circuit at the time instant $t > 0$, $R > 0$ is the resistance, $C > 0$ is the capacitance, V_0 is the initial voltage, and V_s represents an external voltage exciting the circuit. For simplicity, we introduce the deterministic parameters $r = -\dfrac{1}{RC}$, $\beta = \dfrac{\alpha}{RC}$ and $b = \dfrac{V_s}{RC}$ into Eq. (2). Also, V_0 is a random variable; and $C(t)$ represents a stochastic force

$$C(t) = \sum_{i=1}^{N(t)} F_i \delta(t - t_i), \qquad t > 0, \tag{3}$$

where F_i stand for independent and identically distributed random variables, representing concentrated voltage that allows modifying voltage level of an RC circuit acting at the time $t_i \in (0, t)$, $N(t)$ denotes a Poisson counting process with rate $\lambda > 0$ [3], which is the expected number of random voltage applied, F_i, that randomly apply per unit of time and $\delta(t - t_i)$ denotes the Dirac delta function [4].

[1] joavasu2@teleco.upv.es

As $\delta(t-t_i)=0$ when $t<t_i$, Eq. (3) can be replaced by

$$C(t)=\sum_{i=1}^{N(T)} F_i\delta(t-t_i), \qquad \forall T\geq t>0. \tag{4}$$

The objective of this paper is to present a probabilistic analysis of an RC circuit with additive delta-correlated processes from synthetic data obtained from the deterministic model of the RC circuit, in order to obtain the main statistics of the stochastic process solution via the first moments, including the variance function.

2 Probabilistic Model Study

The Dirac delta function represents a concentrated impulse acting at time t_i. Thus, the stochastic process $C(t)$ can also be written as

$$C(t)=\sum_{i=1}^{N(t)} F_i\delta(t-t_i)=\sum_{i=1}^{N(t)} F_iR_{-1}(t-t_i), \quad t>0, \tag{5}$$

where we define

$$R_{-1}(t-t_i):=\delta(t-t_i). \tag{6}$$

Integrating Eq. (5), we obtain

$$g_0(t):=\int_0^t C(s)\,\mathrm{d}s=\sum_{i=1}^{N(t)} F_iR_0(t-t_i), \quad t>0, \tag{7}$$

where $R_0(t-t_i)$ denotes the Heaviside (unit step) function.

The mean function of $g_0(t)$ can be written by

$$\mathbb{E}\left\{g_0(t)\right\}=\mathbb{E}\left\{\sum_{i=1}^{N(t)} F_iR_0(t-t_i)\right\}=\mathbb{E}\left\{F_i\right\}\lambda t. \tag{8}$$

The cross-correlation function [3] of $g_0(s)$ and $g_0(t)$ is expressed by

$$\begin{aligned}\mathbb{E}\left\{g_0(s)g_0(t)\right\}&=\lambda\mathbb{E}\left\{F_i^2\right\}\int_0^{\min(s,t)} R_0(s,\tau)R_0(t,\tau)\mathrm{d}\tau+\mathbb{E}\left\{g_0(s)\right\}\mathbb{E}\left\{g_0(t)\right\}\\&\overset{s\leq t}{=}\mathbb{E}\left\{F_i^2\right\}\lambda s+\mathbb{E}^2\left\{F_i\right\}\lambda^2 st.\end{aligned} \tag{9}$$

The second-order moment of $g_0(t)$ is given by

$$\mathbb{E}\left\{g_0^2(t)\right\}=\mathbb{E}\left\{F_i^2\right\}\lambda t+\mathbb{E}^2\left\{F_i\right\}\lambda^2t^2. \tag{10}$$

The solution of (2) is given by

$$V(t)=V_0\exp(rt)-\frac{b}{r}\left[1-\exp(rt)\right]+\beta\exp(rt)\int_0^t C(s)\exp(-rs)\,\mathrm{d}s, \quad t>0. \tag{11}$$

Integrating by parts, the integral appearing in the third addend in (11) and taking into account (7), we obtain

$$\begin{aligned}V(t)&=V_0\exp(rt)-\frac{b}{r}\left[1-\exp(rt)\right]+\beta\exp(rt)\left[\exp(-rt)\,g_0(t)+r\int_0^t\exp(-rs)\,g_0(s)\mathrm{d}s\right]\\&=V_0\exp(rt)-\frac{b}{r}\left[1-\exp(rt)\right]+\beta\left[g_0(t)+r\exp(rt)\int_0^t\exp(-rs)\,g_0(s)\mathrm{d}s\right], \quad t>0.\end{aligned} \tag{12}$$

Using these results, the first moment of the solution $V(t)$ is given by

$$\begin{aligned}
\mathbb{E}\left\{V(t)\right\} &= \mathbb{E}\left\{V_0\right\}\exp\left(rt\right) - \frac{b}{r}\left[1-\exp\left(rt\right)\right] + \beta\,\mathbb{E}\left\{g_0(t) + r\exp\left(rt\right)\int_0^t \exp\left(-rs\right)g_0(s)\;\mathrm{ds}\right\} \\
&= \mathbb{E}\left\{V_0\right\}\exp\left(rt\right) - \frac{b}{r}\left[1-\exp\left(rt\right)\right] + \beta\left\{\mathbb{E}\left\{g_0(t)\right\} + r\exp\left(rt\right)\int_0^t \exp\left(-rs\right)\mathbb{E}\left\{g_0(s)\right\}\;\mathrm{ds}\right\} \\
&= \mathbb{E}\left\{V_0\right\}\exp\left(rt\right) - \frac{b}{r}\left[1-\exp\left(rt\right)\right] + \beta\left\{\mathbb{E}\left\{F_i\right\}\lambda t + r\exp\left(rt\right)\mathbb{E}\left\{F_i\right\}\lambda\int_0^t \exp\left(-rs\right)s\;\mathrm{ds}\right\} \\
&= \mathbb{E}\left\{V_0\right\}\exp\left(rt\right) - \frac{b}{r}\left[1-\exp\left(rt\right)\right] + \beta\left\{\frac{\mathbb{E}\left\{F_i\right\}\lambda}{r}\left[\exp\left(rt\right)-1\right]\right\}.
\end{aligned} \tag{13}$$

Similarly, the second moment of $V(t)$ is

$$\begin{aligned}
\mathbb{E}\left\{V(t)^2\right\} =& \mathbb{E}\left\{V_0^2\right\}\exp\left(2rt\right) - 2\mathbb{E}\left\{V_0\right\}\exp\left(rt\right)\left[\frac{b}{r}\left[1-\exp\left(rt\right)\right]\right] + \left[\frac{b}{r}\left[1-\exp\left(rt\right)\right]\right]^2 \\
&+ 2\beta\left\{\left[\mathbb{E}\left\{V_0\right\}\exp\left(rt\right) - \frac{b}{r}\left[1-\exp\left(rt\right)\right]\right]\left[\frac{\mathbb{E}\left\{F_i\right\}\lambda}{r}\left[\exp\left(rt\right)-1\right]\right]\right\} \\
&+ \beta^2\left\{\mathbb{E}\left\{F_i^2\right\}\lambda\left[\frac{\exp\left(2rt\right)-1}{2r}\right] + \mathbb{E}^2\left\{F_i\right\}\lambda^2\left[\frac{\left[\exp\left(rt\right)-1\right]^2}{r^2}\right]\right\}.
\end{aligned} \tag{14}$$

Finally, the variance of $V(t)$ is

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left\{V(t)\right\} &= \mathbb{E}\left\{V^2(t)\right\} - \mathbb{E}^2\left\{V(t)\right\} \\
&= \mathbb{V}\mathrm{ar}\left\{V_0\right\}\exp(2rt) + \beta^2\left\{\frac{\mathbb{E}\left\{F_i^2\right\}\lambda}{2r}\left[\exp\left(2rt\right)-1\right]\right\}.
\end{aligned} \tag{15}$$

2.1 Results and Discussion

In this example, we will obtain synthetic data from an RC circuit. For this purpose, we first start from the deterministic model, namely, $v(t) = v_0\exp\left(rt\right) - \dfrac{b}{r}\left[1-\exp\left(rt\right)\right]$. Next, we will perturb the RC circuit model to obtain several simulations for each t_i, namely, $i = 0, 1, \ldots, N$. At this point, we have obtained M synthetic simulations, namely, $j = 1, 2, \ldots, M$, for each t_i, for $\left\{v_i^j\right\}_{j=1}^M = v(t_i) + \left\{\eta_i^j\right\}_{j=1}^M$, where $\eta = \eta^j$ is an additive noise that follows a Gaussian distribution $\eta \sim \mathrm{N}(0, 0.05)$. It is important to mention that the simulations $\left\{\eta^j\right\}_{j=1}^M$ are independent. With this information, it is possible to obtain the first moments of the synthetic simulations, namely, $m_i^l = \dfrac{1}{M}\sum_{j=1}^M (v_i^j)^l$, $l = 1, \cdots, L$.

Finally, we proceed to adjust the stochastic parameters of the proposed model, namely, $V_0 \sim \mathrm{N}(\mu_{V_0}, \sigma_{V_0})$, $F_i \sim \mathrm{N}(\mu_F, \sigma_F)$ and λ, for $N = 10$ and $M = 30$, in order to minimize the following expression

$$\begin{aligned}
&\mathbb{E}\left\{\mu_{V_0}, \sigma_{V_0}, \mu_F, \sigma_F, \lambda\right\} \\
&\quad = \sum_{i=0}^{N}\left\{\left[\mathbb{E}\left\{V(t_i, \mu_{V_0}, \mu_F, \sigma_F, \lambda)\right\} - m_i^1\right]^2 + \left[\mathbb{E}^2\left\{V(t_i, \mu_{V_0}, \sigma_{V_0}, \mu_F, \sigma_F, \lambda)\right\} - m_i^2\right]^2\right\},
\end{aligned} \tag{16}$$

where $\mathbb{E}\left\{V(t_i, \mu_{V_0}, \mu_F, \sigma_F, \lambda)\right\}$ and $\mathbb{E}^2\left\{V(t_i, \mu_{V_0}, \sigma_{V_0}, \mu_F, \sigma_F, \lambda)\right\}$ are the first and second moments of the solution.

Let us consider as deterministic parameters of the linear RC circuit model are as follows: the resistance of the RC circuit is $R = 4000$ ohms, the capacitance of the RC circuit is $C = 0.0005$ farads, the external voltage is $V_s = 3$ volts, and the intensity of the delta-correlated process is $\alpha = -2$. With this information, it is possible to obtain the auxiliary parameters, namely, $r = -\dfrac{1}{RC} = -0.5\ \frac{1}{s}$, $b = \dfrac{V_s}{RC} = 1.5\ \dfrac{V}{s}$, where V denotes the volts and s denotes the seconds, and $\beta = \dfrac{\alpha}{RC} = -1\ \frac{1}{s}$, respectively. Finally, let us consider $v_0 = 1.5$ to perturb the model.

The parameters obtained after of minimizing the expression given in (16) are $V_0 \sim \mathrm{N}(\mu_{V_0}, \sigma_{V_0})$, where

- $\mathbb{E}\{V_0\} = \mu_{V_0} = 1.59376336$,
- $\mathbb{E}\left\{V_0^2\right\} = \mu_{V_0}^2 + \sigma_{V_0}^2 = 1.93628551$

and the delta-correlated process $C(t)$ follows a compound Poisson process as follows:

- F_i follows a Gaussian distribution, $F_i \sim \mathrm{N}(\mu_F, \sigma_F)$, where
  - $\mathbb{E}\{F_i\} = \mu_{F_i} = 0.01617528$.
  - $\mathbb{E}\left\{F_i^2\right\} = \mu_{F_i}^2 + \sigma_{F_i}^2 = 0.28679470$.
- $N(t)$ follows a Poisson process, with rate $\lambda = 0.82559186$.

In Figures 1, 2 and 3, we show the graphical representation between the simulated (using 10^4 simulations) and exact of the first-order, second-order moments and variance of the solution, respectively. As we can see, the agreement between simulated and exact functions of the main statistics are good.

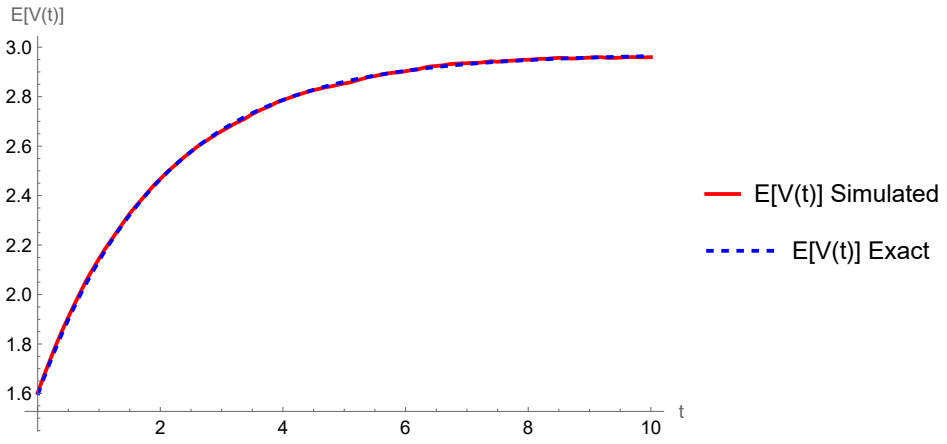


Figure 1: Comparison of the simulated and exact first-order moment of the solution $\mathbb{E}\{V(t)\}$.

3 Conclusions and Future work

In this paper, we have performed a probabilistic analysis of an RC circuit modeled by a first-order linear differential equation with an additive delta-correlated processes. Additionally, we have obtained the main statistics of the solution, namely, the first order moments and the variance functions. Finally, we have compared the main statistics between the simulation and the exact functions obtained showing a good agreement.

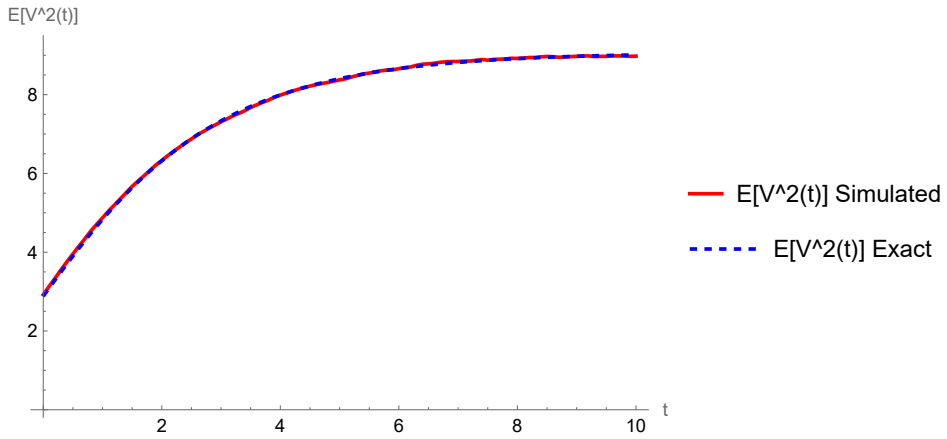


Figure 2: Comparison of the simulated and exact first-order moment of the solution $\mathbb{E}\left\{V^2(t)\right\}$.

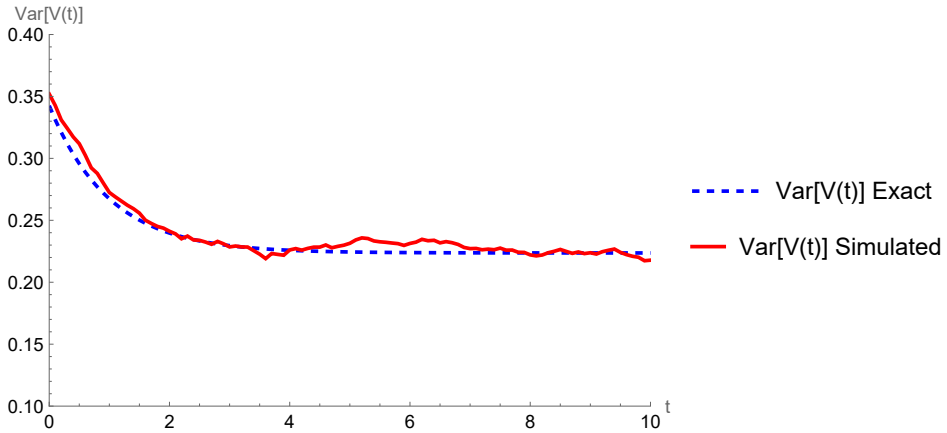


Figure 3: Comparison of the simulated and exact variance of the solution $\mathbb{V}ar\left\{V(t)\right\}$.

Acknowledgments

This work has been supported by the grant PAID-11-25 funded by Universitat Politècnica de València and by grant PID2025-172735NB-I00 funded by MCIN/AEI.

References

[1] E. Kolářová and L. Brančík, Vector stochastic differential equations used to electrical networks with random parameters. International Journal of Advances in Telecommunications, Electrotechnics, Signals and System, **2**(1) (2012), 1–8.

[2] T. T. Soong, Random differential equations in science and engineering (Book). New York, Academic Press, Inc. (Mathematics in Science and Engineering., **103** (1973).

[3] Y.K. Lin, Probabilistic Theory of Structural Dynamics. McGraw-Hill, New York, (1967).

[4] J.-C. Cortés, E. López-Navarro, P. Martínez-Rodríguez, J.-V. Romero, and M.-D. Roselló, Extending the analysis of the euler-bernoulli model for a stochastic static cantilever beam: Theory and simulations. Probabilistic Engineering Mechanics, **74**(1) (2023), 103493.

A new type of multi-Sugeno-inspired aggregation functions

José Carlos R. Alcantud [♭1]

(♭) Unidad de Excelencia Gestión Económica para la Sostenibilidad GECOS and IME, University of Salamanca, Spain.

1 Introduction

Within the domain of aggregation functions [3], fuzzy integrals hold significant importance. Calculating them requires eliciting a fuzzy measure, a process that can be computationally demanding. Different applications necessitate different formulations of fuzzy integrals. The Sugeno integral operates solely with ordinal data [13]. It is designed to model situations involving uncertainty and imprecision, particularly when the primary concern is the relative position of the different options. When it is impossible to consistently measure the intensity of preferences, no numerical values derived from the evaluation of alternatives should be assigned meaning, thus prohibiting the use of arithmetic operations like addition or multiplication. Applications of Sugeno integration include scientometrics [10, 12] and fuzzy rule-based classification [15]. Several extensions have been explored [5, 6, 7, 11].

Relatedly, a Sugeno-inspired aggregation function [8] draws upon the principles of discrete Sugeno integration. However, it replaces the fuzzy measure with a single symmetric, monotonically increasing function that compares each input to the tuple formed by the remaining inputs. This structure avoids the computationally intensive precalculation of a fuzzy measure. Recent research [1] has introduced another variation of the Sugeno integral, which can be described as a component-sensitive extension of these Sugeno-inspired aggregation functions [8]. This model is referred to as multi-Sugeno-inspired aggregation functions. The use of multiple comparative functions (one for each component of the multivalued input) allows for comparisons independent of position, thereby enabling the use of more reliable Sugeno-inspired evaluations. Meanwhile, the model in [1] still avoids the precomputation of fuzzy measures and maintains symmetry.

The multi-Sugeno-inspired aggregation functions defined in [1] are a direct extension of the Sugeno-inspired aggregation function [8], which, in turn, are directly inspired by the disjunctive form of the classical discrete Sugeno integrals. While a conjunctive formula also exists [8, Proposition 3.4], it is not directly comparable to the standard conjunctive form of Sugeno integrals. We investigate the multi-Sugeno-inspired aggregation functions that arise from the generalization of such a conjunctive form upon the principles outlined in [8]. Comparisons with the model investigated in [1] are made too.

[1] jcr@usal.es

2 Preliminary concepts and background

We adhere to the notation of [8]. We refer to $n \in \mathbb{N}$ with $n \geq 2$, and the finite set $N = \{1, \ldots, n\}$. The set of subsets of N is $\mathcal{P}(N)$. Wectors in $[0,1]^n$ will be denoted by bold symbols, for example, $\mathbf{a} = (a_1, ..., a_n) \in [0,1]^n$.

The notation $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$ applies when $a, b \in [0,1]$. It is exported to vectors in a natural way: we let $\mathbf{a} \wedge \mathbf{b} = (a_1 \wedge b_1, \ldots, a_n \wedge b_n)$, $\mathbf{a} \vee \mathbf{b} = (a_1 \vee b_1, \ldots, a_n \vee b_n)$, forr $\mathbf{a} = (a_1, \ldots, a_n), \mathbf{b} = (b_1, \ldots, b_n) \in [0,1]^n$.

We associate necessary concepts from every $\mathbf{a} = (a_1, \ldots, a_n) \in [0,1]^n$. First, we let () be a permutation of N for which $a_{(1)} \leq a_{(2)} \leq \ldots \leq a_{(n)}$. This permutation produces $\mathbf{a}_{\nearrow} = (a_{(1)}, a_{(2)}, \ldots, a_{(n)})$, the non-decreasing sorted vector of $\mathbf{a}$. Then for each $k \in N$, the vector $\hat{\mathbf{a}}_{(k)} \in [0,1]^{n-1}$ is the result of removing $\mathbf{a}_{\nearrow}$'s k-th component, i.e.,

$$(\hat{\mathbf{a}}_{(k)})_j = \begin{cases} a_{(j)} & \text{when } j < k \\ a_{(j+1)} & \text{when } j \geq k. \end{cases} \tag{1}$$

We will refer to aggregation functions defined as follows:

Definition [3, 9]. A mapping $f : [0,1]^n \to [0,1]$ is an aggregation function if it satisfies boundedness (i.e., $f(0, \ldots, 0) = 0$ and $f(1, \ldots, 1) = 1$) and monotonicity (i.e., $f(\mathbf{b}) \geq f(\mathbf{a})$ when $\mathbf{a}, \mathbf{b} \in [0,1]^n$, $\mathbf{b} \geq \mathbf{a}$).

The following concept precedes the definition of fuzzy integrals:

Definition [4]. A mapping $m : \mathcal{P}(N) \to [0,1]$ is a fuzzy measure (also, a capacity) on N if it is set-monotonic (i.e., $m(A) \leq m(A')$ when $A \subseteq A' \subseteq N$), $m(N) = 1$ (a requirement called the normalization condition), and $m(\varnothing) = 0$.

The Sugeno integral was designed for applications in settings defined by ordinal data [13]. Its discrete formula is as follows:

Definition [4, Sect. 5.3]. The Sugeno integral defined by m, a fuzzy measure on N, is the mapping $S_m : [0,1]^n \to [0,1]$ that with each $\mathbf{a} = (a_1, \ldots, a_n) \in [0,1]^n$ associates

$$S_m(\mathbf{a}) = \bigvee_{j=1}^{n} a_{(j)} \wedge m(H_j) = \bigwedge_{j=1}^{n} a_{(j)} \vee m(H_{j+1}), \tag{2}$$

where $H_j = \{(j), \ldots, (n)\}$ for $j = 1, \ldots, n$, and $H_{n+1} = \varnothing$.

The formulas in Eq. (2) are the classical disjunctive and conjunctive formulations of the discrete Sugeno integral, respectively. They produce monotonic or non-decreasing evaluations wih an ordinal spirit. Remarkably, there is not yet a clear interpretation of their operational principles [14].

A crucial limitation of Eq. (2) is that it necessitates a fuzzy measure, but this concept is computationally demanding. It is well known that learning fuzzy measures is difficult in practical applications. This reliance is avoided in the model defined by [8], namely, Sugeno-inspired functions. This model keeps the spirit of the Sugeno integral, although now a single non-decreasing (i.e., monotonic) function F captures the importance of each input compared with the group formed by the other inputs. This strategy reduces computational complexity and simultaneously allows enough modeling flexibility through the choice of F, which still ensures that all inputs are involved in the calculations:

Definition [8]. Each non-decreasing function $F : [0,1]^{n-1} \to [0,1]$ defines a Sugeno-inspired function $SI_F : [0,1]^n \to [0,1]$ by the expression

$$SI_F(\mathbf{a}) = \bigvee_{j=1}^{n} a_{(j)} \wedge F(\hat{\mathbf{a}}_{(j)}), \text{ for each } \mathbf{a} = (a_1, \ldots, a_n) \in [0,1]^n. \tag{3}$$

Eq. (3) defines a symmetric function, therefore it extends OWMAX [8, Remark 3.2 (iii) and Sect. 3.3], but non-symmetric Sugeno integrals are never Sugeno-inspired functions.

A multi-Sugeno-inspired type of functions is defined and investigated in [1]. This model can be considered a component-dependent generalization of Sugeno-inspired functions that preserves the inspiration of Sugeno integrals in the disjunctive form:

Definition [1]. Let $\mathbf{M}_j : [0,1]^{n-1} \to [0,1]$ be a non-decreasing function ($j = 1, \ldots, n$), and define $\mathbf{M} = (\mathbf{M}_1, \ldots, \mathbf{M}_n)$. The multi-Sugeno-inspired function $SI_{\mathbf{M}} : [0,1]^n \to [0,1]$ is

$$SI_{\mathbf{M}}(\mathbf{a}) = \bigvee_{j=1}^{n} a_{(j)} \wedge \mathbf{M}_j(\hat{\mathbf{a}}_{(j)}), \text{ for each } \mathbf{a} = (a_1, \ldots, a_n) \in [0,1]^n. \tag{4}$$

Eq. (4) defines a monotonic and symmetric function too. It defines an aggregation function under mild conditions, for example, when $\min \leq \mathbf{M}_j$ for some $j \in N$ [1]. In this statement, min denotes the standard minimum function. The case $\mathbf{M}_1 = \ldots = \mathbf{M}_n = F$ produces the Sugeno-inspired function SI_F.

3 Results

By inspiration of the alternative formulas in Eq. (2), we propose to investigate the following conjunctive variation of Sugeno-inspired functions [8] and their component-dependent generalization with a disjunctive structure [1]:

Definition. Let $\mathbf{M}_j : [0,1]^{n-1} \to [0,1]$ be a non-decreasing function ($j = 1, \ldots, n$). Then $\mathbf{M} = (\mathbf{M}_1, \ldots, \mathbf{M}_n)$ defines a conjunctive multi-Sugeno-inspired function $SI^{\mathbf{M}} : [0,1]^n \to [0,1]$ by

$$SI^{\mathbf{M}}(\mathbf{a}) = \bigwedge_{j=1}^{n} a_{(j)} \vee \mathbf{M}_j(\hat{\mathbf{a}}_{(j)}), \text{ for each } \mathbf{a} = (a_1, \ldots, a_n) \in [0,1]^n. \tag{5}$$

Proposition. Eq. (5) defines a monotonic and symmetric function.

Proof. Symmetry is obvious. Proving monotonicity is routine. Let $\mathbf{a}, \mathbf{b} \in [0,1]^n$ such that $\mathbf{a} \geq \mathbf{b}$. Fix $(\cdot)$ and $(\cdot)'$, respective permutations of N such that $\mathbf{a}_{\nearrow} = (a_{(1)}, a_{(2)}, \ldots, a_{(n)})$ and $\mathbf{b}_{\nearrow} = (b_{(1)'}, b_{(2)'}, \ldots, b_{(n)'})$. Because $\mathbf{a}_{\nearrow} \geq \mathbf{b}_{\nearrow}$, one has $\mathbf{M}_j(\hat{\mathbf{a}}_{(j)}) \geq \mathbf{M}_j(\hat{\mathbf{b}}_{(j)'})$ $(j = 1, \ldots, n)$. Therefore $a_{(j)} \vee \mathbf{M}_j(\hat{\mathbf{a}}_{(j)}) \geq b_{(j)'} \vee \mathbf{M}_j(\hat{\mathbf{b}}_{(j)'})$ for $j = 1, \ldots, n$, and $SI^{\mathbf{M}}(\mathbf{a}) = \bigwedge_{j=1}^{n} a_{(j)} \vee \mathbf{M}_j(\hat{\mathbf{a}}_{(j)}) \geq \bigwedge_{j=1}^{n} b_{(j)'} \vee \mathbf{M}_j(\hat{\mathbf{b}}_{(j)'}) = SI^{\mathbf{M}}(\mathbf{b})$. □

Unlike the case of Eq. (2), the disjunctive and conjunctive variations of multi-Sugeno-inspired functions generally do not coincide.

Lemma. Fix $n = 2$ and $\mathbf{M}_j : [0,1]^{2-1} \to [0,1]$ $(j = 1, 2)$ such that $a \leq \mathbf{M}_2(a) \leq \mathbf{M}_1(a)$, for all $a \in [0,1]$. Then

$$SI^{\mathbf{M}}(\mathbf{a}) = \begin{cases} a_2 \vee \mathbf{M}_2(a_1), & \text{if } a_1 < a_2, \\ a_1 \vee \mathbf{M}_2(a_2), & \text{if } a_1 \geq a_2. \end{cases} \tag{6}$$

Equivalently, $SI^{\mathbf{M}}(\mathbf{a}) = \max\Big\{\mathbf{M}_2(\min\{a_1, a_2\}), \max\{a_1, a_2\}\Big\}$.

Proof. It suffices to prove the first case, due to symmetry. When $a_1 < a_2$,

$$SI^{\mathbf{M}}(\mathbf{a}) = (a_1 \vee \mathbf{M}_1(a_2)) \wedge (a_2 \vee \mathbf{M}_2(a_1)) = a_2 \vee \mathbf{M}_2(a_1)$$

because $a_2 \leq \mathbf{M}_1(a_2) \leq a_1 \vee \mathbf{M}_1(a_2)$ and $\mathbf{M}_2(a_1) \leq \mathbf{M}_2(a_2) \leq \mathbf{M}_1(a_2) \leq a_1 \vee \mathbf{M}_1(a_2)$. □

Example. Fix $n = 2$ and define $\mathbf{M}_j : [0,1]^{2-1} \to [0,1]$ $(j = 1, 2)$ by $\mathbf{M}_1(a) = \sqrt[3]{a}$, $\mathbf{M}_2(a) = \sqrt{a}$ for each $a \in [0,1]$. We can use the Lemma because $a \leq \mathbf{M}_2(a) \leq \mathbf{M}_1(a)$ for each $a \in [0,1]$. Therefore, when $\mathbf{a} = (a_1, a_2) \in [0,1]^2$, $SI^{\mathbf{M}}(\mathbf{a}) = \max\left\{\sqrt{\min\{a_1, a_2\}}, \max\{a_1, a_2\}\right\}$

However, [1] proves $SI_{\mathbf{M}}(\mathbf{a}) = \min\left\{\sqrt{\min\{a_1, a_2\}}, \max\{a_1, a_2\}\right\}$.

The case $\mathbf{M}_1 = \ldots = \mathbf{M}_n = F$ produces a conjunctive multi-Sugeno-inspired function that is a variation of the Sugeno-inspired function SI_F, namely, the new function

$$SI^F(\mathbf{a}) = \bigwedge_{j=1}^{n} a_{(j)} \vee F(\hat{\mathbf{a}}_{(j)}), \text{ for each } \mathbf{a} = (a_1, \ldots, a_n) \in [0,1]^n. \tag{7}$$

Definition. We can consider Eq. (3) and (7) as disjunctive and conjunctive Sugeno-inspired functions defined from the non-decreasing function $F : [0,1]^{n-1} \to [0,1]$.

Remark. A conjunctive-type alternative expression for Sugeno-inspired functions exists [8, Propos. 3.4], namely, the function that for each $\mathbf{a} = (a_1, \ldots, a_n) \in [0,1]^n$ computes

$$SI_F(\mathbf{a}) = \begin{cases} \bigwedge_{j=1}^{n} a_{(j)} \vee F(\hat{\mathbf{a}}_{(j+1)}), & \text{if } a_{(1)} \leq F(\hat{\mathbf{a}}_{(1)}), \\ F(\hat{\mathbf{a}}_{(1)}), & \text{otherwise,} \end{cases}$$

where $F(\hat{\mathbf{a}}_{(n+1)}) = 0$. Therefore if $F : [0,1]^{n-1} \to [0,1]$ is a non-decreasing function such that $F \geq \min$, then $SI_F(\mathbf{a}) = \bigwedge_{j=1}^{n} a_{(j)} \vee F(\hat{\mathbf{a}}_{(j+1)})$ with the convention $F(\hat{\mathbf{a}}_{(n+1)}) = 0$. Even though this formulation emphasizes the structural similarity between Sugeno integrals and Sugeno-inspired functions, the resulting expression is not a satisfactory Sugeno-inspired function. This is because it fails to meet the fundamental requirement that the fuzzy measure is replaced with a function that compares each input to the tuple formed by the remaining inputs. We believe that the solution presented in Eq. (7) fully satisfies all the requirements to operate faithfully as a Sugeno-inspired function in a conjunctive form.

