IntechOpen

Nonlinear Systems and Matrix Analysis

Recent Advances in Theory and Applications

Edited by Peter Y.P. Chen and Victor Martinez-Luaces

Nonlinear Systems and Matrix Analysis - Recent Advances in Theory and Applications

Edited by Peter Y.P. Chen and Victor Martinez-Luaces

Published in London, United Kingdom

Nonlinear Systems and Matrix Analysis - Recent Advances in Theory and Applications
http://dx.doi.org/10.5772/intechopen.1001634
Edited by Peter Y. P. Chen and Victor Martinez-Luaces

Contributors
Alexander A. Huang, Alice Eraud, Alvaro Humberto Salas, Ana I. Julio, Anderson Pablo Freitas Evangelista, Armando Martínez-Pérez, Bruno Carpentieri, Catherine Bruneau, F. Setoudeh, Gabino Torres-Vega, Ginalber Luiz de Oliveira Serra, Iuliana Matei, L. Gerard Van Willigenburg, M. M. Dezhdar, Mudassir Shams, Najmadeen Saeed, Peter Y. P. Chen, Ricardo L. Soto, Samuel Y. Huang, Sergio Callegari, Shna Abdulkarim, Victor Martinez-Luaces

© The Editor(s) and the Author(s) 2024

The rights of the editor(s) and the author(s) have been asserted in accordance with the Copyright, Designs and Patents Act 1988. All rights to the book as a whole are reserved by INTECHOPEN LIMITED. The book as a whole (compilation) cannot be reproduced, distributed or used for commercial or non-commercial purposes without INTECHOPEN LIMITED's written permission. Enquiries concerning the use of the book should be directed to INTECHOPEN LIMITED rights and permissions department (permissions@intechopen.com).

Violations are liable to prosecution under the governing Copyright Law.

CC BY

Individual chapters of this publication are distributed under the terms of the Creative Commons Attribution 3.0 Unported License which permits commercial use, distribution and reproduction of the individual chapters, provided the original author(s) and source publication are appropriately acknowledged. If so indicated, certain images may not be included under the Creative Commons license. In such cases users will need to obtain permission from the license holder to reproduce the material. More details and guidelines concerning content reuse and adaptation can be found at http://www.intechopen.com/copyright-policy.html.

Notice

Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in London, United Kingdom, 2024 by IntechOpen
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, 167-169 Great Portland Street, London, W1W 5PF, United Kingdom

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Additional hard and PDF copies can be obtained from orders@intechopen.com

Nonlinear Systems and Matrix Analysis - Recent Advances in Theory and Applications
Edited by Peter Y. P. Chen and Victor Martinez-Luaces
p. cm.
Print ISBN 978-1-83769-449-5
Online ISBN 978-1-83769-448-8
eBook (PDF) ISBN 978-1-83769-450-1

We are IntechOpen, the world's leading publisher of Open Access books

Built by scientists, for scientists

7,300+
Open access books available

192,000+
International authors and editors

210M+
Downloads

156
Countries delivered to

Our authors are among the

Top 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Meet the editors

Dr. Peter Chen was born in China, migrated to Australia, and received his tertiary education at the University of New South Wales, Sydney. Presently, he has retired and continues as an independent researcher. After his BSc, MEngSc, and ME degrees, he obtained his Ph.D. at the same university in 1974. He was a senior research scientist at the School of Mechanical and Manufacturing Engineering UNSW. His responsibility was to provide academic advice to research students and faculty members. His own research interests include many topics in mechanical engineering and optical fibers. He is specialized in solving nonlinear problems in those fields. His most recent research is in the electromagnetic wave propagation theory for cosmic redshifts.

Dr. Martinez-Luaces is a researcher in mathematics, chemistry, engineering, and education. He obtained three degrees from the State University of Uruguay (UdelaR) in chemistry, chemical engineering, and mathematics. Later, he obtained a master's degree and a Ph.D. in mathematics education from the University of Granada (UGR), Spain. He has worked for 25 years in four separate faculties within UdelaR (Chemistry, Economics, Engineering, and Sciences), and he held the position of head of the Mathematics Department in the Chemistry Faculty (1996–2002). At present, he is a researcher of the ProfeSTEAM Project at UGR, Spain. He has 24 books or book chapters published in the U.S.A., Europe, Argentina, Singapore, and Uruguay.

Contents

Preface

Research on nonlinear analysis is not restricted to theoretical development. Judging from many publications in a widely diverged field of international journals, more and more attention has been centered on the practical application of nonlinear system analysis in a large variety of disciplines. The successes are due largely to advances in mathematical modeling and simulation and the development of solution methods, including general-purpose computer packages that help to solve some complicated computational requirements.

The contributing chapters of Section 1 are reports of recent advances in theory and applications. Chapter 1 is an introduction to nonlinear system analysis. Chapter 2 is a review of nonlinear control strategies for shapes in structural engineering. Chapter 3 presents an incremental learning method with noise handling for nonstationary dynamic system identification. Chapter 4 describes how to bring data converter pairs into chaotic oscillation for built-in self-test and entropy generation. Chapter 5 considers nonlinear systems involving behavioral analysis and noise effect. Chapter 6 explores the nonlinear relationship between economics and its driving causes. Chapter 7 introduces simplification to properties of nonlinear and dynamical structures and shows how controllability and observability can be computed efficiently. Analytical solution techniques are the topics considered in Chapter 8, while Chapters 9 and 10 employ numerical methods to solve nonlinear systems. Chapter 11 tries to enhance regression analysis with splines and machine learning to capture complex nonlinear multidimensional variables.

There is no limit on how a system can be modeled. But there are some undeniable simulation rules, including (1) assumptions and hypotheses used must not violate any established principles; (2) a simpler model, including linear simulations, is preferred, providing it meets all expectations; (3) reliable observations are superior to computed data; and (4) there are limits on the reliability of the models. Various numerical examples given in each chapter in this section provide illustrations of how those rules could be applied.

Solution methods also play an important role in nonlinear system analysis. There is a current trend to find analytical solutions by treating the system as an inverse problem. In this approach, the solutions are assumed to be a combination of selected analytical functions together with a small number of system parameters. Working backward to satisfy the governing nonlinear equations, these parameters must satisfy a specific number of conditions. Cases of this inverse approach include the inverse differential and integral methods, as well as inverse scattering methods, together with many variants. While this analytical approach could be used to study system characteristics, its practical application is limited as the system must be designed based on the preimposed conditions that may not be feasible in a physical system.

Numerical solutions have the potential to solve any mathematical problems embedded in a nonlinear system. The first step involved is to reduce the mathematical

model to a set of nonlinear matrix equations that could then be solved by an iterative algorithm using linear matrix operations. By choosing the appropriate boundary and initial conditions, and, if needed, extra terms, the same set of equations could be solved to cover a wide range of practical applications.

While there are extensive advances in the theory of nonlinear system analysis, chapters of this section represent only a small number of recent achievements. Formulating a better model and overcoming some inherent limitations remain challenges for system analysts. Not only in sciences and technologies but also in humanities, there is unlimited scope for further research and development of nonlinear system analysis.

The second section of this book is devoted to matrix theory. Indeed, matrix theory applications are present in other branches of mathematics and in the experimental sciences, engineering, and technology.

The importance of matrices, their historical origin, and several of their applications are described in the introductory chapter of this section (Chapter 12). After that, in Chapter 13, the eigenvalues and eigenvectors of matrices that appear in chemical kinetic problems are analyzed to determine the shape of the concentration curves and predict their qualitative behavior, emphasizing the stability of the solutions. Chapter 14 analyzes pairs of matrices that give rise to a diagonal commutator when applied to a given, arbitrary vector, which has an interesting connection with the discrete approximations of derivatives and integrals of a function. Finally, in Chapter 15 an interesting inverse problem is studied—the realizability problem—which consists of determining whether, for a given list of complex numbers, it is possible to find a nonnegative matrix whose spectrum coincides with that list.

The results of the previous chapters are relevant to areas as diverse as chemical kinetics, quantum mechanics, and matrix algebra and show only a small part of the many contributions of matrix theory to scientific knowledge.

Peter Y. P. Chen
Former School of Mechanical and Manufacturing Engineering,
University of New South Wales,
Sydney, NSW, Australia

Victor Martinez-Luaces
ProfeSTEAM Project,
University of Granada,
Granada, Spain

Section 1

Nonlinear System

Chapter 1

Introductory Chapter: Nonlinear System Analysis – An Overview of Historical and Recent Advances

Peter Y.P. Chen

1. Introduction

The wide spectrum of topics in nonlinear system analysis are including but not limited to the following:

- Nonlinear operator theory
- Multi-functional problems
- Approximation theory
- Stability of functional equations
- Fractional calculus
- Abstract metric space
- Nonlinear modeling and simulation
- Chaos theory, noise, and complex dynamics
- Nonlinear control and stability analysis

The theories of nonlinear system analysis have been used in a large variety of disciplines including not only in applied sciences, engineering, and technologies but also in social sciences, economic, environmental science, and other non-science but related studies. The successes are due largely to advances in mathematical modeling and simulation, and the development of solution methods, including general-purpose computer packages that help to solve some complicated computational requirements.

2. A historical overview of nonlinear system analysis

Historically, nonlinear system analysis has evolved from linear system analysis over the later decades of the last century. In linear analysis, a set of linear ordinary

differential equations is used to describe a system. As linear calculus had already been well developed, it is often possible to find solution in closed-form expressions. Although all physical systems are virtually nonlinear in nature, the linear approach has taken full advantage of the fact that over a limited range of the independent variables in such a nonlinear system could be approximated by a linear one. Finite differences or finite elements are used to replace the differential operators in the equations used to simulate the system. Although the resultant set of matrix equations could be solved by standard classical linear algebra, the computer capacity and operating speed available in the earlier days had imposed a practical limit on the size of the matrix. As the responses of the system are linear to the independent variables, solutions for a problem could be obtained as a series of orthogonal functions. Therefore, in general, with the increases in computer power now available, solutions for a linear system no longer impose a limit on its applications.

Even with those inherent limitations due to the linearization approximation, linear system analysis and its applications have been used widely in different fields and for many practical problems. Over the time before 2000, there were already numerous examples of applications such as in rotor dynamics [1], that led to the design and control of turbines for jetliners. Through stress analysis, prestressed concrete was used to change the design and operation of all civil structures completely. In fact, linear system analysis had occupied an important role in post-World War II's construction, manufacturing, and technological developments. Undoubtedly, the single most important influence of those advances is the ability to "crunch" numbers at an exponential rate with later generations of computers. At the same time, commercial computer packages became available such as MATLAB for finite element coding, and FLOW3D for fluid flow. Those packages provide valuable relief to programming needs.

Nonlinear system analysis is not just a natural or logical progression from the linear one. There are systems, both natural and conceptual, that can only be represented by nonlinear equations. The propagation of light waves through an optical fiber is a well-known example. It is through theoretical studies of the solutions for nonlinear Schrödinger equation that the capacity of optical communication systems has been developed from a few hundred in the early times to potentially 20,000 or more telephone calls per fiber currently. Modern manufacturing often involves heating and nonlinear material properties that can be simulated only by nonlinear mathematical equations.

System analysis, both linear and nonlinear, uses the scientific method to identify goals or questions, form hypotheses and/or mathematical simulations, conduct experiments or applications, and analyze data. Historically, the beginning of using the scientific method could be traced back to the seventeenth century. But it was not until the second half of the last century and following the success of its applications in sciences and technologies that the scientific method has been adopted to most of the non-science disciplines. Accompanied by these changes, there are often some subtle modifications to system analysis. Science is objective based, while non-science could be subjective based. For example, based on the same set of economic data, opposing political parties could use their financial system analysis to reach completely different conclusions. The reason behind this outcome is due to the different values or weights putting, subjectively, on the goals and achievements of the models. Notwithstanding these limitations, economic models are needed for the purposes of setting up financial policies, planning administrative mechanisms, and implementing control procedures. However, to achieve the best outcome, both linear and nonlinear system analysis

employs an adaptive predictive approach. Historically, advances were made whenever beneficial changes or new creations in system analysis were deemed necessary.

3. Recent advances in mathematical simulations

Since the beginning of the twenty-first century, it was found that mathematical simulation of many controls and stability analysis for real-world problems need to use fractional operators [2]. Fractional calculus was proposed by mathematicians as early as 1695 [3]. But it is in recent years, fractional calculus began to attract a large attention in research popularity and applications. Fractional calculus consists of fractional differentiation and fractional integration. However, the precise mathematical implications for these terms need to be defined [3] if they are to be used in system simulation.

Fractional models are essential in some viscoelastic flow systems both for control and stability analysis. The fractional acoustic wave equation was found to agree better with the experimental results. A fractional system analysis could be used for control, signal and image processing, mechanics and dynamic systems, biology, environmental science, material science, economics, and multidisciplinary in engineering fields [2].

4. Recent advances in solution methods

The success of nonlinear systems together with linear systems analysis is due to their ability to be applied to a wide spectrum of practical problems. However, it is important that solution methods are available for solving complex set of mathematical equations. Recent advances in solution methods are aiming at more effective ways of solving dynamic systems, stochastic systems with random noise, and stability analysis. Although well-proven methods for all these systems are available [4], researchers are looking for efficiency and reliability.

Since the beginning of the twenty-first century, analytical solutions for a nonlinear system have attracted a great deal of attention from researchers. But, because of the nonlinear nature, little success has been achieved in solving them directly. That is, starting from the nonlinear system itself and finding the solutions analytically in a forward direction. However, for methods starting from some assumed solutions and working out how to satisfy the nonlinear system analytically as an inverse problem, there are many successes [5]. Some of these inverse examples include the inverse differential and integral methods such as those for wave propagation [6], and the G'/G expansion method [7]. Generally, as an inverse problem, there is no limit to how many solutions can be found because there are infinite number of choices for the set of system parameters that define the chosen starting solutions. But the need to have a matching background medium is a notable limitation [8]. From the nature of those solutions, it could be concluded that this inverse approach is more suitable for qualitative analysis that the performance of a design, or the characteristics of a system could be assessed qualitatively.

Far more numerical methods have been used in linear system analysis [9]. For a nonlinear system, the nonlinear terms could be replaced by linearization approximations and solved linearly. An iterative scheme is then used to ensure that the solutions have converged. Among all the known numerical methods, the collocation method that uses Chebyshev polynomials is the most economical due to its minimax property.

However, for computational economics, the Lanczos-Chebyshev collocation method [10, 11] is superior because ordinary power series is used. It is worthwhile to note that over the last 50 years, many papers and conference proceedings have been published on all those numerical methods.

5. Concluding remarks

It is obvious that past advances in nonlinear system analysis have widened the practical applications of linear system analysis and provided added means to develop new materials, new designs, and new control mechanisms. It is important that researchers and analysts follow the traditional scientific methods and validate their models so that those advances would not be misused.

Author details

Peter Y.P. Chen
Former School of Mechanical and Manufacturing Engineering, University of New South Wales, Sydney, NSW, Australia

*Address all correspondence to: peterypchen@yahoo.com.au

IntechOpen

© 2024 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

[1] Chen PYP et al. Recent development in turbomachinery modelling improved balancing and vibration response analysis. Journal of Engineering for Gas Turbines and Power. 2005;**127**:646-653. DOI: 10.1115/1.1850942

[2] Sun HG et al. A new collection of real world applications of fractional calculus in science and engineering. Communications in Nonlinear Science and Numerical Simulation. 2018;**64**:213-231. DOI: 10.1016/j.cnsns.2018.04.019

[3] Baleanu D, Fernandez A. On fractional operators and their classifications. Mathematics. 2019;**7**:830-839

[4] Sastry S. Nonlinear Systems: Analysis, Stability, and Control. New York: Springer-Verlag; 1999. DOI: 10.1007/978-1-4757-3108-8. ISBN 978-1-4419-3132-0 ISBN 978-1-4757-3108-8 (eBook)

[5] Abdelrahman MAE, Sohaly MA, Alharbi A. The new exact solutions for the deterministic and stochastic (2+1)-dimensional equations in natural sciences. Journal of Taibah University for Science. 2019;**13**(1):834-843. DOI: 10.1080/16583655.2019.1644832

[6] Yagle AE. Differential and integral methods for three-dimensional inverse scattering problems with a non-local potential. Inverse Problems. 1988;**4**(2):549-566. DOI: 10.1088/0266-5611/4/2/017

[7] Wang M, Li X, Zhang J. The (G'/G)-expansion method and traveling wave solutions of nonlinear evolution equations in mathematical physics. Physics Letters A. 2008;**372**(4):417-423. DOI: 10.1016/j.physleta.2007.07.051

[8] Afsari A, Abbosh A, Rahmat-Samii Y. A novel differential inverse scattering methodology in biomedical imaging. In: 2017 IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting. 2017. DOI: 10.1109/APUSNCURSINRSM.2017.8072055

[9] Hammad DA, Semary MS, Khattab AG. Parametric quintic spline for time fractional Burger's and coupled burgers' equations. Fixed point theory algorithms. The Sciences and Engineering. 2023;**2023**:9. DOI: 10.1186/s13663-023-00740-3

[10] Chen PYP. Linear and non-linear transient heat conduction problems by the Lanczos-Chebyshev method. Nuclear Engineering and Design. 1981;**64**(2):225-232. DOI: 10.1016/0029-5493(81)90006-6

[11] Chen PYP, Malomed BA. Lanczos–Chebyshev pseudospectral methods for wave-propagation problems. Mathematics and Computers in Simulation. 2012;**82**:1056-1068. DOI: 10.1016/j.matcom.2011.05.013

Chapter 2

A Review of Nonlinear Control Strategies for Shape and Stress in Structural Engineering

Najmadeen Saeed and Shna Abdulkarim

Abstract

Structural engineering plays a pivotal role in ensuring the safety, stability, and longevity of civil infrastructure. As the demand for innovative and efficient structural designs grows, the need for advanced control strategies becomes increasingly apparent. This comprehensive review examines the state-of-the-art nonlinear control strategies for shape and stress in structural engineering. Recognizing the limitations of conventional linear approaches, the chapter systematically explores diverse methodologies such as adaptive control, neural networks, fuzzy logic, and model predictive control. It analyzes their individual and integrated applications in shaping structural form and managing stress levels. The review considers the intricate interplay between shape and stress control strategies, addresses challenges, and proposes future research directions. Case studies and a comparative analysis offer practical insights into the performance and adaptability of these strategies. By emphasizing advances in materials, technologies, and sustainability, this chapter provides a holistic perspective on the evolving landscape of nonlinear control in structural engineering. This synthesis aims to guide researchers and practitioners toward innovative solutions that enhance the safety, resilience, and efficiency of structural systems.

Keywords: nonlinear control, structural engineering, shape strategies, stress management, adaptive control, sustainability

1. Introduction

At the nexus of innovation and resilience, structural engineering pursues the continuous development of structures that maximize longevity and performance while also withstanding external stresses. In this pursuit, the increasing understanding of the innate nonlinearities in structural systems is reshaping the traditional paradigm of linear control techniques. This in-depth study, "A Review of Nonlinear Control Strategies for Shape and Stress in Structural Engineering," looks at the newest developments in using nonlinear control methods to deal with shape and stress, two important parts of structural design.

Traditional linear control schemes [1–16], although useful, are not always able to capture the complex behaviors that are inherent in structural systems [17]. In structural engineering, nonlinearities can originate from a number of factors, including

geometric configurations, material properties, large deflections or rotations, and dynamic loading scenarios [18–21]. These nonlinear phenomena frequently make it difficult to precisely control shape and manage tension in structures [22]. However, linear control models depend on assumptions and simplifications, for instance, small deformations and elastic material behavior, that might not be factual for all structural states, potentially resulting in imprecisions in performance optimization. As a result of these drawbacks, scientists are now more frequently using nonlinear control techniques to manage the complexity of structural behavior [23, 24]. Employing the nonlinear controlling strategy improves structural efficiency and tolerates the construction of a more competent and lighter system due to decreasing material consumption while sustaining essential safety boundaries. Besides, nonlinear control techniques offer more flexibility in monitoring and adjusting performance in structures exhibiting nonlinear behavior, leading to more accurate modeling of actual structural responses.

1.1 Most important nonlinear control approaches

The field of nonlinear control strategies comprises a wide range of techniques, each with specific benefits for controlling stress [25–27], forming structural shapes [16, 28–30], or both simultaneously [30, 31]. Adaptive control has demonstrated promise for improving adaptability because of its capacity to modify parameters in response to changing structural conditions [32, 33]. Fuzzy logic offers a strong framework for forming structural configurations because of its ability to deal with uncertainty and imprecision [34]. Additionally, stress management in structural systems is greatly aided by model predictive control, which is well known for its capacity to maximize performance based on predictive models [35–37]. Korkmaz [38] divided the concept of structural control into three subdomains: active control, adaptive control, and intelligent control (see **Figure 1**). Active structural control utilizes sensors and actuators to alter the deformability and internal stress by modifying the structural response. In adaptive structural control, the alteration process improves the structural response regardless of the previous condition of loadings and actions. In intelligent structural control, on the

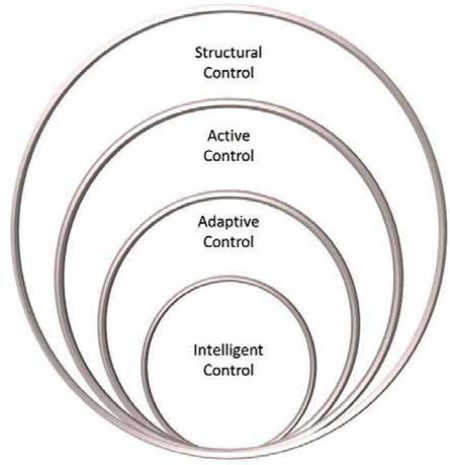


Figure 1.
Structural control and controlling subdomains [38].

other hand, the controlling process ensures the preservation and improvement of the structural performance by remembering the changes in behavior and action, adapting to the current target, and using the earlier events for improvement in future responses.

1.2 Scope of the review

To clarify these nonlinear control techniques' separate and combined uses in regulating stress levels and forming structural shapes, this review methodically investigates them. The subsequent sections will explore the subtleties of nonlinearities in structural engineering in Section 2. It is followed by shape control strategies in Section 3 and stress management tactics in Section 4, as well as the complex interactions between shape and stress in Section 5. Section 6 will address the field's challenges and future directions, while Section 7 will include case studies and applications from the actual world. The review will close with a summary of the most important discoveries and a focus on how nonlinear control strategies may influence structural engineering in the future in Section 8.

2. Nonlinearities in structural engineering

Although structural engineering is based on the concepts of equilibrium and stability, it deals with intrinsic nonlinearities that have a big impact on how structures behave [39, 40]. These nonlinearities originate from different causes, including geometric nonlinearities [41–43], material nonlinearities [19, 44], large deflection nonlinearities [18, 45, 46], boundary condition nonlinearities [47, 48], and dynamic nonlinearities [49].

2.1 Geometric nonlinearities

Nonlinearities are brought about by geometric complications, particularly when working with thin structures or substantial deformations [41–43]. The effect of geometric nonlinearities may be ignored by traditional linear analysis, which could result in inaccurate predictions of structural reactions [4, 6, 8, 45, 50–52]. To accurately represent the behavior of structures under different loads, sophisticated geometrically nonlinear models are essential [53].

2.2 Material nonlinearities

The fundamental components of any structure, materials, frequently behave nonlinearly in a variety of situations [19, 44]. For example, strain-strain correlations in concrete are not linear, especially in the post-yield zone, where strains may not be precisely proportionate to stresses [54, 55]. Furthermore, the nonlinearity of steel materials presents difficulties for linear analysis techniques, particularly in the plastic deformation region [56, 57].

2.3 Large deflection nonlinearities

Large deflection nonlinearity in structures refers to the behavior where deformations become significant enough to cause nonlinear responses, deviating from linear elastic assumptions [18, 45, 46]. Under large deflections, structural

elements undergo considerable distortion, altering their stiffness and load-carrying capacity [58]. This phenomenon commonly occurs in slender structures under high loads or flexible materials [46]. Nonlinear structural analysis techniques are employed to accurately predict deformations and stresses in such scenarios, crucial for ensuring structural integrity and safety [45, 59].

2.4 Boundary condition nonlinearities

Nonlinearities also stem from the boundary conditions imposed on structures [47, 48]. The rigidity of connections and supports can influence the overall structural response. In cases where supports are not perfectly rigid or exhibit nonlinear behavior, the overall structural response becomes intricate and necessitates sophisticated analysis methods [48, 60].

2.5 Dynamic nonlinearities

Dynamic nonlinearities are introduced by dynamic loading situations, such as seismic or wind-induced forces, which are difficult for conventional linear approaches to describe [49]. Dynamic force magnitude and frequency might result in nonlinear responses, necessitating specialized tactics for precise forecasting [61, 62].

Developing sophisticated numerical models and simulation methods to better comprehend and measure these nonlinearities has been the main focus of recent research projects [19, 26, 60, 63–67]. Computational methods and finite element analysis (FEA) have helped shed light on dynamic loading situations, geometric configurations, and nonlinear behavior of materials [21, 68–71]. Experimental experiments have also helped capture real-world nonlinear reactions and validate numerical models [72, 73].

The nonlinearity of structural engineering in general—which is required to comprehend how stress and shape control are managed in structures—was covered in this part. We will now talk about the latest advances in nonlinear control methods and how they can be used to manage stress and change the shape of structures. This study takes into account the complicated issues that come up because structural engineering is not a straight-line subject.

3. Shape control strategies

Shape control schemes have become essential elements in the field of nonlinear control in structural engineering, as designers strive for exact structural configurations and esthetically pleasing designs. This section delves into many approaches that support the dynamic shaping of structural shapes.

3.1 Adaptive control for shape modification

Among the many methods for dynamically sculpting structural shapes, adaptive control is particularly important. Adaptive control ensures that different external loads and environmental impacts are continuously adapted by changing control parameters based on real-time structural conditions useful [74–76]. This adaptability is especially useful in situations where structures that are subject to shifting loading circumstances or deployable structures need to have their structural configurations change dynamically over time [38].

3.2 Neural networks for data-driven shape learning

A data-driven paradigm is introduced when shape control algorithms incorporate neural networks [77]. Neural networks may learn and adapt to complicated structural behaviors because they are inspired by complex learning mechanisms [78]. Neural networks, through analyzing large datasets and identifying nonlinear patterns, provide a reliable way to shape structures according to past performance and interactions with the environment [79].

3.3 Fuzzy logic for managing uncertainties in form

Fuzzy logic is used to shape structural configurations because of its reputation for handling uncertainties and imperfect information [80, 81]. Fuzzy logic offers a framework for decision-making that takes uncertainties into account in situations where exact mathematical models may be difficult to develop [82]. When working with materials that have changing characteristics or complex structural geometries, this is especially helpful [35, 83].

3.4 Model predictive control for dynamic form optimization

Model Predictive Control (MPC) is a technique that has gained popularity for optimizing performance using predictive models to shape structural shapes [84, 85]. MPC takes into account restrictions and objectives by using a predictive model of the structure and iteratively adjusting control inputs to attain desired forms [85, 86]. When sustaining ideal structural arrangements requires real-time alterations, this approach works well.

3.5 Various examples of applications of structure-based shape control

The pursuit of geometric perfection is essential in the field of structural engineering for a variety of architectural compositions. Determining nodal points is the first step toward the reality of architectural forms, from the conception of design to the fulfillment of esthetic quality. This requirement is demonstrated by beams [16, 29, 87–95], trusses [9, 96–103], and frames [104–106] by linear or nonlinear methods, where the exact placement of structural components guarantees the effective distribution of loads while maintaining structural integrity. In addition, the sphere [6, 8, 52, 107, 108], antenna structures [100, 109, 110], egg-shaped structure [4], and dome [3, 5, 15, 111] constructions are examples of architectural achievements where geometric precision combines with esthetic appeal and practicality to create memorable areas and famous structures. While cable structures [3, 10–13, 30, 31, 103, 112–119] challenge preconceived concepts of stability and balance with their intricate designs, cable structures, with their elegant curves and tensioned forms, epitomize the union of engineering genius with artistic harmony.

4. Stress control strategies

To guarantee the longevity, safety, and structural integrity of designed systems, effective stress control techniques are essential. Various approaches used in nonlinear control to control stress in structural engineering are discussed in this section.

4.1 Adaptive control for stress management

Adaptive control techniques are essential for dynamically regulating stress inside structural parts. Adaptive control makes sure that structures can adapt to shifting loads and environmental variables by continuously modifying control settings depending on real-time stress levels, preventing excessive stress concentrations [120, 121]. When structural elements are subjected to fluctuating and unpredictable stresses, this adaptability is very valuable.

4.2 Neural networks for stress prediction and mitigation

A data-driven approach to stress management is offered by the incorporation of neural networks into stress control techniques. Since neural networks are very good at learning complicated patterns, they can be used to anticipate the distribution of stress inside structures [122–124]. Neural networks have a role in stress concentration mitigation and structural performance optimization through the utilization of real-time feedback and historical data.

4.3 Fuzzy logic for stress mitigation in uncertain environments

Stress control systems use fuzzy logic, which can handle uncertainties, to govern structural reactions in unpredictable settings [125, 126]. Fuzzy logic helps decision-makers reduce stress concentrations and improve structural resilience when external variables contribute to inaccurate information [127, 128]. This strategy is especially important in places where environmental uncertainty is common.

4.4 Model predictive control for optimal stress regulation

One effective method for controlling stress in structural parts is Model Predictive Control (MPC) [129]. MPC uses predictive models to repeatedly improve control inputs to produce optimal stress distributions while taking goals and constraints into account [130]. When accurate stress modulation is essential for the longevity and safety of structures, this approach works well.

4.5 Various examples of applications of structure-based stress control

Various structural domains can benefit from the practical implementation of stress control systems. These techniques have been used to optimize stress distributions, improve structural safety, and lengthen the lifespan of crucial infrastructure, ranging from buildings to bridge structures. For instance, trusses [2, 7, 26, 103], cables stayed bridges [2] and cable structures [26, 131] by linear [2, 7] or nonlinear [26, 131] methods. The mentioned examples highlight successful implementations and offer insightful information on the efficacy and practicality of stress control techniques.

5. Integration of shape and stress control

One of the most important aspects of managing structural integrity and performance holistically is the relationship between stress distribution and structural shape [2, 7, 14]. The integration of shape and stress control measures is

examined in this section, emphasizing the benefits that result from examining these two factors together.

5.1 Simultaneous shape and stress control strategies

A major development in the nonlinear control of structural engineering is the merging of form and stress control techniques [31]. Combining control over stress management with structural form manipulation enables a holistic strategy for maximizing both performance and safety [7, 10, 11, 31, 132]. To accomplish this dual goal, a combination of neural networks, fuzzy logic, model predictive control, and adaptive control can be used [7, 11, 132].

5.2 Dynamic adaptability for form and stress optimization

The integration of form and stress management is largely dependent on adaptive control techniques [35]. Structures are capable of real-time adaptation to changing external loads and environmental factors by dynamically adjusting control settings based on both form and stress conditions [38, 133, 134]. By avoiding both high-stress concentrations and undesired deformations, this dynamic adaptability guarantees both optimal performance and safety [135, 136].

5.3 Learning-based approaches for simultaneous control

Shape and stress are simultaneously controlled by neural networks, which are renowned for their capacity to learn intricate patterns [137, 138]. Neural networks can optimize the distribution of stress and the shape of the structure by using both past data and current feedback [137, 139]. This learning-based strategy works especially well in situations where form and stress have a complex and nonlinear relationship [140].

5.4 Uncertainty management through fuzzy logic

Fuzzy logic is included to help manage the uncertainties in form and stress control [125, 141]. In the face of imperfect information, fuzzy logic offers a framework for decision-making that guarantees the robustness of structural adjustments for shape and stress in unpredictable situations [141]. This method improves a structure's resistance to changing and erratic circumstances [78, 142].

5.5 Optimal predictive control for form and stress harmony

Through iterative adjustments of control inputs based on predictive models, Model Predictive Control (MPC) excels in maximizing both form and stress [143]. By taking into account both form and stress objectives at the same time, trade-offs between stress management and structural configurations are avoided throughout the optimization process [144, 145].

5.6 Various examples of application of structure-based simultaneous shape and stress control

In structural engineering, combined form and stress control have many real-world uses. These strategies demonstrate the versatility and adaptability of

concurrent shape and stress control methodologies, ranging from adaptive building facades that dynamically respond to environmental conditions [7, 10, 11, 31, 132] while managing stress to aerospace structures that optimize both aerodynamics and structural integrity [146, 147]. There are some examples of combined form and stress control, such as beams, trusses [2, 9, 97, 98, 101, 103, 148], spheres [6, 8, 52], antenna structures, cable structures [30, 31, 103, 115, 116], and domes [5, 111] by linear [6, 8, 9, 52, 97, 98, 101, 148] or nonlinear [30, 31] methods. The mentioned examples highlight successful implementations and offer insightful information on the efficacy and practicality of stress control techniques [2, 5, 6, 8, 9, 30, 31, 52, 97, 98, 101, 115, 116, 119, 120, 148–150].

6. Challenges and future directions

The progression of nonlinear control systems for managing structural shape and stress within structural engineering will give rise to a multitude of opportunities and challenges. The challenges currently encountered by the field's practitioners and researchers are discussed in this part, along with possible future paths.

6.1 Computational complexity

More computing power is frequently required for the application of complex nonlinear control schemes [151]. Managing computational complexity becomes an increasingly important difficulty as systems become more complicated and real-time responses are required [35]. One persistent issue is manipulating the accuracy of control algorithms with the effectiveness of computing procedures [152].

6.2 Robustness in the face of uncertainties

Nonlinear control systems face difficulties due to the inherent uncertainties in structural engineering, which arise from variations in the environment, material qualities, and load conditions [35, 153–155]. Although adaptive control and fuzzy logic attempt to manage uncertainties, it is still difficult to guarantee that control algorithms will stay robust in a variety of strange and unpredictable situations [156].

6.3 Advancements in sensing technologies

Sufficient and timely data from sensing technologies are essential for nonlinear control techniques to work well [35, 78]. Continuous sensor advances are essential for improving the accuracy and dependability of control activities [157–160]. Examples of these sensors include vision-based systems and strain gauges [160]. The development of nonlinear control techniques depends critically on ongoing research in sensor technology.

6.4 Interplay between shape and stress control

Although the integration of stress control mechanisms with shape is a promising option, there are obstacles to comprehending the complex interplay between these factors [2, 7, 14, 17]. A thorough understanding of the complex relationship between

stress distribution and structural form is required to maximize synergy without compromising individual objectives [17].

6.5 Future directions

Nonlinear control in structural engineering has a bright future ahead of it. Scholars are presently investigating novel approaches, like the integration of machine learning algorithms, to augment the flexibility of control techniques. Opportunities for autonomous optimization of structural configurations under variable situations are presented by advances in artificial intelligence, specifically in reinforcement learning. More future objectives for studying nonlinear shape and stress control are materials innovation, which includes investigating innovative materials as a means of enhancing nonlinear control techniques. Adaptive smart materials can work in concert with control systems to create new opportunities for materials that dynamically adjust to stress situations, shape memory alloys, and self-healing structures [160].

7. Case studies and applications

Analyzing nonlinear control systems for form and stress in structural engineering in real-world applications offers important insights into the applicability, effectiveness, and flexibility of these approaches. The impact of successful implementations on different structural domains is examined in this section through a variety of case studies.

7.1 Deployable structures for adaptive environments

Deployable structures—whose form dynamically adjusts to changing environmental conditions—have become more and more popular. Examples of case studies demonstrate how adaptive control systems allow structures to adjust to Environments for instance Goliath umbrellas at Nabawi Mosque compound in Medina [161] as shown in **Figure 2**. With its deployable structures for adaptable situations, Goliath umbrellas at the Nabawi Mosque compound in Medina exhibit the inventiveness of structural engineering. It expands prayer space during busy times, like Ramadan and Hajj, by using modular platforms and temporary umbrellas. Its adaptable layout guarantees a smooth transition with the Prophet's Mosque, allowing for different audience sizes to be accommodated without compromising the sacredness of the location. It is evidence of creative structural engineering solutions.

7.2 Adaptive morphable structures

Advanced movable structures, which involved altering and controlling the geometric shape of the structures with dynamic motion and altering the behavior of the structures concurrently, were presented at International Expo 2005, Aichi, Japan [162]. He displayed the massive, mobile monument depicted in **Figure 3**. Three similar movable towers with four moving truss components make up this monument. As a result of the ease with which shape morphing from well-known traditional truss structures can be achieved, as demonstrated in **Figure 4**, the monument's shape can be altered to a variety of truss shapes by replacing some of the trusses with linear

Figure 2.
Goliath umbrellas at Nabawi mosque compound in Medina.

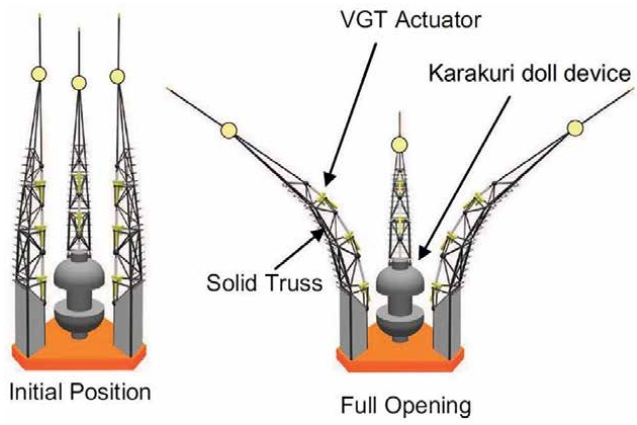


Figure 3.
Illustration of the towers displayed during Aichi, Japan's international Expo 2005 [162].

displacement actuators [163] and adjusting the length of each extendable member (extensible actuator) [164, 165].

7.3 Aerospace structures with dynamic morphing

In the aerospace industry, optimizing aerodynamics through the shaping of aircraft wings and surfaces is largely dependent on nonlinear control systems [166]. For example, Commercial Aircraft Morphing demonstrates how dynamic morphing, increased fuel efficiency [167], and improved overall performance of aeronautical structures may be achieved with the help of adaptive control, neural networks, and model predictive control as shown in **Figure 5**.

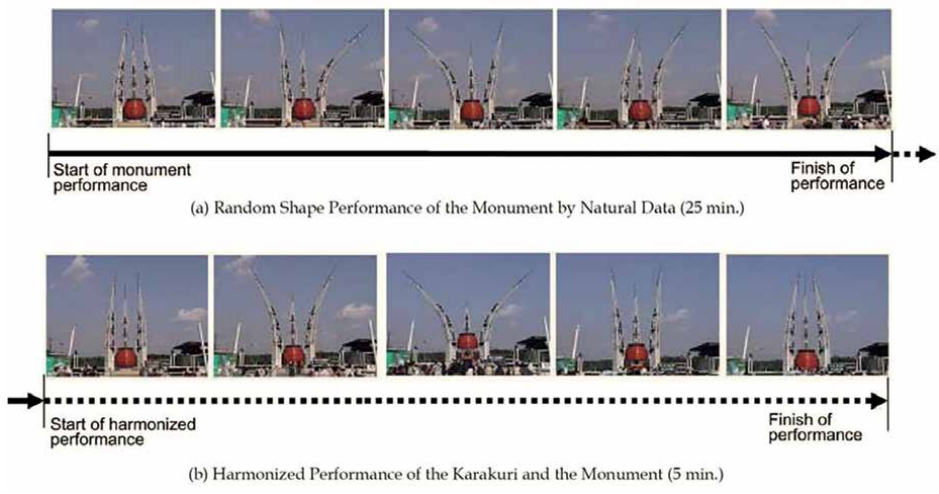


Figure 4.
Monument shape is altered based on performance trends [162].

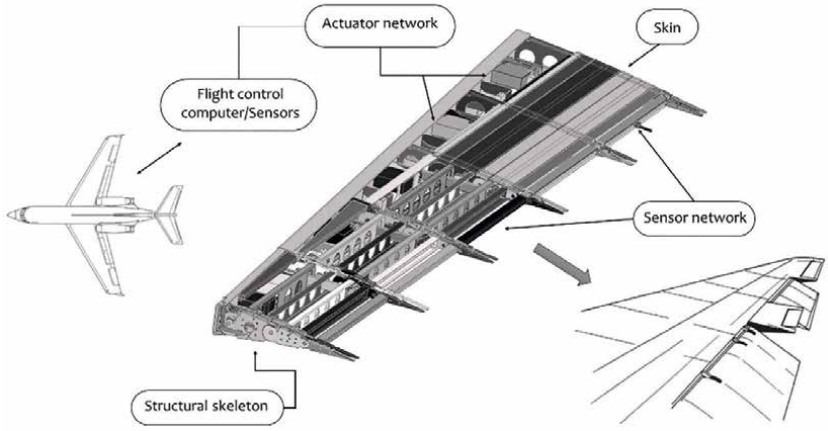


Figure 5.
Diagrams showing the subsystems of the morphing wing gadget [167].

7.4 Bridges with shape and stress optimization

Bridges are an example of vital infrastructure where shape and stress control must be integrated. The case studies were not available; however, the model was demonstrated in the lab to demonstrate how control systems maximize the form and stress distribution of bridge structures [13].

7.5 Tetragonal lattice structure

A shape-morphing control for the space model of a tetragonal lattice system was a case study test validating the nonlinear force method for large deformation control and comparing it to the linear force method [168]. The shape-morphing target was examined in two cases. The first targeting case was approaching the doubly curved

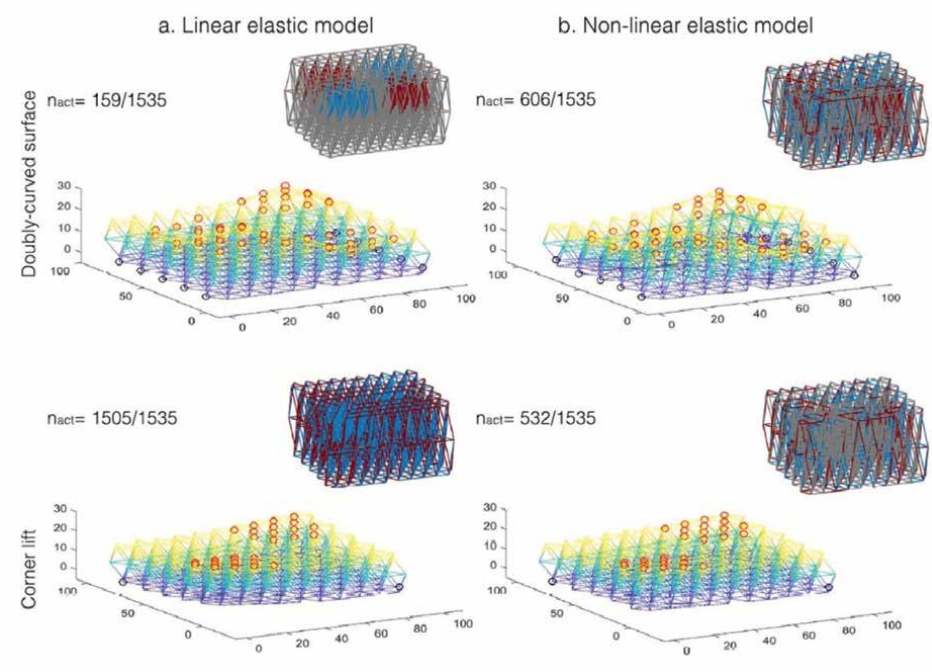


Figure 6.
Shape morphing control of the tetragonal lattice structure for the (a) linear and (b) nonlinear methods; nodes that are to be moved or pinned are shown by the red circles, and the undeformed structures in the top left corner of each deformation reveal where the actuators are placed; components in gray are fixed, while those in blue are expanding, and those in red are shrinking [168].

surface of the morphing, while the other case was the corner lift of the assembly. In the doubly curved scenario for approaching the target, it required 159 actuators (n_{act}) by linear control but 606 actuators by nonlinear control out of 1535 members. This great difference from used actuators refers to neglecting the member stress caused by the elongation of other elements. Likewise, for the corner lift scenario, the linear technique overestimated employing almost the whole body of the system as actuators (n_{act} = 1505), while the nonlinear used 532 actuators for the shape morphing control. The cases demonstrated that compared to the linear technique, the nonlinear controlling approach yields the most fitting results for large deformations of complex assemblies. The findings of both cases are presented in **Figure 6** [168].

8. Conclusions and recommendations

A thorough examination of structural engineering's nonlinear control techniques for form and stress reveals a vibrant field full of opportunities, difficulties, and advancements. This conclusion summarizes the main conclusions and shows the direction for future research.

8.1 Conclusions

The chapter presents a comprehensive overview of the latest developments in nonlinear control techniques applied to form and stress management in structural

engineering. By synthesizing research findings, case studies, challenges, and potential directions, it serves as a roadmap for both scholars and professionals seeking to propel structural engineering into new realms. The findings of this chapter inspire the structural engineering community to deepen their grasp and application of nonlinear control, fostering advancements that will enhance the built environment significantly.

In essence, the chapter offers an in-depth perspective on the evolving landscape of nonlinear control within structural engineering. A pivotal approach highlighted is the fusion of shape and stress control methodologies, which lays the groundwork for resilient, adaptable, and human-centric structural designs. As we navigate through this evolving terrain, collaborative efforts among researchers, practitioners, and experts from diverse fields become indispensable.

The presence of nonlinearities poses significant challenges to traditional control systems, often built upon linear assumptions. Linear control techniques may inadequately capture the true behavior of structures, leading to compromised performance and potential safety hazards. Hence, there is a growing imperative to explore nonlinear control frameworks capable of effectively managing the intricate nonlinear dynamics inherent in structural systems.

8.2 Recommendation

Nonlinear control in structural engineering has a bright future ahead of it. Advances in machine learning, material innovation, and sustainability are areas that researchers are encouraged to investigate. The unification of reinforcement learning, artificial intelligence, and smart materials promises revolutionary discoveries that are in line with the changing needs of the built environment.

Author details

Najmadeen Saeed[1,2*] and Shna Abdulkarim[1,3]

1 Civil Engineering Department, University of Raparin, Kurdistan Region, Iraq

2 Faculty of Engineering, Civil Engineering Department, Tishk International University, Erbil, Kurdistan Region, Iraq

3 Civil Engineering Department, Technical Engineering, Erbil Polytechnic University, Erbil, Kurdistan Region, Iraq

*Address all correspondence to: najmadeen_qasre@uor.edu.krd

IntechOpen

© 2024 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. (cc) BY

References

[1] Ziegler F. Computational aspects of structural shape control. Computers & Structures. 2005;**83**(15):1191-1204. DOI: 10.1016/j.compstruc.2004.08.026

[2] Saeed NM. Prestress and Deformation Control in Flexible Structures [PhD Dissertation]. Cardiff, UK: Cardiff University; 2014. Available from: http://orca.cf.ac.uk/69777/

[3] Manguri A, Saeed N, Haydar B. Optimal shape refurbishment of distorted dome structure with safeguarding of member stress. In: 7th International Engineering Conference "Research & Innovation amid Global Pandemic" (IEC). Erbil, Iraq: IEEE; 2021. pp. 90-95. DOI: 10.1109/IEC52205.2021.9476107

[4] Saeed N, Manguri A, Abdulkarim S, Shekha A. Shape restoration of deformed egg-shaped single layer space frames. In: 2019 International Conference on Advanced Science and Engineering (ICOASE), Duhok, Kurdistan Region, Iraq. Manhattan, New York, U.S: IEEE; 2019. pp. 220-225. DOI: 10.1109/ICOASE.2019.8723714

[5] Saeed N, Manguri A, Al-Zahawi S. Optimum geometry and stress control of deformed double layer dome for gravity and lateral loads. In: 2021 7th International Engineering Conference "Research & Innovation amid Global Pandemic" (IEC). Erbil, Iraq: IEEE; 2021. pp. 84-89. DOI: 10.1109/IEC52205.2021.9476094

[6] Mahmood A, Katebi J, Saeed N, Manguri A. Optimized stress and geometry control of spherical structures under lateral loadings. In: 2022 8th International Engineering Conference on Sustainable Technology and Development (IEC). Erbil, Iraq: IEEE; 2022. pp. 142-148. DOI: 10.1109/IEC54822.2022.9807455

[7] Saeed NM, Kwan ASK. Simultaneous displacement and internal force prescription in shape control of pin-jointed assemblies. AIAA Journal. 2016;**54**(8):2499-2506. DOI: 10.2514/1.J054811

[8] Manguri A, Saeed N, Mahmood A, Katebi J, Jankowski R. Optimal reshaping and stress controlling of double-layer spherical structures under vertical loadings. Archives of Civil Engineering. 2022;**68**:591-606. DOI: 10.24425/ace.2022.143056

[9] Saeed NM, Manguri AA, Szczepanski M, Jankowski R, Haydar BA. Static shape and stress control of trusses with optimum time, actuators and actuation. International Journal of Civil Engineering. 2023;**21**(3):379-390. DOI: 10.1007/s40999-022-00784-3

[10] Manguri AA, Kwan ASK, Saeed NM. Adjustment for shape restoration and force control of cable arch stayed bridges. International Journal of Computational Methods and Experimental Measurements. 2017;**5**(4):514-521. DOI: 10.2495/CMEM-V5-N4-514-521

[11] Saeed NM, Manguri AAH, Adabar AM. Shape and force control of cable structures with minimal actuators and actuation. International Journal of Space Structures. 2021;**36**(3):241-248. DOI: 10.1177/09560599211045851

[12] Saeed NM. Simultaneous force and deformation control of cable arch stayed bridges. Kufa Journal of Engineering. 2019;**10**(4):66-75. DOI: 10.30572/2018/kje/100406

[13] Saeed NM, Kwan ASK. Displacement and internal force control in cable-stayed bridges. Proceedings of the Institution of Civil Engineers-Bridge Engineering. 2018;**171**(1):63-76. DOI: 10.1680/jbren.16.00010

[14] Saeed NM, Kwan ASK. Displacement and force control of complex element structures by matrix condensation. Structural Engineering and Mechanics. 2016;**59**(6):973-992. DOI: 10.12989/sem.2016.59.6.973

[15] Abdulkarim SJ, Saeed NM, Haji HA. Direct displacement control of deformed double layer dome. UKH Journal of Science Engineering. 2020;**4**(1):1-14. DOI: 10.25079/ukhjse.v4n1y2020.pp1-14

[16] Ang KK, Achuthan A, Wang CM. Linear and nonlinear actuations in shape control of beams. In: Smart Structures and Devices. Vol. 4235. Melbourne, Australia: International Society for Optics and Photonics; 2001. pp. 509-520. DOI: 10.1117/12.420895

[17] Saeed NM. Displacement control of nonlinear pin-jointed assemblies based on force method and optimization. AIAA Journal. 2022;**60**(2):1024-1031. DOI: 10.2514/1.J060568

[18] Zhang Z-J, Chen B-S, Bai R, Liu Y-P. Non-linear behavior and design of steel structures: Review and outlook. Buildings. 2023;**13**(8):2111. DOI: 10.1016/S0143-974X(01)00050-5

[19] Sathyamoorthy M. Nonlinear Analysis of Structures (1997). Boca Raton, Florida, USA: CRC Press; 2017

[20] Spacone E, El-Tawil S. Nonlinear analysis of steel-concrete composite structures: State of the art. Journal of Structural Engineering. 2004;**130**(2):159-168. DOI: 10.1061/(ASCE)0733-9445(2004)130:2(159)

[21] Mondkar D, Powell G. Finite element analysis of non-linear static and dynamic response. International Journal for Numerical Methods in Engineering. 1977;**11**(3):499-520. DOI: 10.1002/nme.1620110309

[22] Lewis WJ. Tension Structures: Form and Behaviour. London, UK: Thomas Telford; 2003

[23] Olfati-Saber R. Nonlinear Control of Underactuated Mechanical Systems with Application to Robotics and Aerospace Vehicles. Cambridge, MA, USA: Massachusetts Institute of Technology; 2001

[24] Slotine J-JE, Li W. Applied Nonlinear Control. Vol. 199. NJ: Prentice hall Englewood Cliffs; 1991

[25] Kwan ASK. A simple technique for calculating natural frequencies of geometrically nonlinear prestressed cable structures. Computers & Structures. 2000;**74**(1):41-50. DOI: 10.1016/S0045-7949(98)00318-6

[26] Abdulkarim SJ, Saeed NM. Nonlinear technique of prestressing spatial structures. Mechanics Research Communications. 2023;**127**:104040. DOI: 10.1016/j.mechrescom.2022.104040

[27] Zhu B, Ren Z, Xie W, Guo F, Xia X. Active nonlinear partial-state feedback control of contacting force for a pantograph-catenary system. ISA Transactions. 2019;**91**:78-89. DOI: 10.1016/j.isatra.2019.01.033

[28] Sun D, Tong L. Static shape control of structures using nonlinear piezoelectric actuators with energy constraints. Smart Materials and Structures. 2004;**13**(5):1059. DOI: 10.1088/0964-1726/13/5/012

[29] Achuthan A, Keng AK, Ming WC. Shape control of coupled nonlinear piezoelectric beams. Smart Materials and Structures. 2001;**10**(5):914. DOI: 10.1088/0964-1726/10/5/308

[30] Xu X, Luo Y. Non-linear displacement control of prestressed cable structures. Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering. 2009;**223**(7):1001-1007. DOI: 10.1243/09544100JAERO455

[31] Yuan X, Liang X, Li A. Shape and force control of prestressed cable-strut structures based on nonlinear force method. Advances in Structural Engineering. 2016;**19**(12):1917-1926. DOI: 10.1177/1369433216652411

[32] Bar-Kana I, Kaufman H, Balas M. Model reference adaptive control of large structural systems. Journal of Guidance, Control, and Dynamics. 1983;**6**(2):112-118. DOI: 10.2514/3.8544

[33] Baz A, HONG JT. Adaptive control of flexible structures using modal positive position feedback. International Journal of Adaptive Control and Signal Processing. 1997;**11**(3):231-253. DOI: 10.1002/(SICI)1099-1115(199705)11:3%3C231::AID-ACS435%3E3.0.CO;2-8

[34] Biondini F, Bontempi F, Malerba PG. Fuzzy reliability analysis of concrete structures. Computers & Structures. 2004;**82**(13-14):1033-1052. DOI: 10.1016/j.compstruc.2004.03.011

[35] Housner G, Bergman LA, Caughey TK, Chassiakos AG, Claus RO, Masri SF, et al. Structural control: past, present, and future. Journal of Engineering Mechanics. 1997;**123**(9):897-971. DOI: 10.1061/(ASCE)0733-9399(1997)123:9(897)

[36] Bathaei A, Zahrai SM. Vibration control of an eleven-story structure with MR and TMD dampers using MAC predictive control, considering nonlinear behavior and time delay in the control system. Structures. 2024;**60**:105853. DOI: 10.1016/j.istruc.2024.105853

[37] Bossi L, Rottenbacher C, Mimmi G, Magni L. Multivariable predictive control for vibrating structures: An application. Control Engineering Practice. 2011;**19**(10):1087-1098. DOI: 10.1016/j.conengprac.2011.05.003

[38] Korkmaz S. A review of active structural control: Challenges for engineering informatics. Computers & Structures. 2011;**89**(23):2113-2132. DOI: 10.1016/j.compstruc.2011.07.010

[39] Smith SA, Brake MR, Schwingshackl CW. On the characterization of nonlinearities in assembled structures. Journal of Vibration and Acoustics. 2020;**142**(5):051105. DOI: 10.1115/1.4046956

[40] Carrella A, Ewins D. Identifying and quantifying structural nonlinearities in engineering applications from measured frequency response functions. Mechanical Systems and Signal Processing. 2011;**25**(3):1011-1027. DOI: 10.1016/j.ymssp.2010.09.011

[41] Paimushin V. Problems of geometric non-linearity and stability in the mechanics of thin shells and rectilinear columns. Journal of Applied Mathematics and Mechanics. 2007;**71**(5):772-805. DOI: 10.1016/j.jappmathmech.2007.11.012

[42] Pagani A, Carrera E. Unified formulation of geometrically nonlinear refined beam theories. Mechanics of Advanced Materials and Structures. 2018;**25**(1):15-31. DOI: 10.1080/15376494.2016.1232458

[43] Vassilopoulou I, Petrini F, Gantes CJ. Nonlinear dynamic behavior of cable nets subjected to wind loading. Structures. 2017;**10**:170-183. DOI: 10.1016/j.istruc.2017.03.004

[44] Chili W, Galea S, Jones R. The role of material nonlinearities in composite structures. Composite Structures. 1997;**38**(1-4):71-81. DOI: 10.1016/S0263-8223(97)00043-3

[45] Krawinkler H. Importance of good nonlinear analysis. The Structural Design of Tall and Special Buildings. 2006;**15**(5):515-531. DOI: 10.1002/tal.379

[46] Ghuku S, Saha KN. A review on stress and deformation analysis of curved beams under large deflection. International Journal of Engineering and Technologies. 2017;**11**:13-39. DOI: 10.56431/p-48538j

[47] Mao X-Y, Ding H, Chen L-Q. Vibration of flexible structures under nonlinear boundary conditions. Journal of Applied Mechanics. 2017;**84**(11):111006. DOI: 10.1115/1.4037883

[48] Amabili M. Nonlinear vibrations of rectangular plates with different boundary conditions: Theory and experiments. Computers & Structures. 2004;**82**(31-32):2587-2605. DOI: 10.1016/j.compstruc.2004.03.077

[49] Ali HMH. Nonlinear Dynamic Analysis of Tall Buildings under Wind Loads. Khartoum, Sudan: Sudan University of Science and Technology; 2016

[50] Xie C, An C, Liu Y, Yang C. Static aeroelastic analysis including geometric nonlinearities based on reduced order model. Chinese Journal of Aeronautics. 2017;**30**(2):638-650. DOI: 10.1016/j.cja.2016.12.031

[51] Saeed N, Manguri A, Szczepanski M, Jankowski R. Non-linear analysis of structures utilizing load-discretization of stiffness matrix method with coordinate update. Applied Sciences. 2022;**12**(5):2394. DOI: 10.3390/app12052394

[52] Saeed N, Katebi J, Manguri A, Mahmood A, Szczepanski M, Jankowski R. Using minimum actuators to control shape and stress of a double layer spherical model under gravity and lateral loadings. Advances in Science Technology. Research Journal. 2022;**22**(6):1-13. DOI: 10.12913/22998624/155214

[53] Levy R, Spillers WR. Analysis of Geometrically Nonlinear Structures. Berlin, Germany: Springer Science & Business Media; 2013

[54] Maekawa K, Okamura H, Pimanmas A. Non-linear Mechanics of Reinforced Concrete. Boca Raton, Florida, USA: CRC Press; 2003

[55] Foster SJ. The Structural Behaviour of Reinforced Concrete Deep Beams. Sydney, Australia: UNSW Sydney; 1992

[56] Wong MB. Plastic Analysis and Design of Steel Structures. Oxford, United Kingdom: Butterworth-Heinemann; 2011

[57] Jones RM. Deformation Theory of Plasticity. Blacksburg, Virginia, USA: Bull Ridge Corporation; 2009

[58] Kuder IK, Arrieta AF, Raither WE, Ermanni P. Variable stiffness material and structural concepts for morphing applications. Progress in Aerospace Sciences. 2013;**63**:33-55. DOI: 10.1016/j.paerosci.2013.07.001

[59] Christensen J, Bastien C. Nonlinear Optimization of Vehicle Safety Structures: Modeling of Structures

Subjected to Large Deformations. Oxford, United Kingdom: Butterworth-Heinemann; 2015

[60] Sekulovic M, Salatic R. Nonlinear analysis of frames with flexible connections. Computers & Structures. 2001;**79**(11):1097-1107. DOI: 10.1016/S0045-7949(01)00004-9

[61] Bachynski EE. Design and dynamic analysis of tension leg platform wind turbines [PhD Dissertation]. Trondheim, Norway: Norwegian University of Science and Technology; 2014. Available from: http://hdl.handle.net/11250/238768

[62] Maguire JR. Dynamics: An Introduction for Civil and Structural Engineers. London, UK: Thomas Telford; 2002

[63] Abad MSA, Shooshtari A, Esmaeili V, Riabi AN. Nonlinear analysis of cable structures under general loadings. Finite Elements in Analysis and Design. 2013;**73**:11-19. DOI: 10.1016/j.finel.2013.05.002

[64] Rezaiee-Pajand M, Mohammadi-Khatami M. Nonlinear analysis of cable structures using the dynamic relaxation method. Frontiers of Structural Civil Engineering. 2021;**15**(1):253-274. DOI: 10.1007/s11709-020-0639-y

[65] Coarita E, Flores L. Nonlinear analysis of structures cable-truss. International Journal of Engineering technology. 2015;**7**(3):160. DOI: 10.7763/IJET.2015.V7.786

[66] Thai H-T, Kim S-E. Nonlinear static and dynamic analysis of cable structures. Finite Elements in Analysis and Design. 2011;**47**(3):237-246. DOI: 10.1016/j.finel.2010.10.005

[67] Raju NRBK, Nagabhushanam J. Nonlinear structural analysis using integrated force method. Sadhana. 2000;**25**(4):353-365. Available from: https://link.springer.com/article/10.1007/BF03029720

[68] Coda HB, Silva APDO, Paccola RR. Alternative active nonlinear total Lagrangian truss finite element applied to the analysis of cable nets and long span suspension bridges. Latin American Journal of Solids Structures. 2020;**17**:1-30. DOI: 10.1590/1679-78255818

[69] Skorpen, S.A. and Dekker, "The application and interpretation of linear finite element analysis results in the design and detailing of hogging moment regions in reinforced concrete flat plates," Journal of the South African Institution of Civil Engineering, Vol. 56, No. 1, 2014, pp. 77-92. Available from: https://hdl.handle.net/10520/EJC154136

[70] Gambhir ML, Batchelor B. A finite element for 3-D prestressed cablenets. International Journal for Numerical Methods in Engineering. 1977;**11**(11):1699-1718. DOI: 10.1002/nme.1620111106

[71] Sabouni-Zawadzka A, Zawadzki A. Simulation of a deployable tensegrity column based on the finite element modeling and multibody dynamics simultions. Archives of Civil Engineering. 2020;**66**(4). DOI: 10.24425/ace.2020.135236

[72] Catalfamo S, Smith SA, Morlock F, Brake MR, Reuß P, Schwingshackl CW, et al. Effects of experimental methods on the measurements of a nonlinear structure. In: Dynamics of Coupled Structures, Volume 4: Proceedings of the 34th IMAC, a Conference and Exposition on Structural Dynamics 2016. Cham, Switzerland: Springer; 2016. pp. 491-500. DOI: 10.1007/978-3-319-29763-7_48

[73] Peeters M, Kerschen G, Golinval J-C. Modal testing of nonlinear vibrating structures based on nonlinear normal modes: Experimental demonstration. Mechanical Systems and Signal Processing. 2011;**25**(4):1227-1247. DOI: 10.1016/j.ymssp.2010.11.006

[74] Austin F, Rossi MJ, Van Nostrand W, Knowles G, Jameson A. Static shape control for adaptive wings. AIAA Journal. 1994;**32**(9):1895-1901. DOI: 10.2514/3.12189

[75] Dimino I, Schueller M, Gratias A. An adaptive control system for wing TE shape control. In: Industrial and Commercial Applications of Smart Structures Technologies 2013. Vol. 8690. San Diego, California, United States: SPIE; 2013. pp. 112-122. DOI: 10.1117/12.2012187

[76] Landau ID, Lozano R, M'Saad M. Adaptive control. Vol. 51. New York: Springer; 1998

[77] Zhang R, Liu Y, Sun H. Physics-guided convolutional neural network (PhyCNN) for data-driven seismic response modeling. Engineering Structures. 2020;**215**:110704. DOI: 10.1016/j.engstruct.2020.110704

[78] Salehi H, Burgueño R. Emerging artificial intelligence methods in structural engineering. Engineering Structures. 2018;**171**:170-189. DOI: 10.1016/j.engstruct.2018.05.084

[79] Hsu TY, Loh CH. Damage detection accommodating nonlinear environmental effects by nonlinear principal component analysis. Structural Control and Health Monitoring: The Official Journal of the International Association for Structural Control and Monitoring and of the European Association for the Control of Structures. 2010;**17**(3):338-354. DOI: 10.1002/stc.320

[80] Pourzeynali S, Lavasani H, Modarayi A. Active control of high rise building structures using fuzzy logic and genetic algorithms. Engineering Structures. 2007;**29**(3):346-357. DOI: 10.1016/j.engstruct.2006.04.015

[81] Ayyub BM, Guran A, Haldar A. Uncertainty Modeling in Vibration, Control and Fuzzy Analysis of Structural Systems. Vol. 10. Tuck Link, Singapore: World Scientific; 1997

[82] Sousa JMC, Kaymak U. Fuzzy Decision Making in Modeling and Control. Vol. 27. Tuck Link, Singapore: World Scientific; 2002

[83] Bouaziz O, Bréchet Y, Embury JD. Heterogeneous and architectured materials: A possible strategy for design of structural materials. Advanced Engineering Materials. 2008;**10**(1-2):24-36. DOI: 10.1002/adem.200700289

[84] Chen Y, Zhang S, Peng H, Chen B, Zhang H. A novel fast model predictive control for large-scale structures. Journal of Vibration and Control. 2017;**23**(13):2190-2205. DOI: 10.1177/1077546315610033

[85] Takács G, Rohaľ-Ilkiv B. Model Predictive Vibration Control: Efficient Constrained MPC Vibration Control for Lightly Damped Mechanical Structures. Berlin, Germany: Springer Science & Business Media; 2012

[86] Goorts K, Narasimhan S. Adaptive model predictive control for deployable control systems with constraints. Journal of Structural Engineering. 2019;**145**(10):04019110. DOI: 10.1061/(ASCE)ST.1943-541X.0002392

[87] Hadjigeorgiou EP, Stavroulakis GE, Massalas CV. Shape control and damage identification of beams using piezoelectric actuation and genetic

optimization. International Journal of Engineering Science. 2006;**44**(7):409-421. DOI: 10.1016/j.ijengsci.2006.02.004

[88] Weeks CJ. Static shape determination and control of large space structures: I. The flexible beam. Journal of Dynamic Systems, Measurement, and Control. 1984;**106**(4):261-266. DOI: 10.1115/1.3140683

[89] Yang S, Ngoi B. Shape control of beams by piezoelectric actuators. AIAA Journal. 2000;**38**(12):2292-2298. DOI: 10.2514/2.898

[90] Yu Y, Zhang XN, Xie SL. Optimal shape control of a beam using piezoelectric actuators with low control voltage. Smart Materials and Structures. 2009;**18**(9):095006. DOI: 10.1088/0964-1726/18/9/095006

[91] Chaudhry Z, Rogers CA. Bending and shape control of beams using SMA actuators. Journal of Intelligent Material Systems and Structures. 1991;**2**(4):581-602. DOI: 10.1177/1045389X9100200410

[92] Agrawal BN, Treanor KE. Shape control of a beam using piezoelectric actuators. Smart Materials and Structures (Bristol, United Kingdom). 1999;**8**(6):729. DOI: 10.1088/0964-1726/8/6/303

[93] Chandrashekhara K, Varadarajan S. Adaptive shape control of composite beams with piezoelectric actuators. Journal of Intelligent Material Systems and Structures. 1997;**8**(2):112-124. DOI: 10.1177/1045389X9700800202