4 Conclusions

Research into Sugeno-type evaluative functions represents a promising area for future development. We have presented a comprehensive review of existing literature and new results that incorporate symmetric aggregation methods. Furthermore, it is possible to investigate non-symmetric Sugeno-inspired functions for aggregating unbalanced ordinal data. This is the focus of the pioneering work [2], which explores the trade-offs between disjunctive and conjunctive expressions. We anticipate continuing our investigation along these lines in the future.

Acknowledgments

Financial support of the Department of Education of the Junta de Castilla y León and FEDER Funds (Reference: CLU-2025-2-03) is gratefully acknowledged.

References

[1] José Carlos R. Alcantud. Multi-sugeno-inspired aggregation functions. In Vicenç Torra et al., editor, *MDAI 2026. Proceedings of the 23rd International Conference on Modeling Decisions for Artificial Intelligence*, Lecture Notes in Artificial Intelligence. Springer Nature, 2026.

[2] José Carlos R. Alcantud. Non-symmetric Sugeno-inspired functions for aggregation of unbalanced ordinal data. Submitted, 2026.

[3] Gleb Beliakov, Ana Pradera, and Tomasa Calvo. *Aggregation Functions: A Guide for Practitioners*, volume 221 of *Studies in Fuzziness and Soft Computing*. Springer-Verlag Berlin Heidelberg, Heidelberg, 2007. doi: 10.1007/978-3-540-73721-6.

[4] Gleb Beliakov, Simon James, and Jian-Zhang Wu. *Choquet Capacities and Fuzzy Integrals*. Theory and Applications of Computability. Springer, 2026. doi: 10.1007/978-3-031-97070-2.

[5] Michał Boczek and Marek Kaluszka. On the extended Choquet-Sugeno-like operator. *International Journal of Approximate Reasoning*, 154:48–55, 2023. doi: https://doi.org/10.1016/j.ijar.2022.12.006.

[6] Jana Borzová, Ondrej Hutník, and Miriam Kleinová. Extending the conjunctive form of the Sugeno integral via chains. In *2025 IEEE International Conference on Fuzzy Systems (FUZZ)*, pages 1–6, 2025. doi: 10.1109/FUZZ62266.2025.11152200.

[7] Didier Dubois, Henri Prade, Agnès Rico, and Bruno Teheux. Generalized Sugeno integrals. In J. Carvalho, M. J. Lesot, U. Kaymak, S. Vieira, B. Bouchon-Meunier, and R. Yager, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems, Proc. 16th Int. Conf. IPMU 2016, Eindhoven, Part II*, number 610 in Communications in Computer and Information Science, pages 363–374, 2016. doi: 10.1007/978-3-319-40596-4 31.

[8] Xabier González-García, Lubomíra Horanská, Zdenko Takáč, Juan Tinguaro Rodríguez, Daniel Gómez, and Humberto Bustince. Sugeno-inspired aggregation functions. *Fuzzy Sets and Systems*, 523:109616, 2026. doi: 10.1016/j.fss.2025.109616.

[9] M. Grabisch, Jean-Luc Marichal, R. Mesiar, and E. Pap. *Aggregation Functions*. Cambridge University Press, Cambridge, 2009. doi: 10.1017/CBO9781139644150.

[10] Barbora Hennelová. Generalized Sugeno integrals applied to scientometric indices. In *2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY)*, pages 601–606, 2024. doi: 10.1109/SISY62279.2024.10737520.

[11] Lubomíra Horanská and Alexandra Siposová. A generalization of the discrete Choquet and Sugeno integrals based on a fusion function. *Information Sciences*, 451-452:83–99, 2018. doi: https://doi.org/10.1016/j.ins.2018.03.059.

[12] Radko Mesiar and Marek Gagolewski. H-index and other Sugeno integrals: Some defects and their compensation. *IEEE Transactions on Fuzzy Systems*, 24(6):1668–1672, 2016. doi: 10.1109/TFUZZ.2016.2516579.

[13] Michio Sugeno. *Theory of fuzzy integrals and applications.* PhD thesis, Tokyo Institute of Technology, 1974.

[14] Vicenç Torra and Yasuo Narukawa. The interpretation of fuzzy integrals and their application to fuzzy systems. *International Journal of Approximate Reasoning*, 41(1):43–58, 2006. doi: https://doi.org/10.1016/j.ijar.2005.08.001.

[15] Jonata Wieczynski, Giancarlo Lucca, Eduardo Borges, Asier Urio-Larrea, Carlos Lṕez Molina, Humberto Bustince, and Graçaliz Dimuro. Application of the Sugeno integral in fuzzy rule-based classification. *Applied Soft Computing*, 167:112265, 2024. doi: https://doi.org/10.1016/j.asoc.2024.112265.

On the use of the delta function for the computation of probability density functions in the random Burgers-Riemann problem

Marc Jornet[♭],[1]

(♭) Escuela Superior de Ingeniería y Tecnología, Universidad Internacional de La Rioja, Av. La Paz 137, 26006 Logroño, Spain.

1 Problem

We review the computation of the probability density function in the context of the random Riemann problem for the inviscid Burgers' equation:

$$\frac{\partial}{\partial t}U(x,t)+U(x,t)\frac{\partial}{\partial x}U(x,t)=0,\quad x\in\mathbb{R},\ t>0, \tag{1}$$

$$U(x,0)=\begin{cases}U_L, & x<0,\\ U_R, & x>0.\end{cases} \tag{2}$$

The initial conditions U_L and U_R are random variables that form an absolutely continuous random vector, because of some estimation error or biased measurement. In particular, we are interested in the foundations that permitted the use of the delta function in the calculations of [1], [2], which were the first references to obtain joint density functions. It is a theoretical contribution in the field of uncertainty quantification.

2 Probabilistic context

Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, where Ω is the sample space formed by outcomes $\omega\in\Omega$, $\mathcal{F}\subseteq 2^\Omega$ is the σ-algebra of events, and $\mathbb{P}:\mathcal{F}\to[0,1]$ is the probability measure. Let $\mathbb{E}$ be the expectation operator. Let $A:\Omega\to\mathbb{R}$ be a random variable with probability distribution or law $\mathbb{P}_A=\mathbb{P}\circ A^{-1}$.

If the law of A is given by $\mathrm{d}\mathbb{P}_A=f_A(a)\mathrm{d}a$, where f_A is a non-negative Borel measurable function such that $\int_{-\infty}^{\infty}f_A(a)\mathrm{d}a=1$, then f_A is called the probability density function of A. It is the Radon-Nikodym derivative $\mathrm{d}\mathbb{P}_A/\mathrm{d}a$. In this case, A is absolutely continuous and $\mathbb{P}_A$ is absolutely continuous with respect to the Lebesgue measure.

However, the law of A may be expressed more generally as

$$\mathrm{d}\mathbb{P}_A=\tilde{f}_A(a)\mathrm{d}a+\sum_{k=1}^{m}p_k\mathrm{d}\delta_{a_k}(a),$$

[1]marc.jornet@unir.net

where $p_k \in (0,1]$ are probabilities, $a_k \in \mathbb{R}$ are different mass points, and δ_{a_k} are Dirac measures: $\int_{\mathcal{B}} \mathrm{d}\delta_{a_k}(a)$ is 1 if $a_k \in \mathcal{B}$ and 0 if $a_k \notin \mathcal{B}$, where $\mathcal{B} \subseteq \mathbb{R}$ is a Borel set. The function $\tilde{f}$ is a non-negative Borel measurable function, but non-normalized. In this case, $\mathbb{P}_A$ is not absolutely continuous with respect to the Lebesgue measure. For example, if $\tilde{f}_A \equiv 0$, then A is discrete.

When $\tilde{f}_A$ is not identically zero, then A is a mixed random variable, containing an absolutely continuous part and a discrete part. Operationally, it might be convenient to assume that $\mathrm{d}\delta_{a_k}(a) = \delta(a-a_k)\mathrm{d}a$, that is, the Dirac measure is absolutely continuous with respect to the Lebesgue measure, with Radon-Nikodym derivative $\delta(a-a_k)$. This is an abuse of notation that incorporates the delta "function" as if it existed as a standard probability density function, with the properties $\delta(0)=\infty$, $\delta(\mathbb{R}\backslash\{0\})=\{0\}$, and $\int_{-\infty}^{\infty}\delta(a)\mathrm{d}a = 1$. It is just an idealized object. The generalized density function of A is $f_A(a) = \tilde{f}_A(a) + \sum_{k=1}^{m} p_k \delta(a-a_k)$.

3 Properties of the delta function

For any Borel measurable function $g:\mathbb{R}\to\mathbb{R}$, the property $\mathbb{E}[g(b)] = g(b)$ is obvious, for each point b. Thus, in terms of the delta function $\delta(\cdot - b)$, which acts as the generalized density function of the deterministic variable b, we have $\int_{-\infty}^{\infty} g(b)\delta(a-b)\mathrm{d}a = g(a)$. Observe that, in contrast to Physics, we do not need g to be continuous or any other type of regularity. Our point of view is based on measures, rather than distributions in functional analysis.

In particular,

$$f_A(a) = \mathbb{E}[\delta(a-A)], \tag{3}$$

because

$$\mathbb{E}[\delta(a-A)] = \int_{-\infty}^{\infty} \delta(a-\tilde{a}) f_A(\tilde{a})\mathrm{d}\tilde{a} = f_A(a). \tag{4}$$

This property holds irrespective of the continuity or discontinuity of f_A. It is an equality for almost every $a \in \mathbb{R}$. Even if f_A contains delta-function parts, we can perform the same computation $\int_{-\infty}^{\infty}\delta(a-\tilde{a})f_A(\tilde{a})\mathrm{d}\tilde{a} = f_A(a)$, because the product of delta functions exists as the idealized Radon-Nikodym derivative of the product of Dirac measures. That is, from $\mathrm{d}(\delta_{a_1}\otimes\delta_{a_2})(b_1,b_2) = \mathrm{d}\delta_{a_1}(b_1)\mathrm{d}\delta_{a_2}(b_2)$ by Fubini's theorem, we have $\delta(b_1-a_1, b_2-a_2) = \delta(b_1-a_1)\delta(b_2-a_2)$. In probability, $\delta(b_1-a_1,b_2-a_2)$ is the joint density of the constants a_1 and a_2, and since they are independent, we obtain the product of marginal densities: $\delta(b_1-a_1)\delta(b_2-a_2)$.

In measure theory, the representation (3) is

$$\mathrm{d}\mathbb{P}_A = \int_{\Omega} \mathrm{d}\delta_{A(\omega)}(a)\mathrm{d}\mathbb{P}(\omega), \tag{5}$$

which can be proved by integration:

$$\begin{aligned}\int_{-\infty}^{\infty} g(a)\int_{\Omega}\mathrm{d}\delta_{A(\omega)}(a)\mathrm{d}\mathbb{P}(\omega) &= \int_{\Omega}\int_{-\infty}^{\infty} g(a)\mathrm{d}\delta_{A(\omega)}(a)\mathrm{d}\mathbb{P}(\omega)\\ &= \int_{\Omega} g(A(\omega))\mathrm{d}\mathbb{P}(\omega)\\ &= \int_{-\infty}^{\infty} g(a)\mathrm{d}\mathbb{P}_A(a),\end{aligned}$$

if g is a Borel measurable function such that $\mathbb{E}[|g(A)|] < \infty$.

Note that, if we condition on an event $\mathcal{B} \in \mathcal{F}$, then

$$f_A(a|\mathcal{B}) = \mathbb{E}[\delta(a-A)|\mathcal{B}] = \frac{\mathbb{E}[\delta(a-A)\mathcal{X}_{\mathcal{B}}(A)]}{\mathbb{P}[A\in\mathcal{B}]}, \tag{6}$$

where $\mathcal{X}_{\mathcal{B}}$ is the indicator function. In terms of measure theory,

$$\mathrm{d}\mathbb{P}_A^{\mathcal{B}} = \int_{\Omega} \mathrm{d}\delta_{A(\omega)}(a)\mathrm{d}\mathbb{P}^{\mathcal{B}}(\omega),$$

where the conditional probabilities are being used: $\mathbb{P}^{\mathcal{B}}[\cdot] = \mathbb{P}[\cdot|\mathcal{B}]$ and $\mathbb{P}_A^{\mathcal{B}}[\cdot] = \mathbb{P}_A[\cdot|\mathcal{B}]$.

Finally, we observe that with the multivariate delta function, for example $\delta(b_1 - a_1, b_2 - a_2) = \delta(b_1 - a_1)\delta(b_2 - a_2)$ in the two dimensional scenario, one can develop the previous discussion for random vectors A.

We summarize these points as follows. In a probability context, the delta function is the idealized Radon-Nikodym derivative of the Dirac measure and can be integrated with respect to Borel measurable functions (for example, Lebesgue integrable functions). When being integrated with respect to continuous functions, it coincides with the distributional definition often used in Physics.

We finally emphasize why the representation (3) holds for any generalized density function. It is based on (4) in the probability context (which was noted in [1]), and (5) in the physics or measure-theory contexts (which was noted in [2]).

4 Uncertainty quantification

Suppose that

$$A = A_1\mathcal{X}_{\mathcal{B}_1} + a_2\mathcal{X}_{\mathcal{B}_2},$$

where $\mathcal{B}_1$ and $\mathcal{B}_2$ are disjoints sets of $\mathcal{F}$, A_1 is an absolutely continuous random variable, and α_2 is a constant. Using (6), we have

$$\begin{aligned} f_A(a) &= \mathbb{E}[\delta(a - A)] = \mathbb{E}[\delta(a - A)|\mathcal{B}_1]\mathbb{P}[\mathcal{B}_1] + \mathbb{E}[\delta(a - A)|\mathcal{B}_2]\mathbb{P}[\mathcal{B}_2] \\ &= \mathbb{E}[\delta(a - A_1)|\mathcal{B}_1]\mathbb{P}[\mathcal{B}_1] + \mathbb{E}[\delta(a - a_2)|\mathcal{B}_2]\mathbb{P}[\mathcal{B}_2] \\ &= f_{A_1}(a|\mathcal{B}_1)\mathbb{P}[\mathcal{B}_1] + \delta(a - a_2)\mathbb{P}[\mathcal{B}_2]. \end{aligned}$$

With indicator functions, we arrive at the same result:

$$\begin{aligned} f_A(a) &= \mathbb{E}[\delta(a - A)\mathcal{X}_{\mathcal{B}_1}(a)] + \mathbb{E}[\delta(a - A)\mathcal{X}_{\mathcal{B}_2}(a)] \\ &= \mathbb{E}[\delta(a - A_1)\mathcal{X}_{\mathcal{B}_1}(a)] + \mathbb{E}[\delta(a - a_2)\mathcal{X}_{\mathcal{B}_2}(a)] \\ &= \mathbb{E}[\delta(a - A_1)|\mathcal{B}_1]\mathbb{P}[\mathcal{B}_1] + \delta(a - a_2)\mathbb{P}[\mathcal{B}_2] \\ &= f_{A_1}(a|\mathcal{B}_1)\mathbb{P}[\mathcal{B}_1] + \delta(a - a_2)\mathbb{P}[\mathcal{B}_2]. \end{aligned}$$

Note that the expression $\mathbb{E}[\delta(a - A)\mathcal{X}_{\mathcal{B}_j}(a)]$ requires our probabilistic interpretation of the delta function, because the indicator function is not continuous.

A statistic is

$$\begin{aligned} \mathbb{E}[h(A)] &= \int_{-\infty}^{\infty} h(a)f_A(a)\mathrm{d}a \\ &= \mathbb{P}[\mathcal{B}_1]\int_{-\infty}^{\infty} h(a)f_{A_1}(a|\mathcal{B}_1)\mathrm{d}a + \mathbb{P}[\mathcal{B}_2]\int_{-\infty}^{\infty} h(a)\delta(a - a_2)\mathrm{d}a \\ &= \int_{\mathcal{B}_1} h(a)f_{A_1}(a)\mathrm{d}a + \mathbb{P}[\mathcal{B}_2]h(a_2), \end{aligned}$$

where h is any Borel measurable function, not necessarily regular: h can be a square for a variance, an exponential for a characteristic function, or an indicator function for a probability over a set. Observe that if $h(b) = \mathcal{X}_{(-\infty,a_2]}(b)$, then $h(a_2) = 1$, but if $h(b) = \mathcal{X}_{(-\infty,a_2)}(b)$, then $h(a_2) = 0$,

so $\mathbb{E}[h(A)]$ changes depending on the representative for a discontinuous h; this fact makes sense, because a_2 is a mass point of the discrete part of A.

In the general situation of

$$A = \sum_{j=1}^{r} A_j \mathcal{X}_{\mathcal{B}_j} + \sum_{i=1}^{s} \alpha_i \mathcal{X}_{\mathcal{C}_i},$$

where A_j are absolutely continuous random variables, α_i are constants, and $\mathcal{B}_j$ and $\mathcal{C}_i$ belong to $\mathcal{F}$, we analogously obtain

$$f_A(a) = \sum_{j=1}^{r} f_{A_j}(a|\mathcal{B}_j)\mathbb{P}[\mathcal{B}_j] + \sum_{i=1}^{s} \delta(a-\alpha_i)\mathbb{P}[\mathcal{C}_i]$$

and

$$\mathbb{E}[h(A)] = \sum_{j=1}^{r} \int_{\mathcal{B}_j} h(a) f_{A_j}(a)\mathrm{d}a + \sum_{i=1}^{s} \mathbb{P}[\mathcal{C}_i] h(\alpha_i).$$

In the case of the random Burgers-Riemann problem, we have a particular situation of $r = 2$ and $s = 1$. The pathwise entropy solution of (1)–(2) is

$$U(x,t) = U_L \mathcal{X}_{\mathcal{R}^-(\beta)}(U_L, U_R) + \beta \mathcal{X}_{\mathcal{R}^0(\beta)}(U_L, U_R) + U_R \mathcal{X}_{\mathcal{R}^+(\beta)}(U_L, U_R),$$

where $\beta = x/t$ and the sets of the indicator function give a partition of $\mathbb{R}^2$:

$$\mathcal{R}_r^-(\beta) = \{(u_L, u_R) : u_L < u_R \text{ and } \beta < u_L\},$$

$$\mathcal{R}_r^0(\beta) = \{(u_L, u_R) : u_L < u_R \text{ and } u_L \le \beta \le u_R\},$$

$$\mathcal{R}_r^+(\beta) = \{(u_L, u_R) : u_L < u_R \text{ and } \beta > u_R\},$$

$$\mathcal{R}_s^-(\beta) = \{(u_L, u_R) : u_L > u_R \text{ and } \beta < (u_L + u_R)/2\},$$

$$\mathcal{R}_s^+(\beta) = \{(u_L, u_R) : u_L > u_R \text{ and } \beta > (u_L + u_R)/2\};$$

$$\mathcal{R}^-(\beta) = \mathcal{R}_r^-(\beta) \cup \mathcal{R}_s^-(\beta), \quad \mathcal{R}^0(\beta) = \mathcal{R}_r^0(\beta), \quad \mathcal{R}^+(\beta) = \mathcal{R}_r^+(\beta) \cup \mathcal{R}_s^+(\beta).$$

We illustrate results for the one-dimensional (or first) probability density functions:

Lemma 1 *The density function of $U(x,t)$ is*

$$f_{U(x,t)}(u) = \int_{\mathbb{R}} f_{U_L,U_R}(u,u_R)\mathcal{X}_{\mathcal{R}^-(\beta)}(u,u_R)\mathrm{d}u_R + \int_{\mathbb{R}} f_{U_L,U_R}(u_L,u)\mathcal{X}_{\mathcal{R}^+(\beta)}(u_L,u)\mathrm{d}u_L + \delta(u-\beta)\mathbb{P}[U_L \le \beta \le U_R].$$

Theorem 1 *The density function of $U(x,t)$ is*

$$f_{U(x,t)}(u) = \mathcal{X}_{(\beta,\infty)}(u) \int_{2\beta-u}^{\infty} f_{U_L,U_R}(u,u_R)\mathrm{d}u_R + \mathcal{X}_{(-\infty,\beta)}(u) \int_{-\infty}^{2\beta-u} f_{U_L,U_R}(u_L,u)\mathrm{d}u_L + \delta(u-\beta)\mathbb{P}[U_L \le \beta \le U_R].$$

Corollary 1 *If $h : \mathbb{R} \to \mathbb{R}$ is a Borel measurable deterministic function, then*

$$\begin{aligned}\mathbb{E}[h(U(x,t))] &= \int_{\mathbb{R}} h(u) f_{U(x,t)}(u)\mathrm{d}u \\ &= \int_{\beta}^{\infty} h(u) \int_{2\beta-u}^{\infty} f_{U_L,U_R}(u,u_R)\mathrm{d}u_R\mathrm{d}u\end{aligned}$$

$$+\int_{-\infty}^{\beta} h(u)\int_{-\infty}^{2\beta-u} f_{U_L,U_R}(u_L,u)\mathrm{d}u_L\mathrm{d}u$$
$$+h(\beta)\mathbb{P}[U_L\le\beta\le U_R],$$

whenever the expectation exists.

For the proof, we use the decomposition

$$\begin{aligned} f_{U(x,t)}(u) &= \mathbb{E}[\delta(u-U(x,t))] \\ &= \mathbb{E}[\delta(u-U_L)\mathcal{X}_{\mathcal{R}^-(\beta)}(U_L,U_R)] \\ &\quad +\mathbb{E}[\delta(u-\beta)\mathcal{X}_{\mathcal{R}^0(\beta)}(U_L,U_R)] \\ &\quad +\mathbb{E}[\delta(u-U_R)\mathcal{X}_{\mathcal{R}^+(\beta)}(U_L,U_R)]. \end{aligned} \tag{7}$$

The second addend in (7) is

$$\mathbb{E}[\delta(u-\beta)\mathcal{X}_{\mathcal{R}^0(\beta)}(U_L,U_R)] = \delta(u-\beta)\mathbb{E}[\mathcal{X}_{\mathcal{R}^0(\beta)}(U_L,U_R)] = \delta(u-\beta)\mathbb{P}[U_L\le\beta\le U_R].$$

The first addend in (7) is

$$\begin{aligned} \mathbb{E}[\delta(u-U_L)\mathcal{X}_{\mathcal{R}^-(\beta)}(U_L,U_R)] &= \mathbb{E}[\delta(u-U_L)\mathcal{X}_{\mathcal{R}^-(\beta)}(u,U_R)] \\ &= \mathbb{E}\left[\int_{\mathbb{R}}\delta(u-U_L,u_R-U_R)\mathcal{X}_{\mathcal{R}^-(\beta)}(u,u_R)\mathrm{d}u_R\right] \\ &= \int_{\mathbb{R}}\mathbb{E}\left[\delta(u-U_L,u_R-U_R)\right]\mathcal{X}_{\mathcal{R}^-(\beta)}(u,u_R)\mathrm{d}u_R \\ &= \int_{\mathbb{R}} f_{U_L,U_R}(u,u_R)\mathcal{X}_{\mathcal{R}^-(\beta)}(u,u_R)\mathrm{d}u_R. \end{aligned}$$

The third addend in (7) is computed analogously. By substituting the three terms into (7), we prove Lemma 1 and consequently Theorem 1 and Corollary 1.

The references [1], [2] actually focused on two- and n-dimensional density functions. We illustrate two of the results for the two-dimensional case, with the proofs available in the cited articles:

Theorem 2 *The density function of* $(U(x_1,t_1),U(x_2,t_2))$*, with* $\beta_1=x_1/t_1$ *and* $\beta_2=x_2/t_2$*, is*

$$\begin{aligned} &f_{U(x_1,t_1),U(x_2,t_2)}(u_1,u_2) \\ &=\delta(u_2-u_1)\mathcal{X}_{(\max\{\beta_1,\beta_2\},\infty)}(u_1)\int_{2\max\{\beta_1,\beta_2\}-u_1}^{\infty} f_{U_L,U_R}(u_1,u_R)\mathrm{d}u_R \\ &+\delta(u_2-\beta_2)\mathcal{X}_{(\beta_1,\beta_2]}(u_1)\int_{\beta_2}^{\infty} f_{U_L,U_R}(u_1,u_R)\mathrm{d}u_R \\ &+f_{U_L,U_R}(u_1,u_2)\left(\mathcal{X}_{\{\beta_1<u_1<u_2<\beta_2\}}(u_1,u_2)+\mathcal{X}_{\{u_1>u_2,\beta_1<(u_1+u_2)/2<\beta_2\}}(u_1,u_2)\right) \\ &+\delta(u_1-\beta_1)\mathcal{X}_{(\beta_2,\beta_1]}(u_2)\int_{\beta_1}^{\infty} f_{U_L,U_R}(u_2,u_R)\mathrm{d}u_R \\ &+\delta(u_1-\beta_1,u_2-\beta_2)\mathbb{P}[U_L\le\beta_1\le U_R,U_L\le\beta_2\le U_R] \\ &+\delta(u_1-\beta_1)\mathcal{X}_{[\beta_1,\beta_2)}(u_2)\int_{-\infty}^{\beta_1} f_{U_L,U_R}(u_L,u_2)\mathrm{d}u_L \\ &+f_{U_L,U_R}(u_2,u_1)\left(\mathcal{X}_{\{\beta_2<u_2<u_1<\beta_1\}}(u_1,u_2)+\mathcal{X}_{\{u_2>u_1,\beta_2<(u_1+u_2)/2<\beta_1\}}(u_1,u_2)\right) \\ &+\delta(u_2-\beta_2)\mathcal{X}_{[\beta_2,\beta_1)}(u_1)\int_{-\infty}^{\beta_2} f_{U_L,U_R}(u_L,u_1)\mathrm{d}u_L \end{aligned}$$

$$+\delta(u_2-u_1)\mathcal{X}_{(-\infty,\min\{\beta_1,\beta_2\})}(u_2)\int_{-\infty}^{2\min\{\beta_1,\beta_2\}-u_2} f_{U_L,U_R}(u_L,u_2)\mathrm{d}u_L.$$

Corollary 2 *If $h:\mathbb{R}^2\to\mathbb{R}$ is a Borel measurable deterministic function, then*

$$\begin{aligned}
&\mathbb{E}[h(U(x_1,t_1),U(x_2,t_2))] = \int_{\mathbb{R}^2} h(u_1,u_2) f_{U(x_1,t_1),U(x_2,t_2)}(u_1,u_2)\mathrm{d}u_1\mathrm{d}u_2\\
&= \int_{\max\{\beta_1,\beta_2\}}^{\infty} h(u,u)\int_{2\max\{\beta_1,\beta_2\}-u}^{\infty} f_{U_L,U_R}(u,u_R)\mathrm{d}u_R\mathrm{d}u\\
&+\int_{\beta_1}^{\beta_2} h(u,\beta_2)\int_{\beta_2}^{\infty} f_{U_L,U_R}(u,u_R)\mathrm{d}u_R\mathrm{d}u\mathcal{X}_{\{\beta_1<\beta_2\}}(\beta_1,\beta_2)\\
&+\left(\int_{\beta_1}^{\beta_2}\int_{\beta_1}^{u_2} h(u_1,u_2) f_{U_L,U_R}(u_1,u_2)\mathrm{d}u_1\mathrm{d}u_2\right.\\
&\left.+\int_{-\infty}^{\beta_2}\int_{\max\{2\beta_1-u_2,u_2\}}^{2\beta_2-u_2} h(u_1,u_2) f_{U_L,U_R}(u_1,u_2)\mathrm{d}u_1\mathrm{d}u_2\right)\mathcal{X}_{\{\beta_1<\beta_2\}}(\beta_1,\beta_2)\\
&+\int_{\beta_2}^{\beta_1} h(\beta_1,u)\int_{\beta_1}^{\infty} f_{U_L,U_R}(u,u_R)\mathrm{d}u_R\mathrm{d}u\mathcal{X}_{\{\beta_2<\beta_1\}}(\beta_1,\beta_2)\\
&+h(\beta_1,\beta_2)\mathbb{P}[U_L\le\beta_1\le U_R, U_L\le\beta_2\le U_R]\\
&+\int_{\beta_1}^{\beta_2} h(\beta_1,u)\int_{-\infty}^{\beta_1} f_{U_L,U_R}(u_L,u)\mathrm{d}u_L\mathrm{d}u\mathcal{X}_{\{\beta_1<\beta_2\}}(\beta_1,\beta_2)\\
&+\left(\int_{\beta_2}^{\beta_1}\int_{\beta_2}^{u_1} h(u_1,u_2) f_{U_L,U_R}(u_2,u_1)\mathrm{d}u_2\mathrm{d}u_1\right.\\
&\left.+\int_{-\infty}^{\beta_1}\int_{\max\{2\beta_2-u_1,u_1\}}^{2\beta_1-u_1} h(u_1,u_2) f_{U_L,U_R}(u_2,u_1)\mathrm{d}u_2\mathrm{d}u_1\right)\mathcal{X}_{\{\beta_2<\beta_1\}}(\beta_1,\beta_2)\\
&+\int_{\beta_2}^{\beta_1} h(u,\beta_2)\int_{-\infty}^{\beta_2} f_{U_L,U_R}(u_L,u)\mathrm{d}u_L\mathrm{d}u\mathcal{X}_{\{\beta_2<\beta_1\}}(\beta_1,\beta_2)\\
&+\int_{-\infty}^{\min\{\beta_1,\beta_2\}} h(u,u)\int_{-\infty}^{2\min\{\beta_1,\beta_2\}-u} f_{U_L,U_R}(u_L,u)\mathrm{d}u_L\mathrm{d}u,
\end{aligned}$$

whenever the expectation exists.

Under independence between U_L and U_R, these formulae can be simplified.

Interesting new problems are the following: determine expressions for the density functions under specific random inputs (for example, uniformly distributed); use alternative techniques that do not rely on the delta function operational calculus; and apply these methods to other stochastic systems with mixed distributions.

References

[1] M. Jornet, "Two-dimensional probability distribution of the solution to the random Burgers–Riemann problem," *Physica D: Nonlinear Phenomena*, vol. 453, no. 133800, 2023.

[2] M. Jornet, "Finite-dimensional probability distributions in the random Burgers-Riemann problem," *Communications in Nonlinear Science and Numerical Simulation*, vol. 130, no. 107786, 2024.

Mathematical modeling of crime data with stochastic differential equations

Julia Calatayud[♭],[1]

(♭) Departament de Matemàtiques, Universitat Jaume I,
12071 Castellón, Spain.

1 Introduction

Criminality is a serious problem for any region, which risks its economy, security and quality of life. Mathematical criminology is a current field of research which uses mathematical methods for understanding and predicting the incidence of crime.

There are many works on statistical modeling of crime. There are also works on mathematical modeling of crime, based on high-dimensional ordinary differential equations and partial differential equations [1], [2]. However, although useful for analyzing and simulating system dynamics, these mathematical models are not usually amenable to fitting real crime data, due to the complexities involved in the formulations.

We only found three works on differential equations with real-data fitting: Manchester [3], Cape Town [4], and Los Angeles and Houston [5]. We then investigate differential equation-based models for crime, focusing on real-data fitting. The main objective was the spatio-temporal modeling of crime records. We started with usual models from mathematical epidemiology, considering interactions in crime [6], but we did not find a good way to calibrate parameters in general. Hence we moved to stochastic differential equations, based on quantitative-finance methodologies. We focus on [7], [8]: we show models based on stochastic differential equations to fit crime data, both temporal and spatial, in Spain and in the city of Valencia.

2 Compartmental models

The population is divided into offender and non-offender individuals. An offender is defined as a person who commits an illegal act that should require attention by authorities (for instance, with a fine, community services, or lack of liberty). At a certain time instant, offenders may or may not be in prison. The latter situation may occur when the person has not yet been caught by the police, he/she is waiting for trial or decisions but out of jail, or he/she has been found guilty and is serving sentence but not with penalty of deprivation of liberty.