[94] Bendine K, Wankhade RL. Optimal shape control of piezolaminated beams with different boundary condition and loading using genetic algorithm. International Journal of Advanced Structural Engineering. 2017;**9**(4):375-384. DOI: 10.1007/s40091-017-0173-x

[95] Ganguli A, Jhawar S, Seshu P. Shape control of curved beams using piezoelectric actuators. In: Smart Materials, Structures, and Systems. Vol. 5062. San Diego, California, United States: SPIE; 2003. pp. 297-304. DOI: 10.1117/12.514406

[96] Furuya H, Haftka RT. Static shape control of space trusses with partial measurements. Journal of Spacecraft and Rockets. 1995;**32**(5):856-865. DOI: 10.2514/3.26696

[97] Kawaguchi K-I, Hangai Y, Pellegrino S, Furuya H. Shape and stress control analysis of prestressed truss structures. Journal of Reinforced Plastics and Composites. 1996;**15**(12):1226-1236. DOI: 10.1177/073168449601501204

[98] Manguri A, Saeed N, Szczepanski M, Jankowski R. Buckling and shape control of prestressable trusses using optimum number of actuators. Scientific Reports. 2023;**13**(1):3838. DOI: 10.1038/s41598-023-30274-y

[99] Matunaga S, Onoda J. Actuator placement with failure consideration for static shape control of truss structures. AIAA Journal. 1995;**33**(6):1161-1163. DOI: 10.2514/3.12540

[100] Mitsugi J, Yasaka T, Miura K. Shape control of the tension truss antenna. AIAA Journal. 1990;**28**(2):316-322. DOI: 10.2514/3.10391

[101] Saeed NM, Manguri AA, Szczepanski M, Jankowski R, Haydar BA. Static shape and stress control of trusses with optimum time, actuators and actuation. International Journal of Civil Engineering. 2023;**21**(3):379-390. DOI: 10.1007/s40999-022-00784-3

[102] Trak AB, Melosh RJ. Passive shape control of space antennas with truss support structures. Computers

& Structures. 1992;**45**(2):297-305. DOI: 10.1016/0045-7949(92)90413-T

[103] Xue Y, Wang Y, Xu X, Wan H-P, Luo Y, Shen Y. Comparison of different sensitivity matrices relating element elongations to structural response of pin-jointed structures. Mechanics Research Communications. 2021;**118**:103789. DOI: 10.1016/j.mechrescom.2021.103789

[104] Huang H, Li M, Yuan Y, Bai H. Experimental research on the seismic performance of precast concrete frame with replaceable artificial controllable plastic hinges. Journal of Structural Engineering. 2023;**149**(1):04022222. DOI: 10.1061/JSENDH.STENG-1164

[105] Gams M, Saje M, Planinc I, Kegl M. Optimal size, shape, and control design in dynamics of planar frame structures under large displacements and rotations. Engineering Optimization. 2010;**42**(1):69-86. DOI: 10.1080/03052150902998552

[106] Kameyama K, Shimoda M, Morimoto T. Shape identification for controlling the static deformation of frame structures. In: International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Vol. 46285. Buffalo, New York, USA: American Society of Mechanical Engineers; 2014. p. V01AT02A024. DOI: 10.1115/DETC2014-34265

[107] Chang X, Haiyan X. Using sphere parameters to detect construction quality of spherical buildings. In: 2nd International Conference on Advanced Computer Control. Shenyang, China: IEEE; 2010. DOI: 10.1109/ICACC.2010.5487073

[108] Wang K, Preumont A. Shape control of an adaptive spherical shell reflector under space environment. In: Eighth Symposium on Novel Photoelectronic Detection Technology and Applications. Vol. 12169. San Diego, California, United States: SPIE; 2022. pp. 2462-2467. DOI: 10.1117/12.2625829

[109] Zhang S, Du J, Duan B, Yang G, Ma Y. Integrated structural–electromagnetic shape control of cable mesh reflector antennas. AIAA Journal. 2015;**53**(5):1395-1399. DOI: 10.2514/1.J053726

[110] Song X, Tan S, Wang E, Wu S, Wu Z. Active shape control of an antenna reflector using piezoelectric actuators. Journal of Intelligent Material Systems and Structures. 2019;**30**(18-19):2733-2747. DOI: 10.1177/1045389X19873422

[111] Zhang H, Lu J, Lu M, Li N. Active control experiments on a Levy cable dome. Engineering Structures. 2023;**278**:115450. DOI: 10.1016/j.engstruct.2022.115450

[112] Dajian H, Cheng S. Construction control of the yamen cable-stayed bridge. In: Proceedings of Proceedings of the 3rd International Conference on Current and Future Trends in Bridge Design, Construction and Maintenance. Shanghai, China: Thomas Telford Limited; 2003. DOI: 10.1680/pot3icocaftibdcam.42018.0029

[113] Farahmand-Tabar S, Barghian M. Response control of cable-stayed arch bridge using modified hanger system. Journal of Vibration Control. 2020;**26**(23-24):2316-2328. DOI: 10.1177/1077546320921635

[114] Sharabash AM, Andrawes BO. Application of shape memory alloy dampers in the seismic control of cable-stayed bridges. Engineering Structures. 2009;**31**(2):607-616. DOI: 10.1016/j.engstruct.2008.11.007

[115] Shen LY, Li GQ, Luo YF. Displacement control of prestressed cable structures (in Chinese). Journal of Tongji University (Natural Science). 2006;**34**(3):291-295. Available from: http://en.cnki.com.cn/article_en/cjfdtotal-tjdz200603001.htm

[116] Tanaka H, Natori M. Shape control of cable-network structures based on concept of self-equilibrated stresses. JSME International Journal Series C. 2006;**49**(4):1067-1072. DOI: 10.1299/jsmec.49.1067

[117] Tanaka H, Natori MC. Shape control of space antennas consisting of cable networks. Acta Astronautica. 2004;**55**(3):519-527. DOI: 10.1016/0045-7949(92)90413-T

[118] Shon S, Kwan AS, Lee S. Shape control of cable structures considering concurrent/sequence control. Structural engineering and mechanics: An international journal. 2014;**52**(5):919-935. DOI: 10.12989/sem.2014.52.5.919

[119] You Z. Displacement control of prestressed structures. Computer Methods in Applied Mechanics and Engineering. 1997;**144**(1):51-59. DOI: 10.1016/S0045-7825(96)01164-4

[120] Schnellenbach-Held M, Steiner D. Self-tuning closed-loop fuzzy logic control algorithm for adaptive prestressed structures. Structural Engineering International. 2014;**24**(2):163-172. DOI: 10.2749/101686614X13830790993528

[121] Basu B, Bursi OS, Casciati F, Casciati S, Del Grosso AE, Domaneschi M, et al. A European Association for the Control of structures joint perspective. Recent studies in civil structural control across Europe. Structural Control and Health Monitoring. 2014;**21**(12):1414-1436. DOI: 10.1002/stc.1652

[122] Chatterjee S, Sarkar S, Hore S, Dey N, Ashour AS, Balas VE. Particle swarm optimization trained neural network for structural failure prediction of multistoried RC buildings. Neural Computing and Applications. 2017;**28**:2005-2016. DOI: 10.1007/s00521-016-2190-2

[123] Hait P, Sil A, Choudhury S. Seismic damage assessment and prediction using artificial neural network of RC building considering irregularities. Journal of Structural Integrity and Maintenance. 2020;**5**(1):51-69. DOI: 10.1080/24705314.2019.1692167

[124] Sepasdar R, Karpatne A, Shakiba M. A data-driven approach to full-field nonlinear stress distribution and failure pattern prediction in composites using deep learning. Computer Methods in Applied Mechanics and Engineering. 2022;**397**:115126. DOI: 10.1016/j.cma.2022.115126

[125] Abdulateef WS, Hejazi F. Fuzzy logic based adaptive vibration control system for structures subjected to seismic and wind loads. Structure. 2023;**55**:1507-1531. DOI: 10.1016/j.istruc.2023.06.108

[126] Casciati F, Faravelli L, Yao T. Control of nonlinear structures using the fuzzy control approach. Nonlinear Dynamics. 1996;**11**:171-187. DOI: 10.1007/BF00045000

[127] Faust B. Evaluation of the Residual Load-Bearing Capacity of Civil Structures Using Fuzzy-Logic and Decision Analysis. Universitätsbibliothek: Universtät der Bundeswehr Müchen; 2002

[128] Bektaş N, Kegyes-Brassai O. Development in fuzzy logic-based rapid

visual screening method for seismic vulnerability assessment of buildings. Geosciences. 2022;**13**(1):6. DOI: 10.3390/geosciences13010006

[129] Mei G, Kareem A, Kantor JC. Real-time model predictive control of structures under earthquakes. Earthquake engineering & structural dynamics. 2001;**30**(7):995-1019. DOI: 10.1002/eqe.49

[130] Riverso S, Mancini S, Sarzo F, Ferrari-Trecate G. Model predictive controllers for reduction of mechanical fatigue in wind farms. IEEE Transactions on Control Systems Technology. 2016;**25**(2):535-549. DOI: 10.1109/TCST.2016.2572170

[131] Krishnan S. Structural design and behavior of prestressed cable domes. Engineering Structures. 2020;**209**:110294. DOI: 10.1016/j.engstruct.2020.110294

[132] Senatore G, Reksowardojo A. Force and shape control strategies for minimum energy adaptive structures. Frontiers in Built Environment. 2020;**6**:105. DOI: 10.3389/fbuil.2020.00105

[133] Reksowardojo AP, Senatore G, Smith IF. Design of structures that adapt to loads through large shape changes. Journal of Structural Engineering. 2020;**146**(5):04020068. DOI: 10.1061/(ASCE)ST.1943-541X.0002604

[134] Veuve N, Sychterz AC, Smith IF. Adaptive control of a deployable tensegrity structure. Engineering Structures. 2017;**152**:14-23. DOI: 10.1016/j.engstruct.2017.08.062

[135] Mwafy A, Elnashai AS. Static pushover versus dynamic collapse analysis of RC buildings. Engineering Structures. 2001;**23**(5):407-424. DOI: 10.1016/S0141-0296(00)00068-7

[136] Cavagna L, Ricci S, Riccobene L. A fast tool for structural sizing, aeroelastic analysis and optimization in aircraft conceptual design. In: 50th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference 17th AIAA/ASME/AHS Adaptive Structures Conference 11th AIAA No. Sunrise Valley, United States: American Institute of Aeronautics and Astronautics (AIAA); 2009. p. 2571. DOI: 10.2514/6.2009-2571

[137] Adeli H. Neural networks in civil engineering: 1989-2000. Computer-Aided Civil and Infrastructure Engineering. 2001;**16**(2):126-142. DOI: 10.1111/0885-9507.00219

[138] Papadrakakis M, Lagaros ND, Tsompanakis Y. Structural optimization using evolution strategies and neural networks. Computer Methods in Applied Mechanics and Engineering. 1998;**156**(1-4):309-333. DOI: 10.1016/S0045-7825(97)00215-6

[139] Hajela P, Berke L. Neural networks in structural analysis and design: An overview. Computing Systems in Engineering. 1992;**3**(1-4):525-538. DOI: 10.1016/0956-0521(92)90138-9

[140] Yu G, Xiao L, Song W. Deep learning-based heterogeneous strategy for customizing responses of lattice structures. International Journal of Mechanical Sciences. 2022;**229**:107531. DOI: 10.1016/j.ijmecsci.2022.107531

[141] Ross TJ. Fuzzy Logic with Engineering Applications. Hoboken, New Jersey, United States: John Wiley & Sons; 2009

[142] Adeli H, Sarma KC. Cost Optimization of Structures: Fuzzy

Logic, Genetic Algorithms, and Parallel Computing. Hoboken, New Jersey, United States: John Wiley & Sons; 2006

[143] Benedettelli M. Optimization of building performance via model-based predictive control [PhD Dissertation]. Ancona, Italy: Università Politecnica Delle Marche; 2018. Available from: https://iris.univpm.it/retrieve/e18b8790-947c-d302-e053-1705fe0a27c8/Tesi_Benedettelli.pdf

[144] Kirsch U. Structural Optimization: Fundamentals and Applications. Berlin, Germany: Springer Science & Business Media; 2012

[145] Steven G, Li Q, Xie Y. Multicriteria optimization that minimizes maximum stress and maximizes stiffness. Computers & Structures. 2002;**80**(27-30):2433-2448. DOI: 10.1016/S0045-7949(02)00235-3

[146] Zhu J-H, Zhang W-H, Xia L. Topology optimization in aircraft and aerospace structures design. Archives of Computational Methods in Engineering. 2016;**23**:595-622. DOI: 10.1007/s11831-015-9151-2

[147] Raveh DE, Levy Y, Karpel M. Structural optimization using computational aerodynamics. AIAA Journal. 2000;**38**(10):1974-1982. DOI: 10.2514/2.853

[148] Sener M, Utku S, Wada BK. Geometry control in prestressed adaptive space trusses. Smart Materials and Structures. 1994;**3**(2):219. DOI: 10.1088/0964-1726/3/2/018

[149] Chen W-M, Wang D-J, Li M. Static shape control employing displacement–stress dual criteria. Smart Materials and Structures. 2004;**13**(3):468. DOI: 10.1088/0964-1726/13/3/003

[150] Xu X, Luo YZ. Multi-objective shape control of prestressed structures with genetic algorithms. Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering. 2008;**222**(8):1139-1147. DOI: 10.1243/09544100JAERO394

[151] Lu Q, Sun Y, Mei S. Nonlinear Control Systems and Power System Dynamics. Vol. 10. Berlin, Germany: Springer Science & Business Media; 2013

[152] Ohtori Y, Christenson R, Spencer B Jr, Dyke S. Benchmark control problems for seismically excited nonlinear buildings. Journal of Engineering Mechanics. 2004;**130**(4):366-385. DOI: 10.1061/(ASCE)0733-9399(2004)130:4(366)

[153] Lu Z, Wang Z, Zhou Y, Lu X. Nonlinear dissipative devices in structural vibration control: A review. Journal of Sound and Vibration. 2018;**423**:18-49. DOI: 10.1016/j.jsv.2018.02.052

[154] Worden K, Farrar CR, Haywood J, Todd M. A review of nonlinear dynamics applications to structural health monitoring. Structural Control and Health Monitoring: The Official Journal of the International Association for Structural Control and Monitoring and of the European Association for the Control of Structures. 2008;**15**(4):540-567. DOI: 10.1002/stc.215

[155] Iqbal J, Ullah M, Khan SG, Khelifa B, Ćuković S. Nonlinear control systems-a brief overview of historical and recent advances. Nonlinear Engineering. 2017;**6**(4):301-312. DOI: 10.1515/nleng-2016-0077

[156] Shaw IS. Fuzzy Control of Industrial Systems: Theory and Applications. Vol. 457. Cham, Switzerland: Springer; 2013

[157] Spencer B Jr, Ruiz-Sandoval ME, Kurata N. Smart sensing technology: Opportunities and challenges. Structural

Control and Health Monitoring. 2004;**11**(4):349-368. DOI: 10.1002/stc.48

[158] Sony S, Laventure S, Sadhu A. A literature review of next-generation smart sensing technology in structural health monitoring. Structural Control and Health Monitoring. 2019;**26**(3):e2321. DOI: 10.1002/stc.2321

[159] Nagayama T, Spencer BF Jr. Structural health monitoring using smart sensors. Newmark Structural Engineering Laboratory Report Series. Report No. NSEL-001. 2007. pp. 1-186

[160] Saeed NM. Recent advances in structural health monitoring: Techniques, applications and future directions. International Journal of Reliability and Safety. 2024;**18**(1):55-85. DOI: 10.1504/IJRS.2023.10061436

[161] Liu Y, Han Y, Li B. Geometric characteristics, deployment mechanisms, and digital fabrication methods of a free-form deployable membrane system based on CNC-knitted fabrics and CNC-bent frames. Journal of Asian Architecture and Building Engineering. 2023:1-16. DOI: 10.1080/13467581.2023.2267645

[162] Inoue F. Development of adaptive construction structure by variable geometry truss. In: Proceedings of International Symposium Shell and Spatial Structures-Architectural Engineering-Towards the Future Looking to the Past. Venice, Italy: IntechOpen; 2007. pp. 253-272. DOI: 10.5772/5543

[163] Sofla AYN, Elzey DM, Wadley HNG. Shape morphing hinged truss structures. Smart Materials and Structures. 2009;**18**(6):065012. DOI: 10.1088/0964-1726/18/6/065012

[164] Inoue F, Moroto R, Kurita K, Furuya N. Development of adaptive structure by variable geometry truss (application of movable monument in EXPO 2005). In: Proceedings of 23th International Symposium on Automation and Robotics in Construction. Tokyo, Japan: The International Association for Automation and Robotics in Construction (ISARC); 2006. pp. 704-709. DOI: 10.22260/ISARC2006/0131

[165] Del Grosso AE, Basso P. Adaptive building skin structures. Smart Materials and Structures. 2010;**19**(12):124011. DOI: 10.1088/0964-1726/19/12/124011

[166] Eren U, Prach A, Koçer BB, Raković SV, Kayacan E, Açıkmeşe B. Model predictive control in aerospace systems: Current state and opportunities. Journal of Guidance, Control, and Dynamics. 2017;**40**(7):1541-1566. DOI: 10.2514/1.G002507

[167] Giuliani M, Dimino I, Ameduri S, Pecora R, Concilio A. Status and perspectives of commercial aircraft morphing. Biomimetics. 2022;7(1):11. DOI: 10.3390/biomimetics7010011

[168] du Pasquier C, Shea K. Validation of a nonlinear force method for large deformations in shape-morphing structures. Structural and Multidisciplinary Optimization. 2022;**65**(3):87. DOI: 10.1007/s00158-022-03187-z

Chapter 3

A Type-2 Fuzzy State Observer Model for Non-Stationary Dynamic System Identification: An Incremental Learning Method with Noise Handling

Anderson Pablo Freitas Evangelista
and Ginalber Luiz de Oliveira Serra

Abstract

Real-world identification involves dealing with challenges such as system complexity, noise, and uncertainties. In this context, a method for incremental learning is suggested, utilizing an evolving type-2 state observer fuzzy model. The process involves structure learning through an evolving type-2 multiscaling clustering approach, eliminating the need for data normalization. The estimation of linear state observer models for each rule is achieved using observer Markov parameters computed via a Type-2 Instrumental Variable (T2-IV) algorithm. For obtaining the instruments for the T2-IV algorithm, a recursive moving-average filter is used. Benchmark and online identification tasks are conducted to demonstrate the practicality and robustness of the proposed methodology, with performance comparisons against existing methodologies.

Keywords: type-2 fuzzy state-space modeling, incremental type-2 fuzzy learning, multidimensional learning approach, markov parameters, type-2 instrumental variables

1. Introduction

Machine learning-based identification system is a relevant approach for modeling nonlinear, uncertain, multivariable, and complex systems. This approach aims to estimate a model that represents accurately the dynamic behavior of a real plant [1]. In these terms, fuzzy identification arises as a relevant method for obtaining models which represent nonlinear systems. These techniques are shown a powerful tool in practical problems such as uncertainty, unpredictable dynamics, and noisy measurements. One reason is that fuzzy logic systems (FLS) have the ability to integrate information from different sources, such as physical laws, empirical models, or

measurements [2]. For identification of problems with white noise signals, the classical fuzzy sets (type-1) present satisfactory results. However, when colored noise is considered, type-1 fuzzy sets are not able to mitigate the noise effects. In this case, type-2 fuzzy systems were proposed. The first mention of type-2 fuzzy sets as an extention of classical fuzzy sets was made by Zadeh [3], where theoretical concepts were addressed in the 1990s by Karnik [4]. In practical application, the interval type-2 fuzzy sets gained prominence in problems such as control [5] and modeling [6], as it presents less computational load. However, the development of methodologies for the experimental data analysis in order to obtain the rule-base for an interval type-2 fuzzy model in order to use the advantages of type-2 fuzzy sets described in the literature is still an open research field. In the literature, approaches such as heuristic methods [7] and incremental learning [8] have been used for this purpose. In Aissa Bencherif and Fatima Chouireb's work [9], an incremental learning algorithm for type-2 recurrent Takagi-Sugeno neural-fuzzy network is proposed. For structure learning, the rule firing strength-based approach is used. For a new data, the type-2 firing strength is computed for each rule, where the rule with the highest firing strength is considered for creation rule mechanism. The parameter update is perfomed by gradient descent algorithm. The mobile robot trajectory tracking problem is used to show the applicability of the methodology. In [10], Morteza Montazeri-Gh and Shabnam Yazdani introduce the use of interval type-2 fuzzy logic systems for gas turbine fault diagnosis, aiming to reduce maintenance costs and downtime. Fuzzy Rule Base is estimated using Interval Type-2 Fuzzy C-Means clustering, and parameters of the IT2FLSs are optimized with a metaheuristic algorithm. The performance of the IT2FL-based FDI system is compared to other classification techniques, showing promising results in terms of online applicability, accuracy, and robustness.

In literature, linear models are commonly used in consequent part in Takagi-Sugeno models, such as vector auto-regressive and state-space models. The state-space models present an interesting feature: a compact formula that that shows the relationship between internal variables and the experimental data (output and input signals) [11]. In this context, methodologies based on fuzzy state-space models have been proposed [12, 13]. In Gil et al.'s work [14], a recurrent state-space neural-fuzzy network is introduced. For parameter adjustment of the antecedent/consequent parts, a recursive learning method based on the constrained unscented Kalman filter is employed. The applicability of this methodology is demonstrated through the online identification of a three-tank system. In Yancho et al.'s work [15], a fuzzy state-space model predictive control approach is proposed. The learning algorithm is founded on gradient descent, which is used to fit the modeling structure parameters. The trained model is subsequently applied in model-based predictive control. The applicability of this methodology is demonstrated through computational experiments.

In identification problems, efficiency in mitigating the effects of noise to adjust the consequent parameters must be ensured. In a noisy environment, the Instrumental Variable (IV) method is considered a relevant tool for system modeling [11]. When compared to other identification methods, it is noted that, in the IV method, the requirement for an accurate noise model is not essential [16]. According to literature, the accuracy of the IV method relies on the selection of an appropriate instrument, which must guarantee non-polarized estimation [11]. The fuzzy version of the IV method was introduced by Yancho. Therefore, with the aim of integrating the type-2 state-space fuzzy modeling and non-polarized consequent estimation, in this paper, an incremental learning for evolving interval type-2 state observer fuzzy model based on instrumental variables approach is proposed. For estimating the consequent-part,

fuzzy observer Markov parameters are computed via type-2 fuzzy version of instrumental variable (T2-IV) identification algorithm. The observer Markov parameters are then used to compute the local system states at the current instant, which are subsequently used to compute the matrices of the local linear state observer model.

1.1 Contributions

The proposed methodology presents the following main contributions:

- Proposal of an evolving interval type-2 fuzzy state observer modeling approach with non-polarizing consequent estimation.
- Proposal for a novel composition of type-2 fuzzy rules for estimating an uncertain region. The adjustment of the uncertain region is accomplished through a proportional-integral-based adaptation rule.
- Structure learning based on the multiscale approach, where the data normalization is not required in clustering algorithm.
- Novel state-space online nonlinear identification based on interval type-2 fuzzy observer Markov parameters.

2. Overview of interval type-2 state observer fuzzy model

The proposed methodology is based on an evolving interval type-2 state observer fuzzy model (eIT2-SOFM), where its rules can be described as follows:

$$\textbf{Rule}^i: \textbf{IF } z_{1,k} \text{ is } \tilde{Z}_1^i \textbf{ AND} \cdots \textbf{AND } z_{n_z,k} \text{ is } \tilde{Z}_{n_z}^i \quad \textbf{THEN} \quad \begin{cases} \mathbf{x}_{k+1}^i = \mathbf{A}^i \mathbf{x}_k^i + \mathbf{B}^i \mathbf{u}_k + \mathbf{K}^i \mathbf{e}_k \\ \mathbf{y}_k^i = \mathbf{C}^i \mathbf{x}_k^i + \mathbf{D}^i \mathbf{u}_k \end{cases} \tag{1}$$

where i = 1, 2..., c represents the rule number, and $z_{1,k}, z_{2,k}, \ldots, z_{n_z,k}$ correspond to the antecedent input variables, where n_z is the number of antecedent variables. Additionally, $\mathbf{A}^i \in \Re^{n \times n}$, $\mathbf{B}^i \in \Re^{n \times m}$, $\mathbf{C}^i \in \Re^{p \times n}$, $\mathbf{D}^i \in \Re^{p \times m}$, and $\mathbf{K}^i \in \Re^{n \times p}$ represent the state-space matrices of the local linear model for each rule. The local state vector for the i-th rule is denoted as $\mathbf{x}_k^i = \left[x_{1,k}^i, x_{2,k}^i, \ldots, x_{n,k}^i\right] \in \Re^n$, and the local output vector is $\mathbf{y}_k^i = \left[y_{1,k}^i, y_{2,k}^i, \ldots, y_{p,k}^i\right] \in \Re^p$. The input signal vector is represented by $\mathbf{u}k = [u1, k, u_{2,k}, \ldots, u_{m,k}] \in \Re^m$, and $\mathbf{e}_k \in \Re^p$ is the error vector given by

$$\mathbf{e}_k = \mathbf{y}_k - \hat{\mathbf{y}}_k \tag{2}$$

where $\hat{\mathbf{y}}_k$ is the type-0 eIT2-SOFM output estimation in the eIT2-SOFM. An interval type-2 Gaussian membership function is adopted, represented as $\tilde{\mu}_j^i = \left[\overline{\mu}_j^i, \underline{\mu}_j^i\right]$, with uncertain dispersion denoted as $\tilde{\sigma} = \left[\overline{\sigma}^i, \underline{\sigma}^i\right]$, described by

$$\overline{\mu}_j^i(z_{j,k}) = \exp\left[-\frac{1}{2}\left(\frac{z_j^{i,*} - z_j}{\overline{\sigma}_j^i}\right)^2\right] \tag{3}$$

$$\underline{\mu}_j^i(z_{j,k}) = \exp\left[-\frac{1}{2}\left(\frac{z_j^{i,*} - z_j}{\underline{\sigma}_j^i}\right)^2\right] \tag{4}$$

where $\overline{\mu}_j^i$ and $\underline{\mu}_j^i$ are the upper and lower membership functions, respectively, $\overline{\sigma}_j^i$ is the upper dispersion, $\underline{\sigma}_j^i$ is the lower dispersion, and $z_j^{i,*}$ is the center of i-th cluster and j-th input axis. The proposed eIT2-SOFM adopts interval output estimation $\tilde{\mathbf{y}}_k = \left[\overline{\mathbf{y}}_k, \underline{\mathbf{y}}_k\right]$ as the output model. To compute the eIT2-SOFM output, first, the interval firing strength $\left[\overline{f}_k^i, \underline{f}_k^i\right]$ is computed as follows:

$$\overline{f}_k^i = \prod_{j=1}^{n_z} \overline{\mu}_j^i(z_{j,k}) \tag{5}$$

$$\underline{f}_k^i = \prod_{j=1}^{n_z} \underline{\mu}_j^i(z_{j,k}) \tag{6}$$

and from interval firing strength, the upper and lower normalized firing strength $\left[\overline{\gamma}_k^i, \underline{\gamma}_k^i\right]$ is computed as follows:

$$\overline{\gamma}_k^i = \frac{\overline{f}_k^i}{\sum_{j=1}^{c_k} \overline{f}_k^j} \tag{7}$$

$$\underline{\gamma}_k^i = \frac{\underline{f}_k^i}{\sum_{j=1}^{c_k} \underline{f}_k^j} \tag{8}$$

and $\left[\mathbf{y}_k^r, \mathbf{y}_k^l\right]$ is computed as follows:

$$\mathbf{y}_k^r = \sum_{i=1}^{c_k} \overline{\gamma}^i{}_k \mathbf{y}_k^i \tag{9}$$

$$\mathbf{y}_k^l = \sum_{i=1}^{c_k} \underline{\gamma}_k^i \mathbf{y}_k^i \tag{10}$$

Thus, in the function of $\left[\mathbf{y}_k^r, \mathbf{y}_k^l\right]$, the upper and lower outputs are computed using the following equations:

$$\overline{\mathbf{y}}_k = \max\left(\mathbf{y}_k^r, \mathbf{y}_k^l\right) + \dot{\mathbf{y}}_k \tag{11}$$

$$\underline{\mathbf{y}}_k = \min\left(\mathbf{y}_k^r, \mathbf{y}_k^l\right) - \dot{\mathbf{y}}_k \tag{12}$$

where $\dot{\mathbf{y}}_k$ is the adaptive output degree of uncertainty, which is adjusted based on the digital PI control algorithm, given by

DOI: http://dx.doi.org/10.5772/intechopen.1004751

$$\dot{\mathbf{y}}_k = \begin{cases} \dot{\mathbf{y}}_k \cdot f_f & \text{for } \underline{\mathbf{y}}_k \le \mathbf{y}_k\big|_f \le \overline{\mathbf{y}}_k \\ \dot{\mathbf{y}}_k + g_p|\mathbf{e}_k| + g_i\left|\sum_{j=k-w}^{k}\mathbf{e}_k\right| & \text{otherwise} \end{cases} \tag{13}$$

such that

$$\mathbf{e}_k = \mathbf{y}_k - \frac{\overline{\mathbf{y}}_k + \underline{\mathbf{y}}_k}{2} \tag{14}$$

where $1 \le f_f \le 0.90$ is a adjustment factor, g_p and g_i are the proportional gain and integral gain, respectively, w is the window size, and $\mathbf{y}_k\big|_f$ is the filtered output in instant k computed in the filtering process (Section 3). The proposed incremental learning is performed by the following steps: 1) filtering process, 2) structure learning via evolving method, and 3) submodels updating via type-2 state observer fuzzy identification. In **Figure 1**, the block diagram of the proposed methodology is shown, where $\mathbf{y}_k\big|_r$ is the corrupted output data in instant k. In the next sections, the mathematical formulation of each step is presented.

3. Filtering process

Consider a dynamic system where its experimental data are corrupted by correlated noise. For a data-driven learning algorithm, the noise in the database is a problem,

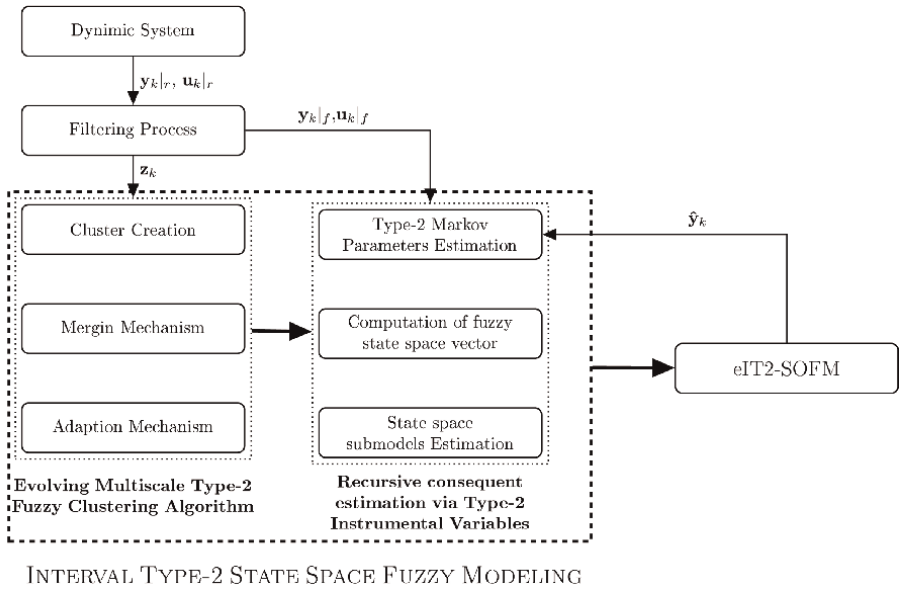


Figure 1.
Block diagram of the proposed methodology. From the dynamic system, output $\mathbf{y}_k\big|_r$ and input $\mathbf{u}_k$ data are obtained and filtered. For the evolving process, the antecedent input vector $\mathbf{z}_k$ is generated from the filtered input and/or output, that is, $\mathbf{z}_k = \left[y_{k-1}\big|_f \cdots y_{k-l}\big|_f \;\; u_{k-1}\big|_f \cdots j_{k-l}\big|_f\right]$. From vector $\mathbf{z}_k$, the evolving mechanism (creation, type-2 adaptation and merging) performs the structure learning, which chances the number of rules in each incoming data. In the sequel, the submodel of each rule is updated by a type-2 fuzzy state observer identification algorithm.

once inconsistent cluster (fuzzy partition) and polarized submodel parameter can be computed. Therefore, in the proposed methodology, a filtering process is performed in order to compute data highly correlated with system dynamic and independent of noise. In this step, a recursive moving-average filter are used, where

$$y_k\Big|_f = \left(1-a_f\right)^4 y_k\Big|_r + 4a_f y_{k-1}\Big|_f - 6a_f^2 y_{k-2}\Big|_f + 4a_f^3 y_{k-3}\Big|_f - a_f^4 y_{k-4}\Big|_f \tag{15}$$

where $a_f \in (0,1)$ is the filtering coefficient chosen by the user. For input data, the filtering process is similar, as follows:

$$u_k\Big|_f = \left(1-a_f\right)^4 u_k\Big|_r + 4a_f u_{k-1}\Big|_f - 6a_f^2 u_{k-2}\Big|_f + 4a_f^3 u_{k-3}\Big|_f - a_f^4 u_{k-4}\Big|_f \tag{16}$$

From $y_k\big|_f$ and $u_k\big|_f$, the vector $\mathbf{z}_k$ is generated, which is used in structure learning, and the regressor vector $\boldsymbol{\delta}_k$, which is used for consequent estimation.

4. Structure learning via evolving type-2 fuzzy clustering method

The structure of the eIT2-SOFM is updated with each new incoming dataset, and the adopted learning method does not necessitate prior knowledge. In other words, the rule base initializes with zero rules. The structure learning relies on an evolving type-2 fuzzy clustering (eT2FC), which is employed to create a fuzzy partition in the input variable space. This clustering method projects an interval type-2 fuzzy set onto each input space axis, characterized by interval type-2 Gaussian membership functions with uncertain dispersion. The eT2FC algorithm is based on a multidimensional scaling approach, eliminating the need for data normalization and providing improved handling of non-stationary problems [17].

Initially, for instant $k = 1$ and the number of rules $c_k = 0$, the antecedent input vector $\mathbf{z}_k = \left[z_{1,k}\, z_{2,k} \cdots z_{n_z,k}\right] \in \Re^{n_z}$ becomes the center of the first cluster, with an initial type-1 dispersion σ_0 defined by the user. The dispersions $\overline{\sigma}_j^i$ and $\underline{\sigma}_j^i$ are computed as a function of σ^i as follows:

$$\overline{\sigma}_j^i = \sigma + \zeta\sigma \tag{17}$$

$$\underline{\sigma}_j^i = \sigma - \zeta\sigma \tag{18}$$

For $k > 1$, the interval type-2 membership values $\tilde{\mu}_j^i = \left[\overline{\mu}_j^i, \underline{\mu}_j^i\right]$ are computed using Eqs. (3) and (4), and the interval firing strengths $\left[\overline{f}_k^i, \underline{f}_k^i\right]$ are calculated using Eqs. (5) and (6). The mean between $\overline{f}^i$ and $\underline{f}^i$ is computed by

$$f_k^i = \frac{\overline{f}_k^i + \underline{f}_k^i}{2} \tag{19}$$

and it is used for the cluster creation (rule creation) mechanism, which described in the sequel.

4.1 Cluster creation rule

To determine the necessity of creating a new cluster, initially, examine the cluster χ with the highest membership value, that is,

$$\chi = \arg\max_{i \in [1, c_k]} f_k^i \tag{20}$$

Therefore, the condition for rule creation is defined as follows:

$$\textbf{IF} f_k^{\chi} < T_f \textbf{ THEN } \mathbf{z}^{c_k+1*} = \mathbf{z}_k \tag{21}$$

where T_f represents the firing strength threshold. When the condition for rule creation is satisfied, the vector $\mathbf{z}_k$ becomes the center of a new cluster (cluster $c_k + 1$). The type-1 dispersion for the new cluster is determined by

$$\sigma_j^{c_k+1} = \alpha \left| z_{j,k} - z_{j,k}^{\varepsilon,*} \right| \tag{22}$$

where dispersion $\overline{\sigma}_j^i$ and $\underline{\sigma}_j^i$ are computed by Eqs. (17) and (18), respectively.

4.2 Merging mechanism

Once the creation rule (21) is satisfied, the merging condition is checked. This mechanism verifies if the new membership function $\tilde{\mu}_j^{c_k+1}$ is redundant. First, it determines the closest membership function to $\tilde{\mu}_j^{c_k+1}$, i.e.,

$$\varepsilon = \arg\max_{i \in [1, c_k]} \exp\left[-\frac{1}{2} \left(\frac{z_j^{i,*} - z_j}{\sigma_j^i} \right)^2 \right] \tag{23}$$

where $i \neq c_k + 1$ and ε is the index of the closest membership function. Therefore, for the membership functions $c_k + 1$ and ε along the j-th axis, the similarity degree is verified as follows:

$$S\left(z_j^{c_k+1,*}, z_j^{\varepsilon,*} \right) = \max\left(\mu_j^{c_k+1}\left(z_j^{\varepsilon,*} \right), \mu_j^{\varepsilon}\left(z_j^{c_k+1,*} \right) \right) \tag{24}$$

where $\mu_j^{c_k+1}\left(z_j^{\chi,*} \right)$ and $\mu_j^{\chi}\left(z_j^{c_k+1,*} \right)$ are computed as follows:

$$\mu_j^{c_k+1}\left(z_j^{\varepsilon,*} \right) = \exp\left[-\frac{1}{2} \left(\frac{z_j^{\varepsilon,*} - z_j^{c_k+1,*}}{\sigma_j^{\varepsilon}} \right)^2 \right] \tag{25}$$

$$\mu_j^{\varepsilon}\left(z_j^{c_k+1,*} \right) = \exp\left[-\frac{1}{2} \left(\frac{z_j^{c_k+1,*} - z_j^{\varepsilon,*}}{\sigma_j^{c_k+1}} \right)^2 \right] \tag{26}$$

From a upper threshold T_u and lower threshold T_l defined by user, the following conditions are verified:

1. If $S > T_u$, the new membership function is replaced by $\mu_j^{c_k+1}\left(z_j^{\epsilon,*}\right)$.

2. If $T_l < S < T_u$, the two membership function must be merged.

3. If $S < T_l$, The new membership function is maintained

If condition 2 is satisfied, the following equations are used for computing the new center and new dispersion

$$z^{new} = \frac{z_j^{c_k+1} + N_p^i z_j^i}{1 + N_p^i} \tag{27}$$

$$\sigma^{new} = \frac{\sigma_j^{c_k+1} + \sigma_j^i}{\sqrt{\pi}} \tag{28}$$

where N_p^i is the number of points $(\mathbf{z})$ associated with the cluster i.

4.3 Cluster adaptation mechanism

In the antecedent parameters adaptation, the approach adopted is to update the cluster center with the highest membership value when the new cluster condition is not satisfied. Thus, the updating of the center $\mathbf{z}^{\chi,*}$ is given by

$$\Delta z = \frac{z_{j,k}^{\chi,*} N_p^{\chi}}{N_p^{\chi} + 1} + \frac{z_{j,k}}{N_p^{\chi} + 1} \tag{29}$$

$$z_{j,k+1}^{\chi,*} = z_{j,k}^{\chi,*} + \rho\left(\Delta z - z_{j,k}^{\chi,*}\right) \tag{30}$$

where ρ is a learning rate defined by user and Δz is the center adjustment.

5. Submodels updating via IV-based type-2 fuzzy state observer identification

The state-space equations are regarded as the consequent part in the proposed eIT2-SOFM. The estimation of the matrices $\mathbf{A}^i, \mathbf{B}^i, \mathbf{C}^i, \mathbf{D}^i,$ and $\mathbf{K}^i$ is based on the fuzzy observer Markov parameters, where its mathematical foundations for the type-1 version are detailed in the works of [13, 18, 19]. This paper presents a type-2 state-space fuzzy modeling approach that utilizes the observer Markov parameters estimated through the IV fuzzy method. The subspace approach is employed to estimate the matrices $\mathbf{A}^i, \mathbf{B}^i, \mathbf{C}^i, \mathbf{D}^i,$ and $\mathbf{K}^i$ from the observer Markov parameters of each rule.

5.1 Mathematical definition of type-2 fuzzy observer markov parameters

Considering a vector auto-regressive model to estimate the local state-space model, as follows:

$$\mathbf{y}_k^i = \sum_{j=0}^{q_p} \dot{\boldsymbol{\Xi}}_{k-j}^{i,(u)} \mathbf{u}_{k-j} + \sum_{j=1}^{q_p} \dot{\boldsymbol{\Xi}}_{k-j}^{i,(y)} \mathbf{y}_{k-j}^i \tag{31}$$

According to [20], if the local state-space model is asymptotically stable, the matrices $\dot{\boldsymbol{\Xi}}_{k-j}^{i,(u)}$ and $\dot{\boldsymbol{\Xi}}_{k-j}^{i,(y)}$ in Eq. (31) are the observer Markov parameters of i-th local state-space model, being

$$\dot{\boldsymbol{\Xi}}_{k-j}^{i,(u)} = \begin{cases} \mathbf{D}^i & \text{if } j = 0 \\ \mathbf{C}^i \left(\mathbf{A}^i\right)^{j-1} \mathbf{B}^i & \text{if } j > 0 \end{cases} \tag{32}$$

$$\dot{\boldsymbol{\Xi}}_{k-j}^{i,(y)} = \mathbf{C}^i \left(\mathbf{A}^i\right)^{j-1} \mathbf{K}^i \tag{33}$$

where $\mathbf{K}^i \in \Re^{n \times p}$ is the local observer matrix [13]. Thus, the matrix composed by the observer Markov parameters matrix is given by

$$\dot{\boldsymbol{\Xi}}^i = \left[\dot{\boldsymbol{\Xi}}_{k-q_p}^{i,(u)}, \ \cdots, \ \dot{\boldsymbol{\Xi}}_k^{i,(u)}, \ \dot{\boldsymbol{\Xi}}_{k-q_p}^{i,(y)} \ \cdots \ \dot{\boldsymbol{\Xi}}_{k-1}^{i,(y)}\right] \tag{34}$$

and Eq. (31) is rewritten in matrix form as follows:

$$\left(\mathbf{y}_k^i\right)^T = \left(\boldsymbol{\delta}_k^i\right)^T \left(\dot{\boldsymbol{\Xi}}^i\right)^T + \boldsymbol{\xi}_k \tag{35}$$

with $\left(\boldsymbol{\delta}_k^i\right)^T = \left[\breve{\mathbf{u}}_{k-q_p}^T \ \ \mathbf{u}_k^T \ \ \left(\breve{\mathbf{y}}_{k-q_p}^i\right)^T\right]$, such that

$$\breve{\mathbf{u}}_{k-q_p} = \begin{bmatrix} \mathbf{u}_{k-q_p} \\ \mathbf{u}_{k-q_p+1} \\ \vdots \\ \mathbf{u}_{k-1} \end{bmatrix} \quad \breve{\mathbf{y}}_{k-q_p}^i = \begin{bmatrix} \mathbf{y}_{k-q_p}^i \\ \mathbf{y}_{k-q_p-1}^i \\ \vdots \\ \mathbf{y}_{k-1}^i \end{bmatrix} \tag{36}$$

Assuming k samples, where q_p is the past time-window and $k > q_p$, from Eq. (35), the following batch equation is derived:

$$\mathbf{Y}_k^i = \boldsymbol{\Delta}_k^i \left(\dot{\boldsymbol{\Xi}}^i\right)^T \tag{37}$$

where

$$\mathbf{Y}_k^i = \begin{bmatrix} \left(\mathbf{y}_{q_p+1}^i\right)^T \\ \left(\mathbf{y}_{q_p+2}^i\right)^T \\ \vdots \\ \left(\mathbf{y}_k^i\right)^T \end{bmatrix}, \quad \boldsymbol{\Delta}_k^i = \begin{bmatrix} \left(\boldsymbol{\delta}_{q_p+1}^i\right)^T \\ \left(\boldsymbol{\delta}_{q_p+2}^i\right)^T \\ \vdots \\ \left(\boldsymbol{\delta}_k^i\right)^T \end{bmatrix} \tag{38}$$

Thus, considering the TSK fuzzy theory, the batch computation fuzzy model output history is given by

$$\mathbf{Y}_k = \sum_{i=1}^{c} \tilde{\mathbf{\Gamma}}_k^i \mathbf{\Delta}_k^i \left(\dot{\mathbf{\Xi}}^i\right)^T \tag{39}$$

where $\tilde{\mathbf{\Gamma}}_k^i = \text{diag}\left(\left[\gamma_{q_p+1}^i, \tilde{\gamma}_{q_p+2}^i, \cdots, \tilde{\gamma}_k^i\right]\right)$, so that $\tilde{\gamma}_{q_p+1}^i$ is type-2 normalized firing strength computed by

$$\tilde{\gamma}_k^i = \frac{\overline{f}_k^i + \underline{f}_k^i}{\sum_{j=1}^{c_k}\left(\overline{f}_k^j + \underline{f}_k^j\right)} \tag{40}$$

Thus, from the batch estimation approach, the outputs from instant q_p to k can be computed by:

$$\tilde{\mathbf{\Gamma}}_k^i \mathbf{Y}_k = \tilde{\mathbf{\Gamma}}_k^i \mathbf{\Delta}_k \left(\dot{\mathbf{\Xi}}^i\right)^T \tag{41}$$

where

$$\mathbf{Y}_k = \begin{bmatrix} \left(\mathbf{y}_{q_p+1}\right)^T \\ \left(\mathbf{y}_{q+2}\right)^T \\ \vdots \\ \left(\mathbf{y}_k\right)^T \end{bmatrix} \tag{42}$$

5.2 Fuzzy observer markov parameters estimation via IV approach

Assuming experimental data are corrupted by correlated noise, the vector $\boldsymbol{\delta}_k$ presents noisy data, that is,

$$\boldsymbol{\delta}_k^r = \boldsymbol{\delta}_k^i + \boldsymbol{\upsilon}_k \tag{43}$$

where $\boldsymbol{\upsilon}_k$ is the correlated noise vector related to $\boldsymbol{\delta}_k$.

According to literature [21, 22], from an instrument vector $\left(\boldsymbol{\delta}_k^f\right)^T$, which is highly correlated with output and/or input data and not correlated with the noise, the following solution for $\dot{\mathbf{\Xi}}^i$, from Eq. (41), is derived:

$$\left(\dot{\mathbf{\Xi}}^i\right)^T = \left(\mathbf{\Delta}_k^f \tilde{\mathbf{\Gamma}}_k^i \mathbf{\Delta}_k^r\right)^{-1} \mathbf{\Delta}_k^f \tilde{\mathbf{\Gamma}}_k^i \mathbf{Y}_k^f \tag{44}$$

where $\mathbf{\Delta}_k^f \in \mathfrak{R}^{\left(k-q_p-1\right)\times\left(q_p(m+p)+m\right)}$ is the batch instruments matrix. Extended the recursive type-1 fuzzy IV algorithm presented in [21, 22] for the type-2 case, the updation of the interval type-2 fuzzy observer Markov parameters is performed by following equations:

$$\mathbf{L}_{k+1}^i = \tilde{\gamma}_k^i \frac{\mathbf{P}_k^i \boldsymbol{\delta}_k^f}{\tilde{\gamma}_k^i + \left(\boldsymbol{\delta}_k^r\right)^T \mathbf{P}_k^i \boldsymbol{\delta}_k^r} \tag{45}$$

$$\dot{\mathbf{\Xi}}_{k+1}^i = \dot{\mathbf{\Xi}}_k^i + \tilde{\gamma}_k^i \mathbf{e}_k \mathbf{L}_{k+1}^i \tag{46}$$

$$\mathbf{P}_{k+1}^{i} = \frac{1}{\beta}\left[\mathbf{I} - \mathbf{L}_{k+1}^{i}\left(\boldsymbol{\delta}_{k}^{r}\right)^{T}\right]\mathbf{P}_{k}^{i} \tag{47}$$

where $1 \leq \beta \leq 0.9$ is the forgetting factor, $\mathbf{L}$ is the IV gain matrix, and $\mathbf{P}$ is the IV variance matrix.

5.3 Computation of space state matrices

According to [23, 24], the local state vector $\mathbf{x}_k^i$ can be computed by the following close-formula:

$$\mathbf{x}_{k}^{i} = \mathbf{S}\left[\boldsymbol{\Lambda}_{k}^{i}\overline{\mathbf{u}}_{k-q_p}^{f} + \boldsymbol{\Upsilon}_{k}^{i}\overline{\mathbf{y}}_{k-q_p}^{f}\right] \tag{48}$$

where $\mathbf{S} \in \Re^{n\times(mf-n)}$ is a positive-defined matrix, and $\boldsymbol{\Lambda}_k^i$ and $\boldsymbol{\Upsilon}_k^i$ are formed by the observer Markov parameters, such that

$$\boldsymbol{\Lambda}_{k}^{i} = \begin{bmatrix} \dot{\boldsymbol{\Xi}}_{k-q}^{i,(u)} & \dot{\boldsymbol{\Xi}}_{k-q_p+1}^{i,(u)} & \cdots & \dot{\boldsymbol{\Xi}}_{k-q_p+q_f-1}^{i,(u)} & \cdots & \dot{\boldsymbol{\Xi}}_{k-1}^{i,(u)} \\ \mathbf{0} & \dot{\boldsymbol{\Xi}}_{k-q_p}^{i,(u)} & \cdots & \dot{\boldsymbol{\Xi}}_{k-q_p+q_f-2}^{i,(u)} & \cdots & \dot{\boldsymbol{\Xi}}_{k-2}^{i,(u)} \\ \vdots & \ddots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \dot{\boldsymbol{\Xi}}_{k-q_p}^{i,(u)} & \cdots & \dot{\boldsymbol{\Xi}}_{k-q_f}^{i,(u)} \end{bmatrix} \tag{49}$$

and

$$\boldsymbol{\Upsilon}_{k}^{i} = \begin{bmatrix} \dot{\boldsymbol{\Xi}}_{k-q}^{i,(y)} & \dot{\boldsymbol{\Xi}}_{k-q_p+1}^{i,(y)} & \cdots & \dot{\boldsymbol{\Xi}}_{k-q_p+q_f-1}^{i,(y)} & \cdots & \dot{\boldsymbol{\Xi}}_{k-1}^{i,(y)} \\ \mathbf{0} & \dot{\boldsymbol{\Xi}}_{k-q_p}^{i,(y)} & \cdots & \dot{\boldsymbol{\Xi}}_{k-q_p+q_f-2}^{i,(y)} & \cdots & \dot{\boldsymbol{\Xi}}_{k-2}^{i,(y)} \\ \vdots & \ddots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \dot{\boldsymbol{\Xi}}_{k-q_p}^{i,(y)} & \cdots & \dot{\boldsymbol{\Xi}}_{k-q_f}^{i,(y)} \end{bmatrix} \tag{50}$$

Once the local state vectors $\mathbf{x}_k^i$ are computed, the fuzzy state vector $\tilde{\mathbf{x}}_k$ is obtained as follows:

$$\tilde{\mathbf{x}}_{k} = \sum_{i=1}^{c_k} \tilde{\gamma}_{k}^{i}\mathbf{x}_{k}^{i} \tag{51}$$

From Eq. (51), the matrices $\left[\mathbf{A}_k^i, \mathbf{B}_k^i, \mathbf{K}_k^i\right]$ can be estimated using the QR solution. Thus, the state equation is formulated as follows:

$$\left(\mathbf{x}_{k}^{i}\right)^{T} = \begin{bmatrix} \left(\mathbf{x}_{k-1}^{i}\right)^{T} & \mathbf{u}_{k-1}^{T} & \mathbf{e}_{k-1}^{T} \end{bmatrix} \begin{bmatrix} \left(\mathbf{A}_{k}^{i}\right)^{T} \\ \left(\mathbf{B}_{k}^{i}\right)^{T} \\ \left(\mathbf{K}_{k}^{i}\right)^{T} \end{bmatrix} = \left(\boldsymbol{\nu}_{k-1}^{x,i}\right)^{T}\left(\boldsymbol{\Theta}_{k}^{i,x}\right)^{T} \tag{52}$$

Let the least square solution of Eq. (52), given by:

$$\left(\boldsymbol{\Theta}_k^{i,x}\right)^T = \left(\mathbf{V}_k \tilde{\boldsymbol{\Gamma}}_k^i \mathbf{V}_k\right)^{-1} \mathbf{V}_k \tilde{\boldsymbol{\Gamma}}_k^i \mathbf{X}_k \tag{53}$$

$$\left(\boldsymbol{\Theta}_k^{i,x}\right)^T = \left(\mathbf{P}_k^{\boldsymbol{\Theta}_x^i}\Big|_{\nu\nu}\right)^{-1} \mathbf{P}_k^{\boldsymbol{\Theta}_x^i}\Big|_{\nu y^f} \tag{54}$$

where

$$\mathbf{X}_k = \begin{bmatrix} \mathbf{x}_{k-q_p}^T \\ \mathbf{x}_{k-q_p+1}^T \\ \vdots \\ \mathbf{x}_k^T \end{bmatrix} \tag{55}$$

Rewritten Eq. (53), it has

$$\mathbf{P}_k^{\boldsymbol{\Theta}_x^i}\Big|_{\nu\nu} \left(\boldsymbol{\Theta}^{i,x}\right)^T = \mathbf{P}_k^{\boldsymbol{\Theta}_x^i}\Big|_{\nu y^f} \tag{56}$$

where the following recursion is derived:

$$\mathbf{P}_k^{\boldsymbol{\Theta}_x^i}\Big|_{\nu\nu} = \mathbf{P}_{k-1}^{\boldsymbol{\Theta}_x^i}\Big|_{\nu\nu} + \tilde{\gamma}_k^i \boldsymbol{\nu}_{k-1} \left(\boldsymbol{\nu}_{k-1}^{x,i}\right)^T \tag{57}$$

$$\mathbf{P}_k^{\boldsymbol{\Theta}_x^i}\Big|_{\nu y} = \mathbf{P}_{k-1}^{\boldsymbol{\Theta}_x^i}\Big|_{\nu y} + \tilde{\gamma}_k^i \boldsymbol{\nu}_{k-1}^{x,i} \left(\mathbf{y}_{k-1}^f\right)^T \tag{58}$$

Thus, by applying QR factorization to $\mathbf{P}^{\boldsymbol{\Theta}_x^i} k|\nu\nu$, it has

$$\mathbf{R}\left(\boldsymbol{\Theta}^{i,x}\right)^T = \mathbf{Q}^T \mathbf{P}_k^{\boldsymbol{\Theta}_x^i}\Big|_{\nu y^f} \tag{59}$$

The matrix $\boldsymbol{\Theta}^{i,x}$ is computed through the backward substitution method in Eq. (59) [21]. To compute the $\left[\mathbf{C}_k^i, \mathbf{D}_k^i\right]$ matrices, the output equation can be formulated as follows:

$$\left(\mathbf{y}_k^i\right)^T = \begin{bmatrix} \left(\mathbf{x}_k^i\right)^T & \mathbf{u}_k^T \end{bmatrix} \begin{bmatrix} \left(\mathbf{C}_k^i\right)^T \\ \left(\mathbf{D}_k^i\right)^T \end{bmatrix} = \left(\boldsymbol{\nu}_{k-1}^{y,i}\right)^T \left(\boldsymbol{\Theta}_k^{i,y}\right)^T \tag{60}$$

Using the same steps for computing $\left(\boldsymbol{\Theta}_k^{i,x}\right)^T$, it has

$$\mathbf{P}_k^{\boldsymbol{\Theta}_y^i}\Big|_{\nu\nu} \left(\boldsymbol{\Theta}^{i,y}\right)^T = \mathbf{P}_k^{\boldsymbol{\Theta}_y^i}\Big|_{\nu y^f} \tag{61}$$

such that

$$\mathbf{P}_k^{\boldsymbol{\Theta}_y^i}\Big|_{\nu\nu} = \mathbf{P}_{k-1}^{\boldsymbol{\Theta}_y^i}\Big|_{\nu\nu} + \tilde{\gamma}_k^i \boldsymbol{\nu}_{k-1}^{y,i} \left(\boldsymbol{\nu}_{k-1}^{y,i}\right)^T \tag{62}$$

$$\mathbf{P}_k^{\boldsymbol{\Theta}_y^i}\Big|_{\nu y} = \mathbf{P}_{k-1}^{\boldsymbol{\Theta}_y^i}\Big|_{\nu y} + \tilde{\gamma}_k^i \boldsymbol{\nu}_{k-1}^{y,i} \left(\mathbf{y}_{k-1}^f\right)^T \tag{63}$$

where $\mathbf{\Theta}_k^{i,y}$ can be computed by applying QR factorization to $\mathbf{P}^{\mathbf{\Theta}_y^i}k|\nu\nu$, such that

$$\mathbf{R}\left(\mathbf{\Theta}^{i,y}\right)^T = \mathbf{Q}^T\mathbf{P}_k^{\mathbf{\Theta}_y^i}\Big|_{\nu y^f} \tag{64}$$

followed by backward substitution applied in Eq. (64).

6. Computational results

Considering the mathematical formulation of the proposed algorithm described in Sections 4 and 5, two case studies are presented: the identification of a SISO nonlinear system and online identification of a time-varying MIMO dynamic system. To validate the results, the following metrics were used:

- Non-dimensional error index (NDEI):

$$\text{NDEI} = \frac{\sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\tilde{e}_{r,k}\right)^2}}{std\left(\mathbf{y}\right)} \tag{65}$$

  where $std(\bullet)$ is the standard deviation.

- Variance accounted for (VAF%):

$$\text{VAF}(\%) = \left[1 - \frac{\tilde{\mathbf{e}}_r}{var\left(\mathbf{y}\right)}\right] \times 100 \tag{66}$$

  where $var(\bullet)$ is the variance.

  where $\tilde{e}_k$ is the confidence region error for interval type-2 estimation, which is described as follows:

$$\tilde{e}_k = \begin{cases} 0 & \text{if } \underline{y}_k < y_k < \hat{\bar{y}}_k \\ y_k - \hat{\bar{y}}_k & \text{if } \bar{e}_k < \underline{e}_k \\ y_k - \underline{y}_k & \text{otherwise} \end{cases} \tag{67}$$

such that

$$\bar{e}_k = |y_k - \bar{\hat{y}}_k|, \;\; \underline{e}_k = |y_k - \underline{\hat{y}}_k| \tag{68}$$

6.1 Nonlinear dynamic system

The identification problem under consideration is a SISO nonlinear dynamic system, commonly utilized as a benchmark in the type-2 fuzzy modeling literature. It is described by the following equation:

$$y_k = \frac{y_{k-1}y_{k-2}\left(y_{k-1} + 0.5\right)}{1 + y_{k-1}^2 + y_{k-2}^2} + u_{k-1} \tag{69}$$

where the input signal is given by $u_k = \sin\left(\frac{2k\pi}{25}\right)$. For the identification process, a dataset consisting of 1300 samples was generated. Among these samples, 1000 were allocated for the training step, while 300 were used for the validation step. The algorithm parameters were configured with the following values: $a_f = 0$; $T_f = 0.002$; $T_u = 0.7$; $T_l = 0.5$; $q_p = 11$; $q_f = 6$; $n = 2$; $w = 5$; $g_p = 10^{-2}$; $f_f = 0.99$; and $g_i = 10^{-3}$. The rule structure adopted in this experiment is

$$\textbf{Rule}^i : \textbf{IF } z_{1,k} \text{ is } \tilde{Z}_1^i \textbf{ AND } z_{2,k} \text{ is } \tilde{Z}_2^i \quad \textbf{THEN} \quad \begin{cases} \mathbf{x}_{k+1}^i = \mathbf{A}^i\mathbf{x}_k^i + \mathbf{B}^i\mathbf{u}_k + \mathbf{K}^i\mathbf{e}_k \\ \mathbf{y}_k^i = \mathbf{C}^i\mathbf{x}_k^i + \mathbf{D}^i\mathbf{u}_k \end{cases} \tag{70}$$

where $z_{1,k} = u_k$ and $z_{2,k} = y_k$.

For comparative analysis, the models eTS [25], xTS (cited in [26]), DENFIS [27], eTF [28], eMG [29], and RIV-NFM [22] are considered. The performances, as assessed by the **NDEI** metric, are presented in **Table 1**. **Figure 2** illustrates the uncertain region estimated by eIT2-SOFM for the validation dataset.

Model	Rules	NDEI
eTS [25]	7	0.1038
xTS cited in [26]	7	0.0936
DENFIS [27]	7	0.0842
eFT [28]	7	0.0653
eMG [29]	7	0.0501
RIV-NFM [22]	6	0.0413
Proposed	*3*	*0.0203*

Table 1.
Comparative analysis of the proposed methodology with other relevant methodologies for the nonlinear dynamic system problem.

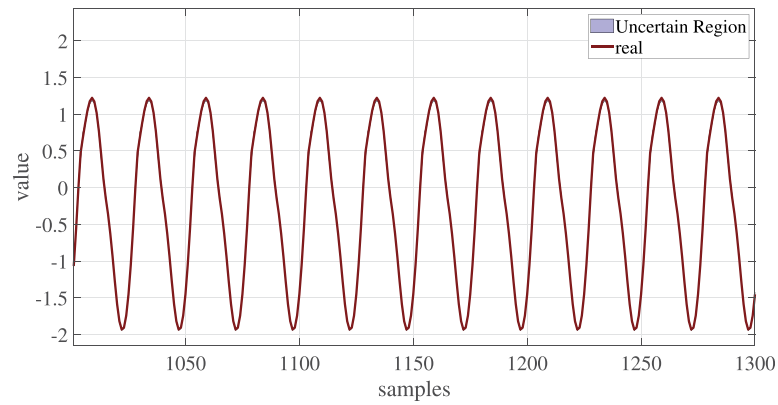


Figure 2.
Uncertain region estimation for nonlinear dynamic system identification.

6.2 Time-varying MIMO dynamic system

A time-varying nonlinear MIMO dynamic system is considered to demonstrate the adaptability of the proposed methodology for time-varying dynamic systems. The nonlinear MIMO dynamic system is described by the following equations:

$$\mathbf{v}_k = \begin{cases} v_{1,k+1} = \dfrac{v_{1,k-1}}{v_{1,k}^2 + 1} + \dfrac{1}{v_{2,k}^2 + 5} + u_{1,k}^3 \\ v_{2,k+1} = 0.1 v_{2,k} v_{1,k-1} - 0.2 v_{2,k} u_{2,k} \end{cases} \tag{71}$$

$$\mathbf{y}_k = \mathbf{G}_k \mathbf{v}_k \tag{72}$$

where

$$\mathbf{G}_k = \begin{cases} \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \text{for } k < 1400 \\ \begin{bmatrix} 3 & 2 \\ -1 & 5 \end{bmatrix} \text{otherwise} \end{cases} \tag{73}$$

and $\mathbf{u}_k = [u_{1,k} \;\; u_{2,k}]^T$, where $u_{1,k}$ is a multistep signal with a uniform distribution between $[-2, 2]$, $u_{2,k}$ is a multistep signal with a uniform distribution between $[-1\ 1]$, and $\mathbf{v}_k = [v_{1,k} \;\; v_{2,k}]^T$ represents the vector of intermediate signals. The outputs signals were corrupted by correlated noises, which are given by

$$\nu_k^{y_1} = \frac{1 + 0.2z^{-1}}{1 + 0.6z^{-1} + 0.2z^{-2}} e_k \tag{74}$$

$$\nu_k^{y_2} = \frac{1 + 0.2z^{-1}}{1 + 0.3z^{-1} + 0.1z^{-2}} e_k \tag{75}$$

where e_k represents white noise with a mean of zero and a variance of σ_e^2. The dataset comprises 2800 samples, with 500 used for initializing the eIT2-SOFM and 2400 samples used for online identification.

The algorithm parameters were set to the following values: $a_f = 0.9$; $T_f = 0.001$; $T_u = 0.7$; $T_l = 0.5$; $q_p = 12$; $q_f = 7$; $n = 4$; $w = 1$; $g_p = 5 \times 10^{-3}$; $f_f = 0.985$; and $g_i = 3 \times 10^{-5}$. The rule structure adopted in this experiment is

$$\begin{aligned} \mathbf{Rule}^i : \mathbf{IF}\ z_{1,k} \text{ is } \tilde{Z}_1^i \ \mathbf{AND}\ z_{2,k} \text{ is } \tilde{Z}_2^i \quad & \mathbf{AND}\ z_{3,k} \text{ is } \tilde{Z}_3^i \ \mathbf{AND}\ z_{4,k} \text{ is } \tilde{Z}_4^i \\ & \mathbf{THEN} \begin{cases} \mathbf{x}_{k+1}^i = \mathbf{A}^i \mathbf{x}_k^i + \mathbf{B}^i \mathbf{u}_k + \mathbf{K}^i \mathbf{e}_k \\ \mathbf{y}_k^i = \mathbf{C}^i \mathbf{x}_k^i + \mathbf{D}^i \mathbf{u}_k \end{cases} \end{aligned} \tag{76}$$

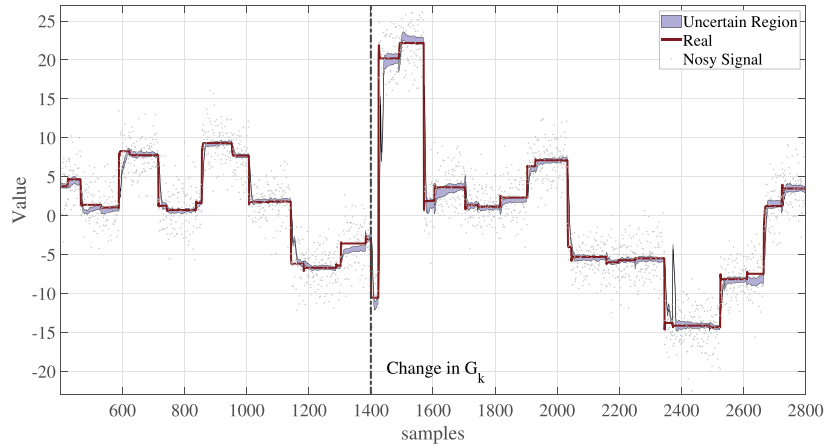
where $z_{1,k} = u_{1,k-1}$, $z_{2,k} = u_{2,k-1}$, $z_{3,k} = y_{1,k-1}$, and $z_{4,k} = y_{2,k-1}$.

In this case, the Monte Carlo method was employed, involving 50 experiment realizations to compute the means of the VAF% and NDEI criteria. The interval estimations of y_1 and y_2 according to the SNR variation, are shown in **Table 2**. The online estimations of the time-varying nonlinear MIMO dynamic system outputs, are shown in **Figure 3**, with an SNR of 10 dB.