The compartmental model is

$$\begin{cases} \frac{\mathrm{d}S}{\mathrm{d}t} = \lambda - \beta SL + rP + a_1 L - \mu S, \\ \frac{\mathrm{d}L}{\mathrm{d}t} = \beta SL - a_1 L - a_2 L - \mu L, \\ \frac{\mathrm{d}P}{\mathrm{d}t} = a_2 L - rP - \mu P, \end{cases} \tag{1}$$

[1] calatayj@uji.es

where $S(t)$ is the amount of susceptible individuals, $L(t)$ is the amount of offenders in liberty, and $P(t)$ is the amount of offenders in prison, at instant $t \geq 0$. The total population $N = S + L + P$ is constant. The total number of individuals convinced by an offender is given by the basic reproduction number,

$$\mathcal{R}_0 = \frac{\beta N}{a_1 + a_2 + \mu}. \tag{2}$$

A result is the following:

- If $\mathcal{R}_0 \leq 1$, then the CFE is globally asymptotically stable.
- If $\mathcal{R}_0 > 1$, then the CEE is globally asymptotically stable.

It indicates that if $\mathcal{R}_0 \leq 1$, then criminality is eradicated at long term. Although our proposed model is an idealization of the complex process of criminality, it demonstrates how suitable policies that manipulate the parameters (i.e. the transitions between states) eventually affect the proportion of criminally active people in the population. When $\mathcal{R}_0 > 1$, criminality is not descending and concrete policies should be implemented.

In Spain, with real data on imprisonment, the basic reproduction number is $\mathcal{R}_0 = 0.998$. Since it is less than 1, one concludes that, if things remain the same and the parameters keep time invariant, criminality will become extinct in the future. However, $\mathcal{R}_0$ is actually near the threshold 1, so any variation of the parameters may make it greater than 1 and make criminality endemic. There should be control policies to ensure that $\mathcal{R}_0$ does not grow and, even better, reduces.

The sensitivity index of a parameter λ is

$$\Gamma_\lambda = \frac{\partial \mathcal{R}_0}{\partial \lambda} \times \lambda.$$

In Spain, the sensitivity indices are

$$\Gamma_\beta = 0.998, \quad \Gamma_{a_1} = -0.247, \quad \Gamma_{a_2} = -0.751.$$

Thus, the most important parameter to control is the force of influence, followed by the transitions from L to P and from L to S.

To account for the unexplained variability of the data and obtain probabilistic intervals, a random error is incorporated into the model. This random error is a Gaussian white noise with bounded homogeneous variance. Nonlinear regression is employed as a generalization of linear regression. A 95% confidence interval for $\mathcal{R}_0$ is $[0.968, 1.030]$. See the details in [7].

3 Stochastic models of Itô type

The reader is referred to [8]. Our dataset contains information about reported criminal events in the city of Valencia (Spain) for ten complete years, from 2010 to 2020. We have a total of 90247 events communicated to the 112-emergency phone, split into *aggression* (55610 cases), *stealing* (25342 cases), *woman alarm* (454 cases) and *others* (8841 cases). Each incident is located in Valencia by its latitude and longitude position. We assign a zip code (a patch) to each latitude-longitude location, among the 26 codes existing in Valencia, from 46001 until 46026. This procedure gives 26 time series, along eleven years.

The proposed model of Itô stochastic differential equations, for $n = 26$, is

$$\mathrm{d}x_{1,t} = \mu_1 x_{1,t}\,\mathrm{d}t + \sigma_1 x_{1,t}\,\mathrm{d}B_{1,t}, \quad \ldots, \quad \mathrm{d}x_{n,t} = \mu_n x_{n,t}\,\mathrm{d}t + \sigma_n x_{n,t}\,\mathrm{d}B_{n,t}.$$

The Brownian motions satisfy $\text{corr}[B_{i,t}, B_{j,t}] = \rho_{ij} \in [-1, 1]$. The construction is as follows: given the correlation matrix $\text{A} = (\rho_{ij})_{i,j}$ and given auxiliary independent Brownian motions $\tilde{B}_{1,t} = B_{1,t}$, $\tilde{B}_{2,t}, \ldots, \tilde{B}_{n,t}$, just define

$$\begin{pmatrix} B_{1,t} \\ \vdots \\ B_{n,t} \end{pmatrix} = \text{L} \begin{pmatrix} \tilde{B}_{1,t} \\ \vdots \\ \tilde{B}_{n,t} \end{pmatrix},$$

where L is a lower-triangular matrix and $\text{A} = \text{LL}^\top$ is the Cholesky decomposition of A. A relevant property is $\text{corr}[\text{d}x_{i,t}, \text{d}x_{j,t} | x_{i,t}, x_{j,t}] = \rho_{ij}$.

The key role of correlation: It measures how similar the increasing and decreasing patterns of crime incidence around the expected value are between zip codes. In practice, knowledge of spatial correlations permits reallocating police personnel, given an unexpected escalation of criminal activity in a specific district. The proposed methodology may be applied to fit whole trend time series, obtain the correlations between the zip codes, and forecast incidences at short term.

Acknowledgments

The author Julia Calatayud was funded by "Acció 1.1. Pla de Promoció de la Investigació i Transferència de Coneixement de la Universitat Jaume I de Castelló", Spain, under the project with reference UJI-2024-03.

References

[1] D. McMillon, C. P. Simon, and J. Morenoff, "Modeling the underlying dynamics of the spread of crime," *PLoS ONE*, vol. 9, no. 4, e88923, 2014.

[2] M. Short, A. Bertozzi, and P. Brantingham, "Nonlinear patterns in urban crime: Hotspots, bifurcations, and suppression," *SIAM Journal on Applied Dynamical Systems*, vol. 9, no. 2, pp. 462–483, 2010.

[3] A. A. Lacey and M. N. Tsardakas, "A mathematical model of serious and minor criminal activity," *European Journal of Applied Mathematics*, vol. 27, no. 3, pp. 403–421, 2016.

[4] K. A. J. White, E. Campillo-Funollet, F. Nyabadza, *et al.*, "Towards understanding crime dynamics in a heterogeneous environment: A mathematical approach," *Journal of Interdisciplinary Mathematics*, vol. 24, no. 8, pp. 2139–2159, 2021.

[5] Y. Cao, K. Dong, B. Siercke, and M. Wilber, "Final report: Crime modeling," UCLA, LA, USA, Tech. Rep., Aug. 2013.

[6] F. Brauer, P. Driessche, and J. Wu, *Mathematical Epidemiology*. Berlin Heidelberg: Springer-Verlag, 2008.

[7] J. Calatayud, M. Jornet, and J. Mateu, "A dynamical mathematical model for crime evolution based on a compartmental system with interactions," *International Journal of Computer Mathematics*, vol. 102, no. 1, pp. 44–59, 2025.

[8] J. Calatayud, M. Jornet, and J. Mateu, "Spatio-temporal stochastic differential equations for crime incidence modeling," *Stochastic Environmental Research and Risk Assessment*, vol. 37, pp. 1839–1854, 2023.

Epistemic and Aleatoric Uncertainty Quantification in Bayesian Convolutional Neural Networks: Application to Food Image Classification

Raúl López-Romero[♭], Joaquín Martínez-Minaya[♭], Cristóbal Giménez-Devis[♮], Fernando López-Mir[♮] and Adrián Colomer[♮1]

(♭) Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València (UPV), Valencia, Spain.

(♮) Instituto Universitario de Investigación e Innovación en Tecnología Centrada en el Ser Humano (HUMAN-tech), Universitat Politècnica de València (UPV), Valencia, Spain.

1 Introduction

The classification of food products through computer vision has become a significant research area at the intersection of deep learning and quality assurance [1]. Convolutional neural networks (CNNs) have demonstrated remarkable predictive accuracy in tasks such as defect detection, species identification, and freshness assessment in the food industry [2]. However, a fundamental limitation of standard deterministic CNNs is their inability to quantify the uncertainty associated with their predictions: the model yields a single point estimate without expressing its confidence or, critically, acknowledging what it does not know [3].

This distinction becomes practically relevant in high-stakes scenarios. In the pistachio production sector, the varieties *Kirmizi* and *Siirt* serve markedly different markets: the former is a protected-designation product used in premium confectionery such as Gaziantep baklava, while the latter is oriented towards direct consumption. Misclassifying one for the other carries legal, economic, and reputational consequences for the producing regions [2]. In such contexts, knowing *when not to trust* the model is as important as the model's raw accuracy.

The field of uncertainty quantification (UQ) in deep learning addresses this gap [4]. Two principal sources of uncertainty are distinguished: *aleatoric* uncertainty, arising from inherent variability or noise in the data and irreducible by design; and *epistemic* uncertainty, reflecting the model's lack of knowledge and reducible with additional data. Capturing both requires probabilistic formulations that go beyond standard maximum likelihood training.

In this work, we present a systematic comparison of three modelling approaches of increasing probabilistic richness applied to pistachio image classification: (i) a deterministic CNN with probabilistic output interpretation via *softmax*; (ii) Monte Carlo Dropout (MCDropout) [3], an approximate Bayesian method; and (iii) a fully Bayesian CNN with variational inference [5]. All three are built upon the VGG16 architecture [6] via transfer learning, enabling a fair comparison.

[1] adcogra@upv.es

Models are evaluated not only on standard classification metrics but also on uncertainty-specific measures including negative log-likelihood (NLL), predictive entropy, and predictive variance, and are further tested on out-of-distribution (OOD) data — a critical probe of whether uncertainty estimates are practically meaningful.

2 Mathematical Framework

2.1 Bayesian Learning for Neural Networks

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ denote the training set of image–label pairs. In a standard CNN, the network weights $\mathbf{w}$ are fixed parameters estimated by minimising a loss function. In the Bayesian approach, $\mathbf{w}$ are treated as random variables with prior $p(\mathbf{w})$, and learning consists of computing the posterior [7]:

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w})\, p(\mathbf{w})}{p(\mathcal{D})}, \tag{1}$$

where $p(\mathcal{D} \mid \mathbf{w}) = \prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i, \mathbf{w})$ is the likelihood and $p(\mathcal{D}) = \int p(\mathcal{D} \mid \mathbf{w})\, p(\mathbf{w})\, d\mathbf{w}$ is the intractable marginal likelihood. Predictions for a new input $\mathbf{x}^*$ are obtained via the posterior predictive:

$$p(y \mid \mathbf{x}^*, \mathcal{D}) = \int p(y \mid \mathbf{x}^*, \mathbf{w})\, p(\mathbf{w} \mid \mathcal{D})\, d\mathbf{w}, \tag{2}$$

approximated in practice, this integral is approximated using Monte Carlo sampling: one draws T samples $w^{(t)}$ from the approximate posterior and computes:

$$p(y \mid x^*, D) \approx \frac{1}{T} \sum_{t=1}^{T} p(y \mid x^*, w^{(t)}). \tag{3}$$

drawn from a posterior approximation.

2.2 Variational Inference and the ELBO

Variational inference [5] approximates the intractable posterior by a tractable distribution $q_{\boldsymbol{\lambda}}(\mathbf{w})$, typically a mean-field Gaussian parameterised by $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \boldsymbol{\sigma}\}$. The optimal $\boldsymbol{\lambda}$ minimises $\mathrm{KL}[q_{\boldsymbol{\lambda}}(\mathbf{w}) \| p(\mathbf{w} \mid \mathcal{D})]$, equivalent to maximising the Evidence Lower BOund (ELBO):

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q_{\boldsymbol{\lambda}}}[\log p(\mathcal{D} \mid \mathbf{w})] - \mathrm{KL}[q_{\boldsymbol{\lambda}}(\mathbf{w}) \| p(\mathbf{w})]\,. \tag{4}$$

The first term rewards data fidelity; the second regularises the variational posterior towards the prior and controls model complexity. Gradients with respect to $\boldsymbol{\lambda}$ are computed via the reparameterisation trick [8]: one draws $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and sets $\mathbf{w} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\varepsilon}$, making the sampling step differentiable. A Gaussian prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{0},\, 0.09\,\mathbf{I})$ is used throughout.

2.3 Monte Carlo Dropout as Approximate Bayesian Inference

Gal and Ghahramani [3] showed that a network trained with dropout and evaluated with dropout *active at test time* constitutes approximate variational inference in a deep Gaussian process. Performing T stochastic forward passes yields a sample set $\{\hat{\mathbf{p}}^{(t)}\}_{t=1}^{T}$ from which moments of the predictive distribution can be estimated, without any modification to the training procedure.

2.4 Uncertainty Metrics

Three complementary metrics characterise predictive uncertainty [4, 7].

Negative log-likelihood (NLL):

$$\mathrm{NLL}(\mathbf{x}) = -\log \hat{p}_y. \tag{5}$$

Predictive entropy (total uncertainty):

$$H[\hat{\mathbf{p}}] = -\sum_k \bar{p}_k \log \bar{p}_k, \qquad \bar{\mathbf{p}} = \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{p}}^{(t)}. \tag{6}$$

Predictive variance (epistemic uncertainty):

$$\mathrm{Var}[\hat{p}_k] = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{p}_k^{(t)} - \bar{p}_k\right)^2. \tag{7}$$

Predictive variance is only available for MCDropout and Bayesian models, since deterministic networks produce a single fixed output. Table 1 summarises the uncertainty types each approach captures.

Table 1: Uncertainty types modelled by each approach.

Model	Aleatoric	Epistemic
Deterministic CNN	—	—
Probabilistic CNN	✓	—
MCDropout	✓	✓
Bayesian CNN (VI)	✓	✓

3 Methodology

3.1 Dataset and Preprocessing

The Pistachio Image Dataset [9] comprises 2,148 RGB images (600×600 px) of two Turkish pistachio varieties: 1,232 *Kirmizi* and 916 *Siirt*. A stratified split yielded 70% training (1,203 images), 10% validation (515), and 20% test (430). The mild class imbalance (57%/43%) was addressed by data augmentation on the minority class (rotations $\leq 25°$, horizontal flips, zoom $\leq 20\%$, brightness $\pm 20\%$), applied exclusively to training data. For OOD evaluation, 20 additional images were compiled: 10 California pistachios and 10 cashews, combining real photographs and generative-AI-produced images.

3.2 Architecture and Transfer Learning

Three pretrained backbones were evaluated (VGG16, InceptionV3, EfficientNetB0). VGG16 [6] consistently outperformed the others (accuracy 0.94, F1 ≥ 0.94 for both classes), which we attribute to its stack of small convolutional filters being especially sensitive to the subtle textural and morphological differences between the two varieties. All subsequent experiments use VGG16 as the backbone. Transfer learning followed two phases: (i) frozen backbone, training only the classification head; (ii) fine-tuning of the last two convolutional blocks with a reduced learning rate ($\eta = 10^{-5}$, SGD, momentum $\mu = 0.9$).

3.3 Model Configurations

(i) Probabilistic CNN. Standard VGG16 fine-tuned with cross-entropy loss. *Softmax* outputs serve as aleatoric probability estimates; no epistemic uncertainty is captured.
(ii) MCDropout. VGG16 with dropout ($p = 0.5$) kept active at inference. $T = 100$ stochastic forward passes per sample estimate the predictive mean, variance (Eq. 7), and entropy (Eq. 6).
(iii) Bayesian CNN (VI). The dense classification head is replaced by variational layers in which each weight follows $\mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$. The ELBO (Eq. 4) is minimised via the reparameterisation trick with prior $\mathcal{N}(\mathbf{0},\, 0.09\,\mathbf{I})$. The convolutional backbone retains its pretrained weights. Both transfer-learning phases use 50 epochs to allow the variational parameters to stabilise. $T = 100$ MC samples are drawn at inference. All models were implemented in Python with Keras/TensorFlow and TensorFlow Probability (v0.25.0), and trained on Google Colab with GPU acceleration.

4 Results and Discussion

4.1 Classification Performance

Table 2 reports the main classification metrics on the held-out test set.

Table 2: Classification performance on the test set (430 images). P = Precision; R = Recall; F1 = F1-score.

	Kirmizi			**Siirt**			**Acc.**
Model	P	R	F1	P	R	F1	
Probabilistic CNN	0.97	0.94	0.95	0.92	0.96	0.94	0.94
MCDropout	0.95	0.96	0.95	0.95	0.94	0.94	0.95
Bayesian CNN (VI)	0.91	0.92	0.91	0.89	0.88	0.88	0.90

All three models achieve competitive accuracy. MCDropout marginally improves upon the deterministic baseline (95% vs. 94%), while the Bayesian CNN trades a moderate accuracy reduction (90%) for substantially richer uncertainty information. The accuracy reduction is expected: modelling each weight as a distribution introduces a regularisation effect that, given the limited dataset relative to the variational parameter count, mildly constrains predictive capacity. No significant bias between classes is observed in any model, confirming that the augmentation strategy successfully addressed the class imbalance.

4.2 Uncertainty Analysis on the Test Set

Table 3 summarises the uncertainty metrics disaggregated by prediction outcome.

In the probabilistic CNN, 75% of correct predictions are made with confidence exceeding 90% (NLL < 0.1). Incorrect predictions, however, show a bimodal behaviour: about half occur with moderate confidence (genuine ambiguity, NLL < 0.9), while the remainder are overconfident errors (NLL > 1.5), particularly on the *Siirt* class — precisely the failure mode that probabilistic models aim to address.

MCDropout successfully differentiates correct from incorrect predictions through elevated variance and entropy in the latter group ($\overline{\text{Var}}$: 0.025 vs. 0.007; $\overline{H}$: 0.44 vs. 0.23). The model is therefore aware of its uncertainty in difficult cases. Absolute variance values remain low, however, reflecting the limitation of applying dropout exclusively to the classification head. Extending MCDropout

Table 3: Mean uncertainty metrics on the test set by prediction outcome.

Model	**Subset**	$\overline{\text{NLL}}$	$\overline{H}$	$\overline{\text{Var}}$
Prob. CNN	Overall	0.153	0.250	—
	Correct	0.099	0.230	—
	Incorrect	1.187	0.600	—
MCDropout	Overall	0.168	0.250	0.0076
	Correct	—	0.230	0.0066
	Incorrect	—	0.440	0.0251
Bayesian (VI)	Overall	0.326	0.021	0.301
	Correct	—	0.019	—
	Incorrect	—	0.032	0.450

to block 4 onward of VGG16 yields a fourfold increase in epistemic variance on incorrect predictions without significant accuracy degradation — a finding directly relevant to deployment-ready calibration.

The Bayesian CNN (VI) exhibits a qualitatively different uncertainty profile: very high predictive variance ($\overline{\text{Var}} = 0.301$) reflecting substantial epistemic uncertainty, consistent with the limited dataset relative to the variational parameter space. When the model errs, variance rises sharply (up to 0.45), confirming that epistemic uncertainty correlates with prediction failure — a desirable and practically useful property.

4.3 Out-of-Distribution Behaviour

Figure 1 shows the distributions of uncertainty metrics for in-distribution test images (*known*: pistachios) versus OOD images (*unknown*: California pistachios and cashews) for each model. The probabilistic CNN assigns higher NLL to OOD samples on average, yet a non-negligible proportion of OOD images receive low NLL, indicating overconfident misclassifications — the classic failure mode of deterministic networks.

MCDropout provides a partial improvement: OOD samples exhibit higher entropy and variance than in-distribution ones, suggesting the model detects some distributional shift. The known–unknown separation remains moderate, with noticeable overlap between both distributions, consistent with the head-only dropout limitation discussed above.

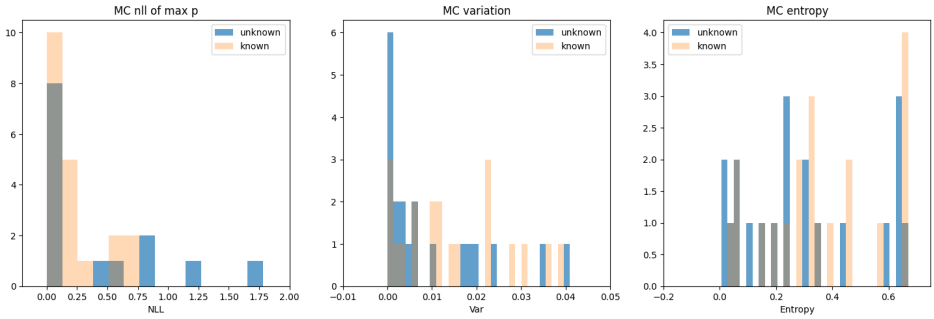
The Bayesian CNN reveals the most nuanced behaviour. Some OOD samples — particularly California pistachios — receive *lower* variance than in-distribution ones, likely because their visual appearance activates feature representations learned for *Kirmizi*. This highlights that, despite its theoretical advantages, the model can still produce overconfident predictions on unfamiliar data. This illustrates an important general principle: Bayesian epistemic uncertainty is not equivalent to OOD detection — it can remain low for inputs that happen to align with learned representations despite belonging to an unknown class.

5 Conclusions

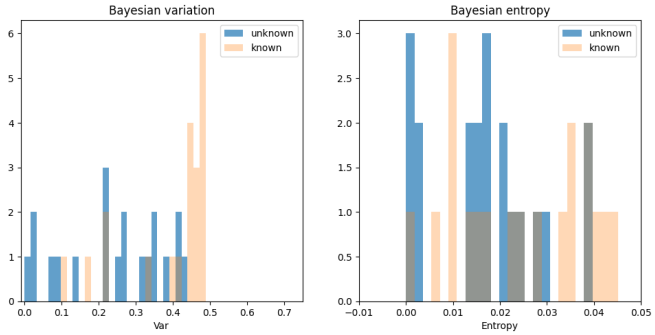
We have presented a rigorous probabilistic framework for uncertainty-aware image classification in a food industry setting. The key contributions are: (i) a unified comparison of probabilistic CNN, MCDropout, and variational Bayesian CNN on the same VGG16 backbone, characterising the accuracy–uncertainty trade-off; (ii) quantitative evidence that MCDropout's epistemic uncertainty

Figure 1: Distributions of uncertainty metrics for in-distribution test images versus OOD images

(a) Metrics for comparing performance with both image sets.



(b) Metrics for comparing the *MC Dropout* approach.



(c) Metrics for comparing the Bayesian approach.

improves significantly when dropout is extended to deeper convolutional blocks, without accuracy loss; (iii) empirical demonstration that the Bayesian CNN captures greater epistemic uncertainty than MCDropout, and that both show limited OOD detection for visually similar unknown classes; and (iv) an illustration that uncertainty-aware models provide actionable confidence estimates that purely deterministic models cannot, which is critical in high-stakes food authentication settings. Future work will extend variational layers to convolutional blocks and investigate temperature scaling for improved calibration.

Acknowledgments

This work is part of the Master's Thesis of Raúl López Romero in the *Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones* at Universitat Politècnica de València.

References

[1] L. Zhou, C. Zhang, F. Liu, Z. Qiu, and Y. He, "Application of deep learning in food: A review," *Comprehensive Reviews in Food Science and Food Safety*, vol. 18, no. 6, pp. 1793–1811, 2019.

[2] D. Singh, Y. S. Taspinar, R. Kursun, I. Cinar, M. Koklu, I. A. Ozkan, and H.-N. Lee, "Classification and analysis of pistachio species with pre-trained deep learning models," *Electronics*, vol. 11, no. 7, p. 981, 2022.

[3] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, vol. 48, pp. 1050–1059, 2016.

[4] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5574–5584, 2017.

[5] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, vol. 37, pp. 1613–1622, 2015.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[7] O. Dürr, B. Sick, and E. Murina, *Probabilistic Deep Learning with Python, Keras and TensorFlow Probability*. Shelter Island, NY: Manning Publications, 2020.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[9] M. Köklü, "Pistachio image dataset," 2021. Available at: `https://www.muratkoklu.com/datasets/`.

Uncertainty Quantification in Inverse Identification Problem via Polynomial Chaos Expansions

Roberta Lima[♭],[1] Rubens Sampaio[♭], Eduardo S. de Cursi[♮], and Emmanuel Pagnacco[♮]

(♭) Laboratório de Dinâmica e Vibrações, Departamento de Engenharia Mecânica, PUC-Rio, Rua Marquês de São Vicente, 255, Gávea, Rio de Janeiro, 22451-00, RJ, Brazil.

(♮) Laboratoire de Mécanique de Normandie, Normandie Université, INSA Rouen Normandie 685, Avenue de l'Université Saint-Etienne du Rouvray, Rouen, France.

1 Introduction

Considering uncertainties becomes a major issue in current engineering design and in mechanical inverse problems. In this work, we consider the identification of the elastic modulus of a homogeneous beam which is subjected to a four points static bending test. But, if there is uncertainties in measurements, the parameter identification which results from the solution of an optimization problem shows that it has a statistical distribution that is not known *a priori*. Thus, we aim at taking into account some uncertainties throughout the identification process, in order to model the desired elastic modulus.

More precisely, the uncertainties are taken into account by modeling each of them as a random variable. The input random vector is then transformed into a random variable given as the output of the identification procedure, based on the minimization of a residual involving the input and the analytical model. Then, the compound model is represented by a chosen polynomial chaos expansion. Two simple examples are used to show the methodology, each considering different hypotheses about the uncertainty in the loading.

This study is organized as follows. The deterministic four point bending test is described and a model of it is presented in the first Section. In this section, the linear Euler-Bernoulli model for the static response of a beam in bending is introduced, taking into account the boundary conditions of the four point bending test. The deterministic optimization problem for the identification of the elastic modulus is presented, when the measurements, the test, and the analytical model, are free from errors, but this is not shown due to lack of space. Then, the problem is recast by considering there are uncertainties in the measured displacements as well as in the total load measurements. Uncertainties in the positions of the loads was also done but are not shown in this paper because they give no new results. To represent the results of this identification problem, the random elastic modulus, representations by polynomial chaos expansion are used. We then focus on the numerical strategies that we have used to determine the coefficients of the polynomial expansion, and the results are compared. Unfortunately, only parts of the results are shown.

As a control of the results we have used Monte Carlo simulations [1], [2], but no discussion is presented. This methodology was used since it is standard and it is good enough for our proposes.

[1] robertalima@puc-rio.br

2 Toy problem: four point bending test description and analytical model

The problem we want to model is a three dimensional body, a slab, say, but here we discuss a toy problem to simplify the description. Let a uniform homogeneous specimen be subjected to a static four-point bending test. It is a classical experiment in mechanics to determine the elastic modulus e of the specimen. This test consists in placing the beam specimen, of length l, supported by its ends, in contact with two pin, or roller supports, where two concentrated loads, q_1 and q_2, as sketched below, will act (see Fig. 1). The distances between the left side support and the actuator pins are l_1 and l_2, respectively. This geometry is often used in material testing due to its convenience, as it avoids the need to clamp the specimen to the testing apparatus.

Modeling the beam as an one-dimensional object, and using the Euler-Bernoulli theory, we get the relationship between the beam's deflection and the applied load as:

$$e\, i_{yy} \frac{\partial^4 w}{\partial x^4} = q_1 \delta\left(x - l_1\right) + q_2 \delta\left(x - l_2\right) \qquad \text{for} \quad 0 < x < l \tag{1}$$

where $w\left(x\right)$ is the deflection of the beam in the transverse direction at some position x, $\delta\left(x\right)$ is the Dirac function and i_{yy} is the second moment of area of the beam's cross-section. It is calculated with respect to the axis which passes through the centroid of the cross-section being perpendicular to the applied loading. For the four-point bending test, boundaries conditions are:

$$\begin{cases} w = 0 \quad \text{and} \quad \frac{\partial^2 w}{\partial x^2} = 0 \qquad \text{for} \quad x = 0 \\ w = 0 \quad \text{and} \quad \frac{\partial^2 w}{\partial x^2} = 0 \qquad \text{for} \quad x = l \end{cases} \tag{2}$$

The application of two point loads at different locations leads to $w\left(x\right)$ being a sum of piecewise

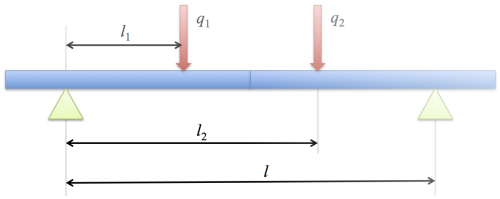


Figure 1: Sketch of the four point bending test

functions:

$$w\left(x;\, e,\, q_1, q_2\right) = \frac{\left(l - l_1\right) q_1}{6 e\, i_{yy} l} \left(\frac{l}{l - l_1} \left\langle x - l_1 \right\rangle^3 - x^3 + l_1 \left(2l - l_1\right) x \right) + \frac{\left(l - l_2\right) q_2}{6 e\, i_{yy} l} \left(\frac{l}{l - l_2} \left\langle x - l_2 \right\rangle^3 - x^3 + l_2 \left(2l - l_2\right) x \right)$$

where $\langle \bullet \rangle$ are singularity brackets such that $\langle x - b \rangle^n = \frac{(x-b)^n}{n+1} H\left(x - b\right)$ for integer n, $n \geq 0$ and $H\left(\bullet\right)$ is the Heaviside step function: $H\left(x\right) = 0$ for $x < 0$ and $H\left(x\right) = 1$ for $x > 0$.

The four point bending test is classically designed to have a symmetry, when one chooses $q_1 = q_2$ and $l_2 = l - l_1$.

Identification problem in presence of uncertainties

We are now interested in the identification results obtained when there are uncertainties in the displacement measurements and in the loading. Since the loading is driven to respect the requested load in the experiments, it is assumed that the error comes from a difference between the true

loading and the load measurement. Then, denoting random variables by capital letters, the applied loads Q_1 and Q_2 are now modeled as random variables, such that:

$$Q_1 = \frac{1}{2}\left(\bar{q} + \xi_q\right) \qquad \text{and} \qquad Q_2 = \frac{1}{2}\left(\bar{q} + \xi_q\right) \tag{3}$$

where ξ_q is a random variable which follows an uniform law:

$$\xi_q \sim \mathcal{U}\left(-\delta_q,\, +\delta_q\right) \qquad \text{for} \qquad \delta_q = c_q \bar{q} \tag{4}$$

Moreover, the measured displacements vector becomes a random variable vector $\mathbf{W}_{\text{meas}}$ such that:

$$\mathbf{W}_{\text{meas}} = \mathbf{w}\left(\mathbf{x};\, e_{\text{true}},\, Q_1,\, Q_2\right) + \boldsymbol{\xi}_w \tag{5}$$

where $\boldsymbol{\xi}_w$ is a random vector which follows a multi-normal law:

$$\boldsymbol{\xi}_w \sim \mathcal{N}_5\left(0,\, \sigma_w\right) \qquad \text{for} \qquad \sigma_w = c_w \max\left(\left|\mathbf{w}_{\text{meas}}\right|\right), \tag{6}$$

assuming the noise in measurements have no correlation. Hence, the coefficients of variation c_q and c_w enable to drive the level of noise in measurements.

Then, assuming the known information is identical to the one of the deterministic situation, we recast the identification problem as:

$$E^* = \arg_{e \in \mathbb{R}^+} \min \left(\sum_{k=1}^{5} \left(\left\{\mathbf{W}_{\text{meas}}\right\}_k - w\left(x_k;\, e,\, \tfrac{1}{2}\bar{q},\, \tfrac{1}{2}\bar{q}\right)\right)^2 \right) \tag{7}$$

where the optimal solution E^* becomes now a random variable having an unknown distribution [3]. In other words, we have defined here the stochastic model that leads to the random variable output E^* from an input random variable vector $\mathbf{Z} = \{Q_1; \mathbf{W}_{\text{meas}}\}$, in the form $E^* = \mathcal{H}\left(\mathbf{Z}\right)$, $\mathcal{H}$ being the mathematical function $\mathcal{H} : \mathbb{R}^6 \mapsto \mathbb{R}$ that model the inverse problem. Its output depends on the coefficients of variation c_q and c_w. At the limit, when all these coefficients of variation tend to be null, the previous deterministic situation would be recovered, to produce $E^* = e_{\text{true}}$ for the distribution of the optimal solution. Finally, one has to note that the centered statistical moments defined by:

$$\mu_i = \mathrm{E}\left[\left(E^* - \mathrm{E}\left[E^*\right]\right)^i\right] \qquad \text{for } i \geq 2 \tag{8}$$

are also valuable quantities to characterize the RV E^* and they will be used in the following sections.

3 Modeling of the random variable optimal solution

3.1 Polynomial chaos expansion

Let the stochastic quantities considered to be in the probability space (Ω, A, Γ) with space of events Ω, σ -algebra A, and probability measure Γ. Then, to model the random variable (RV) optimal solution E^*, we choose to represent it using a polynomial expansion having the form :

$$E_p^* = \sum_{k=0}^{p} a_k \psi_k\left(y\right), \qquad p \in \mathbb{N} \tag{9}$$

where p denotes the order of expansion, a_k are the expansion coefficients to be determined and $\psi_k\left(y\right)$ are the polynomials forming the orthogonal basis $\{\psi_0, \ldots, \psi_p\}$ with respect to the measure

Γ. Hence, for Y a suitable random variable having f_{Y} as the probability density function (PDF) defined onto the support $\mathcal{S}$, the following useful property holds:

$$\int_{y\in\mathcal{S}} \psi_l(y)\,\psi_k(y)\,f_{\mathrm{Y}}(y)\,\mathrm{d}y = \mathrm{E}\left[\psi_l(y)\,\psi_k(y)\right] = \delta_{kl} \tag{10}$$

where $\mathrm{E}[\bullet]$ denotes the expected operator and δ_{kl} the Kronecker symbol.

Moreover, for a proper Monte-Carlo (MC) simulation based strategy, it is interesting to express the link between RVs. Then, denoting F_{Y} the cumulative distribution function (CDF) of the RV Y and F_{E^*} the CDF of E^*, we introduce the isoprobabilist transformation $E^* \longmapsto Y$ such that $F_{E^*}(e) = F_Y(y)$, leading to:

$$e = F_{E^*}^{-1}\left(F_{\mathrm{Y}}(y)\right) \qquad \text{or} \qquad y = F_Y^{-1}\left(F_{E^*}(e)\right) \tag{11}$$

3.2 Numerical strategies

Monte-Carlo simulations enable us to have access to "measurements" vector data $\hat{\mathbf{E}}^*$ which is a sampling of E^* enabling us to construct an empirical CDF $\widehat{F}_{E^*}(e)$. Next, we can use numerical strategies to obtain the expansion coefficients, when limiting the order of expansion, p, to a small, practical value. For this, we consider the variational method, the collocation or least squares method, and the moment method in their straightforward form since the interest here is not in efficiency but in applicability.

The variational method exploits the orthogonality of the polynomial basis. From relations Eq. (9) and (10) we have:

$$\int_{y\in\mathcal{S}} \psi_k(y)\,E^* f_{\mathrm{Y}}(y)\,\mathrm{d}y = a_k \int_{y\in\mathcal{S}} \psi_k(y)\,\psi_k(y)\,f_{\mathrm{Y}}(y)\,\mathrm{d}y \tag{12}$$

and, using the relation Eq. (11):

$$a_k = \frac{\mathrm{E}\left[\psi_k\left(y\left(e\right)\right)E^*\right]}{\mathrm{E}\left[\psi_k^2\left(y\left(e\right)\right)\right]}. \tag{13}$$

The collocation method consists in solving directly in the least squares sense the relation Eq. (9) for the over-determined set of collected data. Hence:

$$a_k = \arg\min\left(\|R\|\right) \qquad \text{where} \qquad R = E^* - \sum_{k=0}^{p} a_k\psi_k\left(y\left(e\right)\right). \tag{14}$$

But, for the data considered here, it is obvious that both the variational method and the collocation method will produce the same results, since relations Eq. (13) and (14) leads to the same expressions when using the vector data sampling $\hat{\mathbf{E}}^*$.

At last, the moment method consists in minimizing the gap between the center moments of E^* and the ones given by the expansion:

$$a_k = \arg\min\left(\sum_{i=2}^{p+1}\left(\mathrm{E}\left[\left(E^*\right)^i\right] - \mathrm{E}\left[\left(\sum_{k=0}^{p} a_k\psi_k\left(y\left(e\right)\right)\right)^i\right]\right)^2\right) \tag{15}$$

However, for a better numerical conditioning (this is a crucial point, the components of the vector must have equal weights in the optimisation otherwise there is problems), we express this relation as:

$$\begin{cases} a_0 = \mathrm{E}\left[E^*\right] \\ a_k = \arg\min\left(\displaystyle\sum_{i=2}^{p+1}\left(1 - \tfrac{1}{\mu_i}\mathrm{E}\left[\left(\sum_{k=1}^{p} a_k\psi_k\left(y\left(e\right)\right)\right)^i\right]\right)^2\right) & \text{for } 1 \le k \le p. \end{cases} \tag{16}$$

Since this leads to a nonlinear set of equations, a tolerance value ϵ has to be defined in a relative stopping criterion.

Table 1: Statistical moments and major quantiles of E^* for the two applications

App.	Mean E $[E^*]$	Variance μ_2	Skewness $\mu_3\mu_2^{-\frac{3}{2}}$	Kurtosis $\mu_4\mu_2^{-2}$	Quantile 5%	Quantile 95%
1	200	1.57	0.02	2.13	198	202
2	201	1.19	-0.34	4.5	199	202

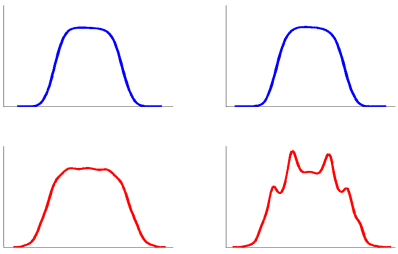


Figure 2: Results for the first application: left, empirical PDF of E^* ; right, empirical PDF issued from the polynomial expansion at the order 15; blue curves are for the Hermite polynomial basis while red curves are for the Chebyshev of 2nd kind polynomial basis; left curves are from the variational method while right curves are from the moments method

4 Conclusions

The paper addresses two main issues. 1) Uncertainty quantification problem: how the distributions of different kind of noise (on both the input and output) affect the distribution of the minimizer, the elasticity modulus, in a static bending test with several discrete measurements of the deformation of a beam?; 2) A modelization problem: how to model the uncertainty using a polynomial chaos expansion in order to avoid using only the two firsts moments to plot an envelope graph, as one sees frequently in the literature.

Acknowledgments

The authors thanks the support of Faperj, CNPq, and INSA-Rouen.

References

[1] J. E. Souza de Cursi and R. Sampaio, *Uncertainty quantification and stochastic modeling with Matlab*. London: Elsevier, 2015, ISBN: 978-1-78548-005-8. DOI: 10.1016/C2014-0-04713-2

[2] R. Sampaio and R. Lima, *Modelagem estocástica e geração de amostras de variáveis e vetores aleatórios*. São Carlos, Brazil: Notas em Matemática Aplicada, SBMAC, 2012, vol. 70, ISBN: 2175-3385.

[3] R. Lima and R. Sampaio, "What is uncertainty quantification?" *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, vol. 40, p. 155, 2018.

Bayesian Inference of Contagion Parameters in Branching Processes under Additive Noise

João Pedro Freitas[♭], Roberta Lima[♭],[1] Rubens Sampaio[♭], and Eduardo S. de Cursi[♯]

(♭) Pontifícia Universidade Católica do Rio de Janeiro,
Rua Marquês de São Vicente, 255, Gávea, Rio de Janeiro, 22451-00, RJ, Brazil.

(♯) Laboratoire de Mécanique de Normandie, Normandie Université, INSA Rouen Normandie
685, Avenue de l'Université Saint-Etienne du Rouvray, Rouen, France.

1 Introduction

Individual-level models study the propagation of epidemics over time. These models assume that individual behavior plays a major role in the infected population size. Transmissibility [1] is the key factor. In order to model this concept, randomness must be considered due to the inherent uncertainty, for example, in genomic structures or human actions. Therefore, a proper way is to model transmissibility as a discrete random variable. As a consequence, population randomness over time relies on individual's uncertainty. This is indeed the concept of demographic stochasticity, which can be modeled as a stochastic process. The Bienaymé-Galton-Watson (BGW) process is in agreement with this perspective. It has been used to study deductive reasoning problems in the epidemiological context [2]. Moreover, a plausible reasoning problem of understanding transmissibility with observations from this branching process is presented in Freitas et al. (2026). More specifically, this reference shows that it is possible to perform parametric Bayesian inference of the random variable that models transmissibility in light of data from the BGW. This paper focuses on quantifying the impact of noisy data when introduced in the proposed Bayesian approach.

Data come from different realizations of a specific generation's random variable from the BGW. The Bayesian updating process ends when tolerance, given by L2-Wasserstein distance, is reached. Since data generation is stochastic, the number of updates required for convergence is also a random variable. Therefore, different updating processes are simulated in order to study the effect of additive noise. Noisy data is introduced in different levels for each one of the Bayesian updating process according to a discrete uniform distribution, whose support relies on the standard deviation of the generation's random variable. The mode of each last posterior distribution is analyzed to understand the impact of noise intensity and the robustness of the Bayesian approach.

2 Epidemic propagation over time

The BGW process $\mathcal{X} = \{X_t\}$ is used to model epidemic propagation over time. Each random variable X_t from the family of random variables models the infected population size at discrete time t. In this paper, the discrete random variable named contagion C represents the number of new successful infections per previously infected individual. It is assumed that the contagion follows

[1] robertalima@puc-rio.br

a Binomial law, $C \sim \text{Binomial}(m, p)$, where in the epidemiological context m is the maximum number of people a single infector can meet and p is the probability of infecting each one of them. The population size of a generation according to [3] is given by

$$X_{t+1} = \sum_{k=0}^{X_t} C_k,\ t \in \mathbb{N}. \tag{1}$$

This stochastic process assumes that the contagions C_k are independent and identically distributed (i.i.d.) random variables. Moreover, when the size of any generation X_k, $k \in \mathbb{N}$ is zero, the extinction of the disease occurs. In this scenario, for any further value $t > k$, X_t is also zero.

3 Bayesian parametric inference problem

Is it possible to learn any information from C with observations of the branching process? [2] introduces two strategies to use data from the BGW to do a Bayesian parametric inference problem. The dataset $\{X_t\}_k$ from the first comes from k different realizations of the same generation's random variable X_t. The other strategy observes a realization of the stochastic process up to a certain generation. The former strategy is applied in this paper. Figure 1 shows different realizations of X_2, leading to the set $\{X_2\}_3 = \{3, 4, 2\}$.

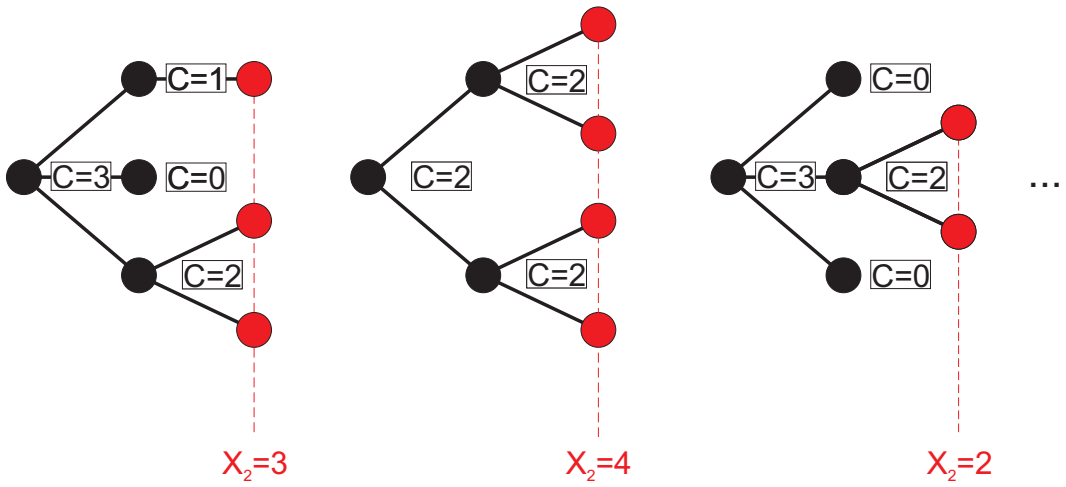


Figure 1: Different realizations of the second generation from a branching process.

The fundamental aspect of the Bayesian parametric inference problem is how to build proper likelihood functions based on the dataset. According to [2], Eq. (2) builds a likelihood function in terms of the probability generating function (pgf) of the contagion random variable G_C:

$$L(\boldsymbol{w}) = \frac{1}{\hat{x}_t!} \frac{d^{(\hat{x}_t)}\left[G_C\left(G_C\left(\ldots\left(G_C\left(s, \boldsymbol{w}\right)\right)\right)\right)\right]}{ds^{(\hat{x}_t)}}\bigg|_{s=0}, \tag{2}$$

where $\boldsymbol{w}$ is the vector representing the target parameters in the inferential process and $\hat{x}_t$ is the observed realization of the random variable X_t.

This paper focuses on the one-dimensional inference problem of the parameter p from $C \sim \text{Binomial}(2, p)$, whose pgf is given by

$$G_C(s, p) = [(1 - p) + ps]^2. \tag{3}$$

Finally, the iterative Bayesian updating process is represented as

$$f_P\left(p \mid \{\hat{x}_t\}\right) \propto \frac{1}{\hat{x}_t!} \left. \frac{d^{(\hat{x}_t)} \overbrace{\left[(1-p)+p\left[(1-p)+p\left[\ldots\left[(1-p)+p\,s\right]^2\right]^2\right]^2\right]^2}^{t-1 \text{ recurrences}}}{ds^{(\hat{x}_t)}} \right|_{s=0} f_P\left(p \mid \{\hat{x}_{t-1}\}\right). \tag{4}$$

There is no coherent information available to ascribe features to the first prior distribution $f_P(p)$. Hence, it follows a uniform distribution $\mathcal{U}[0,1]$.

Since the data in this paper are artificially generated, a convergence criterion $\zeta_{\mathcal{W}_2}$ is required to stop the updating process. This criterion is based on L_2-Wasserstein distance $\mathcal{W}_2(\bullet,\bullet)$ [4], [5] between $f_P(p|\{\hat{x}_t\})$ and $f_P(p|\{\hat{x}_{t-1}\})$,

$$\mathcal{W}_2\left(f_P\left(p \mid \{\hat{x}_t\}\right), f_P\left(p \mid \{\hat{x}_{t-1}\}\right)\right) = \left[\int_0^1 \left|F_P\left(p \mid \{\hat{x}_t\}\right)^{-1}(z) - F_P\left(p \mid \{\hat{x}_{t-1}\}\right)^{-1}(z)\right|^p dz\right]^{1/2}, \tag{5}$$

for this aim. In Eq. (5), $F_P(p|\{\hat{x}_t\})$ and $F_P(p|\{\hat{x}_{t-1}\})$ are the cumulative distribution functions of the posterior and prior distributions respectively.

As a consequence of the uncertainty from the dataset generation, the total number of updates until convergence is also a random variable N.

4 Additive noise

The aim of this paper is to clarify the impact of noisy data in the proposed Bayesian updating process for the parameter estimation of p and check the robustness of this approach. Noise is introduced here according to Eq. (6),

$$\hat{x}_t^{\epsilon} = \hat{x}_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{U}_{\{-\lfloor\sigma_{X_t}\rceil,\ldots,\lfloor\sigma_{X_t}\rceil\}}, \tag{6}$$

where $\hat{x}_t^{\epsilon}$ is the noisy datum, $\hat{x}_t$ is the original observed value and ϵ_t is the realization of a discrete uniform distribution, whose support begins with $-\lfloor\sigma_{X_t}\rceil$, increases a unit until it ends with $\lfloor\sigma_{X_t}\rceil$, in which σ_{X_t} is the standard deviation of the generation X_t.

The book [6] provides a solution to evaluate the statistic σ_{X_t} through

$$\sigma_{X_t}^2 = G''_{X_t}(s)\big|_{s=1} + G'_{X_t}(s)\big|_{s=1} - G_{X_t}(s)^2\Big|_{s=1}. \tag{7}$$

5 Results

The chosen parametric inference problem concerns the parameter p of the model $C \sim \text{Binomial}(2,p)$. From a Bayesian perspective, this parameter is treated as a random variable P. The updating process here uses data according to a BGW ruled by a $C \sim \mathit{Binomial}\,(2, 0.7)$. Therefore, it is expected that the posterior distribution will have greater values around $p = 0.7$ once the updates are complete. As previously mentioned, the number of updates N until convergence in this framework is stochastic. Hence, it is not possible to properly study the additive noise effect on the dataset with a unique observation of the updating process. This is why 1000 different updating processes

are executed for each data generation source, which are the third generation X_3, and the fourth generation X_4.

Each updating process finishes when the L_2-Wasserstein distance, according to Eq. (5), is less than $\xi_{\mathcal{W}_2} = 0.0005$. The original dataset $\{X_t\}_n$, where n is the number of updates, is gradually changed with the additive noise. Initially, the first 10% of data is affected by noise according to Eq. (6). Then, the following 10% of the remaining original data is also corrupted by noise. The contamination stops when 50% of the original dataset is changed. With the noisy data per level, the Bayesian update is again performed, given that some likelihood functions according to Eq. (2) might have changed. At each level of contamination, the aim is to capture the fluctuation in the mode of the last posterior distribution.

5.1 Dataset from the third generation

The original dataset from the third generation is increasingly contaminated by additive noise from the realization of a discrete uniform $\mathcal{U}_{\{-2,\ldots,2\}}$ random variable, given that $\sigma_{X_3} = 1.8945$. Figure 2 displays histograms of the modes of each last posterior distribution from different realizations of the Bayesian updating process. Each histogram represents a level of noise. Table 1 sums up some observations from the histograms in Fig. 2.

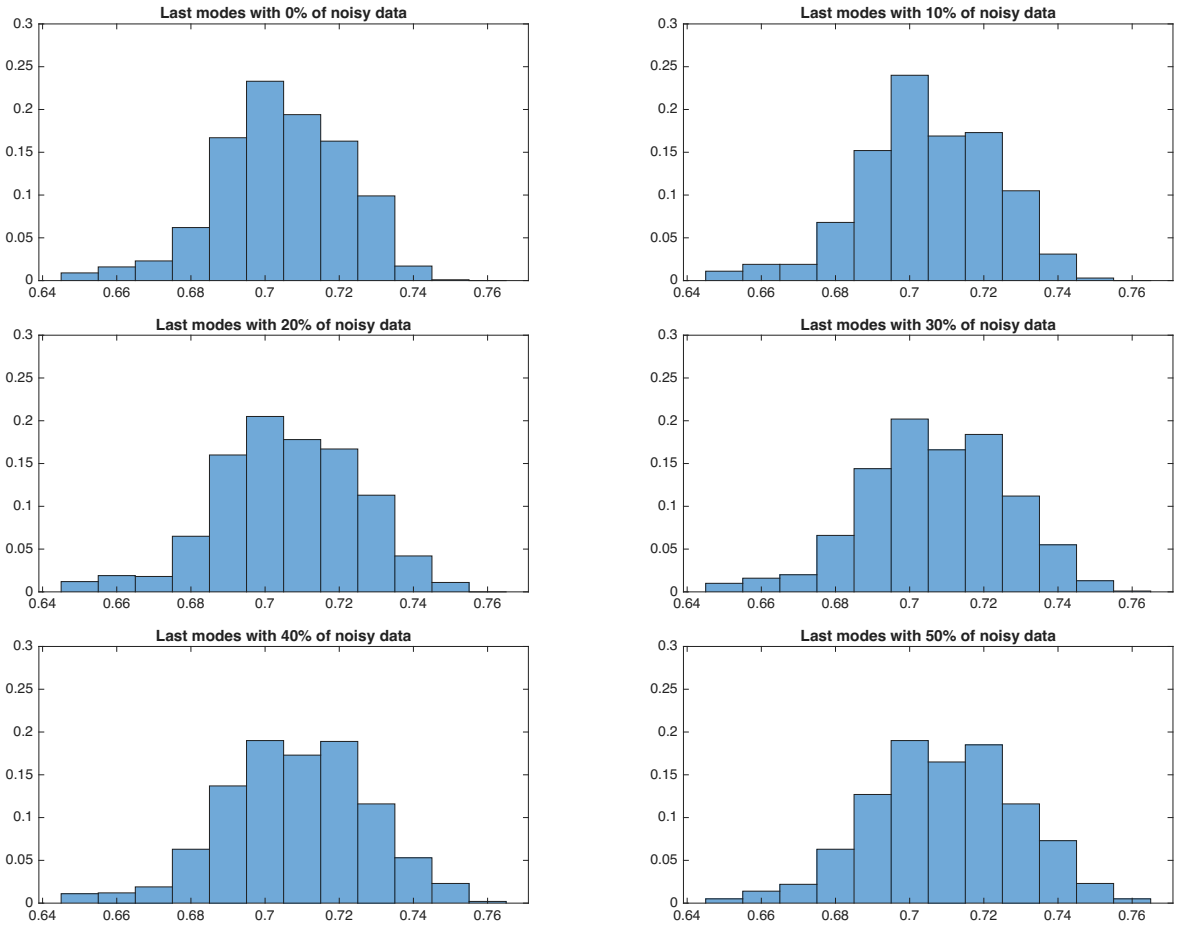


Figure 2: Additive noisy effect on modes of last posteriors for different levels of contamination when dataset's source is the third generation.

Table 1: Frequencies related to the dataset from third generation.

Noise level	**Frequency [0.645, 0.695]**	**Frequency [0.695, 0.705]**	**Frequency [0.705, 0.765]**
0%	0.277	0.233	0.474
10%	0.269	0.240	0.481
20%	0.274	0.205	0.511
30%	0.256	0.202	0.531
40%	0.242	0.190	0.556
50%	0.231	0.190	0.567

Figure 2 shows that the distribution shifts toward greater values as the noise level increases. For noise levels of 30% and above, some modes appear in the bin centered at $p = 0.76$, and its relative frequency increases slightly with the noise level. When 40% of the dataset is contaminated, the relative frequency of the bin centered at $p = 0.70$ is almost the same as that of the bin centered at $p = 0.72$, close to 0.190. The cumulative frequency of the bins below the one centered at $p = 0.7$ decreases by 16.61% when 50% of the data is noisy. On the other hand, at the same noise level, the cumulative frequency above the bin centered at $p = 0.7$ increases by 19.62%.

5.2 Dataset from the fourth generation

The last dataset is originally generated by realizations of the fourth generation. Additive noise is a realization of $\mathcal{U}_{\{-3,\ldots,3\}}$, because $\sigma_{X_4} = 2.8613$. The corresponding histograms for this scenario are displayed in Fig. 3 and cumulative frequencies are shown in Tab. 2.

Table 2: Frequencies related to the dataset from fourth generation.

Noise level	**Frequency [0.645, 0.695]**	**Frequency [0.695, 0.705]**	**Frequency [0.705, 0.765]**
0%	0.336	0.238	0.424
10%	0.308	0.246	0.442
20%	0.291	0.236	0.469
30%	0.283	0.220	0.494
40%	0.263	0.208	0.526
50%	0.256	0.211	0.532

The same shifting behavior is observed in this case. The cumulative frequency of the bins below the one centered at $p = 0.7$ decreases by 23.81% when 50% of the data is noisy. Conversely, at the same noise level, the cumulative frequency above the bin centered at $p = 0.7$ increases by 19.62%. However, the shifting frequencies are more resistant to moving to farther bins compared to the results from previous generations. This observation is supported by the fact that, in the absence of noise, 1.8% of the modes from the third generation are greater than 0.735, compared to 3.1% from the fourth generation. Yet, for noise levels of 30% and above, this percentage is higher for the third generation. For instance, at the highest noise level, 10.1% of its modes exceed 0.735, versus 6.3% for the fourth generation. Furthermore, unlike in the third generation, the bin centered at $p = 0.7$ remains the most frequent across all noise levels.

6 Conclusion

The aim of this paper is to evaluate the robustness of the proposed Bayesian approach. To this end, additive noise is introduced into the original datasets at varying levels, and its effect is analyzed. For contamination levels below 30%, no significant changes are observed in the modes of the

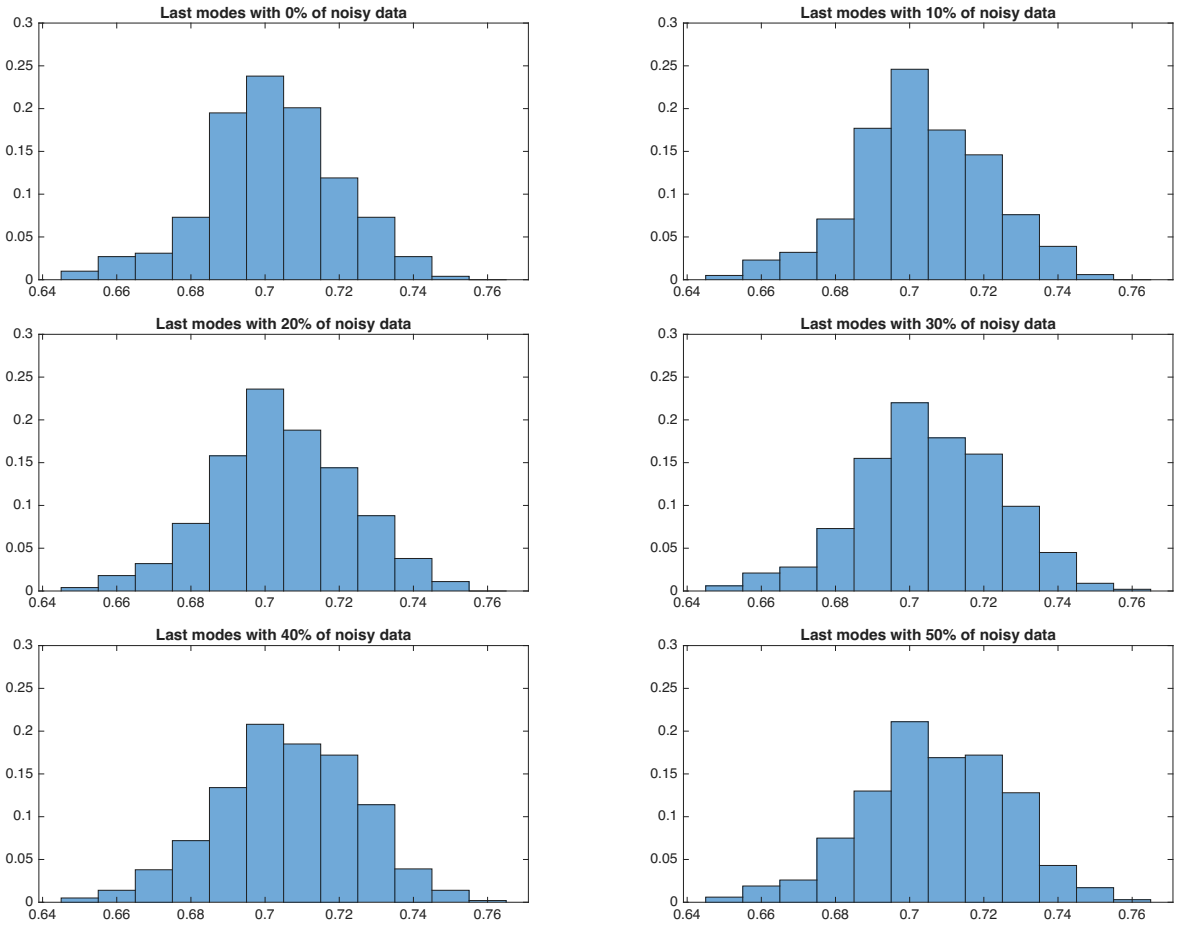


Figure 3: Additive noisy effect on modes of last posteriors for different levels of contamination when dataset's source is the fourth generation.

final posterior distributions of P after the updating process. These results illustrate a robustness property, at least for the benchmark considered.

However, when 40% of the dataset from the third generation is contaminated, the relative frequency of the bin centered at $p = 0.7$ is almost the same as the one cenetered at $p = 0.72$. The approach loses then its accuracy. For the fourth generation, by contrast, only negligible changes are observed.

Although the additive noise follows a uniform distribution, its consequences are not symmetric on the proposed Bayesian approach. Specifically, there is a tendency to overestimate the mode of P. This observation is supported by the fact that, as the noise level increases, the cumulative frequency of bins centered below $p = 0.7$ decreases while the cumulative frequency of bins above it increases.

Acknowledgments

The authors thanks the support of Faperj and CNPq.

References

[1] N. H. L. Leung, "Transmissibility and transmission of respiratory viruses," *Nature Rev Microbiol*, vol. 19, pp. 528–545, 2021. DOI: 10.1038/s41579-021-00535-6

[2] J. P. Freitas, R. Lima, and R. Sampaio, "Bayesian parametric inference problem applied on branching processes," *Computational and Applied Mathematics*, vol. 45, p. 179, 2025. DOI: https://doi.org/10.1007/s40314-025-03572-6

[3] R. B. Schinazi, *Classical and Spatial Stochastic Processes*, 2nd. New York: Springer, 2014.

[4] S. Vallender, "Calculation of the wasserstein distance between probability distributions on the line," *Theory of Probability & Its Applications*, vol. 18, no. 4, pp. 784–786, 1974. DOI: 10.1137/1118101

[5] M. Deza and E. Deza, *Encyclopedia of Distances*, 4th. Berlin: Springer, 2016.

[6] G. Grimmett and D. Welsh, *Probability: An Introduction*, 2nd. New York: Oxford University Press, 2014.

Part IX

Mathematical models in population dynamics

On parameter estimation of differential equations with incomplete data

Benito Chen-Charpentier[♭],[1]

(♭) Department of Mathematics, University of Texas at Arlington
Arlington, Texas 76019, USA.

1 Introduction

Many mathematical models are written in terms of ordinary, delay and partial differential equations. In this paper we will deal with models given by initial value ordinary differential equations (ODEs) but the results presented can be extended to other types of models. Models based on ordinary differential equations are widely used in Mathematical Biology, especially in population dynamics and ecology. These ordinary differential equations are usually nonlinear and their solution requires approximations. There are many different numerical methods for solving ODEs, for example, Runge-Kutta and predictor corrector methods [1], [2]. An alternative way of solving ODEs is to use neural networks [3], [4]. Briefly a neural network consists of cells or nodes arranged in layers. Each cell receives an input from each cell in the previous layer, takes a weighted average of the inputs and applies an activation function to produce an output. There are several ways in which a neural network can be used to solve ODEs. The first is Physics informed neural networks (PINNs) where the solutions of the ODEs are written as neural networks and the loss function used to evaluate the weights is a mean square error of the derivative of the network with the right hand side of the equation evaluated at the network. That is, the loss is a measure of how well the neural network satisfies the ODEs at given points in time [5], [6]. A second method called neural ODEs uses as the first layer of the neural network the initial conditions, and as the second the ODEs [7]. Mathematical models involve parameters which many times are not known. Parameter estimation is the process of using experimental or observed data to determine the specific values of the unknown constants in a mathematical model. The goal is to find the set of parameters p^* that minimizes the difference between the model's predictions and the observed data. First it is necessary to give an objective or loss function that measures how well the predicted values agree with the observed data. This is usually the sum of square residuals or a likelihood function. Then an optimization algorithm is used to find the minimum (or maximum) value of the objective function. For ODEs the equations are solved with initial values given to the unknown parameters and then a numerical algorithm like steepest descent or Levenberg-Marquardt is used to find the parameter values that optimize the objective function. There are routines that will do the parameter estimation of ODEs by solving numerically the differential equation and minimizing the loss function. However, the problem of estimating the parameters of ODEs using observed data is not a simple one. Models involve simplifications, data is usually noise, the objective function may have local minima or be non-convex. There is also the issue of parameter identifiability, that is, whether the parameters

[1] bmchen@uta.edu

can be uniquely identified from the observed data [9], [10]. There are alternative methods for parameter estimation: Bayesian methods [11], [12], spline-based [13], [14], and PINNs [15], [16]. An important issue with parameter estimation is incomplete data. Some missing data in a time series of observations can cause error but the most serious situation is when data is only available for some populations. For example in a vector transmitted disease there may be no data on the vectors, or in a within-host infection they usually is only data for the pathogen. In this paper we concentrate on studying the behavior of two methods of parameter estimation for ODEs when there is data available for only one population. The two methods are the standard one of solving the ODEs numerically and adjusting the parameters to minimize the sum of square residuals, and the second a PINN based one. In Section 2 these two methods are explained, and three simple models, Newton's law of cooling, predator-prey interaction, and within-host infection, are presented. In Section 3 the two parameter estimation methods are applied to several scenarios based on the three models, both with complete and incomplete data. For incomplete data there may not be a unique solution. Finally in Section 4 some conclusions are given.