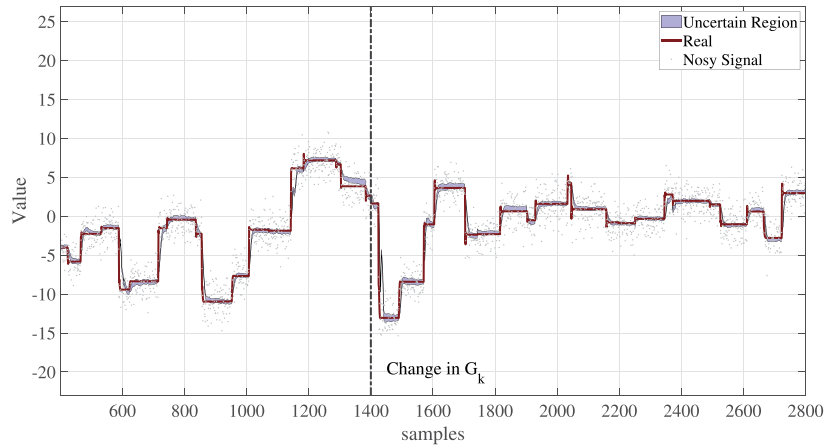
	VAF% (±std[a])		NDEI	
SNR	y_1	y_2	y_1	y_2
10	95.83 (±1.34%)	95.10 (±1.91%)	0.210	0.202
15	96.34 (±0.86%)	95.61 (±0.92%)	0.202	0.191
20	96.62 (±0.57%)	95.80 (±0.58%)	0.193	0.188
25	97.01 (±0.41%)	95.93 (±0.53%)	0.186	0.181
30	97.27 (±0.23%)	96.16 (±0.44%)	0.179	0.169

[a]*std: standard deviation.*

Table 2.
Estimation performance of the proposed methodology for different levels of noise for VAF and NDEI metrics in time-varying nolinear MIMO dynamic system modeling.



(a)



(b)

Figure 3.
Upper and lower estimation of time-varying nonlinear MIMO dynamic system: (a) y_1 and b) y_2. The SNR for this experiment was 10 dB. It is noted that the estimation accuracy of the proposed methodology even in noise environments. The purple region was estimated by the eIT2-SOFM based on the experimental data.

7. Discussions

In this paper, aspects of the proposed methodology were presented. The application of eIT2-SOFM for the nonlinear identification of SISO and MIMO dynamic systems was discussed. In Section 6.1, a modeling benchmark problem was used to compare the performance of eIT2-SOFM with other methodologies presented in the literature. Upon reviewing **Table 1**, it becomes evident that significantly improved results are achieved by the proposed methodology. Additionally, it is noteworthy that only three rules were created by eIT2-SOFM during the identification process, making it the model with the fewest number of rules among the compared methodologies. This result highlights the methodology's ability to track the nonlinear behavior of dynamic systems.

In Section 6.2, a case study involving a nonlinear MIMO system was presented to demonstrate the tracking capabilities of the proposed methodology in dealing with time-varying problems. The performance results, as shown in **Table 2**, indicate that the eIT2-SOFM achieved a performance exceeding 95% for each SNR value. This underlines the adaptability of the proposed learning algorithm, even in the presence of correlated noise within the dataset.

8. Conclusions

Considering the experimental results and the methodological aspects of the proposed modeling approach based on eIT2-SOFM, the following concluding remarks are made:

- The proposed method demonstrates robustness to outliers and noise through the incorporation of a filtering process, type-2 fuzzy sets, and the T2-IV algorithm. The filtering process precedes the structure learning step to prevent the creation of nonrelevant rules, while the T2-IV algorithm provides a nonpolarized estimation of the local state observer model parameters.

- Numerical robustness is ensured, as the QR-decomposition is applied to compute the local state observer models.

- The computational results have demonstrated that the proposed methodology is effective for modeling complex dynamic systems characterized by uncertainty, nonlinearity, and both single and multivariable aspects, even in the presence of colored noise.

Among practical projects and problems that can be solved by the algorithm, the following has been widely considered for research:

- Black-box model-based control, where the plant presents nonlinearity, uncertain behavior, and correlated noise, as satellite positioning [30], multimobile manipulator, and induction motor.

- Computational modeling of experimental data, where the data are nonlinear, uncertain, and/or corrupted by correlated noise, such as parameter estimation of vehicle dynamics, mechatronic systems, mobile robot navigation, and nonstationary processes [31].

Author details

Anderson Pablo Freitas Evangelista[1*†] and Ginalber Luiz de Oliveira Serra[2†]

1 Federal Institute of Education, Science and Tecnology of Maranhao, Açailândia, MA, Brazil

2 Federal Institute of Education, Science and Tecnology of Maranhao, Sao Luis, MA, Brazil

*Address all correspondence to: anderson.evangelista@ifma.edu.br

† These authors contributed equally.

IntechOpen

© 2024 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

References

[1] Serra G, Bottura C. An IV-QR algorithm for neuro-fuzzy multivariable online identification. IEEE Transactions on Fuzzy Systems. 2007;**15**(2):200-210

[2] Babuska R. Fuzzy Modeling for Control. Netherlands: Springer; 2012

[3] Zadeh LA. The concept of a linguistic variable and its application to approximate reasoning. Information Sciences. 1975;**8x**:199-249

[4] Karnik NN, Mendel JM, Liang Q. Type-2 fuzzy logic systems. IEEE Transactions on Fuzzy Systems. 1999;**7**(6):643-658

[5] Zhang D, Zhang L, Yu Z, Shu L, Swain AK. A sum-based discrete event-triggered dynamic output feedback control for interval type-2 fuzzy systems. ISA Transactions. Oct 2022;**129**(Pt A):44-55. DOI: 10.1016/j.isatra.2021.12.031. Epub 2021 December 28. PMID: 35016801

[6] Zhao Z, Li J. Identification of continuous stirred tank reactor based on PCA-interval type-2 fuzzy logic system method. Procedia Computer Science. 2021;**183**:230-236

[7] Antonelli M, Bernardo D, Hagras H, Marcelloni F. Multiobjective evolutionary optimization of type-2 fuzzy rule-based systems for financial data classification. IEEE Transactions on Fuzzy Systems. 2017;**25**(2):249-264

[8] Luo C, Tan C, Wang X, Zheng Y. An evolving recurrent interval type-2 intuitionistic fuzzy neural network for online learning and time series prediction. Applied Soft Computing. 2019;**78**:150-163

[9] Bencherif A, Chouireb F. A recurrent TSK interval type-2 fuzzy neural networks control with online structure and parameter learning for mobile robot trajectory tracking. Applied Intelligence. 2019;**49**(11):3881-3893

[10] Montazeri-Gh M, Yazdani S. Application of interval type-2 fuzzy logic systems to gas turbine fault diagnosis. Applied Soft Computing. 2020;**96**: 106703

[11] Ljung L. System Identification: Theory for the User. 2nd ed. Upper Saddle River: Prentice Hall PTR; 1999

[12] Han D. A study on application of fuzzy adaptive unscented Kalman filter to nonlinear turbojet engine control. International Journal of Aeronautical and Space Sciences. 2018;**19**(2):399-410

[13] Torres LMM, Serra GLO. State-space recursive fuzzy modeling approach based on evolving data clustering. Journal of Control, Automation and Electrical Systems. 2018;**29**(4):426-440

[14] Gil P, Oliveira T, Brito Palma L. Online non-affine nonlinear system identification based on state-space neuro-fuzzy models. Soft Computing. 2018;**23**(16):7425-7438

[15] Todorov YV, Terziyska MN, Petrov MG. NEO-fuzzy state-space predictive control. IFAC-PapersOnLine. 2015;**48**(24):99-104

[16] Soderstrom T, Stoica P. Instrumental variable methods for system identification. Circuits, Systems, and Signal Processing. 2002;**21**(1):1-9

[17] Ashrafi M, Prasad DK, Quek C. IT2-GSETSK: An evolving interval type-II TSK fuzzy neural system for online modeling of noisy data. Neurocomputing. 2020;**407**:1-11

[18] Pires DS, De Oliveira Serra GL. Nonlinear dynamic system identification based on fuzzy Kalman filter. In: 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE); Vancouver, BC, Canada. 2016. pp. 17-23. DOI: 10.1109/FUZZ-IEEE.2016.7737662

[19] Pires D, Serra G. An approach for fuzzy Kalman filter modeling based on evolving clustering of experimental data. Journal of Intelligent & Fuzzy Systems. 2018;**35**(2):1819-1834

[20] Chiuso A, Picci G. Consistency analysis of some closed-loop subspace identification methods. Automatica. 2005;**41**(3):377-391

[21] Serra GLO, Bottura CP. Fuzzy instrumental variable approach for nonlinear discrete-time systems identification in a noisy environment. Fuzzy Sets and Systems. 2009;**160**(4): 500-520

[22] Filho ODR, Serra GLO. Recursive fuzzy instrumental variable based evolving neuro-fuzzy identification for non-stationary dynamic system in a noisy environment. Fuzzy Sets and Systems. 2018;**338**:50-89

[23] Ni Z, Liu J, Wu Z, Shen X. Identification of the state-space model and payload mass parameter of a flexible space manipulator using a recursive subspace tracking method. Chinese Journal of Aeronautics. 2019;**32**(2):513-530. DOI: 10.1016/j.cja.2018.05.005. ISSN 1000-9361

[24] Ni Z, Liu J, Wu Z. Identification of the time-varying modal parameters of a spacecraft with flexible appendages using a recursive predictor-based subspace identification algorithm. Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering. 2019;**233**(6): 2032-2050. DOI: 10.1177/0954410018770560

[25] Angelov P, Filev D. Simpl_eTS: A simplified method for learning evolving Takagi-Sugeno fuzzy models. In: Proceedings of the 2005 IEEE International Conference on Fuzzy Systems FUZZ-IEEE – 2005. Reno; 2005. pp. 1068-1073

[26] Angelov P, Zhou X. Evolving fuzzy systems from data streams in real-time. In: 2006 International Symposium on Evolving Fuzzy Systems. IEEE; 2006

[27] Kasabov NK, Song Q. DENFIS: Dynamic evolving neural-fuzzy inference system and its application for time-series prediction. IEEE Transactions on Fuzzy Systems. 2002;**10** (2):144-154

[28] Angelov P, Lughofer E, Klement EP. Two approaches to data-driven design of evolving fuzzy systems: eTS and FLEXFIS. In: NAFIPS 2005 – 2005 Annual Meeting of the North American Fuzzy Information Processing Society. IEEE; 2005

[29] Lemos A, Caminhas W, Gomide F. Multivariable gaussian evolving fuzzy modeling system. IEEE Transactions on Fuzzy Systems. 2011;**19**(1):91-104

[30] Juang J-N, Phan MQ. Identification and Control of Mechanical Systems. Cambridge, UK: Cambridge University Press; 2001. 334 p

[31] de Oliveira Serra GL. Kalman Filters – Theory for Advanced Applications. London, UK: IntechOpen; 2018

Chapter 4

Bringing Data Converter Pairs into Chaotic Oscillation for Built-in Self-Test and Entropy Generation

Sergio Callegari

Abstract

A pair comprising an analog-to-digital converter (ADC) and a digital-to-analog converter (DAC) can enter chaotic oscillation when closed in a feedback loop with a limited set of additional elements. This phenomenon can be employed for entropy generation in true random number generators (TRNGs). Additionally, the oscillation can expose defects in the components' operation, providing an opportunity for built-in self-test (BIST). Reconfigurable loops sustaining self-oscillation characterize the oscillation-based test (OBT) approach, appealing for not requiring resources to excite the blocks under test (BUTs). While OBT has been applied to various signal processing primitives, its use in data converters has been mostly confined to specific subsystems or variations of servo-testing. Here, it is shown that chaotic OBT of data converter pairs may offer insights on their input-output characteristics, while providing entropy generation at the same time. A PIC microcontroller, together with an external DAC and some operational amplifiers, is used as a test bed to validate the approach's scope and demonstrate its applicability to real-world systems.

Keywords: analog to digital converter (ADC), digital to analog converter (DAC), oscillation based test (OBT), built-in self test (BIST), entropy source, chaotic map

1. Introduction

The possibility of purposely obtaining complex dynamics from electronic systems dates back to the early 80s, marked by the introduction of the Chua's circuit [1]. Notwithstanding previous evidences of complex behaviors in oscillators and filters [2], the absence of simple, reproducible examples had previously relegated chaotic phenomena in electronic circuits to the realm of curiosities or misbehaviors. Soon after, another milestone was marked by the advent of discrete-time circuits utilizing *chaotic maps* [3, 4] which enabled a drastic simplification of the adopted models. Specifically, the restriction to the subclass of piece-wise affine Markov (PWAM) maps enabled the application of advanced mathematical tools for analysis and (to some extent) design [5, 6].

Early applications of chaos in electronics encompassed stochastic artificial neural models [7], secure and broadband communication [8], the synthesis of spreading

sequences for CDMA systems [9], and the disruption of periodic behavior to reduce electro-magnetic interference in switched-mode and digital systems [10].

Relevant to this chapter, another early envisioned application of chaos was in synthesizing excitations for testing frameworks, especially for analog and mixed-signal circuits [11, 12]. In this context, chaotic waveforms may offer distinct advantages, including their broad frequency content, which can expedite evaluations and unveil issues not apparent with classic test signals. Almost contemporary to this, the testing field witnessed the introduction of the oscillation-based test (OBT) framework [13, 14]. OBT entails minor circuit modifications activated during the test phase, inducing sustained oscillations in the block under test (BUT) to identify faults or defects. A key advantage lies in resource savings, eliminating the need for a test signal generation unit. OBT has been successfully applied to various signal processing blocks, such as filters, amplifiers, and modulators [15]. Shortly after its introduction, the proposal of chaotic OBT came as an appealing extension [16].

A further significant application of chaotic dynamics is *entropy generation* for true random number generators (TRNGs), which gained prominence amid increasing concerns regarding the vulnerability of pseudorandom number generators (PRNGs) in security applications [17, 18]. Curiously, there is evidence that some TRNGs circuits ultimately deriving their properties from chaotic dynamics have been proposed even without realizing this fact [19]. Unfortunately, a significant hurdle in incorporating chaos-based TRNGs into computer systems lies in their inherently analog nature. This challenge can be mitigated by basing their architecture on analog or mixed-mode building blocks that are already well accepted in predominantly digital chipsets. Subsystems derived from data converters emerge as excellent candidates given the established practice of integrating analog-to-digital converter (ADC) and digital-to-analog converter (DAC) IP blocks in systems on a chip (SOCs). Indeed, at the beginning of the century, the recognition that the *quantization error function* of ADCs could serve as a foundation for chaotic maps led to TRNG proposals based on coupled ADC and DAC stages [20, 21]. The use of *complete* ADC and DAC pairs is also feasible. Even if it might not result in the most streamlined architectures, it has garnered attention for the convenience in repurposing building blocks that may be readily available with spare capacity [22].

The premises presented so far suggest the existance of an interesting research area at the convergence between different applications of chaotic dynamics in relation to data converters. The use of ADC and DAC pairs as chaotic oscillators can at the same time establish the foundation for both entropy generation and the OBT of these components. In fact, a versatile three-way operation may be obtainable by a reconfigurable architecture where ADCs and DACs pairs can serve either in their conventional data conversion roles, operate in an OBT-type built-in self-test (BIST) mode, or contribute to populating entropy buffers for TRNGs. The latter two tasks can, to some extent, be performed concurrently.

While OBT has already been proposed for data converters, it remained mostly confined to specific subsystems thereof or to variations of the so called *servo-testing* technique [23, 24]. In this work, the goal is to get insight into their overall input-output relationship. Furthermore, chaotic excitation has been previously proposed for the test of ADCs [12]. Now, the challenge is to have an excitation that gets itself conditioned by the non-idealities of the converter, potentially amplifying the possibility to reveal them. A possible limitation of the concept is that ADCs and DACs need to be tested in pairs, complicating the attribution of the observed effects to defects in the one or in the other. Furthermore, once one has a converter pair, traditional testing

based on digital vectors might appear a straightforward choice. Yet, OBT can save the resources required for storing or computing test sequences. Additionally, the proposed concept has the unique advantage of joining testing with entropy generation. The chapter is organized as follows: Section 2 illustrates the theoretical principles inherent in bringing ADC and DAC pairs in chaotic oscillation; Section 3 illustrates applicability to entropy generation; Section 4 describes how to take advantage of the oscillations for OBT and built-in self-test (BIST); finally, Section 5 provides experimental results obtained on a microcontroller SOC. A PIC microcontroller, together with an external DAC and a few other components, is employed as a flexible test bed both to provide a tangible example and to prove applicability to real-world systems. Some conclusive remarks are eventually drawn together with hints at open problems and possible developments.

2. Chaotic oscillation of data converter pairs

To induce chaotic oscillation in a data converter pair, it is sufficient to establish an autonomous discrete-time one-dimensional (1D) dynamical system, as depicted in **Figure 1**. The current state, stored in the *analog register* AR, is processed into the next state by a map derived from the quantization error function $e(\cdot)$ of the ADC. Its computation involves the DAC, as shown in the upper signal path in the figure. Formally, the model is:

$$x_{n+1} = M(x_n) = a\,e(x_n) + b = a(q(x_n) - x_n) + b, \tag{1}$$

where x_n is the state x at time n. The quantization function $q(\cdot)$ is obtained as $f_{\mathrm{DAC}}(d_n)$, where $d_n = f_{\mathrm{ADC}}(x_n)$ is the *digital* (discrete) version of x_n, and $f_{\mathrm{ADC}}(\cdot)$ and $f_{\mathrm{DAC}}(\cdot)$ are the input-output characteristics of the ADC and DAC, respectively. The two converters share compatible resolutions and the same analog range. The parameters a (a gain such that $|a| > 1$) and b (an offset) need to be selected to ensure that an interval $[x_L, x_R]$ remains *invariant* through $M(\cdot)$, acting as an attractor for system trajectories. To exemplify the concept, let the ADC and DAC analog range be normalized to $[0, 1]$, with $l = 2^r$ digital levels, r being the bit resolution of the converters. **Figure 2** shows the static characteristic of the converters, along with $q(\cdot)$ and $M(\cdot)$. The plots consider a scenario where x can slightly exceed the nominal ADC input range. In the latter plot, the invariant set (IS) of $M(\cdot)$ is emphasized, as its features will be explored shortly.

Meanwhile, it is important to note that the trajectories within the interval $[x_L, x_R]$ are inherently chaotic. While they remain confined to the interval, stable p-period cycles (including equilibrium points with $p = 1$) are not possible. To be a point on a

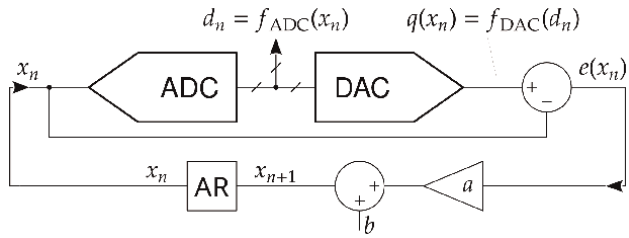


Figure 1.
Autonomous, discrete-time 1D system based on a data converter pair, capable of chaotic behavior.

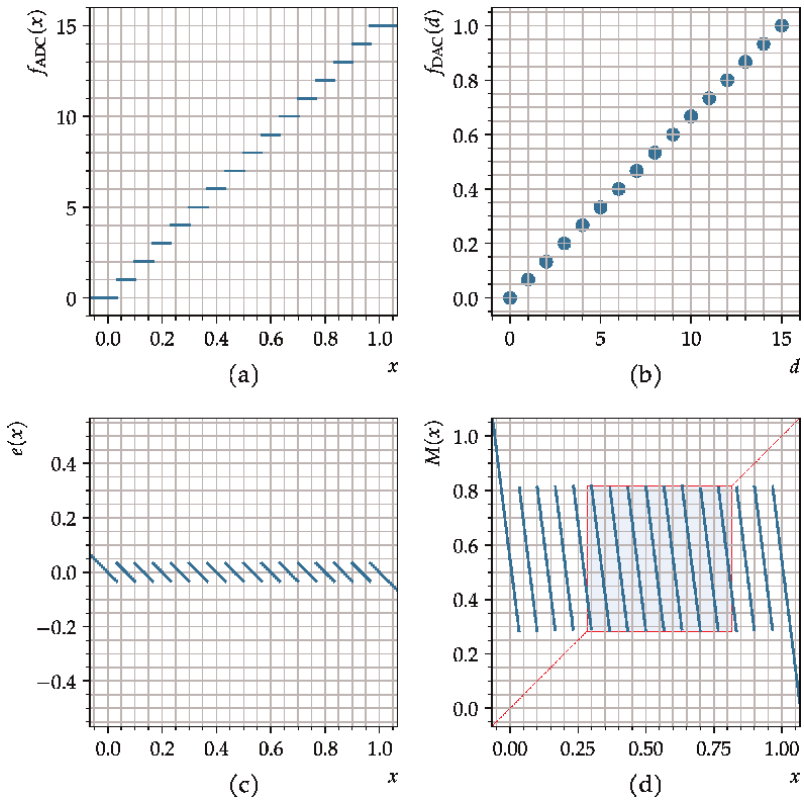


Figure 2.
Operation of the architecture in ***Figure 1****. For representation clarity, a sample case with only 16 quantization levels is considered. Loop parameters are set at $a = 8$ and $b = 0.55$. Individual plots: (a) ADC characteristic function; (b) DAC characteristic function; (c) quantization error function; and (d) map determining the state evolution in the closed loop system.*

p-period orbit, some $\hat{x}$ must be an equilibrium point of $M^p(\cdot)$. In this case, the root of the characteristic equation at $\hat{x}$ is

$$z = \left.\frac{\mathrm{d}M^p(x)}{\mathrm{d}x}\right|_{x=\hat{x}} = (-a)^p. \tag{2}$$

However, this value inevitably falls outside the unit circle, as $|a| > 1$, indicating instability. Additionally, the system exhibits sensitivity to initial conditions. Two trajectories starting at a closely spaced distance $|\delta_0|$ after n steps become separated by $|\delta_n|$, which grows exponentially as $e^{\lambda n}$, where $\lambda = \ln|a| > 0$ represents the Lyapunov exponent. To exemplify this sensitivity, **Figure 3** shows two trajectories originating from almost overlapping values.

To gain a deeper understanding of the IS, it is important to observe that its endpoints are mutually defined as

$$x_L = \min_{x \in [x_L, x_R]} M(x) \qquad \text{and} \qquad x_R = \max_{x \in [x_L, x_R]} M(x) \tag{3}$$

and that among intervals $[x_L, x_R]$ satisfying these criteria, the goal is to find the smallest one preventing the system state from escaping. Let $\Delta = 1/(l-1)$ represent

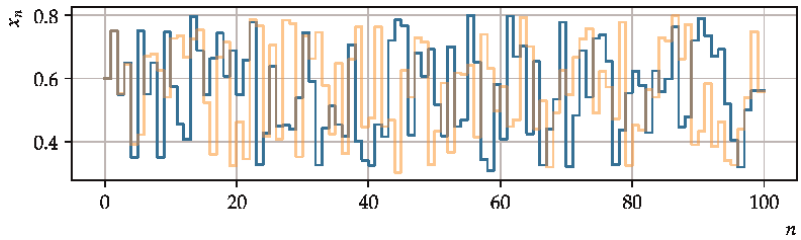


Figure 3.
Sample trajectories starting from nearby points, obtained from the architecture depicted in ***Figure 1*** *matching the configuration presented in* ***Figure 2****.*

the quantization step, and note that all branches of $q(\cdot)$ span at least $[-\Delta/2, \Delta/2]$. Given $|a|>1$, it follows that $x_L \leq -|a|\Delta/2+b$ and $x_R \geq |a|\Delta/2+b$. Two cases must be considered.

For $a<0$, the branches of $M(\cdot)$ have a positive slope, which implies that the only possibilty for x_L to be strictly less than $-|a|\Delta/2+b$ is if $x_L = M_L(x_L)$, where $M_L(\cdot)$ is the left-most branch of $M(\cdot)$. However, this scenario must be excluded as it would place an unstable equilibrium point at the boundary of the IS. This, in turn, would allow minor perturbations to trigger an escape from the IS. Hence, it must be $M_L(x_L)>x_L$. Similar reasoning leads to the conclusion that x_R cannot be strictly larger than $|a|\Delta/2+b$ and that it must be $M_R(X_R)<X_R$ with $M_R(\cdot)$ representing the right-most branch of $M(\cdot)$. Altogether,

$$\begin{cases} x_L < M_L^{-1}(x_L) \\ x_R > M_R^{-1}(x_R) \end{cases} \Rightarrow \begin{cases} -|a|\dfrac{\Delta}{2}+b > -\dfrac{\Delta}{2} \\ |a|\dfrac{\Delta}{2}+b < 1+\dfrac{\Delta}{2} \end{cases} \tag{4}$$

and

$$\begin{cases} x_L = -|a|\dfrac{\Delta}{2}+b \\ x_R = \quad |a|\dfrac{\Delta}{2}+b \end{cases}. \tag{5}$$

The inequalities can be rewritten as

$$\frac{1}{2}\frac{|a|-1}{l-1} = (|a|-1)\frac{\Delta}{2} < b < 1-(|a|-1)\frac{\Delta}{2} = 1-\frac{1}{2}\frac{|a|-1}{l-1}. \tag{6}$$

Moreover, by subtracting one of the inequalities in (4) from the other, one gets $|a|\Delta<1+\Delta$, that is

$$|a| < 1+\frac{1}{\Delta} = l, \tag{7}$$

this bound being reachable for $b=1/2$.

For $a>0$, the scenario is slightly more complex, due to the negative slope of the branches of $M(\cdot)$. For x_L to be $\leq -a\Delta/2+b$, one needs $x_L = M_R(x_R)$, with $x_R > 1+\Delta/2$. Should x_L be also $< -\Delta/2$, then x_R would be $M_L(x_L) > a\Delta/2+b$.

However, this would lead to $x_L = M_R(M_L(x_L))$, that is unacceptable, as it would place an unstable period-2 orbit at the boundary of the IS. This, in turn, would allow minor perturbations to trigger an escape from the IS. Consequently, getting both $x_L < -a\Delta/2 + b$ and $x_R > a\Delta/2 + b$ simultaneously is not possible. If $x_R = a\Delta/2 + b$, as long as $x_R < 1 + \Delta/2$, one gets $x_L = -|a|\Delta/2 + b$; otherwise, $x_L = M_R(x_R)$. Additionally, it must be $M_R(x_R) > -\Delta/2$ to avoid also having $x_R = M_L(x_L)$. By similar reasoning if $x_L = -a\Delta/2 + b$, as long as $x_L > -\Delta/2$, one gets $x_R = |a|\Delta/2 + b$; otherwise, $x_R = M_L(x_L)$. Furthermore, one must satisfy $M_L(x_L) < 1 + \Delta/2$ to avoid having $x_L = M_R(x_R)$. Altogether,

$$\begin{cases} x_R < M_R^{-1}\left(-\dfrac{\Delta}{2}\right) \\ x_L > M_L^{-1}\left(1+\dfrac{\Delta}{2}\right) \end{cases} \Rightarrow \begin{cases} a\dfrac{\Delta}{2} + b < 1 + \dfrac{b}{a} + \dfrac{\Delta}{2a} \\ -a\dfrac{\Delta}{2} + b > -\dfrac{1}{a} + \dfrac{b}{a} - \dfrac{\Delta}{2a} \end{cases} \tag{8}$$

and

$$\begin{cases} x_L = \min\left(-a\dfrac{\Delta}{2} + b, -a^2\dfrac{\Delta}{2} - ab + a + b\right) \\ x_R = \max\left(a\dfrac{\Delta}{2} + b, a^2\dfrac{\Delta}{2} - ab + b\right). \end{cases} \tag{9}$$

The inequalities can be rewritten as

$$1 - \frac{a}{a-1} - \frac{\Delta}{2}(a+1) < b < \frac{a}{a-1} - \frac{\Delta}{2}(a+1) \tag{10}$$

that is,

$$1 - \frac{a}{a-1} - \frac{1}{2}\frac{a+1}{l-1} < b < \frac{a}{a-1} - \frac{1}{2}\frac{a+1}{l-1}. \tag{11}$$

Moreover, subtracting one of the two inequalities in (8) from the other, again one gets the inequality in (7).

While the case with $a > 0$ seems to be tolerant to a larger parameter range than the one with $a < 0$, it also leads to more complex relationships. Yet, adopting the stricter condition (6) in place of (11) simplifies (9) back into (5) also in this condition.

The considerations proposed so far reveal that a determines the size of the IS (the larger $|a|$, the larger the set), while b shifts it along the converters' scale. When there is no attractive IS, in principle, the system should diverge. In practice, components in the feedback path of **Figure 1** may experience saturation, leading to the clipping of $M(\cdot)$. Trajectories passing through clipping points inevitably result in stable periodic orbits (for $a > 0$) or equilibrium points (for $a < 0$), as illustrated in **Figure 4**, preventing chaotic behavior. Adhering to the guidelines above can assist in configuring a and b to establish an invariant interval within the clipping range, thereby avoiding this condition.

Until now, the discussion has assumed ideal behavior from the components in **Figure 1**. In reality, there will be unavoidable deviations from it. While these deviations will be the primary focus of Section 4, it is worth anticipating that the model in Eq. (1) transforms into:

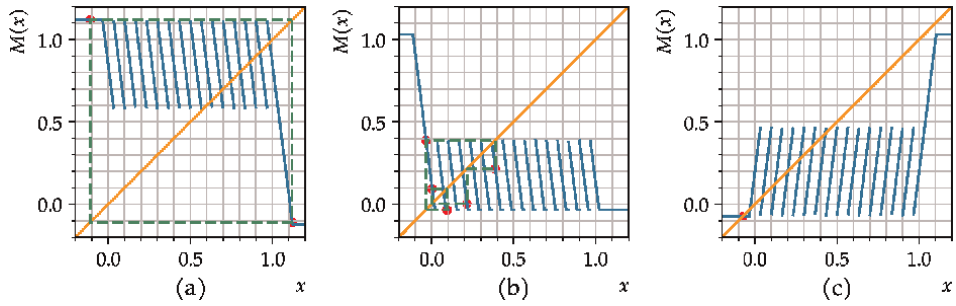


Figure 4.
Stable orbits as a result of an incorrect configuration of parameters a and b, given the presence of saturation in the architecture depicted in ***Figure 1****. In the individual plots: (a) period-2 cycle with $a > 0$ and saturation occurring out of the data converter nominal range; (b) longer cycle with $a > 0$ and saturation occurring in the data converter nominal range; (c) stable equilibrium point with $a < 0$.*

$$x_{n+1} = \tilde{M}(x_n) + \nu(n) \tag{12}$$

where $\tilde{M}(\cdot)$ is a modified version of $M(\cdot)$ influenced by the static errors in the converters and in the other analog components in **Figure 1**, while $\nu(n)$ is a noise term condensing the effects of the ADC input-referred noise $\nu_{\mathrm{ADC}}(n)$, the DAC output referred noise $\nu_{\mathrm{DAC}}(n)$, as well as noise in the other analog components. Expectably, errors and noise effects from the data converters will tend to dominate, both because these are the most complex components, and due to the fact that they get amplified by a.

For illustrative purposes, **Figure 5a** hints at the type of relationship that can be obtained from x_n to x_{n+1}, based on artificially synthesized data converter characteristics with a very low resolution, including missing codes, non-monotonicity, and nonlinearity. Note that, in this scenario, $\tilde{e}(\cdot)$, analogous to $e(\cdot)$ in the ideal case, exhibits a significantly larger image set. As a consequence, the choice of parameters a and b must be much more conservative to maintain $M(\cdot)$ within the clipping range. Specifically, only reduced values of $|a|$ can be tolerated, particularly when one deals with data converters with an effective number of bits (ENOB) non-negligibly smaller than the nominal resolution. On the other hand, $\nu(n)$ may end up partially disrupting the periodicity expected to arise when the clipping limits are reached.

As a final note, observe that an effective way to restore a more regular behavior when non-idealities are involved can be to *pretend* that the ADC and DAC have a lower bit-resolution $\hat{r}$ than their real one r. To this aim, one passes only the $\hat{r}$ most significant bits (MSBs) from the ADC output to the corresponding MSBs in the DAC input. **Figure 5b** hints at the behavior that can be achieved by this approach for the same test conditions as in **Figure 5a**. Clearly, also this strategy requires a reduction in $|a|$, as $|a|$ now needs to be related to $\hat{l} = 2^{\hat{r}}$. Additionally, one should consider that "not passing" some bits from the ADC to the DAC is a form of *truncation* that introduces an offset as large as $\left(2^{(r-\hat{r})} - 1\right)\Delta/2$ in $e(\cdot)$ that needs to be compensated acting on b by a times that amount. On the other hand, this strategy enables the usage of a DAC with an inherently lower resolution than the ADC, or, alternatively, it lets the DAC least significant bits (LSBs) be used as a degree of freedom to achieve an effect equivalent to varying b in steps.

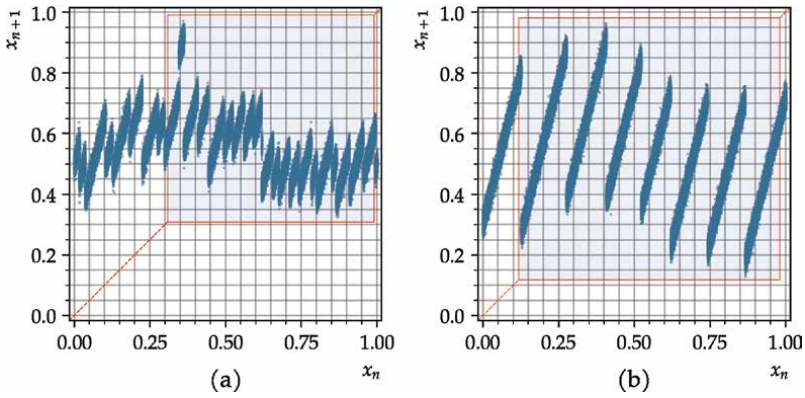


Figure 5.
*Sample relationship from x_n to x_{n+1} achieved from the architecture in **Figure 1** in presence of ADC/DAC non-idealities and noise. For illustration, 5-bit data converters are employed with artificially synthesized errors (nonlinearity, as well as missing values and non-monotonicity in the ADC). In both cases, $a = -4$. In the individual plots: (a) all the bits in the ADC output are passed to the DAC, $b = 0.5$. Visually, non-monotonicity results in "stacked" branches (e.g., at $x_n \approx 0.35$) while missing codes cause long branches that appear "more spaced" than usual (e.g., at $x_n = 0.45$); (b) only $\hat{r} = 3$ bits from the ADC are passed to the DAC, restoring a more regular behavior. Here, $b = 0.5 - 0.19$ to compensate for the data truncation.*

3. Entropy generation from data converter pairs

TRNGs are systems capable of delivering bit-streams made of independent and identically distributed (IID) bits from the observation of completely unpredictable physical phenomena. Because it is not always possible to rely on a physical process *directly* capable of delivering IID symbols (e.g., as the toss of an unbiased coin), TRNGs are typically built as the cascade of multiple subsystems, as sketched in **Figure 6** [25].

In this architecture, the *entropy distiller* acts as a post-processor block taking the bits B_s of digital samples from an unpredictable process and deriving from them IID bits B_r notwithstanding uneven distributions or correlations on B_s. The *entropy buffer* lets the distiller operate on multiple B_s samples at once, as well as mix data from multiple entropy sources and deliver a sustained throughput even when the source(s) operates discontinuously or at a non-constant rate. In this work, we consider a single source.

From an information theory point of view, the information rate before the distiller ends up as a bound for the bit rate at its output, given that a perfectly random binary stream has an entropy rate H of one bit per symbol [26]. Namely, if H_s is the entropy rate at B_s and f_s is its bit rate, the bit rate f_r at B_r should be less than $f_s H_s$ [27].

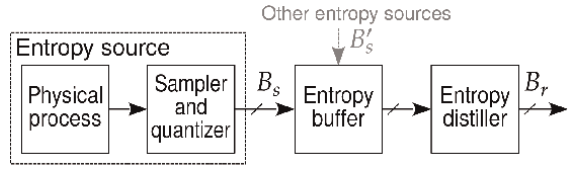


Figure 6.
Architecture of a typical TRNG.

In practice, the situation is somehow fuzzy. On one hand, one should have f_r significantly lower than f_s to accommodate for less than ideal efficiency in the post-processing and uncertainty in the knowledge of the uphill H_s. On the other hand, when the *randomness* of B_r is validated by means of statistical *test suites* [28, 29], the tests can potentially be passed even when B_r delivers little or or no entropy at all. This is proven by the very existance of PRNG that succeed in the tests. Consequently, distillers where $f_f < f_s H_s$ is not (strictly) respected may appear to deliver "random" data, and care is evidently required.

In the current discussion, the idea is evidently to use an (analog) chaotic circuit as the physical process, and in this view, it is important do establish a few points. First of all, statistical tests suites are not alone an appropriate way of testing TRNGs [28]. Secondly, systems that solely rely on the output of the distiller for testing purposes are insecure. In fact, the conditioning process might obscure any deficiencies in the data stream originating from the entropy source. In the worst scenario, one might end up using a TRNG that has degenerated into a PRNG without realizing it. For this reason, a secure TRNG should incorporate an interface to inspect the output of its entropy source (possibly even at a sampling rate or at a resolution better than those used to deliver B_s). As a third point, validation should include entropy estimation at this interface but possibly other health checks too. Finally, having entropy sources that are inherently charaterized by both a high bit rate and a high entropy rate is desirable as these properties simplify the design of the distiller while enabling a high data throughput.

From these points, the appeal of the ADC-DAC-based chaotic systems as an entropy source should be evident. The presence of the ADC in the loop makes the physical source capable of sampling and self-quantizing its state at a high resolution, inherently providing the interface needed for validation and health checks. Furthermore, with reference to entropy estimation and measurements, a chaotic system based on a 1D map has distinguished advantages, as mathematical tools exist enabling a computation of the entropy rates to be expected. A complete theoretical discussion is available in [6, 26, 30]. Here, the key points shall be reviewed, sacrificing some formality for brevity.

A first, notable concept is represented by the Perron–Frobenius operator (PFO) $\mathbf{P}_M$ associated to the map that lets one observe how it transforms probability density functions (PDFs) [6]. Namely, if the initial state x_0 is drawn in the map IS according to a PDF $\rho_0(\cdot)$, then one can compute the PDF associated to x_1 as $\rho_1 = \mathbf{P}_M[\rho_0]$. As long as x_0 of the system is not known exactly ($\rho_0(\cdot)$ has bounded variation), the sequence $\{x_n\}$ forms a Markov process whose statistics can be studied via the PFO. The operator is linear and in many practical cases (viz, for mixing maps) admits a unique invariant density $\hat{\rho}(\cdot)$ to which PDF sequences such as $(\rho_0, \rho_1, \rho_2, \ldots)$ converge at an exponential rate. The invariant density describes the distribution of points in typical trajectories.

Working with the PFO, one faces two difficulties: the first is that the PFO is infinitely dimensional and thus not easily manageable; the second is that, in the proposed application, one is not ultimately interested in the statistics of $\{x_n\}$, rather in some discrete version of it from which the bits B_s are obtained. Let $\mathcal{S}_n = g(x_n)$, where $g(\cdot)$ is an *output function* obtaining discrete symbols $\mathcal{S}$ from the system state x. In the proposed system, $g(\cdot)$ will be naturally derived from $f_{\mathrm{ADC}}(x_n)$ as some $\hat{g}\big(f_{\mathrm{ADC}}(x_n)\big)$, where $\hat{g}(\cdot)$ is a mapping implementable by purely digital means. Being derived from $\{x_n\}$, the sequence $\{\mathcal{S}_n\}$ forms a hidden Markov process.

The difficulties above can be overcome by approximating the PFO over a finite dimensional space. This can be done by partitioning the IS into a finite number of

non-overlapping intervals (let I_i be the generic interval) and observing the *coarse* dynamics by which the system state jumps from one interval (viz., discrete state) to another. By doing so, the PFO gets approximately expressible via a *kneading matrix* $\boldsymbol{K}$ whose generic entry $k_{i,j}$ returns the fraction of the interval I_i that gets mapped via $M(\cdot)$ into I_j. With this, the probability vector that is invariant under $\boldsymbol{K}$ ends up as an approximation of $\hat{\rho}(\cdot)$.

Interestingly, for a subclass of piece-wise affine maps, called PWAM maps, this reduction can be practiced with no approximation at all. A piece-wise affine map is PWAM if a partition can be defined so that: (i) partition points include the map breakpoints; and (ii) partition points map into partition points. With PWAM maps, picking an output function $g(\cdot)$ such that each partition interval I_i is mapped into an output symbol S_i assures that the sequence $\{\mathcal{S}_n\}$ taking values in $\{S_0, S_1, \ldots\}$ forms a Markov process. Expectably, in the corresponding Markov chain, the transition probability from state S_i to state S_j is given by the entry $k_{i,j}$ of $\boldsymbol{K}$. Furthermore, in PWAM maps, the invariant density $\hat{\rho}(\cdot)$ is known to be uniform (flat) within each partition interval. Hence, a perfect derivation of $\hat{\rho}(\cdot)$ is possible from the kneading matrix.

Relevant to this work, for ideal ADC and DAC pairs, the architecture in **Figure 1** provides PWAM maps as long as a is integer and Eq. (5) holds. In this case, the map gets composed of $|a|$ identical, linear branches within its IS. This statement may appear in conflict with examples such as that in **Figure 2d**, where the branches at the extremes of the IS are different from the others, but becomes clear if you "wrap around" the invariant interval onto itself so that these "incomplete" branches merge together into a single one. The partition is then obtained by considering the domains of the individual branches as partition intervals. Let these be I_0 to $I_{|a|-1}$ for growing values of x_n with I_0 possibly including also the last branch when the extreme branches are incomplete. An example of the resulting kneading matrix and Markov chain is illustrated in **Figure 7**, with reference to the setup illustrated in **Figure 2d**.

In general, for an integer a, one should ideally get an $|a| \times |a|$ kneading matrix where all the entries are $1/|a|$, leading to a Markov chain with $|a|$ states that is fully connected and where all the transition probabilities are $1/|a|$. This chain corresponds to the toss of a fair $|a|$-faced dice. Furthermore, the Markov states can be obtained from the ADC output, since the partition intervals are necessarily aligned with its quantization thresholds. Finally, when $|a|$ is a power of two, that is, $|a| = 2^w$, the Markov states (viz., the symbols $\mathcal{S}$) can be perfectly encoded in w-bits binary words that can be serialized into a bitstream. In other words, when all the components obey to their ideal behavior and $|a| = 2^w$, the architecture in **Figure 1** represents a *perfect* entropy source, delivering a bistream B_s that is already perfectly random with no need for post-processing.

Unfortunately, real-world analog hardware is always subject to deviations from nominal behavior, and one must consider what happens in this case. As shown in **Figure 5**, the map will be significantly distorted and perturbed by noise, making it non-PWAM and characterized by more than $|a|$ branches in its IS. In order to get insight into this situation, a preliminary consideration is due: when a chaotic map is employed as an entropy source, the achievable entropy per cycle will be a function of *both* the map and the output function (i.e., the way in which the analog state is transformed in output bits). Specifically, the map is going to set a *fundamental limit* on the achievable entropy (from its *metric* entropy [31]), and the output function will let it be approached to various degrees. Intuitively, the inherent limit set by the map will be related to its Lyapunov exponent λ that measures the amount of information (about

$$k = \begin{pmatrix} \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \end{pmatrix}$$

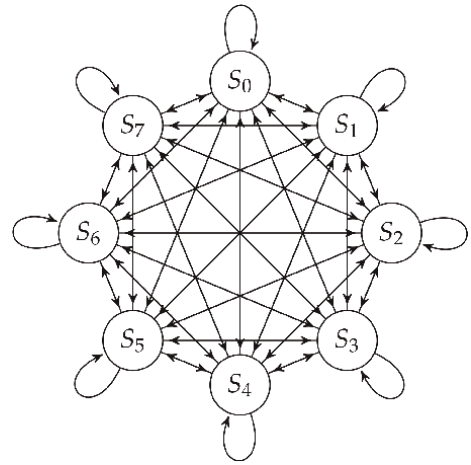


Figure 7.
Kneading matrix and Markov chain corresponding to the map in ***Figure 2d*** *(obtained for $a = 8$) restricted to its IS. In the chain, all transition probabilities are 1/8.*

the system initial condition) that is lost at every cycle [26]. Even in the non-ideal case, the map under exam is going to be made of branches with a local slope always close to the nominal a. Hence, one can expect the entropy rate to be eventually bounded by $\log_2(|a|)$.

For what concerns the output function, the finer the partition of the IS, that is, the larger the dictionary of the symbols produced by the output function, the more information will be preserved in the translation of the continuous state into the output symbols, leading to a better approximation of the fundamental limit set by the map. However, this will also result in requiring a large number of bits to encode each symbol and thus in a poor output entropy per bit at the output, complicating the design of the distiller. Clearly, the optimal situation in terms of achievable output bit rate and entropy rate would be a symbol dictionary comprising $|a|$ symbols and having $|a| = 2^w$ so that the symbols can be exactly mapped into binary words. The question is how this output function should be designed. A first critical aspect in this sense is that even if the domain of the output function should be the IS, in the proposed system, the latter cannot be known precisely in advance as it depends on the non-idealities in the ADC, DAC, and other analog components as shown in Section 2.

The results in [26] provide some notable aid in this sense. In fact, such paper suggests that even when seeking an output function offering a *coarse* discretization, it is convenient to base it on a much *finer* partitioning of the domain, by associating each output symbol to the union of many small partition intervals, uniformly scattered across the domain, as illustrated in **Figure 8a** and **b**.

This is noteworthy. First of all, if the fine partitioning is fine enough, and poor alignment with the actual IS will not be very important, thus allowing the quantization function to be defined for the whole of $[0, 1]$ rather than for the IS, as illustrated in **Figure 8c**. Secondly, in the case at hand, $g(\cdot)$ must be based on $f_{\mathrm{ADC}}(\cdot)$ which inherently provides a fine partitioning of $[0, 1]$. Specifically, when $|a|$ is a power of 2, as in 2^w, $g(\cdot)$ can be implemented by simply extracting the w LSBs from the ADC output at each cycle, with those bits directly providing a binary encoding of the discrete state.

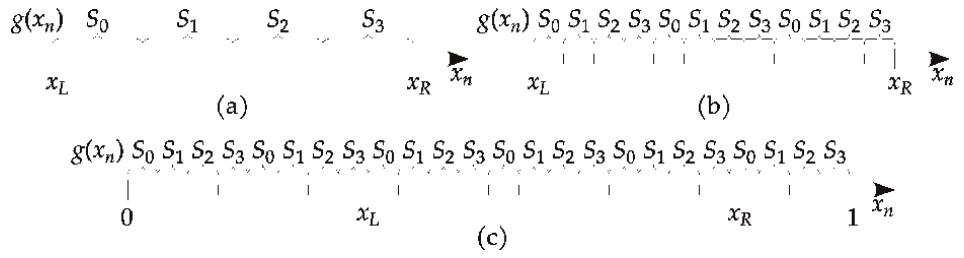


Figure 8.
Comparison of output functions. In all these example cases, an output function $g(\cdot)$ resulting in an alphabet of four output symbols (S_0 to S_3) is considered. In (a), $g(\cdot)$ is based on a coarse partitioning of the IS. In (b), it is based on a finer partitioning. Plot (c) shows that as long as the fine partitioning is fine enough, and $g(\cdot)$ can be defined over the whole range $[0, 1]$ of the converters and operate robustly even when the actual $[x_L, x_R]$ is not known in advance.

4. Built-in self-test of data converter pairs

Established the possibility of obtaining an entropy source by driving an ADC and DAC pair in self-oscillation, it is time to consider how this configuration can help evaluating the performance of the data converters.

In the architecture in **Figure 1**, the possibility to observe the system state x via its coarse (digital) version d is inherent. The analysis shall thus be carried out by analyzing sequences $\{d_n\}$ to appreciate deviations of the converters from their nominal operation. To contain data management costs, tools that can operate on d_n values without memorizing large sequences are desirable. Under this premise, the estimation of the probability distribution $P(d)$ of $\{d_n\}$ via histogram methods represents a first, obvious choice.

From $P(d)$, missing codes are immediately recognizable. Furthermore, $P(d)$ can hint at the span of the *effective* IS obtained in oscillation which in turn is related to the *excess range* in the image set of the error function. Refer to Section 2 for the notation and let $[\tilde{x}_L, \tilde{x}_R]$ be the effective IS. Also, let $\tilde{e}_T$ and $\tilde{e}_B$ be the bounds of $\tilde{e}(x)$ for $x \in [\tilde{x}_L, \tilde{x}_R]$, when the bounds of $e(x)$ in the same domain should evidently be $-\Delta/2$, $\Delta/2$. Clearly, $x_L - \tilde{x}_L$ and $\tilde{x}_R - x_R$ owe to $\tilde{e}_T - \Delta/2$ and $\tilde{e}_B + \Delta/2$ via a. The relationship is straightforward when $a < 0$, and the branches of $M(\cdot)$ take a positive slope. In this case, $x_L - \tilde{x}_L \approx |a|\tilde{e}_B + \Delta/2$ and $\tilde{x}_R - x_R \approx |a|\tilde{e}_T - \Delta/2$. Established this point, consider that $\tilde{x}_L$ and $\tilde{x}_R$ can be known (with some approximation) from $\tilde{d}_L$ and $\tilde{d}_R$, the latter two being the extremes of the support of the estimated $P(d)$. Consequently approximate values of $\tilde{e}_B + \Delta/2$ and $\tilde{e}_T - \Delta/2$ are derivable. These values are in close relationship to the positive and negative peaks of the difference between the ideal and effective characteristics of the ADC + DAC chain in $[\tilde{x}_L, \tilde{x}_R]$ plus some noise amplitude (in fact they also incorporate the errors and noise of the other analog elements in the loop). Similar considerations clearly be made also in the case where DAC is used whose resolution is smaller than that of the ADC, as proposed at the end of Section 2.

As an example, and for mere representation purpose, **Figure 9a** and **d** show histograms obtained from the ADC output sequence $\{d_n\}$ corresponding to the maps in **Figure 5a** and **b**. In the plots, the very large artificially synthesized errors and the extremely low converter resolution make the missing codes and the enlargement in the IS from the expected one stand out to the human eye, but the approach is clearly applicable at any resolution.

Additionally, one can look at sequences made of the (d_n, d_{n+1}) pairs rather than d_n alone. In this case, two-dimensional (2D) histograms with estimations of $P(x_{n+1}, d_n)$

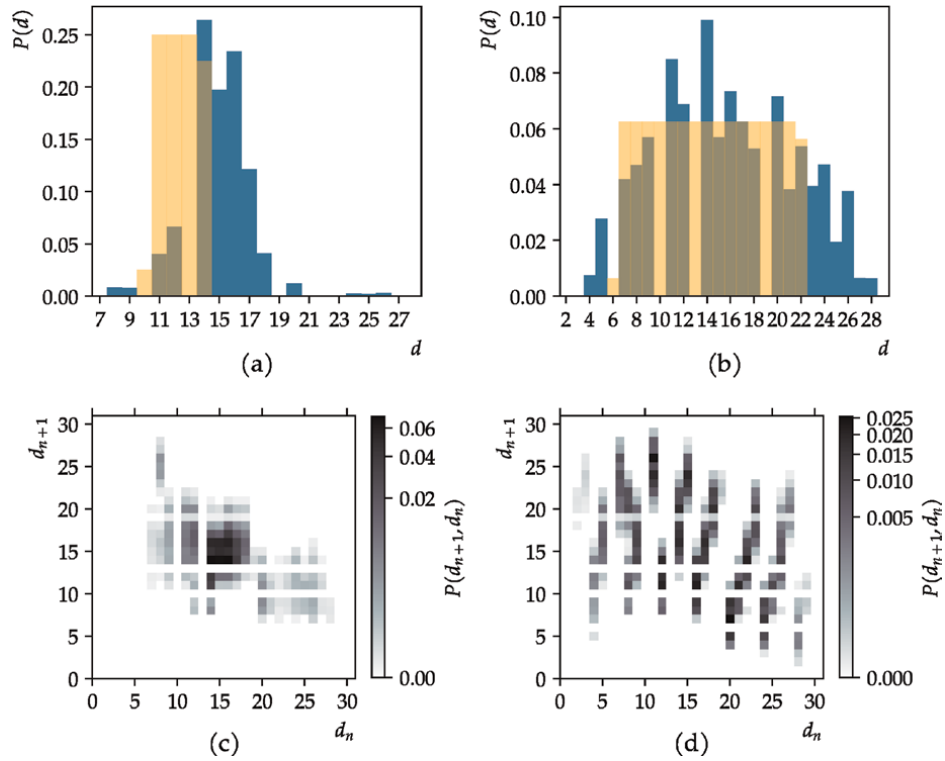


Figure 9.
Histograms from $\{d_n\}$ (data sequence from the ADC output in oscillation).

(or $P(d_{n+1}|d_n = d)$ for all scanned d values) would immediately reflect the shape of $\tilde{M}(\cdot)$ and $\tilde{e}(\cdot)$, as illustrated in **Figure 9c** and **d** that again correspond to the maps in **Figure 5a** and **b**. To some extent, this enables an estimation of the overall converter pair error at specific codes. In fact, what is observed at some abscissa d in the 2D histograms is the consequence of a times the ADC + DAC overall error in the neighborhood of the analog value corresponding to that code, plus the ADC error at d_{n+1}. As long as $|a|$ is sufficiently large, the first component can be expected to dominate.

Note that because the collection of full two-dimensional histograms can be expensive, one can just track of the extremes of the support of $P(d_{n+1}|d_n = d)$ for all d values, rather than building the full histograms.

To summarize, by looking at the loop *digital output* one can see if the ADC + DAC chain error gets *too large* either on the span of the effective IS of the system (with a 1D histogram) or in the neighborhood of some code (with a 2D histogram).

The main limitations of the proposed technique are obviously twofold:

1. the converters end up being tested limited to the IS of $\tilde{M}(\cdot)$ rather than in their full range; and

2. the collected data provide information, where the ADC and DAC errors get "mixed" in a complicated and hard to untangle way.

With respect to the first point, it is clear that the strategy described at the end of Section 2, based on the usage of just a few of the MSBs of the DAC to close the loop,

can be helpful since it regularizes the shape of $\tilde{M}(\cdot)$ so enabling a better control of the actual IS. In addition to that, rather than trying to make the IS as large as possible, it may be convenient to operate at a reduced value of $|a|$ that lets the IS be smaller and then shift the IS across the converters' range by varying b (to some extent an equivalent effect can be obtained by digitally offsetting the output of the ADC before passing it to the DAC). Clearly, being able to digitally control both a and b (e.g., by having a provided by a programmable gain amplifier and b by a programmable reference) would get the best flexibility, while coming at the cost of a complication in the architecture.

For what concerns the second point, expectably it makes a punctual evaluation of the data converter errors difficult. However, it does not rule out the existence of a monotonic relationship, where larger errors correspond to larger histogram excess spans. In turn, this can enable an empiric determination of a threshold on the histogram excess spans above which the ADC + DAC chain can be considered suspicious. Additionally, adopting the strategy described at the end of Section 2 and using 2D histograms can hint at the specific sub-range of the data converter, where the threshold is exceeded.

An open problem is clearly to make the best possible usage of the information provided by the $\{d_n\}$ sequences during oscillation, going beyond a mere "good enough"/"not good enough" thresholding mechanism. We conjecture that *diversity* might be exploitable in this sense in conjunction with 2D histograms. Namely, being able to rely on multiple test runs having multiple converters to mix and match or adopting different b values could provide sufficient data to better localize and quantify errors in the data converter characteristics. In other words, multiple measures taken in different conditions may allow decoupling the error sources that contribute to the histogram excess spans.

As a final consideration, it is worth noticing that even without histogram methods, the mere testing of the entropy source described in the previous section evidently happens to be a proxy for validating the data converters (because a high entropy per symbol can only be obtained if the chaotic map is close to the ideal one, which in turn is a condition satisfied only if the data converters behave accurately).

5. Experimental results from a microcontroller-based prototype

For the validation of the approach, data collected from a prototype system based on a low-cost PIC microcontroller has been used. The prototype system uses the microcontroller ADC as part of the loop. In principle, it could have been possible to also take advantage of the on-board DAC or even to use a SOC with programmable analog primitives to further increase the overall level of integration of the prototype. However, in prototyping, preference was given to close the loop with external components, in order to enable a direct observation of the analog state via bench instrumentation to simplify debugging. Overall, the prototype follows the setup in [22].

The overall architecture of the prototype is shown in **Figure 10**. The microcontroller incorporates a 10-bit ADC paired to an external 10-bit DAC. Communication between the microcontroller and the DAC is achieved via the microcontroller serial peripheral interface bus. A *linear combiner* implemented by means of an operational amplifier circuit provides the gain a and the offset b. In the prototype, both are fixed in the sense that they are determined by resistor ratios, so that their variation requires

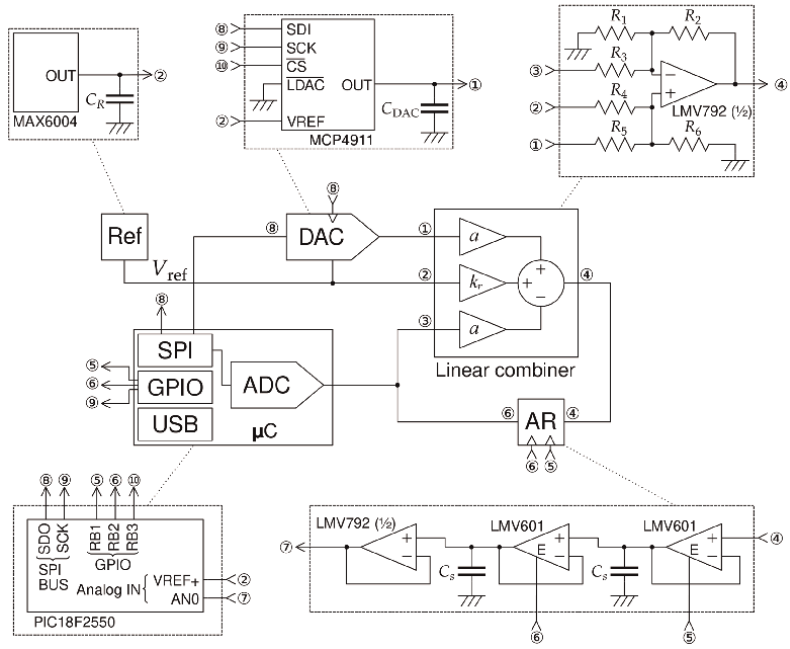


Figure 10.
Architecture of the test system.

changing the resistors. Specifically, b is obtained by scaling a reference voltage, where the scale is indicated as k_r in the diagram. The analog register is implemented by the cascade of two track-and-hold blocks operating in phase opposition, the phases being generated by the microcontroller itself.

Data collection has been practiced running the system in a configuration in which only a portion of the ADC digital output gets transmitted to the DAC. In fact, the strategy illustrated at the end of Section 2 has been adopted to enable a more accurate control or the IS. Specifically, the lower seven bits of the ADC output have been masked to zero before passing the ADC output to the DAC input. Gain a has then been set to -4, with offset b at 0.18. The system has been operated at cycle rates up to some tens kilocycles/s. This limit has been adopted both to avoid introducing errors due to the settling of the analog register and to simplify the transmission of the experimental data to a personal computer via the microcontroller USB port for subsequent analysis.

Before commenting the experimental data, it is worth underlining that even if the latter is obtained from the actual operation of a prototype circuit, only a single instance has been tested, with a single tuning of the system parameters. Furthermore, the nature of the prototype made it impossible to simulate or inject faults in the data converters. As a consequence, the results in the following cannot comprise comparative analyses, but shall be interpreted as an illustration of the information that can be gathered from a real system.

5.1 Operation as an entropy source

Following Section 3, having $|a| = 4$ would, in ideal conditions, enable the extraction from the system of one random symbol $\mathcal{S}$ per cycle defined on a four-valued

alphabet $\{S_0, S_1, S_2, S_3\}$ and thus suitable for perfect encoding in two bits. In other words, in ideal conditions, the test system should be able to deliver $\log_2 4 = 2$ random bits per cycle.

For building the output symbols, an output function $\hat{g}(\cdot)$ has thus been defined operating on the ADC output data, which can be done by merely selecting the two LSBs at each cycle. This is in agreement with the considerations at the end of Section 3. Even if better performing output functions can in some cases be obtained [22], this is also the simplest and most obvious choice. Clearly, in any real case, the extracted symbols will not be perfectly random. This is to be expected and consistent with the goal to obtain an entropy source and not a TRNG in its own. Consequently, in the following, the validation of the output data has not been based on randomness tests, rather on quantitative entropy estimations and on distribution plots.

Figure 11a shows the distribution of the two-bit symbols produced by the system at each cycle. This is a visual indication of *balance* (or lack thereof) in the output data. Evidently, although modest, some unbalance is present. To get an ideal of *correlations*, whole output sub-sequences need to be examined. To this aim, it is convenient to preliminary pack the output bits B_s in binary words W_s. For instance, if the words W_s are *bytes*, in the test system, four cycles will be used to generate a word, and the analysis of two consecutive words $W_{\hat{n}}$ and $W_{\hat{n}+1}$ will be equivalent to consider eight-cycle sub-sequences. **Figure 11b** shows an histogram estimation of the probability of extracting any two consecutive bytes, namely, $P(W_{\hat{n}+1}, W_{\hat{n}})$. Ideally, this diagram should be perfectly uniform. In practice, a texture is evident. The fact that the texture is orthogonal to the axes and stays the same when observed from both axes is a visual indication that the probability of $W_{\hat{n}+1}$ is not significantly conditioned by $W_{\hat{n}}$. In other words, this plot confirms unbalance, but does not show visible correlations. This is in agreement to the general knowledge that in chaotic systems correlations are quickly (exponentially) vanishing [6].

From a quantitative point of view, entropy estimators applied on the system output bits B_s revealed entropy rates always in excess of 0.96 entropy bits per output bit, which is a perfectly acceptable (and in fact quite favorable) value for the subsequent entropy distillation. This value has been obtained by processing data obtained by running the prototype for some million cycles.

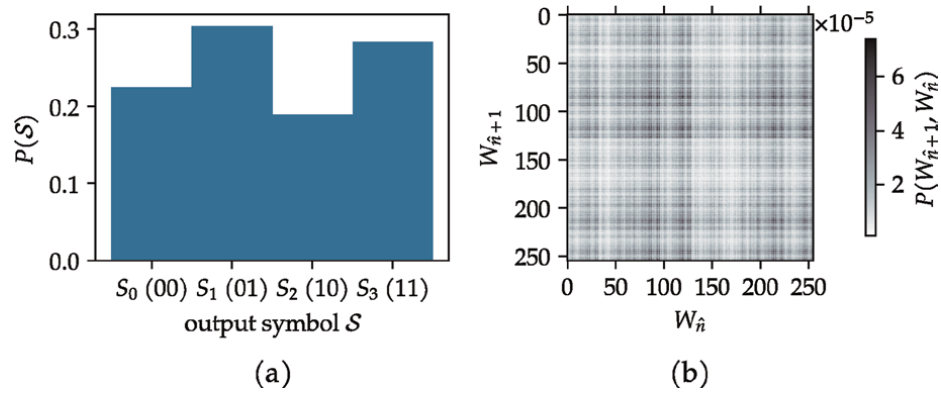


Figure 11.
Visual analysis of the quality of the output bitstreams produced by the test system. In (a), the distribution of the output symbols can be observed (each symbol is encoded on two bits in the output bistreams). In (b), the 2D distribution of consecutive output bytes is observed.

5.2 Operation for OBT

To appreciate data converter errors from the oscillating system, one can build and observe 1D and 2D histograms such as those in Section 4. The plots in **Figure 12** are obtained from the experimental data collected at the ADC output on 32×10^6 cycles. In particular, **Figure 12a** shows the estimated probability distribution of d and compares it with similar data estimated from the simulation of an ideal system on the same number of cycles. The *fuzziness* on the top of both plots is an artifact of the estimation, due to the finite length of the adopted sequence and the large number of bins. It should not be of concern (indeed, also the plot based on the synthetic ideal data looks the same in this respect). What is to be observed is the extra span of the experimental plot in comparison with the ideal one. This is approximately as large as 30 code points. Thererore, it indicates a peak deviation of the full (ADC + DAC) characteristics from the ideal one as large as $30/|a| - 1/2 \approx 7$ quantization steps *in dynamic conditions* (including noise and the errors of the other elements in the signal processing chain). This corresponds to errors involving the lower $\log_2 7 \approx 2 - 3$ bits of the data converters. This number seems reasonable for a setup incorporating an ADC integrated inside a microcontroller. In fact, microcontroller data converters often have ENOBs significantly lower than the tabled resolution. Furthermore, one should not forget that errors in the components used to close the loop and caused by dynamic behavior are also incorporated in this figure.

Figure 12b provides a 2D histogram in the same lines of the illustrative one in **Figure 9d**. Again the experimental data are overlapped to data from an ideal system. The quality by which the data converters quantization error plot is reconstructed is surprisingly remarkably good. It and lets one appreciate how errors occur around specific code-points. For instance, the real converter-pair characteristics tend to be evidently above the nominal one around code 430 and evidently below it at codes above 650. This information is to be taken with some care because the visible deviations at some d_n are actually due to errors on the whole loop (including not just the data converters but also other elements). Furthermore, as noted in Section 4, ADC errors are accounted twice, both at code d_n (scaled by $|a|$) and at code d_{n+1} since the system state x is observed through the ADC reading. Indeed, this mixing of different

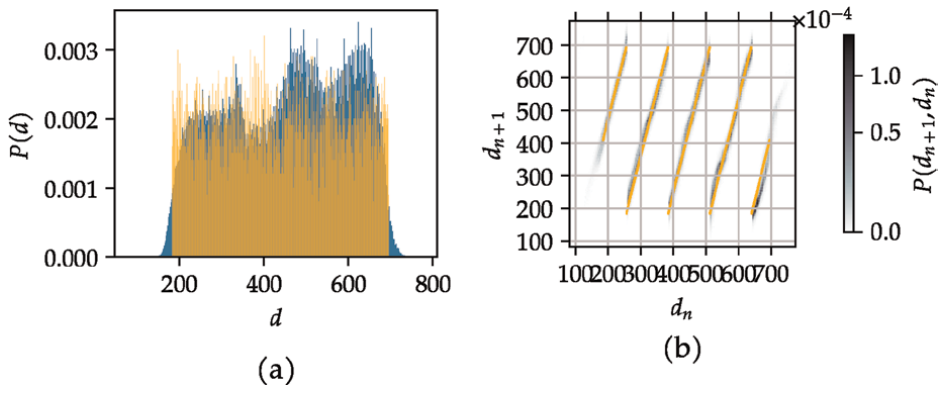


Figure 12.
Visual analysis of data from the ADC output in oscillation, for the purpose of evaluating the correct operation of the data converters. In (a), the distribution of the data points at the ADC output (in blue), compared with equivalent data for a synthetic chain using ideal data converters (in orange). In (b), a 2D histogram permitting reconstruction of the actual quantization error characteristics. Also in this case the real data (blue-black) is overlapped to the ideal one (orange).

error sources is the main limit of the method. A second limit is that only a portion of the data converter range can be tested at once: for instance in the test case, only codes from 160 to 720 get tested. Changing the b value would let the tested range be moved around across the whole converter scale.

6. Conclusions

The possibility of bringing a data converter pair into chaotic self-oscillation has been discussed both by the setup of a theoretical framework and via experimental tests on a prototype system, with the aim of showing the suitability of this setup both for entropy generation (in view of true random number generation) and for the built-in self-test of the data converters themselves in an OBT arrangement.