2 Methods

Parameter estimation is an inverse problem: the observed data is used to determine the model. More specifically, the data is used to find some, or all, the values of the parameters of the model. In the case of ODE models, the initial data can also be fitted from the data. Not all models are based on physical, biological, chemical, economical or social processes. Such models include regression and neural networks, and will not be considered here. First we describe the two methods of parameter estimation for ODEs that we will use.

2.1 Least squares

The first method is the most widely used and for simplicity we will call it the method of least squares.

Consider a system of ODEs defined by:

$$\frac{d\mathbf{u}}{dt} = f(\mathbf{u}, t, \mathbf{p}), \qquad \mathbf{u}(0) = \mathbf{u}_0,$$

where $\mathbf{u}$ is the state vector of dimension N and $\mathbf{p}$ is the vector of parameters of dimension M. t is time and will be taken between 0 and T. The observations will be at the P times t_i with $0 \leq t_1 \ldots < t_1 < \ldots \leq T$. The parameters are the unknown constants in the equations such as growth and decay rates, or transmission coefficients. The least squares parameter estimation method consists of the following steps: 1. start with an initial guess for the parameters $\mathbf{p}_0$; 2. numerically integrate the ODEs on the interval $[0, T]$ and generate a trajectory $\hat{\mathbf{u}}(t, \mathbf{p}_0)$; 3. compare the numerical trajectory to the observed data y_i by calculating the loss $L_{ls}(p)$ that we will take as the sum of the squares of the residuals

$$L_{ls}(\mathbf{p}) = \sum_{i=1}^{P} \|\hat{\mathbf{u}}(t_i, \mathbf{p}_0) - \mathbf{y}_i\|^2;$$

4. use an optimizer to find new values for $\mathbf{p}$ that reduce the loss; repeat steps 2 to 4 until an acceptable approximation of the minimum of $L_{ls}(\mathbf{p})$ is found.

The loss function usually has many local minima, so it is necessary to run simulations with different initial guesses of the parameter values. Also the fitted parameters should be studied to determine if the values make biological sense. The data may be weighted to give more relevance to the more reliable points. Also if the state variables have large ranges of variation it may be

necessary to scale them, for example, by taking the logarithm. Scaling all variables between 0 and 1 may also help the convergence. See [17] and the references therein.

2.2 Physics-informed neural networks

While the least squares method is standard, Physics-Informed Neural Networks (PINNs) are gaining popularity. In mathematical terms a neural network approximates a function $f(x)$ just as the function can be approximated using combinations of polynomials or basis functions [3], [4]. A neural network consists of nodes arranged in layers. Each node receives inputs $\mathbf{x}$ from each of the nodes in the previous layer, makes a linear combination them using a weight matrix W and a bias vector $\mathbf{b}$, $\mathbf{z} = W\mathbf{x} + \mathbf{b}$, and then applies an activation function σ, for example a sigmoid or an hyperbolic tangent, to introduce non-linearity to an output which will be the input of the next layer. A network with L layers is the composition of such functions

$$f(\mathbf{x}) = \sigma_L(W_L)\sigma_{L-1}(W_{L-1}\dots\sigma_1(W_1\mathbf{x} + \mathbf{b}_1)\dots + \mathbf{b}_L).$$

where $\mathbf{x}$ is the input to the network. This process is called forward propagation. Then a loss function is used to measure how close the output of the neural network is to the target values. An optimization routine is used to find the values of the matrices W and the vectors $\mathbf{b}$ that minimize the loss function. The derivatives of the loss function with respect to the weights and biases are calculated symbolically to insure high accuracy and keep the method mesh free.

PINNs incorporate physical, biological, or other information into the neural network. Instead of using a numerical integrator to produce an approximate solution to the ODEs, a neural network is trained to represent the solution $\hat{\mathbf{u}}(t, \mathbf{p})$. The loss function includes the residual of the ODE, forcing the network to satisfy the physics given by the differential equations. First we consider that the parameters are known and omit $\mathbf{p}$ from the argument of $\mathbf{u}$. The inputs to the network are points on the time interval of interest $[0, T]$. The loss function has a physics (or biology) part

$$L_{phys} = \sum_{i=1}^{P} \| \frac{d\hat{\mathbf{u}}(t_i)}{dt} - \mathbf{f}(\hat{\mathbf{u}}) \|^2.$$

It may also have a data part, if observations are given,

$$L_{obs} = \sum_{i=1}^{P} \| \hat{\mathbf{u}}(t_i) - \mathbf{y}_i \|^2,$$

and an initial condition part

$$L_{IC} = \| \hat{\mathbf{u}}(0) - \mathbf{u}_0 \|.$$

By minimizing a weighted sum of the three losses an approximation to the solution of the ODEs is found. This solution is valid for all values of t in the interval of interest. It can also be differentiated and manipulated analytically. The downside is that it is more expensive and difficult to evaluate than just solving the ODEs by integrating numerically.

To use PINNs for parameter estimation of ODEs, minor modifications need to be made. First it is necessary to have observable data $\mathbf{y}_i$. Second the loss function needs to include the L_{obs} part. The unknown parameters in the ODEs are then treated as trainable variables by the optimizer, and are updated together with the neural network's weights and biases. The PINN finds the solution of the ODEs and the unknown parameters in one optimization process. If using the standard least squares procedure produces parameters that are realistic and solutions that approximate the observed values, then there is no need to consider PINNs that are more expensive to evaluate. But there are results that PINNs deal better when the data is noisy [18], [19]. There is also the

question of which method works better when there is incomplete or hidden data, as when there is data available for only one of the state variables. We will explore this questions by using three simple models: Model 1 for Newton's law of cooling, Model 2 based on the Lotka-Volterra equations, and Model 3 which is a simple within-host infection model.

2.3 Mathematical models

Model 1 consist of a single ODE with one unknown parameter $k > 0$

$$\frac{du}{dt} = -k(u - u_a), \tag{1}$$

with u the temperature and u_a the known ambient temperature. For the parameter estimation it is necessary to give observed values of the temperature, so there is no incomplete data in our sense, even though there may be missing values in the observed time series. We will use this model to compare the results for noisy data.

Model 2 is a predator-prey model based on the Lotka-Volterra equations [20]

$$\begin{aligned} \frac{dx}{dt} &= ax - bxy \\ \frac{dy}{dt} &= -cy + exy, \end{aligned} \tag{2}$$

where x is the prey, y the predator, and a, b, c, e are positive parameters.

We will use both methods on this model using first data for both prey and predator and then only data for the prey. The data will be noisy. The state variables are S, healthy or target cells, I infected cells, and v a virus or other parasite. The immune system is only considered implicitly in the death term of the virus. The equations are

$$\begin{aligned} \frac{dS}{dt} &= -\beta Sv \\ \frac{dI}{dt} &= \beta Sv - \delta I \\ \frac{dv}{dt} &= N_r \delta I - cv. \end{aligned} \tag{3}$$

Here $\beta > 0$ is the infection rate, $\delta > 0$ the killing rate of immune cells by the virus, $N_r > 0$ the number of copies the virus makes of itself after replicating inside the infected cell, and $c > 0$ the dead rate of the virus that includes the killing rate by the immune system. It is assumed that the epidemic is short lived, such as influenza, so that there is no need to add demographics.

For this model the only data available is for the number of virus particles. We fitted the parameters using both methods using only noisy virus data.

3 Results

The numerical simulations were done using the programming language julia [21]. Julia is free, has a mathematics friendly syntax similar to that of Matlab or python, and it uses just in time compilation to make it comparable in performance to compiled languages like C, C++, or Fortran, and much faster than interpreted languages like Matlab, python or R. While its ecosystem of routines is not as wide as those of older languages, it is enough for most scientific coding problems. [22] explains the advantages of using julia in Mathematical Biology.

For our simulations we used the following julia packages: DifferentialEquations.jl for solving numerically the ODEs [23], Optim.jl for the optimizer part of both methods [24], lux.jl for the neural network [25], and other utility packages like plotting routines.

The data for all simulations was created by solving numerically the ODEs with initial given parameter values and then perturbing the solution by add a random noise of different strengths. The noise used has a normal distribution with mean 0 and standard deviation 1. The numerical method use for solving the ODEs is a Rosenbrock type one suitable for stiff systems, and with adaptive step size [26]. After the parameters were estimated, a numerical solution was calculated using the converged parameters and plotted together with the data.

Model 1 consists of one equation and one unknown parameter k. So there is no incomplete data. Three simulations were done all with the "true" solution calculated with $k = .5$ on the interval $[0, 100]$. The first one was with no noise, and starting guess of $k = .1$ for both methods. The neural network consisted of two layers, the first one with one node and 16 outputs with the activation function a sigmoid, and the second layer with 16 nodes and one output. Both methods converged to a value of k that agrees with the exact k to more than three decimal places. For the second simulation, a noise with amplitude 1.5 was added. Least squares gives $k = .520$ and PINN $k = .526$. Both simulated curves approximate the data very well but towards the end of the time interval, the PINN result is slightly better. For the third simulation, the noise was very large with amplitude 10. Now the optimal parameters are very different, least squares $k = .681$, PINN $k = .241$. The PINN curve is graphically closer to the data points. This result agrees with previous results about PINNs being better for noisy data.

Model 2 has two equations and two populations, prey and predator. The neural network had three layers. The first one with 1 input and 32 outputs with tanh activation, the second with 32 nodes. 32 outputs and tanh activation, and the third layer with 32 nodes and 2 outputs. The first simulation used $a = 1.1, b = 1.0, c = 0.4, e = 0.1$ as the parameters for the "true" solution. The data points for both prey and predator were used in both methods. The strength of the noise in the data was 0.1. Both methods gave very good results. A second simulation was done increasing the noise in the data to 0.6. Both methods produce very good results but least squares misses a small peak on the data. The parameters for the true solution were then changed to $[a, b, c, e] = [1.5, 1.0, 3.0, 1.0]$ so the "true" solution has a third peak. Least squares now misses the second and third peaks, and PINN misses the third peak and has an extra small peak at the beginning.

When using only data for the prey in the parameter estimation, there is a lack of identifiability. For a given solution of the state variables and parameter values, it is possible to keep the prey population unaltered while scaling the prey one and adjusting the parameter values correspondingly. Without any values for the prey population, least squares and PINNs can give good fits for the prey and different values for the prey and parameters. But least squares may fail if the initial guess if the parameter values is not close, while PINN may still converge.

For Model 3, the incomplete or "hidden" data is more important since usually only data for the virus is available. The situation is more serious for infection models that include the immune cells explicitly and more processes. Comparison between the two methods is done with only the virus data. The identifiability issue affects now two of the three populations.

4 Conclusions

In many Mathematical Biology problems there is incomplete data, meaning that there is data only for some of the state variables and not all. A simple example is given by Model 3, since the number of virus particles can be estimated from blood tests, but not the number of healthy or infected cells. Also, there are not usually reliable values for all the parameters in the model. So parameter estimation needs to be done for predictions and exploring different scenarios. The standard way

of giving an estimate of the values, solving the ODEs numerically and the adjusting the values by minimizing the sum of the squares of the residuals, is the simplest method and the one that should be tried first. If the fit to the observed values is good, the parameters make biological sense, and the graphs of the hidden variables (without experimental values) are reasonable, then the problem can be considered solved. Since there is no data for the hidden variables, it is not possible to determine if the curves are correct, but they should have a shape that can be inferred from the observed data. If the converged solution is not as good as needed, a different method should be tried. PINNs are more expensive but known to work better with noisy data. It involves a different optimization problem with different convergence properties. PINNs can also force constraints such as positivity of solutions. On the downside is that there is not much known about the convergence of PINNs. There are some empirical rules on how to choose the network, from the number of nodes to the number of layers to the activation functions, but no theoretical results. Also, how do you weight the different parts of the loss function is an open question. A large weight of the physics part ensures that the network simulates well the ODEs while a large weight of the parameter part gives good fit of the parameters. Our simulations of different scenarios confirm that least squares is a good method that may fail for very noisy data while PINNs may still converge. But for missing data, both least squares and PINNs may give solutions for the hidden variables that are not close to the ones that were used to create them while still giving a good fit to the populations with data. Since in practice the hidden solutions cannot be known, any solution that gives good fitting of the known variables is good as long as the parameters obtained and the hidden variables make sense. Finally, real data may also include processes that are not included in the mathematical model. So there may not even exist a set of parameter that gives good fit, and neither method will produce good results.

References

[1] R. Burden, J. D. Faires, and A. Reynolds, "Numerical analysis, brooks/cole," *Boston, Mass, USA*, 2010.

[2] J. C. Butcher, *Numerical methods for ordinary differential equations*. John Wiley & Sons, 2016.

[3] L. Fausett, *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications* (Prentice-Hall International Editions). Prentice-Hall, 1994, ISBN: 978-0-13-334186-7.

[4] K. Gurney, *An introduction to neural networks*. CRC press, 2018.

[5] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational physics*, vol. 378, pp. 686–707, 2019.

[6] T. Knoke and T. Wick, "Solving differential equations via artificial neural networks: Findings and failures in a model problem," *Examples and Counterexamples*, vol. 1, p. 100 035, Nov. 2021, ISSN: 2666657X. DOI: 10.1016/j.exco.2021.100035 Accessed: Apr. 6, 2026.

[7] C. Rackauckas, M. Innes, Y. Ma, J. Bettencourt, L. White, and V. Dixit, *DiffEqFlux.jl - A Julia Library for Neural Differential Equations*, Feb. 2019. arXiv: 1902.02376 [cs, stat]. Accessed: Jul. 29, 2024.

[8] T. M. Inc., *Matlab version: 9.13.0 (r2022b)*, Natick, Massachusetts, United States, 2022. [Online]. Available: https://www.mathworks.com

[9] M. Van Noort et al., "Navigating the landscape of parameter identifiability methods: A workflow recommendation for model development," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 13, no. 7, pp. 1170–1179, Jul. 2024, ISSN: 2163-8306, 2163-8306. DOI: 10.1002/psp4.13148 Accessed: Apr. 7, 2026.

[10] M. J. Simpson and R. E. Baker, "Parameter Identifiability, Parameter Estimation, and Model Prediction for Differential Equation Models," *SIAM Review*, vol. 68, no. 1, pp. 153–171, Feb. 2026, ISSN: 0036-1445, 1095-7200. DOI: 10.1137/24M1667968 Accessed: Apr. 7, 2026.

[11] W. C. Roda, "Bayesian inference for dynamical systems," *Infectious Disease Modelling*, vol. 5, pp. 221–232, Jan. 2020, ISSN: 2468-2152. DOI: 10.1016/j.idm.2019.12.007 Accessed: Apr. 7, 2026.

[12] M. Barzegari and L. Geris, "An open source crash course on parameter estimation of computational models using a bayesian optimization approach," *Journal of Open Source Education*, vol. 4, no. 40, 2021.

[13] J. M. Varah, "A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations," *SIAM Journal on Scientific and Statistical Computing*, vol. 3, no. 1, pp. 28–46, Mar. 1982, ISSN: 0196-5204, 2168-3417. DOI: 10.1137/0903003 Accessed: Apr. 7, 2026.

[14] H. Liang and H. Wu, "Parameter Estimation for Differential Equation Models Using a Framework of Measurement Error in Regression Models," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1570–1583, Dec. 2008, ISSN: 0162-1459. DOI: 10.1198/016214508000000797 Accessed: Apr. 7, 2026.

[15] S. Cuomo, V. S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli, "Scientific machine learning through physics–informed neural networks: Where we are and what's next," *Journal of Scientific Computing*, vol. 92, no. 3, p. 88, 2022.

[16] S. A. Frank, "Optimizing differential equations to fit data and predict outcomes," *Ecology and Evolution*, vol. 13, no. 3, e9895, Mar. 2023, ISSN: 2045-7758. DOI: 10.1002/ece3.9895 Accessed: Apr. 6, 2026.

[17] A. F. Villaverde, F. Fröhlich, D. Weindl, J. Hasenauer, and J. R. Banga, "Benchmarking optimization methods for parameter estimation in large kinetic models," *Bioinformatics*, vol. 35, no. 5, pp. 830–838, Mar. 2019, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty736 Accessed: Apr. 7, 2026.

[18] L. Yang, X. Meng, and G. E. Karniadakis, "B-pinns: Bayesian physics-informed neural networks for forward and inverse pde problems with noisy data," *Journal of Computational Physics*, vol. 425, p. 109 913, 2021.

[19] P. Pilar and N. Wahlström, "Physics-informed neural networks with unknown measurement noise," in *6th Annual Learning for Dynamics & Control Conference*, PMLR, 2024, pp. 235–247.

[20] L. Allen, *An Introduction to Mathematical Biology*. Pearson-Prentice Hall, 2007.

[21] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, 2017. DOI: 10.1137/141000671 [Online]. Available: https://epubs.siam.org/doi/10.1137/141000671

[22] S. Pal, M. Bhattacharya, S. Dash, S.-S. Lee, and C. Chakraborty, "A next-generation dynamic programming language Julia: Its features and applications in biological science," *Journal of Advanced Research*, vol. 64, pp. 143–154, Oct. 2024, ISSN: 2090-1232. DOI: 10.1016/j.jare.2023.11.015 Accessed: Apr. 9, 2026.

[23] C. Rackauckas and Q. Nie, "Differentialequations.jl–a performant and feature-rich ecosystem for solving differential equations in julia," *Journal of Open Research Software*, vol. 5, no. 1, p. 15, 2017.

[24] P. K. Mogensen and A. N. Riseth, "Optim: A mathematical optimization package for Julia," *Journal of Open Source Software*, vol. 3, no. 24, p. 615, 2018. DOI: 10.21105/joss.00615

[25] A. Pal, *Lux: Explicit Parameterization of Deep Neural Networks in Julia*, version v0.5.0, If you use this software, please cite it as below., Apr. 2023. DOI: 10.5281/zenodo.7808904 [Online]. Available: https://doi.org/10.5281/zenodo.7808904

[26] G. Wanner and E. Hairer, *Solving ordinary differential equations II*. Springer Berlin Heidelberg New York, 1996, vol. 375.

Numerical integration of a nonlinear structured population model with two physiological structures.

Luis M. Abia [♭], Óscar Angulo [♭],[1] Juan Carlos López-Marcos [♭], and Miguel Ángel López-Marcos [♭]

(♭) Universidad de Valladolid e IMUVa,
Dpto. de Matemática Aplicada. Facultad de Ciencias. Univ. de Valladolid. Pso. Belén 7, 47011 Valladolid, SPAIN,
lmabia@uva.es, lopezmar@uva.es, malm@uva.es.

1 Introduction

The study of populations dynamics can be made by using the properties of the individual level [5]. By using this hypothesis, individuals are distinguished by means of different p hysiological condi-tions such as age, maturation, size, stage, status, location, etc. These features will be use to determine their birth, growth and death rates, and their each-other or environment interactions. The resulting models are known as structured population ones and the aim of their study is to understand how these conditions affect t he d ynamics of t he population.

With respect to the interest of the characteristic that structures the individuals in the population, the age is one of the most important. For instance, in epidemiology, we can structure individuals by different kind of age: chronological, infection, recovery, etc. The theoretical study of age-structured models is well developed and we refer to classical monographies [1], [3], [6].

However, in the literature, there are many models that combine the effects of two physiological features but there are very few theoretical studies of the main properties of such equations [2], [7], [4].

2 Model and Numerical Method

We consider the numerical integration of a model describing the evolution of a physiologically-structured population which is structured by two features. The following nonlinear model takes the form

$$u_t(x,y,t) + u_x(x,y,t) + u_y(x,y,t) = -\mu(x,y,t,I_\mu(t))\, u(x,y,t), 0 < x < x_M, 0 < y < y_M, t > 0, \tag{1}$$

$$u(0,y,t) = \int_0^{x_M}\int_0^{y_M} \alpha(y,\varepsilon,\sigma,I_\alpha(t))\, u(\varepsilon,\sigma,t)\, d\sigma\, d\varepsilon\ , \quad t > 0, 0 \le y \le y_m. \tag{2}$$

$$u(x,0,t) = \int_0^{x_M}\int_0^{y_M} \beta(x,\varepsilon,\sigma,I_\beta(t))\, u(\varepsilon,\sigma,t)\, d\sigma\, d\varepsilon\ , \quad t > 0, 0 \le x \le x_m. \tag{3}$$

$$u(x,y,0) = \phi(x,y), \quad 0 \le x \le\le x_m,\ 0 \le y \le y_m. \tag{4}$$

[1] oscar.angulo@uva.es

where variable t represents the time, and x and y are two traits correlated with time. The dependent variable u is the density of individuals of traits x and t at time t. The boundary conditions (2)-(3) account for the input at time t of individuals at the initial level of each characteristic. The mortality and recovering rates (μ, α and β) depend on the structuring variables, the time to take into account environmental changes and on a weighted total population to simulate the influence of the population on the individual development. These dependences are given through

$$I_s(t) = \int_0^{x_M} \int_0^{y_M} \gamma_s(\varepsilon, \sigma)\, u(\varepsilon, \sigma, t)\, d\sigma\, d\varepsilon\ , \quad t > 0, 0 \leq x \leq x_m, s = \alpha.\beta, \mu. \tag{5}$$

Finally, function ϕ denots the initial distribution of the population with respect to both characteristics.

Goal of this work is to continue the development on novel numerical methods that are pioneer in this field. We first introduced a first order error finite difference method, next we have just introduce a first numerical scheme that integrate along characteristic curves and, nowadays, the first numerical method for the computation of the solution of this nonlinear population model with two structures. We will demostrate the applicability of the numerical method to obtain the solution and a theoretical study about its convergence.

References

[1] Iannelli, M., Mathematical Theory of Age-Structured Population Dynamics, Pisa, Giardini editori e stampatori, 1995.

[2] Inaba, H., Endemic threshold analysis for the Kermack–McKendrick reinfection model, *Josai Math. Monogr.* 9: 105–133, 2016.

[3] Inaba, H., Age-Structured Population Dynamics in Demography and Epidemiology. New York Springer, 2017.

[4] Kang, H, Huo, X., Ruan, S., On first-order hyperbolic partial differential equations with two internal variables modeling population dynamics of two physiological structures, *Annali di Matematica Pura ed Applicata* 200: 403–452, 2021.

[5] Metz, J.A.J., Diekmann, O., The Dynamics of Physiologically Structured Populations. Lecture Notes in Biomathematics, vol. 86. Berlin, Springer, 1986.

[6] Webb, G.F., Theory of Nonlinear Age-Dependent Population Dynamics. New York, Marcel Dekker, 1984.

[7] Webb, G.F., Dynamics of populations structured by internal variables, *Math. Z.* 189: 319–335, 1985.

A general framework for the convergence analysis of Floquet multipliers of delay equations

Alessia Andò, Giusy Bosco, Dimitri Breda and Davide Liessi[1]

CDLab – Computational Dynamics Laboratory,
Department of Mathematics, Computer Science and Physics,
University of Udine, via delle Scienze 206, 33100 Udine, Italy.

Delayed terms frequently arise in mathematical modeling, in particular, for example, in population dynamics and epidemics in mathematical biology. An interesting problem in applications is determining the asymptotic stability of equilibria, periodic orbits and other kinds of invariant sets. These stability properties can be studied via the principle of linearized stability, which links them to the stability properties of linearizations of the system. These can in turn be derived from the evolution operators, and for periodic orbits, in particular, the ones which advance the state of the system by one period, i.e. the monodromy operators [1]. The elements of the spectrum of the latter are the Floquet multipliers, whose position with respect to the unit circle determines the stability of the periodic orbit.

With the aim of approximating the spectra of evolution operators of delay equations (both of differential and renewal type), we consider a discretization based on a reformulation of the operators by means of a fixed-point equation. We develop a general convergence analysis relying on a list of suitable hypotheses related to the properties of the equation and of the chosen approximation techniques. On the one hand, this framework unifies the proofs for some methods based on pseudospectral discretization (see [2], [3]), which are restated in this new form. On the other hand, it opens the way to rigorously justifying other methods, currently lacking a formal convergence analysis: we exemplify this with a Fourier-based reduction method.

Acknowledgments

AA, DB and DL are members of INdAM Research group GNCS and of UMI research group "Modellistica socio-epidemiologica". This work originates from GB's MSc thesis, supervised by DB and DL. The research collaboration was fostered by the workshop *Towards rigorous results in state-dependent delay equations* held at the Lorentz Center in Leiden on 4–8 March 2024. The work was partially supported by the Italian Ministry of University and Research (MUR) through the PRIN 2022 project (No. 20229P2HEA) "Stochastic numerical modelling for sustainable innovation", Unit of Udine (CUP G53C24000710006).

[1] davide.liessi@uniud.it

References

[1] O. Diekmann, S. A. van Gils, S. M. Verduyn Lunel, and H.-O. Walther, *Delay Equations: Functional-, Complex- and Nonlinear Analysis* (Appl. Math. Sci. 110). New York: Springer, 1995. DOI: 10.1007/978-1-4612-4206-2.

[2] D. Breda, S. Maset, and R. Vermiglio, "Approximation of eigenvalues of evolution operators for linear retarded functional differential equations," *SIAM J. Numer. Anal.*, vol. 50, no. 3, pp. 1456–1483, 2012. DOI: 10.1137/100815505.

[3] D. Breda and D. Liessi, "Approximation of eigenvalues of evolution operators for linear renewal equations," *SIAM J. Numer. Anal.*, vol. 56, no. 3, pp. 1456–1481, 2018. DOI: 10.1137/17M1140534.

On Runge-Kutta Methods for Stochastic Delay Differential Equations

Alessia Ando[♭], Dimitri Breda[♭], Davide Liessi[♭], Stefano Maset[♮] and Faraz William[♭1]

(♭) Università degli Studi di Udine,
Via delle Scienze, 206, 33100 Udine UD, Italy.

(♮) University of Trieste,
Piazzale Europa, 1, 34127 Trieste TS, Italy.

1 Abstract

Stochastic Delay Differential Equations (SDDEs) hold significant importance in the advanced modelling of population dynamics, epidemics outbreaks and other problems in natural sciences. While time integration is essential for simulating and understanding their behaviour and properties, it poses nontrivial challenges due to the presence of both randomness and memory effects. Additionally, a key challenge is the non-availability of higher-order derivative-free methods for SDDEs.

In this talk, we present possible extensions of Runge–Kutta methods for Delay Differential Equations (DDEs) and Stochastic Ordinary Differential Equations (SODEs) to SDDEs. First, we consider functional Euler Methods which were introduced for DDEs in [2] and further analysed in [3]. These methods allow any kind of dependence on the past and any choice of time steps, overcoming the limitations of traditional methods, which require commensurable delays and step sizes. The Euler method is limited to first-order convergence in both weak and strong sense.

To achieve a higher weak order of convergence for SDDEs we then present an extension of the method introduced in [5] for SODEs to SDDEs which achieves a weak order 2. The effectiveness of the proposed methods is demonstrated on a scalar logistic equation [4] and a predator-prey system [1].

References

[1] B. Du, Y. Wang, and X. Lian. A stochastic predator-prey model with delays. Adv. Differ. Equ., 2015:1–16, 2015.
[2] S. Maset, L. Torelli, and R. Vermiglio. Runge–Kutta methods for retarded functional differential equations. Math. Models Methods Appl. Sci., 15(8):1203–1251, 2005. MR 2156640.
[3] S. Maset. Numerical methods for Delay Differential Equations. In D. Breda, R. Vermiglio, and J. Wu (eds.), Delays and Structures in Dynamical Systems: Modelling, Analysis and Numerical Methods. CISM Lecture Notes, Springer, submitted.
[4] Z. Wu, H. Huang, and L. Wang. Stochastic delay logistic model under regime switching. In Abstract and Applied Analysis, volume 2012, page 241702. Wiley Online Library, 2012.

[1] william.faraz@spes.uniud.it

[5] A. Rößler (2009). Second order Runge–Kutta methods for Itô stochastic differential equations. SIAM Journal on Numerical Analysis, 47(3), 1713-1738.

Population-Based Statistical Modeling and Lipschitz Reinforcement Learning for Response Times in Video Games

Ana Coronado Ferrer[♭],[1] [2]

(♭) I.U. Matemática Pura y Aplicada, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

Abstract. We propose a unified mathematical framework for modeling response times in video game environments using probability distributions adapted to population dynamics as input for the application of specific reinforcement learning techniques. Response times are first modeled through parametric distributions, particularly Gamma-type laws, motivated by their asymmetry and suitability for reaction time data. Building upon these distributions, we introduce a predictive model based on Lipschitz extensions within a reinforcement learning scheme, allowing the estimation of individual responses under new stimuli. The approach integrates statistical modeling and artificial intelligence, providing a scalable tool for predicting player behavior in dynamic environments. The final aim is to develop a simulation tool to model user responses when interacting with new video game scenarios.

Keywords: Gamma distribution · Reinforcement Learning · Lipschitz extensions · Population dynamics · Response time modeling

1 Introduction

It is well known that video games constitute complex interactive environments in which users continuously respond to multiple stimuli under time constraints. Given the growing importance of this activity within leisure practices, recent years have witnessed a significant increase in scientific research focused on analyzing player behavior, with the aim of developing interactive models that support the design and evaluation of new products.

A key component in this line of research is the study of physiological response times, particularly those associated with visual stimuli characteristic of video games. The temporal dimension of these responses plays a fundamental role in assessing cognitive performance, decision-making processes, and adaptive strategies. From a statistical perspective, response times are non-negative, typically asymmetric, and exhibit variability across individuals. These features suggest the use of continuous probability distributions capable of capturing such behavior. Inspired by previous approaches that model collective phenomena through distributions derived from Gaussian components [1], we extend this framework to the context of video game interaction. Similar approaches have been developed recently with applications to health sciences and psychology [2], [3].

[1] ancofer1@alumni.upv.es

[2] The author thanks Enrique A. Sánchez Pérez for his collaboration and supervision of this research.

Response time distributions can be interpreted as population-level distributions, since they describe the variability of a given behavioral variable across a set of individuals. In this sense, the fitted Gamma-type laws provides a statistical approximation and also represent the aggregated effect of multiple underlying processes and heterogeneous player profiles. The Gamma distribution can be considered population-based because it naturally models the aggregated waiting times or lifespans of individuals within a population, arising as the sum of k independent exponential random variables, each representing a memoryless inter-event time. These distributions are then incorporated into a reinforcement learning framework to build a predictive metric model based on extensions of Lipschitz functions, enabling simulation of user responses to new situations.

Therefore, the aim of this work is to present the fundamental elements for the construction of a predictive model based on experimental data obtained from response time tests conducted with students enrolled in a video game degree program. Consequently, our objective is twofold. First, to model the empirical distribution of response times across a population of players. Second, to integrate this statistical description into a reinforcement learning (RL) framework based on Lipschitz extensions, enabling the prediction of individual responses under new stimuli.

Let us recall some basic definitions. Let $(\mathcal{S}, d)$ be a metric space. A function $f : \mathcal{S} \to \mathbb{R}$ is Lipschitz if

$$|f(s_1) - f(s_2)| \leq L\, d(s_1, s_2), \quad \forall s_1, s_2 \in \mathcal{S}. \tag{1}$$

Given a set of observed states $\{s_i\}$ with known responses $f(s_i)$, the Lipschitz extension allows defining f over the entire space $\mathcal{S}$. For the application of RL schemes we consider a standard RL setting defined by a state space $\mathcal{S}$ (game situations), an action space $\mathcal{A}$ (possible responses), a reward function $R(s, a)$, a policy π. The key novelty lies in incorporating a metric structure into the state space and using Lipschitz extensions to generalize observed behaviors.

2 Statistical Modeling of Response Times

Let X be a random variable representing the response time of a player to a given stimulus. After analyzing the data collected from our experience with 25 students, teachers, and staff associated with the Video Games degree at Florida Universitaria (associate center to the Universitat Politècnica de València), this information suggests that such variables are well described by asymmetric distributions. In particular, we consider the Gamma distribution

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0, \tag{2}$$

where $\alpha > 0$ is the shape parameter and $\beta > 0$ is the rate parameter.

In order to model the variability of response times across different individuals, we consider a family of Gamma distributions as the ones shown in Figure 1. This formulation allows us to represent different behavioral regimes within a population by varying the parameters. In particular, the curves displayed correspond to the cases $\Gamma(2, 1)$, $\Gamma(3, 2)$ and $\Gamma(4, 0.5)$, which simplify respectively to xe^{-x}, $4x^2e^{-2x}$ and $\frac{1}{96}x^3e^{-0.5x}$. These examples illustrate how changes in α and β modify both the dispersion and the skewness of the distribution, capturing heterogeneous response patterns across players.

Since response times are positive values, the accuracy of the model is justified by the ability to model skewness of the family of proposed functions, as well as its flexibility in capturing different regimes of variability.

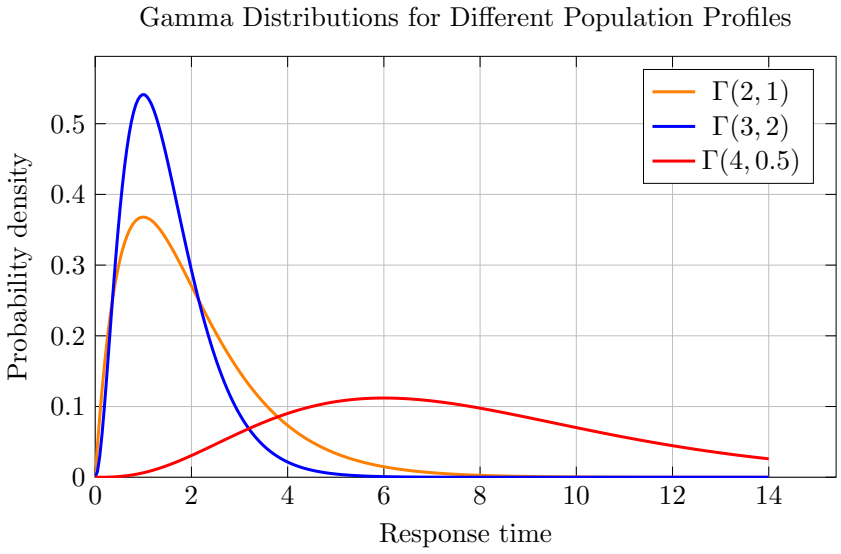


Figure 1: Examples of Gamma distributions representing different population response-time profiles.

2.1 Aggregation of Gaussian processes

Response times can be interpreted as the result of aggregating Gaussian processes, leading to distributions related to the chi family [1]. In this sense, both Gamma and chi-type distributions can be viewed as models arising from population dynamics, where each individual contributes independently to the global statistical behavior.

Indeed, let $\{X_i\}_{i=1}^{k}$ denote independent random variables associated with partial cognitive processes (perception, decision, motor response). Then, a global response variable can be defined as:

$$Z = \sum_{i=1}^{k} X_i, \tag{3}$$

or alternatively through a norm-based aggregation:

$$Z = \left(\sum_{i=1}^{k} X_i^2 \right)^{1/2}. \tag{4}$$

This approach allows the use of two modeling schemes, which will be explained in the next section. Before doing so, let us describe the basic elements of the Lipschitz regression scheme, which can be extended to our reinforcement learning approach.

2.2 Lipschitz regression procedures for the predictive response model

Within this framework, reinforcement learning techniques based on Lipschitz extensions provide a natural and effective tool for prediction in metric spaces [4], [5]. The first step to get the is to define a regression method in the adequate metric context. Given a set of known observations defined on a subset $A \subset XS$ of a metric space (S, d), Lipschitz extension theory allows the construction of functions over the whole space while preserving stability properties. In particular, the classical

McShane and Whitney extensions are given by

$$F_{\text{Whitney}}(x) = \inf_{a\in A}\{f(a) + L\, d(x,a)\}, \quad F_{\text{McShane}}(x) = \sup_{a\in A}\{f(a) - L\, d(x,a)\},$$

where $f : A \to \mathbb{R}$ is a real-valued function and L is its Lipschitz constant. Convex combinations of these operators enable the estimation of responses for previously unseen stimuli based on known data. In our setting, the space S is endowed with a metric that compares both the fitted response time distributions (Wasserstein metric) and the associated performance scores of individual players, thus defining a meaningful notion of similarity between subjects and situations. This approach makes it possible to design predictive models that generalize from known behavioral patterns to new scenarios, providing a robust framework for modeling response distributions under unknown stimuli within a reinforcement learning scheme.

In our case, each exercise ("Level") in the set for which the battery of questions is answered by the participants in the experiment is described by a distribution. Suppose we have $n \in \mathbb{N}$ questions q_k, $k \in \{1, \ldots, n\}$, in the test, and define a distribution function for each question via $q_k \mapsto \Phi_k$. Each such distribution is assumed to be of the form defined in the previous section, for example, a Gamma distribution or any adjusted distribution obtained through a Bayesian scheme.

3 The model: mathematical tools and AI techniques

For the application of the distributions obtained to an RL algorithm, we need to fix a precise procedure.

3.1 Clustering of individuals, Lipschitz regression and RL

Each individual is represented as a vector of response times, which defines a metric space of players. A scoring function assigns a performance value to each individual, allowing the use of Lipschitz extensions to estimate the expected score for new individuals based on proximity in this space. On the other hand, each item of the test provides an independent distribution of the physiological response of the participants. This complementary approach involves two main components.

(i) The first consists of identifying each individual with the vector of reaction times to each question in the test. A clustering procedure is then applied, yielding player profiles for each video game based on the specific cognitive abilities involved in that game.

(ii) The second component consists of defining a metric space on the set of distributions obtained via the Wasserstein distance. A real-valued score is then assigned to each question in the test, measuring the cognitive ability associated with it. For instance, if a question records the reaction time to identify the first circle in a sequence of squares, the corresponding score could reflect the coincidence between fast responses to that question and fast responses to an analogous question in which a diamond breaks the sequence. Such scores may be derived either from integral formulas or from experimental assessments based on other questions in the test.

The approach considered in (i) above allows for the use of clustering techniques for grouping individuals into different behavior classes, which would provide strategic information for the design of new games. However, the interest of (ii) goes in a different direction. Using Lipschitz (metric) regression in the first step, the Lipschitz extension of the scores of known situations would help to simulate the score associated with a different distribution of responses of other phenomena. The Wasserstein distance, together with other metrics computed with different parameters that characterize a given feature of the video game, provide the underlying metric space, on which the

extension of the function defining the new distribution can be obtained. Once such a distribution is known, the system provides an approximation to the score.

As in [4], this procedure can be used to define a sequence of metric spaces that is progressively enriched with successive experiences on the system, which, together with a metric space of actions on the space of states, provides a reinforcement learning procedure to learn the behavior of the player in a given video game.

3.2 The experiment

As part of the research program being developed on the topic explained in this extended abstract, real experiments are being conducted with students of a video game degree. A battery of 20 experiments of aparition of stimule in the screen and the response times of 25 indivuals were performed as primary data. Many other parameterers (size of the stiluli, color, sprawn time...) were also recorded. The specific distribution functions resulting from this population (which resemble those in Figure 1) are presented below for the first three experiments (Level 1, Level 2 and Level 3) in Figures 2 and 3.

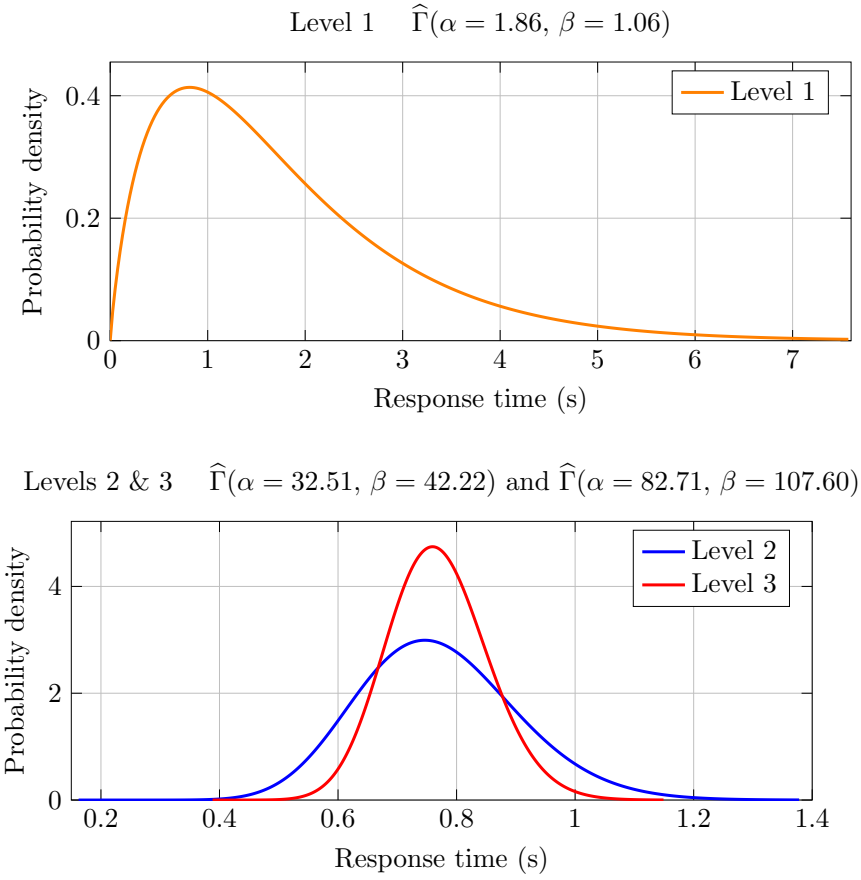


Figure 2: Gamma distributions fitted to empirical response-time data ($n = 24$ per level). Top: Level 1, characterized by high variability and right skew. Bottom: Levels 2 and 3, showing markedly lower dispersion and concentration near the mean. Fitted parameters: Level 1: $\hat{\alpha} = 1.86$, $\hat{\beta} = 1.06$; Level 2: $\hat{\alpha} = 32.51$, $\hat{\beta} = 42.22$; Level 3: $\hat{\alpha} = 82.71$, $\hat{\beta} = 107.60$.

The response time distributions show a clear shift from the high variability and wide dispersion

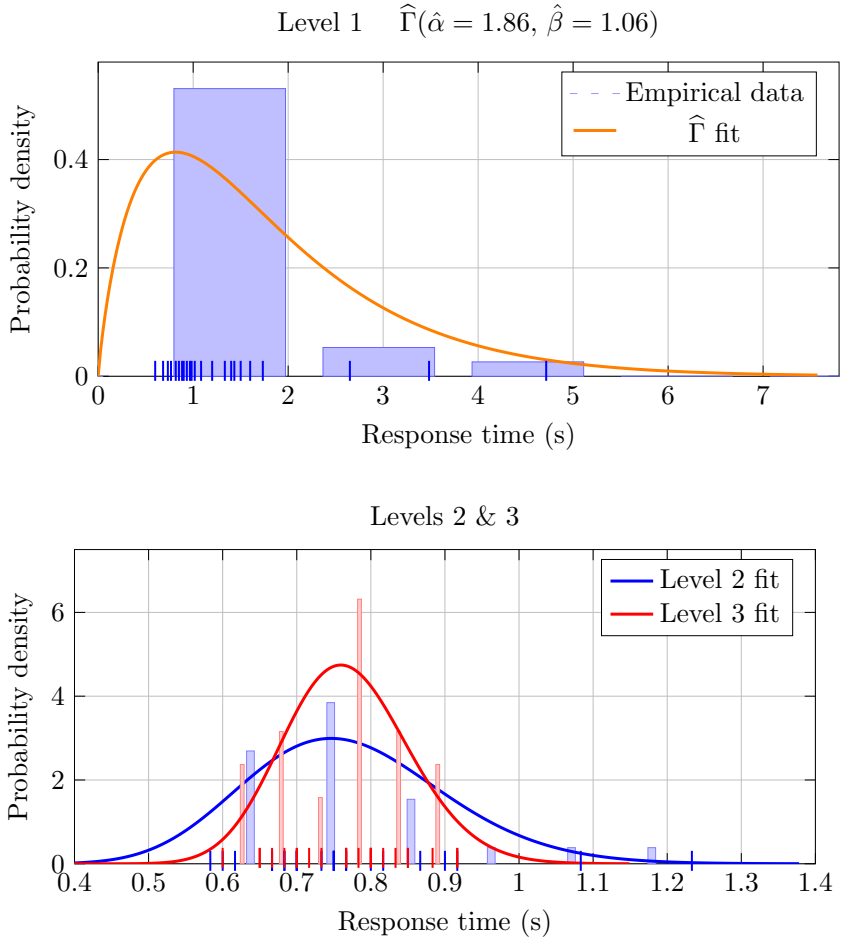


Figure 3: Gamma distributions fitted to empirical response-time data ($n = 24$ per level). Shaded bars show the empirical density histogram; tick marks on the baseline indicate individual observations; solid curves show the fitted $\widehat{\Gamma}$ distribution. Top: Level 1 ($\hat{\alpha} = 1.86$, $\hat{\beta} = 1.06$), characterized by high variability and right skew. Bottom: Level 2 ($\hat{\alpha} = 32.51$, $\hat{\beta} = 42.22$, blue) and Level 3 ($\hat{\alpha} = 82.71$, $\hat{\beta} = 107.60$, red), showing markedly lower dispersion and progressive concentration near the mean.

seen in Level 1 to a highly consistent, quasi-normal performance in Levels 2 and 3. As players gain proficiency, the narrowing of the probability density functions (indicated by significantly higher shape parameters α) reflects a transition from hesitant, exploratory decision-making to automated, reliable execution. This improvement is likely due to the "novelty effect" of Level 1, where initial uncertainty and lack of familiarity with the gaming stimuli led to the observed lack of coordination in user responses.

The distributions obtained in this way serve as the basis for building the model described above. In particular, these distributions form the core of a metric space defined using the Wasserstein metric. Combined with evaluation functions for these distributions, this setup provides a Lipschitz regression structure, which is the first step toward our reinforcement learning algorithm.

4 Conclusions

We have introduced a unified framework combining population-based distributions and Lipschitz-based reinforcement learning for modeling response times in video games. The use of Gamma-type distributions provides a robust statistical foundation. Then, we have explained that Lipschitz extensions enable effective generalization of metric regression, which opens the door to the construction of a reinforcement learning scheme founded on this statistical framework. This integration would allow the prediction of individual behavior in complex environments, opening new perspectives in both artificial intelligence and cognitive modeling.

The work focuses on empirical validation using real gameplay data and the extension of the model to multi-agent environments. Some partial results from these experiments with a group of individuals are presented in this communication.

References

[1] D. Puig Castro, A. Coronado Ferrer, J. C. Castro Palacio, P. Fernández de Córdoba, N. Ortigosa, and E. A. Sánchez Pérez, "Non-centered chi distributions as models for fair assessment in sports performance," *Symmetry*, vol. 17, no. 7, p. 1039, 2025. DOI: 10.3390/sym17071039.

[2] N. Ortigosa, M. Orellana-Panchame, J. C. Castro-Palacio, P. Fernández de Córdoba, and J. M. Isidro, "Monte Carlo simulation of a modified chi distribution considering asymmetry in the generating functions: Application to the study of health-related variables," *Symmetry*, vol. 13, no. 6, p. 924, 2021. DOI: 10.3390/sym13060924.

[3] J. C. Castro-Palacio, P. Fernández-de-Córdoba, J. M. Isidro, S. Sahu, and E. Navarro-Pardo, "Human reaction times: Linking individual and collective behaviour through physics modeling," *Symmetry*, vol. 13, no. 3, p. 451, 2021. DOI: 10.3390/sym13030451. [Online]. Available: https://doi.org/10.3390/sym13030451.

[4] J. M. Calabuig, H. Falciani, and E. A. Sánchez-Pérez, "Dreaming machine learning: Lipschitz extensions for reinforcement learning on financial markets," *Neurocomputing*, vol. 398, pp. 172–184, 2020. DOI: 10.1016/j.neucom.2020.02.056.

[5] H. Falciani and E. A. Sánchez-Pérez, "Semi-Lipschitz functions and machine learning for discrete dynamical systems on graphs," *Machine Learning*, vol. 111, no. 5, pp. 1765–1797, 2022. DOI: 10.1007/s10994-022-06130-x.

A Hybrid Computational Framework for a Fisher-KPP Problem with Nonlocal Free Boundary

Vera Egorova[♮1], M.-C. Casabán[♭], and Rafael Company[♭]

(♮) Depto. de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, Av. de Los Castros, s/n, Santander, Spain.

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València Camí de Vera s/n, València, Spain.

1 Introduction

The application of reaction-diffusion systems has become a cornerstone for simulating the complex spatial patterns found in biological and ecological expansion [1]. While classical Stefan-type problems primarily address the dichotomy between species expansion and extinction, the transition to nonlocal free-boundary formulations allows for the inclusion of cognitive factors such as environmental memory and long-distance sensory perception [2]. Recent theoretical developments have further extended this framework to time-periodic environments and epidemiological contexts, demonstrating that a complex spreading–balancing–vanishing trichotomy governs species survival [3], [4].

In this work, we focus on a Fisher-KPP problem with a nonlocal free boundary that accounts for social interaction mechanisms and nonlocal perception, capabilities that enable individuals to react to distant signals perceived through sight or smell [2]. To solve this system, we combine the front-fixing (FF) technique with an alternating direction explicit (ADE) scheme [5]. Furthermore, we construct a Physics-Informed Machine Learning (PIML) classifier to predict the population's asymptotic fate without the need for continuous numerical simulations [6].

2 Mathematical Model

We consider the following Fisher-KPP problem with a nonlocal free boundary $h(t)$ proposed in [2]:

$$u_t(t,x) = D\,u_{xx}(t,x) + f(u(t,x)), \quad t>0,\ 0<x<h(t)\,, \tag{1}$$

$$h'(t) = \mu \int_0^{h(t)} u(t,x)\,w(h(t)-x)\,\mathrm{d}x\,, \quad t>0\,, \tag{2}$$

$$u_x(t,0) = 0, \quad u(t,h(t)) = 0, \quad t>0\,, \tag{3}$$

$$h(0) = h_0, \quad u(0,x) = u_0(x), \quad 0\le x\le h_0\,, \quad h_0>0. \tag{4}$$

The variable $u(t,x)$ expresses the density of a species at time t and spatial point x and the parameter $D>0$ its diffusion rate. The function $f(u(t,x))$ specifies the local growth rate of the

[1] egorovav@unican.es

population and it is taken to be of logistic type, namely

$$f(u(t,x)) = r\,u(t,x)\,(a - u(t,x))\,, \tag{5}$$

where r is the intrinsic growth rate of population and a is the carrying capacity. The nonlocal equation (2) governs the movement over time of the unknown moving boundary $h(t)$ of the habitat occupied by a single species with $\mu > 0$ being the proportionality constant. The kernel function defined by

$$w(h(t)-x) = c_1 e^{-\alpha_1(h(t)-x)} - c_2 e^{-\alpha_2(h(t)-x)}\,, \quad c_1 > c_2 > 0\,, \quad \alpha_1 > \alpha_2 > 0\,, \tag{6}$$

quantifies the influence of population clusters on the boundary velocity, reflecting how a species prioritizes different regions of its habitat during movement. Then, the population grows and competes locally at each point of the domain, while the free boundary evolves according to a nonlocal perception of the total population, weighted by a kernel function. In this model, the Neumann boundary condition is imposed only at the left endpoint, while at the right endpoint we apply the Dirichlet condition, see (3). The initial population density $u_0(x)$ satisfies

$$u_0(x) \in \mathcal{C}^2\left([0,h_0]\right), \quad u_0'(0) = u_0(h_0) = 0\,, \quad u_0(x) > 0\,, \forall x \in [0,h_0), \tag{7}$$

where h_0 is the initial population range. For the numerical examples the following definition of the initial distribution is used

$$u_0(x) = 0.01(h_0^2 - x^2). \tag{8}$$

3 Methods

3.1 Front-Fixing Transformation and ADE Scheme

Our computational approach addresses the moving interface by coupling coordinate transformation with a highly efficient time-stepping algorithm.

To overcome the difficulties associated with a time-dependent spatial grid, we apply the Landau transformation $z = x/h(t)$, which maps the evolving habitat $[0,h(t)]$ onto the fixed computational domain $z \in [0,1]$. Defining the transformed density as $v(t,z) = u(t,x)$, the original Fisher-KPP equation is reformulated as a diffusion-advection-reaction problem:

$$\begin{aligned}
v_t(t,z) &= D\frac{1}{h^2(t)}\,v_{zz}(t,z) + z\frac{h'(t)}{h(t)}\,v_z(t,z) + f(v(t,z)), \quad t>0, \quad 0<z<1, && (9)\\
h'(t) &= \mu\,h(t)\int_0^1 v(t,z)\,w((1-z)h(t))\,\mathrm{d}z, \quad t>0, && (10)\\
v_z(t,0) &= 0, \quad v(t,1) = 0, \quad t>0\,, && (11)\\
h(0) &= h_0, \quad v(0,z) = u_0(z\,h_0),\ h_0>0, \quad 0\le z\le 1\,, && (12)
\end{aligned}$$

with the transformed kernel

$$w((1-z)h(t)) = c_1\,e^{-\alpha_1\,(1-z)\,h(t)} - c_2\,e^{-\alpha_2\,(1-z)\,h(t)}, \quad c_1>c_2>0, \quad \alpha_1>\alpha_2>0\,. \tag{13}$$

To solve the transformed system (9)–(13), we employ an ADE method of Barakat and Clark [5]. Its primary advantage lies in its unconditional stability for linear diffusion problems while remaining strictly explicit in its execution.

Let us consider a time horizon T, with $t \in [0,T]$. To construct a uniform grid over the fixed domain $[0,1]\times[0,T]$, we divide the spatial interval into M subintervals and the temporal interval

into N time steps. The spatial and temporal step sizes are given by $\Delta z = 1/M$ and $\Delta t = T/N$, respectively. The spatial grid points are defined as $z_j = j\Delta z$ for $j = 0, 1, \ldots, M$, and the temporal grid points as $t^n = n\Delta t$ for $n = 0, 1, \ldots, N$. We denote in the following way the numerical approximations of $v(t, z)$ and $h(t)$, as well as the kernel $w((1-z)h(t))$ given by (13)

$$\begin{aligned} v_j^n &\approx v(z_j, t^n), \\ h^n &\approx h(t^n), \\ w_j^n &= c_1\, e^{-\alpha_1\,(1-z_j)\,h^n} - c_2\, e^{-\alpha_2\,(1-z_j)\,h^n} \approx w((1-z_j)h(t^n)). \end{aligned}$$

We denote the numerical approximation of integral in (10) obtained by using the Simpson's rule, as follows

$$S(p^n) = \frac{\Delta z}{3}\left(p_0^n + 4\sum_{\text{odd}\,i} p_i^n + 2\sum_{\text{even}\,i} p_i^n + p_M^n\right), \tag{14}$$

where p^n is the vector $p^n = [p_0^n, p_1^n, \ldots, p_M^n]$ of components defined by $p_i^n = v_i^n\, w_i^n$. Then, the condition (10) is discretized as follows

$$\frac{h^{n+1} - h^n}{\Delta t} = \mu\, h^n\, S(p^n)\,, \quad 0 \le n \le N-1\,. \tag{15}$$

The core of the ADE method is the splitting of the temporal step into two sequential sweeps. Hence, at each time level n, set $Q_j^n = v_j^n$ and $P_j^n = v_j^n$ for $j = 1, \ldots, M-1$.

1. **Upward Sweep (Q_j^{n+1}):** Calculations proceed from z_0 to z_{M-1}, using the updated value at $j-1$ and the previous value at $j+1$ to compute the state at node j:

$$\begin{cases} Q_1^{n+1} = & [b^n Q_1^n + c_1^n Q_2^n + \Delta t f(Q_1^n)]\left(1 + \frac{\Delta t}{2\Delta z} z_1 \mu S(p^n)\right)^{-1}, \\ Q_j^{n+1} = & \left[a_j^n\, Q_{j-1}^{n+1} + b^n\, Q_j^n + c_j^n\, Q_{j+1}^n + f\left(Q_j^n\right)\Delta t\right]\left(1 + \lambda\dfrac{D}{(h^n)^2}\right)^{-1}, \\ & j = 2, \ldots, M-1, \\ Q_0^{n+1} = & Q_1^{n+1}, \quad Q_M^{n+1} = 0, \end{cases} \tag{16}$$

   with the coefficients:

$$a_j^n = \lambda\frac{D}{(h^n)^2} - \frac{\Delta t}{2\Delta z} z_j\, \mu\, S(p^n)\,, \quad b^n = 1 - \lambda\frac{D}{(h^n)^2}\,, \quad c_j^n = \lambda\frac{D}{(h^n)^2} + \frac{\Delta t}{2\Delta z} z_j\, \mu\, S(p^n)\,, \tag{17}$$

   where $\lambda = \Delta t/(\Delta z)^2$.

2. **Downward Sweep (P_j^{n+1}):** Calculations proceed from z_{M-1} back to z_1, using the updated value at $j+1$ and the previous value at $j-1$:

$$\begin{cases} P_M^{n+1} = & 0, \\ P_j^{n+1} = & \left[c_j^n\, P_{j+1}^{n+1} + b^n\, P_j^n + a_j^n\, P_{j-1}^n + f\left(P_j^n\right)\Delta t\right]\left(1 + \lambda\dfrac{D}{(h^n)^2}\right)^{-1}, \\ & j = M-1, \ldots, 1, \\ P_0^{n+1} = & P_1^{n+1}. \end{cases} \tag{18}$$

The final solution at t^{n+1} is obtained by averaging both sweeps, $v_j^{n+1} = \frac{1}{2}(Q_j^{n+1} + P_j^{n+1})$.

3.2 Physics-Informed Machine Learning

Given the high dimensionality of the parameter space $\mathcal{P} = \{D, r, a, \mu, c_1, c_2, \alpha_1, \alpha_2, h_0\}$ and the nonlinear nature of the problem (9)–(13), analytical exploration of the parameter influence is highly complex [2]. To address this, we develop a PIML framework to map the initial configuration directly to the asymptotic trichotomy: $\mathcal{F} : \mathcal{P} \to \{\text{Vanishing, Balancing, Spreading}\}$.

First, we generate a synthetic dataset of $N = 10^5$ samples using Latin Hypercube Sampling (LHS) across the parameter space. Each sample is labeled by running the ADE scheme until $T = 100$:

$$\text{Regime} = \begin{cases} \text{Vanishing} & \text{if } h(T) < \hat{h} \\ \text{Balancing} & \text{if } h(T) > \hat{h}, \quad |h'(T)| < \epsilon \\ \text{Spreading} & \text{if } h(T) \gg \hat{h}, \quad h'(T) \approx const \end{cases} \tag{19}$$

where $\hat{h} = \frac{\pi}{2}\sqrt{D/a}$ represents the critical habitat size [2]. The derivative $h'(T)$ is estimated via a linear regression on the final 10% of the temporal trajectory.

Instead of using only raw parameters, we embed additional physical constraints of the Fisher-KPP and Stefan problems into the training data:

1. Kernel mass, which dictates the balance between the expansion and contraction forces:

$$\int_0^\infty w(y)\mathrm{d}y = \frac{c_1}{\alpha_1} - \frac{c_2}{\alpha_2} \tag{20}$$

2. Survival ratio $h_0/\hat{h}$.

3. Ratios of the kernel parameters $\frac{c_1}{c_2}$ and $\frac{\alpha_1}{\alpha_2}$.

4. Classic Fisher-KPP steady spreading speed $c^* = 2\sqrt{rD}$.

The balancing regime occupies a narrow bifurcation zone where the population's fate is highly sensitive to parameter fluctuations. To address this, we employ a Bagging Random Forest; by aggregating multiple decision trees, this ensemble approach reduces prediction variance and provides a more stable classification of the system's behavior near these sharp transition boundaries.

4 Results

4.1 Steady State

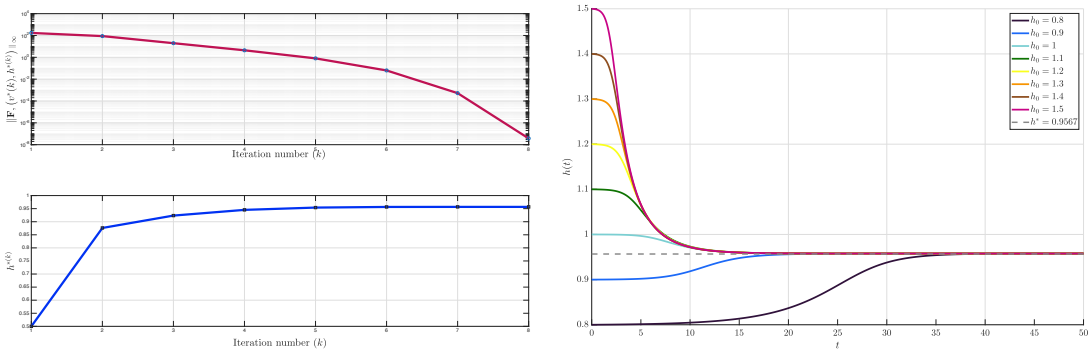
A distinguishing feature of the considered nonlocal free-boundary model is the emergence of a balancing state, where internal growth and perceived boundary conditions reach a spatial equilibrium. To identify this equilibrium, we employ Newton's method to solve the stationary version of system (1)-(7) after applying the FF transformation :

$$\left.\begin{aligned} \frac{D}{(h^*)^2}\, v^*_{zz}(z) + f(v^*(z)) &= 0, \qquad 0 < z < 1, \\ \mu\, h^* \int_0^1 v^*(z)\, w\big(h^*(1-z)\big)\, \mathrm{d}z &= 0, \\ v^*_z(0) = 0, \quad v^*(1) &= 0. \end{aligned}\right\} \tag{21}$$

where $f(v^*(z)) = v^*(z)(a - v^*(z))$. The implementation of the FF transformation provides a significant computational advantage: by mapping the moving boundary onto a fixed interval, the

nonlocal integral is evaluated over the constant domain $[0, 1]$. This approach avoids the complexities of time-dependent integration limits, simplifying the numerical treatment of the perception kernel and allowing the use of the same high-resolution spatial grid for both the stationary Newton solver and the dynamic ADE scheme.

As illustrated in Figure 1(a), the numerical steady-state solver exhibits rapid quadratic convergence, reaching a residual tolerance of 10^{-6} within seven iterations. For the baseline parameter set $(D = 1, a = 5, c_1 = 1.6, c_2 = 1)$, the system yields an equilibrium habitat size of $h^* \approx 0.9567$. To validate this result, we perform dynamical simulations using the ADE scheme starting from various initial habitat sizes $h_0 \in [0.8, 1.5]$ (see Figure 1(b)). Regardless of the initial configuration, provided h_0 exceeds the survival threshold $\hat{h}$, all boundary trajectories converge asymptotically to the same stationary value h^*. This confirms that the balancing regime represents a stable attractor in the system's phase space.



(a) Newton's Method Convergence

(b) Asymptotic Convergence of $h(t)$

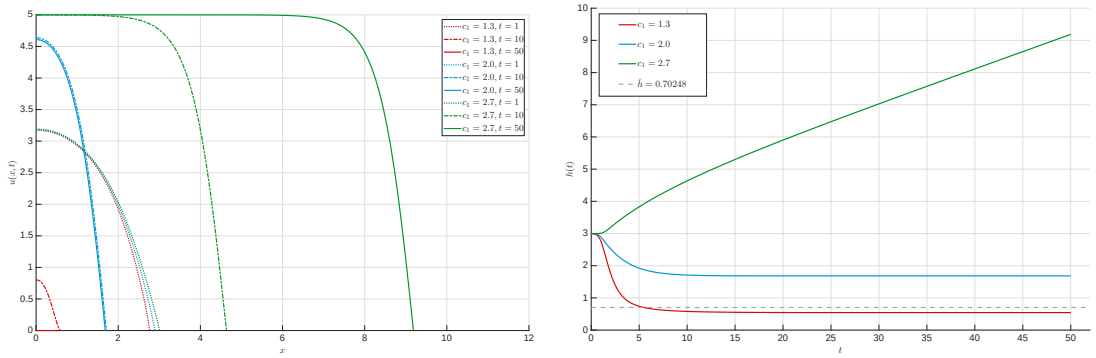
Figure 1: (a) Quadratic decay of the residual norm in the steady-state solver. (b) Temporal evolution of the free boundary for different h_0, converging to the equilibrium $h^* \approx 0.9567$.