The usage of this setup for entropy generation appears reliable and mature enough for practical applications. Furthermore, utilization for data converter testing shows promise, and the techniques described in the paper for examining the collected data appear meaningful when applied on experimental data. Some open problem remains, though. One is in the very validation of the approach that would require simulation setups or prototype systems where errors can be injected to assess the actual ability of the approach to fully reveal them. Another open problem lays in the interpretation of the data that can be collected during oscillation. While these data definitely reveal misbehavior, errors from different origins can get mixed up to the point that decoupling them can be extremely hard. One can expect that exploiting *diversity* by working on multiple data sets (e.g., mixing and matching different ADCs and DAC or working with different parameters in the loop used to establish the oscillating behavior) may help better isolating the individual error sources and quantifying their magnitude. This will be one of the first objectives of future research.

Author details

Sergio Callegari
Advanced Research Center on Electronic systems (ARCES) and Department of Electrical, Electronic and Information Engineering (DEI), University of Bologna, Bologna, Italy

*Address all correspondence to: sergio.callegari@unibo.it

IntechOpen

© 2024 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

References

[1] Chua LO. The genesis of Chua's circuit. AEÜ Archiv für Elektronik und Übertragungstechnik. 1992;**46**(4):250-257

[2] Baillieul J, Brockett R, Washburn R. Chaotic motion in nonlinear feedback systems. IEEE Transactions on Circuits and Systems. 1980;**27**(11):990-997

[3] Rodriguez-Vazquez AB, Huertas JL, Chua L. Chaos in a switched-capacitor circuit. IEEE Transactions on Circuits and Systems. 1985;**32**(10):1083-1085

[4] Rodriguez-Vazquez A, Huertas JL, Rueda A, Perez-Verdu B, Chua LO. Chaos from switched-capacitor circuits: Discrete maps. Proceedings of the IEEE. 1987;**75**(8):1090-1106

[5] Lasota A, Mackey MC. Chaos, fractals and noise. In: Stochastic Aspects of Dynamics. 2nd ed. New York: Springer-Verlag; 1995

[6] Setti G, Mazzini G, Rovatti R, Callegari S. Statistical Modeling of discrete time chaotic processes—Basic finite-dimensional tools and applications. Proceedings of the IEEE. 2002;**90**(5):662-690

[7] Bean JT, Langlois PJ. A noise generator based on chaos for a neural network application. In: Proc. of NDES'93 (Nonlinear Dynamics and of Electronics Systems Conference). Dresden, DE: World Scientific; 1993. pp. 236-243

[8] Kennedy MP, Kolumbán G, Jákó Z. Chaotic modulation schemes. In: Kennedy MP, Rovatti R, Setti G, editors. Chaotic Electronics in Telecommunications. Boca Raton, USA: CRC International Press; 2000. pp. 151-183

[9] Mazzini G, Setti G, Rovatti R. Chaotic complex spreading sequences for asynchronous DS-CDMA. I. System modeling and results. IEEE Transactions on Circuits and Systems, Part I. 1997; **44**(10):937-947

[10] Callegari S, Rovatti R, Setti G. Chaotic modulations can outperform random ones in EMI reduction tasks. Electronics Letters. 2002;**38**(12):543-544

[11] Kolumban G, Vizvari B, Mogel A, Schwartz W. Chaotic systems: A challenge for measurement and analysis. In: Joint Conference: IEEE Instrumentation and Measurement Technology Conference and IMEKO Technical Committee 7. Vol. 2. Brussels, Belgium: IEEE; 1996. pp. 1396-1401

[12] Addabbo T, Fort A, Rocchi S, Vignoli V. Histogram test of ADCs with chaotic samples. In: 2010 IEEE Instrumentation & Measurement Technology Conference Proceedings. Austin, TX, USA: IEEE; 2010. pp. 546-549

[13] Arabi K, Kaminska B, Inventors; Opmax Inc., Assignee. Oscillation-based test method for testing an at least partially analog circuit. US patent 6005407. 1999. Filed in 1995

[14] Arabi K, Kaminska B. Oscillation-test strategy for analog and mixed-signal integrated circuits. In: Proceedings of the 14th IEEE VLSI Test Symposium. Princeton, NJ: IEEE; 1996. pp. 476-482

[15] Huertas Sánchez G, García V, de la Vega D, Rueda Rueda A, Huertas Díaz JL. Oscillation-based test in mixed-signal circuits. In: Frontiers in Electronic Testing. Dordrecht: Springer; 2006

[16] Callegari S, Pareschi F, Setti G, Soma M. Complex oscillation-based test

and its application to Analog filters. IEEE Transactions on Circuits and Systems—Part I: Regular Papers. 2010;**57**(5): 956-969

[17] Bernstein GM, Lieberman MA. Secure random number generation using chaotic circuit. IEEE Transactions on Circuits and Systems. 1990;**37**(9): 1157-1164

[18] Taylor G, Cox G. Digital randomness. IEEE Spectrum. 2011; **48**(9):32-58

[19] Callegari S. Evaluation of a couple of true random number generators with liberally licensed hardware, firmware, and drivers. In: Proc. of ICECS 2015. Cairo, Egypt: IEEE; 2015. pp. 197-200

[20] Gerosa A, Bernardini R, Pietri S. A fully integrated chaotic system for the generation of truly random numbers. IEEE Transactions on Circuits and Systems—Part I: Fundamental Theory and Applications. 2002;**49**(7):993-1000

[21] Callegari S, Rovatti R, Setti G. Embeddable ADC-based true random number generator for cryptographic applications exploiting nonlinear signal processing and chaos. IEEE Transactions on Signal Processing. 2005;**53**(2):793-805

[22] Callegari S, Fabbri M, Beirami A. Very low cost chaos-based entropy source for the retrofit or design augmentation of networked devices. Analog Integrated Circuits and Signal Processing. 2016;**87**(2):155-167

[23] Lechner A, Richardson A. 3. In: Huertas JL, editor. Test of A/D Converters. Boston, MA: Springer US; 2004. pp. 73-98. DOI: 10.1007/978-0-387-23521-9_4

[24] Arabi K, Kaminska B. Efficient and accurate testing of analog-to-digital converters using oscillation-test method. In: Proceedings of the European Design and Test Conference, ED&TC'97. Paris, France: IEEE; 1997. pp. 348-352

[25] Turan MS, Barker E, Kelsey J, McKay KA, Baish ML, Boyle M. Recommendation for the Entropy Sources Used for Random Bit Generation. Gaithersburg, MD: National Institute for Standards and Technology; 2018. SP 800-90B. DOI: 10.6028/NIST.SP.800-90B

[26] Beirami A, Nejati H, Callegari S. Fundamental performance limits of chaotic-map random number generators. In: Proc. of the 52nd Annual Allerton Conference on Communication, Control, and Computing. Monticello, IL, USA: IEEE; 2014. pp. 1126-1131

[27] Beirami A, Nejati H. A framework for investigating the performance of chaotic-map truly random number generators. IEEE transactions on circuits and systems—Part II: Express. Briefs. 2013;**60**(7):446-450

[28] Rukhin A, Soto J, Nechvatal J, Smid M, Barker E, Leigh S, et al. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. Gaithersburg, MD: National Institute for Standards and Technology; 2010. SP 800-22 Rev. 1a. Special publication. Available from: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-22r1a.pdf

[29] NIST. Security Requirements for Cryptographic Modules. Gaithersburg, MD: National Institute for Standards and Technology; 2001. FIPS 140-2. Available from: http://www.itl.nist.gov/fipspubs/by-num.htm

[30] Callegari S, Setti G. ADCs, chaos and TRNGs: A generalized view exploiting

Markov chain Lumpability properties.
In: Proc. of ISCAS. New Orleans, LA
(USA): IEEE; 2007. pp. 213-216

[31] Ott E. Chaos in Dynamical Systems.
New York: Cambridge University Press;
1993

Chapter 5

Perspective Chapter: Behavioral Analysis of Nonlinear Systems and the Effect of Noise on These Systems

F. Setoudeh and M.M. Dezhdar

Abstract

One of the crucial concepts in determining the structure of dynamic systems is to recognize the behavior of nonlinear systems, which is one of the current issues in engineering sciences. In general, nonlinear systems exhibit behaviors such as stability, periodic, quasi-periodic and chaotic. Since in nonlinear systems, changing parameters can have a great effect on changing the behavior of nonlinear systems, for this reason, it has been studied how different parameters affect the behavior of a system. Due to the importance of determining the behavior of nonlinear systems, in this chapter, first, various criteria for estimating the behavior of nonlinear systems are discussed and then the effect of these parameters on these systems is examined.

Keywords: nonlinear system, behavioral analysis, stability, chaos, periodic, noise

1. Introduction

For researchers in engineering sciences, investigating and determining the qualitative results of dynamic systems is of particular importance, which is often studied and examined using theoretical theories, differential equations, and other tools. One of the key concepts in analyzing dynamic systems is identifying the behavior of nonlinear systems, which is among the current topics in engineering sciences. For this purpose, many researchers in engineering sciences are interested in studying how different parameters affect the behavior of a system.

Nonlinear dynamics and complex systems are a very broad subject that essentially falls into an interdisciplinary research field. This issue includes mathematics, physics, chemistry, medicine, engineering sciences, and so on. Complex systems are composed of numerous components, each of which may interact with one another and even external factors, resulting in diverse interactions. For example, water and air, human organs, living organisms, infrastructure such as electrical grids, complex software, electronic systems, ecological systems, cellular systems, and ultimately all kinds of networks can be considered complex systems, each with their own components and external interactions. Modeling the behavior of complex systems is challenging due to

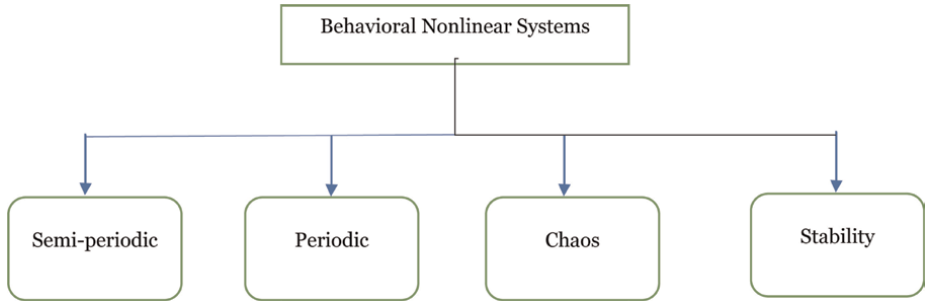


Figure 1.
Behavioral nonlinear systems.

their interdependence and interactions among components or between a specific system and its environment. Complex systems possess distinctive features such as nonlinearity, self-organization, feedback, robustness, adaptability, Emergence, and network-like structures that result from diverse interactions among their components. In the past, researchers discovered that the relationship between nonlinear dynamics and complex systems in engineering sciences is such that simple systems (with a few variables) exhibit stable behavior, while complex systems (with many variables) exhibit unstable behavior. However, it has recently been found that even a simple nonlinear system can exhibit irregular behavior [1–4]. This means that a system displaying irregular behavior is considered a complex system, but it is possible for a simple nonlinear system to exhibit irregular behavior. In general, nonlinear systems exhibit behaviors such as stability, periodic, semi-periodic, and chaotic, which are detailed in this section (refer to **Figure 1**) [1–4].

The distinction between the behaviors of linear and nonlinear systems is about their stability. The behavior of non-switching linear systems is not sensitive to initial conditions, while in some cases of nonlinear systems, the equilibrium point of the system may be stable. However, for some initial conditions, the system response may be converged (stable), while for others, it may be divergent (unstable), and identifying these types of systems is not an easy task [2–6]. One type of nonlinear systems behavior is chaos [3], which is unpredictable and represents order within disorder [1, 2]. Chaotic systems, although they appear random, belong to the category of deterministic systems, which are distinct from stochastic systems that are random by nature. In chaotic systems, a small change in initial conditions can lead to significant changes in the output [2–7], which is a characteristic feature of chaos. Another behavior of nonlinear systems is periodic. In general, this behavior occurs in dynamic nonlinear systems around a fixed point with a specific amplitude and frequency [6–9]. Another behavior of nonlinear systems is semi-periodic, which is formed from the combination of several periodic behaviors [9–14].

2. Detecting chaos in the time domain

One of the common methods for detecting chaos is to use the patterns present in the time series [15]. One of the simplest methods in the time domain is to plot the time series in phase space and then observe the created pattern to determine the signal behavior. Detecting periodic behavior in phase space is very simple, but the behavior

of chaotic and quasi-periodic systems in phase space is very complex and unpredictable. This method alone is not a precise way to detect the presence of chaos, as random systems and real dynamic systems mixed with noise also exhibit similar behavior in phase space.

Lyapunov exponents of a system are a set of non-changing geometric measures that directly express the system's dynamics. One of its applications is in detecting the phenomenon of chaos in a system and also as a measure of the chaotic nature of behavior. The topic of chaotic behavior is qualitative as far as it relates to sensitivity to initial conditions and structural instability. However, in studying a system, we have information about its behavior in the form of a differential equation, a recursive mapping, or a time series, and therefore it is necessary to have analytical or quantitative methods for detecting chaos in any system so that we can distinguish chaotic behavior from random noise-like behavior. Additionally, this method should be able to provide both a measure and a quantity for the degree of chaos in the system. Describing the quantitative sensitivity of a system's behavior to initial conditions in chaotic situations is possible by introducing Lyapunov exponents [5].

Lyapunov exponents are a set of non-changing geometric measures that directly express the dynamics of systems. The Lyapunov exponent is calculated as follows:

Consider two neighboring points in phase space at times zero and t, where the distance between the points in the direction of (i) is $\|\delta x_i(0)\|$ and $\|\delta x_i(t)\|$, respectively. The Lyapunov exponent is defined as follows:

$$\frac{\|\delta x_i(t)\|}{\|\delta x_i(0)\|} = e^{\lambda_i t} \rightarrow \lambda_i = \lim_{t\to\infty} \left(\frac{1}{t} \ln \frac{\|\delta x_i(t)\|}{\|\delta x_i(0)\|} \right) \tag{1}$$

In this equation, λ_i represents the Lyapunov exponent. As can be seen, two points with infinitesimally small proximity in the initial state, diverge significantly from each other in the direction of the (i). This phenomenon is referred to as "sensitivity to initial conditions."

a. If the $\lambda_i < 0$, then we will have a stable fixed point or a stable periodic cycle. In other words, all selected initial points will converge toward a fixed point or a periodic cycle. These systems are called asymptotically stable. With a negative increase, $\lambda_i \rightarrow -\infty$, the stability of the system increases, such that for $\lambda_i = -\infty$, there exists a super-stable fixed point or periodic cycle.

b. If $\lambda_i = 0$, the system only oscillates around a fixed point. In this case, every selected initial point oscillates around a stable limit cycle.

c. If $\lambda_i < 0$, there are no stable fixed points or limit cycles; in fact, the points are unstable, but the system is bounded and chaotic. In other words, if the largest Lyapunov exponent in the system is positive, the system is chaotic; otherwise, it is not chaotic [7, 11, 16, 17].

Chaotic systems have unique properties that distinguish them from other dynamics. One of the distinctive features of chaotic systems is their strong dependence on initial conditions. One powerful tool for detecting chaos is the Lyapunov exponent.

A variety of methods have been presented for calculating the Lyapunov exponent, one of which is the calculation of the Lyapunov exponent using time series [7]. However, Lyapunov exponent is highly sensitive to noise, which is why using

Lyapunov exponent as a criterion for detecting chaos in noisy environments is not a good measure [7–9]. As seen, chaotic dynamics create complex attractors that are limited to a part of space and cannot cover the entire space. Various methods have been proposed to calculate the dimensions of chaos and hidden dynamics, including correlation dimensions and fractal dimensions. One of the disadvantages of these methods is their computational complexity and their dependence on noise. Another method is to use R/S analysis and the Hurst exponent [18]. This method is used to distinguish between random time series and non-random and chaotic time series. The disadvantages of this method include computational complexity and the inability to detect chaos in noisy environments. Another method for analyzing chaotic signals is to use the Kolmogorov-Sinai entropy. This law talks about the disorder in a system. A random signal has the most disorder, but a deterministic system has the most order. This entropy is related to the Lyapunov expressions of the signal. This entropy is the average of the positive Lyapunov exponents of the system. One of the limitations of this method is its high sensitivity to noise [19].

3. Detection of nonlinear systems behavior

Many issues in various fields, including electrical engineering, are inherently nonlinear and are modeled using partial and ordinary differential equations. Only a limited number of these equations have exact solutions, and most of these problems do not provide precise answers, necessitating the use of novel methods for their analysis. Based on the observed time series of a process, detecting the presence of nonlinearity is quite challenging. Therefore, efforts have been made to provide tools for identifying the behavior of nonlinear systems. The behavior of nonlinear systems in phase space is highly intricate. A simple approach to understanding this behavior is to plot the time series in phase space and analyze the patterns created. In this section, various criteria for analyzing the behavior of nonlinear systems are introduced.

3.1 Calculation of the Lyapunov exponent based on the analytical method of differential transformation and behavioral analysis of nonlinear systems with respect to unknown parameters

In order to analyze the behavior of nonlinear dynamic systems, a method for calculating the Lyapunov exponent based on various analytical differential transformation methods has been proposed. In this process, using the differential transformation method, the time series of the desired system is calculated and replaced in relation to the Lyapunov exponent, and then the behavior of the system is determined using the calculated Lyapunov exponents. In this section, the Lyapunov exponent method based on three representative limit cycle algorithms has been proposed, with the aim of clarifying them further. Since the behavior of nonlinear dynamic systems depends on the changes in their parameters, for example, in a Colpitts oscillator, the behavior of the oscillator depends on the changes in parameters such as Inductor and capacitor value, therefore, to ensure that a nonlinear system has the desired behavior, the most effective approach is to accurately adjust the parameters, if possible. Consequently, the analysis of the behavior of nonlinear dynamic systems with respect to changes in various parameters of the systems has been studied using the Lyapunov exponent method based on the differential transformation method.

3.1.1 Calculation of the Lyapunov exponent based on the classical analytical differential transformation method

The aim of this section is to determine the different behaviors of nonlinear dynamic systems using the Lyapunov exponent method based on time series. Accordingly, changes in the Lyapunov exponent for unknown parameters have also been investigated. This process, using the classical differential transformation method and the Lyapunov exponent method based on time series, is described in the context of method-1.

Method-1: Let us assume that the time series $\{x_1(t), x_2(t), \ldots, x_n(t)\}$, is the solution of the main Eq. (1), and the following expression,

$$\lambda_j = \frac{1}{T} ln \left| \frac{x_j(r+T) - x_j(s+T)}{x_j(r) - x_j(s)} \right|, \quad j = 1, 2, \ldots, N. \tag{2}$$

λ_j represents the Lyapunov exponent of the system under consideration, in which T, the time of evolution, and r and s are two selected nearby sample points on the path. In this case, by applying the differential transformation method, the Lyapunov exponent of the system in the time series space will be as follows.

$$\lambda_j = \frac{1}{T} ln \left| \frac{X_j(1)(r-s) + \sum_{k=2}^{n} X_j(k)\left[(r+T)^k - (s+T)^k\right]}{X_j(1)(r-s) + \sum_{k=2}^{n} X_j(k)[r^k - s^k]} \right|, \quad j = 1, 2, \ldots, N \tag{3}$$

Proof: Clearly, by substituting the point "N" from the time series into the differential transformation method $\left(x(N) = \sum_{k=0}^{n} X_j(k) N^k\right)$ for the definition of the corresponding Lyapunov coefficient, the desired relationship is obtained.

In continuation, with the help of presenting the algorithmic steps, we describe the use of method-1.

Algorithm 1: Calculation of the Lyapunov Exponent Based on the Classical Differential Transformation Method.

First step: Applying the differential transformation method to the system state equations (calculating the recursive relationships).

$$X_j(k) = \frac{1}{k!} \left[\frac{d^k x_j(t)}{dt^k} \right]_{t=t_0}, \quad j = 1, 2, \ldots, N. \tag{4}$$

Second step: Calculating the limited time series using the differential transformation method.

$$x_j(t) = \sum_{k=0}^{N} X_j(k)(t - t_0)^k, \quad j = 1, 2, \ldots, N. \tag{5}$$

Third step: Calculating the Lyapunov exponent based on the differential transformation method using method-1.

$$\lambda_j = \frac{1}{T} ln \left| \frac{X_j(1)(r-s) + \sum_{k=2}^{n} X_j(k)\left[(r+T)^k - (s+T)^k\right]}{X_j(1)(r-s) + \sum_{k=2}^{n} X_j(k)[r^k - s^k]} \right|, \quad j = 1, 2, \ldots, N \tag{6}$$

Fourth step: Analysis of the changes in Lyapunov exponents with respect to unknown parameters and identification of different system behaviors [20–22].

3.2 Identification of behavior based on analysis in the frequency domain

One method for detecting chaotic behavior from periodic behavior involves using frequency domain analysis. This process entails plotting the frequency spectrum of a time series, revealing features that may not be easily observed in the time domain. In periodic signals, energy concentrates at specific frequencies, while chaotic behaviors exhibit a frequency spectrum with non-zero values across various frequencies, creating a wide band. In deterministic systems, a wide band spectrum can signal the onset of chaos. However, it's essential to note that relying solely on frequency analysis is not always accurate for determining the presence of chaos. For instance, power spectrum characteristics are also employed in frequency domain analysis. Additionally, it's worth mentioning that the frequency spectrum of a random time series or a time series from real dynamic systems affected by noise will also exhibit a wide band, making it challenging to distinguish between these scenarios based solely on the frequency spectrum [23].

3.3 Nonlinear system behavior detection based on time-frequency analysis

As mentioned, the methods for detecting chaos in the time and frequency domains are not robust against noise. Another method used for chaos detection is time-frequency analysis. One of these methods is based on the short-time Fourier transform. The main idea of the short-time Fourier transform is to multiply the input signal x(t) by a window function $w(\tau)$, which changes its location with time. In other words, the signal is divided into short-time segments, and the Fourier transform is applied to each segment. In this way, each frequency spectrum shows the frequency content in a short-time period. Such a spectrum includes the change of frequency content with time. The short-time Fourier transform is defined as follows [24, 25]:

$$STFT(t,f) = \int_{-\infty}^{\infty} x(t+\tau)w(\tau)e^{-j2\pi f\tau}d\tau \tag{7}$$

where $x(t)$ is the input signal, $w(\tau)$ is the window function with the width of the T, and X(τ,ω) is the complex-valued spectrum.

$$w(\tau) = \begin{cases} 1 & \text{for} \quad m(t) \geq c(t) \\ 0 & \text{for} \quad m(t) \leq c(t) \end{cases} \tag{8}$$

Short-time Fourier transform determines which frequency components and at what times are present in the signal. The algorithm for detecting chaos based on short-time Fourier transform is as follows:

1. Calculate the short-time Fourier transform

2. Estimate the dominant spectral components frequency

The frequency at which the short-time Fourier transform frequency of the signal is maximized is calculated according to the following equation:

$$f_m(t) = \arg\max_f \mathrm{STFT}(t, \mathrm{f}) \tag{9}$$

where, f represents the frequency and t the time.

3. We define the $u_\Omega(t,f)$ as follows:

$$u_\Omega(t,f) = \begin{cases} 1 & \mathrm{STFT}(t, \mathrm{f}) \geq 0.01 max_f |\mathrm{STFT}(t, \mathrm{f})| \\ 0 & \text{elsewhere,} \end{cases} \tag{10}$$

4. In order to detect chaos, we define the following function:

$$\mathrm{m(t)} = \int_0^{f_m(t)} u_\Omega(t,f)\mathrm{d}f \tag{11}$$

5. To detect chaos, we operate as follows:

$$d(t) = \begin{cases} 1 & \text{for } \mathrm{m(t)} \geq \mathrm{c(t)} \\ 0 & \text{for } \mathrm{m(t)} \leq \mathrm{c(t)} \end{cases} \tag{12}$$

If $d(t) = 1$, it indicates a chaotic signal and if $d(t) = 0$, it indicates a periodic signal. In this case, $\mathrm{c(t)}$ is the threshold value that depends on the selected window width. This method is resistant to noise and is used to detect chaos in a noisy environment. One of the drawbacks of this method is its strong dependence on the selected window width. On the other hand, this method cannot detect random signals such as noise. Another new method that has been recently used is the use of energy distribution based on continuous wavelet transform in chaotic systems [26]. Continuous wavelet transforms of $x(t)$ signal is defined as:

$$W_x(a,b) = \langle x, W \rangle = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t)\overline{\Psi}\left(\frac{t-b}{a}\right)dt \tag{13}$$

In which a and b are the scale and shift parameters, respectively and $\overline{\Psi}\left(\frac{t-b}{a}\right)$ is the complex conjugate of the $\Psi\left(\frac{t-b}{a}\right)$ function.

The concept of shift in the wavelet transform is similar to the concept of time shift in the short-time Fourier transform. The shift represents the amount of window displacement and contains the time information of the transform. However, unlike the short-time Fourier transform, the wavelet transform does not have a direct frequency parameter. Instead, it uses a scale parameter that is inversely related to the frequency. In other word: $a = \frac{1}{f}$

A function ψ is called a wavelet if:

a. It has a nonzero value only in a certain range (small wave).

b. It has a limited frequency range.

In this case, the energy distribution based on continuous wavelet transform to detect chaos is used. The algorithm for detecting chaos is as follows:

First, we sample the signal. In this case, we divide the signal into N parts which are equal. Then, for each part, we apply the continuous wavelet transform and calculate the wavelet coefficients and the energy of each part from the coefficients of the continuous wavelet transform according to Parseval's theorem. In chaotic systems, the energy distribution changes irregularly. This method is not resistant to noise [26].

4. A recent criterion for distinguishing chaos from noise

In recent years, various nonlinear electronic systems that exhibit chaotic behavior have been studied. Some behaviors that are hidden in normal conditions due to the presence of noise are real examples of chaotic behavior of a completely deterministic nature. One of these systems is oscillators. The output signal of ideal oscillators is a periodic function in the time domain and especially at high frequencies is usually sinusoidal. Also, regardless of factors that are usually negligible such as heat and wear, it can be said that the frequency and amplitude of this signal are always constant. But anyway, in practical oscillators, these undesirable noise effects cause minor disturbances in the frequency, phase, and amplitude of the output signal. On the other hand, oscillators can show chaotic behavior for some parameters. For this reason, chaos analysis in oscillators and its separation from noise is of great importance.

Chaotic systems exhibit behavior that resembles random processes, yet they are non-random. Chaotic series are a subset of nonlinear processes known for their high complexity and irregular behavior. Although chaotic time series may appear random, they possess distinct properties that set them apart from truly random series. One key characteristic of chaotic processes, which differentiates them from random processes, is their sensitivity to initial conditions. Even a slight error in measuring the initial state can lead to exponential growth in the Lyapunov exponent in future values of the time series. In most cases, the frequency spectrum and autocovariance function of chaotic series resemble white noise. In fact, chaotic processes often share first and second-moment properties with white noise and colored noise. The frequency spectrum obtained from a random time series and a time series associated with real dynamic systems mixed with noise both exhibit a wide band. Therefore, distinguishing between these cases based solely on the frequency spectrum is not possible. A criterion for detecting and distinguishing chaos from noise is presented below [27].

Theorem 1: The variance of the autocorrelation coefficient of the energy signal of chaos is greater than the variance of the autocorrelation coefficient of the energy signal of noise.

Proof: Consider the following continuous-time system:

$$\dot{x} = f(x, \varphi) \tag{14}$$

In this case, according to Lyapunov exponent, it can be written

$$\|\delta x(t)\| = \|\delta x(0)\| e^{\lambda t} \overset{\|ax\|=\|a\|\|x\|}{\rightarrow} \|\delta x(t)\| = \left\|\delta x(0) e^{\lambda t}\right\| \rightarrow \delta x(t) = \delta x(0) e^{\lambda t} \tag{15}$$

where, λ is the largest Lyapunov exponent and $\delta x(t)$ is the variation of the signal x. In chaotic systems, as we know, $\lambda > 0$.

According to the definition of the norm of the signal and energy, we can write:

$$\|x\|^2 = E_x(t) \rightarrow E_x(t) = \left(\|\delta x(0)\|e^{\lambda t}\right)^2 \tag{16}$$

On the other hand, the spectrum of color noise power in general can be considered as follows:

$$S(\alpha) = \frac{A}{f^{\alpha}} \rightarrow X(f) = \frac{\sqrt{A}}{f^{\frac{\alpha}{2}}} \rightarrow \mathrm{E}_n(t) = \int X(t)^2 dt \rightarrow \mathrm{E}_n(t) = 0.31Aln(t) \tag{17}$$

To show that $E_n(t) \leq \mathrm{E}_x(t)$, we must show that $E_n(t) - E_x(t) \leq 0$. For this purpose, the Taylor series of the multivariable function $E_n(t) - E_x(t) \leq 0$ around the variables $A, t, \|\delta x(0)\|, \lambda$ is written as follows:

$$E_n(t) - E_x(t) = 0.31A\ln(t) - \|\delta x(0)\|^2 - 2\|\delta x(0)\|^2\lambda t - 4\|\delta x(0)\|^2\lambda^2 t^2 \tag{18}$$

As can be seen, for t<1 this relation is negative and for t > 1 for $A < \frac{\|\delta x(0)\|^2}{0.31\ln(t)}$ we can conclude: $E_n(t) \leq \mathrm{E}_x(t)$

By multiplying both sides of the above relation by, e^{-st}, we have:

$$\begin{aligned} &E_n(t) \leq \mathrm{E}_x(t) \rightarrow E_n(t)e^{-st} \leq \mathrm{E}_x(t)e^{-st} \rightarrow \\ &\int_0^{\infty} E_n(t)e^{-st} \leq \int_0^{\infty} \mathrm{E}_x(t)e^{-st} \rightarrow E_n(s) \leq \mathrm{E}_x(s) \end{aligned} \tag{19}$$

According to the definition of autocorrelation coefficient $R_{\mathrm{E}_x(t)} = L^{-1}\left(\mathrm{E}_x(s)^2\right)$, the above relation can be written as subscript:

$$R_{\mathrm{E}_n(t)} \leq R_{\mathrm{E}_x(t)} \tag{20}$$

Using the properties of the probability distribution function:

$$E_n(t) \leq \mathrm{E}_x(t) \rightarrow f_n \leq f_x \tag{21}$$

Where f_n is the noise energy distribution function and f_x is the chaotic signal energy distribution function. From the two relations (20) and (21):

$$\int R_{\mathrm{E}_n} f_n \leq \int R_{\mathrm{E}_x} f_x \rightarrow \mu_{R_{\mathrm{E}_n}} \leq \mu_{R_{\mathrm{E}_x}} \tag{22}$$

Where $\mu_{R_{\mathrm{E}_x}}$ is the mean of the autocorrelation coefficient of chaos and $\mu_{R_{\mathrm{E}_n}}$ is the mean of the autocorrelation coefficient of noise signal. From the two relations (20) and (22) we can conclude:

$$R_{\mathrm{E}_n}(t) - \mu_{R_{\mathrm{E}_n}} \leq R_{\mathrm{E}_x}(t) - \mu_{R_{\mathrm{E}_x}} \rightarrow \left(R_{\mathrm{E}_n}(t) - \mu_{R_{\mathrm{E}_n}}\right)^2 f_n \leq \left(R_{\mathrm{E}_x}(t) - \mu_{R_{\mathrm{E}_x}}\right)^2 f_x \tag{23}$$

Therefore, according to the definition of variance, it can be written:

$$\int \left(R_{\mathrm{E}_n}(t) - \mu_{R_{\mathrm{E}_n}}\right)^2 f_n dt \leq \int \left(R_{\mathrm{E}_x}(t) - \mu_{R_{\mathrm{E}_x}}\right)^2 f_x dt \rightarrow \delta^2_{R_{\mathrm{E}_n}} \leq \delta^2_{R_{\mathrm{E}_x}} \tag{24}$$

This means that in a noisy environment, the variance of the autocorrelation coefficient of the energy signal of chaos is greater than the variance of the autocorrelation coefficient of the energy signal of noise.

Method 1: Compare the variance of the autocorrelation coefficient of the energy signal of chaos in a noisy environment to the variance of the autocorrelation coefficient of the energy signal of noise.

Proof: For the chaotic signal $x(t)$ in a noisy environment as $y(t) = x(t) + n(t)$, according to the variance properties, we can write:

$$\mathrm{var}(x + y) = \mathrm{var}(x) + \mathrm{var}(y) + 2\mathrm{cov}(x, y) \tag{25}$$

The energy of the chaotic signal in the noise medium is calculated as follows:

$$E(x(t) + n(t)) = \|x(t) + n(t)\| = E(x(t)) + E(n(t)) + 2\int x(t)n(t)dt \tag{26}$$

Where, $n(t)$ is a noise signal.

So:

$$\mathrm{var}(E(x(t) + n(t))) > \mathrm{var}(E(n(t))) \tag{27}$$

In this relation, $E(\ldots)$ represents energy.

As seen in Theorem 1 and Method 1, due to the dependence of the chaotic signal on the nonlinear dynamics of the system, it can be said that chaos has smoother changes than noise, but due to the random nature of noise, noise follows more severe changes. For this reason, it can be said that the variance of the energy of the chaotic signal is greater than the variance of the energy of the noise signal in different frequency subbands. As it is clear, by separating the high-frequency part in the chaotic signal mixed with noise, the effect of noise on the signal is reduced. Because noise often appears in high frequencies, by taking the signal information in the frequency domain and attenuating the high frequencies, which is equivalent to attenuating noise, chaos detection in a noisy environment can be better followed by examining the variance of energy in frequency sub-bands.

5. Chaos detection based on autocorrelation coefficient of energy distribution using static discrete wavelet transform

This method is based on a time-frequency analysis of the signal. By using static violet transformation, low-frequency components (signal general) and high-frequency components (signal details) are separated. The static wavelet transform is the same as the discrete wavelet transform. **Figure 2** shows the general structure of the signal decomposition algorithm into low and high-frequency components using the discrete wavelet transform.

Stationary wavelet transform is similar to discrete wavelet transform with the difference that sampling is not used in it (like **Figure 3**).

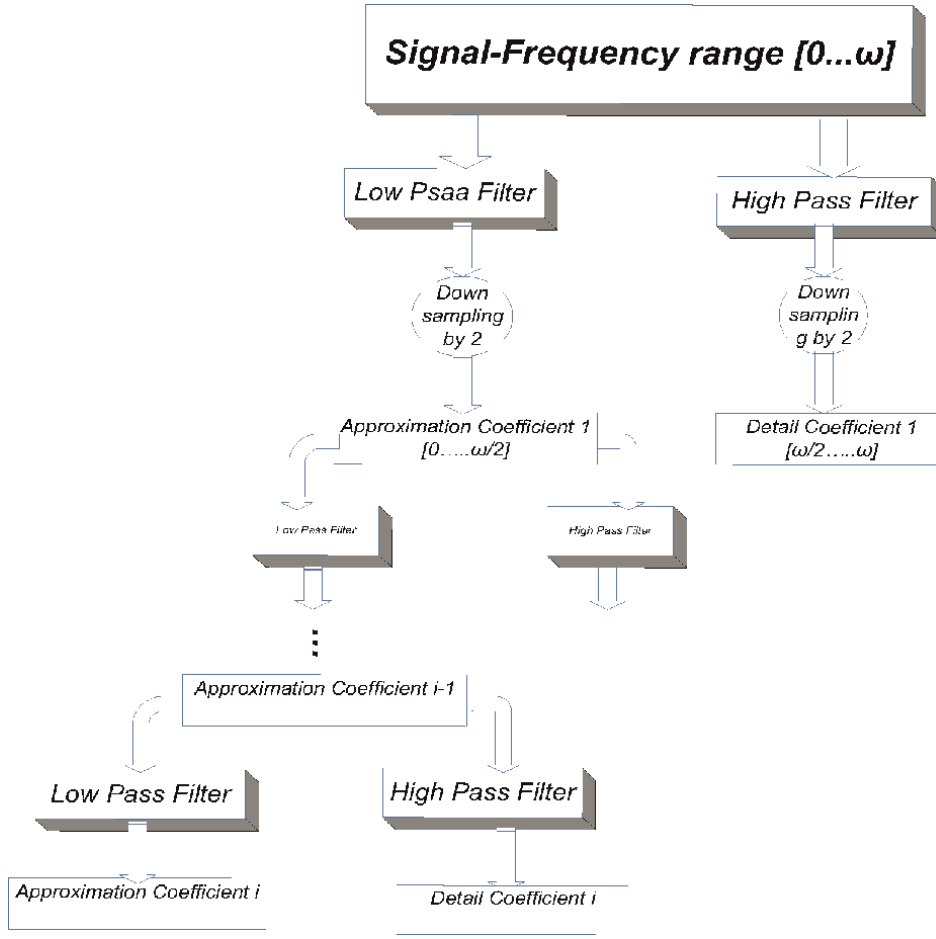


Figure 2.
General structure of signal decomposition using discrete wavelet transform.

The algorithm for detecting chaos is used to detect chaos.

1. The high-frequency components of the signal (signal details) are decomposed in several steps using the static wavelet transform (according to **Figure 2**).

2. Calculation of energy from detail coefficients (calculation of energy distribution in different frequency sub-bands) using Parswal's relation.

$$E(d_i) = \sum_{i=1}^{m} |d_i|^2 \tag{28}$$

3. Calculation of autocorrelation coefficient of energy distribution in different frequency sub-bands.

4. The variance of the autocorrelation coefficient of the energy distribution in different frequency sub-bands of the chaotic signal is more than the variance of the autocorrelation coefficient of the energy distribution in the frequency sub-bands of the noise signal.

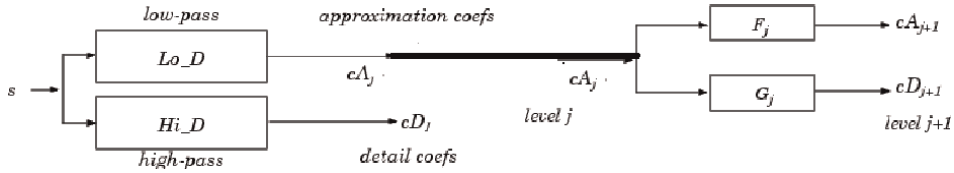


Figure 3.
General structure of signal decomposition using static wavelet transform.

For example, the fourth-order chaotic oscillator circuit based on memristor is shown in **Figure 4**. In **Figures 5** and **6**, the autocorrelation coefficient of energy distribution in different frequency sub-bands of chaos oscillator and the autocorrelation coefficient of energy distribution in frequency sub-bands of Gaussian white noise are considered.

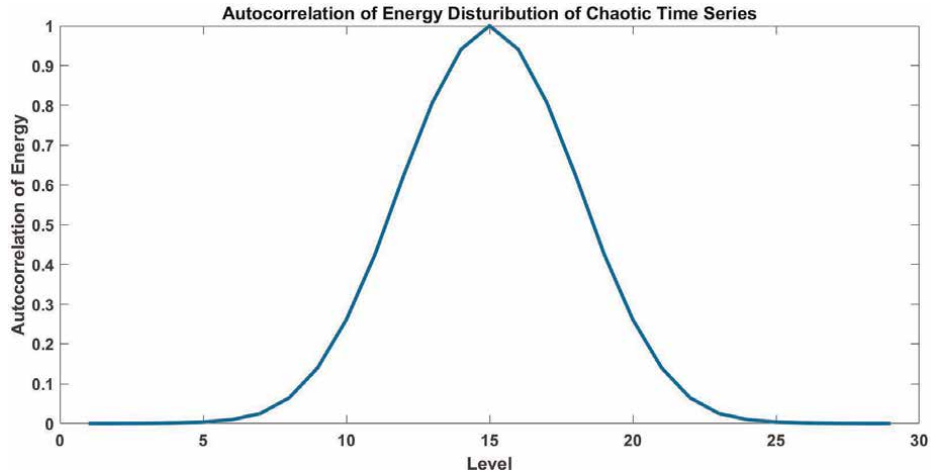


Figure 4.
The circuit schematic of chaotic oscillator based on memristor.

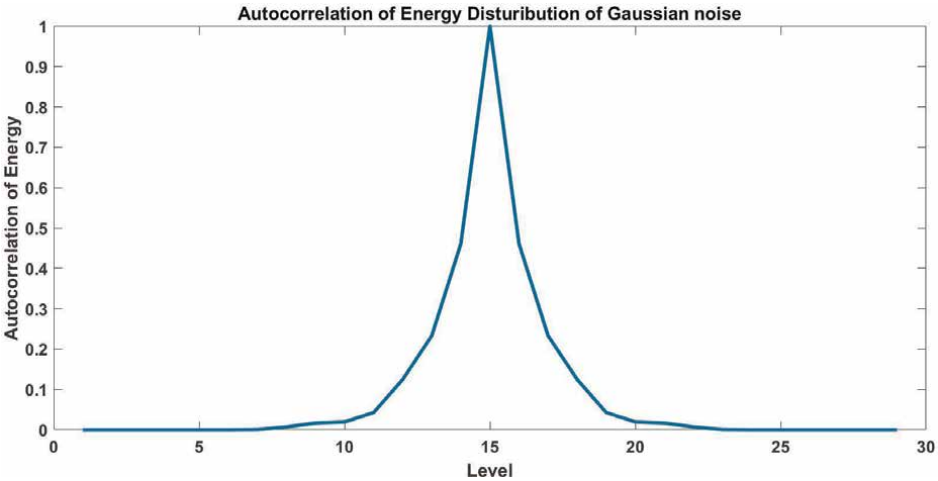


Figure 5.
Autocorrelation coefficient of energy distribution in frequency sub-bands of chaos signal.

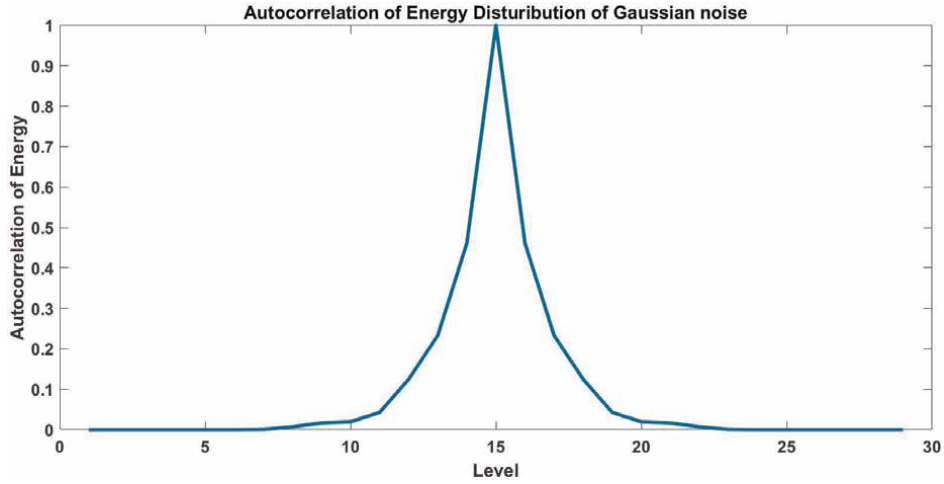


Figure 6.
Autocorrelation coefficient of energy distribution in frequency sub-bands of Gaussian white noise signal with mean 1 and variance 5.

6. Conclusions

In this chapter, we introduce a method for calculating the Lyapunov exponents using the differential transformation method to explore how various system parameters influence system behavior. We also propose a new criterion, based on the static violet transform, to differentiate between noise and chaos. Additionally, a method for detecting chaos based on energy distribution in various frequency sub-bands is described. Simulation results demonstrate the effectiveness of these methods in distinguishing chaos from noise.

Conflict of interest

There is no conflict of interest between the authors.

Declarations

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. This research did not receive any specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author details

F. Setoudeh* and M.M. Dezhdar
Electrical Engineering Department, Arak University of Technology, Iran

*Address all correspondence to: f.setoudeh@arakut.ac.ir

IntechOpen

© 2024 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

References

[1] Lathrop D. Nonlinear Dynamics and chaos: With Applications to Physics, Biology, Chemistry, and Engineering. American Institute of Physics; 2015

[2] Hilborn RC. Chaos and Nonlinear Dynamics: An Introduction for Scientists and Engineers. Oxford University Press; 2000

[3] Alligood KT, Sauer TD, Yorke JA, Chillingworth D. Chaos: An introduction to dynamical systems. SIAM Review. 1998;**40**(3):732-732

[4] Skokos CH, Gottwald GA, Laskar J. Chaos Detection and Predictability. Springer; 2016

[5] Maldonado J, Hernandez J. Chaos theory applied to communications–part I: Chaos generators. In: Electronics, Robotics and Automotive Mechanics Conference (CERMA 2007). IEEE; 2007. pp. 50-55

[6] Corron NJ, Hahs DW. A new approach to communications using chaotic signals. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications. 1997;**44**(5): 373-382

[7] Abarbanel HD, Brown R, Kennel M. Lyapunov exponents in chaotic systems: Their importance and their evaluation using observed data. International Journal of Modern Physics B. 1991;**5**(09): 1347-1375

[8] Hanac E. The phase plane analysis of nonlinear equation. Journal of Mathematical and Analytical. 2018;**9**: 89-97

[9] Baranovski AL, Schwarz W. Chaotic and random point processes: Analysis, design, and applications to switching systems. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications. 2003;**50**(8):1081-1088

[10] Thompson JMT, Stewart HB, Turner R. Nonlinear dynamics and chaos. Computers in Physics. 1990;**4**(5): 562-563

[11] Caponetto R, Fazzino S. A semi-analytical method for the computation of the Lyapunov exponents of fractional-order systems. Communications in Nonlinear Science and Numerical Simulation. 2013;**18**(1):22-27

[12] Deshmukh V, Sinha S. Control of dynamic systems with time-periodic coefficients via the Lyapunov-Floquet transformation and backstepping technique. Journal of Vibration and Control. 2004;**10**(10):1517-1533

[13] Mihajlovic N. Literature Study on Periodic Solutions in Nonlinear Dynamic Systems. 2002

[14] Saeed NA, Mohamed MS, Elagan SK. Periodic, quasi-periodic, and chaotic motions to diagnose a crack on a horizontally supported nonlinear rotor system. Symmetry. 2020;**12**(12):2059

[15] Pánis R, Kološ M, Stuchlík Z. Detection of chaotic behavior in time series. 2020. arXiv preprint arXiv: 2012.06671

[16] Sandri M. Numerical calculation of Lyapunov exponents. Mathematica Journal. 1996;**6**(3):78-84

[17] Ding R, Li J, Li B. Determining the spectrum of the nonlinear local Lyapunov exponents in a multidimensional chaotic system. Advances in Atmospheric Sciences. 2017; **34**:1027-1034

[18] Weron R, Przybyłowicz B. Hurst analysis of electricity price dynamics. Physica A: Statistical Mechanics and its Applications. 2000;**283**(3–4):462-468

[19] Klimek S, Leśniewski A. Quantized chaotic dynamics and non-commutative KS entropy. Annals of Physics. 1996; **248**(2):173-198

[20] Taheri AG, Setoudeh F, Najafi M, Feizi E. A new sufficient condition for stability analysis of nonlinear systems based on differential transform method (DTM). Journal of Control Engineering and Applied Informatics. 2020;**22**(4): 3-12

[21] Taheri AG, Setoudeh F, Tavakoli MB, Feizi E. Nonlinear analysis of memcapacitor-based hyperchaotic oscillator by using adaptive multi-step differential transform method. Chaos, Solitons & Fractals. 2022;**159**:112122

[22] Ghomi Taheri A, Setoudeh F, Tavakoli M. Nonlinear analysis of Colpitts oscillator using on differential transform method. Journal of Electrical and Computer Engineering Innovations (JECEI). 2020;**9**(2):127-142

[23] Chen D, Shi S, Gu X, Shim B. Weak signal frequency detection using chaos theory: A comprehensive analysis. IEEE Transactions on Vehicular Technology. 2021;**70**(9):8950-8963

[24] Rubežić V, Djurović I, Daković M. Time–frequency representations-based detector of chaos in oscillatory circuits. Signal Processing. 2006;**86**(9):2255-2270

[25] Djurović I, Rubežić V. Multiple STFT-based approach for chaos detection in oscillatory circuits. Signal Processing. 2007;**87**(7):1772-1780

[26] Zhu Q, Liang S. A method for detecting chaotic vibration based on continuous wavelet transform. International Journa Sensing, Computing and Control. 2011;**1**(2):125-132

[27] Setoudeh F, Sedigh AK, Najafi M. A novel method for chaos detection in heavy noisy environments based on distribution of energy. International Journal of Bifurcation and Chaos. 2019; **29**(13):1950179

Chapter 6

Exploring the Non-Linear Relationship between Economic Growth and Its Main Drivers over the Last Decade in EU: Evidence from a Panel Smooth Transition Regression

Catherine Bruneau, Alice Eraud and Iuliana Matei

Abstract

Rising oil, coal, and natural gas prices linked to the conflict between Russia and Ukraine have raised concerns about global economic growth and inflationary trends (International Monetary Fund (IMF), 2023). It is therefore interesting to examine the possible impact of oil prices on the relationship between economic growth and its determinants, including inflation. This article addresses this issue, using a panel dataset of 26 EU countries over the period 2011–2023 and studying the evolution of their growth within a Panel Smooth Transition Regression (PSTR) framework. Our empirical findings show that the real oil price is a significant transition variable between two extreme regimes and, accordingly, reveal that the determinants of economic growth have a time-varying intensity; notably, domestic investment, government spending, budget deficit, energy consumption of (non)renewable energy, trade balance, population growth, monetary policy as captured by the term spread and the M2 money growth, as well as the energy-related inflation.

Keywords: economic growth, oil price, inflation, energy use, monetary policy, PSTR models, European Union countries

1. Introduction

The escalating conflict between Russia and Ukraine has triggered a surge in oil, coal, and natural gas prices, prompting concerns about their potential impact on global economic growth and inflationary trends, as highlighted by the International Monetary Fund (IMF) in 2023. In response to these heightened uncertainties, there has been a notable uptick in academic interest surrounding the causes of inflation trends observed over the past decade and their potential repercussions on economic

growth. Advanced econometric models such as the Panel Smooth Transition Regression (PSTR) are employed to provide a nuanced understanding of the multifaceted relationships at play and their potential non-linear features.

More specifically, this paper explores and provides evidence of the non-linearity of the relationship between economic growth and its main determinants by identifying two different regimes linked to changes in the real price of oil within the Eurozone between 2012 and 2022.

Following Hansen's [1] approach, incorporating the concept of smooth transitions between regimes, allowing for a more flexible representation of non-linear relationships, one uses the PSTR model to explore non-linearities in the relationship between economic growth and its determinants to identify threshold points where the relationship between these variables changes, providing a more nuanced understanding of how economic dynamics evolve under different conditions in the particular context one mentioned.

The paper builds on Ben Cheikh et al. [2] approach investigating the relationship between energy consumption, income, and environmental pollution, with a focus on the impact of CO_2 emissions, using a non-linear regime-switching model to identify endogenous turning points in the relationship between economic development and environmental quality. The analysis, applied to Middle East and North African (MENA) countries, uses a non-linear panel smooth transition regression (PSTR) model to capture heterogeneity in pollutant emissions. The findings emphasize the importance of considering non-linear relationships for a nuanced understanding of environmental sustainability, economic growth, and energy consumption.

Based on these findings, this paper examines the complex relationship between economic growth and its determinants, including the common factors that explain the business cycle as well as certain monetary factors, and, more specifically, energy-related inflation. Noteworthy contributors to this non-linearity include a country's level of investment and energy consumption from non-renewable sources.[1] Special focus will also be paid on the real oil price as one source of non-linearity.

Our main contribution is to show how a PSTR approach can provide interesting insights into the complex interplay of economic growth and its determinants by highlighting the role of real oil prices in this non-linearity.

The results we obtain, while they need to be interpreted with caution due to the short period studied, could offer policymakers a more nuanced understanding, enabling them to develop targeted strategies for navigating the complex economic landscape, particularly in the face of geopolitical conflicts and crises.

The paper is organized as follows. The first part is devoted to a short literature review, and the second one is devoted to the presentation of the methodology. The data are described in part 3. The results are commented on in part 4. Part 5 concludes.

2. Related literature review

The inquiry into the impact of inflation on economic growth is a topic of considerable interest and discussion in the academic literature, as evidenced by works such

[1] The contribution of the renewable energy consumption is found insignificant.

as Gillman and Kejak [3]. While ongoing debate exists, there is a general consensus that inflation has a globally adverse effect on medium and long-term growth [4–8]. However, it has been proposed that the connection between economic growth and inflation is not a straightforward linear relationship; rather, it is influenced by the level of inflation. Fischer [9] introduces the idea of a threshold above and below which the growth effects of inflation differ. Specifically, he suggests a positive relationship between inflation and growth for low inflation levels but a negative or insignificant one for high levels. Additionally, in cases of negative impact, the marginal growth costs appear to vary with inflation; the effect is stronger at lower inflation rates than at higher ones [10–12].

While the non-linear nature of the inflation–growth relationship is widely acknowledged, controversies persist regarding the inflation level acting as the threshold, the sensitivity of this non-linear relationship to factors such as data frequency, analytical framework, methodology, country classification (developed/developing), and the presence of high-inflation observations.

PSTR model has found application in a diverse range of economic modeling problems. These applications encompass investigations into the connection between pollution and economic growth [13, 14], the inflation-growth relationship [15–17], the impact of oil prices on the current account of oil-exporting nations [18, 19], borrowing costs of European countries during the recent financial crisis [20–22] or the behavior of exchange rates [23], among others [2, 24, 25]. These diverse studies highlight the PSTR model's capability to effectively capture heterogeneity in panel data.

3. Methodology: A PSTR approach

The multi-regime non-dynamic panel smooth transition regression (PSTR) model with individual (μ_i) and possible time (λ_t) effects are specified as:

$$Y_{it} = \mu_i + \lambda_t + \beta^{(0)\prime} X_{it} + \beta^{(1)\prime} X_{it} * g\left(q_{it}; \gamma, c\right) + u_{it} \tag{1}$$

for $i = 1, \ldots N$ (cross-section dimension) and $t = 1, \ldots T$ (time dimension).

Y_{it} denotes the dependent variable, μ_i denotes the individual effect (which does not depend on time), and λ_t denotes the time effect (which does not depend on the individuals).

X_{it} is the K-dimensional vector of time-varying explanatory variables.

q_{it} is the transition variable and u_{it} the error term.

$g\left(q_{it}; \gamma, c\right)$ is the (scalar) transition function normalized to vary between 0 and 1.

In the simplest case, just one transition is supposed to occur between two extreme regimes ($m = 1$ transition). In this case, for the first (extreme) regime, the specification of Y is a linear function of X with parameter $\beta^{(0)}$. It is observed when $g\left(q_{it}; \gamma, c\right) \to 0$, while, in the second extreme regime, observed when $g\left(q_{it}; \gamma, c\right) \to 1$, Y is another linear function of X with parameters $\beta^{(0)} + \beta^{(1)}$.

It is worth emphasizing that the regime, at date t, is an intermediate one characterized by a linear function of X, with parameters $\beta^{(0)} + \beta^{(1)} g\left(q_{it}; \gamma, c\right)$ between $\beta^{(0)}$and $\beta^{(0)} + \beta^{(1)}$. Accordingly, the model has time-varying coefficients.

The transition function is generally specified as a logistic function:

$$g(q_{it};\gamma,c) = \left(1 + \exp\left(-\gamma\prod_{j=1}^{m}(q_{it} - c_j)\right)\right)^{-1} \quad (2)$$

with m denoting the number of transitions.

γ is the slope parameter supposed to be strictly positive and $c = (c_1, c_2, \ldots, c_m)$ with $(c_1 < c_2 < \ldots < c_m)$ is the set of the threshold parameters.

With m = 1, the simplest specification of the panel logistic smooth transition regression (PLSTR) model is obtained with the transition function specified as follows: $g(q_{it};\gamma,c_1) = \frac{1}{1+\exp\left(-\gamma(q_{it}-c_1)\right)}$.[2]

In this case, the LSTR model implies that the two extreme regimes are associated with low and high values of $g(q_{it};\gamma,c)$ with a single monotonic transition of the coefficients form $\beta^{(0)}$ to $\beta^{(0)} + \beta^{(1)}$ as q_{it} increases, where the change is centered around c_1.

When $\gamma \to +\infty, g(q_{it};\gamma,c)$ becomes an indicator function $1_{q_{it}>c_1}$ (equal to1 if $q_{it} > c_1$, and 0 otherwise). In that case, the PSTR model is similar to the two-regime panel threshold model of Hansen [1].

A generalization of the PSTR model to allow for more than two different regimes is the additive PSTR model [26]:

$$Y_{it} = \mu_i + \lambda_t + \beta^{(0)\prime}X_{it} + * \sum_{j=1}^{r}\beta^{(j)\prime}X_{it}\, g\left(q_{it}^{(j)};\gamma_j,c_j\right) + u_{it} \quad (3)$$

where the transition functions $g\left(q_{it}^{(j)};\gamma_j,c_j\right)$ are defined as in (2) with $c_j = \left(c_{j,1}, c_{j,2}, \ldots, c_{j,m_j}\right)$.

If, $\forall j = 1, .., r, m_j = 1$ and $q_{it}^{(j)} = q_{it}$, the model in (3) becomes a PSTR model with $r+1$ regime. Accordingly, the additive PSTR model can be viewed as a generalization of the multiple regime panel threshold model as shown by Hansen [1].

When the largest model that can fit the data is a two-regime PSTR model (1) with $r = 1$ and $m = 1$, as in the present study, model (3) plays a role in the evaluation of the estimated model as explained below.

Estimating the $\beta^{(0)}, \beta^{(j)}, \gamma_j, c_j, j = 1, \ldots, r$ parameters in the additive PSTR model is a relatively straightforward application of the fixed effects estimator and non-linear least squares (NLS).

Finally, the model is evaluated by using two misspecification tests.

- A test of parameter constancy over time with an alternative specifying that the parameters change smoothly over time;

[2] Note: another specification is possible giving the panel exponential smooth transition regression (PESTR) model with the transition function specified as:

$g(q_{it};\gamma,c) = 1 - \exp\left(-\gamma\prod_{j=1}^{m}(q_{it} - c_j)^2\right)$ still with $\gamma > 0$,

- A test of no remaining non-linearity where the alternative is that the parameters change smoothly over time.

For both tests, an extension of the PSTR model is proposed in the form (3).[3,4]

In order to investigate the potential non-linear effect exerted by the situation on the oil market on the relationship between the gross domestic product (GDP) growth and its usual determinants for the 26 countries of the Eurozone, the panel smooth transition regression methodology appears to be particularly well adapted. More precisely, the PSTR analysis of the problem, referring to specification (1), will include the following variables. For each country i and year t, the dependent variable is the GDP growth rate, $GDPG_{it}$ and the explanatory variables are as follows:

- Growth rate of the nominal oil price $\Delta LnOIL_t$
- Year-to year inflation rate $\Delta LnCPI_{it}$
- Initial GDP, GDP_{i0}
- Domestic investment growth, $INVG_{it}$
- Population growth, $POPG_{it}$
- Government expenditures growth, $GEXPG_{it}$
- Non-renewable energy consumption growth, $NRECG_{it}$
- Terms of trade growth, $TOTG_{it}$
- Budget deficit, BD_{it}
- Term spread variation, TS_{it}
- M2 money growth, $M2G_{it}$

The transition variable is the real oil price, in logarithm, $LnROIL_t$, whose dynamics can be considered stationary (see **Figure 1a** of Appendix).

---

[3] The model under the alternative may be called a Time Varying Panel Smooth Transition Regression (TV-PSTR) model, and it is specified as:

$$Y_{it} = \mu_i + \lambda_t + \beta^{(0)'} X_{it} + * \sum_{j=1}^{r} \beta^{(j)\prime} X_{it}\, g\left(q_{it}^{(j)}; \gamma_j, c_j\right) + \left(\beta^{(0)\prime} X_{it} + * \sum_{j=1}^{r} \beta^{(j)\prime} X_{it}\, g\left(q_{it}^{(j)}; \gamma_j, c_j\right)\right) f\left(\tfrac{t}{T}; \gamma_{r+1}, c_{r+1}\right) + u_{it}$$

with a logistic specification for f similar the one of g and with time as transition variable. The null hypothesis is then H_0: $\{\ \gamma_{r+1} = 0\ \}$ since $f\left(\frac{t}{T}; \gamma_{r+1}, c_{r+1}\right) = 1/2$ when $\gamma_{r+1} = 0$.

[4] In the PSTR framework it is a natural idea to consider an additive PSTR model with $r+1$ transitions as an alternative, that is:

$$Y_{it} = \mu_i + \lambda_t + \beta^{(0)'} X_{it} + * \sum_{j=1}^{r+1} \beta^{(j)\prime} X_{it}\, g\left(q_{it}^{(j)}; \gamma_j, c_j\right) + u_{it}$$

The null hypothesis of no remaining heterogeneity can then be formulated as:

H_0: $\{\ \gamma_{r+1} = 0\ \}$

As usual, the oil price and the related inflation rate are considered as exogenous variables, as well as the initial GDP and the population growth given the short period (only 13 years). In addition, the budget deficit dynamics display sufficient inertia to be assumed as an exogenous variable. Likewise, the monetary policy as captured by M2 money growth is expected to have somewhat delayed effects on economic growth. However, potential endogeneity issues can be expected for the inflation rate, domestic investment growth, government expenditures growth, terms of trade growth as well as term spread which are, therefore, introduced with one lag.

Finally, the model which will be estimated becomes:

$$\begin{aligned} GDPG_{it} = \mu_i + (&\beta_1^{(0)}\Delta LnOIL_t + \beta_2^{(0)}\ \Delta LnCPI_{i,t-1} + \beta_3^{(0)}GDP_{i,0} + \beta_4^{(0)}INVG_{i,t-1} + \beta_5^{(0)}GEXPG_{i,t-1} + \\ &+\beta_6^{(0)}POPG_{i,t} + \beta_7^{(0)}NRECG_{i,t-1} + \beta_8^{(0)}TOTG_{i,t-1} + \beta_9^{(0)}BD_{i,t} + \beta_{10}^{(0)}TS_{i,t-1} + \beta_{11}^{(0)}M2G_{i,t}) \\ &+[(\beta_1^{(1)}\Delta LnOIL_t + \beta_2^{(1)}\ \Delta LnCPI_{i,t-1} + \beta_3^{(1)}GDP_{i,0} + \beta_4^{(1)}\ INVG_{i,t-1} + \beta_5^{(1)}\ GEXPG_{i,t-1} \\ &+\beta_6^{(1)}POPG_{i,t} + +\beta_7^{(1)}NRECG_{i,t-1} + \beta_8^{(1)}TOTG_{i,t-1} + \beta_9^{(1)}BD_{i,t} + \beta_{10}^{(1)}TS_{i,t-1} \\ &+\beta_{11}^{(1)}M2G_{i,t}))g(LnOIL_t;\gamma_1,c_1) + u_{i,t} \end{aligned} \tag{4}$$

4. Data and variables

Our panel data sample covers the period 2011–2023 and includes the 26 European Union (EU) countries: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, and Sweden. We collect the data at a yearly frequency from Eurostat, European Central Bank, and World Development Indicators Databases. We aim to tackle the non-linear contribution of the determinants of economic growth (particularly inflation) depending on the conditions in the oil market; accordingly, we choose the real oil price (in logarithm) as the transition (threshold) variable. We retain a set of explanatory variables inspired by the empirical literature, including Barro [5], Sala-i-Martin [27], Ozturk [28], López-Villavicencio and Mignon [17], and Eggoh and Khan [24].

Our monetary and financial variables vector comprises the term spread and the money growth. We compute yearly averages using monthly data on government bond yields coming from the European Central Bank (ECB). Term spreads are the difference between each country's 10-year bond rates and the 3-month Euribor rate. This spread is commonly used to study yield curves (e.g., [29]). A 10-year-3-month term spread approaching zero suggests a "flattening" yield curve. In addition, a negative value of the spread, which is observed for inverted yield curves, is generally considered a sign of recession. Annual M2 money growth (as a broader measure of the money supply in an economy) accounts for inflation rate changes and their implications on economic growth.

The data related to the macroeconomic variables comes from Eurostat and refers to inflation, terms of trade, budget balance (deficit or surplus), and primary energy consumption. We use the harmonized Consumer Price Index (CPI) to measure inflation. The export-import price index ratio is taken to determine the terms of trade. Current prices divided by chain-linked quantities using 2015 as the reference year provided. This variable quantifies the percentage change over 5 years (year Y to year Y–5) and reflects how much an economy can import per unit of export products and services, suggesting its trade competitiveness. The government budget balance,

whether deficit or surplus, is taken as a percentage of GDP and serves as an indicator of fiscal policy. In addition, the World Development Indicators database provides data on annual economic growth, energy consumption, domestic investment, government consumption expenditures, and population growth. We use growth rate of real GDP per capita (at constant 2015 US prices) as a dependent variable. Among the explanatory variables vector, we also consider the initial level of GDP per capita measured by the natural logarithm of the value of GDP per capita every 5 years. This variable captures Solow's [30] convergence process in which countries with a lower initial capital stock per capita (or production per capita) expand faster. According to the literature, the coefficient of this variable should be negative. In line with neoclassical growth theory, our PSTR models contain both population growth and domestic investments (via the annual growth of gross fixed capital formation). The first variable is expected to negatively impact GDP growth, while increased investment rates should have a favorable impact on the evolution of economic activity. The government spending growth rate (the general government's final consumption expenditure growth) is expected to be positively or negatively linked to economic growth. Alesina et al. [31], for example, show that fiscal corrections relying mostly on spending cuts that are concentrated on government wages and transfers tend to be expansionary, whereas those relying mainly on tax increases are contractionary.

In addition to these traditional variables influencing economic growth, we also consider the non-renewable energy consumption growth (in kg of oil equivalent per capita). Indeed, fossil fuels account for a large part of the energy mix, at least, 74.2%, observed in 2023. Recent growth models (e.g., Stern [32], Soytas and Sary [33]) emphasize the role of energy in economic growth, whereas neoclassical growth models (e.g., Solow [30]) focus solely on exogenous technological changes. These theoretical findings inspired empirical research (e.g., Kraft and Kraft [34], Ozturk [28], Apergis and Payne [35]) on causality between these variables to guide environmentally friendly energy strategies. Energy consumption is expected to stimulate economic growth, as proposed by the "growth hypothesis" in the related literature. Its validation means that energy is essential to economic growth; hence, strong energy policies are needed to boost growth or constrain energy consumption to decelerate growth.

As indicated before, we aim to determine the extent to which fluctuations in oil energy prices contribute to non-linearities in the relationship between economic growth and its determinants. To this end, we consider the natural logarithm of the real oil energy price from the World Bank Commodity Price Data. It refers to the average annual organization of the petroleum exporting countries (OPEC) crude real oil price: the crude oil, the average spot price of Brent, Dubai, and West Texas Intermediate, equally weighed/$ per bbl. This variable, $LnROILP_t$, whose dynamics can be considered as stationary over the period of study, is the threshold variable in our PSTR models.

As indicated before, for each country i and year t, the dependent variable is the GDP growth rate, $GDPG_{it}$ and the explanatory variables are:

- Growth rate of the nominal oil price $\Delta LnOIL_t$
- Year-to year inflation rate $\Delta LnCPI_{it}$
- Initial GDP, GDP_{i0}
- Domestic investment growth, $InvG_{it}$
- Population growth, $POPG_{it}$

- Government expenditures growth, $GEXPG_{it}$
- Non-renewable energy consumption growth, $NRECG_{it}$
- Terms of trade growth, $TOTG_{it}$
- Budget deficit, BD_{it}
- Term spread variation, TS_{it}
- M2 money growth, $M2W_{it}$

Table 1 of Appendix shows us the matrix correlation between the explanatory variables. Since there is no substantial correlation between these variables (except for the link between initial GDP and population growth), they can be included in the model simultaneously.

Furthermore, **Tables 2** and **3** of the Appendix provide definitions and main descriptive statistics of variables in our growth regression analysis. Regarding our variable of interest, **Table 3** indicates that the lower real oil price was 41.14 $/barrel while the highest level corresponds to 95.29 $/barrel. OPEC's nominal oil price averaged 68.44 $/barrel. In the EU economies, between 2011 and 2022, inflation averaged 2.13% per year, while real GDP growth averaged 2.9%, respectively. In addition, **Figures 2–4** display the trends of inflation, real GDP growth, and real oil prices (see Appendix) among EU countries. **Figure 2** shows the scatter plot for the whole sample on the link between economic growth and inflation. Globally, there is a positive relationship between inflation and growth. This relationship seems to break down, however, midway between 4 and 8% inflation; above that threshold, there is a negative relationship between these two variables. However, the inverted U-shaped relationship between inflation and economic growth (observed in EU economies) has been documented in the recent empirical literature (see, e.g., [17]). The next **Figure 5** illustrate these patterns by country.

To avoid spurious results, tests for cross-sectional independence in the errors and variable stationarity checks were performed. **Table 4** in the appendix significantly rejects the null hypothesis of no cross-sectional dependency at the 1% level of significance for all variables, indicating reliable interdependencies between the countries. Considering the cross-dependence results, the Pesaran [2007] CIPS panel unit root findings show that the most part of our variables are stationary in level (except for the logarithm of budget deficit and the terms of trade). Some variables, such as oil prices do not have enough observations (11 in total) to test for stationarity, but, as previously indicated, the dynamics of the oil price can be considered graphically as stationary. Considering these findings, we may confidently move forward with the PSTR estimations regarding the relationship between economic growth and its determinants.

5. Results and discussion

Before estimating Eq. (4), we checked for linearity (homogeneity) in the relationship between GDP growth and inflation, conditioned by the oil price transition variable. Thus, testing whether the model has nonlinearity features is a necessary step before performing PSTR model estimation.

The results are detailed in **Table 5** of the Appendix. For the test of linearity, we check whether the order m is one or not. We find that null hypothesis of linearity is

rejected at the 1% significance level meaning that there exists a non-linear relationship between inflation and growth when the real oil price is considered as a transition variable. According to the two statistics (LR and LMF statistics), 2 regimes are fund (i.e., there is evidence on the existence of one threshold in the model). Additionally, the logistic specification is preferred over the exponential one since the logistic model had lower LM and LMF p-values.

Table 6 provides real oil price thresholds for the EU-26 countries as a whole. The real oil price threshold for the EU countries is 4.03. Since our data on real oil price are in natural logarithm, to compute the corresponding threshold value in dollar, we applied an exponential function to the constant value (4.03). This transformation informs us that the threshold for real oil price is 56 $ for the EU-26 countries.

In both regimes, the effect of oil-related inflation has a negative and statistically significant influence on economic growth at a 5% level of significance. It dominates the effect of consumption-based inflation whose effect is found non-significant in both regimes.

As expected, initial GDP has a positive impact on GDP growth, whatever the regime, but population growth has no significant impact.

Delayed investment growth has a positive and significant impact in the second regime; not surprisingly, this determinant of GDP growth plays no significant role in the critical and highly uncertain periods that primarily determine the first regime. Similarly, the lagged budget deficit has a positive impact on GDP growth in the second regime but not in the first. The same applies to terms trade's delayed growth, whose positive impact is only significant in the second regime.

Interestingly, the growth rate of non-renewable energy appears as a positive determinant of GDP growth for both regimes, meaning that the "growth hypothesis" is validated; the non-renewable energy being a key ingredient for the economic growth in the EU countries.

The growth rate of money supply (M2) should have a positive effect on GDP growth over the decade 2011–2023, due to the quantitative easing decided by the ECB to deal with the consequences of the sovereign debt crisis, as well as the Covid crisis. However, overall, the relationship between M2 and the growth rate is negative during the studied period.

Of course, the central bank's balance sheet is clearly related to phases of sustained growth and structural change, but it above all reflects the only ability of central banks to react very quickly to critical shocks by implementing stabilizing measures. Although the two waves of support programs implemented by the ECB (the public sector purchase programs-PSPP) following the sovereign debt crisis and pandemic emergency purchase programs (PEPP), following the pandemic, resulted in a sharp acceleration in the growth of the money supply, they were implemented in response to major recessions, which have a negative impact on this relationship over a relatively short and specific period.

The results must therefore be qualified with regard to the link between growth and money supply, even if this control variable remains significant at the scale of the regime. Delayed effects on the role of the money supply in stimulating growth may need to be investigated further over a longer period.

Finally, the lagged term spread has a negative and positive impact respectively in the first and second regimes. However, once again, the role of the term spread as an early indicator of recessions is difficult to highlight in the period studied, where we observe only one recession during the period covered by the study.

6. Conclusion

The real oil price is found a significant transition variable between two regimes of relations between growth and macroeconomic variables, distinctively affecting their intensity. While the oil price maintains a negative relationship with the annual GDP growth rate, erasing the expected impact of inflation; this relationship is non-linear and depends on a minimum oil price threshold, affected mainly by supply/demand imbalances in the oil market through an exogenous geopolitical context.

However, as we have pointed out, the period studied is short, quite heavily impacted by very critical periods, namely the sovereign debt crisis and the Covid event. What is more, the annual frequency is too low to really assess the impact of the financial variables traditionally introduced to explain economic growth. Consequently, these results cannot provide a solid basis for extrapolating what may happen in the future, by comparing the oil price to a reference value.

All in all, we consider that the results obtained are encouraging to develop a more in-depth analysis, notably by using a higher frequency database, over a longer period, distinguishing sub-panels of countries, in order to obtain a finer and more robust analysis of the non-linear relationship between economic growth and its determinants, as a function of oil market conditions.

Appendix

	GCEXG	INFLH	INVG	LBD	LGDP0	ROILP	M2G	NECG	POPG	TOTG	TS	DLNOILP
GCEXG	1.00											
INFLH	−0.02	1.00										
INVG	0.21	−0.03	1.00									
LBD	0.10	−0.18	0.11	1.00								
LGDP0	0.06	−0.10	0.05	0.27	1.00							
ROILP	−0.21	0.58	−0.21	−0.15	−0.03	1.00						
M2G	0.10	−0.07	0.04	−0.01	−0.22	−0.47	1.00					
NECG	−0.10	−0.30	−0.11	0.16	−0.18	−0.09	−0.03	1.00				
POPG	0.19	−0.04	0.11	0.29	0.64	−0.03	−0.15	−0.26	1.00			
TOTG	0.04	−0.27	0.03	0.27	−0.21	−0.46	0.37	0.14	−0.15	1.00		
TS	−0.40	0.13	−0.22	−0.28	−0.40	0.33	−0.03	−0.08	−0.24	0.03	1.00	
DLNOILP1	0.09	0.16	−0.13	0.11	0.04	0.37	−0.24	0.33	−0.07	−0.16	−0.13	1.00

Note: INFL—inflation; LROIL—real oil price (ln); DLNOIL—nominal oil price inflation, LGDP0—initial GDP (ln); INVG—domestic investment; POPG—population growth; GEXG—gov. spending growth rate; NECG—energy consumption growth; TOTG—terms of trade growth rate; LBD—budget deficit (ln); M2G—M2 growth rate; and TS—term spread.

Table 1.
Correlation matrix for explanatory variables: 2012–2022.

Variable	Explanation, computation	Data sources
GDPG	The annual growth rate of GDP per capita based on constant local currency	World Bank
InvG	The Gross fixed capital formation (annual % growth)	World Bank
TOTG	The terms of trade growth is computed based on the export-import price index. Current prices divided by chain-linked quantities using 2015 as the reference year provide these. This variable quantifies the percentage change over five years (year Y to year Y–5) and reflects how much an economy can import per unit of export products and services, suggesting its trade competitiveness.	Eurostat
INFHC	Annual increase in harmonized Consumer Price Index (CPI), 2015=100	World Bank
POPG	Population growth (derived from total population, annual)	World Bank
$LGDP_{io}$	The logarithm of the value of GDP per capita every five years	World Bank
GEXG	Gov. spending growth rate based on the general government final consumption expenditure (annual % growth)	World Bank
LBD	The logarithm of the budget balance (LBD): Net lending (+) /net borrowing (−) as percentage of GDP	Eurostat, World Bank
OILP	The average annual OPEC crude oil price. It refers to the crude oil, average spot price of Brent, Dubai and West Texas Intermediate, equally weighed/ $ per bbb.	World Bank Commodity Price Data
NECG	The non-renewable energy consumption growth. It is computed based on the energy use (tonnes of oil equivalent per capita, TOE) - It refers to use of primary energy before transformation to other end-use fuels, which is equal to indigenous production plus imports and stock changes, minus exports.	Eurostat, World Bank
M2G	The M2 money growth rate	ECB
TS	The term spread computed as the difference between each country's 10-year gov. bond rates and the 3-month Euribor rate. This spread is commonly used to study yield curves.	ECB

Note: GDPG—the growth rate of GDP per capita; GEXG—domestic investment growth; INFL—inflation; LOIL—nominal oil price(ln); $LGDP_{io}$—initial GDP (ln); InvG—domestic investment; POPG—population growth; NECG—energy consumption growth; TOTG—terms of trade growth rate; BD—budget deficit (ln); M2G—M2 growth rate.

Table 2.
Data description and sources (summary).

	GCEXG	INFH	INVG	LBD	LGDPI0	ROILP	M2G	NECG	POPG	TOTG	TS	DLNOILP1
Mean	1.63	1.47	3.03	−1.12	10.1	68.4	5.65	−0.94	0.22	0.84	2.33	−0.01
Median	1.35	1.30	2.64	−1.53	10.1	63.1	5.20	−0.54	0.23	0.90	1.56	−0.01
Maximum	14.9	5.80	101.	2.12	11.6	95.3	21.1	12.7	3.93	12.8	22.2	0.52
Minimum	−7.47	−1.60	−40.4	−3.38	8.81	42.1	−9.37	−19.3	−6.19	−11.3	−0.09	−0.64
Std. Dev.	2.94	1.40	11.4	1.31	0.65	19.1	3.61	5.11	0.97	3.47	2.36	0.32
Skewness	0.79	0.32	3.12	0.62	0.08	0.21	0.20	−0.53	−0.75	−0.06	3.29	−0.31
Kurtosis	6.13	2.78	25.8	2.30	2.35	1.52	6.12	3.83	11.4	3.48	22.3	2.53
Observations	286	286	286	286	286	286	286	286	286	286	286	286

Note: authors computation.

Table 3.
Main descriptive statistics.

Variable	Pesaran CD – Stats (probability)	Pesaran CIPS – t-Stat (model with constant)	Pesaran CIPS – t-Stat (model with constant and trend)
GDPG	44.58 (0.000)	−2.290**	−2.719
InvG	16.592 (0.000)	−3.514***	−2.874*
TOTG	22.328 (0.000)	−1.964	−1.742
INFHC	56.776 (0.000)	−2.300**	−2.655
POPG	3.106 (0.0019)	−2.416**	−2.611
Ln GDP_{io}	55.020 (0.000)	—	—
GEXG	15.603 (0.000)	−3.356***	−3.834***
Ln BD	37.091 (0.000)	−1.928	−3.594***
Ln OILP	59.791 (0.000)	—	—
NECG	32.890 (0.000)	−4.090***	−6.312***
M2 G	41.361 (0.000)	−3.621***	−5.320***
TS	45.397 (0.000)	−3.016***	—

*Note: (i) ***, **, * significant at 1% level, 5% level, and 10% level, respectively; (ii) ln BD is weakly stationary as well as TOTG (around 12%).*

Table 4.
Cross-section dependence test (Pesaran – CD) and related panel unit root test (Pesaran CIPS) results: 2012-2022.

Model	Hypotheses	Test	Stat	(p-value)
PSTR UE	H0 : m= 0 vs H1 : m=1	LM	101.604***	(0.000)
		LMF	12.473***	(0.000)
	H0 : m= 1 vs H1 : m=2	LM	18.788	(0.065)
		LMF	1.451	(0.152)

*Notes: (i) ***, **, * - significant at 1% level, 5% level and 10% level, respectively; (ii) LM and LMF tests are the Lagrange Multiplier and Fischer tests for linearity; (iii) H0: linear model; H1: PSTR model; iv) m=1 and m=2 are the logistic and the exponential transition functions, respectively; v) *** - indicates the strongest rejection of the linearity.*

Table 5.
LM and LMF tests of linearity (p-values).

Panel models	PSTR UE_26			
	Regime 1		Regime 2	
	Coef.	t-Stat	Coef.	t-Stat
Transition parameters				
Speed of transition – γ1	13.9944			
Threshold parameter – c1	4.0305		**56.28**	

Panel models	**PSTR UE_26**			
	Regime 1		**Regime 2**	
	Coef.	**t-Stat**	**Coef.**	**t-Stat**
Expanatory variables				
Dlog Nominal Oil price	−0.8034	−0.4899	−4.2965	−0.2413
Inflation(−1)	−0.1145	−0.356	−0.3214	−0.8241
GDP initial	5.7499*	1.7673	−0.0122	−0.1341
Domestic Investment growth (−1)	−0.0007	−0.0187	0.0744*(1)	−1.4607
Pop. Growth	−0.8447	−0.7964	−1.3829	−1.2651
Gov. spending growth(−1)	−0.1057	0.2799	−0.8283*	1.6617
Non-renew. Cons. Growth	0.1915	−0.1241	1.6120	−0.9788
Terms of trade growth(−1)	−0.1986	−1.4467	0.2378*	1.6416
Budget deficit $_{t-1}$	0.4248*	1.7860	−0.3009	−1.0284
Term spread (−1)	−6.8085***	−4.6942	7.7431***	4.6546
M2 growth	−0.3680	0.0357	−2.2983	0.2131
No. Obs	286x13		286x13	
No. Countries	26		26	
RSS	1017.382		1017.382	
AIC criterion	1.528		1.528	
BIC criterion	1.835		1.835	
Nb. parametters	24		24	
Opti no. transit. fc.	1		1	

*Note: i) ***, **, * - significant at 1% level, 5% level and 10% level, respectively; ii) the threshold 1.46 indicates a coefficient which is significant at 7% risk level for a unilateral test (a positive effect of investment growth on GDP growth is expected).*

Table 6.
PSTR estimates with the OPEC nominal oil price as the threshold variable: 2012–2022.

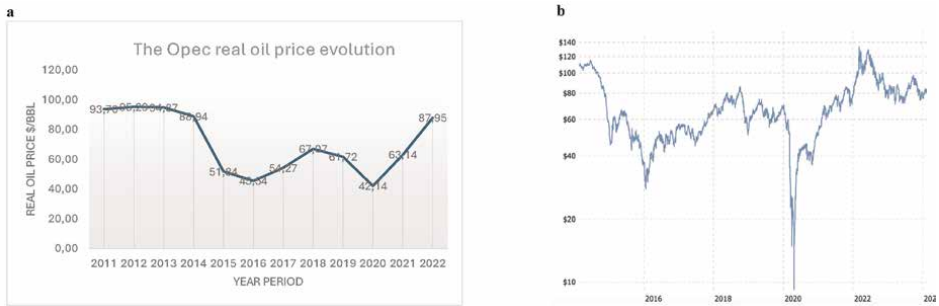


Figure 1.
(a) The evolution of the real OPEC oil price: 2012–2022; (b) BRENT crude oil price. Note: Authors computation.

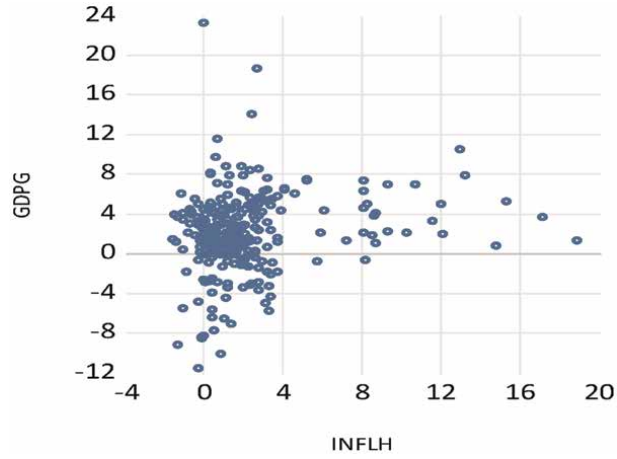


Figure 2.
The scatterplot on the link between inflation and GDP growth. Note: Authors computation.

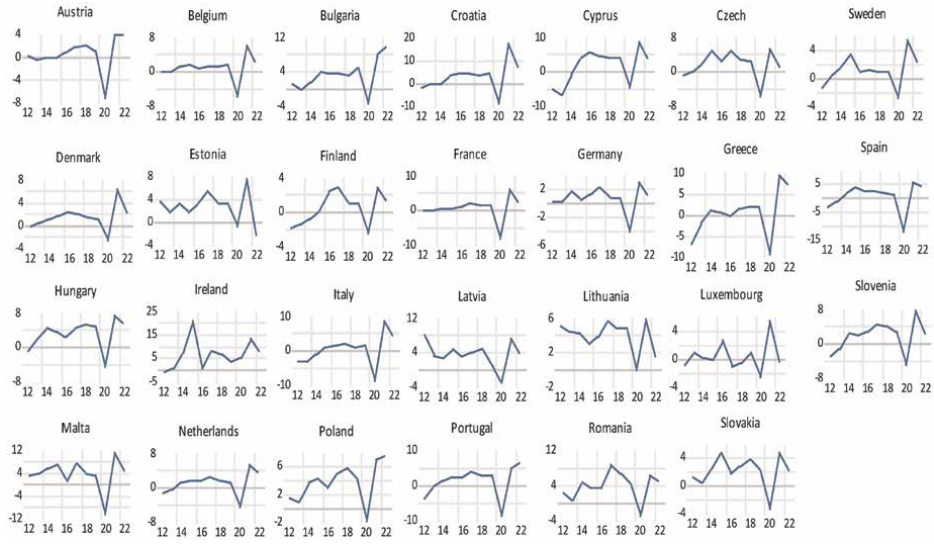


Figure 3.
GDP growth rates by EU member state: 2012-2022. Note: Authors computation.

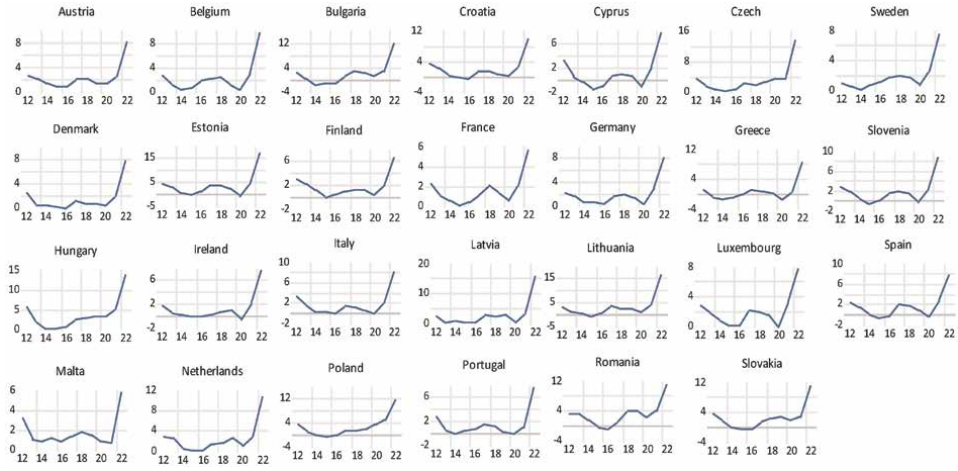


Figure 4.
The evolution of inflation in the EU countries: 2012–2022. Note: Authors computation.

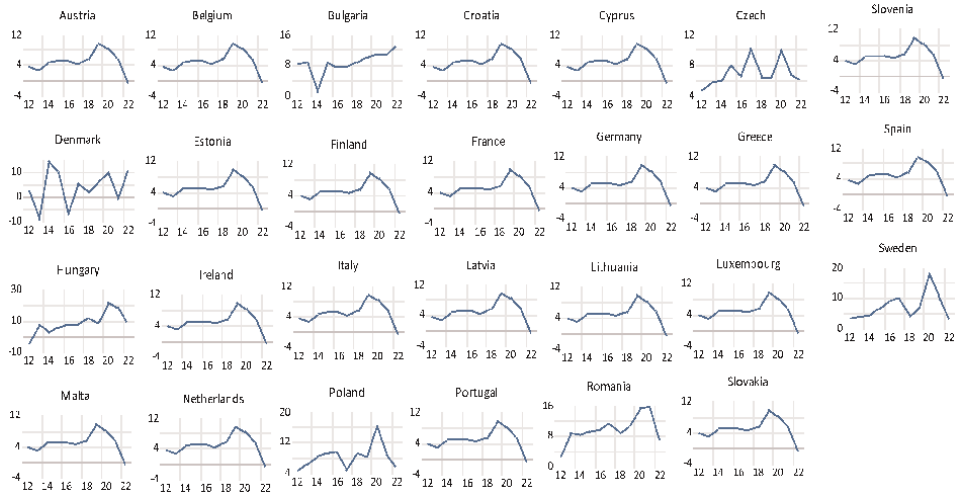


Figure 5.
The evolution of M2 growth in the EU countries: 2012–2022. Note: Authors computation.

Author details

Catherine Bruneau[1], Alice Eraud[2] and Iuliana Matei[1*]

1 CES-University of Paris 1 Panthéon-Sorbonne, Paris, France

2 University of Paris 1, Paris, France

*Address all correspondence to: iuliana.matei@univ-paris1.fr

IntechOpen

© 2024 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

References

[1] Hansen BE. Threshold effects in non-dynamic panels: Estimation, testing, and inference. Journal of Econometrics. 1999;**93**(2):345-368

[2] Ben Cheikh N, Ben Naceur S, Kanaan O, Rault C. Investing the asymetric impact of oil prices on GCC stock markets. Economic Modelling. 2021;**102**(2C)

[3] Gillman M, Kejak M. Inflation and balanced-path growth with alternative payment mechanisms. Economic Journal. 2005;**115**:247-270

[4] Barro RJ. Human capital and growth. American Economic Review. 2001;**91**(2): 12-17

[5] Barro RJ. Economic growth in a cross section of countries. The Quarterly Journal of Economics. 1991;**106**(2):407-443

[6] Chari VV, Christiano J, Lawrebce J, Kehoe PJ. Optimality of the Friedman rule in economies with distorting taxes. Journal of Monetary Economics. 1996;**37** (2):203-223. ISSN 0304-3932

[7] Gylfason T, Herbertsson TP. Does inflation matter for growth? Japan and the World Economy. 2001;**13**(4):405-428

[8] Kormendi R, Meguire P. Macroeconomic determinants of growth: Cross-country evidence. Journal of Monetary Economics. 1985;**16**:141-163

[9] Fischer S. The role of macroeconomic factors in growth. Journal of Monetary Economics. 1993;**32**:485-512

[10] Burdekin RCK, Denzau AT, Keil MW, Sitthiyot T, Willett TD. When does inflation hurt economic growth? Different nonlinearities for different economies. Journal of Macroeconomics. 2004, 2004;**26**(3):519-532

[11] Ghosh A, Phillips S. Warning: Inflation may be harmful to your growth. IMF Staff. 1998;**45**:672-710

[12] Harris LC. Market orientation and performance: Objective and subjective empirical evidence from U.K. companies. Journal of Management Studies. 2001;**38**:17-43

[13] Nektarios A, Xepapadeas A. Smooth transition pollution-income paths. Ecological Economics. 2006;**57**(2):182-189

[14] Aslanidis N, Xepapadeas A. Regime switching and the shape of the emission–income relationship. Economic Modelling. 2008;**25**:731-739

[15] Espinoza R, Leon H, Prasad A. When should we worry about inflation? The World Bank Economic Review. 2012;**26** (1):100-127

[16] Seleteng M, Bittencourt M, van Eyden R. Non-linearities in inflation–growth nexus in the SADC region: A panel smooth transition regression approach. Economic Modelling. 2013;**30** (C):149-156

[17] López-Villavicencio A, Mignon V. On the impact of inflation on output growth: Does the level of inflation matter? Journal of Macroeconomics. 2011;**33**(3):455-464

[18] Allegret JP, Couharde C, Dramane C, Mignon V. Current accounts and oil price fluctuations in oil-exporting countries: The role of financial development. Journal of International Money and Finance. 2014;**47**(C):185-201

[19] Nusair SA. Oil price and inflation dynamics in the gulf cooperation council countries. Energy. 2019;**181**(C):997-1011

[20] Bruneau C, Delatte AL, Fouquau J. Was the European sovereign crisis self-fulfilling? Empirical evidence about the drivers of market sentiments. Journal of Macroeconomics. 2014;**42**:38-51

[21] Delatte AL, Gex M, López-Villavicencio A. Has the CDS market influenced the borrowing cost of European countries during the sovereign crisis? Journal of International Money and Finance. 2012;**31**(3):481-497

[22] Delatte AD, Fouquau J, Portes R. Regime-dependent sovereign risk pricing during the euro crisis. Review of Finance, European Finance Association. 2017;**21**(1):363-385

[23] Bereau S, López Villavicencio A, Mignon V. Currency misalignments and growth: A new look using nonlinear panel data methods. Applied Economics. 2012;**44**(27):3503-3511

[24] Eggoh J, Khan M. On the nonlinear relationship between inflation and economic growth. Research in Economics. 2014;**68**. DOI: 10.1016/j.rie.2014.01.001

[25] Campello ADVC, Lins LN. Metodologia de análise e tratamento da evasao e retencao em cursos de graduacao de instituicoes federais de ensino superior. RJ: XXVIII Encontro Nacional de Engenharia De Producao; 2008. p. 13

[26] Gonzalez A, Teräsvirta T, van Dijk VD, Yang Y. Panel smooth transition regression models. In: No 604, SSE/EFI Working Paper Series in Economics and Finance. Stockholm School of Economics; 2017

[27] Sala-I-Martin X. I just ran two million regressions. The American Economic Review. 1997;**87**(2):178-183. Available from: http://www.jstor.org/stable/2950909 [Accessed: 20 February 2024]

[28] Ozturk I. A literature survey on energy-growth nexus. Energy Policy. 2010;**38**:340-349

[29] Estrella A, Hardouvelis GA. The term structure as a predictor of real economic activity. The Journal of Finance. 1991;**46**(2):555-576

[30] Solow RM. A contribution to the theory of economic growth. The Quarterly Journal of Economics. 1956;**70**(1):65-94

[31] Alesina A, Perotti R, Tavares J, Obstfeld M, Eichengreen B. The political economy of fiscal adjustments. Brookings Papers on Economic Activity. 1998;**1998**(1):197-266

[32] Stern P. Toward a coherent theory of environmentally significant behavior. Journal of Social Issues. 2000;**56**:407-424. DOI: 10.1111/0022-4537.00175

[33] Soytas U, SarI R. Energy consumption, economic growth, and carbon emissions: Challenges faced by an EU candidate member. Ecological Economics. 2009;**68**:1667-1675

[34] Kraft J, Kraft A. On the relationship between energy and GNP. Journal of Energy Development. 1978;**3**:401-403

[35] Apergis N, Payne JE. The renewable energy consumption-growth nexus in Central America. Applied Energy. 2011;**88**(1):343-347

Chapter 7

To Be or Not to Be Connected: Reconstructing Nonlinear Dynamical System Structure

L. Gerard Van Willigenburg

Abstract

On the one hand, controllability and observability relate to the ability to control and observe the state of a dynamical system. On the other, controllability and observability are known as structural properties relating to internal connections of dynamical systems. If the dynamical system is nonlinear, subtle differences between these two occur and defining and computing these properties becomes very much more complicated, because they rely on differential geometry instead of linear algebra. One contribution of this chapter is to define and compute controllability and observability of analytical dynamical systems in a particularly *simple*, *unifying* manner, based on *connectivities* and *sensitivities*. A second contribution is to present *a new canonical form of controllability and observability singularities*, showing that these are essentially initial states that permanently switch-off connections to the input and output of the system. The third and final contribution is to show that by considering these singularities as different systems, *nonlinear system structure becomes a global property*, instead of a local one. What does remain local are state-transformations transforming dynamical systems into canonical forms revealing system structure. By using these canonical forms *as the starting point*, our simple, unifying definitions of controllability and observability are obtained. Examples are presented to illustrate these results.

Keywords: canonical forms, controllability, observability, accessibility, reachability, Kalman decomposition, structural singularities, lie algebraic rank conditions (LARC), sensitivity rank conditions (SERC), sensitivity-based algorithms

1. Introduction

Initiated by Kalman, between 1955 and 1970 the use of state-space representations and time-domain analysis led to a series of discoveries of fundamental concepts and design methodologies for the control of both linear time-invariant and linear time-varying dynamical systems having multiple input- and output-variables. Until then, most analyses were limited to the frequency domain and linear time-invariant systems with only a single input-variable and output-variable. Notable discoveries were the controllability and observability properties of linear systems [1, 2], and that these are dual as well as structural properties [3]. They play an important role in the solution of

the linear quadratic state and output feedback design problems, as well as the realization problem of input–output maps [1, 4, 5]. Around the same time, Bellman [6] and Pontryagin [7] laid major foundations for optimal control theory applicable to multivariable nonlinear systems. Together with the development of computers, this facilitated the design and implementation of optimal feedback control systems for nonlinear dynamical systems on computers available at the time [8].

Around 1970, attempts started to generalize the theory and concepts developed for linear systems to nonlinear systems. The nonlinearity of systems significantly complicates concepts. System properties generally become local instead of global, and the corresponding mathematics requires differential geometry instead of linear algebra. Differential geometry very much complicates definitions, derivations, and computations involving Lie algebras. Still, nonlinear system theory managed to generalize most aspects of linear system theory [9–11]. Despite the many complications associated with nonlinear system theory, the Kalman decomposition and other canonical representations of nonlinear dynamical systems turn out to posses the same simple structure as those obtained for linear systems [2, 3, 9, 11]. This important observation will be exploited in this chapter.

More recently, controllability and observability of large complex networks have become an important research topic. Although large networks are very often modeled by linear dynamics, chemical networks are generally nonlinear, requiring analysis of what is sometimes called nonlinear controllability and nonlinear observability [12–16]. Sensitivity-based algorithms are a promising development to determine these properties, especially for large-scale nonlinear dynamical systems [17, 18]. They reveal the importance of connectivities and sensitivities in defining and computing controllability and observability as explained and illustrated in this chapter.

As opposed to ordinary dynamical system representations, canonical representations reveal connections of state-variables to the input and output in a straightforward manner that can therefore be visualized using directed graphs. An important contribution of linear and nonlinear system theory was to discover these canonical representations that can be obtained for any dynamical system by a suitable change of state-space coordinates. This change of coordinates is realized by a state-transformation that may hold only locally. This situation is sketched in **Figure 1**.

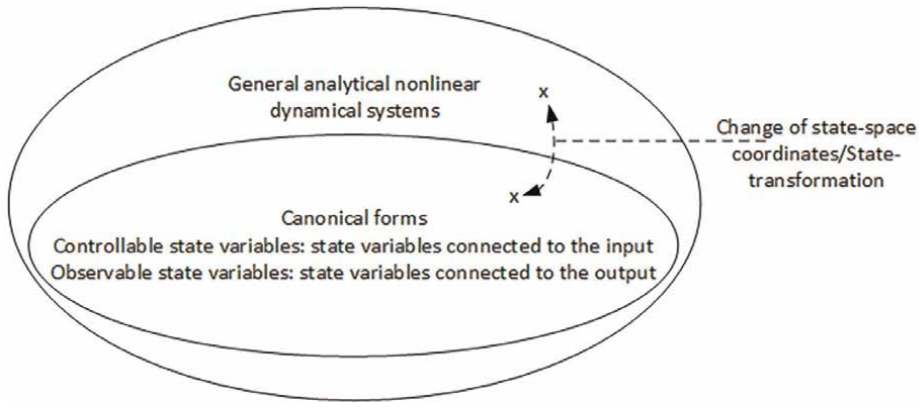


Figure 1.
Canonical forms facilitating simple definitions/explanations of controllability and observability as connectivities to the input and output representing structural properties of dynamical systems. Changes of coordinates/state-transformations connect general analytical nonlinear dynamical systems to their canonical forms.

Given the situation sketched in **Figure 1**, we asked ourselves the following question: "*When considering nonlinear dynamical system structure, would it not be better to start from canonical representations*"?

This chapter provides a positive answer by showing that canonical representations reveal structural properties easily and naturally. This allows us to define controllability and observability based on these connectivities. By first considering the structure of canonical forms, the mathematical complexity only comes in *at the very end*, when extending canonical representations to ordinary ones by means of state-transformations (see **Figure 1**). We will also show how these state-transformations and their associated Lie algebraic computations can be avoided completely by using *sensitivity-based algorithms* to establish controllability and observability. Avoiding these is especially important for large-scale systems. The algorithms compute a sensitivity rank condition (SERC) and uncontrollable/unobservable state-variables or modes, if any [17–19].

Remarkably, a canonical representation related to controllability/observability singularities, being points in the state-space where controllability/observability properties change, seems not to have been considered in the literature. A canonical form of controllability/observability singularities will be presented here and shown to be the key to considering nonlinear system structure as a global property, instead of a local one.

The terminology used in this chapter coincides with that commonly used in nonlinear system theory with one notable exception. What comes out as controllability in this chapter, is commonly known as local strong accessibility if the system is nonlinear and affine in the input [9, 10, 18–20]. We reflect on this notable exception and other results of this chapter in the conclusion section.

2. State-space representation of dynamical systems

To facilitate their analysis, numerical solution and control, dynamical systems described by ordinary differential equations are often represented in the so-called state-space form given by

$$\dot{x}(t) = f(x(t), u(t)), x \in \mathbb{R}^{n_x}, u \in \mathbb{R}^{n_u}, \tag{1}$$

$$y(t) = h(x(t)), y \in \mathbb{R}^{n_y}. \tag{2}$$

Within this formulation t denotes time, $x \in \mathbb{R}^{n_x}$ is the *state-vector* or *state* collecting all *state-variables* $x_i,\ i = 1, 2, .., n_x$, $u \in \mathbb{R}^{n_u}$ is the *input-vector* or *input* collecting all *input-variables* $u_i \in \mathbb{R},\ i = 1, 2, .., n_u$, and $y \in \mathbb{R}^{n_y}$ is the *output-vector* or *output* collecting all *output-variables* $y_i,\ i = 1, 2, .., n_y$. For convenience, we will generally drop the argument t. Eq. (1) describes how the state x propagates and depends on the input u and is called the *state-equation*. Eq. (2) describes how the state x maps on the output y and is called the *output-equation*. Several results from nonlinear system theory, used in this chapter, rely on differential geometry that applies to systems Eqs. (1) and (2) that are affine in the input, i.e.

$$f(x, u) = f_0(x) + \sum_{k=1}^{n_u} f_k(x) u_k, f_k(x) \in \mathbb{R}^{n_x}, k = 0, 1, .., n_u. \tag{3}$$

In Eq. (3), $f_k(x), k = 0, 1, .., n_u$, are vector functions with $f_0(x)$ called the drift term. Throughout this chapter f, h in (1)–(3) are assumed to be *analytic* vector functions.

3. Canonical state-space representations of dynamical systems

3.1 Controllability as obtained from its canonical form

Reconsider **Figure 1** and let

$$x' = \Psi(x), x', x, \Psi \in \mathbb{R}^{n_x} \tag{4}$$

represent the state-transformation that locally puts the system (1) into what will be called the *controllability canonical form* in this chapter, in line with the early development for linear systems as presented in [21], while appearing in [9–11] under different names associated with controllability.

$$\dot{x}' = f'(x', u),\ x' = \begin{bmatrix} x'^{u} \\ x'^{\overline{u}} \end{bmatrix} = \begin{bmatrix} \Psi^{u}(x) \\ \Psi^{\overline{u}}(x) \end{bmatrix},\ f'(x', u) = \begin{bmatrix} f'^{u}\left(x'^{u}, x'^{\overline{u}}, u\right) \\ f'^{\overline{u}}\left(x'^{\overline{u}}\right) \end{bmatrix}. \tag{5}$$

In Eq. (5), the transformed state x' separates into x'^{u} containing state-variables that are *connected* and $x'^{\overline{u}}$ containing state-variables that are *disconnected* from the input. A corresponding separation of the state-transformation $\Psi(x)$ into $\Psi^{u}(x)$ and $\Psi^{\overline{u}}(x)$ is specified in (5). Each $\Psi_i(x)$, $i = 1, 2, .., n_x$ is a scalar function of the state-variables x and equal to the transformed state-variable x'_i. Obviously, parts of the system denoted by the uppercase $\overline{u}$ that are disconnected from the input cannot be controlled.

Definition 1.

In the controllability canonical form (5), if $x'_i \in x'^{\overline{u}}$ then x'_i is called an uncontrollable state-variable of system (5) and $\Psi_i(x) \in \Psi^{\overline{u}}(x)$ is called an uncontrollable mode of system (1), (3). If $x'_i \in x'^{u}$, then x'_i is called a controllable state-variable of system (5) and $\Psi_i(x) \in \Psi^{u}(x)$ is called a controllable mode of system (1). $\Psi_i(x)$ not necessarily depends on all state-variables x_i, $i = 1, 2, .., n_x$. The state-variables of the set $\{x_i | \Psi_i(x)$ depends on $x_i\}$ are called state-variables making up the controllable/uncontrollable mode $\Psi_i(x)$.

Theorem 1.

1) Along trajectories of analytical systems (1), (3) the number of controllable modes $n_x^u = \dim(\Psi^{u}(x))$ and the number of uncontrollable modes $n_x^{\overline{u}} = \dim\left(\Psi^{\overline{u}}(x)\right)$ is constant but may depend on the initial state. For most initial states, n_x^u, $n_x^{\overline{u}}$ are identical. For exceptional initial states, called singularities in Section 4, $n_x^u = n_x - n_x^{\overline{u}}$ is reduced. 2) Along trajectories of system (1), (3) the set of state-variables $\left\{x_i | \Psi^{\overline{u}}(x)\right.$ depends on $\left.x_i\right\}$ making up all uncontrollable modes is invariant.

Proof.

1) and 2) follow from the Hermann-Nagano theorem in [22] according to which the state-space of an analytical dynamical system (1), (3) foliates into manifolds of dimension n_x^u, as specified in Theorem 1. Trajectories of system (1), (3) stay on a single manifold, the manifold being determined by the initial condition. These manifolds can be described locally using the coordinates $x' = \Psi(x)$ given by state-transformation (4) into the controllability canonical form.

Corollary 1.

Within the controllability canonical form (5), uncontrollable state-variables $x'^{\overline{u}}$ and controllable state-variables x'^{u} correspond one-to-one with uncontrollable

modes $\Psi^{\overline{u}}(x)$ and controllable modes $\Psi^{u}(x)$ of the system (1). The uncontrollable state-variables and modes are *disconnected* from the input, whereas the controllable state-variables and modes are *connected* to the input. Along trajectories of analytical systems (1), (3) the number of controllable and uncontrollable modes $n_x^u, n_x^{\overline{u}}$ are invariant but may depend on the initial state of the trajectory. For most initial states, $n_x^u, n_x^{\overline{u}}$ have the same value. For exceptional initial states, called controllability singularities in Section 4, the number of controllable modes $n_x^u = n_x - n_x^{\overline{u}}$ is reduced. Along trajectories of analytical systems (1), (3) the set of state-variables making up all uncontrollable modes is invariant.

Definition 1 together with Corollary 1 are graphically represented by **Figure 2**. From them the following alternative definition of controllability in terms of connectivities is obtained.

Definition 2.

Analytical dynamical systems (1), (3) are controllable along a trajectory if in the controllability canonical form (5) no state-variable x_i', $i = 1, 2, .., n_x$, or equivalently no mode $\Psi_i(x)$ of system (1), (3), is *disconnected* from the input.

Remark 1.

Computation of the state-transformation $\Psi(x)$ is generally performed using Lie algebraic computations [9–11]. These generally become problematic and time-consuming for large-scale systems. Sensitivity-based algorithms, especially developed for large-scale systems, provide a very attractive alternative [17–19]. In Section 5 we will elaborate on this.

Theorem 2.

Without having to compute the state-transformation (4), sensitivity-based algorithms very efficiently compute $n_x^u = \dim(x'^u)$ and $n_x^{\overline{u}} = \dim\left(x'^{\overline{u}}\right)$ as well as the set of state-variables making up all uncontrollable modes $x'^{\overline{u}}$ within the controllability canonical form (5), along trajectories of analytical systems (1), (3).

Proof.

Follows from [18] in which $n_x^u = \dim(x'^u)$, $n_x^{\overline{u}} = \dim\left(x'^{\overline{u}}\right)$ and the set of state-variables making up all uncontrollable modes are all obtained from a singular value decomposition (SVD) of a sensitivity matrix $S \in \mathbb{R}^{n_r \times n_x}, n_r \geq n_x$. Each zero singular value represents an uncontrollable mode, and each nonzero singular value a controllable mode. The nonzero components of the corresponding right singular vectors indicate the state-variables making up the corresponding mode.

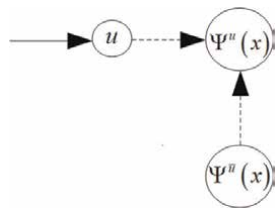


Figure 2.
Graphical representation and partitioning of a system with an input u along a trajectory. The state-space naturally partitions into a controllable part represented by the controllable modes $\Psi^{u}(x)$ and an uncontrollable part represented by the uncontrollable modes $\Psi^{\overline{u}}(x)$. $\Psi^{\overline{u}}(x)$ is disconnected from the input whereas $\Psi^{u}(x)$ is connected to the input. Connections internal to the system are represented by arrows with broken lines.

3.2 Observability as obtained from its canonical form

A development very similar to that of controllability in the previous section applies to observability. Because of this similarity, this section focuses on the differences. Reconsider **Figure 1** and let

$$x' = \Psi(x), x', x, \Psi \in \mathbb{R}^{n_x} \tag{6}$$

now represent the state-transformation that locally puts the system (1)–(3) into what will be called the *observability canonical form* in this chapter while appearing in [9–11] under different names associated with observability.

$$\begin{aligned} &\dot{x}' = f(x', u), y' = h(x'), \\ &x' = \begin{bmatrix} x'^{y} \\ x'^{\bar{y}} \end{bmatrix} = \begin{bmatrix} \Psi^{y}(x) \\ \Psi^{\bar{y}}(x) \end{bmatrix}, f(x', u) = \begin{bmatrix} f^{y}(x'^{y}, u) \\ f^{\bar{y}}\left(x'^{y}, x'^{\bar{y}}, u\right) \end{bmatrix}, h(x') = h\left(x'^{y}\right) \end{aligned} \tag{7}$$

In Eq. (7), the state x' separates into x'^{y} containing state-variables that are connected and $x'^{\bar{y}}$ containing state-variables that are disconnected from the output. A corresponding separation of the state-transformation $\Psi(x)$ into $\Psi^{y}(x)$ and $\Psi^{\bar{y}}(x)$ is specified in (7). Given the similarities with controllability in the previous section Definition 1, Theorem 1, Definition 2, Corollary 1 and Theorem 2 in the previous section apply if controllability is replaced by observability, Eq. (5) by (7) and input u by output y. **Figure 2** then turns into **Figure 3**.

Remark 2.

To *construct* analytical systems having certain controllability/observability properties, one can select the corresponding canonical form and *choose* the system parts arbitrarily. This *generically* realizes the corresponding controllability/observability properties, since there is the possibility, having zero probability, that an arbitrary choice causes additional uncontrollable/unobservable modes. Having realized the appropriate controllability/observability properties this way, we may subsequently "hide" them the by performing a state-transformation.

Remark 3.

Following Remark 2, all canonical forms have the property that the system parts do not cause additional uncontrollable/unobservable modes.

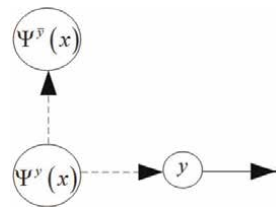


Figure 3.
Graphical representation and partitioning of a system with an output y along a trajectory. The state-space naturally partitions into an observable part represented by the observable modes $\Psi^{y}(x)$ and an unobservable part represented by the unobservable modes $\Psi^{\bar{y}}(x)$. $\Psi^{\bar{y}}(x)$ is disconnected from the output whereas $\Psi^{y}(x)$ is connected to the output. Connections internal to the system are represented by arrows with broken lines.

3.3 The Kalman canonical form of analytical nonlinear dynamical systems

Partitioning of systems along trajectories into parts that do and do not connect to the system input and output were obtained in sections 3.1, 3.2. These parts are represented by controllable/uncontrollable and observable/unobservable modes. These two separations lead naturally to a separation into four system parts, as represented for linear systems by the Kalman decomposition [3]. A similar decomposition for nonlinear system exists [9–11]. As before, the latter decomposition is obtained from a suitable state-transformation

$$x' = \Psi(x), x', x, \Psi \in \mathbb{R}^{n_x} \tag{8}$$

that now transforms the system into the form

$$\dot{x}' = f'(x', u), x' = \begin{bmatrix} x'^{u\bar{y}} \\ x'^{uy} \\ x'^{\bar{u}\bar{y}} \\ x'^{\bar{u}y} \end{bmatrix} = \begin{bmatrix} \Psi^{u\bar{y}}(x) \\ \Psi^{uy}(x) \\ \Psi^{\bar{u}\bar{y}}(x) \\ \Psi^{\bar{u}y}(x) \end{bmatrix}, f'(x,u) = \begin{bmatrix} f'^{u\bar{y}}\left(x'^{u\bar{y}}, x'^{uy}, x'^{\bar{u}\bar{y}}, x'^{\bar{u}y}, u\right) \\ f'^{uy}\left(x'^{uy}, x'^{\bar{u}y}, u\right) \\ f'^{\bar{u}\bar{y}}\left(x'^{\bar{u}y}, x^{\bar{u}\bar{y}}\right) \\ f'^{\bar{u}y}\left(x'^{\bar{u}y}\right) \end{bmatrix},$$

$$y' = h'(x) = h'\left(x'^{uy}, x'^{\bar{u}y}\right) \tag{9}$$

where $\Psi^{u\bar{y}}(x)$ are controllable and unobservable modes that are connected to the input and disconnected from the output and similarly for $\Psi^{uy}(x), \Psi^{\bar{u}\bar{y}}(x)$ and $\Psi^{\bar{u}y}(x)$. The decomposition (9) will be called the *controllability observability canonical form* or *Kalman canonical form* in this chapter. It is graphically represented by **Figure 4**.

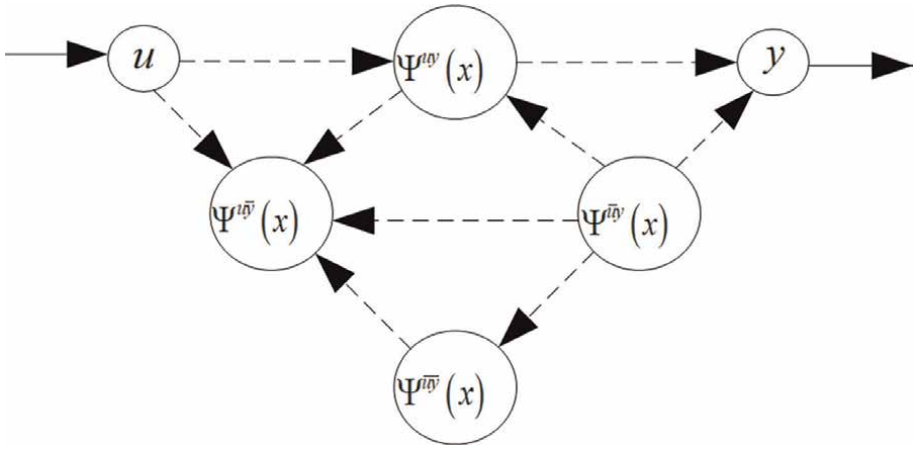


Figure 4.
Graphical representation and partitioning of a system with input $u \in \mathbb{R}^{n_u}$ and output $y \in \mathbb{R}^{n_y}$ along a trajectory. The state-space partitions into four parts represented by the modes $\Psi^{u\bar{y}}(x)$, $\Psi^{uy}(x)$, $\Psi^{\bar{u}\bar{y}}(x)$ and $\Psi^{\bar{u}y}(x)$. The partitioning is based on whether or not system parts connect to the input and output.

Remark 4.

Not all system parts in **Figures 2–4** have to be present. Also, not all internal connections have to be present as long as the connectivity of system parts with the system input and output remains unchanged. In **Figure 4** for example, the connection from $\Psi^{uy}(x)$ to $\Psi^{u\bar{y}}(x)$ may be absent as well as the one from $\Psi^{\bar{u}y}(x)$ to $\Psi^{\bar{u}\bar{y}}(x)$.

Remark 5.

From **Figure 4** observe that $\Psi^{uy}(x)$ in the only part that may be controlled by output feedback.

4. Controllability/observability singularities: Initial states affecting nonlinear system structure

Linear systems are described by (1)–(3) with

$$f(x,u) = Fx + Gu, h(x) = Hx. \tag{10}$$

F, G and H are real matrices that fully determine the interconnections and system structure. The entries of F, G and H do not depend on x nor on u. Therefore, the time-dependency of the state x and input u does not change the system structure [3].

For analytical nonlinear dynamical systems (1)–(3) however, this need no longer be the case. Initial states may switch-off, i.e. disconnect, connections to the input and output, thereby changing the system structure [16, 23]. To illustrate this, we start with an example presented in the next section.

4.1 Examples and definition of controllability/observability singularities

Example 1

$$\dot{x} = f(x,u), f(x,u) = \begin{bmatrix} -2x_1 - x_3 + u_1 \\ (1-x_2)(x_1+u_1) \\ -x_1 \end{bmatrix}, x, f \in \mathbb{R}^3, u \in \mathbb{R}^1, \tag{11}$$

$$y = h(x),\ h(x) = \begin{bmatrix} (1-x_2)x_3 \\ x_2 \end{bmatrix}, y, h \in \mathbb{R}^2. \tag{12}$$

If, in Eq. (11), we take $x_2(0) = 1$, then $\dot{x}_2 = 0$ and thus $x_2 = 1$ over the entire time-domain. In this way, *the constant state-variable x_2 disconnects itself from the other state-variables and input*. From the output-Eq. (12) and $x_2 \equiv 1$, we observe that both x_3 and x_1 are *disconnected from the output*. These disconnections *reduce* the number of controllable state-variables connected to the input as well as the number of observable state-variables connected to the output. Accordingly, *the system structure is changed*. Application of a state-transformation to system (11), (12) and initial states satisfying $x_2(0) = 1$, does not change the system structure, but does change state-variables into modes and constant state-variables into non-constant ones. To stress the role of initial conditions in determining system structure [23], we introduce the following definition.

Definition 3.

Initial states of the analytical dynamical system (1)–(3) that *reduce* the number of controllable/observable modes as compared to initial states in their neighborhood we call *controllability/observability singularities* of the analytic system (1)–(3).

Controllability singularities occur for instance in chemical systems when zero initial concentrations of some species prevent subsequent chemical reactions to occur [15]. They are different from what are mostly called *singular states* of dynamical systems which have a different degree of non-holonomy as compared to neighboring states [24]. As to our Example 1, according to Definition 3, initial states satisfying $x_2(0) = 1$ are both controllability as well as observability singularities of system (11), (12).

In Example 1, if $x_2 = 1$ would hold at isolated times only, this does not affect system structure since the disconnections from the input and output disappear immediately. But if $x_2 = 1$ holds along some part of a trajectory, the system structure changes along that part of the trajectory causing what is called *temporal system structure* [25–27]. Because analytical dynamical systems do not allow state-variables to be constant on a time-interval and time-varying outside this time-interval, the structure of analytic systems is fixed along trajectories [22]. But analytic systems do allow state-variables to be *very close* to being constant along part of a trajectory. In Example 1, when x_2 becomes very close to 1, one may say that the analytic system (11), (12) "almost changes structure" [27]. But for x_2 to really change the analytic system structure, it needs to be exactly 1 over the entire time domain. For arbitrary inputs $u(t)$ this can only happen if state-variable $x_2 = 1$ is disconnected from the input and other state-variables, so when $x_2(0) = 1$.

We deliberately *constructed* system (11), (12), starting from both the controllability canonical form (5) and the observability canonical form (7), while letting the *constant state-variable* $x_2 = 1$, *that is disconnected from the other state-variables and input*, switch-off state-variables from the input and output causing the controllability and observability singularities. The next theorem states that this type of switching is the *basic mechanism* causing controllability and observability singularities.

Theorem 3.

For analytical dynamical systems (1)–(3), *canonical representations* of controllability/observability singularities exist in which *constant state-variables that are disconnected from the input and the remaining state-variables* switch-off state-variables from the input/output causing the controllability/observability singularities.

The canonical representations of controllability/observability singularities will be given in the next section and the proof in Appendix 2. As to controllability, observe that Definition 3 and Theorem 3 comply with the Hermann-Nagano theorem [22], stating that for analytic systems (1)–(3) the number of uncontrollable modes $n_x^{\overline{u}} = \dim\left(x^{\overline{u}}\right)$ is fixed along trajectories, but may depend on the initial state.

4.2 Canonical state-space representations of controllability/observability singularities

To obtain the canonical representation of controllability singularities, we start from the controllability canonical representation (5) dropping accents of transformed states. We denote by $x^{s\overline{u}}$ the state-vector containing the constant state-variables that realize the switching-off. The switching-off occurs if $x^{s\overline{u}}(0) = \overline{x}^{s\overline{u}}$, in which $\overline{x}^{s\overline{u}}$ is *a steady state of* $x^{s\overline{u}}$ *that is unaffected by the input and state-variables not contained in* $x^{s\overline{u}}$. We denote by $x^{u\overline{u}}$ the vector of state-variables that become uncontrollable because they are switched-off from the input and by vector x^{uu} the controllable state-variables that are not switched-off from the input, and so:

$$x^{u} = \begin{bmatrix} x^{uu} \\ x^{u\overline{u}} \end{bmatrix}. \tag{13}$$

To the controllability singularities $x^{s\overline{u}}(0) = \overline{x}^{s\overline{u}}$ the following *canonical singular controllability form* corresponds:

$$\begin{bmatrix} \dot{x}^{uu} \\ \dot{x}^{u\overline{u}} \\ \dot{x}^{\overline{u}} \end{bmatrix} = \begin{bmatrix} f^{uu}\left(x^{uu}, x^{u\overline{u}}, x^{\overline{u}}, u\right) \\ f^{u\overline{u}}\left(x^{uu}, x^{u\overline{u}}, x^{\overline{u}}, u\right) \\ f^{\overline{u}}\left(x^{\overline{u}}\right) \end{bmatrix}. \tag{14}$$

$$x^{s\overline{u}}(0) = \overline{x}^{s\overline{u}} \Rightarrow x^{s\overline{u}}(t) = \overline{x}^{s\overline{u}}, f^{u\overline{u}}\left(x^{uu}, x^{u\overline{u}}, x^{\overline{u}}, u\right) = f^{u\overline{u}}\left(x^{u\overline{u}}, x^{\overline{u}}\right), \ -\infty < t < \infty. \tag{15}$$

Eq. (15) describes that if $x^{s\overline{u}}(0) = \overline{x}^{s\overline{u}}$, $x^{s\overline{u}}$ are constant state-variables, unaffected by the input and state-variables not captured by $x^{s\overline{u}}$, that realize the switching-off. Therefore,

$$x^{s\overline{u}} \supset \left\{x^{u\overline{u}}, x^{\overline{u}}\right\} \tag{16}$$

In a similar fashion, starting from the observability canonical representation (7), observability singularities $x^{s\overline{y}}(0) = \overline{x}^{s\overline{y}}$ switch-off state-variables from the output. We denote the vector of state-variables that become unobservable because they are switched-off from the output by $x^{y\overline{y}}$. Vector x^{yy} represents the observable state-variables that are not switched-off from the output, and therefore:

$$x^{y} = \begin{bmatrix} x^{yy} \\ x^{y\overline{y}} \end{bmatrix}. \tag{17}$$

To the observability singularities $x^{s\overline{y}}(0) = \overline{x}^{s\overline{y}}$ the following *canonical singular observability form* corresponds:

$$\dot{x} = \begin{bmatrix} \dot{x}^{yy} \\ \dot{x}^{y\overline{y}} \\ \dot{x}^{\overline{y}} \end{bmatrix} = \begin{bmatrix} f^{yy}\left(x^{yy}, x^{y\overline{y}}\right) \\ f^{y\overline{y}}\left(x^{yy}, x^{y\overline{y}}\right) \\ f^{\overline{y}}\left(x^{yy}, x^{y\overline{y}}, x^{\overline{y}}\right) \end{bmatrix}, y = h\left(x^{yy}, x^{y\overline{y}}\right). \tag{18}$$

$$x^{s\overline{y}}(0) = \overline{x}^{s\overline{y}} \Rightarrow x^{s\overline{y}}(t) = \overline{x}^{s\overline{y}}, \ f^{yy}\left(x^{yy}, x^{y\overline{y}}\right) = f^{yy}(x^{yy}), h\left(x^{yy}, x^{y\overline{y}}\right) = h(x^{yy}), \ -\infty < t < \infty. \tag{19}$$

Eq. (19) describes that if $x^{s\overline{y}}(0) = \overline{x}^{s\overline{y}}$, $x^{s\overline{y}}$ are constant state-variables, unaffected by the input and state-variables not captured by $x^{s\overline{y}}$, that realize the switching-off.

Theorem 4.

By considering controllability/observability singularities as *different systems*, the structural properties of analytical dynamical systems (1)–(3) become global.

Proof.

From Theorem 1, the number of controllable and observable modes is constant along any trajectory of an analytical dynamical system (1)–(3). Therefore, these only depend on the initial state of a trajectory. By Definition 3, controllability/observability singularities are the only ones changing system structure.

5. Determining system structure through sensitivity-based algorithms

5.1 Determining controllability/observability of systems and individual state-variables

Because uncontrollable state-variables and modes are disconnected from the input, their sensitivity to input-variables vanishes. Because unobservable state-variables and modes are disconnected from the output, the sensitivity of the output to them, vanishes. Sensitivity-based algorithms capture these insights by calculating a sensitivity matrix $S \in \mathbb{R}^{n_r \times n_x}, n_r \geq n_x$ [17–19] along a trajectory of system (1)–(3). As stated by Theorem 2, a singular value decomposition (SVD) of this matrix provides the number of uncontrollable/unobservable modes as the number of zero singular values. In other words, if the matrix is full-rank i.e. having no zero singular values, the system is controllable/observable along the trajectory and satisfies what is called a sensitivity rank condition (SERC) in [17–19]. Moreover, the state-variables making up each controllable/observable and uncontrollable/unobservable mode are indicated by the nonzero components of the corresponding right singular vectors. The state-variables making up the uncontrollable/unobservable modes are represented by what is called a controllability/observability signature in [17–19]. Thereby, without calculating state-transformations and canonical forms, the sensitivity-based algorithm provides almost all information system and control engineers are interested in. Specifically, they determine the controllability and observability of individual state-variables of the system (1)–(3) when applying the following definition.

Definition 4.

A state-variable $x_i, i = 1, 2, .., n_x$ is called controllable/observable if it does not make up any uncontrollable/unobservable mode. Otherwise, state-variable x_i is called uncontrollable/unobservable.

5.2 A challenging small-scale example containing two controllability singularities

Although being small-scale, the example presented next is a challenging one, because it contains two controllability singularities. Also, close to the singularity, transformed state-variables that have to be computed by the sensitivity-based algorithm, tend to grow very large. The example illustrates the Hermann-Nagano theorem in [22], which is used and described in the proof of Theorem 1, as well as the canonical form of controllability singularities.

Example 2: An uncontrollable system with two controllability singularities.

$$\dot{x} = f(x,u) = \begin{bmatrix} -x_2 \\ x_1 \\ 0 \end{bmatrix} u_1 + \begin{bmatrix} 2x_3x_1 \\ 2x_3x_2 \\ x_3^2 + 1 - x_1^2 - x_2^2 \end{bmatrix} u_2 \tag{20}$$

System (20) originates from [11], example 3.8. From the analysis of this example in [11], we conclude that system (20) has a single uncontrollable mode foliating the state-space into submanifolds with dimension 2. These submanifolds are tori given by the equation

$$\left(x_1^2 + x_2^2 + x_3^2 + 1\right) / \sqrt{x_1^2 + x_2^2} = c, \tag{21}$$

with $c > 2$ a constant that is determined by the initial conditions. This constant tends to infinity when approaching the singularity $x_1 = x_2 = 0$, where the torus degenerates into the x_3 axis. The other singularity occurs for $x_1^2 + x_2^2 = 1$, $x_3 = 0$, resulting in $c = 2$, where the torus degenerates into a circle (Appendix 1 provides further details). The dimensions of this foliation are thus two for the tori and one for the x_3 axis and circle $x_1^2 + x_2^2 = 1,\ x_3 = 0$.

Figure 5 concerns the controllability of system (20). It graphically represents the singular values $\sigma_i, i = 1, 2, 3$ (left panel) determining SERC and the components of the right singular vector ν_3(right panel) corresponding to the only (numerically) zero singular value σ_3 making up the controllability signature [18]. From the right panel of **Figure 5**, we observe that all the components of ν_3 are nonzero, implying that all three state-variables together make up the single uncontrollable mode. Therefore, according to Definition 4, no state-variable is controllable. When represented in the controllability canonical form (5), obtained after state-transformation (22), to be presented in the next section, the single uncontrollable mode is transformed into the single uncontrollable state-variable $x_3' = 1/c$. This is confirmed by the controllability signature in the right panel of **Figure 6**. Then, according to Definition 4, the other two state-variables $x_1' = x_3$, $x_2' = x_1$ are controllable.

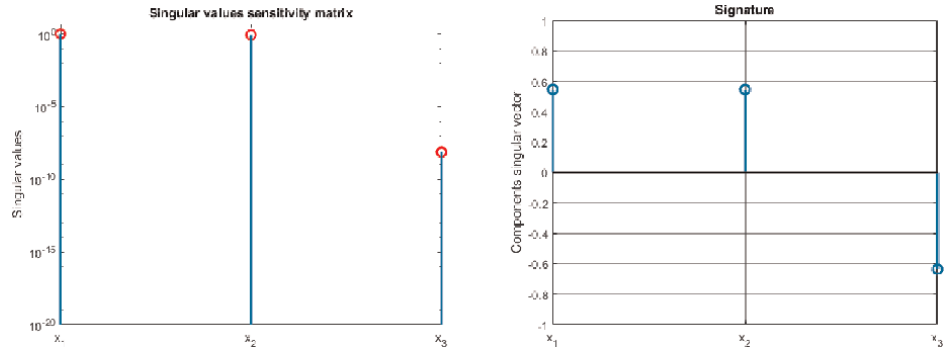


Figure 5.
Singular values (left panel) and controllability signature (right panel) of system (20) confirming the existence of one uncontrollable mode involving all three state-variables.

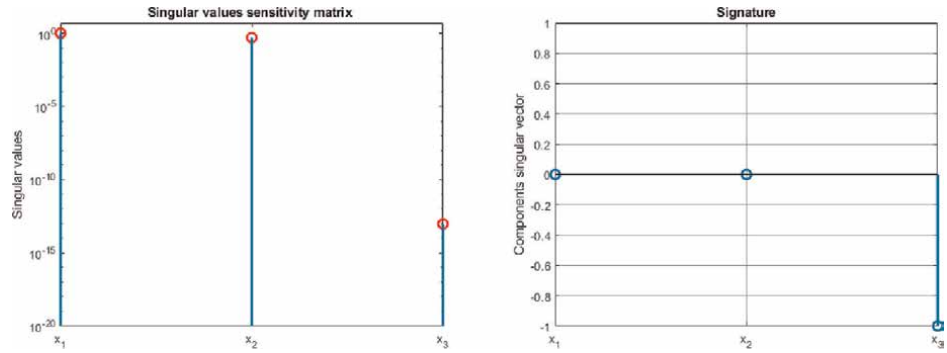


Figure 6.
Singular values and controllability signature after transformation (22) into the controllability canonical form (5) showing $x_3' = 1/c$ as the only uncontrollable mode and state-variable.

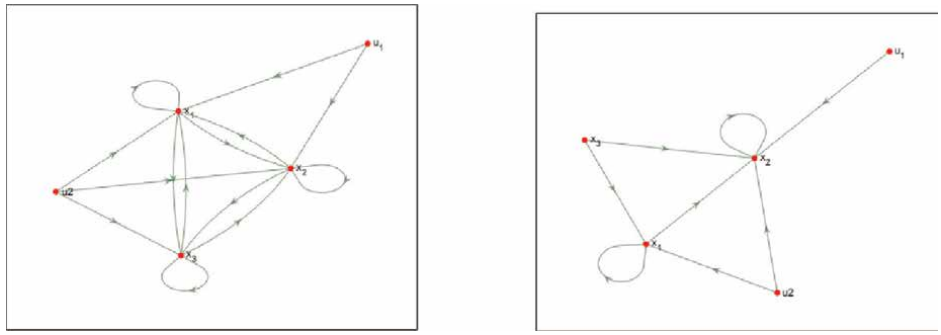


Figure 7.
Directed graph of system (20) (left panel) and its controllability canonical form (right panel). Only the latter reveals uncontrollability of state-variable $x_3' = 1/c$.

Figure 7 shows directed graphs of the original system (20) (left panel) and its controllability canonical form (right panel). Observe that only the latter directed graph reveals uncontrollability, illustrating that directed graphs only reveal uncontrollability/unobservability, when the system is represented in canonical form (minus permutations of state-variables).

In the next section we will show how each of the two controllability singularities can be made to match the canonical singular controllability form (14), (15). Note that this canonical form is obtained as a special case of the controllability canonical form (5). The latter canonical form will therefore also be obtained in the next section.

5.2.1 Canonical representations of the two controllability singularities

For system (20), the controllability singularity $x_1(0) = x_2(0) = 0$, implies $x_1(t) = x_2(t) = 0, t \geq 0$, which gives rise to two uncontrollable modes involving state-variables x_1 and x_2. This leaves state-variable x_3 as the single controllable state-variable, as confirmed by **Figure 8**.