4.2 Trichotomy

The capacity of the proposed method to resolve the model's complex asymptotic landscape is demonstrated in Figure 2. By varying the kernel expansion amplitude c_1, we observe how the system traverses the three distinct regimes of the trichotomy. For these simulations, we maintain a baseline configuration with $D = 1, r = 1, a = 5, \mu = 1, \alpha_1 = 1.9, \alpha_2 = 1, c_2 = 1$, and an initial habitat size $h_0 = 3$.

The results identify c_1 as a critical bifurcation parameter that dictates the long-term survival of the species. In the vanishing regime ($c_1 = 1.3$, red lines), the contractionary forces encoded in the kernel overwhelm the expansionary drive, causing a spatial retreat of the boundary and a total collapse of the population density toward extinction. Conversely, the spreading regime ($c_1 = 2.7$, green lines) represents an invasion scenario where the expansionary component dominates; here, the species successfully occupies the domain as the interior density converges to the carrying capacity $a = 5$, while the boundary undergoes indefinite linear growth.

The balancing regime ($c_1 = 2.0$, blue lines) provides the most stringent test for the numerical scheme. In this intermediate case, the competitive nonlocal forces reach a precise equilibrium, resulting in a stationary, non-trivial population profile $u^*(x)$ supported on a finite habitat h^*. Notably, the profile overlap observed at $t = 10$ and $t = 50$ in Figure 2(a) underscores the numerical stability and accuracy of the ADE scheme, confirming its ability to capture the fine-scale details of the stationary state without the introduction of artificial oscillations or numerical diffusion.



(a) Population density (b) Free boundary

Figure 2: Numerical simulation of the Fisher-KPP problem illustrating the spreading-balancing-vanishing trichotomy for different expansion parameters $c_1 \in \{1.3, 2.0, 2.7\}$.

4.3 Sensitivity and Classification

The PIML classifier achieves a 94% accuracy rate in predicting the trichotomy. The balancing regime, where the boundary stabilizes at a finite h^*, represents a narrow bifurcation zone between spreading and vanishing.

Using the trained PIML surrogate, we generate high-resolution phase maps. The relationship between Kernel Mass and the Survival Ratio (Fig. 3(c)) provides the most distinct classification of the system outcomes. There is a strict horizontal boundary at a Survival Ratio of approximately 1.0. Below this critical survival threshold ($h_0 < \hat{h}$), the population vanishes regardless of the kernel mass. For populations that exceed the survival threshold, the outcome is determined by the Kernel Mass ($\frac{c_1}{\alpha_1} - \frac{c_2}{\alpha_2}$). A negative mass typically leads to vanishing, while a positive mass supports spreading. The balancing state exists as a narrow vertical transition zone where the kernel mass is near zero.

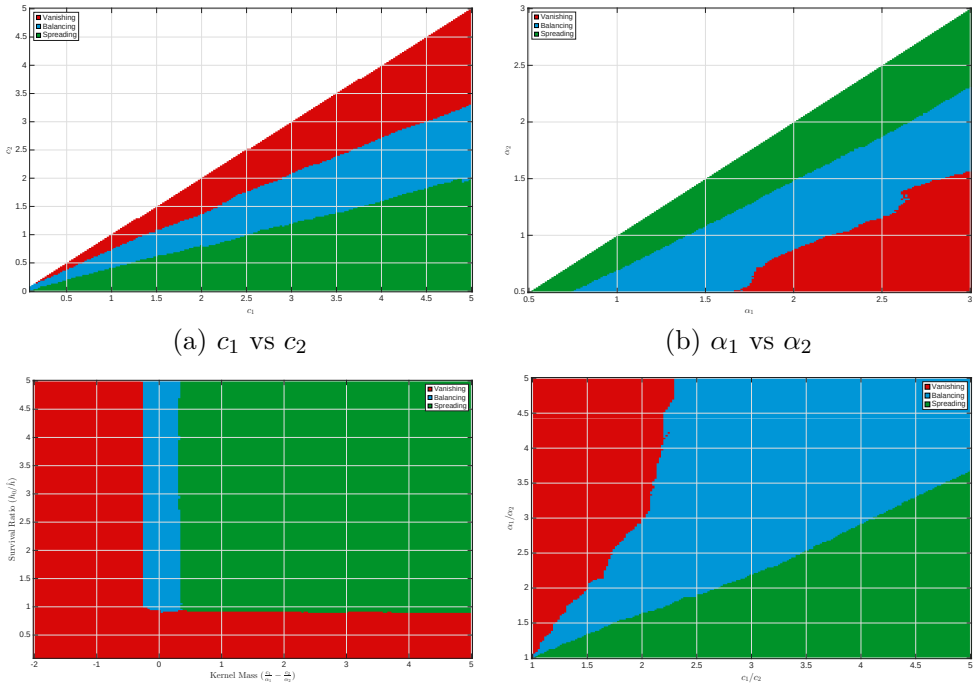
5 Conclusions

In this study, we have presented a comprehensive computational framework to analyze the complex spatiotemporal dynamics of a Fisher-KPP model with a nonlocal free boundary. By integrating advanced numerical discretization with machine learning, we have addressed the challenges inherent in characterizing the spreading–balancing–vanishing trichotomy that governs species survival.

The proposed PIML architecture serves as a high-fidelity surrogate, bridging the gap between rigorous mathematical theory and real-time ecological prediction. By explicitly embedding theoretical constructs, such as the survival ratio ($h_0/\hat{h}$) and the net kernel mass, into the feature space, the model achieved a 94% accuracy rate in predicting asymptotic outcomes. This surrogate model effectively captures the "critical slowing-down" phenomena near bifurcation boundaries, enabling the instantaneous generation of high-resolution phase-space maps without the PDE simulations.

Acknowledgments

This research has been supported by the Government of Cantabria through the Program Contract between the Government of Cantabria and the University of Cantabria (Contrato Programa Gobierno de Cantabria–UC), under the project "Sistemas Diferenciales Aleatorios en Dinámica de Poblaciones: Métodos Numéricos, Análisis y Simulaciones" (Ref: 20.VQ19.64662).



(a) c_1 vs c_2 (b) α_1 vs α_2

(c) Kernel Mass vs Survival Ratio (d) Ratio c vs Ratio α

Figure 3: Comprehensive PIML phase space maps for the trichotomy outcomes.

References

[1] Y. Du and Z. Lin, "Spreading-vanishing dichotomy in the diffusive logistic model with a free boundary," *SIAM Journal on Mathematical Analysis*, vol. 42, pp. 377–405, 1 Jan. 2010, ISSN: 0036-1410. DOI: 10.1137/090771089.

[2] C. Feng, M. A. Lewis, C. Wang, and H. Wang, "A Fisher–KPP model with a nonlocal weighted free boundary: Analysis of how habitat boundaries expand, balance or shrink," *Bulletin of Mathematical Biology*, vol. 84, 3 Mar. 2022, ISSN: 15229602. DOI: 10.1007/s11538-022-00995-8.

[3] R. B. Salako and S. Xue, "Dynamics of a nonlocal weighted moving boundary model in time-periodic environments," *Journal of Differential Equations*, vol. 464, 2026, ISSN: 10902732. DOI: 10.1016/j.jde.2026.114246.

[4] Q. Zhou, Z. Lin, and C. A. Santos, "Vanishing, balancing and spreading for an impulsive faecal-oral model," *Communications in Nonlinear Science and Numerical Simulation*, vol. 158, Jul. 2026, ISSN: 10075704. DOI: 10.1016/j.cnsns.2026.109814.

[5] H. Z. Barakat and J. A. Clark, "On the solution of the diffusion equations by numerical methods," *Journal of Heat Transfer*, vol. 88, pp. 421–427, 4 Nov. 1966, ISSN: 0022-1481. DOI: 10.1115/1.3691590.

[6] N. Malekjani, A. Bück, A. Kharaghani, and E. Tsotsas, "SDD-PINN: Physics-informed neural network for single droplet drying," *Digital Chemical Engineering*, vol. 19, p. 100 306, Jun. 2026, ISSN: 2772-5081. DOI: 10.1016/J.DCHE.2026.100306.

On the existence and positivity of almost periodic solutions of a plankton model with seasonality

Abraham J. Arenas[a,b,1] Gilberto González-Parra[b]
and Myladis R. Cogollo[a]

(a) Departamento de Matemáticas y Estadística, Universidad de Córdoba.
(b) Department of Mathematics, New Mexico Tech, Socorro, New Mexico, USA.

1 Introduction

In this work we study the existence of a positive almost periodic solutions for a compartmental model of plankton ecosystems. We investigate the impact of introducing Holling Type III functional forms and quadratic mortality. We include these terms into a generalized plankton model with seasonality. The model is given by the following non-autonomous system of ordinary differential equations,

$$\begin{aligned}
\dot{P}_1(t) &= P_1(t)\left[\frac{\mu_1(t)N(t)}{N(t)+\kappa_1} - \frac{\phi_1(t)Z(t)P_1^a(t)}{\Delta_1(t)} - \sigma_1(t)P_1^a(t)\right],\\
\dot{P}_2(t) &= P_2(t)\left[\frac{\mu_2(t)N(t)}{N(t)+\kappa_2} - \frac{\phi_2(t)Z(t)P_2^a(t)}{\Delta_2(t)} - \sigma_2(t)P_2^a(t)\right],\\
\dot{Z}(t) &= Z(t)\left[\frac{\phi_1(t)(1-\psi_1)P_1^{a+1}(t)}{\Delta_1(t)} + \frac{\phi_2(t)(1-\psi_2)P_2^{a+1}(t)}{\Delta_2(t)} - \sigma_Z(t)Z^a(t)\right],\\
\dot{N}(t) &= -\frac{\mu_1(t)P_1(t)N(t)}{N(t)+\kappa_1} - \frac{\mu_2(t)P_2(t)N(t)}{N(t)+\kappa_2} + \frac{\phi_1(t)\psi_1 Z(t)P_1^{a+1}(t)}{\Delta_1(t)}\\
&\quad + \frac{\phi_2(t)\psi_2 Z(t)P_2^{a+1}(t)}{\Delta_2(t)} + \sigma_1(t)P_1^{a+1}(t) + \sigma_2(t)P_2^{a+1}(t) + \sigma_Z(t)Z^{a+1}(t),
\end{aligned} \tag{1}$$

where $\Delta_1(t) = 1+\epsilon_1(t)P_1^{a+1}(t)+\rho\epsilon_2(t)P_2^{a+1}(t), \Delta_2(t) = 1+\epsilon_2(t)P_2^{a+1}(t)+\rho\epsilon_1(t)P_1^{a+1}(t)$ are subject to the mass conservation constraint for nutrient N,

$$\dot{P}_1(t) + \dot{P}_2(t) + \dot{Z}(t) + \dot{N}(t) = 0, \tag{2}$$

with $P_1 + P_2 + Z + N = 1$, $P_1(0) > 0, P_2(0) > 0, Z(0) > 0, N(0) > 0$ and $\mu_1(t), \mu_2(t)$ are T-periodic functions where there are positives constants such that $\mu_i^l \leq \mu_i(t) \leq \mu_i^u$, for $i = 1, 2$.

[1] aarenas@correo.unicordoba.edu.co

2 Methods

The functions $\phi_1(t), \phi_2(t)$ represent the attack coefficients or consumption rates of the predator on the preys P_1 and P_2. They indicate the intensity with which an individual predator captures or consumes prey per unit of time in the absence of saturation. Next, the functions $\sigma_1(t), \sigma_2(t), \sigma_Z(t)$ represent mortality or loss rates independent of predation:

- σ_1, σ_2: natural mortality rate of prey P_1 and P_2.
- σ_Z: natural mortality rate of predator Z.

Moreover, $\kappa_1, \kappa_2, \psi_1, \psi_2 \in (0, 1)$, and when $a = 0$ we obtain the Holling Type II in system (1), whereas $a = 1$ yields the fully quadratic system equivalent to the system proposed by Bates et. al. in [1]. We will assume $0 < a \leq 1$. The meaning of each of the parameters is presented in Table 1.

Table 1: Interpretation of parameters given in (1), see [2]

Parameter	Process
κ_1	Half-saturation constant for N uptake by P_1
κ_2	Half-saturation constant for N uptake by P_2
ϕ_1	Z grazing rate on P_1
ϕ_2	Z grazing rate on P_2
ϵ_1	Half-saturation constant for Z uptake of P_1
ϵ_2	Half-saturation constant for Z uptake of P_2
σ_1	P_1 specific mortality rate
σ_2	P_2 specific mortality rate
σ_Z	Z specific mortality rate
ψ_1	Proportion of P_1 uptake excreted by Z
ψ_2	Proportion of P_2 uptake excreted by Z

.

We analyze the system dynamics based on the principles of permanence, stability, existence, and uniqueness of a quasi-periodic positive solution. To prove the global attractiveness of a positive solution, we use a suitable Lyapunov function. In addition, we prove the existence and uniqueness of the solution. We show that the sufficient conditions obtained for a globally attractive unique positive solution depend on the model parameters. The analytical results are illustrated with numerical simulations.

It is important to mention that the method used to investigate the positive T-periodic solution of the nonlinear ecosystem, using the degree of coincidence theory (see [3]) or Brower's fixed-point theorem (see [4]), cannot be used to investigate the almost periodic solution of system (1). The assumption of periodicity for the parameters μ_1 and μ_2 is one way to incorporate the seasonality of the environment. A very important problem in the study of mathematical models with population-growth in a periodic environment, is the existence and global stability of positive periodic solutions. This plays a similar role as a globally stable equilibrium does in the autonomous model. Seasonal forcing has been shown to be important in generating the cycles observed in many ecological and epidemiological systems. For instance, by including seasonal forcing to represent changes in the transmission of viruses during the school year. By doing this numerical simulations of mathematical models have been shown to be consistent with observations of measles case reports, [5], [6].

Phytoplankton blooms are population explosions in the presence of sunlight and nutrients, lasting from a few days to several weeks, until the nutrient has been consumed or they are grazed down by zoo-plankton. In temperate and polar oceans phytoplankton blooms are closely associated

with seasonal changes, [2], [7]. Thus, it is reasonable to seek a positive periodic solution that is globally asymptotically stable for system (1).

3 Mathematical definitions and results

For convenience, we will use the following notations in the discussion: for a continuous function $f(t)$ defined on $[0,T]$, we denote by $\overline{f}$ its mean value,

$$\overline{f} = \frac{1}{T}\int_0^T f(t)dt.$$

Let us denote $\mathbb{R}^-$ the set of real negative, and $\mathbb{R}^+$ the set of real non-negative. Thus, from the point of view of mathematical biology, we choose

$$\mathcal{D} = \{(x_1, x_2, x_3, x_4)/0 < x_i,\, i = 1, \cdots, 4\} \subset \mathbb{R}^4_+ \tag{3}$$

to study the dynamical of the solutions of model (1).

A biological property of ecological models is permanence, which means that the species involved in the system persist regardless of the initial population conditions. Mathematically, this implies that the solutions of the model, represented by differential equations, remain strictly positive within the invariant region and possess no limit points on its boundary, [8].

Definition 3.1. *System (1) is called permanent if there are positive constants m, M such that*

$$\min\{\liminf_{t\longrightarrow\infty} P_1(t), \liminf_{t\longrightarrow\infty} P_2(t), \liminf_{t\longrightarrow\infty} Z(t), \liminf_{t\longrightarrow\infty} N(t)\} \geq m,$$
$$\max\{\limsup_{t\longrightarrow\infty} P_1(t), \limsup_{t\longrightarrow\infty} P_2(t), \limsup_{t\longrightarrow\infty} Z(t), \limsup_{t\longrightarrow\infty} N(t)\} \leq M,$$

see [9].

The following lemma is important to prove the permanence of model (1).

Lemma 3.2. *Let a, b, α positive numbers and $w'(t) \leq (\geq) w(t)\left(b - a\, w(t)^\alpha\right)$, $w(0) > 0$. When $t \geq t_1$ and $w(t_1) > 0$, then*

$$\limsup_{t\to\infty} w(t) \leq \left(\frac{b}{a}\right)^{\frac{1}{\alpha}} \left(\liminf_{t\to\infty} w(t) \geq \left(\frac{b}{a}\right)^{\frac{1}{\alpha}}\right) \tag{4}$$

Proof. See [10]. □

3.1 Existence of solutions, positivity and permanence

Since that we put $a \in (0,1]$, from now on we will use the following $0 < P_i(t), Z(t), N(t) \leq 1$ for all $t \in (0,\infty)$, $P_i^{a+1}(t), Z^{a+1}(t), N^{a+1}(t) \leq P_i(t), Z(t), N(t) \leq P_i^a(t), Z^a(t), N^a(t) \leq 1$. Moreover, we define the following real numbers

$$\sigma_m = \min\{\sigma_1, \sigma_2, \sigma_Z\},\ \sigma_M = \max\{\sigma_1, \sigma_2, \sigma_Z\}.$$

Let $\mathcal{X}(t) = (P_1(t), P_2(t), Z(t), N(t))$, and denote the right–hand side of (1) by

$$F(\mathcal{X}(t)) = \Big(f_1(\mathcal{X}(t)),\, f_2(\mathcal{X}(t)),\, f_3(\mathcal{X}(t)),\, f_4(\mathcal{X}(t))\Big)$$

Since $F \in C^1(\mathcal{D})$, the mapping $F : \mathcal{D} \to \mathbb{R}^4_+$ is locally Lipschitz on $\mathcal{D}$. Therefore, by the fundamental existence–uniqueness theorem, system (1) admits a unique solution for any prescribed initial condition. The following lemma guarantees the positivity of solutions.

Lemma 3.3. *For system* (1)*, the set* $\mathcal{D}$ *constitutes a positively invariant region.*

Theorem 3.4. *For any solution* $\mathcal{X}$ *evolving in the* $\mathcal{D}$*, system* (1) *satisfies*

$$\min\left\{\liminf_{t\to\infty} P_1(t),\ \liminf_{t\to\infty} P_2(t),\ \liminf_{t\to\infty} Z(t),\ \liminf_{t\to\infty} N(t)\right\} \geq m_{min},$$
$$\max\left\{\limsup_{t\to\infty} P_1(t),\ \limsup_{t\to\infty} P_2(t),\ \limsup_{t\to\infty} Z(t),\ \limsup_{t\to\infty} N(t)\right\} \leq M_{max}, \tag{5}$$

where $m_{min} = \{I_{P_1}, I_{P_2}, I_Z, I_N\}$*, and* $M_{max} = \{S_{P_1}, S_{P_2}, S_Z, S_N\}$ *for* $t > T$*. Therefore, system* (1) *holds the Definition 3.1.*

3.2 Numerical simulations

In this section we show a few numerical simulations that show the global stability of a almost periodic solution. Figure 1 shows the global attractivity of the solutions of system (1). Figure 2 shows the existence of an almost periodic solution of system (1).

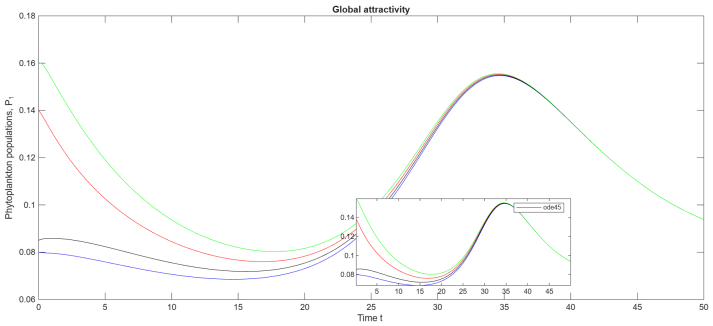


Figure 1: Global attractivity of the solutions of system (1).

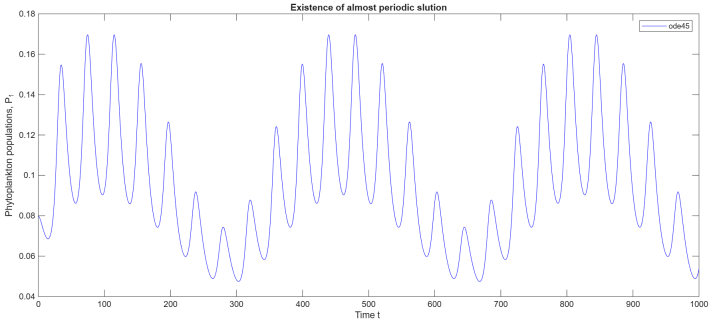


Figure 2: Existence of an almost periodic solution of system (1).

4 Conclusions

In this work, a non-autonomous phytoplankton, zooplankton, and nutrient model is generalized. The global attractivity of quasi-periodic positive solutions is studied using a suitable Lyapunov function. The Arzela-Ascoli Theorem and Bochner's definition of a quasi-periodic solution are used. Numerical simulations that support the analytical results are shown.

Acknowledgments

The first author and third authors wish to thank the University of Cordoba for its support.

References

[1] M. L. Bates, S. M. B. Nash, D. W. Hawker, J. Norbury, J. S. Stark, and R. A. Cropp, "Construction of a trophically complex near-shore Antarctic food web model using the Conservative Normal framework with structural coexistence," *Journal of Marine Systems*, vol. 145, pp. 1–14, 2015.

[2] I. M. Moroz, R. Cropp, and J. Norbury, "Chaos in plankton models: Foraging strategy and seasonal forcing," *Ecological Modelling*, vol. 332, pp. 103–111, 2016.

[3] R. E. Gaines and J. L. Mawhin, *Coincidence degree and nonlinear differential equations*. Springer, 2006.

[4] H. Freedman and J. Wu, "Periodic solutions of single-species models with periodic delay," *SIAM Journal on Mathematical Analysis*, vol. 23, no. 3, pp. 689–701, 1992.

[5] R. A. Taylor, A. White, and J. A. Sherratt, "Seasonal forcing in a host–macroparasite system," *Journal of Theoretical Biology*, vol. 365, pp. 55–66, 2015.

[6] R. Cropp and J. Norbury, "Simple predator–prey interactions control dynamics in a plankton food web model," *Ecological Modelling*, vol. 220, no. 13–14, pp. 1552–1565, 2009.

[7] M. Gao, H. Shi, and Z. Li, "Chaos in a seasonally and periodically forced phytoplankton–zooplankton system," *Nonlinear Analysis: Real World Applications*, vol. 10, no. 3, pp. 1643–1650, 2009.

[8] A. Mandal, P. K. Tiwari, S. Samanta, E. Venturino, and S. Pal, "A nonautonomous model for the effect of environmental toxins on plankton dynamics," *Nonlinear Dynamics*, vol. 99, no. 4, pp. 3373–3405, 2020.

[9] H. Li and Y. Takeuchi, "Dynamics of the density dependent predator–prey system with Beddington–DeAngelis functional response," *Journal of Mathematical Analysis and Applications*, vol. 374, no. 2, pp. 644–654, 2011.

[10] C. Shen, "Permanence and global attractivity of the food-chain system with Holling IV type functional response," *Applied Mathematics and Computation*, vol. 194, no. 1, pp. 179–185, 2007.

A Generic Model of CD8 T Cell Dynamics: Mathematical and Sensitivity Analysis

Oscar Angulo♭, Fabien Crauste♮, [1]

(♭) Departamento de Matemática Aplicada, Facultad de Ciencias, Universidad de Valladolid
Pso. Belén 7, 47011 Valladolid, Spain

(♮) Université Paris Cité, CNRS, MAP5
F-75006 Paris, France.

1 Introduction

Acquired immune responses are triggered by the organism in order to specifically target a pathogen. CD8 T cell responses correspond to the activation of naive CD8 T cells, located in lymph nodes, by antigen-presenting cells, and their subsequent proliferation and differentiation in cytotoxic cells, called effector cells, and memory cells.

Two distinct responses exist: so-called primary responses, when the organism has never encountered the pathogen, and secondary responses—or memory responses—when the organism has already encountered the pathogen and performed a primary response. Vaccines are supposed to induce a primary response so memory cells are already present in the organism when the "true" pathogen infects it.

Mathematical models of the CD8 T cell response mostly focused on the description of a primary response [1-4], in lab animals, when the entire response can be controled by the experimentalist: the amount of antigen used to induce the response is mostly controled, the naive cells that respond are identified. On the contrary, very few models focused on the secondary CD8 T cell response [5].

2 Methods

We will introduce a new model of the CD8 T cell response describing both primary and secondary immune responses. The model builts on a standard approach using partial differential equations and introducing a structured variable defined as the maturity of cells. In addition, the model accounts for both proliferating (effector) and non-proliferating (naive, memory) cells.

The resulting nonlinear and nonlocal model couples both ordinary differential equations accounting for dynamics of resting cells and the immunogen, and a transport particla differential equation describing dynamcis of proliferating cells. Depending on the level of maturity of resting and proliferating cells, the two stages can exchange cells: resting cells are activated by the immunogen through antigen-presenting cells and start proliferating, while proliferating cells can differentiate in resting memory cells.

[1] fabien.crauste@math.cnrs.fr

Linear stability of the coupled ODE-PDE model is investigated with standard methods. Parameter estimation and model fitting to data is performed numerically on mice data. Sensitivity analysis is performed by computing first and second order Sobol indices.

3 Results

The model is introduced based on biological assumptions, and existence of steady state solutions is established through mathematical analysis. It shows that the system has two steady states that both describe an extinction of the CD8 T cell population in the absence of CD8 T cell production, one describing the extinction of the pathogen and the other maintenance of it.

Stability analysis of the steady states proves that the steady state where all cell populations go extinct is locally asymptotically stable, while the other steady state is unstable.

Model fitting on experimental data of a CD8 T cell immune response in mice and estimation of parameter values highlights the ability of the model to describe a real CD8 T cell immune response. Choices of the various functions characterizing the model are discussed since they impact the results.

The model fitting step motivates a local sensitivity analysis, performed using Sobol indices, that identifies the main contributors to the dynamics of the CD8 T cell response.

4 Conclusions

The model we introduced shows equivalent properties than previously published models to describe a primary immune response. It must still be challenged on secondary immune response, for which data are scarce.

Identifiability issues were raised while performing model fitting, these issues need to be addressed. However, the Sobol analysis highlights key parameters that could ensure identifiability.

This is still a work in progress.

References

Bibliography

[1] Antia R, Bergstrom C, Pilyugin S, Kaech S, Ahmed R (2003) Models of CD8+ Responses: 1. What is the antigen-independent proliferation program. J Theor Biol 221:585–598

[2] Crauste F, Mafille J, Boucinha L, Djebali S, Gandrillon O, Marvel J, Arpin C (2017). Identification of Nascent Memory CD8 T Cells and Modeling of Their Ontogeny. Cell Syst 4(3):306-317.

[3] Crauste F, Terry E, Mercier I, Mafille J, Djebali S, Andrieu T, Mercier B, Kaneko G, Arpin C, Marvel J, Gandrillon O (2015) Predicting pathogen-specific CD8 T cell immune responses from a modeling approach. J Theor Biol 374:66-82

[4] De Boer RJ, Oprea M, Antia R, Murali-Krishna K, Ahmed R, Perelson AS (2001) Recruitment times, proliferation, and apoptosis rates during the CD8(+) T-cell response to lymphocytic choriomeningitis virus. J Virol 75(22):10663-9.

[5] Zarnitsyna VI, Handel A, McMaster SR, Hayward SL, Kohlmeier JE, Antia R (2016) Mathematical Model Reveals the Role of Memory CD8 T Cell Populations in Recall Responses to Influenza. Front Immunol 7:165.

Part X

Student's Projects

Parallel solution of the Schrödinger eigenvalue problem via real rational schemes

Sergio Blanes[♭], José L. Ruiz-Benito[♮1]

(♭) I.U. de Matemática Multidisciplinar, Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

(♮) Institut de Matemàtiques i Aplicacions de Castelló, Universitat Jaume I,
Av. Vicent Sos Baynat, s/n, Castelló de la Plana, Spain.

1 Introduction

The Schrödinger eigenvalue problem can be solved using the Imaginary Time Propagation (ITP) technique. This approach transforms the problem into a parabolic differential equation. While numerical methods suited for this kind of equations work reasonably well, most fail to exploit the parallel processing power of modern computers.

Rational approximations, such as Chebyshev ones, offer a solution. By using partial fraction decomposition, the exponential function is split into a sum of simple fractions (a number divided by a first-degree polynomial). This allows each fraction to be computed on a separate process simultaneously and then combined.

However, this decomposition usually introduces complex numbers, which increases the computational cost. In this work, we overcome this drawback by designing new rational approximations based entirely on real numbers. This preserves the benefits of parallel computing while eliminating the extra cost of complex arithmetic.

Main references: [1], [2].

2 Methods

Consider the time-dependent Schrödinger equation

$$i\frac{\partial}{\partial t}\psi(x,t) = H\psi(x,t), \quad \psi(x,0) = \psi_0(x), \tag{1}$$

where the Hamiltonian

$$H = T + V(x) = -\frac{1}{2}\Delta + V(x)$$

is a Hermitian operator, Δ is the Laplacian, and $V(x)$ denotes the interaction potential.

We aim to compute the eigenvalues $0 \le E_0 \le E_1 \le E_2 \le \dots$ of H, with particular interest in E_0. Applying the imaginary-time transformation $t = -i\tau$ converts (1) into the parabolic equation

$$-\frac{\partial}{\partial \tau}\psi(x,\tau) = H\psi(x,\tau), \quad \psi(x,0) = \psi_0(x), \tag{2}$$

[1] jbenito@uji.es

whose formal solution is $\psi(x,\tau) = e^{-\tau H}\psi_0(x)$. Expanding $\psi(x,\tau)$ in the eigenfunction basis $\phi_i(x)$ of H yields

$$\psi(x,\tau) = \sum_i e^{-\tau E_i} c_i \phi_i(x), \quad c_i = \langle \phi_i(x), \psi_0(x) \rangle,$$

where $\langle \cdot, \cdot \rangle$ represents the standard L^2 scalar product. For a long time integration we have that

$$\psi(x,\tau) \longrightarrow e^{-\tau E_0} c_0 \phi_0(x).$$

The first eigenfunction $\phi_0(x)$ is then recovered by normalization, and its corresponding eigenvalue is computed via $E_0 = \langle \phi_0, H\phi_0 \rangle$.

Consequently, the problem reduces to executing the time integration of (2). This can be achieved by constructing a robust rational approximation of the exponential function to evaluate $\psi(x,\tau) = e^{-\tau H}\psi_0(x)$. After spatial discretization, the task becomes computing

$$e^{-\tau M} v,$$

where M is a matrix with positive eigenvalues, v is a vector, and τ is a large positive parameter. We thus seek an approximating function $g(x) \approx e^x$ for $x \leq 0$ of the form

$$g(x) = \sum_{k=1}^{N} \frac{b_k}{1 - c_k x},$$

with real coefficients b_k and $c_k \geq 0$.

Taking $g(x)$ as the approximation of e^x that minimizes $\max_{x \leq 0}\{|e^x - g(x)|\}$ results in complex coefficients c_k. To avoid the computational overhead of complex arithmetic, we implement a minimization strategy to find real parameters that minimize a chosen norm $\|e^x - g(x)\|$. A practical ansatz defines this error as an integral

$$\|e^x - g(x)\|^2 = \int_{-\infty}^{0} \left(e^x - g(x)\right)^2 w(x)\, dx,$$

using a suitable weight function $w(x)$ such as e^x or $1/(x^2+1)$. We also enforce the consistency condition $g(0) = 1$ and the optimal function $g(x)$ can be obtained via numerical optimization.

Once $g(x)$ is determined, the action of the matrix exponential is approximated as

$$e^{-\tau M} v \approx g(-\tau M) v = \sum_{k=1}^{N} b_k \left(I + c_k \tau M\right)^{-1} v,$$

where each component $y_k = \left(I + c_k \tau M\right)^{-1} v$ satisfies the linear system

$$\left(I + c_k \tau M\right) y_k = v.$$

As a result, the time integration is reduced to solving N independent linear systems. Because these systems can be distributed simultaneously across different processing cores, the effective computational time of the algorithm scales down to that of a single linear system solve.

3 Results and conclusions

The scalar functions g obtained through this minimization strategy exhibit a balanced behavior when approximating the exponential function. They are highly accurate for values of τ close to zero and degrade gracefully for larger values, while stability is inherently guaranteed by construction. When applying g to the eigenvalue problem, this robust performance is preserved, enabling accurate long-time integration in a single step to easily reach a reliable approximation of the ground state. If desired, a local algorithm can subsequently be applied to achieve high-precision results for the ground state.

Acknowledgments

This work has been supported by Ministerio de Ciencia e Innovación (Spain) through project PID2022-136585NB-C21, MCIN/AEI/10.13039/501100011033/FEDER, UE.

References

[1] P. Bader, S. Blanes, and F. Casas, "Solving the schrödinger eigenvalue problem by the imaginary time propagation technique using splitting methods with complex coefficients," *J. Chem. Phys.*, vol. 139, 2013.

[2] S. Blanes, "Parallel computation of functions of matrices and their action on vectors for exponential integrators," *J. Comput. Appl. Math.*, vol. 476, 2026.

Numerical Methods for Solving the Black–Scholes Equation

Carlos Sevilla Cabedo[♭],[1] Manuel Sellés Armendia[♮], [2] Roger Rosado Santos[♮],[3]

Universitat Politècnica de València
Camí de Vera s/n, València, Spain.

1 Introduction

This paper explores the application of numerical methods to solve partial differential equations (PDEs) within financial markets. Guided by professors Alicia Cordero and Juan Ramón Torregrosa, this work bridges foundational numerical analysis with real-world quantitative finance, aiming to evaluate the practical implementation of these techniques in modern derivative pricing.

Applying PDEs to finance requires understanding foundational derivative instruments. Options, a primary class of derivatives, grant the holder the right to buy (Call) or sell (Put) an underlying asset at a predetermined strike price K. These contracts are generally classified as European (exercisable strictly at expiration) or American (exercisable at any time prior to expiration). The underlying asset's price evolution is typically modeled using Geometric Brownian Motion, which is heavily influenced by market volatility σ and the risk-free interest rate r. The terminal value of these contracts is strictly defined by their payoff functions: $\max(S_T - K, 0)$ for Calls and $\max(K - S_T, 0)$ for Puts, where S_T represents the asset price at maturity.

While the classical Black-Scholes framework provides exact, closed-form analytical solutions under idealized assumptions, these formulas fall short when evaluating exotic options or incorporating realistic market constraints (such as early exercise features). Consequently, numerical methods become indispensable. Although they require significant computational resources and yield approximate solutions, they offer the flexibility and robustness necessary to solve complex financial PDEs. By analyzing finite difference schemes and stochastic simulations, this study highlights the critical role of numerical methods as a cornerstone of quantitative finance.

2 Analyzed Aspects

2.1 The 1-D Black–Scholes Equation

The valuation of financial derivatives in continuous time is governed by the Black–Scholes Partial Differential Equation (PDE). To solve this problem numerically as an initial value problem, we apply a time-reversal transformation $\tau = T - t$. The resulting diffusion-convection-reaction equation is:

$$\frac{\partial V}{\partial \tau} = \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV, \tag{1}$$

[1] csevcab@teleco.upv.es
[2] mselarm@teleco.upv.es
[3] rrossan@teleco.upv.es

defined on the domain $(S, \tau) \in [0, S_{max}] \times [0, T]$. The payoff function, which is the initial condition is:

$$V(S, 0) = \text{payoff}(S) \tag{2}$$

The boundary conditions are of Dirichlet Type and are given by:

$$V(0, \tau) = f(\tau), \quad V(S_{\max}, \tau) = g(\tau) \tag{3}$$

Numerical Schemes and Formulations

Explicit Scheme

The explicit scheme approximates the time derivative using forward differences and spatial derivatives with central differences. The discrete recurrence relation is:

$$V_{i,j+1} = \frac{\lambda}{2}(\sigma^2 S_i^2 - hrS_i)V_{i-1,j} + (1 - \sigma^2 S_i^2 \lambda - rk)V_{i,j} + \frac{\lambda}{2}(\sigma^2 S_i^2 + hrS_i)V_{i+1,j}, \tag{4}$$

where $\lambda = k/h^2$.

Implicit Scheme

Using backward differences in time results in an **unconditionally stable** scheme.
Firstly, the coefficients are defined as:

$$\begin{aligned} a_i' &= -\frac{\lambda}{2}\sigma^2 S_i^2 + \frac{rS_i k}{2h} \\ b_i' &= 1 + \lambda\sigma^2 S_i^2 + rk \\ c_i' &= -\frac{\lambda}{2}\sigma^2 S_i^2 - \frac{rS_i k}{2h} \end{aligned}$$

where $\lambda = k/h^2$. The implicit scheme can be elegantly expressed in matrix form as:

$$AV^{(j)} = V^{(j-1)} + d^{(j)}, \quad j = 1, 2, \ldots, n_t \tag{5}$$

where $V^{(j)}$ is the vector of unknowns at the current time level τ_j, and $V^{(j-1)}$ is the vector of known values from the previous time level τ_{j-1}: The coefficient matrix A is a sparse, tridiagonal matrix of size $(n_x - 1) \times (n_x - 1)$:

$$A = \begin{pmatrix} b_1' & c_1' & 0 & \ldots & \ldots & 0 \\ a_2' & b_2' & c_2' & 0 & \ldots & 0 \\ 0 & a_3' & b_3' & c_3' & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & a_{n_x-2}' & b_{n_x-2}' & c_{n_x-2}' \\ 0 & \ldots & \ldots & 0 & a_{n_x-1}' & b_{n_x-1}' \end{pmatrix}_{(n_x-1)\times(n_x-1)}$$

The vector $d^{(j)}$ contains the contributions from the known Dirichlet boundary conditions at $S = 0$ and $S = S_{max}$:

$$d^{(j)} = \begin{pmatrix} -a_1' V_{0,j} \\ 0 \\ \vdots \\ 0 \\ -c_{n_x-1}' V_{n_x,j} \end{pmatrix}_{(n_x-1)\times 1}$$

Crank–Nicolson Scheme

The Crank-Nicolson method averages the explicit and implicit spatial operators, achieving second-order temporal accuracy $O(k^2) + O(h^2)$. The system is defined as:

$$AV^{(j+1)} = BV^{(j)} + d^{(j)}. \tag{6}$$

This scheme is **unconditionally stable** and serves as the standard for efficient financial PDE resolution.

2.2 Comparison of Finite Difference Methods

The following table summarizes the convergence and stability properties of the implemented schemes:

Table 1: Comparison of Explicit, Implicit, and Crank-Nicolson Finite Difference Schemes

Property	**Explicit Scheme**	**Implicit Scheme**	**Crank-Nicolson Scheme**
Order of Accuracy	$O(k) + O(h^2)$	$O(k) + O(h^2)$	$\mathbf{O(k^2) + O(h^2)}$
Matrix System (per step)	$V^{(j+1)} = AV^{(j)} + d^{(j)}$	$AV^{(j)} = V^{(j-1)} + d^{(j)}$	$AV^{(j+1)} = BV^{(j)} + d^{(j)}$
Computation per step	Direct Matrix Multiplication	Solve Tridiagonal System	Solve Tridiagonal System
Stability	**Conditional** (CFL req.)	**Unconditional**	**Unconditional**
CFL Constraint ($\lambda = k/h^2$)	$\lambda \leq \frac{2-rk}{2\sigma^2 S_{\max}^2}$	No Restriction	No Restriction
Convergence	Conditionally Convergent	Unconditionally Convergent	**Unconditionally Convergent**

3 Results

To evaluate the accuracy and computational efficiency of the derived finite difference schemes (Explicit, Implicit, and Crank-Nicolson), we analyze the pricing of both European and American Put options. The domain is defined as $(S, \tau) \in [0, S_{\max}] \times [0, T]$, where $\tau = T - t$.

Both options share the initial payoff condition $V(S, 0) = \max(K - S, 0)$ and the upper boundary condition $V(S_{\max}, \tau) = 0$. However, they differ at the lower boundary ($S \to 0$): the European Put is discounted to $V(0, \tau) = Ke^{-r\tau}$, whereas the American Put yields immediate cash flow $V(0, \tau) = K$.

The numerical experiments are conducted using the following benchmark parameters:

Parameter	**Symbol**	**Value**
Spot Price	S_0	100
Strike Price	K	100
Risk-free Rate	r	0.05 (5%)
Volatility	σ	0.20 (20%)
Time to Maturity	T	1.0 year
Max Asset Price	$S_{\max}$	300

Table 2: Parameters for the numerical simulation.

3.1 Pricing a European Put Option

For European options, the numerical results are benchmarked against the exact analytical Black-Scholes formula and a stochastic Monte Carlo simulation (10^5 paths).

Table 3: European Put Option Pricing and Runtime Comparison at $S_0 = 100$

Method	**Option Price**	**Abs. Error**	**Runtime (s)**
Black-Scholes (Exact)	5.5735	–	–
Explicit FD	**5.5778**	**4.31x10^{-3}**	**0.2977**
Implicit FD	5.5769	3.38x10^{-3}	0.5558
Crank-Nicolson FD	5.5774	3.84x10^{-3}	0.6079
Monte Carlo	5.5749	1.40x10^{-3}	0.5601

Our comparative study demonstrates that for a single underlying asset, the **Explicit Finite Difference Scheme is the most computationally efficient method**. As detailed in Table 3, it is nearly twice as fast as the Monte Carlo simulation while maintaining an absolute error on the order of 10^{-3}.

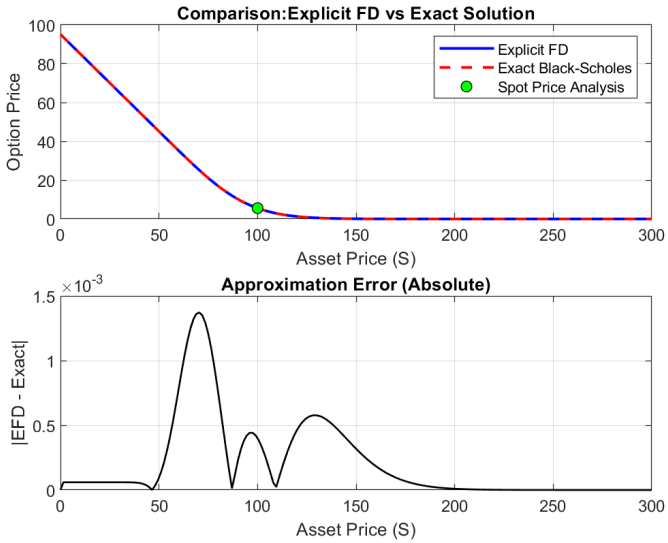


Figure 1: Comparison between Explicit FD and Exact Black-Scholes solution (Top) and Absolute Approximation Error (Bottom).

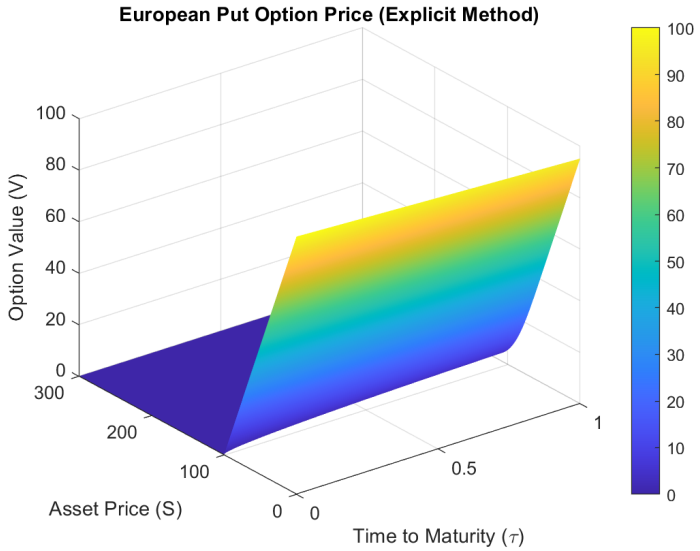


Figure 2: Option Value Surface $V(S, \tau)$ calculated using the Explicit Finite Difference Method.

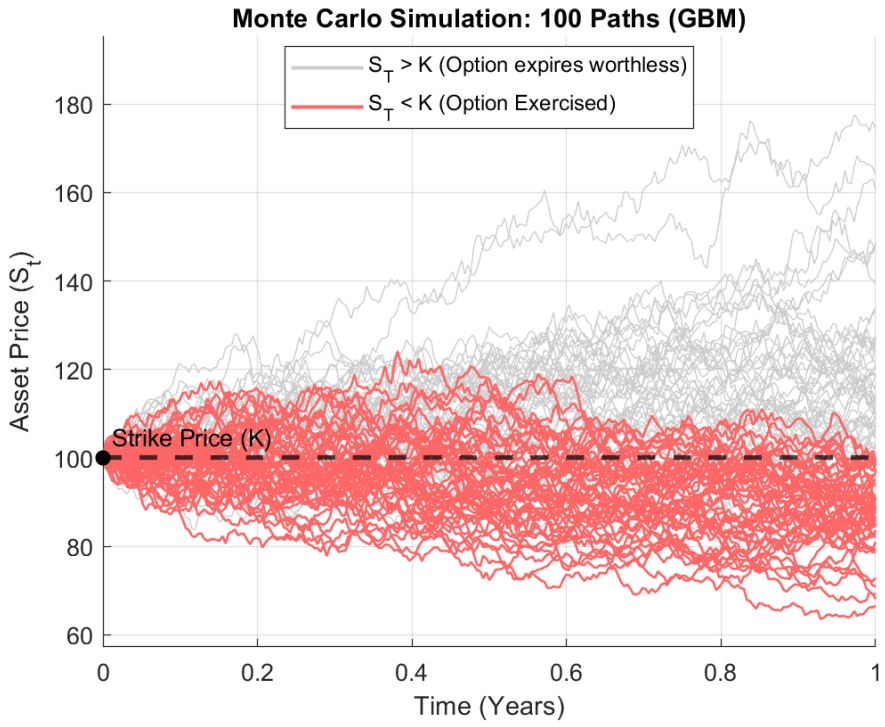


Figure 3: Monte Carlo Simulation: 100 Paths showing exercised (red) vs. worthless (grey) outcomes.

3.2 Pricing an American Put Option

Unlike European options, American options can be exercised at any time, transforming the PDE into a free-boundary linear complementarity problem. The solution must satisfy the Early Exercise

Constraint: $V(S,\tau) \geq \max(K-S,0)$. Because no analytical closed-form solution exists, numerical methods are strictly required. For the stochastic benchmark, we utilize the Longstaff-Schwartz (LSM) algorithm.

Table 4: American Put Option Pricing and Runtime Comparison at $S_0 = 100$

Method	Option Price	Runtime (s)
Explicit FD	**6.0933**	**0.2152**
Implicit FD	6.0914	0.5148
Crank-Nicolson FD	6.0923	0.6233
Longstaff-Schwartz	5.9940	0.6856

The results verify the financial principle that early exercise flexibility adds value. Comparing the Explicit FD results, the American Put is priced at **6.0933**, yielding an **Early Exercise Premium** of **0.5198** over the European price (5.5735).

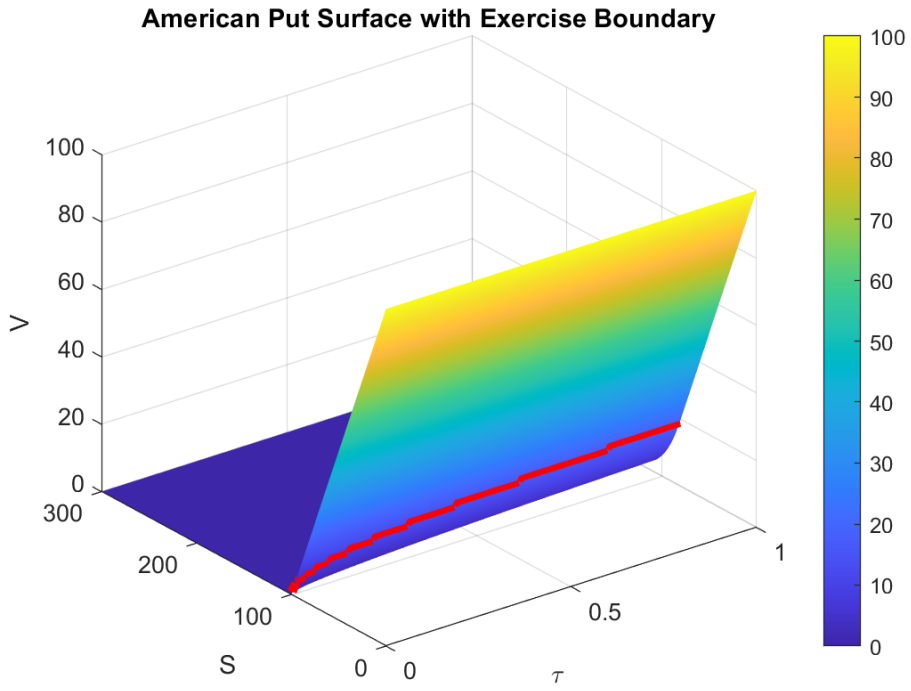


Figure 4: American Put Price Surface with the Optimal Exercise Boundary.

Figure 4 visually captures this free-boundary dynamic. The red line represents the **Optimal Early Exercise Boundary** $S^*(\tau)$. To its right ($S \leq S^*$), the surface forms a steeply inclined flat plane where $V = K - S$; time value has vanished, making immediate exercise optimal. To its left ($S > S^*$), the surface curves strictly above the intrinsic payoff ($V > K - S$), indicating positive time value and forming the continuation region where the investor should hold the asset.

4 Conclusions

This work has provided a comprehensive exploration of numerical methods for pricing financial derivatives. The main findings are summarized as follows:

- **Numerical Dominance in Low Dimensions ($d = 1$):** For single-asset problems, the Explicit Finite Difference scheme proves to be computationally dominant, providing deterministic accuracy and superior speed (0.21s) compared to stochastic simulations, provided the CFL stability condition is met.
- **Efficiency in American Options:** Finite Difference Methods demonstrated a significant advantage over Monte Carlo for pricing American options. By operating backwards in time, the early exercise constraint (Linear Complementarity Problem) is integrated naturally via Operator Splitting, capturing the early exercise premium accurately.

Acknowledgments

We wish to extend our sincere gratitude to our esteemed teachers, Juan Ramón Torregrosa and Alicia Cordero. Their invaluable guidance and assistance were key to resolve numerous doubts and questions that arose throughout the duration of this project. Furthermore, they imparted comprehensive explanations of the underlying theory. The privilege of presenting this paper at the congress is a direct result of the opportunity they afforded us to participate in this esteemed event.

References

[1] Black, F. & Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*.

[2] Merton, R. C. (1973). *Theory of Rational Option Pricing*. The Bell Journal of Economics and Management Science.

[3] Hull, J. C. (2023). *Options, Futures, and Other Derivatives* (11th ed.). Pearson.

[4] Wilmott, P., Howison, S., & Dewynne, J. (1995). *The Mathematics of Financial Derivatives: A Student Introduction*. Cambridge University Press.

[5] Shreve, S. E. (2004). *Stochastic Calculus for Finance II: Continuous-Time Models*. Springer.

[6] Morton, K. W., & Mayers, D. F. (2005). *Numerical Solution of Partial Differential Equations: An Introduction*. Cambridge University Press.

[7] Duffy, D. J. (2006). *Finite Difference Methods in Financial Engineering: A Partial Differential Equation Approach*. John Wiley & Sons.

[8] Hundsdorfer, W., & Verwer, J. G. (2003). *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*. Springer Science & Business Media.

[9] Brennan, M. J., & Schwartz, E. S. (1977). *The Valuation of American Put Options*. The Journal of Finance, 32(2), 449–462.

[10] Han, H., & Wu, X. (2003). A Fast Numerical Method for the Black-Scholes Equation of American Options. *SIAM Journal on Numerical Analysis*, 41(6), 2081-2095.

A Reduced-Order Digital-Twin Framework for MOST-Assisted Thermal Buffering in Biological CubeSat Payloads

E. Vega-Fleitas[1], S. Moll-López[2],[1] J. A. Moraño-Fernández[3], A. Herrero-Debón[3] and O. Portero-Sáez[4]

[1] Instituto de Diseño y Fabricación, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.

[2] Departamento de Matemática Aplicada, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.

[3] Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.

[4] Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.

1 Introduction

Biological CubeSat payloads require a stable thermal microenvironment because orbital thermal cycling can alter sample viability, metabolic activity, stress-response dynamics and experimental reproducibility. Space-biology CubeSats have therefore incorporated dedicated payload thermal-control hardware, particularly in missions involving microorganisms, yeast cultures, cells or small biological systems [1], [2], [3]. In Low Earth Orbit (LEO), repeated eclipse periods impose a recurrent cold-phase demand on the thermal-control system. Conventional resistive heaters are robust, but they consume scarce spacecraft power, especially in small platforms where payload operation, communication and thermal regulation share the same limited energy budget [4], [5].

The aim of this contribution is to introduce a requirements-driven digital-twin framework for assessing whether a Molecular Solar Thermal (MOST) material can act as a semi-passive thermal-assist branch for biological CubeSat payloads. MOST systems store photon energy through reversible molecular photoisomerisation and can later release heat through thermal or catalytic triggering [6], [7], [8]. In the proposed architecture, the MOST module is not treated as a replacement for electric heating. Instead, it is modelled as a solar-charged auxiliary heat reservoir whose release is constrained by biological safety requirements.

The central control and modelling question is whether stored solar heat can reduce electric-heater demand without increasing biologically meaningful cold exposure. To answer it, we combine orbital forcing, a lumped thermal network, MOST charge–release kinetics and biological thermal-stability metrics in a single reduced-order digital twin.

[1] sanmollp@mat.upv.es

2 Methods

The methodological core is a reduced-order thermal model of the coupled spacecraft–payload system. It considers a representative 3U CubeSat in a circular 500 km LEO. The thermal environment includes direct solar input, albedo, Earth infrared radiation and radiative rejection to deep space. Eclipse is represented by a binary sunlight indicator, $s(t)$, with $s(t) = 1$ during sunlit arcs and $s(t) = 0$ during eclipse. This allows the model to resolve repeated sunlit and cold orbital phases. The payload is represented by a five-node resistive–capacitive thermal network composed of the biological chamber, spacecraft structure, radiator, MOST module and electronics/battery node. In compact form, the nodal energy balance can be written as

$$C_i \dot{T}_i = \sum_j K_{ij} \left(T_j - T_i\right) + Q_i(t) + u_i(t), \tag{1}$$

where i and j index the five thermal nodes, T_i is the temperature of node i, C_i is its thermal capacitance, K_{ij} is the thermal conductance between nodes i and j, $Q_i(t)$ contains environmental and internal heat terms, and $u_i(t)$ represents controlled heater or MOST heat input.

The MOST subsystem is described by a state of charge $x(t) \in [0, 1]$, with total stored energy $E_{\mathrm{MOST}} = x m_{\mathrm{act}} e_{\mathrm{MOST}}$, where m_{act} is the active MOST mass and e_{MOST} is the specific stored energy of the material. Its dynamics combine optical charging during sunlit arcs, temperature-dependent self-discharge and controlled catalytic release:

$$\dot{x} = \frac{\eta_c P_{\mathrm{opt}} s(t)}{m_{\mathrm{act}} e_{\mathrm{MOST}}} - k_{\mathrm{sd}}(T_m) x - k_{\mathrm{rel}}(T_m, u_d) x, \tag{2}$$

where η_c is the charging efficiency, P_{opt} is the available optical charging power, T_m is the MOST-module temperature, $k_{\mathrm{sd}}(T_m)$ is the temperature-dependent self-discharge rate, $k_{\mathrm{rel}}(T_m, u_d)$ is the controlled release rate, and u_d is the discharge command.

Three families of controllers are considered. The baseline controller activates the electric heater when the biological chamber temperature $T_{\mathrm{bio}}(t)$ falls below the lower tolerance boundary $T_{\mathrm{low}} = T_{\mathrm{target}} - \Delta T$, where T_{target} is the target biological temperature and ΔT is the half-width of the admissible tolerance band. A MOST-priority controller uses stored molecular heat preferentially and activates the heater only when the MOST reserve is depleted. Finally, a safety-constrained MOST controller discharges MOST below T_{low} but activates the heater as an override whenever the chamber approaches a prescribed safety threshold. This last controller is the biologically conservative option.

Because biological relevance depends on both the magnitude and duration of thermal excursions, the main severity metric is cumulative cold exposure:

$$A_{\mathrm{cold}} = \int_0^{t_{\mathrm{sim}}} \max\left(0, T_{\mathrm{low}} - T_{\mathrm{bio}}(t)\right)\, dt, \tag{3}$$

reported in K h, where t_{sim} is the total simulation time. This metric complements heater energy and the fraction of time outside the tolerance band, which can overestimate risk when deviations are only sub-Kelvin threshold crossings caused by bang-bang switching.

3 Results

The nominal case corresponds to a microbial payload with $T_{\mathrm{target}} = 25\,^{\circ}\mathrm{C}$ and a ± 2 K tolerance band over ten consecutive LEO orbits. Table 1 reports a restricted subset of the simulation results. The heater-only baseline consumes 36.24 Wh and shows a very small cumulative cold exposure of 0.086 K h. The strict safety-constrained MOST controller reduces heater demand to 34.26 Wh,

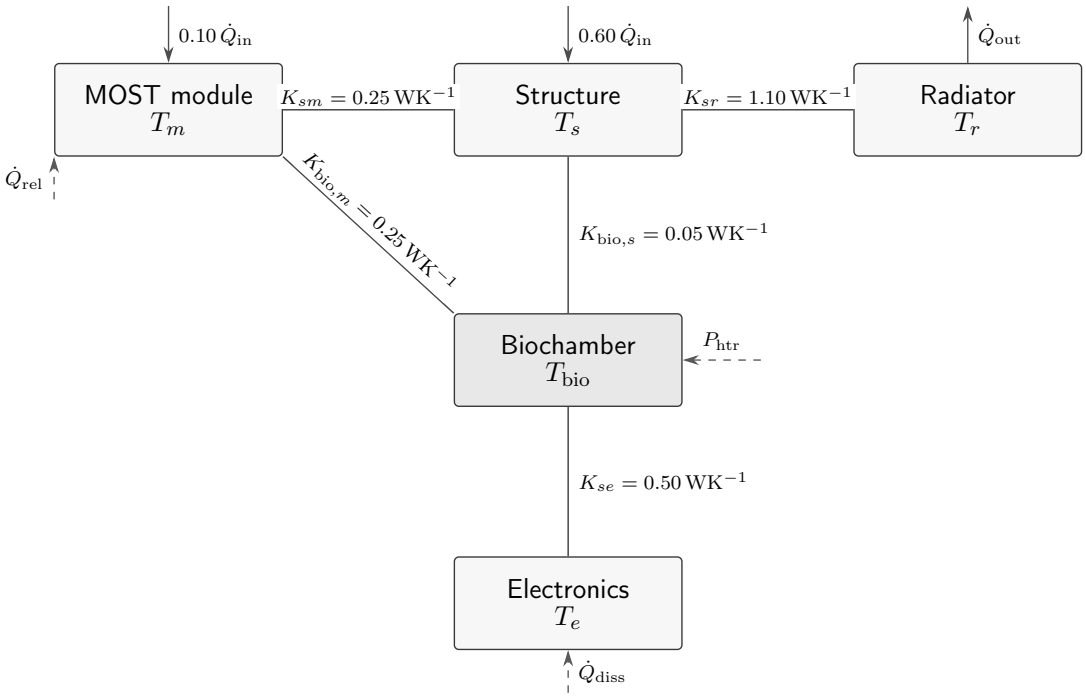


Figure 1: Reduced-order thermal network used in the digital-twin framework. The biological chamber is the controlled node, while the MOST module provides a solar-charged auxiliary heat branch.

corresponding to a 5.5% saving, while keeping cumulative cold exposure at the same practical level as the baseline. In contrast, the MOST-priority strategy produces larger apparent energy savings but allows a strong cold excursion, making it unsuitable for biological payload operation.

Table 1: Representative nominal results over ten LEO orbits. Only the minimum set of metrics needed to communicate the modelling insight is reported here; extended design-space and uncertainty analyses are deliberately left outside this conference abstract.

Controller	E_{htr} [Wh]	A_{cold} [K h]	$T_{\mathrm{bio,min}}$ [°C]	Saving [%]
Heater-only baseline	36.24	0.086	22.97	–
MOST-safety, strict override	34.26	0.077	22.97	5.5
MOST-priority	29.69	46.34	15.36	18.1

The main modelling result is thus not that MOST can replace conventional thermal control. Rather, the framework shows that MOST can be useful only when embedded in a safety-constrained hybrid controller. In this regime, stored solar heat slightly reduces heater duty cycle while the electric heater remains responsible for guaranteeing the biological temperature envelope. Relaxing the safety constraint increases energy savings but rapidly increases cumulative cold exposure, indicating a steep energy–stability trade-off.

4 Conclusions

This research introduces a reduced-order digital-twin framework for early-stage assessment of MOST-assisted thermal buffering in biological CubeSat payloads. The framework links orbital thermal

forcing, a five-node lumped thermal model, MOST charge–release kinetics and biological cold-exposure metrics. Representative simulations suggest that a strict safety-constrained controller can provide modest heater-energy savings without increasing biologically meaningful cold exposure, whereas a MOST-priority strategy is unsafe despite its larger energy saving.

The present contribution focuses on the modelling architecture and on a representative nominal comparison between heater-only thermal control and MOST-assisted thermal buffering. Emphasis is placed on how the reduced-order digital-twin framework links orbital environmental forcing, auxiliary heat storage and biological payload stability in a compact form suitable for early-stage design assessment.

Acknowledgments

This work was developed as part of a Final Degree Project in the Bachelor's Degree in Aerospace Engineering at the School of Aerospace Engineering and Industrial Design (ETSIADI), Universitat Politècnica de València (UPV), Spain.

References

[1] S. F. Harandi, E. Baniasadi, and A. Tajik, "Review of biological CubeSat missions and onboard life-support systems for micro-organisms," *Life Sciences in Space Research*, vol. 35, pp. 24–36, 2022. DOI: 10.1016/j.lssr.2022.07.001.

[2] M. R. Padgen, M. P. Lera, A. J. Ricco, and S. Bhattacharya, "The EcAMSat fluid system to study antibiotic resistance in microgravity: development and lessons learned," *Acta Astronautica*, vol. 173, pp. 449–459, 2020. DOI: 10.1016/j.actaastro.2020.04.009.

[3] S. Bhattacharya, P. Bhatt, D. Bhattacharya, J. Gim, et al., "BioSentinel: A Nanosatellite for Deep-Space Biological Experiments Beyond Low Earth Orbit," *Astrobiology*, vol. 23, no. 6, pp. 647–663, 2023. DOI: 10.1089/ast.2022.0043.

[4] D. G. Gilmore, *Spacecraft Thermal Control Handbook*, 2nd. El Segundo, CA, USA: The Aerospace Press, 2002, vol. 1.

[5] European Cooperation for Space Standardization, *ECSS-E-HB-31-03A: Thermal Design Handbook*, ECSS, Noordwijk, The Netherlands, 2016.

[6] K. Moth-Poulsen et al., "Molecular solar thermal energy storage and release system," *Energy & Environmental Science*, vol. 5, pp. 8534–8537, 2012. DOI: 10.1039/C2EE22426G.

[7] J. Orrego-Hernández, A. Dreos, and K. Moth-Poulsen, "Engineering of norbornadiene/quadricyclane photoswitches for molecular solar thermal energy storage applications," *Accounts of Chemical Research*, vol. 53, pp. 1478–1487, 2020. DOI: 10.1021/acs.accounts.0c00235.

[8] Z. Wang, R. Losantos, N. Mora-Diez, and K. Moth-Poulsen, "Macroscopic heat release in a molecular solar thermal energy storage system," *Energy & Environmental Science*, vol. 12, pp. 187–193, 2019. DOI: 10.1039/C8EE01011K.