Since x_3 is the single controllable state-variable, in the canonical singular controllability representation (14), (15) x^{uu} can only involve state-variable x_3 and we take $x^{uu} = x_3$. Since c in Eq. (21) is constant it may serve as $x^{\overline{u}}$. However, $c \to \infty$ as $x_1 \to 0$,

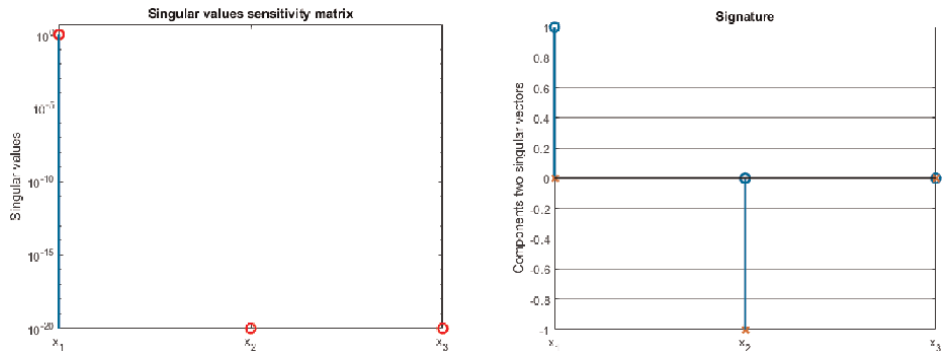


Figure 8.
Singular values (left panel) and controllability signature (right panel) of the controllability singularity $x_1(0) = x_2(0) = 0$ *of system (20). These confirm two uncontrollable modes involving state-variables* x_1 *and* x_2*, leaving one controllable state-variable* x_3 .

$x_2 \to 0$. This is overcome by taking $x^{\overline{u}} = 1/c$. Finally, we may choose either x_1 or x_2 as $x^{u\overline{u}}$. For $x^{u\overline{u}} = x_1$, the state-transformation into the canonical singular controllability form (14) becomes

$$x' = \begin{bmatrix} x'_1 \\ x'_2 \\ x'_3 \end{bmatrix} = \begin{bmatrix} x'^{uu} \\ x'^{u\overline{u}} \\ x'^{\overline{u}} \end{bmatrix} = \begin{bmatrix} x_3 \\ x_1 \\ 1/c \end{bmatrix} = \Psi(x), \tag{22}$$

with $x'^{s\overline{u}} = \left[x'_2, x'_3\right]^T = [x_1, 1/c]^T$ and $\overline{x'^{s\overline{u}}} = [0, 0]^T$. Appendix 1 reveals that the inverse $x = \Psi^{-1}(x')$ is only one-to-one locally. **Figure 9** confirms that $x'_1 = x_3$ is the single controllable mode and state-variable. The two uncontrollable modes involve the other two state-variables $x'_2 = x_1 = 0, x'_3 = 1/c = 0, -\infty < t < \infty$.

The second controllability singularity of system (20) concerns initial states satisfying $x_1^2(0) + x_2^2(0) = 1, x_3(0) = 0$. Then $x_1^2(t) + x_2^2(t) = 1, x_3(t) = 0, -\infty < t < \infty$ and the torus (21) degenerates into a circle which is obtained for $c = 2$. **Figure 10** confirms that we obtain the canonical singular controllability form (14), (15) by taking $x'^{uu} = x_1, x'^{u\overline{u}} = x_3, x'^{\overline{u}} = 1/c, x'^{su} = \left[x'_2,\ x'_3\right]^T$ and $\overline{x'^{su}} = [0,\ 1/2]^T$, i.e. by means of the state-transformation

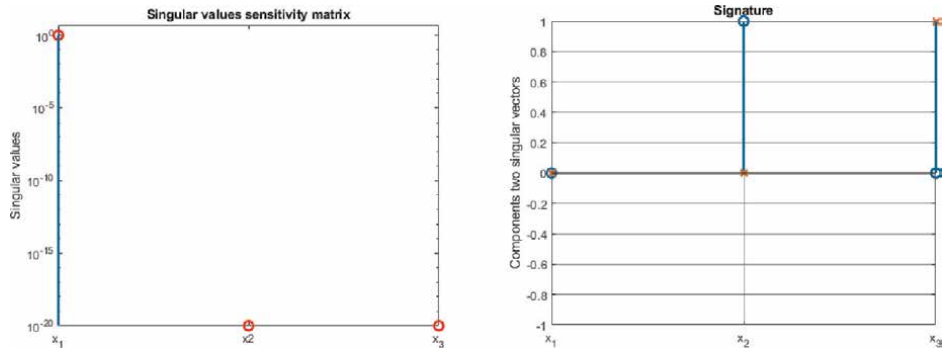


Figure 9.
Singular values and controllability signature after transformation (22) into the canonical singular controllability form (14) showing $x'_1 = x_3$ as the only controllable mode and state-variable.

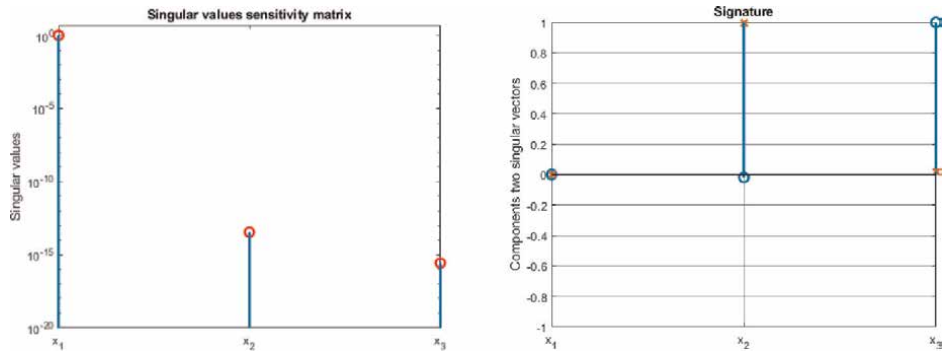


Figure 10.
Singular values and signature after transformation (23) into the canonical singular controllability form (14) showing $x'_1 = x_1$ as the only controllable mode and state-variable.

$$x' = \begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix} = \begin{bmatrix} x'^{uu} \\ x'^{u\overline{u}} \\ x'^{\overline{u}} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_3 \\ 1/c \end{bmatrix} = \Psi(x). \tag{23}$$

As a second example we reconsider system (11), (12) of Example 1. From our sensitivity-based algorithm we find system (11) to be controllable, since the singular values obtained are 3.2782e + 00, 7.5852e − 01 and 2.6537e − 02. This system has a controllability singularity $x_2(0) = 1$. **Figure 11** displays the singular values and controllability signature of this canonical controllability singularity, showing that only the 2nd component is nonzero confirming that x_2 is the only uncontrollable state-variable. From our sensitivity-based algorithm we also find system (11), (12) to be observable, since the singular values obtained are 2.0052e + 00, 8.5798e − 01 and 3.4738e − 01. As explained in Section 4.1 $x_2(0) = 1$ is also an observability singularity. **Figure 12** confirms this, showing that x_2 is the only state-variable that remains observable.

To summarize, for system (11), (12) of Example 1, $x^{s\overline{u}}(0) = x_2(0) = 1 = \overline{x}^{s\overline{u}}$ is a canonical controllability singularity satisfying (14) with $x^u = x_1$, $x^{u\overline{u}} = [x_2, x_3]^T$ and $x^{\overline{u}} = \varnothing$, as well as a observability canonical singularity $x^{s\overline{y}}(0) = x_2(0) = 1 = \overline{x}^{s\overline{y}}$ satisfying (17) with $x^y = x_2, x^{y\overline{y}} = [x_1, x_3]^T$ and $x^{\overline{y}} = \varnothing$.

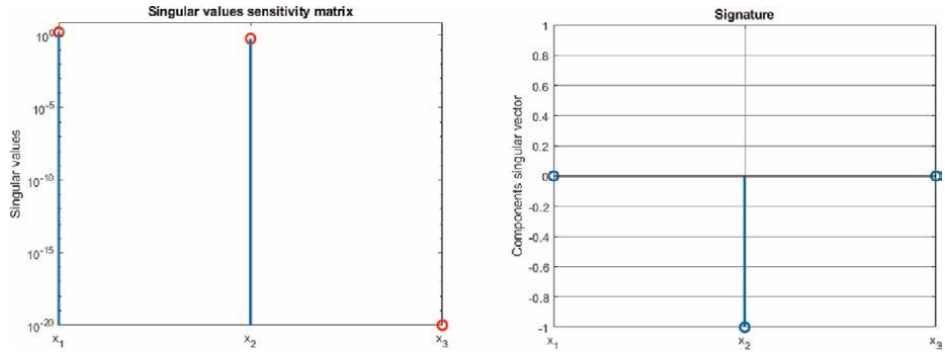


Figure 11.
Singular values (left panel) and controllability signature (right panel) of the controllability singularity $x_2(0) = 1$ of system (11).

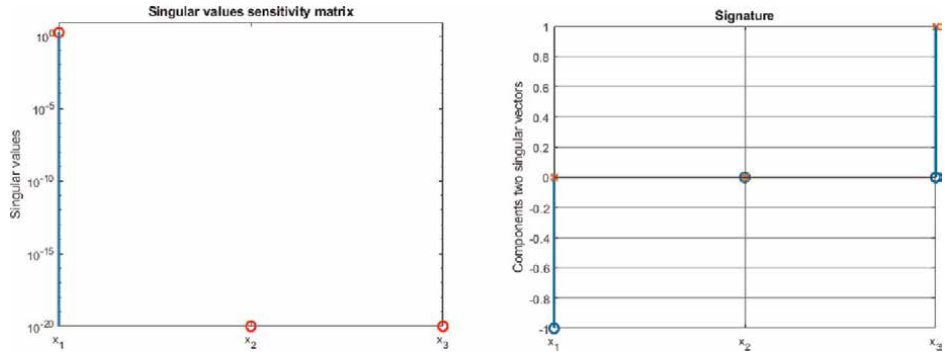


Figure 12.
Singular values (left panel) and observability signature (right panel) of the observability singularity $x_2(0) = 1$ of system (11), (12).

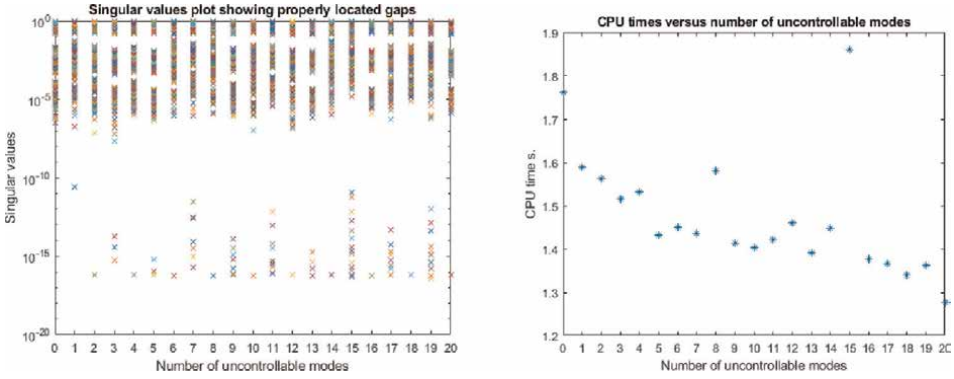


Figure 13.
The sensitivity-based algorithm correctly (left panel) and efficiently (right panel) establishes the number of uncontrollable modes of systems with 200 state-variables and 25 input-variables: The number of uncontrollable modes in each case equals the number of numerically zero singular values because some of these overlaps.

5.3 Large-scale examples

To illustrate and challenge the capability of sensitivity-based algorithms to solve high-dimensional problems efficiently, we generated large-scale nonlinear dynamical systems having 200 state-variables and 25 input-variables, following Remark 2 at the end of Section 3.3. Within the controllability canonical form of linear systems [3, 21], we selected the nonzero parts of the time-invariant system matrices random, while taking different values for the number of uncontrollable state-variables: $n_x^{\overline{u}} = 0, 1, 2, .., 20$. To change these linear time-invariant systems into nonlinear systems with $n_x^{\overline{u}}$ uncontrollable modes, we applied the following nonlinear state-transformation

$$x_i' = x_i, x_{i+1}' = e^{x_i + x_{i+1}}, i = 1, 3, 5, 7, .., 199. \tag{24}$$

We applied the sensitivity-based algorithm to the nonlinear systems with state x'. The left panel of **Figure 13** shows the singular values obtained from the sensitivity-based algorithm.

It shows that all gaps in the singular values are properly located (recognizing that several singular values overlap), because the number of singular values below this gap should be considered numerically zero, each one corresponding to an uncontrollable mode. The right panel shows the very short CPU times required to compute each result that is based on the concatenation of sensitivity matrices of three short trajectories. For details concerning the sensitivity-based algorithm we refer to [17, 18, 28]. We only mention here that, by exploiting duality, the sensitivity-based algorithm is also able to establish observability of nonlinear systems. The computations we performed on an ordinary PC using MATLAB.

6. Conclusions

We showed how *canonical representations* and *sensitivity-based algorithms* simplify and unify the definition, analysis and computation of controllability and observability of analytical, nonlinear and dynamical systems. For dynamical

systems represented in canonical form, controllability/observability simply translate into whether state-variables connect to the input/output. For systems not represented by canonical forms, we showed that controllability/observability translates into scalar functions of all state-variables, called *modes*, being connected to the input/output. Controllable/observable and uncontrollable/unobservable modes, as well as the state-variables involved in these modes, we computed very efficiently, using sensitivity-based algorithms. These algorithms nicely circumvent Lie algebraic computations, as well as state-transformations into canonical forms, which may both give rise to computational difficulties, especially for large-scale systems.

As for the restriction in this chapter to only study analytical dynamical systems, we remark that systems not belonging to this class are usually piecewise analytic. Then the analysis and results of this chapter apply to each separate interval over which the system is analytic. We also remark that by augmenting the system state with constant parameters, we can include the structural property identifiability as a special case of observability.

Originally, controllability is the ability to steer the system from any state to another, by means of the input. According to the analysis and definitions presented here, controllability relates to the connectivity of internal state-variables and modes to the input. For linear systems they are equivalent. If the system is nonlinear and affine in the input, our definition of controllability corresponds to what in the literature is usually called local strong accessibility, that is a slightly weaker property if the drift term is nonzero. As for the observability of dynamical systems, no such subtle difference occurs.

Starting from conventional canonical forms, we constructed *new canonical forms of structural singularities*, obtained from the insight that these are caused by initial conditions that permanently switch-off connections to the input/output. This insight also suggests to consider structural singularities as different systems. We showed how this turns system structure, determined by the *dimensions* of subsystems within corresponding canonical forms, into a global property. On the other hand, state-transformations into canonical forms may hold only locally.

If the state-space model has been developed from first principles (e.g. energy conservation, Newton's laws), state-variables have a clear meaning and interpretation. Since sensitivity-based algorithms provide the state-variables involved in the uncontrollable and unobservable modes, they then immediately provide the exact information a system modeler, designer or engineer is interested in.

Acknowledgements

The author likes to thank Hans Stigter, Jaap Molenaar, Dominique Joubert and Andrew Laidlaw for providing valuable ideas, discussions and suggestions that improved, and partly inspired, this chapter.

Declarations

The author has no conflicts of interest to disclose.
The author has no relevant financial or non-financial interests to disclose.
Data will be made available on reasonable request.

A. Appendix 1. Example 2 and the local character of its state-transformations.

The state-space of system (20) in Example 2, according to [11] example 3.8, is foliated as represented by **Figure 14**.

In Section 5.2.1, we reasoned and showed that state-transformations (22) and (23) transform the two controllability singularities of system (20) into the corresponding two canonical singular controllability forms. The inverse of state-transformation (23) corresponding to the singularity $c = 2$, where the torus degenerates into the circle $x_1^2 + x_2^2 = 1, x_3 = 0$, requires recovery of x_2 from $x_1' = x_1, x_2' = x_3, x_3' = 1/c$. We find two possible solutions: $x_2 = \pm\sqrt{1 - x_1'^2}$. This reveals that state-transformation (23) and its inverse are only one-to-one *locally*.

The inverse of state-transformation (22) corresponding to the singularity $c = \infty$, $x_1 = x_2 = 0$, recovers $x_2 = 0$ from $x_3' = 1/c = 0, x_2' = x_1 = 0$ as the limiting case $c = \infty$ of eq. (21). Finally, for initial states that are not controllability singularities, both transformation (22) and (23) provide the controllability canonical form (5) with $x^{\overline{u}} = x_3' = 1/c$, $x^u = \left[x_1', x_2'\right]^T$. To recover x_2 from x_1', x_2', x_3', 4 solutions apply, as shown by the 4 dots in **Figure 14**. Again this demonstrates that in general, the state-transformations into canonical forms are only one-to-one locally.

B. Appendix 2. Proof of theorem 3.

To proof Theorem 3 we will need the following lemma.

Lemma A2.1.

For analytic systems (1)–(3) a controllability canonical form (5) exists in which all uncontrollable state-variables are constant. This also applies to controllability singularities.

Proof.

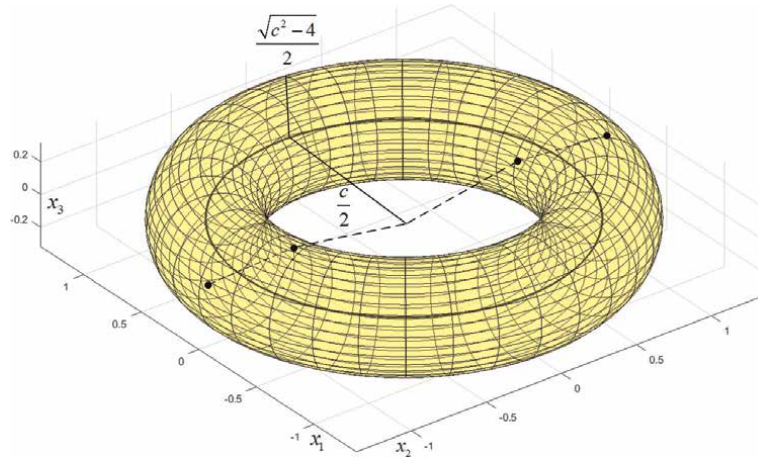


Figure 14.

The manifolds of example 2 are tori described by $\left(x_1^2 + x_2^2 + x_3^2 + 1\right)/\sqrt{x_1^2 + x_2^2} = c > 2$ *with c a constant. The two singularities are the circle* $c = 2, x_1^2 + x_2^2 = 1, x_3 = 0$ *and the* x_3 *axis* $c = \infty, x_1 = x_2 = 0$. *Knowing* x_1, x_3, $1/c$ *the 4 dots represent 4 solutions for* x_2.

For analytic systems (1)–(3), having $n_{\overline{u}} \geq 1$ uncontrollable modes, the state-space foliates into submanifolds characterized by

$$F : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_x^{\overline{u}}}, S(c) = \{x | F(x) = c\}, \tag{25}$$

with $c \in R^{n_x^{\overline{u}}}$ a different constant vector for each submanifold $S(c)$ that depends only on the initial state $x(0)$. $S(c)$ are also referred to as level sets [11]. For state x to be part of the corresponding submanifold $S(c)$, the transformation $F(x)$ may be considered as $n_x^{\overline{u}}$ constraints to be satisfied by x. Starting from (1)–(3), chose as state-transformation one with $x'^{\overline{u}} = \Psi^{\overline{u}}(x) = F(x)$ while taking $x'^{u} = \Psi^{u}(x)$ such that $\Psi(x)$ is a state-transformation. Then the dynamics of the new state x' is represented by the controllability canonical form (5) satisfying $x'^{\overline{u}}(t) = c$, i.e. with constant uncontrollable state-variables equal to $x'^{\overline{u}}(0)$.

As to controllability singularities, i.e. when $x^{s\overline{u}}(0) = \overline{x}^{s\overline{u}}$ is satisfied, the only thing that changes is that the dimension of $F(x)$ increases from $n_x^{\overline{u}}$ to $n_x^{\overline{u}} + n_x^{u\overline{u}}$, where $n_x^{u\overline{u}} \geq 1$ is the number of additional uncontrollable modes due to the controllability singularity.

Proof of Theorem 3.

The controllability canonical form of Lemma A2.1 applied to controllability singularities $x^{s\overline{u}}(0) = \overline{x}^{s\overline{u}}$ of system (1)–(3), complies with the canonical singular controllability form (14), (15) because the uncontrollable state-variables $x^{\overline{u}} \cup x^{u\overline{u}}$ will all be constant. Since the switching state-variables are among them, i.e. $x^{s\overline{u}} \supset \left(x^{\overline{u}} \cup x^{u\overline{u}}\right)$, the state transformation will therefore have $x^{s\overline{u}}$ as constant uncontrollable state-variables. Moreover, when $x^{s\overline{u}}(0) = \overline{x}^{su}$ is not satisfied, i.e. in a regular point close to the singularity, we reobtain the canonical form (5) since the components of $F(x)$ corresponding to $n_{u\overline{u}} = \dim\left(x^{u\overline{u}}\right)$ are no longer constant, so no longer switching off connections to the input.

As to the canonical singular observability form, the situation is slightly more complicated. A Kalman decomposition of the system (1)–(3), given by (9), may be applied at regular points close to the singularity. From this canonical form, consider the part containing the observable state-variables

$$\begin{bmatrix} \dot{x}^{uy} \\ \dot{x}^{\overline{u}y} \end{bmatrix} = \begin{bmatrix} f^{uy}\left(x^{uy}, x^{\overline{u}y}, u\right) \\ f^{\overline{u}y}\left(x^{\overline{u}y}\right) \end{bmatrix}, y = h\left(x^{uy}, x^{\overline{u}y}\right). \tag{26}$$

Because the reduced system (26) captures all observable modes, which are turned into observable state-variables, it will still contain the observability singularity. Also, it will still contain the switching state-variables $x^{s\overline{y}}$, because these influence the output since they realize the switching-off when $x^{s\overline{y}}(0) = \overline{x}^{s\overline{y}}$. Applying the canonical form of Lemma A2.1 to the reduced system (26), provides a canonical representation in which the switching state-variables $x^{s\overline{y}}$, that are uncontrollable, will be constant. This representation may be extended with the parts that have been dropped in (26) to obtain a canonical representation that complies with the canonical singular observability form (18), (19). Moreover, when $x^{s\overline{y}}(0) = \overline{x}^{s\overline{y}}$ is not satisfied, i.e. in any regular point close to the singularity, we reobtain the canonical form (7) because $x^{s\overline{y}}$ is no longer constant, so no longer switching off connections to the output.

Author details

L. Gerard Van Willigenburg
Mathematical and Statistical Methods Group, Wageningen University,
The Netherlands

*Address all correspondence to: gerard.vanwilligenburg@wur.nl

IntechOpen

© 2024 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

References

[1] Kalman RE. Contributions to the theory of optimal control. Boletin De La Sociedad Matematica Mexicana. 1960;**5**: 102-119

[2] Kalman RE. Mathematical description of linear dynamical systems. Journal of the Society for Industrial and Applied Mathematics - Series A. 1963;**1**(2): 152-192. DOI: 10.1137/0301010

[3] Kalman RE. Canonical structure of linear dynamical systems. Proceedings of the National Academy of Sciences of the United States of America. 1962;**48**(4): 596-600

[4] Joseph DP, Tou TJ. On linear control theory. Transactions of the American Institute of Electrical Engineers, Part II: Applications and Industry. 1961;**80**(4): 193-196. DOI: 10.1109/TAI.1961.6371743

[5] Ho BL, Kalman RE. Effective construction of linear state-variable models from input/output functions. at-Automatisierungstechnik. 1966; **14**(1–12):545-548. DOI: 10.1524/auto.1966.14.112.545

[6] Bellman R. Dynamic programming and Lagrange multipliers. Proceedings of the National Academy of Sciences of the United States of America. 1956;**42**(10): 767-769

[7] Kopp RE. Pontryagin maximum principle. In: Leitmann G, editor. Optimization Techniques, vol. 5, Mathematics in Science and Engineering. Amsterdam: Elsevier; 1962. pp. 255-279. DOI: 10.1016/S0076-5392(08)62095-0

[8] Athans M. The role and use of the stochastic linear-quadratic-Gaussian problem in control system design. IEEE Transactions on Automatic Control. 1971;**16**(6):529-552. DOI: 10.1109/TAC.1971.1099818

[9] Nijmeijer H, Van der Schaft AJ. Nonlinear Dynamical Control Systems. Vol. 175. New York: Springer; 1990

[10] Isidori A. Nonlinear Control Systems. London: Springer Science & Business Media; 2013

[11] Kwatny HG, Blankenship G. Nonlinear Control and Analytical Mechanics: A Computational Approach. Berlin: Springer Science & Business Media; 2000

[12] Angulo MT, Aparicio A, Moog CH. "Structural accessibility and structural observability of nonlinear networked systems." IEEE Transactions on Network Science and Engineering. Jul 2020;**7**(3): 1656-1666. DOI: 10.1109/TNSE.2019.2946535

[13] Kawano Y, Cao M. Structural accessibility and its applications to complex networks governed by nonlinear balance equations. IEEE Transactions on Automatic Control. 2019;**64**(11):4607-4614. DOI: 10.1109/TAC.2019.2901822

[14] Drexler DA, Virágh E, Tóth J. Controllability and reachability of reactions with temperature and inflow control. Fuel. 2018;**211**:906-911. DOI: 10.1016/j.fuel.2017.09.095

[15] Drexler DA, Tóth J. Global controllability of chemical reactions. Journal of Mathematical Chemistry. 2016;**54**(6):1327-1350. DOI: 10.1007/s10910-016-0626-7

[16] Joubert D, Stigter JD, Molenaar J. "Assessing the role of initial conditions

in the local structural identifiability of large dynamic models," Scientific Reports, vol. 11, no. 1, Art. no. 1, 2021, doi: 10.1038/s41598-021-96293-9

[17] Stigter JD, van Willigenburg LG, Molenaar J. An efficient method to assess local controllability and observability for non-linear systems. IFAC-PapersOnLine. 2018;**51**(2):535-540. DOI: 10.1016/j.ifacol.2018.03.090

[18] Van Willigenburg LG, Stigter JD, Molenaar J. Sensitivity matrices as keys to local structural system properties of large-scale nonlinear systems. Nonlinear Dynamics. 2022;**107**(3):2599-2618. DOI: 10.1007/s11071-021-07125-4

[19] Van Willigenburg LG, Stigter JD, Molenaar J. Establishing local strong accessibility of large-scale nonlinear systems by replacing the lie algebraic rank condition. In: Proceedings European Control Conference. The Netherlands: Rotterdam; 2021. pp. 2645-2650

[20] Mir I, Taha H, Eisa SA, Maqsood A. A controllability perspective of dynamic soaring. Nonlinear Dynamics. 2018; **94**(4):2347-2362. DOI: 10.1007/s11071-018-4493-6

[21] Kwakernaak and Sivan. Linear Optimal Control Systems. New York: Wiley; 1972

[22] Hermann R, Krener A. Nonlinear controllability and observability. IEEE Transactions on Automatic Control. 1977;**22**(5):728-740. DOI: 10.1109/TAC.1977.1101601

[23] Saccomani MP, Audoly S, D'Angio L. Parameter identifiability of nonlinear systems: The role of initial conditions. Automatica. 2003;**39**(4): 619-632. DOI: 10.1016/S0005-1098(02)00302-3

[24] Jean F. The car with N trailers: Characterization of the singular configurations. ESAIM. 1996;**1**:241-266. DOI: 10.1051/cocv:1996108

[25] Van Willigenburg LG, De Koning WL. A Kalman decomposition to detect temporal linear system structure. In: 2007 European Control Conference (ECC). Elsevier; 2007. pp. 1721-1726. DOI: 10.23919/ECC.2007.7068259

[26] Van Willigenburg LG, De Koning WL. Temporal linear system structure. IEEE Transactions on Automatic Control. 2008;**53**(5): 1318-1323. DOI: 10.1109/TAC.2008.921033

[27] Van Willigenburg LG, De Koning WL. Temporal and differential stabilizability and detectability of piecewise constant rank systems. Optimal Control Applications and Methods. 2012;**33**(3):302-317. DOI: 10.1002/oca.997

[28] Stigter JD, Molenaar J. A fast algorithm to assess local structural identifiability. Automatica. 2015;**58**: 118-124. DOI: 10.1016/j.automatica.2015.05.004

Chapter 8

Perspective Chapter: Families of Seventh-Order KdV Equations Having Traveling Wave and Soliton Solutions

Alvaro Humberto Salas Salas

Abstract

In this paper, we consider the problem of finding traveling wave solutions to the generalized seventh-order KdV equation (KdV7). Solitons are non-linear waves that exhibit extremely unexpected and interesting behavior—solitary waves that propagate without deformation. We use different approaches in order to find one and multisoliton solutions. Soliton travels through liquid, solid, and gaseous media and even as electron waves through an electromagnetic field. Making use of a traveling wave transformation, we obtain a non-linear ode, which is solved using either hyperbolic or elliptic algorithm. We also use the Hirota method to get the bilinear form, and then we may obtain multisoliton solutions. In the end, we consider the forced KdV7.

Keywords: traveling wave solutions, KdV7, solitons, cnoidal waves, deformed sine-Gordon equation, Sawada-Kotera equation, Kaup–Kupershmidt equation, Ito equation, lLax equation

1. Introduction

One of the most notable achievements in the second half of the twentieth century, which also clearly illustrates the underlying unity of Mathematics and Nonlinear Physics, is the Theory of Solitons. Solitons are nonlinear waves exhibiting extremely unexpected and interesting behavior—solitary waves propagating without deformation.

The other waves, the nonlinear ones, are less familiar and are very different from the linear ones. A wave in the sea approaching the shore is a good example of a nonlinear wave. Note that the amplitude, wavelength, and speed vary as the wave advances, while in linear waves, these are constant. The distance between the crests decreases, the height of the waves increases as they perceive the bottom, and the speed changes. The upper part of the wave overtakes the lower part, falls on it, and the wave breaks. There are even more intricate phenomena such as two waves that

intersect, interact in complicated and nonlinear ways, and give rise to three waves instead of two.

Now we come to solitons. During a horseback ride around Edinburgh, on the Union Canal in Hermiston, very close to the Riccarton campus of Heriot-Watt University, the Scottish engineer John Scott-Russell watched as a barge was towed along a narrow canal by two horses that pulled from land to obtain a more efficient design of boats.

A decisive step in the theory of integrable systems was the integration of the KdV equation. Thus, Gardner, Greene, Kruskal, and Miura observed that if we consider a potential $u(x)$ for the stationary Schrödinger equation on the line, the corresponding scattering data are transformed extremely easily when the potential changes as long as $u(x,t)$ satisfies the KdV equation. Therefore, given an initial condition $u(x)$ for KdV, we can find the associated scattering data and determine its evolution immediately.

In this paper, we consider the following generalized seventh-order KdV equation (KdV7 for short):

$$u_t + au^3u_x + bu_x^3 + cuu_xu_{2x} + du^2\,u_{3x} + \alpha u_{2x}u_{3x} + \beta u_xu_{4x} + \gamma uu_{5x} + u_{7x} = 0. \quad (1)$$

This nonlinear evolution equation describes the behavior of physical phenomena such as shallow water waves and plasmas. Its conservation laws were determined to predict its complete integrability [1, 2]. In Ref. [3], Wazwaz obtained one and two soliton solutions for the following special cases:

- The seventh-order Sawada-Kotera-Ito equation:

$$u_t + 252u^3u_x + 63u_x^3 + 378u_xu_{2x} + 126u^2\,u_{3x} + 63u_{2x}u_{3x} + 42u_xu_{4x} + 21uu_{5x} + u_{7x} = 0. \quad (2)$$

- The seventh-order Lax equation:

$$u_t + 140u^3u_x + 70u_x^3 + 280u_xu_{2x} + 70u^2\,u_{3x} + 70u_{2x}u_{3x} + 42u_xu_{4x} + 14uu_{5x} + u_{7x} = 0. \quad (3)$$

- The seventh-order Kaup-Kuperschmidt equation

$$u_t + 2016u^3u_x + 630u_x^3 + 2268u_xu_{2x} + 504u^2\,u_{3x} + 252u_{2x}u_{3x} + 147u_xu_{4x} + 42uu_{5x} + u_{7x} = 0. \quad (4)$$

These three cases of the seventh-order KdV equation are completely integrable. This means that each of these equations admits an infinite number of conservation laws, and as a result, each gives rise to N-soliton solutions. We aim to describe the families of these KdV7 that admit soliton and cnoidal wave solutions.

2. Cnoidal wave solutions

Let

$$u(x,t) = p + q\wp\left(x - \lambda t + \xi_0; g_2, g_3\right). \quad (5)$$

Then

$$\begin{aligned}
&u_t + au^3u_x + bu_x^3 + cuu_xu_{2x} + du^2\,u_{3x} + \alpha u_{2x}u_{3x} + \beta u_xu_{4x} + \gamma uu_{5x} + u_{7x} = \\
&\frac{1}{2}q\sqrt{-g_2\wp - g_3 + 4\wp^3}[2ap^3 - 2bg_3q^2 - cg_2pq - 36\gamma g_2p - 24\beta g_3q - 1440g_3 - 2\lambda \\
&+ (6ap^2q - 2bg_2q^2 - cg_2q^2 + 24dp^2 - 12\alpha g_2q - 36\beta g_2q - 36\gamma g_2q - 4032g_2)\wp + \\
&6p\left(aq^2 + 2cq + 120\gamma + 8dq\right)\wp^2 + \\
&2\left(aq^3 + 4bq^2 + 6cq^2 + 12dq^2 + 72\alpha q + 120\beta q + 360\gamma q + 20160\right)\wp^3],
\end{aligned} \tag{5a}$$

where $\wp = \wp(x - \lambda t + \xi_0; g_2, g_3)$.
The system to be solved is

$$\begin{aligned}
&2ap^3 - 2bg_3q^2 - cg_2pq - 36\gamma g_2 - 24\beta g_3q - 1440g_3 + 2\lambda = 0.\\
&6ap^2q - 2bg_2q^2 - cg_2q^2 + 24dp^2 - 12\alpha g_2q - 36\beta g_2q - 4032g_2 = 0.\\
&6apq^2 + 12cpq + 720\gamma + 48dpq = 0.\\
&2aq^3 + 8bq^2 + 12cq^2 + 24dq^2 + 144\alpha q + 240\beta q + 40320 = 0.
\end{aligned} \tag{6}$$

We will have a solution for the following parameter values:

$$\begin{aligned}
g_2 &= \frac{6p^2(aq + 4d)}{2bq^2 + cq^2 + 12\alpha q + 36\beta q + 4032}.\\
g_3 &= \frac{2ap^3 - cg_2pq - 36\gamma g_2 + 2\lambda}{2\left(bq^2 + 12\beta q + 720\right)}.\\
p &= -\frac{120\gamma}{q(aq + 2c + 8d)}.\\
&aq^3 + 2(2b + 3c + 6d)q^2 + 24\,(3\alpha + 5\beta)q + 20160 = 0.
\end{aligned} \tag{7}$$

Example. Let

$$\begin{aligned}
&u^{(0,1)}(x,t) - 63.5567u(x,t)^3u^{(1,0)}(x,t) - 5089.7u^{(1,0)}(x,t)^3 + 0.939611u(x,t)u^{(1,0)}(x,t)u^{(2,0)}(x,t)\\
&+0.471886u\,(x,t)^2u^{(3,0)}(x,t) + 0.0310296u^{(2,0)}(x,t)u^{(3,0)}(x,t) + 0.626069u^{(1,0)}(x,t)u^{(4,0)}(x,t)\\
&+0.48252u(x,t)u^{(5,0)}(x,t) + u^{(7,0)}(x,t) = 0.
\end{aligned} \tag{8}$$

See **Figure 1**.
Let now

$$u(x,t) = B + C\,\mathrm{cn}^2(\omega x - \lambda t + \xi_0, m). \tag{9}$$

Then

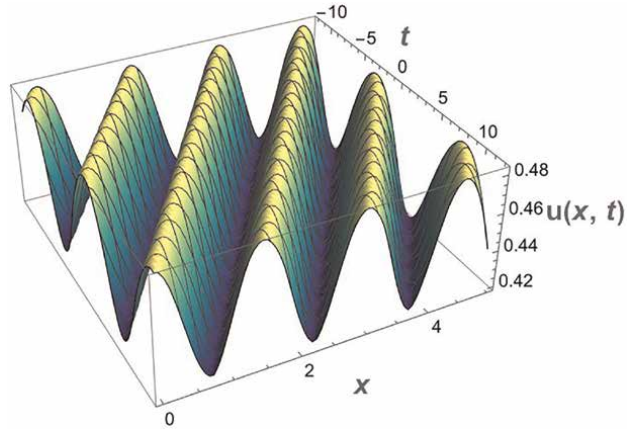


Figure 1.
$u(x,t) = 0.480062 + \wp(-7.21261t - x - (4.70653 \ \ - 5.33465i); 0.0139651, - 0.0000274235)$.

$$u_t + au^3u_x + bu_x^3 + cuu_xu_{2x} + du^2\,u_{3x} + \alpha u_{2x}u_{3x} + \beta u_xu_{4x} + \gamma uu_{5x} + u_{7x} =$$

$$2C\,\mathrm{cn}\,\mathrm{sn}\,\mathrm{dn}\,[-\begin{pmatrix} \lambda - aB^3\omega - 2BcC\omega^3 + 4B^2d\omega^3 + 2BcCm\omega^3 - 8B^2dm\omega^3 + 8C\alpha\omega^5 - 24Cm\alpha\omega^5 + 16Cm^2\alpha\omega^5 \\ +8C\beta\omega^5 - 24Cm\beta\omega^5 + 16Cm^2\beta\omega^5 - 16B\gamma\omega^5 + 136Bm\gamma\omega^5 - 136Bm^2\gamma\omega^5 + 64\omega^7 \\ -2112m\omega^7 + 5952m^2\omega^7 - 3968m^3\omega^7 \end{pmatrix}\mathrm{cn}$$

$$+\omega\begin{pmatrix} 3aB^2C - 4BcC\omega^2 + 4bC^2\omega^2 + 2cC^2\omega^2 - 8BCd\omega^2 + 8BcCm\omega^2 - 4bC^2m\omega^2 - 2cC^2m\omega^2 - 12B^2dm\omega^2 \\ +16BCdm\omega^2 + 16C\alpha\omega^4 - 88Cm\alpha\omega^4 + 88Cm^2\alpha\omega^4 + 16C\beta\omega^4 - 136Cm\beta\omega^4 + 136Cm^2\beta\omega^4 + 16C\gamma\omega^4 \\ +240Bm\gamma\omega^4 - 136Cm\gamma\omega^4 - 480Bm^2\gamma\omega^4 + 136Cm^2\gamma\omega^4 - 4032m\omega^6 + 24192m^2\omega^6 - 24192m^3\omega^6 \end{pmatrix}\mathrm{cn}^3$$

$$+\omega\begin{pmatrix} 3aBC^2 - 4bC^2\omega^2 - 4cC^2\omega^2 - 4C^2d\omega^2 - 6BcCm\omega^2 + 8bC^2m\omega^2 + 8cC^2m\omega^2 - 24BCdm\omega^2 \\ +8C^2dm\omega^2 + 72Cm\alpha\omega^4 - 144Cm^2\alpha\omega^4 + 120Cm\beta\omega^4 - 240Cm^2\beta\omega^4 + 240Cm\gamma\omega^4 + 360Bm^2\gamma\omega^4 \\ -480Cm^2\gamma\omega^4 - 20160m^2\omega^6 + 40320m^3\omega^6 \end{pmatrix}\mathrm{cn}^5$$

$$+\omega\left(aC^3 - 4bC^2m\omega^2 - 6cC^2m\omega^2 - 12C^2dm\omega^2 + 72Cm^2\alpha\omega^4 + 120Cm^2\beta\omega^4 + 360Cm^2\gamma\omega^4 - 20160m^3\omega^6\right)\mathrm{cn}^7].$$

Next, we equate to zero the coefficients of cn^j to obtain an algebraic system of nonlinear equations. This system admits a solution under the condition

$$\Delta_1\Delta_2 = 0, \tag{10}$$

where

$\Delta_1 = 17640a^2 + 27a\alpha^2\gamma + 90a\alpha\beta\gamma + 180a\alpha\gamma^2 + 75a\beta^2\gamma + 300a\beta\gamma^2 - 840ab\gamma + 300a\gamma^3 - 126a\alpha c - 210a\beta c$
$-1260a\gamma c - 504a\alpha d - 840a\beta d - 2520a\gamma d + 10b^2\gamma^2 + 14bc^2 - 3ab\gamma c - 5b\beta\gamma c + 10b\gamma^2 c + 112bcd + 224bd^2$
$-12ab\gamma d - 20b\beta\gamma d - 20b\gamma^2 d + 14c^3 - 3\alpha\gamma c^2 - 5\beta\gamma c^2 + 126c^2 d + 336cd^2 - 15\alpha\gamma cd - 25\beta\gamma cd - 30\gamma^2 cd + 224d^3$
$-12\alpha\gamma d^2 - 20\beta\gamma d^2 - 30\gamma^2 d^2$, and
$\Delta_2 = 28224a^2 + 3a\alpha^3 + 3a\alpha^2\beta + 63a\alpha^2\gamma - 63a\alpha\beta^2 + 234a\alpha\beta\gamma + 297a\alpha\gamma^2 - 135a\beta^3 + 135a\beta^2\gamma + 168a\alpha b$
$+3192a\beta b + 675a\beta\gamma^2 - 3528ab\gamma + 405a\gamma^3 - 252a\alpha c + 588a\beta c - 2772ac\gamma - 1008a\alpha d - 3024a\beta d - 3024a\gamma d$
$+504b^3 + 4\alpha^2 b^2 - 16\alpha\beta b^2 + 84\alpha b^2\gamma - 20\beta^2 b^2 + -60\beta b^2\gamma + 360b^2\gamma^2 + 84b^2 c - 2016b^2 d - 70bc^2 + 2\alpha^2 bc$
$-12\alpha\beta bc + 48\alpha bc\gamma + 10\beta^2 bc - 120\beta bc\gamma + 270bc\gamma^2 - 672bcd + 2016bd^2 - 6\alpha^2 bd + 12\alpha\beta bd - 108\alpha b\gamma d$
$+90\beta^2 bd - 180\beta b\gamma d - 270b\gamma^2 d + 7c^3 - 2\alpha\beta c^2 + 3\alpha c^2\gamma + 10\beta^2 c^2 - 45\beta c^2\gamma + 45c^2\gamma^2 + 168c^2 d + 1008cd^2$
$-3\alpha^2 cd + 6\alpha\beta cd - 54\alpha c\gamma d + 45\beta^2 cd - 90\beta c\gamma d - 135c\gamma^2 d.$

The Eqs. (2)–(4) obey the condition in Eq. (10).
Solving the system we obtain the solutions as follows:

$$\begin{aligned}
\lambda = \omega &\left(\begin{array}{c} aB^3 + 8B^2dm\omega^2 - 4B^2d\omega^2 - 2BcCm\omega^2 + 2BcC\omega^2 \\ +16B\gamma\omega^4 + 136B\gamma m^2\omega^4 - 136B\gamma m\omega^4 - 8\alpha C\omega^4 \\ -8\beta C\omega^4 - 16\alpha Cm^2\omega^4 - 16\beta Cm^2\omega^4 + 24\alpha Cm\omega^4 \\ +24\beta Cm\omega^4 + 3968m^3\omega^6 - 5952m^2\omega^6 + 2112m\omega^6 - 64\omega^6 \end{array} \right). \\
B = &-\frac{4(2m-1)\omega^2}{3\left(aC^2 - 2cCm\omega^2 - 8Cdm\omega^2 + 120\gamma m^2\omega^4\right)}(bC^2 + cC^2 + C^2d - 18\alpha Cm\omega^2 - 30\beta Cm\omega^2 \\
&-60\gamma Cm\omega^2 + 5040m^2\omega^4). \\
aC^3 + &\left(-4bm\omega^2 - 6cm\omega^2 - 12dm\omega^2\right)C^2 + \left(72\alpha m^2\omega^4 + 120\beta m^2\omega^4 + 360\gamma m^2\omega^4\right)C \\
&-20160m^3\omega^6 = 0.
\end{aligned} \tag{11}$$

- Sawada-Kotera-Ito Eq. (2):

$$\begin{aligned}
C &= 2m\omega^2 \\
\lambda &= 4\omega\left(63B^3 + 252B^2m\omega^2 - 126B^2\omega^2 + 336Bm^2\omega^4 - 336Bm\omega^4 + 84B\omega^4\right. \\
&\quad \left. +152m^3\omega^6 - 228m^2\omega^6 + 108m\omega^6 - 16\omega^6\right). \\
u(x,t) &= B + C\mathrm{cn}^2\left(\sqrt{\omega}x - \lambda t\right) + \xi_0|m).
\end{aligned} \tag{12}$$

$$\begin{aligned}
B &= -\frac{4}{3}(2m-1)\omega^2, C = 4m\omega^2. \\
\lambda &= \frac{128}{3}(m-2)(m+1)(2m-1)\omega^7. \\
u(x,t) &= B + C\mathrm{cn}^2\left(\sqrt{\omega}x - \lambda t\right) + \xi_0|m).
\end{aligned} \tag{13}$$

- Lax Eq. (3):

$$\begin{aligned}
&C = 2m\omega^2. \\
&\lambda = 4\omega\left(35B^3 + 140B^2m\omega^2 - 70B^2\omega^2 + 196Bm^2\omega^4 - 196Bm\omega^4 + 56B\omega^4 + 96m^3\omega^6\right. \\
&\quad \left. -144m^2\omega^6 + 80m\omega^6 - 16\omega^6\right)u(x,t) = B + 2m\omega\mathrm{cn}^2\left(\sqrt{\omega}(x - \lambda t) + \xi_0|m\right).
\end{aligned} \tag{14}$$

- Kaup-Kuperschmidt Eq. (4):

$$\begin{aligned}
B &= -\frac{1}{6}(2m-1)\omega^2, C = \frac{m\omega^2}{2}. \\
\lambda &= \frac{2}{3}(m-2)(m+1)(2m-1)\omega^7. \\
u(x,t) &= B + 2m\omega\mathrm{cn}^2\left(\sqrt{\omega}(x - \lambda t) + \xi_0|m\right).
\end{aligned} \tag{15}$$

- Letting m = 1 in Eqs. (2)–(4), we obtain solitonic solutions.

3. Soliton solutions

3.1 First family

Let

$$A = \frac{252}{\alpha+\beta+\gamma}, w = k^7.$$
$$c = -\frac{1}{126}(\alpha+\beta+\gamma)(\alpha-5(\beta+4\gamma)) - b - d. \tag{16}$$
$$d = \frac{42a}{\alpha+\beta+\gamma} + \frac{1}{378}(\alpha+\beta+\gamma)(\alpha-5\beta+10\gamma) + \frac{b}{3}.$$

We make the transformation

$$u = A\partial_{x,x}\log(1+\exp(kx-wt)) \tag{17}$$

to obtain the soliton solution

$$u_{\text{soliton}}(x,t) = \frac{126k^2}{(\alpha+\beta+\gamma)\left(1+\cosh\left(kx-k^7t-\xi\right)\right)}. \tag{18}$$

For a graphical illustration, see **Figure 2**.
We are interested in the existence of two soliton solutions. Let

$$u(x,t) = \frac{252}{\alpha+\beta+\gamma}\partial_{x,x}\log(1+\exp\theta_1+\exp\theta_2+\rho\exp(\theta_1+\theta_2)).$$
$$\theta_1 = k_1x - k_1^7t \text{ and } \theta_2 = k_2x - k_2^7t \tag{19}$$

Making the choices

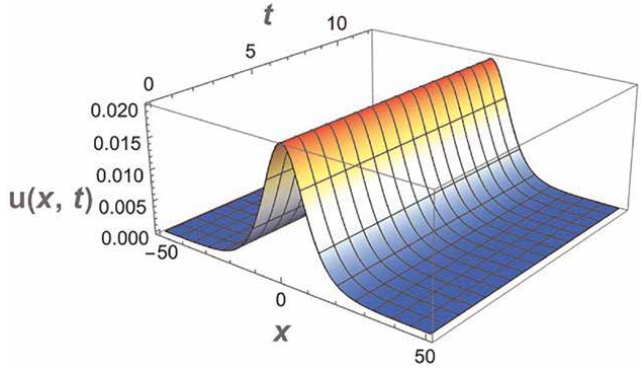


Figure 2.
Soliton solution $u(x,t) = \frac{0.04}{1+\cosh\ (0.0000128t-0.2x)}$ *to the Sawada-Kotera-Ito equation with* k = 0.2.

$$a = -\frac{\gamma^2(\beta+2\gamma)(\beta^3-3\beta^2\gamma-9\beta\gamma^2+31\gamma^3)}{49(\beta-5\gamma)^3}$$

$$b = -\frac{(\beta-\gamma)(\beta^3-6\beta^2\gamma+15\beta\gamma^2-23\gamma^3)}{7(\beta-5\gamma)^2}$$

$$c = \frac{\beta^4-9\beta^3\gamma+25\beta^2\gamma^2-33\beta\gamma^3+76\gamma^4}{7(\beta-5\gamma)^2}$$

$$d = \frac{2\gamma(\beta^3-5\beta^2\gamma+2\beta\gamma^2+17\gamma^3)}{7(\beta-5\gamma)^2}$$

$$\alpha = -\frac{\beta^2-4\beta\gamma+13\gamma^2}{\beta-5\gamma}$$

$$\rho = \frac{\gamma^2(k_1-k_2)^2\left(k_1{}^2-k_1k_2+k_2{}^2\right)^2}{(k_1+extk_2)^2\left(2\beta^2k_1{}^2k_2{}^2-2\beta\gamma k_1k_2\left(k_1{}^2+4k_1k_2+k_2{}^2\right)+\gamma^2\left(k_1{}^4+4k_1{}^3k_2+9k_1{}^2k_2{}^2+4k_1k_2{}^3+k_2{}^4\right)\right)}. \tag{20}$$

We obtain

$$u_t + au^3u_x + bu_x^3 + cuu_xu_{2x} + du^2\,u_{3x} + \alpha\,u_{2x}u_{3x} + \beta u_xu_{4x} + \gamma u_{5x} + u_{7x} = (\beta-\gamma)(\beta-2\gamma)(\beta-3\gamma)R(k_1,k_2,\theta_1,\theta_2).$$

We conclude that the two soliton solutions exist for the parameter values in Eq. (20) under the condition

$$(\beta-\gamma)(\beta-2\gamma)(\beta-3\gamma) = 0. \tag{21}$$

We obtained the following result:

Theorem. The following families of KdV7 admit one and two soliton solutions for any p. The two soliton solutions have the form

$$u_{2-\text{soliton}}(x,t) = \frac{252}{\alpha+\beta+\gamma}\partial_{xx}\log\left(1+\exp(\theta_1)+\exp(\theta_2)+\rho\exp(\theta_1+\theta_2)\right), \tag{22}$$

being

$$\theta_1 = k_1x - k_1^7t, \theta_2 = k_2x - k_2^7t. \tag{23}$$

- First set:

$$\left\{\begin{array}{c} a=\frac{15p^3}{784}, b=0, c=\frac{15p^2}{28}, d=\frac{15p^2}{56}, \alpha=\frac{5p}{2}, \beta=p, \\ \rho=\frac{(k_1-k_2)^2\left(k_1{}^2-k_1k_2+k_2{}^2\right)^2}{(k_1+k_2)^2\left(k_1{}^2+k_1k_2+k_2{}^2\right)^2}, \gamma=p \end{array}\right\} \tag{24}$$

KdV7:

$$u_t + \frac{15}{784}p^3u^3u_x + \frac{15}{28}p^2uu_xu_{2x} + \frac{15}{56}p^2u^2u_{3x} + \frac{5}{2}pu_{2x}u_{3x} + pu_xu_{4x} + puu_{5x} + u_{7x} = 0 \tag{25}$$

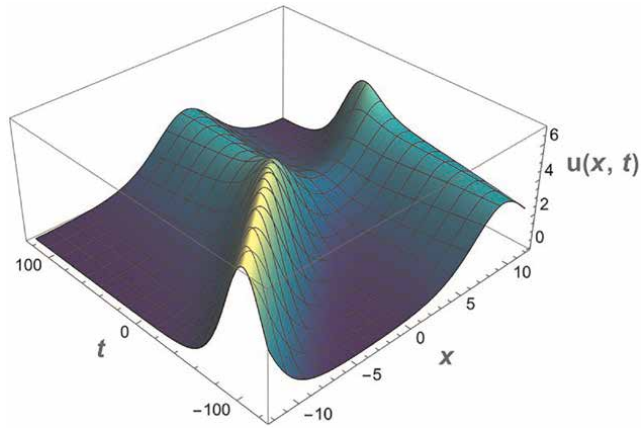


Figure 3.
Two soliton solution for p = 1.

An illustration with $p = 1$ is shown in **Figure 3**. The solution is

$$u(x,t) = \frac{14.e^{0.352855t+0.5x} + 27.44e^{0.278313t+0.7x} + 2.52676e^{0.2705t+1.2x} + 0.0975793e^{0.262688t+1.7x} + 0.0497854e^{0.188146t+1.9x}}{\left(e^{0.172521t+0.5x} + e^{0.0979793t+0.7x} + 0.0035561e^{0.0901668t+1.2x} + e^{0.180334t}\right)^2} \tag{26}$$

- Second set:

$$\left\{ \begin{array}{c} a = \frac{4p^3}{147}, b = \frac{p^2}{7}, c = \frac{6p^2}{7}, d = \frac{2p^2}{7}, \alpha = 3p, \beta = 2p, \\ \rho = \frac{(k_1 - k_2)^2\left(k_1^2 - k_1k_2 + k_2^2\right)}{(k_1 + k_2)^2\left(k_1^2 + k_1k_2 + k_2^2\right)}, \gamma = p \end{array} \right\} \tag{27}$$

KdV7:

$$u_t + \frac{4}{147}p^3u^3u_x + \frac{1}{7}p^2u_x^3 + \frac{6}{7}p^2uu_xu_{2x} + \frac{2}{7}p^2u^2u_{3x} + 3pu_{2x}u_{3x} + 2pu_xu_{4x} + puu_{5x} + u_{7x} = 0 \tag{28}$$

- Third set:

$$\left\{ \begin{array}{c} a = \frac{5p^3}{98}, b = \frac{5p^2}{14}, c = \frac{10p^2}{7}, d = \frac{5p^2}{14}, \alpha = 5p, \beta = 3p, \\ \rho = \frac{(k_1 - k_2)^2}{(k_1 + k_2)^2}, \gamma = p \end{array} \right\} \tag{29}$$

KdV7:

$$u_t + \frac{5}{98}p^3u^3u_x + \frac{5}{14}p^2u_x^3 + \frac{10}{7}p^2uu_xu_{2x} + \frac{5}{14}p^2u^2u_{3x} + 5pu_{2x}u_{3x} + 3pu_xu_{4x} + puu_{5x} + u_{7x} = 0 \tag{30}$$

Now, our aim is to find three soliton solutions for the parameter values in Eq. (20). Assume the ansatz

$$u(x,t) = \frac{252}{\alpha+\beta+\gamma}\partial_{xx}\log\left(1+\sum_{j=1}^{3}\eta_j+\sum_{1\le i<j\le 3}\rho_{i,j}\eta_i\eta_j+\rho_{1,2,3}\eta_1\eta_2\eta_3\right), \tag{31}$$

where

$$\eta_j = \exp\left(k_j x - k_j^7 t\right) \text{ for } j = 1,2,3. \tag{32}$$

We have:

$$\begin{aligned}
&u_t + au^3u_x + bu_x^3 + cuu_xu_{2x} + du^2\,u_{3x} + \alpha\,u_{2x}u_{3x} + \beta u_xu_{4x} + \gamma uu_{5x} + u_{7x} =\\
&98(\beta-5\gamma)k_1k_2(k_1+k_2)\left(\gamma^2k_1^6-\gamma^2\rho_{1,2}k_1^6-6\gamma^2k_2k_1^5+2\beta\gamma k_2k_1^5-4\gamma^2k_2\rho_{1,2}k_1^5\right.\\
&+2\beta^2k_2^2k_1^4+18\gamma^2k_2^2k_1^4-12\beta\gamma k_2^2k_1^4-8\gamma^2k_2^2\rho_{1,2}k_1^4-4\beta^2k_2^3k_1^3-26\gamma^2k_2^3k_1^3+20\beta\gamma k_2^3k_1^3\\
&-10\gamma^2k_2^3\rho_{1,2}k_1^3+2\beta^2k_2^4k_1^2+18\gamma^2k_2^4k_1^2-12\beta\gamma k_2^4k_1^2-8\gamma^2k_2^4\rho_{1,2}k_1^2-6\gamma^2k_2^5k_1+2\beta\gamma k_2^5k_1\\
&\left.-4\gamma^2k_2^5\rho_{1,2}k_1+\gamma^2k_2^6-\gamma^2k_2^6\rho_{1,2}\right)/\gamma^4z_1z_2\\
&+98(\beta-5\gamma)k_1k_3(k_1+k_3)\left(\gamma^2k_1^6-\gamma^2\rho_{1,3}k_1^6-6\gamma^2k_3k_1^5+2\beta\gamma k_3k_1^5-4\gamma^2k_3\rho_{1,3}k_1^5\right.\\
&+2\beta^2k_3^2k_1^4+18\gamma^2k_3^2k_1^4-12\beta\gamma k_3^2k_1^4-8\gamma^2k_3^2\rho_{1,3}k_1^4-4\beta^2k_3^3k_1^3-26\gamma^2k_3^3k_1^3+20\beta\gamma k_3^3k_1^3-\\
&10\gamma^2k_3^3\rho_{1,3}k_1^3+2\beta^2k_3^4k_1^2+18\gamma^2k_3^4k_1^2-12\beta\gamma k_3^4k_1^2-8\gamma^2k_3^4\rho_{1,3}k_1^2-6\gamma^2k_3^5k_1+2\beta\gamma k_3^5k_1\\
&\left.-4\gamma^2k_3^5\rho_{1,3}k_1+\gamma^2k_3^6-\gamma^2k_3^6\rho_{1,3}\right)/\gamma^4z_1z_3\\
&+98(\beta-5\gamma)k_2k_3(k_2+k_3)\left(\gamma^2k_2^6-\gamma^2\rho_{2,3}k_2^6-6\gamma^2k_3k_2^5+2\beta\gamma k_3k_2^5-4\gamma^2k_3\rho_{2,3}k_2^5\right.\\
&+2\beta^2k_3^2k_2^4+18\gamma^2k_3^2k_2^4-12\beta\gamma k_3^2k_2^4-8\gamma^2k_3^2\rho_{2,3}k_2^4-4\beta^2k_3^3k_2^3-26\gamma^2k_3^3k_2^3+20\beta\gamma k_3^3k_2^3\\
&-10\gamma^2k_3^3\rho_{2,3}k_2^3+2\beta^2k_3^4k_2^2+18\gamma^2k_3^4k_2^2-12\beta\gamma k_3^4k_2^2-8\gamma^2k_3^4\rho_{2,3}k_2^2-6\gamma^2k_3^5k_2+2\beta\gamma k_3^5k_2\\
&\left.-4\gamma^2k_3^5\rho_{2,3}k_2+\gamma^2k_3^6-\gamma^2k_3^6\rho_{2,3}\right)/\gamma^4z_2z_3+\text{h.o.t}
\end{aligned}$$

Equating to zero the coefficients of z_1z_2, z_1z_3, and z_2z_3 and solving the resulting system of algebraic equations we obtain

$$\begin{aligned}
\rho_{1,2} &= \frac{(k_1-k_2)^2\left(2\beta^2k_2^2k_1^2+2\beta\gamma k_2k_1^3-8\beta\gamma k_2^2k_1^2+2\beta\gamma k_2^3k_1+\gamma^2k_1^4-4\gamma^2k_2k_1^3+9\gamma^2k_2^2k_1^2-4\gamma^2k_2^3k_1+\gamma^2k_2^4\right)}{\gamma^2(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)^2}.\\
\rho_{1,3} &= \frac{(k_1-k_3)^2\left(2\beta^2k_3^2k_1^2+2\beta\gamma k_3k_1^3-8\beta\gamma k_3^2k_1^2+2\beta\gamma k_3^3k_1+\gamma^2k_1^4-4\gamma^2k_3k_1^3+9\gamma^2k_3^2k_1^2-4\gamma^2k_3^3k_1+\gamma^2k_3^4\right)}{\gamma^2(k_1+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)^2}.\\
\rho_{2,3} &= \frac{(k_2-k_3)^2\left(2\beta^2k_3^2k_2^2+2\beta\gamma k_3k_2^3-8\beta\gamma k_3^2k_2^2+2\beta\gamma k_3^3k_2+\gamma^2k_2^4-4\gamma^2k_3k_2^3+9\gamma^2k_3^2k_2^2-4\gamma^2k_3^3k_2+\gamma^2k_3^4\right)}{\gamma^2(k_2+k_3)^2\left(k_2^2+k_3k_2+k_3^2\right)^2}.
\end{aligned} \tag{33}$$

Next, we equate to zero the coefficient of $z_1z_2z_3$ in order to obtain the value for $\rho_{1,2,3}$. It is given as follows.

- For $\beta = \gamma$:

$\rho_{1,2,3} = P_1/Q_1$, where
$P_1 = \gamma^4(k_2-k_1)^2(k_2-k_3)^2(k_3-k_1)^2(k_1+k_2+k_3)$
$(k_2^4k_1^{13}+k_3^4k_1^{13}-2k_2k_3^3k_1^{13}+3k_2^2k_3^2k_1^{13}-2k_2^3k_3k_1^{13}+k_2^5k_1^{12}+k_3^5k_1^{12}-k_2k_3^4k_1^{12}+k_2^2k_3^3k_1^{12}$
$+k_2^3k_3^2k_1^{12}-k_2^4k_3k_1^{12}+2k_2^6k_1^{11}+2k_3^6k_1^{11}-3k_2k_3^5k_1^{11}+6k_2^2k_3^4k_1^{11}-5k_2^3k_3^3k_1^{11}+6k_2^4k_3^2k_1^{11}$
$-3k_2^5k_3k_1^{11}+2k_2^7k_1^{10}+2k_3^7k_1^{10}-4k_2k_3^6k_1^{10}-10k_2^2k_3^5k_1^{10}-18k_2^3k_3^4k_1^{10}-18k_2^4k_3^3k_1^{10}-10k_2^5k_3^2k_1^{10}$
$-4k_2^6k_3k_1^{10}+3k_2^8k_1^9+3k_3^8k_1^9-6k_2k_3^7k_1^9-7k_2^2k_3^6k_1^9-38k_2^3k_3^5k_1^9-26k_2^4k_3^4k_1^9-38k_2^5k_3^3k_1^9-7k_2^6k_3^2k_1^9$
$-6k_2^7k_3k_1^9+3k_2^9k_1^8+3k_3^9k_1^8-3k_2k_3^8k_1^8-7k_2^2k_3^7k_1^8-35k_2^3k_3^6k_1^8-58k_2^4k_3^5k_1^8-58k_2^5k_3^4k_1^8-35k_2^6k_3^3k_1^8$
$-7k_2^7k_3^2k_1^8-3k_2^8k_3k_1^8+2k_2^{10}k_1^7+2k_3^{10}k_1^7-6k_2k_3^9k_1^7-7k_2^2k_3^8k_1^7-41k_2^3k_3^7k_1^7-60k_2^4k_3^6k_1^7-99k_2^5k_3^5k_1^7$
$-60k_2^6k_3^4k_1^7-41k_2^7k_3^3k_1^7-7k_2^8k_3^2k_1^7-6k_2^9k_3k_1^7+2k_2^{11}k_1^6+2k_3^{11}k_1^6-4k_2k_3^{10}k_1^6-7k_2^2k_3^9k_1^6-35k_2^3k_3^8k_1^6$
$-60k_2^4k_3^7k_1^6-92k_2^5k_3^6k_1^6-92k_2^6k_3^5k_1^6-60k_2^7k_3^4k_1^6-35k_2^8k_3^3k_1^6-7k_2^9k_3^2k_1^6-4k_2^{10}k_3k_1^6+k_2^{12}k_1^5+k_3^{12}k_1^5$
$-3k_2k_3^{11}k_1^5-10k_2^2k_3^{10}k_1^5-38k_2^3k_3^9k_1^5-58k_2^4k_3^8k_1^5-99k_2^5k_3^7k_1^5-92k_2^6k_3^6k_1^5-99k_2^7k_3^5k_1^5-58k_2^8k_3^4k_1^5$
$-38k_2^9k_3^3k_1^5-10k_2^{10}k_3^2k_1^5-3k_2^{11}k_3k_1^5+k_2^{13}k_1^4+k_3^{13}k_1^4-k_2k_3^{12}k_1^4+6k_2^2k_3^{11}k_1^4-18k_2^3k_3^{10}k_1^4-26k_2^4k_3^9k_1^4$
$-58k_2^5k_3^8k_1^4-60k_2^6k_3^7k_1^4-60k_2^7k_3^6k_1^4-58k_2^8k_3^5k_1^4-26k_2^9k_3^4k_1^4-18k_2^{10}k_3^3k_1^4+6k_2^{11}k_3^2k_1^4-k_2^{12}k_3k_1^4$
$-2k_2k_3^{13}k_1^3+k_2^2k_3^{12}k_1^3-5k_2^3k_3^{11}k_1^3-18k_2^4k_3^{10}k_1^3-38k_2^5k_3^9k_1^3-35k_2^6k_3^8k_1^3-41k_2^7k_3^7k_1^3-35k_2^8k_3^6k_1^3$
$-38k_2^9k_3^5k_1^3-18k_2^{10}k_3^4k_1^3-5k_2^{11}k_3^3k_1^3+k_2^{12}k_3^2k_1^3-2k_2^{13}k_3k_1^3+3k_2^2k_3^{13}k_1^2+k_2^3k_3^{12}k_1^2+6k_2^4k_3^{11}k_1^2$
$-10k_2^5k_3^{10}k_1^2-7k_2^6k_3^9k_1^2-7k_2^7k_3^8k_1^2-7k_2^8k_3^7k_1^2-7k_2^9k_3^6k_1^2-10k_2^{10}k_3^5k_1^2+6k_2^{11}k_3^4k_1^2+k_2^{12}k_3^3k_1^2$
$+3k_2^{13}k_3^2k_1^2-2k_2^3k_3^{13}k_1-k_2^4k_3^{12}k_1-3k_2^5k_3^{11}k_1-4k_2^6k_3^{10}k_1-6k_2^7k_3^9k_1-3k_2^8k_3^8k_1-6k_2^9k_3^7k_1-4k_2^{10}k_3^6k_1$
$-3k_2^{11}k_3^5k_1-k_2^{12}k_3^4k_1-2k_2^{13}k_3^3k_1+k_2^4k_3^{13}+k_2^5k_3^{12}+2k_2^6k_3^{11}+2k_2^7k_3^{10}+3k_2^8k_3^9+3k_2^9k_3^8+2k_2^{10}k_3^7$
$+2k_2^{11}k_3^6+k_2^{12}k_3^5+k_2^{13}k_3^4)$,

and

$$Q_1 = \gamma^4(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)^2(k_1+k_3)^2(k_2+k_3)^2(k_1+k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)^2$$
$$\left(k_2^2+k_3k_2+k_3^2\right)^2$$
$$(k_1^4+2k_2k_1^3+2k_3k_1^3+3k_2^2k_1^2+3k_3^2k_1^2+5k_2k_3k_1^2+2k_2^3k_1+2k_3^3k_1+5k_2k_3^2k_1+5k_2^2k_3k_1$$
$$+k_2^4+k_3^4+2k_2k_3^3+3k_2^2k_3^2+2k_2^3k_3).$$

- For $\beta = 2\gamma$:

$\rho_{1,2,3} = P_2/Q_2$, where
$P_2 = \gamma^4(k_2-k_1)^2\left(k_1^2-k_2k_1+k_2^2\right)\left(k_1^2+k_2k_1+k_2^2\right)(k_2-k_3)^2(k_3-k_1)^2(k_1+k_2+k_3)^2$
$\left(k_1^2-k_3k_1+k_3^2\right)\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2-k_3k_2+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)(k_1^4+2k_2k_1^3+2k_3k_1^3$
$+3k_2^2k_1^2+3k_3^2k_1^2+5k_2k_3k_1^2+2k_2^3k_1+2k_3^3k_1+5k_2k_3^2k_1+5k_2^2k_3k_1+k_2^4+k_3^4+2k_2k_3^3+3k_2^2k_3^2+2k_2^3k_3)$,

and

$$Q_2 = \gamma^4(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)^2(k_1+k_3)^2(k_2+k_3)^2(k_1+k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)^2$$
$$\left(k_2^2+k_3k_2+k_3^2\right)^2$$
$$(k_1^4+2k_2k_1^3+2k_3k_1^3+3k_2^2k_1^2+3k_3^2k_1^2+5k_2k_3k_1^2+2k_2^3k_1+2k_3^3k_1+5k_2k_3^2k_1+5k_2^2k_3k_1+k_2^4+k_3^4$$
$$+2k_2k_3^3+3k_2^2k_3^2+2k_2^3k_3).$$

- For $\beta = 3\gamma$:

$\rho_{1,2,3} = P_3/Q_3$, where
$P_3 = \gamma^4(k_2-k_1)^2\left(k_1^2-k_2k_1+k_2^2\right)\left(k_1^2+k_2k_1+k_2^2\right)(k_2-k_3)^2(k_3-k_1)^2(k_1+k_2+k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)$
$\left(k_1^2+k_3k_1+k_3^2\right)$
$\left(k_2^2-k_3k_2+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)(k_1^4+2k_2k_1^3+2k_3k_1^3+3k_2^2k_1^2+3k_3^2k_1^2+5k_2k_3k_1^2+2k_2^3k_1+2k_3^3k_1+5k_2k_3^2k_1$
$+5k_2^2k_3k_1+k_2^4+k_3^4+2k_2k_3^3+3k_2^2k_3^2+2k_2^3k_3)$,

and

$Q_3 = \gamma^4(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)^2(k_1+k_3)^2(k_2+k_3)^2(k_1+k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)^2\left(k_2^2+k_3k_2+k_3^2\right)^2$
$(k_1^4+2k_2k_1^3+2k_3k_1^3+3k_2^2k_1^2+3k_3^2k_1^2+5k_2k_3k_1^2+2k_2^3k_1+2k_3^3k_1+5k_2k_3^2k_1+5k_2^2k_3k_1+k_2^4+k_3^4$
$+2k_2k_3^3+3k_2^2k_3^2+2k_2^3k_3)$.

We have three soliton solutions only when $\beta = 2\gamma$ or $\beta = 3\gamma$. Thus, the KdV7 has two soliton solutions for the parameter values in Eq. (20), but it does not have three soliton solutions for $\gamma = \beta$.

3.2 Second family

Let

$$\begin{aligned} a &= -\frac{5(7\alpha+5\beta-6\gamma)(\alpha+\beta+\gamma)^2}{7938}. \\ b &= \frac{1}{63}(\alpha+\beta+\gamma)(36\alpha+35\beta+10\gamma). \\ c &= -\frac{1}{63}(\alpha+\beta+\gamma)(37\alpha+35\beta+15\gamma). \\ d &= \frac{1}{126}(\alpha+\beta+\gamma)(\alpha+5\beta+30\gamma). \end{aligned} \tag{34}$$

We make the transformation

$$u(x,t) = \frac{126}{\alpha+\beta+\gamma}\partial_{x,x}\log\left(1+\exp\left(kx-k^7t\right)+\rho\exp\left(2kx-2k^7t\right)\right) \tag{35}$$

to obtain the soliton solution

$$u_{\text{soliton}}(x,t) = \frac{504k^2e^{k^7t+kx}}{(\alpha+\beta+\gamma)\left(2e^{k^7t}+e^{kx}\right)^2}\text{for}\rho = \frac{1}{4}. \tag{36}$$

A cnoidal wave solution is

$$u_{\text{cnoidal}}(x,t) = p+\frac{\left(1-2m\pm\sqrt{m^2-m+1}\right)p}{m-1}\text{cn}^2\left(\sqrt{\frac{q(\alpha+\beta+\gamma)}{252m}}(x-\lambda t)+\xi_0,m\right), \tag{37}$$

where

$$\lambda = -\frac{1}{500094m^3}[(\alpha+\beta+\gamma)^2(2205\alpha m^3p^3+1575\beta m^3p^3-1890\gamma m^3p^3-126\alpha m^3p^2q$$
$$-630\beta m^3p^2q-3780\gamma m^3p^2q-2331\alpha m^3pq^2-2205\beta m^3pq^2-2016\gamma m^3pq^2+2\alpha m^3q^3$$
$$+2\beta m^3q^3-124\gamma m^3q^3+63\alpha m^2p^2q+315\beta m^2p^2q+1890\gamma m^2p^2q+2331\alpha m^2pq^2+2205\beta m^2pq^2$$
$$+2016\gamma m^2pq^2-3\alpha m^2q^3-3\beta m^2q^3+186\gamma m^2q^3-126\gamma mpq^2-3\alpha mq^3-3\beta mq^3$$
$$-66\gamma mq^3+2\alpha q^3+2\beta q^3+2\gamma q^3)].$$

3.3 Third family

Let

$$\begin{aligned} a &= \frac{8(\alpha+\beta+\gamma)^2(4\alpha-10\beta+25\gamma)}{453789}.\\ b &= \frac{1}{882}\left(-16\alpha^2+22\alpha\beta-23\alpha\gamma+38\beta^2+31\beta\gamma-7\gamma^2\right).\\ c &= \frac{8\alpha^2+2\alpha\beta+107\alpha\gamma-6\beta^2+93\beta\gamma+99\gamma^2}{1029}.\\ d &= -\frac{2\left(2\alpha^2+4\alpha\beta-31\alpha\gamma+2\beta^2-31\beta\gamma-33\gamma^2\right)}{1029}. \end{aligned} \tag{38}$$

We make the transformation

$$u(x,t) = \frac{441}{2(\alpha+\beta+\gamma)}\partial_{x,x}\log\left(1+\exp\left(kx-k^7t\right)+\rho\exp\left(2kx-2k^7t\right)\right) \tag{39}$$

to obtain the soliton solution

$$u_{\text{soliton}}(x,t) = \frac{3528k^2\left(16e^{k^7t-kx}+e^{kx-k^7t}+4\right)}{(\alpha+\beta+\gamma)\left(16e^{k^7t-kx}+e^{kx-k^7t}+16\right)^2} \text{ for}\rho = \frac{1}{16}. \tag{40}$$

A cnoidal wave solution is

$$u(x,t) = p+\frac{3mp}{1-2m}\text{cn}^2\left(\sqrt{\frac{2p(\alpha+\beta+\gamma)}{147(1-2m)}}(x-\lambda t)+\xi_0,m\right), \tag{41}$$

where

$$\lambda = -\frac{4(m-2)(m+1)p^3(48\alpha-50\beta-99\gamma)(\alpha+\beta+\gamma)^2}{3176523(2m-1)^2}. \tag{42}$$

Let us investigate the existence of two soliton solutions in the ansatz form

$$u(x,t) = \frac{441}{2(\alpha+\beta+\gamma)}\partial_{xx}\log(1+\exp\theta_1+\exp\theta_2+\frac{1}{16}[\exp(2\theta_1)+\exp(2\theta_2)]+$$
$$\rho[\exp(\theta_1+2\theta_2)+\exp(2\theta_1+\theta_2)]+\kappa\exp(\theta_1+\theta_2)+\rho^2\exp(2\theta_1+2\theta_2)),$$
where $\theta_1 = k_1x-k_1^7t$ and $\theta_2 = k_2x-k_2^7t$.

We have three soliton solutions for the following choices:

$$a=\frac{4\gamma^3}{147},\ b=\frac{5\gamma^2}{14},\ c=\frac{9\gamma^2}{7},\ d=\frac{2\gamma^2}{7},\ \alpha=6\gamma,\ \beta=\frac{7\gamma}{2}. \tag{43}$$

$$\kappa=\frac{2k_1^4-k_2^2k_1^2+2k_2^4}{2(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)},\quad \rho=\frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)}{16(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)} \tag{44}$$

Letting $\gamma=42$ gives the Kaup–Kuperschmidt seventh-order equation:

$$u_t+2016u^3u_x+630u_x^3+2268u_xu_{2x}+504u^2\,u_{3x}+252u_{2x}u_{3x}+147u_xu_{4x}+42uu_{5x}+u_{7x}=0. \tag{45}$$

3.4 Three soliton solutions

The three soliton solutions have the form

$$u_{3-\text{soliton}}(x,t)=1/2\partial_{xx}\log\left(1+\sum_{i,j,l=0}^{2}\rho_{i,j,l}\exp\left(i\left(k_1x-k_1^7t\right)+j\left(k_2x-k_2^7t\right)+l\left(k_3x-k_3^7t\right)\right)\right), \tag{46}$$

where $\rho_{i,j,l}=1$ when $i+j+l=1$. The parameter values are obtained as follows. First, we set

$$k_1x-k_1^7t=\log(z_1),k_2x-k_2^7t=\log(z_2)\text{ and }k_3x-k_3^7t=\log(z_3) \tag{47}$$

to get

$$\begin{aligned}&u_t+2016u^3u_x+630u_x^3+2268u_xu_{2x}+504u^2\,u_{3x}\\&+252u_{2x}u_{3x}+147u_xu_{4x}+42u_{5x}+u_{7x}=\Psi(z_1,z_2,z_3).\end{aligned}$$

Next, we solve the equation

$$\frac{\partial^{i+j+l}}{\partial z_1^i\partial z_2^j\partial^l z_3}\Psi(z_1,z_2,z_3)\big|_{z_1=z_2=z_3=0}=0 \tag{48}$$

for $\rho_{i,j,l}$. The parameter values are:

$$\rho_{0,0,2}=\frac{1}{16},\rho_{0,1,1}=\frac{2k_2^4-k_3^2k_2^2+2k_3^4}{2(k_2+k_3)^2\left(k_2^2+k_3k_2+k_3^2\right)},\rho_{0,1,2}=\frac{(k_2-k_3)^2\left(k_2^2-k_3k_2+k_3^2\right)}{16(k_2+k_3)^2\left(k_2^2+k_3k_2+k_3^2\right)}.$$

$$\rho_{0,2,0}=\frac{1}{16},\rho_{0,2,1}=\frac{(k_2-k_3)^2\left(k_2^2-k_3k_2+k_3^2\right)}{16(k_2+k_3)^2\left(k_2^2+k_3k_2+k_3^2\right)},\rho_{1,0,1}=\frac{2k_1^4-k_3^2k_1^2+2k_3^4}{2(k_1+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)}.$$

$$\rho_{1,0,2}=\frac{(k_1-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)}{16(k_1+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)},\rho_{1,1,0}=\frac{2k_1^4-k_2^2k_1^2+2k_2^4}{2(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)},\ \rho_{1,2,0}=\frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)}{16(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)}.$$

$$\rho_{2,0,0}=\frac{1}{16},\rho_{2,0,1}=\frac{(k_1-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)}{16(k_1+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)},\rho_{2,1,0}=\frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)}{16(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)}.$$

$$\rho_{0,2,2}=\frac{(k_2-k_3)^4\left(k_2^2-k_3k_2+k_3^2\right)^2}{256(k_2+k_3)^4\left(k_2^2+k_3k_2+k_3^2\right)^2}.$$

$$\rho_{1,1,1}=\frac{1}{4(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)}$$

$$\times\left[4k_2^4k_1^8+4k_3^4k_1^8-2k_2^2k_3^2k_1^8-2k_2^6k_1^6-2k_3^6k_1^6-k_2^2k_3^4k_1^6-k_2^4k_3^2k_1^6+4k_2^8k_1^4+4k_3^8k_1^4\right.$$
$$\left.-k_2^2k_3^6k_1^4-6k_2^4k_3^4k_1^4-k_2^6k_3^2k_1^4-2k_2^2k_3^8k_1^2-k_2^4k_3^6k_1^2-k_2^6k_3^4k_1^2-2k_2^8k_3^2k_1^2+4k_2^4k_3^8-2k_2^6k_3^6+4k_2^8k_3^4\right].$$

$$\rho_{1,1,2}=\frac{\left(2k_1^4-k_2^2k_1^2+2k_2^4\right)(k_1-k_3)^2(k_2-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)\left(k_2^2-k_3k_2+k_3^2\right)}{32(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)}.$$

$$\rho_{1,2,1}=\frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_2-k_3)^2\left(k_2^2-k_3k_2+k_3^2\right)\left(2k_1^4-k_3^2k_1^2+2k_3^4\right)}{32(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)}.$$

$$\rho_{1,2,2}=\frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_1-k_3)^2(k_2-k_3)^4\left(k_1^2-k_3k_1+k_3^2\right)\left(k_2^2-k_3k_2+k_3^2\right)^2}{256(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^4\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)^2}.$$

$$\rho_{2,0,2}=\frac{(k_1-k_3)^4\left(k_1^2-k_3k_1+k_3^2\right)^2}{256(k_1+k_3)^4\left(k_1^2+k_3k_1+k_3^2\right)^2}.$$

$$\rho_{2,1,1}=\frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_1-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)\left(2k_2^4-k_3^2k_2^2+2k_3^4\right)}{32(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)}.$$

$$\rho_{2,1,2}=\frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_1-k_3)^4(k_2-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)^2\left(k_2^2-k_3k_2+k_3^2\right)}{256(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^4(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)^2\left(k_2^2+k_3k_2+k_3^2\right)}.$$

$$\rho_{2,2,0}=\frac{(k_1-k_2)^4\left(k_1^2-k_2k_1+k_2^2\right)^2}{256(k_1+k_2)^4\left(k_1^2+k_2k_1+k_2^2\right)^2}.$$

$$\rho_{2,2,1}=\frac{(k_1-k_2)^4\left(k_1^2-k_2k_1+k_2^2\right)^2(k_1-k_3)^2(k_2-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)\left(k_2^2-k_3k_2+k_3^2\right)}{256(k_1+k_2)^4\left(k_1^2+k_2k_1+k_2^2\right)^2(k_1+k_3)^2(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)}.$$

$$\rho_{2,2,2}=\frac{(k_1-k_2)^4\left(k_1^2-k_2k_1+k_2^2\right)^2(k_1-k_3)^4(k_2-k_3)^4\left(k_1^2-k_3k_1+k_3^2\right)^2\left(k_2^2-k_3k_2+k_3^2\right)^2}{4096(k_1+k_2)^4\left(k_1^2+k_2k_1+k_2^2\right)^2(k_1+k_3)^4(k_2+k_3)^4\left(k_1^2+k_3k_1+k_3^2\right)^2\left(k_2^2+k_3k_2+k_3^2\right)^2}.$$

3.5 Four soliton solutions

The four soliton solutions have the form $u(x,t)=1/2\partial_{xx}\log f(x,t)$, where

$$\begin{cases} f(x,t) = \sum_{i_1=0}^{2}\sum_{i_2=0}^{2}\sum_{i_3=0}^{2}\sum_{i_4=0}^{2} B_{i_1 i_2 i_3 i_4} z_1^{i_1} z_2^{i_2} z_3^{i_3} z_4^{i_4}, \\ z_j = \exp\left(k_j x - k_j^7 x\right) \text{ for any} j \end{cases} \tag{49}$$

The 76 coefficients $B_{i_1 i_2 i_3 i_4}$ are given by

$$B_{0002} = \frac{1}{16}, B_{0011} = \frac{2k_3^4 - k_4^2 k_3^2 + 2k_4^4}{2(k_3+k_4)^2\left(k_3^2 + k_4 k_3 + k_4^2\right)}.$$

$$B_{0012} = \frac{(k_3-k_4)^2\left(k_3^2 - k_4 k_3 + k_4^2\right)}{16(k_3+k_4)^2\left(k_3^2 + k_4 k_3 + k_4^2\right)}, B_{0020} = \frac{1}{16}.$$

$$B_{0021} = \frac{(k_3-k_4)^2\left(k_3^2 - k_4 k_3 + k_4^2\right)}{16(k_3+k_4)^2\left(k_3^2 + k_4 k_3 + k_4^2\right)}.$$

$$B_{0022} = \frac{(k_3-k_4)^4\left(k_3^2 - k_4 k_3 + k_4^2\right)^2}{256(k_3+k_4)^4\left(k_3^2 + k_4 k_3 + k_4^2\right)^2}.$$

$$B_{0101} = \frac{2k_2^4 - k_4^2 k_2^2 + 2k_4^4}{2(k_2+k_4)^2\left(k_2^2 + k_4 k_2 + k_4^2\right)}.$$

$$B_{0102} = \frac{(k_2-k_4)^2\left(k_2^2 - k_4 k_2 + k_4^2\right)}{16(k_2+k_4)^2\left(k_2^2 + k_4 k_2 + k_4^2\right)}.$$

$$B_{0110} = \frac{2k_2^4 - k_3^2 k_2^2 + 2k_3^4}{2(k_2+k_3)^2\left(k_2^2 + k_3 k_2 + k_3^2\right)}.$$

$$\begin{aligned} B_{0111} = [&4k_3^4k_2^8 + 4k_4^4k_2^8 - 2k_3^2k_4^2k_2^8 - 2k_3^6k_2^6 - 2k_4^6k_2^6 - k_3^2k_4^4k_2^6 \\ &-k_3^4k_4^2k_2^6 + 4k_3^8k_2^4 + 4k_4^8k_2^4 - k_3^2k_4^6k_2^4 - 6k_3^4k_4^4k_2^4 \\ &-k_3^6k_4^2k_2^4 - 2k_3^2k_4^8k_2^2 - k_3^4k_4^6k_2^2 - k_3^6k_4^4k_2^2 - 2k_3^8k_4^2k_2^2 \\ &+4k_3^4k_4^8 - 2k_3^6k_4^6 + 4k_3^8k_4^4]/[4(k_2+k_3)^2\left(k_2^2+k_3k_2+k_3^2\right)(k_2+k_4)^2 \\ &(k_3+k_4)^2\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)]. \end{aligned}$$

$$B_{0112} = \frac{\left(2k_2^4 - k_3^2k_2^2 + 2k_3^4\right)(k_2-k_4)^2(k_3-k_4)^2\left(k_2^2-k_4k_2+k_4^2\right)\left(k_3^2-k_4k_3+k_4^2\right)}{32(k_2+k_3)^2\left(k_2^2+k_3k_2+k_3^2\right)(k_2+k_4)^2(k_3+k_4)^2\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)}$$

$$B_{0120} = \frac{(k_2-k_3)^2\left(k_2^2-k_3k_2+k_3^2\right)}{16(k_2+k_3)^2\left(k_2^2+k_3k_2+k_3^2\right)}$$

$$B_{0121} = \frac{(k_2-k_3)^2\left(k_2^2-k_3k_2+k_3^2\right)(k_3-k_4)^2\left(k_3^2-k_4k_3+k_4^2\right)\left(2k_2^4-k_4^2k_2^2+2k_4^4\right)}{32(k_2+k_3)^2\left(k_2^2+k_3k_2+k_3^2\right)(k_2+k_4)^2(k_3+k_4)^2\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)}$$

$$B_{0122}=\frac{(k_2-k_3)^2\left(k_2^2-k_3k_2+k_3^2\right)(k_2-k_4)^2(k_3-k_4)^4\left(k_2^2-k_4k_2+k_4^2\right)\left(k_3^2-k_4k_3+k_4^2\right)^2}{256(k_2+k_3)^2\left(k_2^2+k_3k_2+k_3^2\right)(k_2+k_4)^2(k_3+k_4)^4\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)^2}$$

$$B_{0200}=\frac{1}{16},\ B_{0201}=\frac{(k_2-k_4)^2\left(k_2^2-k_4k_2+k_4^2\right)}{16(k_2+k_4)^2\left(k_2^2+k_4k_2+k_4^2\right)}$$

$$B_{0202}=\frac{(k_2-k_4)^4\left(k_2^2-k_4k_2+k_4^2\right)^2}{256(k_2+k_4)^4\left(k_2^2+k_4k_2+k_4^2\right)^2},B_{0210}=\frac{(k_2-k_3)^2\left(k_2^2-k_3k_2+k_3^2\right)}{16(k_2+k_3)^2\left(k_2^2+k_3k_2+k_3^2\right)}.$$

$$B_{0211}=\frac{(k_2-k_3)^2\left(k_2^2-k_3k_2+k_3^2\right)(k_2-k_4)^2\left(k_2^2-k_4k_2+k_4^2\right)\left(2k_3^4-k_4^2k_3^2+2k_4^4\right)}{32(k_2+k_3)^2\left(k_2^2+k_3k_2+k_3^2\right)(k_2+k_4)^2(k_3+k_4)^2\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)}.$$

$$B_{0212}=\frac{(k_2-k_3)^2\left(k_2^2-k_3k_2+k_3^2\right)(k_2-k_4)^4(k_3-k_4)^2\left(k_2^2-k_4k_2+k_4^2\right)^2\left(k_3^2-k_4k_3+k_4^2\right)}{256(k_2+k_3)^2\left(k_2^2+k_3k_2+k_3^2\right)(k_2+k_4)^4(k_3+k_4)^2\left(k_2^2+k_4k_2+k_4^2\right)^2\left(k_3^2+k_4k_3+k_4^2\right)}.$$

$$B_{0220}=\frac{(k_2-k_3)^4\left(k_2^2-k_3k_2+k_3^2\right)^2}{256(k_2+k_3)^4\left(k_2^2+k_3k_2+k_3^2\right)^2}-$$

$$B_{0221}=\frac{(k_2-k_3)^4\left(k_2^2-k_3k_2+k_3^2\right)^2(k_2-k_4)^2(k_3-k_4)^2\left(k_2^2-k_4k_2+k_4^2\right)\left(k_3^2-k_4k_3+k_4^2\right)}{256(k_2+k_3)^4\left(k_2^2+k_3k_2+k_3^2\right)^2(k_2+k_4)^2(k_3+k_4)^2\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)}.$$

$$B_{0222}=\frac{(k_2-k_3)^4\left(k_2^2-k_3k_2+k_3^2\right)^2(k_2-k_4)^4(k_3-k_4)^4\left(k_2^2-k_4k_2+k_4^2\right)^2\left(k_3^2-k_4k_3+k_4^2\right)^2}{4096(k_2+k_3)^4\left(k_2^2+k_3k_2+k_3^2\right)^2(k_2+k_4)^4(k_3+k_4)^4\left(k_2^2+k_4k_2+k_4^2\right)^2\left(k_3^2+k_4k_3+k_4^2\right)^2}.$$

$$B_{1001}=\frac{2k_1^4-k_4^2k_1^2+2k_4^4}{2(k_1+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)},B_{1002}=\frac{(k_1-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)}{16(k_1+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)}.$$

$$B_{1010}=\frac{2k_1^4-k_3^2k_1^2+2k_3^4}{2(k_1+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)}$$

$$\begin{aligned}B_{1011}=&[4k_3^4k_1^8+4k_4^4k_1^8-2k_3^2k_4^2k_1^8-2k_3^6k_1^6-2k_4^6k_1^6-k_3^2k_4^4k_1^6\\&-k_3^4k_4^2k_1^6+4k_3^8k_1^4+4k_4^8k_1^4-k_3^2k_4^6k_1^4-6k_3^4k_4^4k_1^4-k_3^6k_4^2k_1^4-2k_3^2k_4^8k_1^2\\&-k_3^4k_4^6k_1^2-k_3^6k_4^4k_1^2-2k_3^8k_4^2k_1^2+4k_3^4k_4^8-2k_3^6k_4^6+4k_3^8k_4^4]\\&/[4(k_1+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)(k_1+k_4)^2(k_3+k_4)^2\\&\left(k_1^2+k_4k_1+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)].\end{aligned}$$

$$B_{1012}=\frac{\left(2k_1^4-k_3^2k_1^2+2k_3^4\right)(k_1-k_4)^2(k_3-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)\left(k_3^2-k_4k_3+k_4^2\right)}{32(k_1+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)(k_1+k_4)^2(k_3+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)}$$

$$B_{1020}=\frac{(k_1-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)}{16(k_1+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)}$$

$$B_{1021}=\frac{(k_1-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)(k_3-k_4)^2\left(k_3^2-k_4k_3+k_4^2\right)\left(2k_1^4-k_4^2k_1^2+2k_4^4\right)}{32(k_1+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)(k_1+k_4)^2(k_3+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)}$$

$$B_{1022} = \frac{(k_1-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)(k_1-k_4)^2(k_3-k_4)^4\left(k_1^2-k_4k_1+k_4^2\right)\left(k_3^2-k_4k_3+k_4^2\right)^2}{256(k_1+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)(k_1+k_4)^2(k_3+k_4)^4\left(k_1^2+k_4k_1+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)^2}$$

$$B_{1100} = \frac{2k_1^4-k_2^2k_1^2+2k_2^4}{2(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)}$$

$$B_{1101} = [4k_2^4k_1^8+4k_4^4k_1^8-2k_2^2k_4^2k_1^8-2k_2^6k_1^6-2k_4^6k_1^6-k_2^2k_4^4k_1^6-k_2^4k_4^2k_1^6$$
$$+4k_2^8k_1^4+4k_4^8k_1^4-k_2^2k_4^6k_1^4-6k_2^4k_4^4k_1^4-k_2^6k_4^2k_1^4-2k_2^2k_4^8k_1^2-k_2^4k_4^6k_1^2-$$
$$k_2^6k_4^4k_1^2-2k_2^8k_4^2k_1^2+4k_2^4k_4^8-2k_2^6k_4^6+4k_2^8k_4^4]/[4(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)$$
$$(k_1+k_4)^2(k_2+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)\left(k_2^2+k_4k_2+k_4^2\right)].$$

$$B_{1102} = \frac{\left(2k_1^4-k_2^2k_1^2+2k_2^4\right)(k_1-k_4)^2(k_2-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)\left(k_2^2-k_4k_2+k_4^2\right)}{32(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_4)^2(k_2+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)\left(k_2^2+k_4k_2+k_4^2\right)}$$

$$B_{1120} = \frac{\left(2k_1^4-k_2^2k_1^2+2k_2^4\right)(k_1-k_3)^2(k_2-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)\left(k_2^2-k_3k_2+k_3^2\right)}{32(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)}$$

$$B_{1121} = (k_1-k_3)^2(k_2-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)\left(k_2^2-k_3k_2+k_3^2\right)(k_3-k_4)^2\left(k_3^2-k_4k_3+k_4^2\right)$$
$$(4k_4^4k_1^8+4k_4^4k_1^8-2k_2^2k_4^4k_1^8-2k_2^6k_1^6-2k_4^6k_1^6-k_2^2k_4^4k_1^6-k_2^4k_4^2k_1^6+4k_2^8k_1^4+4k_4^8k_1^4+$$
$$-k_2^2k_4^6k_1^4-6k_2^4k_4^4k_1^4-k_2^6k_4^2k_1^4-2k_2^2k_4^8k_1^2-k_2^4k_4^6k_1^2-k_2^6k_4^4k_1^2-2k_2^8k_4^2k_1^2+4k_2^4k_4^8-2k_2^6k_4^6+4k_2^8k_4^4)$$
$$/[64(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)(k_1+k_4)^2$$
$$(k_2+k_4)^2(k_3+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)].$$

$$B_{1122} = \left(2k_1^4-k_2^2k_1^2+2k_2^4\right)(k_1-k_3)^2(k_2-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)\left(k_2^2-k_3k_2+k_3^2\right)(k_1-k_4)^2$$
$$(k_2-k_4)^2(k_3-k_4)^4\left(k_1^2-k_4k_1+k_4^2\right)\left(k_2^2-k_4k_2+k_4^2\right)\left(k_3^2-k_4k_3+k_4^2\right)^2$$
$$/[512(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)$$
$$(k_1+k_4)^2(k_2+k_4)^2(k_3+k_4)^4\left(k_1^2+k_4k_1+k_4^2\right)\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)^2].$$

$$B_{1200} = \frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)}{16(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)}$$

$$B_{1201} = \frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_2-k_4)^2\left(k_2^2-k_4k_2+k_4^2\right)\left(2k_1^4-k_4^2k_1^2+2k_4^4\right)}{32(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_4)^2(k_2+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)\left(k_2^2+k_4k_2+k_4^2\right)}$$

$$B_{1202} = \frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_1-k_4)^2(k_2-k_4)^4\left(k_1^2-k_4k_1+k_4^2\right)\left(k_2^2-k_4k_2+k_4^2\right)^2}{256(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_4)^2(k_2+k_4)^4\left(k_1^2+k_4k_1+k_4^2\right)\left(k_2^2+k_4k_2+k_4^2\right)^2}$$

$$B_{1210} = \frac{(k_{1=}-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_2-k_3)^2\left(k_2^2-k_3k_2+k_3^2\right)\left(2k_1^4-k_3^2k_1^2+2k_3^4\right)}{32(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)}$$

$$B_{1211} = (k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_2-k_3)^2\left(k_2^2-k_3k_2+k_3^2\right)(k_2-k_4)^2\left(k_2^2-k_4k_2+k_4^2\right)$$
$$\left(4k_3^4k_1^8+4k_4^4k_1^8-2k_3^2k_4^2k_1^8-2k_3^6k_1^6-2k_4^6k_1^6-k_3^2k_4^4k_1^6-k_3^4k_4^2k_1^6+4k_3^8k_1^4+4k_4^8k_1^4-k_3^2k_4^6k_1^4\right.$$
$$\left.-6k_3^4k_4^4k_1^4-k_3^6k_4^2k_1^4-2k_3^2k_4^8k_1^2-k_3^4k_4^6k_1^2-k_3^6k_4^4k_1^2-2k_3^8k_4^2k_1^2+4k_3^4k_4^8-2k_3^6k_4^6+4k_3^8k_4^4\right)$$
$$/[64(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)$$
$$(k_1+k_4)^2(k_2+k_4)^2(k_3+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)].$$

$$B_{1212} = [(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_2-k_3)^2\left(k_2^2-k_3k_2+k_3^2\right)\left(2k_1^4-k_3^2k_1^2+2k_3^4\right)$$
$$(k_1-k_4)^2(k_2-k_4)^4(k_3-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)\left(k_2^2-k_4k_2+k_4^2\right)^2\left(k_3^2-k_4k_3+k_4^2\right)]$$
$$/[512(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)$$
$$(k_1+k_4)^2(k_2+k_4)^4(k_3+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)\left(k_2^2+k_4k_2+k_4^2\right)^2\left(k_3^2+k_4k_3+k_4^2\right)].$$

$$B_{1220} = \frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_1-k_3)^2(k_2-k_3)^4\left(k_1^2-k_3k_1+k_3^2\right)\left(k_2^2-k_3k_2+k_3^2\right)^2}{256(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^4\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)^2}.$$

$$B_{1221} = [(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_1-k_3)^2(k_2-k_3)^4\left(k_1^2-k_3k_1+k_3^2\right)\left(k_2^2-k_3k_2+k_3^2\right)^2$$
$$(k_2-k_4)^2(k_3-k_4)^2\left(k_2^2-k_4k_2+k_4^2\right)\left(k_3^2-k_4k_3+k_4^2\right)\left(2k_1^4-k_4^2k_1^2+2k_4^4\right)]$$
$$/[512(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^4\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)^2$$
$$(k_1+k_4)^2(k_2+k_4)^2(k_3+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)].$$

$$B_{1222} = (k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_1-k_3)^2(k_2-k_3)^4\left(k_1^2-k_3k_1+k_3^2\right)\left(k_2^2-k_3k_2+k_3^2\right)^2$$
$$(k_1-k_4)^2(k_2-k_4)^4(k_3-k_4)^4\left(k_1^2-k_4k_1+k_4^2\right)\left(k_2^2-k_4k_2+k_4^2\right)^2\left(k_3^2-k_4k_3+k_4^2\right)^2$$
$$/[4096(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^4\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)^2$$
$$(k_1+k_4)^2(k_2+k_4)^4(k_3+k_4)^4\left(k_1^2+k_4k_1+k_4^2\right)\left(k_2^2+k_4k_2+k_4^2\right)^2\left(k_3^2+k_4k_3+k_4^2\right)^2].$$

$$B_{2000} = \frac{1}{16}, B_{2001} = \frac{(k_1-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)}{16(k_1+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)}, B_{2002} = \frac{(k_1-k_4)^4\left(k_1^2-k_4k_1+k_4^2\right)^2}{256(k_1+k_4)^4\left(k_1^2+k_4k_1+k_4^2\right)^2}.$$

$$B_{2010} = \frac{(k_1-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)}{16(k_1+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)}.$$

$$B_{2011} = \frac{(k_1-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)(k_1-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)\left(2k_3^4-k_4^2k_3^2+2k_4^4\right)}{32(k_1+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)(k_1+k_4)^2(k_3+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)}.$$

$$B_{2012} = \frac{(k_1-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)(k_1-k_4)^4(k_3-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)^2\left(k_3^2-k_4k_3+k_4^2\right)}{256(k_1+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)(k_1+k_4)^4(k_3+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)^2\left(k_3^2+k_4k_3+k_4^2\right)}.$$

$$B_{2020} = \frac{(k_1-k_3)^4\left(k_1^2-k_3k_1+k_3^2\right)^2}{256(k_1+k_3)^4\left(k_1^2+k_3k_1+k_3^2\right)^2}.$$

$$B_{2021} = \frac{(k_1-k_3)^4\left(k_1^2-k_3k_1+k_3^2\right)^2(k_1-k_4)^2(k_3-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)\left(k_3^2-k_4k_3+k_4^2\right)}{256(k_1+k_3)^4\left(k_1^2+k_3k_1+k_3^2\right)^2(k_1+k_4)^2(k_3+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)}.$$

$$B_{2022} = \frac{(k_1-k_3)^4\left(k_1^2-k_3k_1+k_3^2\right)^2(k_1-k_4)^4(k_3-k_4)^4\left(k_1^2-k_4k_1+k_4^2\right)^2\left(k_3^2-k_4k_3+k_4^2\right)^2}{4096(k_1+k_3)^4\left(k_1^2+k_3k_1+k_3^2\right)^2(k_1+k_4)^4(k_3+k_4)^4\left(k_1^2+k_4k_1+k_4^2\right)^2\left(k_3^2+k_4k_3+k_4^2\right)^2}.$$

$$B_{2100} = \frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)}{16(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)}.$$

$$B_{2101} = \frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_1-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)\left(2k_2^4-k_4^2k_2^2+2k_4^4\right)}{32(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_4)^2(k_2+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)\left(k_2^2+k_4k_2+k_4^2\right)}.$$

$$B_{2102} = \frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_1-k_4)^4(k_2-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)^2\left(k_2^2-k_4k_2+k_4^2\right)}{256(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_4)^4(k_2+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)^2\left(k_2^2+k_4k_2+k_4^2\right)}.$$

$$B_{2110} = \frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_1-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)\left(2k_2^4-k_3^2k_2^2+2k_3^4\right)}{32(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)}.$$

$$B_{2111} = [(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_1-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)(k_1-k_4)^2$$
$$\left(k_1^2-k_4k_1+k_4^2\right)(4k_3^4k_2^8+4k_4^4k_2^8-2k_3^2k_4^2k_2^8-2k_3^6k_2^6-2k_4^6k_2^6-k_3^2k_4^4k_2^6$$
$$-k_3^4k_4^2k_2^6+4k_3^8k_2^4+4k_4^8k_2^4-k_3^2k_4^6k_2^4-6k_3^4k_4^4k_2^4-k_3^6k_4^2k_2^4-2k_3^2k_4^8k_2^2$$
$$-k_3^4k_4^6k_2^2-k_3^6k_4^4k_2^2-2k_3^8k_4^2k_2^2+4k_3^4k_4^8-2k_3^6k_4^6+4k_3^8k_4^4)]$$
$$/[64(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)$$
$$\left(k_2^2+k_3k_2+k_3^2\right)(k_1+k_4)^2(k_2+k_4)^2(k_3+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)$$
$$\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)].$$

$$B_{2112} = [(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_1-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)$$
$$\left(2k_2^4-k_3^2k_2^2+2k_3^4\right)(k_1-k_4)^4(k_2-k_4)^2(k_3-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)^2$$
$$\left(k_2^2-k_4k_2+k_4^2\right)\left(k_3^2-k_4k_3+k_4^2\right)]/[512(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^2$$
$$(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)(k_1+k_4)^4(k_2+k_4)^2(k_3+k_4)^2$$
$$\left(k_1^2+k_4k_1+k_4^2\right)^2\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)].$$

$$B_{2120} = \frac{(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_1-k_3)^4(k_2-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)^2\left(k_2^2-k_3k_2+k_3^2\right)}{256(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^4(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)^2\left(k_2^2+k_3k_2+k_3^2\right)}.$$

$$B_{2121} = [(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_1-k_3)^4(k_2-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)^2$$
$$\left(k_2^2-k_3k_2+k_3^2\right)(k_1-k_4)^2(k_3-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)\left(k_3^2-k_4k_3+k_4^2\right)$$
$$\left(2k_2^4-k_4^2k_2^2+2k_4^4\right)]/[512(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^4(k_2+k_3)^2$$
$$\left(k_1^2+k_3k_1+k_3^2\right)^2\left(k_2^2+k_3k_2+k_3^2\right)(k_1+k_4)^2(k_2+k_4)^2(k_3+k_4)^2$$
$$\left(k_1^2+k_4k_1+k_4^2\right)\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)].$$

$$B_{2122} = [(k_1-k_2)^2\left(k_1^2-k_2k_1+k_2^2\right)(k_1-k_3)^4(k_2-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)^2$$
$$\left(k_2^2-k_3k_2+k_3^2\right)(k_1-k_4)^4(k_2-k_4)^2(k_3-k_4)^4\left(k_1^2-k_4k_1+k_4^2\right)^2$$
$$\left(k_2^2-k_4k_2+k_4^2\right)\left(k_3^2-k_4k_3+k_4^2\right)^2]/[4096(k_1+k_2)^2\left(k_1^2+k_2k_1+k_2^2\right)(k_1+k_3)^4$$
$$(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)^2\left(k_2^2+k_3k_2+k_3^2\right)(k_1+k_4)^4(k_2+k_4)^2(k_3+k_4)^4$$
$$\left(k_1^2+k_4k_1+k_4^2\right)^2\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)^2].$$

$$B_{2200} = \frac{(k_1-k_2)^4\left(k_1^2-k_2k_1+k_2^2\right)^2}{256(k_1+k_2)^4\left(k_1^2+k_2k_1+k_2^2\right)^2}.$$

$$B_{2201} = \frac{(k_1-k_2)^4\left(k_1^2-k_2k_1+k_2^2\right)^2(k_1-k_4)^2(k_2-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)\left(k_2^2-k_4k_2+k_4^2\right)}{256(k_1+k_2)^4\left(k_1^2+k_2k_1+k_2^2\right)^2(k_1+k_4)^2(k_2+k_4)^2\left(k_1^2+k_4k_1+k_4^2\right)\left(k_2^2+k_4k_2+k_4^2\right)}$$

$$B_{2202} = \frac{(k_1-k_2)^4\left(k_1^2-k_2k_1+k_2^2\right)^2(k_1-k_4)^4(k_2-k_4)^4\left(k_1^2-k_4k_1+k_4^2\right)^2\left(k_2^2-k_4k_2+k_4^2\right)^2}{4096(k_1+k_2)^4\left(k_1^2+k_2k_1+k_2^2\right)^2(k_1+k_4)^4(k_2+k_4)^4\left(k_1^2+k_4k_1+k_4^2\right)^2\left(k_2^2+k_4k_2+k_4^2\right)^2}$$

$$B_{2210} = \frac{(k_1-k_2)^4\left(k_1^2-k_2k_1+k_2^2\right)^2(k_1-k_3)^2(k_2-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)\left(k_2^2-k_3k_2+k_3^2\right)}{256(k_1+k_2)^4\left(k_1^2+k_2k_1+k_2^2\right)^2(k_1+k_3)^2(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)}$$

$$B_{2211} = [(k_1-k_2)^4\left(k_1^2-k_2k_1+k_2^2\right)^2(k_1-k_3)^2(k_2-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)$$
$$\left(k_2^2-k_3k_2+k_3^2\right)(k_1-k_4)^2(k_2-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)\left(k_2^2-k_4k_2+k_4^2\right)$$
$$\left(2k_3^4-k_4^2k_3^2+2k_4^4\right)]/[512(k_1+k_2)^4\left(k_1^2+k_2k_1+k_2^2\right)^2(k_1+k_3)^2(k_2+k_3)^2$$
$$\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)(k_1+k_4)^2(k_2+k_4)^2(k_3+k_4)^2$$
$$\left(k_1^2+k_4k_1+k_4^2\right)\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)].$$

$$B_{2212} = [(k_1-k_2)^4\left(k_1^2-k_2k_1+k_2^2\right)^2(k_1-k_3)^2(k_2-k_3)^2\left(k_1^2-k_3k_1+k_3^2\right)$$
$$\left(k_2^2-k_3k_2+k_3^2\right)(k_1-k_4)^4(k_2-k_4)^4(k_3-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)^2$$
$$\left(k_2^2-k_4k_2+k_4^2\right)^2\left(k_3^2-k_4k_3+k_4^2\right)]/[4096(k_1+k_2)^4\left(k_1^2+k_2k_1+k_2^2\right)^2(k_1+k_3)^2$$
$$(k_2+k_3)^2\left(k_1^2+k_3k_1+k_3^2\right)\left(k_2^2+k_3k_2+k_3^2\right)(k_1+k_4)^4(k_2+k_4)^4(k_3+k_4)^2$$
$$\left(k_1^2+k_4k_1+k_4^2\right)^2\left(k_2^2+k_4k_2+k_4^2\right)^2\left(k_3^2+k_4k_3+k_4^2\right)].$$

$$B_{2220} = \frac{(k_1-k_2)^4\left(k_1^2-k_2k_1+k_2^2\right)^2(k_1-k_3)^4(k_2-k_3)^4\left(k_1^2-k_3k_1+k_3^2\right)^2\left(k_2^2-k_3k_2+k_3^2\right)^2}{4096(k_1+k_2)^4\left(k_1^2+k_2k_1+k_2^2\right)^2(k_1+k_3)^4(k_2+k_3)^4\left(k_1^2+k_3k_1+k_3^2\right)^2\left(k_2^2+k_3k_2+k_3^2\right)^2}.$$

$$B_{2221} = [(k_1-k_2)^4\left(k_1^2-k_2k_1+k_2^2\right)^2(k_1-k_3)^4(k_2-k_3)^4\left(k_1^2-k_3k_1+k_3^2\right)^2$$
$$\left(k_2^2-k_3k_2+k_3^2\right)^2(k_1-k_4)^2(k_2-k_4)^2(k_3-k_4)^2\left(k_1^2-k_4k_1+k_4^2\right)$$
$$\left(k_2^2-k_4k_2+k_4^2\right)\left(k_3^2-k_4k_3+k_4^2\right)]/[4096(k_1+k_2)^4\left(k_1^2+k_2k_1+k_2^2\right)^2(k_1+k_3)^4$$
$$(k_2+k_3)^4\left(k_1^2+k_3k_1+k_3^2\right)^2\left(k_2^2+k_3k_2+k_3^2\right)^2(k_1+k_4)^2(k_2+k_4)^2(k_3+k_4)^2$$
$$\left(k_1^2+k_4k_1+k_4^2\right)\left(k_2^2+k_4k_2+k_4^2\right)\left(k_3^2+k_4k_3+k_4^2\right)].$$

$$B_{2222} = [(k_1-k_2)^4\left(k_1^2-k_2k_1+k_2^2\right)^2(k_1-k_3)^4(k_2-k_3)^4\left(k_1^2-k_3k_1+k_3^2\right)^2$$
$$\left(k_2^2-k_3k_2+k_3^2\right)^2(k_1-k_4)^4(k_2-k_4)^4(k_3-k_4)^4\left(k_1^2-k_4k_1+k_4^2\right)^2$$
$$\left(k_2^2-k_4k_2+k_4^2\right)^2\left(k_3^2-k_4k_3+k_4^2\right)^2]/[65536(k_1+k_2)^4\left(k_1^2+k_2k_1+k_2^2\right)^2(k_1+k_3)^4$$
$$(k_2+k_3)^4\left(k_1^2+k_3k_1+k_3^2\right)^2\left(k_2^2+k_3k_2+k_3^2\right)^2(k_1+k_4)^4(k_2+k_4)^4(k_3+k_4)^4$$
$$\left(k_1^2+k_4k_1+k_4^2\right)^2\left(k_2^2+k_4k_2+k_4^2\right)^2\left(k_3^2+k_4k_3+k_4^2\right)^2].$$

For a graphical illustration, see **Figure 4**.

3.6 *N*-soliton solutions

The N-soliton solutions have the form $u(x,t) = 1/2\partial_{xx}\log f(x,t)$, where

$$\begin{cases} f(x,t) = \sum_{i_1,i_2,\cdots,i_N=0}^{2} B_{i_1i_2\cdots i_N} \prod_{j=1}^{N} z_j^{i_j}, \\ z_j = \exp\left(k_j x - k_j^7 x\right) \text{ for any} j \end{cases} \tag{50}$$

In order to find the unknown coefficients $B_{i_1i_2\cdots i_N}$, we define

$$\Psi(z_1,z_2,\cdots,z_N) = u_t + 2016u^3u_x + 630u_x^3 + 2268u_xu_{2x} + 504u^2\,u_{3x}$$
$$+252u_{2x}u_{3x} + 147u_xu_{4x} + 42uu_{5x} + u_{7x}.$$

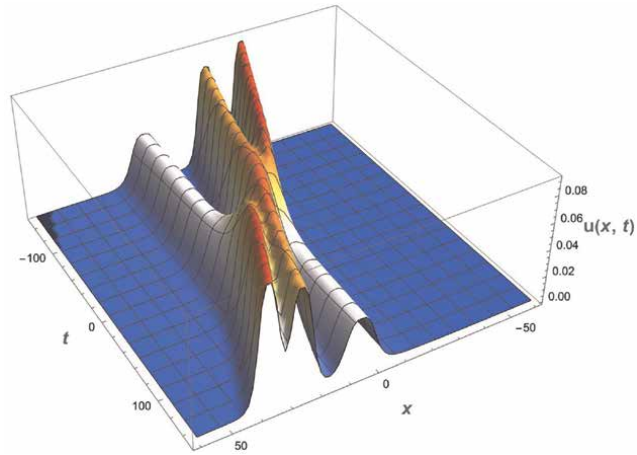


Figure 4.
Four soliton solution: $\{k_1 = 0.489591, k_2 = 0.68479, k_3 = 0.104676, k_4 = 0.733527\}$.

The number $B_{i_1 i_2 \cdots i_N}$ is obtained from the equation

$$\frac{\partial^{i_1+i_2+\cdots+i_N}}{\partial z_1^{i_1} \partial z_2^{i_2} \cdots \partial z_N^{i_n}} \Psi(0,0,0,\ldots,0) = 0 \tag{51}$$

4. Bilinearization

Let us consider the case when $d = c/2 - b$. The family to be considered is

$$u_t + au^3 u_x + bu_x^3 + cuu_x u_{2x} + (c/2 - b)u^2\, u_{3x} + \alpha u_{2x} u_{3x} + \beta u_x u_{4x} + \gamma uu_{5x} + u_{7x} = 0. \tag{52}$$

Let

$$u(x,t) = A\partial_{xx} f(x,t). \tag{53}$$

Plugging the ansatz in Eq. (53) into Eq. (52) and integrating once with respect to x, taking a zero integration constant gives

$$\begin{aligned}
&3f_x(8f^4(4Af_{xxx}^2(Ab + 5\beta - 5\gamma)D^2{}_x(f\cdot f) + 2f(A(\gamma-\beta)f_{xxxxxx}D^2{}_x(f\cdot f) + 2f(D^8{}_x(f\cdot f) + \\
&D_{xt}(f\cdot f))) + (A(\alpha-\beta+\gamma) - 140)D^4{}_x(f\cdot f)^2 + (A(\beta+\gamma) - 112)D^2{}_x(f\cdot f)D^6{}_x(f\cdot f)) \\
&+D^2{}_x(f\cdot f)^4\left(aA^3 + 6A(12\alpha + 4Ab - 2Ac + 3\beta + 77\gamma) - 20160\right) \\
&+4f^2D^2{}_x(f\cdot f)^2D^4{}_x(f\cdot f)\left(A^2(c-2b) - 3A(4\alpha+\beta+19\gamma) + 3360\right)) \\
&-48Af^2f_x^4f_{xxx}(4Ab + 15(\gamma-\beta))D^2{}_x(f\cdot f) - 24Af^2f_x^2f_{xxx}(3(4Ab - 15\beta + 15\gamma)D^2{}_x(f\cdot f)^2 \\
&+10f^2(\beta-\gamma)D^4{}_x(f\cdot f)) + 8Af_x^3(2f^2(\beta-\gamma)(f^2\left(20f_{xxx}^2 - 2ff_{xxxxxx} + D^6{}_x(f\cdot f)\right) \\
&+15D^2{}_x(f\cdot f)D^4{}_x(f\cdot f)) + 9(3Ab + 20(\gamma-\beta))D^2{}_x(f\cdot f)^3) + 4Af^2(\beta-\gamma)f_{xxx}(-4f^4\left(20f_{xxx}^2 + D^6{}_x(f\cdot f)\right) \\
&+8f^5f_{xxxxxx} - 45D^2{}_x(f\cdot f)^3 + 30f^2D^2{}_x(f\cdot f)D^4{}_x(f\cdot f)) + 288Af^2(\beta-\gamma)f_x^6f_{xxx} \\
&+96Af_x^7(Ab + 12(\gamma-\beta))D^2{}_x(f\cdot f) + 48Af_x^5(3(2Ab - 15\beta + 15\gamma)D^2{}_x(f\cdot f)^2 \\
&+5f^2(\beta-\gamma)D^4{}_x(f\cdot f)) + 288A(\gamma-\beta)f_x^9 = 0.
\end{aligned}$$

The choices

$$\left\{\alpha=\frac{5\gamma}{2},\beta=\gamma,a=\frac{15\gamma^3}{784},b=0,c=\frac{15\gamma^2}{28}\right\} \tag{54}$$

will give the following bilinear form

$$D_{xt}^1(f\cdot f)+D_x^8(f\cdot f)=0 \tag{55}$$

This corresponds to the KdV7 $(A=2,\gamma=28)$

$$u_t+420u^3u_x+420u_xu_{2x}+210u^2\,u_{3x}+70u_{2x}u_{3x}+28u_xu_{4x}+28uu_{5x}+u_{7x}=0. \tag{56}$$

This KdV7 admits one and two soliton solutions. However, it does not have three solitons solutions despite the fact that it admits bilinear form.

One soliton solution: $u(x,t)=2\partial_{xx}\log\left(1+\exp\left(k_1x-k_1^7t\right)\right)$..

Two soliton solution:

$u(x,t)=2\partial_{xx}\log\left(1+\exp\left(k_1x-k_1^7t\right)+\exp\left(k_2x-k_2^7t\right)+A_{12}\exp\left(k_1x-k_1^7t\right)\exp\left(k_2x-k_2^7t\right)\right)$,

where

$$A_{1,2}=\frac{\left(k_2-k_1\right)^2\left(k_1^2-k_2k_1+k_2^2\right)^2}{\left(k_1+k_2\right)^2\left(k_1^2+k_2k_1+k_2^2\right)^2}. \tag{57}$$

Breather: $u(x,t)=2\partial_{xx}\log\left(pe^{kx-\lambda t}+qe^{\lambda t-kx}+r\sin(\kappa x-\mu t)\right)$, where

$$\begin{aligned}\lambda&=k\left(-7\kappa^6+k^6-21\kappa^2k^4+35\kappa^4k^2\right).\\ \mu&=\kappa\left(-\kappa^6+7k^6-35\kappa^2k^4+21\kappa^4k^2\right).\\ p&=-\frac{\kappa^2r^2\left(3\kappa^2-k^2\right)^2}{4k^2q\left(\kappa^2-3k^2\right)^2}.\end{aligned} \tag{58}$$

Let us consider a more general than Eq. (56) KdV7

$$\begin{aligned}&u_t+420u^3u_x+420uu_xu_{xx}+210u^2u_{xxx}+70u_{xx}u_{xxx}+28u_xu_{xxxx}\\&+28uu_{xxxxx}+u_{7x}+6puu_x+45qu^2u_x+15qu_xu_{xx}+pu_{xxx}+15quu_{xxx}+quu_{5x}+u_{7x}=0.\end{aligned}$$

This KdV7 admits the bilinear form

$$D_{xt}^1(f\cdot f)+pD_x^4(f\cdot f)+qD_x^6(f\cdot f)+D_x^8(f\cdot f)=0. \tag{59}$$

The one soliton solutions are

$$u(x,t)=\frac{k^2e^{k^3t\left(k^4+k^2q+p\right)+kx}}{\left(e^{k^3t\left(k^4+k^2q+p\right)}+e^{kx}\right)^2}.$$

The two soliton solutions are

$$u(x,t) = \partial_{xx} \log(1 + \exp(k_1 x - w_1 t) + \exp(k_2 x - w_2 t) + A_{1,2} \exp(k_1 x - w_1 t) \exp(k_2 x - w_2 t)),$$

where $w_1 = k_1^3 p + k_1^5 q + k_1^7, w_2 = k_2^3 p + k_2^5 q + k_2^7,$ and

$$A_{1,2} = \frac{(k_1 - k_2)^2 \left(5k_1^2 q - 5k_2 k_1 q + 5k_2^2 q + 7k_1^4 - 14k_2 k_1^3 + 21k_2^2 k_1^2 - 14k_2^3 k_1 + 7k_2^4 + 3p\right)}{(k_1 + k_2)^2 \left(5k_1^2 q + 5k_2 k_1 q + 5k_2^2 q + 7k_1^4 + 14k_2 k_1^3 + 21k_2^2 k_1^2 + 14k_2^3 k_1 + 7k_2^4 + 3p\right)}. \tag{60}$$

On the other hand, direct calculations show that the KdV7

$$u_t + \frac{(3\alpha - 5\beta)(2\beta + \gamma)^2}{1176} u^3 u_x + \frac{1}{56}(\beta - \gamma)(2\beta + \gamma) u_x^3 + \frac{1}{28}(2\beta + \gamma)(2\alpha - 3\beta + 3\gamma) u u_x u_{xx}$$
$$+ \frac{1}{28}(2\beta + \gamma)(\alpha - 2\beta + 2\gamma) u^2 u_{xxx} + \alpha u_{xx} u_{xxx} + \beta u_x u_{xxxx} + \gamma u u_{xxxxx} + u_{xxxxxxx} = 0$$

may be written in the following Hirota's bilinear form [4]:

$$\begin{cases} 2D_{xt}(f \cdot f) + \dfrac{7(\beta - \gamma)}{2\beta + \gamma} D_x^4 (f \cdot g) - \dfrac{3(\beta - 3\gamma)}{2\beta + \gamma} D_x^4 (f \cdot f) + \dfrac{14(6\alpha - 8\beta - 7\gamma)}{2\beta + \gamma} D_x^8 (f \cdot f) \\ + \dfrac{14(6\alpha - 8\beta - 7\gamma)}{2\beta + \gamma} (g \cdot g) = 0. \\ D_x^4 (f \cdot f) - (f \cdot g) = 0. \\ u(x,t) = A \partial_{xx} \log f(x,t), A = \dfrac{168}{\gamma + 2\beta}. \end{cases} \tag{61}$$

The seventh-order Kaup-Kuperschmidt Eq. (4) belongs to this class ($A = 1/2$). Using the obtained bilinear form, we may obtain all the results we presented in previous sections (for the special case $d = c/2 - b$).

5. Forced KdV7

The forced KdV7 is written as

$$u_t + a u^3 u_x + b u_x^3 + c u u_x u_{2x} + d u^2 u_{3x} + \alpha u_{2x} u_{3x} + \beta u_x u_{4x} + \gamma u_{5x} + u_{7x} = f(t). \tag{62}$$

The forced Sawada-Kotera-Ito Eq. (2) and the forced Lax Eq. (3) admit the exact solution.

$$u_t + 252 u^3 u_x + 63 u_x^3 + 378 u_x u_{2x} + 126 u^2 u_{3x} + 63 u_{2x} u_{3x} + 42 u_x u_{4x} + 21 u u_{5x} + u_{7x} = f(t).$$

Exact solution:

$$u(x,t) = B + 2\text{sech}^2(x - \lambda(t)) + F(t), \tag{63}$$

where

$$\lambda(t) = 4\int\Big(63B^3 + 189B^2F(t) + 126B^2 + 189BF(t)^2 + 252BF(t) + 84B + 63F(t)^3 + 126F(t)^2 + 84F(t) + 16\Big)dt. \tag{64}$$

and

$$F(t) = \int f(t). \tag{65}$$

$$u_t + 140u^3u_x + 70u_x^3 + 280u_xu_{2x} + 70u^2\,u_{3x} + 70u_{2x}u_{3x} + 42u_xu_{4x} + 14uu_{5x} + u_{7x} = 0.$$

- Exact solution:

$$u(x,t) = B + 2\mathrm{sech}^2(x - \lambda(t)) + F(t), \tag{66}$$

where

$$\lambda(t) = 4\int\left(\begin{array}{c} 35B^3 + 105B^2F(t) + 70B^2 + 105BF(t)^2 + 140BF(t) \\ +56B + 35F(t)^3 + 70F(t)^2 + 56F(t) + 16 \end{array}\right)dt. \tag{67}$$

and

$$F(t) = \int f(t). \tag{68}$$

For other parameter values, we obtained the following result:
If $\alpha + \beta + \gamma \neq 0$ and

$$\begin{aligned} a &= \frac{1}{63}d(\alpha + \beta + \gamma) \\ b &= \frac{1}{126}\left(-\alpha^2 + 4\alpha\beta - 11\alpha\gamma + 5\beta^2 - 5\beta\gamma - 10\gamma^2 + 126d\right) \\ c &= \frac{1}{21}\left(5\alpha\gamma + 5\beta\gamma + 5\gamma^2 - 42d\right) \end{aligned} \tag{69}$$

then the forced KdV7 in Eq. (2) admits the exact solution

$$u(x,t) = B + \frac{252}{\alpha + \beta + \gamma}\mathrm{sech}^2(x - \lambda(t)) + F(t), \tag{70}$$

where

$$\begin{aligned} \lambda(t) &= \frac{1}{63}\int\left(\begin{array}{c} F(t)\left(3\alpha B^2d + 3\beta B^2d + 3B^2\gamma d + 504Bd + 1008\gamma\right) \\ +F(t)^2(3\alpha Bd + 3\beta Bd + 3B\gamma d + 252d) + F(t)^3(\alpha d + \beta d + \gamma d) \end{array}\right)dt \\ &+\frac{1}{63}t\left(\alpha B^3d + \beta B^3d + B^3\gamma d + 252B^2d + 1008B\gamma + 4032\right), \\ F(t) &= \int f(t). \end{aligned}$$

See also [5].

DOI: http://dx.doi.org/10.5772/intechopen.1004789

Author details

Alvaro Humberto Salas Salas
Fizmako Research Group, Universidad Nacional de Colombia, Manizales, Colombia

*Address all correspondence to: ahsalass@unal.edu.co

IntechOpen

© 2024 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

References

[1] Yao R-X, Li Z-B. Conservation Laws and new exact solutions for the generalized seventh order KdV equation. Chaos, Solitons and Fractals. 2004;**20**: 259-266. DOI: 10.1016/S0960-0779(03) 00373-4

[2] Fan E, Hona YC. Generalized Tanh method extended to special types of nonlinear equations. Zeitschrift fur Naturforschung A. 2002;**57**:692-700. DOI: 10.1515/zna-2002-0809

[3] Wazwaz AM. Soliton solutions for seventh-order Kawahara equation with time-dependent coefficients. Modern Physics Letters B. 2011;**25**:643-648. DOI: 10.1142/S0217984911026012

[4] Optical and Quantum Electronics. 2020;**52**:511. DOI: 10.1007/s11082-020-02628-7

[5] Alvaro H. Salas, computing exact solutions to a generalized lax seventh-order forced KdV equation (KdV7). Applied Mathematics and Computation. 2010;**216**(8):2333-2338

Chapter 9

Numerical Solutions of Nonlinear Schrödinger Equation: An Application Example of Nonlinear Analysis

Peter Y.P. Chen

Abstract

The nonlinear Schrödinger equation is used to show how numerical methods can be used to solve mathematical problems present in nonlinear analysis. The Lanzos-Chevbychev Pseudospectral method is shown to be effective, flexible, and economical to meet various demands in practical applications of mathematical simulations using nonlinear differential equations. The electromagnetic wave propagation through an inhomogeneous, anisotropic, and complex space is used as an example to show how successful mathematical modeling could be used to explain the complex phenomenon of astronomical redshift that is the central issue in the widely debated Hubble tension.

Keywords: application of nonlinear analysis, numerical solution methods, nonlinear Schrödinger equation, pseudospectral method, electromagnetic wave propagation in space, astronomical redshift

1. Introduction

In recent times, nonlinear analysis has been increasingly used in science and technology. Many advanced and innovative applications in those fields include nonlinear effects in their design and development. To be useful to real-world problems, those mathematical models need to be solved by methods developed in nonlinear analysis. Out of many possible mathematical methods, some are developed specifically for nonlinear differential equations (NDEs). This chapter will concentrate on a specific method for the solutions of second order NDEs. As a specific example, nonlinear Schrödinger equation (NLSE) is being chosen. The emphasis is to show how such numerical methods can be used to investigate how electromagnetic waves propagate under various realistic physical conditions in space.

For many years researchers have had extensive interest in how to solve NDEs analytically. But, because of the nonlinear nature, little success has been achieved in solving them directly. That is, starting from NDE itself and finding the solutions analytically in a forward direction. However, for methods starting from some assumed solutions and working out how to satisfy the NDE analytically as an inverse problem,

there are many successes. Some of these inverse examples include the inverse differential and integral methods such as, for wave propagation [1], the *G'/G* expansion method [2] and its various variants [3–5], and inverse scattering methods for antenna design [6]. Generally, as an inverse problem, there is no limit to how many solutions can be found because there are an infinite number of choices for the set of system parameters that define the chosen base functions. However, the need to have a matching background medium is a notable limitation [7]. From the nature of those solutions, it could be concluded that this inverse approach is more suitable for qualitative analysis that the performance of a design, or the characteristics of a system could be assessed qualitatively. For quantitative assessment, the direct approach is a better choice, because the solutions are obtained by satisfying not just the NDE but also the initial/boundary conditions.

Numerical solutions of NDEs start with a scheme to discretize the problem into a set of simultaneous nonlinear algebraic equations. Linear algebraic algorithms are then used to solve those equations with an iterative scheme to cater for the nonlinear terms. For transient problems, especially when a long history involving a large set of equations is needed, the computational efficiency of the chosen method becomes important. As many different models under different prescribed conditions may be encountered, the flexibility of the method is also a factor for consideration.

In Section 2 of this chapter, we describe the Lanczos-Chevbychev Pseudospectral (LCPS) method [8, 9] that we have used to solve many different NDEs. The LCPS method uses an economized power series and has been shown to perform as well as similar orthogonal eigenfunction series expansion methods such as the Chebyshev Pseudospectral method. However, the advantage of using LCPS is that an ordinary power series is involved that would be the simplest and the most economical computing method. The details of the LCPS method are given in this section together with some application examples.

In Section 3, we apply NLSE to electromagnetic wave propagation through space [10, 11], together with two simple examples. We show in Section 4 that long distance, and other characteristic nature of space such as anisotropy, inhomogeneity, and gravitational effect, can be effectively included. How to calibrate our findings with empirical data is also described there. In Section 5, we discuss the usefulness and limitations of mathematical simulations based on examples we have solved. The check list includes items such as the appropriateness of the model, the variable ranges within which the model is applicable, and the implication of any assumptions made. To be realistic, we make use of our findings on astronomical redshift and compare them to popularly accepted theories in that the debate on Hubble tension is receiving considerable attention [12, 13]. We present our conclusions in Section 6.

2. Numerical solution methods for second order nonlinear partial differential equations

As higher order can be reduced to second order by introducing additional second order equations, we can restrict ourselves to consider only a second order differential equation in a dispersive field.

2.1 Nonlinear partial differential equations

Consider a time-dependent two-dimensional boundary problem

$$i\mathbf{u}_t = \mathbf{D}(\mathbf{x},t)\nabla^2\mathrm{u} + \mathbf{F}(\mathbf{x},t,\mathbf{u})\mathbf{u}, \tag{1}$$

$$\alpha\, \boldsymbol{u}'(\boldsymbol{x},t) + \beta\, \boldsymbol{u}\ (\boldsymbol{x},\mathrm{t}) + \gamma = 0, \text{at}\ \boldsymbol{x} = \boldsymbol{x_b}, \tag{2}$$

and,

$$\boldsymbol{u}(\boldsymbol{x},0) = \boldsymbol{u_o}(\boldsymbol{x}), \tag{3}$$

where $\boldsymbol{D}$ is the dispersion coefficient, and the nonlinear $\boldsymbol{F}$ is a spatial and time-dependent potential. α, β, and γ are coefficients associated with the boundary conditions.

As an example, we use a single mode solution $u(X,Y,t)$ in a two-dimensional Cartesian system with spatial variables, X and Y, coefficients D_1 and D_2, and a nonlinear potential F, such that

$$\boldsymbol{x} = \begin{Bmatrix} X \\ Y \end{Bmatrix},\ \boldsymbol{D}(\boldsymbol{x},t) = \begin{bmatrix} D_1(X,Y,t) & 0 \\ 0 & D_2(X,Y,t) \end{bmatrix},\ \boldsymbol{F}(x,t,u) = \begin{Bmatrix} F_u(X,Y,t,u,v) \\ F_v(X,Y,t,u.v) \end{Bmatrix}. \tag{4}$$

For numerical reasons, if a function is not smooth, such as in the case solitons, it is desirable to adopt a multidomain approach. The given rectangular two-dimensional domain of interest is divided into M x N subdomains. The affine transformation is used to scale each subdomain to $[-1, 1]^2$, in the new coordinates {x,y},

$$\Omega := \Omega^{i,j}, \quad i = 1,2,\ \ldots,M;\ \ j = 1,2,\ \ldots,N. \tag{5}$$

In the subdomains, the associated surfaces are

$$\begin{aligned} S_x^{0,j} &= \Omega^{1,j}[-1,y], S_x^{i,j} = \Omega^{i,j}[1,y] \cap \Omega^{i+1,j}[-1,y],\ \ i = 1,2,\ \ldots,M-1, \\ S_x^{M,j} &= \Omega^{M,j}[1,y],\ \ j = 1,2,\ \ldots,N, \\ S_y^{i,0} &= \Omega^{i,1}[x,-1], S_y^{i,j} = \Omega^{i,j}[x,1] \cap \Omega^{i,j+1}[x,-1], j = 1,2,\ \ldots,N-1, \\ S_y^{i,N} &= \Omega^{i,N}[x,1],\ \ i = 1,2,\ \ldots,M. \end{aligned} \tag{6}$$

Boundary conditions specified in Eqs. (2) and (3) apply only on those surfaces that form part of the boundary. For inter-subdomain surfaces, the specified conditions are continuities of both the function and its derivative. These also apply to the four corner points.

Based on the Lanczos-Chebushev Pseudospectral (LCS) method [8, 9], the function $u^{i,j}$ in each subdomain $\Omega^{\mathrm{i,j}}$, is be represented by the tensor product of two truncated power series,

$$u^{i,j}(x,y,t) = \sum_{l=0}^{L}\sum_{k=0}^{K} u_{l,k}^{i,j}(t)x^k y^l \tag{7}$$

and, using term-by-term differentiation, the derivatives

$$(u_x)^{i,j}(x,y,t) = \sum_{l=0}^{L}\sum_{k=1}^{K} k\ \overset{i,j}{u}_{l,k}(t)x^{k-1}y^l,$$

$$\left(u_y\right)^{i,j}(x,y,t) = \sum_{l=1}^{L}\sum_{k=0}^{K} l\ \overset{i,j}{u}_{l,k}(t)x^k y^{l-1},$$

$$(u_{xx})^{i,j}(x,y,t) = \sum_{l=0}^{L}\sum_{k=2}^{K} k(k-1)\,\overset{i,j}{u}_{l,k}(t)x^{k-2}y^{l},$$
$$(u_{yy})^{i,j}(x,y,t) = \sum_{l=2}^{L}\sum_{k=0}^{K} l\,(l-1)\,\overset{i,j}{u}_{l,k}(t)x^{k}y^{l-2}. \tag{8}$$

Based on the approach proposed by Lanczos [14, 15], the discretization of the problem is done by collocation at specially chosen gride points. For example, the grid points for the x variable over the interval $[-1,1]$ in each subdomain are the K - *1* roots of a Chebyshev function, where K is the highest order of the power series used,

$$x_k = -\cos\left\{\frac{(2k+1)\pi}{2(K-2)}\right\}, k = 0, 1, \ldots, K-2, \tag{9}$$

and for the y variable,

$$y_l = -\cos\left\{\frac{(2l+1)\pi}{2(L-2)}\right\}, l = 0, 1, \ldots, L-2. \tag{10}$$

For each subdomain, the function $u^{i,j}(x,y,t)$ as well as its derivatives are substituted into the governing differential equation at the grid points, (x_l, y_k), l = 0, 1, … , L-2 and k = 0, 1, … , K-2 to give (L - 1) x (K – 1) ODEs. On the four surfaces of each subdomain, boundary or interfacial continuity condition is specified on grid points: [$x = \pm 1, y = \pm 1$], [(x_l, l = 0,1 … L-2), ($y = \pm 1$)], and [($x = \pm 1$), (y_k, k = 0, 1 … K-2)] to give a further (2 L + 2 k + 4) ODEs. The assemble of ODEs for the system is in the form,

$$iAU_t - L_1U - H_1(U,t) = 0. \tag{11}$$

In Eq. (11), the unknown coefficients $\boldsymbol{U}$ is a *[M x N x (K + 1) x (L + 1)]*. $\boldsymbol{A}$ and $\boldsymbol{L_1}$ are linear matrices, $\boldsymbol{H_1}$ is a nonlinear vector. But their row dimensions are larger than the length of $\boldsymbol{U}$. We use a discrete least square method to rectify this problem by multiplying Eq. (11) with $\mathbf{A}^{\mathbf{T}}$, the matrix transpose of $\mathbf{A}$. The resultant matrix equations are well-posed to be solve by a linear equation solver with an iterative procedure to carter for the nonlinear term.

For a one-dimensional problem, N = 1 and L = 0, and in the m^{th} subdomain,

$$\mathbf{U}^m = \left\{u_0^1, u_1^1, \ldots, u_k^1, u_0^2, u_1^2, \ldots, u_K^2, \ldots, u_0^M, u_1^M, \ldots, u_K^M\right\}'. \tag{12}$$

The matrix $\boldsymbol{A}$ in Eq. (11) consists of

$$\boldsymbol{A} = \begin{bmatrix} A_1 & & & & & & & \\ & A_2 & & & & & & \\ & & \cdot & & & & & \\ & & & \cdot & & & & \\ & & & & \cdot & & & \\ & & & & & & & A_M \\ S_{1,1} & S_{1,2} & & & & & & \\ S_{2,1} & S_{2,2} & S_{2,3} & & & & & \\ & & \cdot & & & & & \\ & & & \cdot & & & & \\ & & & S_{M-1,M-2} & S_{M-1,M-1} & S_{M-1,M} & & \\ & & & & S_{M,M-1} & S_{M,M} & & \end{bmatrix}, \tag{13}$$

where $A_m = \begin{bmatrix} x_1^0 & x_1^1 & \cdot & \cdot & \cdot & x_1^M \\ x_2^0 & x_2^1 & \cdot & \cdot & \cdot & x_2^M \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ x_K^0 & x_K^1 & \cdot & \cdot & \cdot & x_K^M \end{bmatrix}, m = 1, 2, \ldots M,$

and A_m is independent of m. There are M rows of S's and each S is a 2 x $(K + 1)$ matrix:

$$S_{1,1} = \begin{bmatrix} (-1)^0 & (-1)^1 & \cdot & \cdot & \cdot & (-1)^{K-1} & (-1)^K \\ 0 & 1 & 2 & \cdot & \cdot & (K-1) & K \end{bmatrix},$$

$$S_{1,2} = \begin{bmatrix} 0 & \cdot & & & \cdot & & \cdot \\ 0 & -1 & -2(-1)^1 & \cdot & \cdot & -(K-1)(-1)^{K-2} & -K(-1)^{K-1} \end{bmatrix},$$

and

$$S_{M,M-1} = \begin{bmatrix} 1 & 1 & \cdot & \cdot & \cdot & 1 & 1 \\ 0 & & \cdot & & \cdot & & \cdot \end{bmatrix}$$

$$S_{M,M} = \begin{bmatrix} (-1)^0 & (-1)^1 & \cdot & \cdot & \cdot & (-1)^{K-1} & (-1)^K \\ 1 & 1 & \cdot & \cdot & \cdot & 1 & 1 \end{bmatrix}. \tag{14}$$

Depending on the actual field equation involved, $\boldsymbol{L}_1$ and $\boldsymbol{H}_1$ can be constructed accordingly. Details of how the elements of each matrix could be determined are given in Ref. [8].

2.2 Solution by the real time evolution (RTE) method

This well-known method has been used to solve an initial-boundary problem [16] that evolves into a stationary and periodic solution. This method could be modified to cover cases where the solutions are periodical in time. In this special application, $\boldsymbol{U}$ is the same at the beginning and at the end of a period. To march from the beginning of a period to the end, we have chosen the implicit Crank-Nicholson stepwise formulation that is unconditionally stable. For Eq. (11),

$$iA\left(U^{m+1} - U^m\right) - \frac{\Delta t}{2}\left[L_1\left(U^{m+1} + U^m\right) + H_1\left(U^{m+1}, t^{m+1}\right) + H_1(U^m, t^m)\right], \tag{15}$$

where the superscript m refers to the time step number. With the superscript r refers to the iteration number and the symbol '→' means an integration, step, the iterative approach would be

$$U^{m+1,0} = U^m; \text{then } U^{m+1,r-1} \quad \rightarrow \quad U^{m+1,r}. \tag{16}$$

It should be noted that the iteration approach is needed due to the nonlinear $\boldsymbol{H_1}$ terms. Since $\boldsymbol{A}$ and $\boldsymbol{L_1}$ are both linear, only one inversion is required for all the iterative and time steps:

$$U^{m+1,r+1} = \left[iA - \frac{\Delta t}{2}L_1\right]^{-1}\left(\left[iA + \frac{\Delta t}{2}L_1\right]U^m + \frac{\Delta t}{2}\left[H_1\left(U^{m+1,r}, t^{m+1}\right) + H_1(U^m, t^m)\right]\right). \tag{17}$$

But starting with any pulse energy, a periodic solution may not exist. For this reason, we have developed a version of RTE method that, as shown later, the iteration will converge to an exactly periodic (EP) solution.

For a single mode problem, a term *exp(iμ t_0)* could be factored out from the solution of Eq. (17) at the position of the pulse peak. At a given time, $t = t_o$,

$$\mathrm{u}(\mathrm{x}, \mathrm{y}, \mathrm{t}_0) = \exp(\mathrm{i}\mu\mathrm{t}_0)\hat{\mathrm{u}}(\mathrm{x}, \mathrm{y}, \mathrm{t}_0). \tag{18}$$

Generally, if T is the period,

$$\mathrm{u}(x, y, t_0 + T) = \exp[i\mu(t_0 + T)]\hat{\mathrm{u}}(x, y, t_0 + T), \tag{19}$$

and, for *u(x,y,t_0)* to be periodic,

$$\mu T = 2\pi \tag{20}$$

and

$$\hat{\mathrm{u}}(x, y, t_0) = \hat{\mathrm{u}}(x, y, t_0 + T). \tag{21}$$

As Eq. (17) can only be used to solve for exactly the number of coefficients in the series expansion, the pulse energy needs to be specified so that μ will be unique. For this reason, we have designed a set of iterative algorithms based on the pulse energy being of a specific value. To achieve this purpose, the pulse energy at the end of each iterative step is adjusted to the specific energy, eventually, the procedures lead to the correct μ and the converged pulse shape and pulse energy. The rate of convergence could be improved, if we also use the well-used averaging method [17] for u:

i. Start with

$$\hat{\mathrm{u}}^0(x, y, 0) = u(x, y, 0), \quad \rightarrow \quad u^0(x, y, T). \tag{22}$$

ii. Find $\mu_u T$ from $u^m(x_o, y_o, T)$, then

$$\begin{aligned} \hat{\mathrm{u}}^m(x, y, T) &= \exp[-i\mu_u T]u^m(x, y, T), \\ w(x, y) &= \frac{1}{2}[\hat{u}^m(x, y, T) + u^m(x, y, 0)], \\ u^{m+1}(x, y, 0) &= w^{m+1}(x, y)\sqrt{E/\langle |w| \cdot |w| \rangle} \quad \rightarrow \quad u^{m+1}(x, y, T), \end{aligned} \tag{23}$$

where the superscript m is the iteration number and E is the specified energy. The symbol '→' indicates that, in each iteration, $u^m(x,y,T)$ is obtained from $u^m(x,y,0)$ using Eq. (17).

An obvious pre-condition is that the initial input used must be close to the converged solution. There are no set rules, but a Gaussian pulse is a good start. If error in the input has a negative imaginary component in its eigenvalue, the iteration will not converge due to modulation instability.

2.3 Numerical example of bimodal wave propagation

The governing equation for the spatiotemporal evolution of complex wave $u(z, x, \tau)$ and $v(z, x, \tau)$ in a planar waveguide is known [18] as

$$\begin{aligned} \mathrm{i}u_z + \frac{1}{2}u_{xx} + \frac{1}{2}D_1 u_{\tau\tau} + vu* &= 0, \\ 2\mathrm{i}\,(v_z + cv_\tau) + \frac{1}{2}v_{xx} + \frac{1}{2}D_2 v_{\tau\tau} - qv + \frac{1}{2}u^2 &= 0. \end{aligned} \tag{24}$$

where c is the group velocity mismatch parameter and q the phase mismatch constant. These equations are the same in form to those dealt with previously but with t and y replaced by z and τ respectively.

The system where v has twice the frequency of u is known as second harmonic generation. Traveling waves would split into a fundamental and a second harmonic modes. For such a system, the solution principles used in the RTE method remain the same with both u and v involved [19]:

$$\begin{aligned} \hat{u}^m(x,,z) &= \exp(-i\mu z)u^m(x,,z), \\ \hat{v}^m(x,\tau,z) &= \exp(-2i\mu z)v^m(x,\tau,z). \end{aligned} \tag{25}$$

As there are two pulse energies E_u and E_v now involved, we have three choices for assigning energy: The total energy $E = E_u + E_v$, or E_u and E_v by itself.

A system that supports the copropagating of two pulses of arbitrary frequencies may support also a continuously varying spectrum. For this reason, errors in the initial guess could grow with distance traveled. It is important to design an algorithm to ensure that the iterative procedures will lead only to the ground state solutions for u. It is noted that, at convergence, v will also assume equilibrium state. To implement these ideas into our algorithms, we set $\mu_u = 1$ for u. At the end of each iterative cycle, we re-scale u so that μ could be forced to converge to 1. For v we do not preset μ_v and just let it assume its own value at convergence. The constraint we use for v is a specified energy ratio $R = E_v/E_u$. Again, E_v is not given a value at the beginning, but it will assume a value once a converged E_u is found. The algorithms are as follows:

i. Start with

$$\begin{aligned} \hat{u}^0(x,y,0) = u(x,y,0), \quad &\rightarrow \quad u^0(x,y,T), \\ \hat{v}^0(x,y,0) = v(x,y,0), \quad &\rightarrow \quad v^0(x,y,T). \end{aligned} \tag{26}$$

ii. Find $\mu_u T$ from $u^m(x_o, y_o, T)$ and $\mu_v T$ from $v^m(x_o, y_o, T)$, then

$$
\begin{aligned}
\hat{u}^m(x,y,T) &= \exp[-i\mu_u T]u^m(x,y,T),\\
\hat{v}^m(x,y,T) &= \exp[-i\mu_v T]v^m(x,y,T),\\
uu(x,y) &= \frac{1}{2}[\hat{u}^m(x,y,T) + u^m(x,y,0)],\\
vv(x,y) &= \frac{1}{2}[\hat{v}^m(x,y,T) + v^m(x,y,0)],\\
r_u &= \sqrt{\frac{1}{\mu_u}},\\
u^{m+1}(x,y,0) &= r_u\, uu(x,y) \quad \rightarrow \quad u^{m+1}(x,y,T),\\
r_v &= \sqrt{R\frac{\langle |u^{m+1}(x,y,0)| \cdot |u^{m+1}(x,y,0)| \rangle}{\langle |vv(x,y)| \cdot |vv(x,y)| \rangle}},\\
v^{m+1}(x,y,0) &= r_v\, vv(x,y) \quad \rightarrow \quad v^{m+1}(x,y,T).
\end{aligned}
\tag{27}
$$

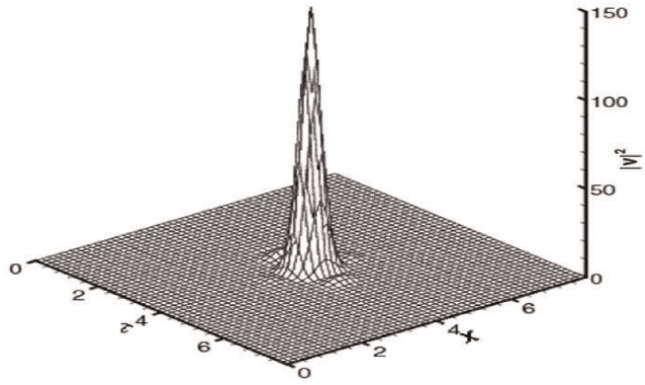


Figure 1.
Stationary solutions found for Eq. (24) with E = 400, D2 = − 0.2 and q = 2. (symmetry in the central x-plane was used with 4 x 2 subdomains and K = L = 8. The initial guesses for u and v were Gaussian pulses in both directions).

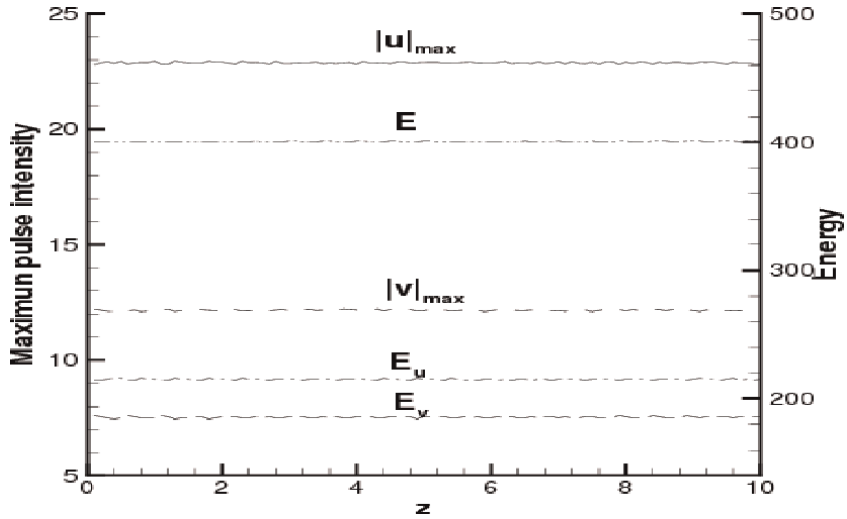


Figure 2.
Stable propagation of the stationary solutions shown in ***Figure 1.***

We have applied the LCPS method to Eq. (24) and obtain a set of ODEs that was solved with RTE method to give a set of stationary solutions. The complicated waveforms found could be seen from **Figure 1**. We then propagate this set of solutions over a distance z = 10. The propagation histories show that there is no change in the pulses amplitude and energy as can be seen from the plots in **Figure 2**.

3. Electromagnetic wave propagation through space

Numerical procedures described in the previous sections have been modified and used to study the propagation characteristics of electromagnetic waves in the form of bright, dark, and anti-dark solitons [10, 11, 20]. The steps needed are to be described below.

3.1 Stable periodic (SP) soliton solutions of NLSE

For a plane wave the governing NLSE and boundary conditions are

$$u_x - \frac{i}{2}D(x)u_{tt} - i\gamma|u|^2u = 0, u(0,x) = u(L,x) = 0, \tag{28}$$

where u is the slow varying envelope of the axial electric field, $D(x)$ and γ represent the dispersion coefficient and self-phase modulation parameters, respectively. x and t are the spatial propagation distance and temporal local time, respectively. L is the width of the numerical window used for t. For application to space, x is a very large number while D and γ are very small. To eliminate possible numerical complications associated with those numbers, scaling factors, x_o and t_o, are introduced so that

$$x^* = \frac{x}{x_o}, t^* = \frac{t}{t_o}, \tag{29}$$

then, together with

$$D^* = \frac{Dx_o^{0.5}}{\gamma^{0.5}}, \quad u^* = (\gamma x_o)^{0.5}u, \tag{30}$$

Eq. (28) becomes dimensionless,

$$u_x - \frac{i}{2}D(x)u_{tt} - i|u|^2u = 0, \tag{31}$$

where the superscript * has been omitted for simplicity.

To solve Eq. (31) numerically, consider pulse propagation as a transient problem along the spatial distance, x, the discretization is one dimensional and only at the temporal local time domain t. Using M subdivisions

$$\Omega := \Omega^i, \quad i = 1, 2, \ldots, M. \tag{32}$$

For each subdivision, the numerical window of length L is mapped into an interval varying from – 1 to +1, and an economized power series is used:

$$u(t,x) = \sum_{k=0}^{K} u_k(x) t^k. \tag{33}$$

Applying the LCPS method to Eq. (31) at the collocation points, t_0, $t_{1\ldots}$ t_K, to each subdomain leads to a set of ODEs. The assembly of all subdomains involves the series expansion coefficient as a vector of length $(K+1) \times M$:

$$\boldsymbol{u} = \left\{u_0^1, u_1^1, \ldots, u_K^1, u_0^2, u_1^2, \ldots, u_K^2, \ldots, u_0^M, u_1^M, \ldots, u_K^M\right\}'. \tag{34}$$

Between two adjacent subdomains, i and i + 1, the continuity conditions are:

$$u^i\big|_1 = u^{i+1}\big|_{-1}; \frac{d}{dx}u^i\bigg|_1 = \frac{d}{dx}u^{i+1}\bigg|_{-1}. \tag{35}$$

The set of transient ODEs obtained is in the form,

$$\boldsymbol{A}\boldsymbol{u}_x(x) - i\boldsymbol{L}\boldsymbol{u}(x) = i\boldsymbol{Q}(x,\boldsymbol{u}). \tag{36}$$

Applying the RTE method described in Section 2.2 to the above equation,

$$\boldsymbol{A}\left(\boldsymbol{u}^{m+1} + \boldsymbol{u}^m\right) - \frac{i\Delta x}{2}\left[\boldsymbol{L}\left(\boldsymbol{u}^{m+1} + \boldsymbol{u}^m\right)\right] = \frac{i\Delta x}{2}\left[\boldsymbol{Q}\left(x,\boldsymbol{u}^{m+1}\right) + \boldsymbol{Q}(x,\boldsymbol{u}^m)\right], \tag{37}$$

where Δx is the step size, and the superscript m refers to the time step number. To carter for the nonlinear nature of Eq. (37) that is associating with $\boldsymbol{Q}$, an iterative algorithm [9, 10] is used.

The initial input pulse for a bright soliton could be

$$u(t,0) = \beta\ \exp\left[-\alpha(t - 0.5L)^2\right], \tag{38}$$

where L is the numerical window used for t, α a chosen constant to give an input pulse as close to the SP soliton as possible, and β an adjusting parameter to give a specified pulse energy, E,

$$E(x) = \int_{-\frac{L}{2}}^{\frac{L}{2}} \left(\left|u(t,x)\right|^2\right) dt. \tag{39}$$

As *u(t,x)* is a truncated soliton pulse, it has been found [9] that to eliminate residual reflection, the following boundary conditions could be used:

$$u(t,x) = 1000\frac{\partial u}{\partial t} - u(t,x)\ at\ x = \pm 0.5L. \tag{40}$$

To find the stable periodical (SP) solution, Eq. (37) is integrated to a selected distance Z, with the first half using a specified dispersion coefficient of – D, and for the second half D. For an SP solution, the input pulse must be the same as the output pulse. We use this fact to design an iterative scheme based on successive halves,

$$\boldsymbol{u}_{in}^{m+1} = 0.5\left(\boldsymbol{u}_{in}^m + \boldsymbol{u}_{out}^m\right), \tag{41}$$

where u_{in}, and u_{out} are the input and output pulse to the dispersion map respectively and the superscript m denote the iteration number. It should be noted that stable periodic solitons are special cases of cyclic solitons in that no phase matching is needed. For the exact periodic (SP) solutions described in Section 2.2. the input and output pulses have the same amplitude and phase.

3.2 A numerical example of a SP bright soliton

Using the procedure described in the previous section and a dispersion map with length Z = 6, **Figure 3** shows how the solutions converged to stable and periodic pulses [10]. The distance, x, shown is the cumulated distance. As the step size is 0.0005, each iteration generates 12,000 pulses. In the last few iterative cycles, the pulse width is

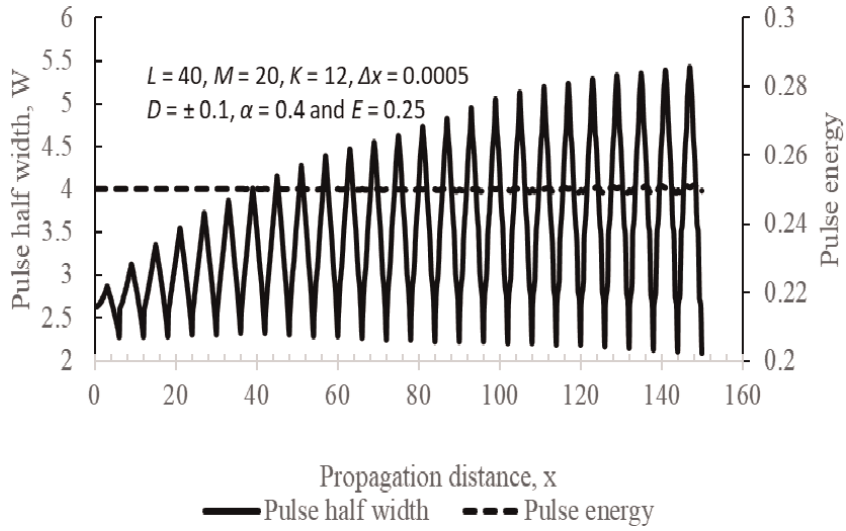


Figure 3.
Iteration convergence for the numerical example.

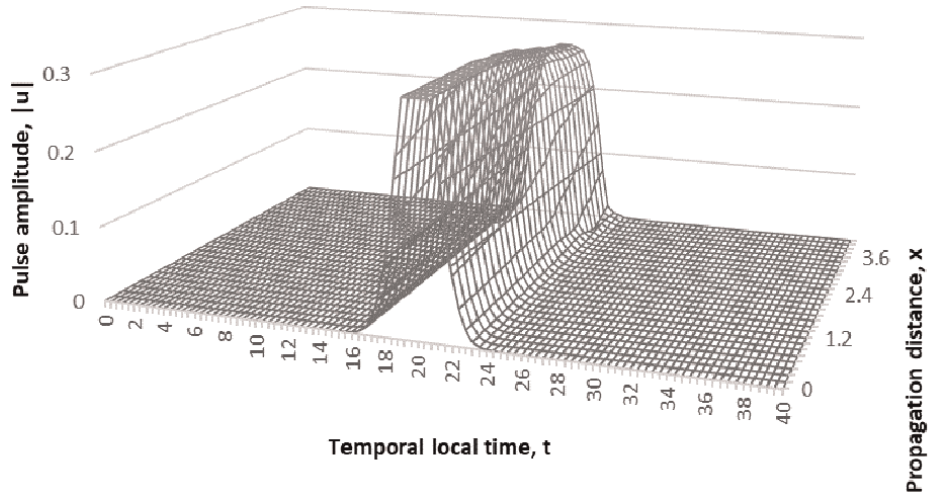


Figure 4.
Pulse changes when propagating through a medium with -D.

changing linearly with the distance traveled in both halves of the dispersion map. The input and output pulses in a dispersion map have the same shape and energy but not the same phase.

Figure 4 shows changes of the pulse shape as an SP soliton is traveling through a medium with negative *D*.

4. Electromagnetic wave propagation in a complex space system

Some of the transmission characteristics of electromagnetic waves through space have been investigated previously [11, 20]. We shall deal with a space that has other complex features in the following sections.

4.1 SP solitons with different pulse energies and in a space with random dispersion coefficient

We have solved for a segment consisting of piecewise continuous dispersion coefficients as given in **Table 1** below. The pulse width histories for cases with pulse energy E = 0.4 and 0.8 found for the above cases are shown in **Figure 5**. The plots show that the overall pulse width change for the random dispersion case is the same as that based on the averaged dispersion. Also in the plots are equations of the trendlines showing the linear relationship between pulse width and distance traveled when the average coefficient D is used. Also, for two times increase pulse energy, the deviation from the average gradient is only +/− 4%.

4.2 A system of multiple SP solitons

The NLQE for multiple solitons is

$$u_x^j - \frac{i}{2}D(x)u_{tt}^j - i\gamma\left\{\sum_{k=1}^{J}\left|u^k\right|^2\right\}u^j = 0, \mathrm{j} = 1, 2, ..\mathrm{J}. \tag{42}$$

Although a set of equations, equal in number to the number of solitons, are involved, the same numerical procedures described previously can be used to obtain a set of SP solitons. For our purpose, however, we only use three well-spaced and

	Case 1			Case 2		
Section	**Propagation distance, *x***	**Dispersion coefficient, D**	**External source, u_o**	**Propagation distance, *x***	**Dispersion coefficient, *D***	**External source, u_o**
1	2	−0.1	0	2	−0.2	0
2	0.5	−0.1	0	0.5	0.2	−0.2
3	1	−0.1	0	1	0.1	0
4	0.5	−0.1	0	0.5	−0.2	0.2
5	2	−0.1	0	2	−0.15	0

Table 1.
Two cases of propagation through different D and u_o.

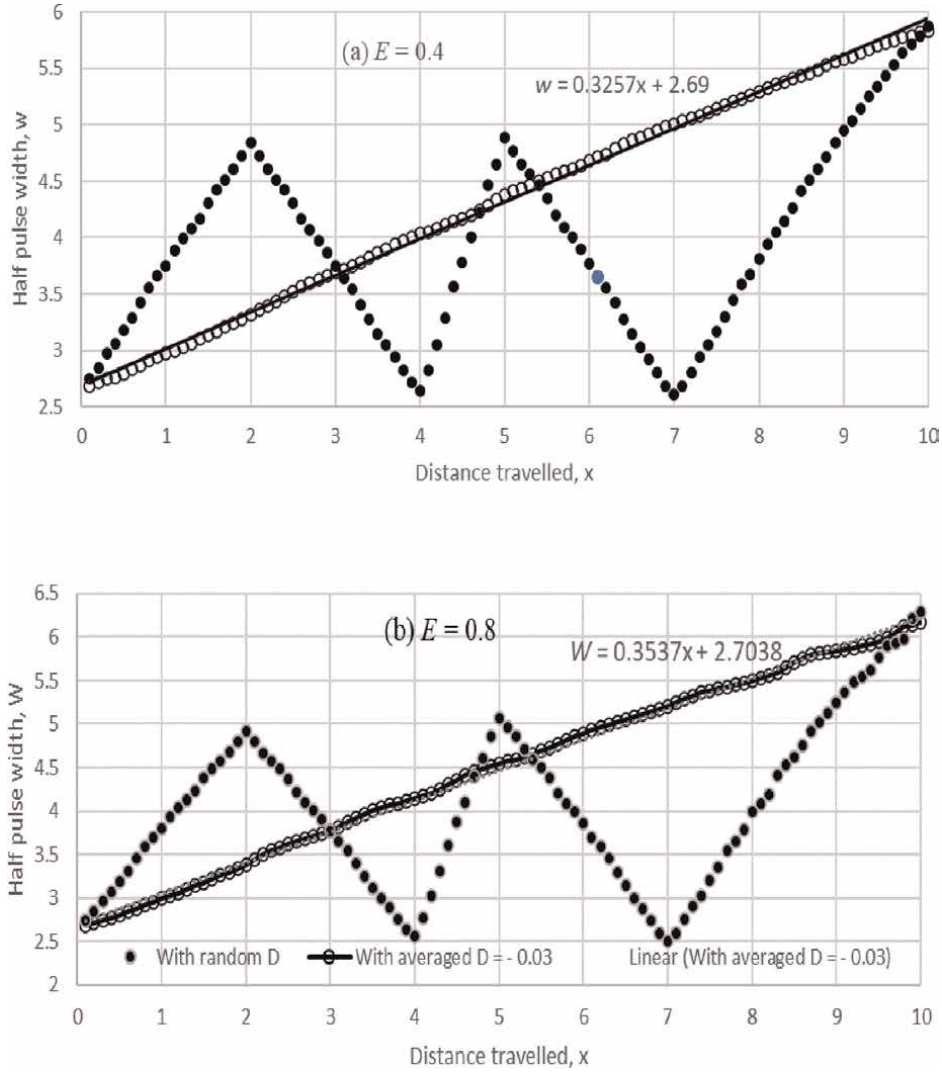


Figure 5.
Propagation of SP soliton with different pulse energy.

identical pulses. **Figure 6** shows how the central pulse has converged to an SP soliton. The same applies to the other two solitons.

How the pulse shape is changing can be seen in **Figure 7**. The gradient of pulse width changes against distance traveled is not sensitive to pulse energy as can be seen in **Figure 8**.

4.3 Propagating through space with a CW background

To include a constant CW background, u_o, into NLSE, let

$$u = v + u_o.$$

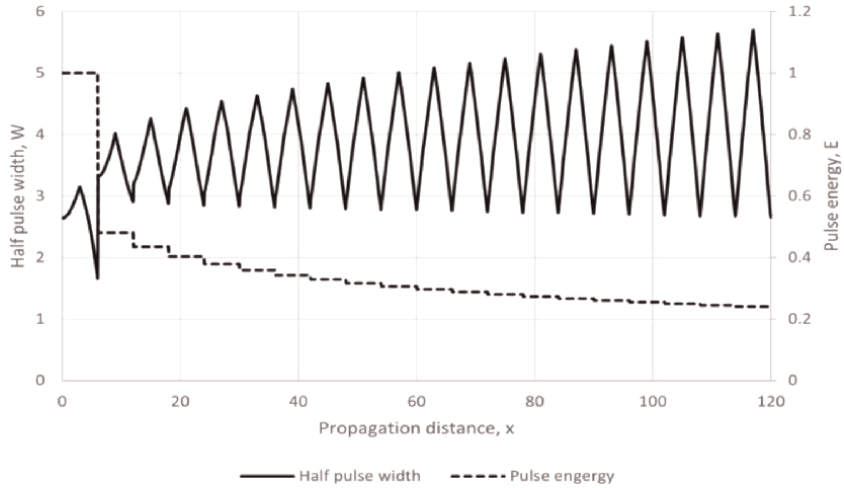


Figure 6.
Iteration convergence for the central pulse.

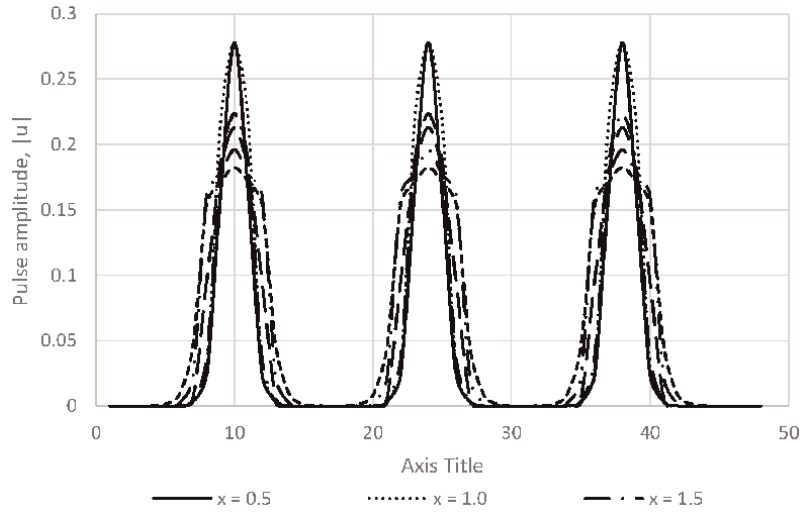


Figure 7.
Changing shape as pulse propagating through a region where D is negative.

Substituting the above into Eq. (31) to give

$$v_x - \frac{i}{2}D(x)v_{tt} - i|v+u_o|^2(v+u_o) = 0. \tag{43}$$

Using the same numerical procedures as described previously, Eq. (43) could be solved to give an SP solution. By propagating this SP soliton along the distance x, the transmission characteristics could be determined. We have solved for two cases with different system parameters as shown in **Table 1**. In fact, Case 1 is using a constant D that is the same as the distance weighted average of D in Case 2. The pulse width and energy histories are plotted out in **Figure 9**. It could be seen that u_o has little influence on the overall pulse width change.

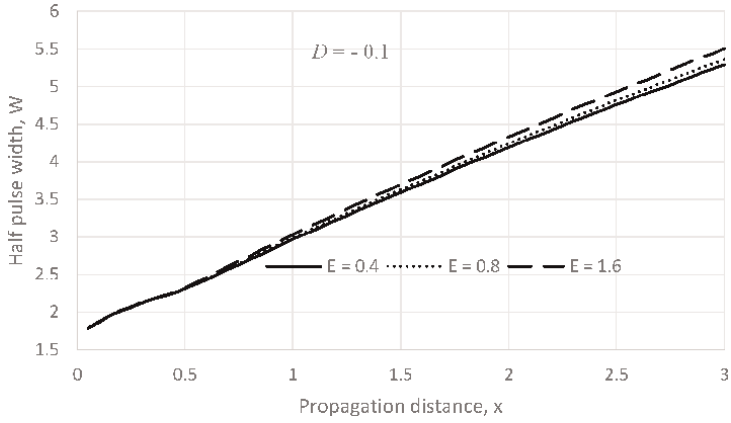


Figure 8.
Half pulse width changes versus distance traveled at different pulse energy.

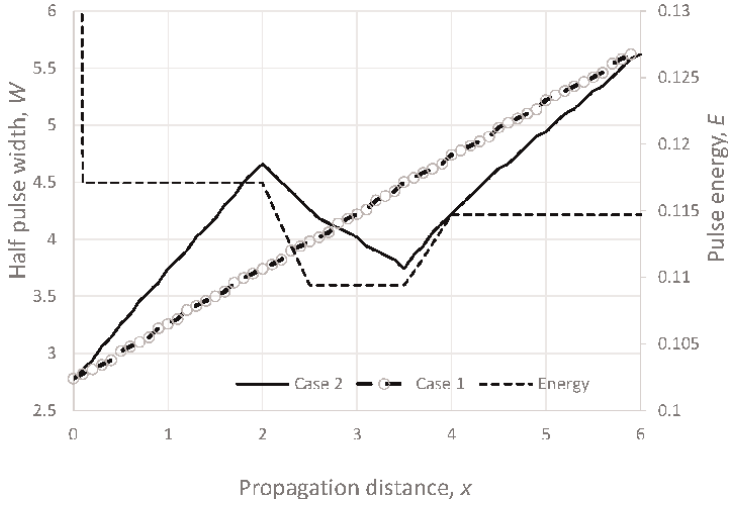


Figure 9.
Two cases of propagation through space with CW background.

4.4 Propagation through an amplifying or attenuating space

The NLS equation for electromagnetic waves (solitons) propagation in dimensionless form and in an attenuating space is

$$u_x - \frac{i}{2}D(x)u_{tt} - i|u|^2u = S(x), \tag{44}$$

where $S(x) = su$ for amplification, and $S(x) = -su(x)$ for attenuation and s is a constant.

To solve the above equation numerical, S must be added to $\boldsymbol{Q}$ in Eq. (36), As a numerical example, a system consists of three sections was used. In all sections,

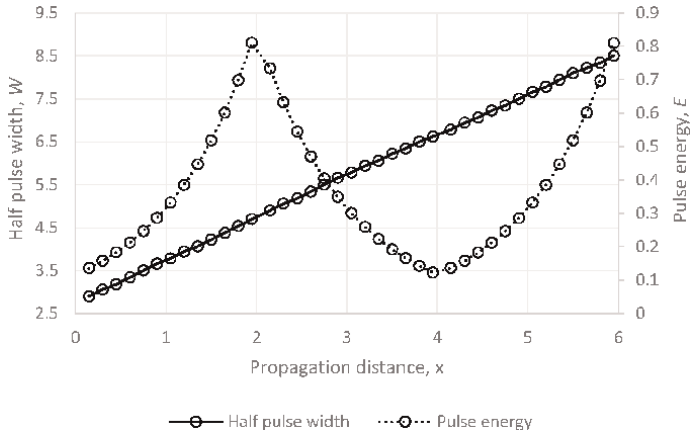


Figure 10.
Propagation histories in the present of an external source.

$D = -0.2$ but s = 0.5, − 0.5 and 0.5, respectively. Solutions are shown in **Figure 10**. Features of the solution histories fond are: (a) based on the sign of s, the pulse energy increases or decreases steadily, and (b) practically, there is no change in the gradient of the wavelength half width versus x curve.

4.5 Propagation with gravitational deflection

Based on the general relativity theory, the approximate light path deflection angle, $\Delta\Theta$ is found to be [21],

$$\Delta\Theta = \frac{4GM}{\Delta}, \tag{45}$$

where d the distance between the light path and the center of the mass, G is the gravitation constant, M is the mass, and Δ is the distance between the wave front and the center of the mass.

Eq. (45) could be used in its dimensionless form,

$$\Delta\Theta = \frac{C}{\mathcal{E}}, \tag{46}$$

where $\mathcal{E} = \frac{C\Delta}{4GM}$. Since the event is taking place in space, we have no way of knowing M and Δ. But we can still track the gravitational deflection history using a single arbitrarily chosen parameter C. Using a new rectangular coordinate system *(x1, x2),* at a particular step, let $((x1)_m, (x2)_m)$ be the position of the mass center and $((x1)_1, (x2)_1)$ the wave front; the straight line connecting the wavefront to the center of the mass is

$$\mathcal{E} = \sqrt{\left((x1)_m - (x1)_1\right)^2 + \left((x2)_m - (x2)_1\right)^2}. \tag{47}$$

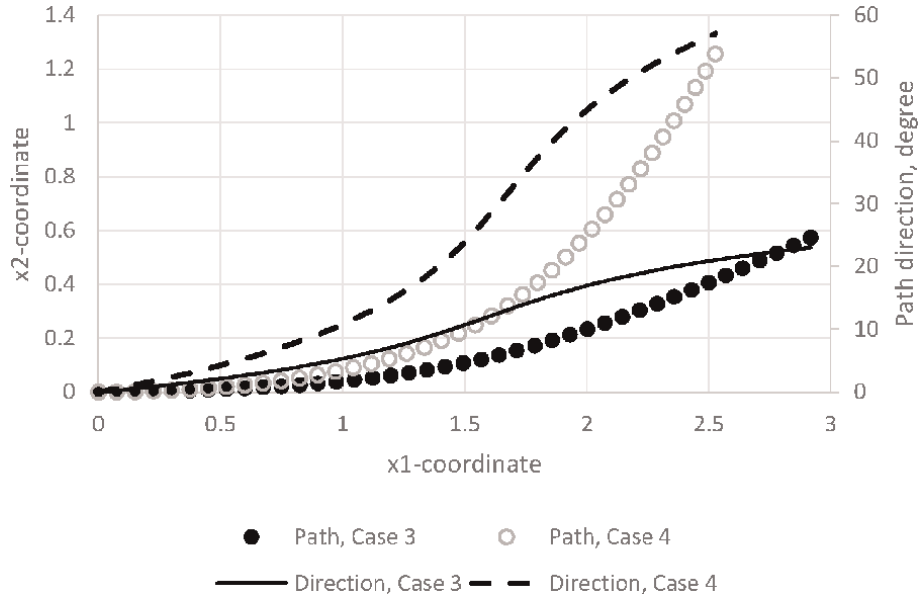


Figure 11.
Propagation in the present of gravitational deflection.

Then, by specifying a C, the deflection angle $\Delta\Theta$ can be found from Eq. (46). If a wavefront is propagating along a light path making an angle Θ with the *x1*-axis. Integrating along *x1* with step Δx, the new wave front position would be $[((x1)_1 + \Delta x)\ cos\ (\Theta + \Delta\Theta)),\ ((x2)_1 + \Delta x)\ sin\ (\Theta + \Delta\Theta)]$. Knowing the new position, Eq. (46) and Eq. (47) could be used to find $\mathcal{E}$ and $\Delta\Theta$, and the next wavefront position in the next integration step.

To implement this deflection scheme, let Z be the length of the propagation distance to be investigated. Let t(0, 0) be the starting position with the mass $\boldsymbol{M}$ located along the straight line $x1 = 0.5\,Z$. If φ is the angel between the line joining the wavefront to the center of the mass and the x1-axis, the mass is located at (0.5 Z, 0.5 Z tan(φ)). For this example, let the initial $\varphi = 20$, $\Theta = 0$, and $Z = 3$, and $D = -0.2$. Considering two cases, Case 3, C = 0.00005, and Case 4, $C = 0.0001$, respectively. After solving for the wavefront histories, the solutions are plotted out in **Figure 11**. It can be seen that, at x1 = 0.5 Z where the wavefront is closest to the mass, the deflection rate is the largest. As expected, a larger C will give a larger deflection. Without deflection, the wavefront will move along the x1-axis. With deflection, the wavefront has traveled the same distance along its path, but shorter in term of x1-coodinate. It should be pointed out x1 and x2 are scaled down to dimensionless quantities chosen according to local conditions. They would be many orders smaller than the entire propagation distance.

4.6 Calibrations with physical systems

The coefficients associated with dispersion and self-phase modulation for space cannot be measured directly. To apply our numerical findings to cosmological red-shift, calibration with measured data must be used as described in Ref. [10, 19]. If the starting and ending wavelength is λ_1 and λ_2, and the half pulse width, W_1 and W_2, and the dimensionless redshift, z, as defined in astronomy, is given by,

$$z = \frac{\lambda_2 - \lambda_1}{\lambda_1} = \frac{W_2 - W_1}{W_1}. \tag{48}$$

Also, in astronomy, the measured redshift-distance relation is given by Hubble's law

$$z = H_o d/c, \tag{49}$$

where H_o is the Hubble constant given in units of km/s/Mpc. The distance between the light source and an observer, d is in Mpc unit, and c is the speed of light in km/s. Since Eq. (49) is linear with specific physical dimensions, and our numerical results are also linear but dimensionless, we could choose any two given points in Eq. (49) for calibration, so that our results agree with Hubble's law.

To use real physical data for calibration, we choose, as an example, H_o = 70 km/sec/Mpc for Eq. (49). We also choose the pulse representing the spectral line due to Lyman-alpha hydrogen that has a wavelength of 121.6 nm., and the corresponding period is 405 ps. For this example, we choose the linear relationship that we have found previously for a random medium and shown in **Figure 5**,

$$W = 0.9539x + 2.5834, \tag{50}$$

and the half pulse width at the two calibration points, A at x_A = 2, and B at x_B = 6, from Eq. (50), are W_A = 4.9286, and W_B = 8.3068. Then, the temporal time multiplying scaling factor to convert W into period in picosecond,

$$t_o = \frac{405}{W_A} = 82.1734 \text{ ps}. \tag{51}$$

Now, the redshift between A and B,

$$z_{AB} = \frac{W_B - W_A}{W_A} = 0.6854. \tag{52}$$

Using Eq. (49), the distance variable between the source and the observer,

$$d/c = \frac{z_{AB}}{H_o} = \frac{0.6854}{70} 0009792 \text{Mpcs/km} \tag{53}$$

$$d = 0.009792c \times \frac{3.08567758 \times 10^{19}}{9.461 \times 10^{12}} = 9.581\, 10^9 \textit{ light year} \tag{54}$$

Using Eq. (53), the distance scaling factor used to convert x to d/c is

$$f_x = \frac{d/c}{x_{A-x_B}} 0002448 \text{Mpcs/km} \tag{55}$$

5. Discussion

We have demonstrated that numerical methods can be used to solve many nonlinear field problems. The general mathematical procedures involve the reduction of the governing differential equations to a set of nonlinear algebraic equations that is

solved by an iterative approach. Depending on the expected characteristics of the solutions, numerical algorithms may need to be modified. For an exactly stationary and periodic traveling wave, the iterative procedures as described in Section 2.2 are needed. But for the studies of redshift, only a stable and periodic solution in the sense, that the traveling pulse can reproduce itself anywhere in the whole propagation history, the iterative procedures needed are less elaborative as described in Section 3.1. Since the system is a nonlocal problem, the finding of the correct initial pulse is vital for the investigation of redshift in starlight.

The soliton solutions' spike-like characteristic needs a very high order of series approximation. We choose to sub-divide the computational domain into zones so that a lower order series could be used for each zone. In theory, the tail of a soliton extends to zero value at infinity. Because we can only use a finite computational window, we are solving for a truncated soliton. Therefore, the use of boundary condition, Eq. (40), is important to limit reflection at the boundary. As the pulse width is changing with distance traveled, the choices for an initial pulse width and the size of the computational window need careful design. Providing a numerical method can satisfactorily reproduce a stable and periodic soliton, there is no reason why such a method cannot also be used.

For mathematical simulation to be an effective tool employed to design, operate, or control a system, a set of well-proven relationships between the dependent and independent variables must be used, together with their system parameters. If not all the parametric data are known, calibration can be carried out to find the missing ones. However, it is important that the simulation must be used within a valid range of all parameters. What has been done in Section 4.6 is special because the linear relationship found in the numerical simulation is the same as the empirical Hubble law.

In Section 4, we have used numerical examples to show that electromagnetic wave can travel through sections of space that have different dispersion coefficient, but the overall wavelength change is equal to that based on the averaged coefficient (See Section 4.1). This characteristic is because soliton behave both as a wave and a particle. Similarly, we have shown that the present of other solitons (See Section 4.2), the present of a CW background (See Section 4.3), or the present of a source, do not significantly change the rate of wavelength changes against distance traveled. The last case, Section 4, is about the effect of gravity; the estimate found is approximate. But, since the space is known to be a virtual empty void even with the present of more than 200 billion galaxies, a journey through the universe would experience gravitational effect only over a miniscule portion of the total length. We can conclude that gravity will contribute to redshift but not significantly.

Hubble tension is a typical case where prediction from the standard model of cosmology is at variant to the experimentally observed empirical Hubble law [12, 13]. As some of the parameters used in the model were based on small redshift data available on the time, it not surprising to see that the model predicts differently when data for larger redshift are available. Based on the cosmological principle, the model uses a homogeneous and isotropic universe. But the real space has large clusters and voids even measured in the cosmological scale. Redshift in starlight traveling through a cluster will be quite different to the one traveling through a void.

The propagation theory used in our investigation is built on NLSE that ensures a balance between linear dispersion and nonlinear self-phase modulation. The space is never assumed as isotropic or homogeneous. By solving the NLSE, the traveling

histories are constructed according to local conditions encountered. In this approach, we can accommodate the complex nature of the universe.

6. Conclusion

1. The LCPS method is an effective and economical method to convert a differential equation into a set of nonlinear algebraic equations that can then be solved numerically.

2. The propagation theory based on the NLSE predicts redshift in electromagnetic wave through space can be calibrated to agree with Hubble law.

3. The complex nature of the universe does alter the particle nature of solitons, allowing changes during their transmission histories to be accumulated.

4. In mathematical simulations, the limits on their ranges of applicability should be recognized.

Conflict of interest

The author declares no conflict of interest.

Author details

Peter Y.P. Chen
Former School of Mechanical and Manufacturing Engineering, University of New South Wales, Sydney, NSW, Australia

*Address all correspondence to: peterypchen@yahoo.com.au

IntechOpen

© 2024 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

References

[1] Yagle AE. Differential and integral methods for three-dimensional inverse scattering problems with a non-local potential. Inverse Problems. 1988;**4**(2): 549-566. DOI: 10.1088/0266-5611/4/2/017 http://hdl.handle.net/2027.42/49089

[2] Wang M, Li X, Zhang J. The (*G*'/*G*)-expansion method and traveling wave solutions of nonlinear evolution equations in mathematical physics. Physics Letters A. 2008;**372**(4):417-423. DOI: 10.1016/j.physleta.2007.07.051

[3] Wen Z, Wang Q. Abundant exact explicit solutions to a modified cKdV equation. Journal of Nonlinear Modeling and Analysis. 2020;**2**:45-56. DOI: 10.12150/jnma.2020.45

[4] Sirisubtawee S, Koonprasert S. Exact traveling wave solutions of certain nonlinear partial differential equations using the (G^{1}'/G^{2}) expansion method. Advances in Mathematical Physics. 2018; **2018**:7628651. DOI: 10.1155/2018/7628651

[5] Kaewta S et al. Applications of the (G0/G2)-expansion method for solving certain nonlinear conformable evolution equations. Fractal and Fractional. 2021;**5**: 88. DOI: 10.3390/fractalfract5030088

[6] Andriychuk M, Yevstyhneiev B. Asymptotic solution of the scattering problem on a set of chaotic placed small particles. In: 2023 IEEE XXVIII International Seminar/Workshop on Direct and Inverse Problems of Electromagnetic and Acoustic Wave Theory (DIPED). Tbilisi, Georgia; 2023. pp. 141-144. DOI: 10.1109/DIPED59408.2023.10269484

[7] Afsari A, Abbosh A, Rahmat-Samii Y. A novel differential inverse scattering methodology in biomedical imaging. In: 2017 IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting. Piscataway, NJ, United States: IEEE. pp. 25-26. DOI: 10.1109/APUSNCURSINRSM.2017.8072055

[8] Chen PYP, Malomed BA. Lanczos–Chebyshev pseudospectral methods for wave-propagation problems. Mathematics and Computers in Simulation. 2012, 2011;**82**:1056-1068. DOI: 10.1016/j.matcom.2011.05.013

[9] Chen PYP. The Lanczos-Chebyshev pseudospectral method for solution of differential equations. Applications of Mathematics. 2020;**7**:927-938. DOI: 10.4236/am

[10] Chen PYP. A mathematical model for redshift. Applications of Mathematics. 2020;**11**:146-156. DOI: 10.4236/am.2020.113013

[11] Chen PYP. Propagation of dispersion-managed dark solitons and the novel application to redshift in starlight. Optik. 2022;**251**:168384. DOI: 10.1016/j.ijleo.2021.168384

[12] Di Valentino E. Cosmology intertwined II: The Hubble constant tension. Astroparticle Physics. 2021;**131**: 102605. DOI: 10.1016/j.astropartphys.2021.102605

[13] Hu J, Wang FY. Hubble tension: The evidence of new physics. Universe. 2023;**9**(2):94. DOI: 10.3390/universe9020094

[14] Lanczos C. Trigonometric interpolation of empirical and analytical functions. Journal of Mathematical Physics. 1938;**17**:123-199. DOI: 10.1002/sapm1938171123

[15] Lanczos C. Trigonometric interpolation of empirical and analytical functions. In: Fox L, editor. Numerical Solution of Ordinary and Partial Differential Equations. New York: Pergamon; 1962

[16] Chen PYP, Chu PL, Malomed BA. An iterative numerical method for dispersion-managed solitons. Journal of Optical Communications. 2005;**245**: 425-435. DOI: 10.1016/j.optcom.2004.10.034

[17] Koch O, Weinmüller EB. The convergence of shooting methods for singular boundary value problems. Mathematics of Computation. 2001;**72**: 289-305

[18] Malomed BA et al. Spatiotemporal solitons in multidimensional optical media with a quadratic nonlinearity. Physical Review E. 1997;**56**:4725-4736. DOI: 10.1103/PhysRevE.56.4725

[19] Chen PYP, Malomed BA. Stabilization of spatiotemporal solutions in second- harmonic-generating media. Journal of Optical Communications. 2009;**282**:3804-3811. DOI: 10.1016/j.optcom.2009.06.027

[20] Chen PYP. Investigation of nonlinear Schrödinger equation for application to astronomical redshift. Optik. 2022;**261**:169181. DOI: 10.1016/j.ijleo.2021.169181

[21] Lerner L. A simple calculation of the deflection of light in a Schwarzschild gravitational field. American Journal of Physics. 1997;**65**:1194-1196. DOI: 10.1119/1.18757

Chapter 10

Perspective Chapter: On Two-Step Hybrid Numerical-Butterfly Optimization Technique for System of Nonlinear Equations in Banach Space

Mudassir Shams and Bruno Carpentieri

Abstract

In this study, we propose a novel hybrid numerical optimization technique that combines iterative methods with a butterfly optimization scheme to solve nonlinear equations. The iterative methods, characterized by cubic convergence order, refine local solutions, while the butterfly optimization scheme enables global search. Our approach aims to improve efficiency and robustness by mitigating sensitivity to initial guesses. We conduct a local convergence analysis in Banach space and estimate convergence radii to guide the selection of initial values. The proposed technique is evaluated through engineering applications, demonstrating superior performance compared to classical methods and other optimization schemes such as particle swarm optimization, sperm swarm optimization, and ant line optimization.

Keywords: nonlinear systems of equations, two-steps hybrid schemes, optimization numerical methods, butterfly optimization schemes, applications of hybrid methods, efficiencies of solution methods

1. Introduction

The solution of systems of nonlinear equations represented as

$$F(\mathbf{x}) = \begin{cases} f_1(x_1, x_2, \ldots, x_n) = 0, \\ f_2(x_1, x_2, \ldots, x_n) = 0, \\ \vdots \\ f_n(x_1, x_2, \ldots, x_n) = 0, \end{cases} \tag{1}$$

is a critical computation required in many scientific disciplines such as physics, engineering, and biology. Nonlinear equations are essential for capturing intricate interactions between variables, often beyond the scope of linear models [1–3].

They are instrumental in describing phenomena such as chaotic systems, fluid dynamics, and quantum mechanics. Engineers utilize nonlinear models to optimize designs, simulate complex structures, and analyze control systems. In economics, nonlinear modeling aids in understanding market dynamics and economic systems with nonlinear feedback loops [4, 5]. Moreover, nonlinear equations can provide useful insights into complex biological processes such as population dynamics and enzyme kinetics [6]. Given that closed-form solutions of Eq. (1) are not always attainable [7–11], numerical methods serve as indispensable tools for approximating solutions [12–15]. Several iterative methods have been proposed in the past decades, including numerous variants of the Newton-Raphson method [16–18], see, for example, [19–23], decomposition methods [24], homotopy analysis methods [25], and references therein. Newton's method for systems of nonlinear equations is generalized as follows:

$$\mathbf{x}_i^{[1]} = \mathbf{x}_i - \frac{F(\mathbf{x}_i)}{F'(\mathbf{x}_i)}, (i = 1, 2, \ldots), F'(\mathbf{x}_i) \neq 0, \tag{2}$$

where $F{:}\Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is a Frechet-differentiable function. It is widely recognized, however, that while iterative root-finding methods can be effective, they are not without limitations [26, 27]. One significant drawback is their sensitivity to initial guesses [28]. Iterative methods typically require starting points that are sufficiently close to the actual root for reliable convergence, and often divergence may occur due to inaccurate initial estimates. One possible strategy that has been explored to enhance the robustness and efficiency of solving nonlinear equations is to combine iterative root-finding algorithms with optimization techniques, such as butterfly optimization [29] ($\mathrm{OM}^{\ominus[1*]}$), particle swarm optimization [30] ($\mathrm{OM}^{\ominus[2*]}$), sperm swarm optimization [31] ($\mathrm{OM}^{\ominus[3*]}$), ant swarm optimization [32] ($\mathrm{OM}^{\ominus[4*]}$), or others. In the first step, an optimization method is employed to search for an initial guess or an approximate solution that is close to the true root of (1). Optimization techniques excel at efficiently exploring the solution space and finding promising candidate solutions. Once the optimization phase has provided an initial guess or an approximate solution, a numerical root-finding algorithm is applied to refine the solution iteratively. This step aims to improve the accuracy and convergence of the solution obtained from the optimization phase. Iterative root-finding algorithms like Newton-Raphson, Secant method, or hybrid methods like Brent's method are commonly used for this purpose. This hybrid approach leverages the strengths of both optimization and root-finding methods to improve convergence properties and address the limitations of each technique individually. By combining optimization techniques for global search of the solution space with iterative root-finding algorithms for local refinement, the hybrid approach can potentially overcome the sensitivity to initial guesses and improve the overall efficiency and robustness of the solution process [33–39].

The effectiveness of the hybrid numerical iterative optimization technique relies heavily on appropriate parameter tuning to achieve optimal results. Difficulties may arise in achieving convergence to global minima, particularly in complex scenarios. Moreover, the suitability of this approach for high-dimensional problems and its sensitivity to errors in the data are significant considerations to take into account. Addressing convergence to local minima can pose difficulties in certain contexts. An accurate local convergence analysis is essential for evaluating the behavior and reliability of iterative root-finding algorithms, including hybrid numerical optimization

techniques. This analysis provides insights into how these methods perform when an approximate root of the equation is given, ensuring convergence within a specified neighborhood of the root [40–42].

In the context of local convergence analysis for numerical schemes used in iterative root-finding algorithms, two primary approaches are commonly employed [43–46]. The first approach involves utilizing Taylor's series expansion to generate iterations, analyzing the behavior of the algorithm near the root of the equation. This method is applicable to both scalar equations and systems of nonlinear equations, allowing for a comprehensive understanding of the local convergence properties of the numerical scheme. Direct application of Taylor's expansion for local convergence analysis may have limited applicability in the context of iterative root-finding algorithms. In many cases, iterative root-finding algorithms are designed to be computationally efficient and may not explicitly involve higher-order derivatives or rigorous Taylor series expansions in their implementation. See, for example, Refs. [3, 47, 48] and the references therein. Many iterative root-finding algorithms, such as Newton's method, primarily rely on the first derivative (or Jacobian matrix for systems of equations) for updating the solution iteratively. Despite not explicitly involving higher-order derivatives, these methods can still converge effectively in practice, even for nonlinear equations that may not be highly differentiable. Consider, for example, the function $F : [-1.5,1.5] \to \mathbb{R}$ defined as

$$F(t) = \begin{cases} at^3 \log(t) + bt^5 - ct^4, t \neq 0, \\ 0, t = 0, \end{cases} \tag{3}$$

where $a \in \mathbb{R} - \{0\}, b, c \in \mathbb{R}$ with $b + c = 0$. Then, $x^* = 1 \in \Omega$ is a root of $F(t) = 0$. However, the third derivative of the function does not exist, since this function is not continuous at $t = 0$. Examining the Jacobian matrix offers an alternative approach to local convergence analysis that can be more robust and applicable in practical settings [49, 50]. By analyzing the eigenvalues of the Jacobian matrix, this approach offers valuable information about the stability and local convergence behavior of the iterative scheme without the need for higher-order derivatives. Understanding the difference between these two approaches allows for a more thorough evaluation of the local convergence properties of numerical schemes, tailored to the problem at hand.

In this study, we propose a novel hybrid numerical optimization technique that combines iterative methods with a butterfly optimization scheme. The iterative methods, characterized by cubic convergence order [43], are employed for local refinement, while the butterfly optimization scheme facilitates global search of the solution space. A local convergence analysis is conducted in Banach space, and convergence radii are estimated to guide the selection of initial values. By leveraging the strengths of both iterative and optimization techniques, our hybrid approach aims to overcome the sensitivity to initial guesses inherent in optimization algorithms and enhance overall efficiency and robustness. We evaluate the proposed numerical schemes using engineering applications and compare their performance with conventional and optimization schemes, including the butterfly optimization technique, particle swarm optimization, sperm swarm optimization scheme, and ant line optimization technique, to illustrate their potential for addressing complex nonlinear equations in various practical domains. The study concludes with closing remarks and recommendations for further work.

2. The $MM^{\supset[1]}$ family of iterative nonlinear root-finding methods

In a previous study, we introduced a new family of methods [43] (abbreviated as $MM^{\supset[1]}$) defined by the recurrence relation:

$$\mathbf{x}_i^{[2]} = \mathbf{x}_i^{[1]} - \left(\frac{F'(\mathbf{x}_i) - F'\left(\mathbf{x}_i^{[1]}\right)}{\alpha F'\left(\mathbf{x}_i^{[1]}\right) + (2-\alpha)F'(\mathbf{x}_i)}\right)\left(\frac{F(\mathbf{x}_i)}{F'(\mathbf{x}_i)}\right), \tag{4}$$

where $\mathbf{x}_i^{[1]} = \mathbf{x}_i - \frac{F(\mathbf{x}_i)}{F'(\mathbf{x}_i)}$. The main convergence result for the iteration scheme (4) was obtained using Taylor's series and is summarized in the theorem below.

Theorem 1: Let $\mathbf{x}^* \in \Omega$ be a simple root of a sufficiently differentiable function $F : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$ in an open interval Ω. If $\mathbf{x}_0$ is sufficiently close to $\mathbf{x}^*$ then the convergence order of the family of iterative method (9) is three and the error equation is given by [43]:

$$\mathbf{e}_{i+1} = \left(2\mathbf{C}_2^2 + \frac{1}{2}\mathbf{C}_3 - \alpha\mathbf{C}_2^2\right)\mathbf{e}_i^3 + O\left(\mathbf{e}_i^4\right), \tag{5}$$

where $\mathbf{C}_m = \frac{F^m(\zeta)}{m!F'(\zeta)}, m \geq 2$.

Our previous research demonstrated that the numerical scheme (4) exhibits superior stability, consistency, and efficiency compared to established methods used for solving nonlinear equations, including those proposed by Singh et al. [51], Huen et al. [52], Amat et al. [53], Chun et al. [54], and Kou et al. [55]. Additionally, our method is designed to solve systems of nonlinear equations. It outperforms existing techniques (see, e.g., [56–58]) in terms of computational order of convergence, rate of convergence, minimizing residual error, solution time, and elapsed CPU time required to generate basins of attraction. Moreover, we extended our scheme (4) to simultaneous methods, which have proven to be more efficient than conventional simultaneous techniques in terms of computational efficiency for locating all roots of nonlinear equations [59]. Unlike some analytical methods, such as power series expansions, iterative methods for solving nonlinear equations typically do not have a well-defined radius of convergence. The convergence behavior of iterative methods can vary depending on various factors such as the initial guess, the characteristics of the function, and the specific implementation of the method. As a result, it may be difficult to determine a precise region in which convergence is guaranteed. In this study, we extend the applicability of these methods and also address the aforementioned limitations by presenting a new local convergence analysis of the $MM^{\supset[1]}$ method (4) in Banach space, which is more rigorous than the local convergence analysis of the first type presented in Ref. [43]. Banach spaces are favored over Taylor series expansions for computing local convergence because Taylor series expansions require functions to be sufficiently smooth and well-behaved within a small region around the point of interest. In contrast, Banach spaces offer a more flexible and comprehensive framework. They enable the study of convergence without requiring strict smoothness, accommodating a broader range of functions, including those that are only continuous or have limited differentiability. In this study, the radius of convergence—the distance from the root within which the iteration function can be approximated by a Taylor series expansion—is determined.

The iterative technique ensures convergence to the root if the initial estimates fall within this convergence radius.

2.1 Local convergence analysis of the MM$^{\ominus[1]}$ method in Banach space

To compute convergence radii, we establish the assumptions below. Let $\varpi_0 : [0,\infty) \to 0,\infty)$ be a continuous function. Assume

i. Equation

$$\varpi_0(t) - 1 = 0, \tag{6}$$

has a minimal zero ξ_0. Set $I_0 = [0, 2\xi_0)$. Let $\varpi : I_0 \to 0,\infty)$ be a continuous and increasing function. Define function $\S_1$onto the interval I_0 in the following way

$$\S_1(t) = \frac{\int_0^1 \varpi\big(1-\theta^{[1]}\big)td\theta^{[1]}}{1-\varpi_0(t)}. \tag{7}$$

ii. Equation

$$\S_1(t) - 1, \tag{8}$$

has a minimal zero $d_1^{[1]} \in I\backslash\{0\}$.

iii. Equation

$$\varpi_0(\S_1(t)) - 1 = 0, \tag{9}$$

has a minimal zero ξ_1. Set

$$\xi_2 = \min\{\xi_0, \xi_1\}. \tag{10}$$

Let $v : I_1 \to 0,\infty)$ be a non-decreasing function, where $I_1 = [0, \xi_2)$, and define a function $\S_2$ on the interval I_1 as follows:

$$\S_2(t) = \S_1(t) - \frac{\int_0^1 \varpi_1\big(\big(\theta^{[1]}(t)\big)td\theta^{[1]}\big)}{1-\varpi_0(t)} \frac{\varpi_1(t) + (\varpi_1(\S_1(t)))t}{\alpha\varpi_1(\S_1(t)) + 2\varpi_1(t) + \alpha\varpi_0(t)}, \tag{11}$$

where

$$P(t) = \int_0^1 \varpi\Big(\theta^{[1]}(t)\Big)td\theta^{[1]} + |2|\int_0^1 \varpi\Big(1-\theta^{[1]}\S_1(t)t\Big)d\theta^{[1]}. \tag{12}$$

iv. Equation

$$\S_2(t) - 1 = 0, \tag{13}$$

has a minimal zero $d_2^{[1]} \in I \backslash \{0\}$.
With these assumptions, the following relations are easily proven:

$$0 \le \varpi_0(t) < 1, \tag{14}$$

$$0 \le \varpi_0(\S_1(t)t) < 1, \tag{15}$$

$$0 \le \varpi_0(\S_2(t)t) < 1, \tag{16}$$

$$0 \le \S_i(t) < 1; i = 1,2,3. \tag{17}$$

The radius of convergence $d^{[1]}$ is represented as

$$d^{[1]} = \min\left\{d_m^{[1]}\right\}, m = 1, 2. \tag{18}$$

Let $\overline{B}\left(p, d^{[1]}\right)$ represent the closure of the open ball $B\left(p, d^{[1]}\right)$ with center p and radius $d^{[1]} > 0$. The convergence analysis relies on the following additional assumptions, where we assume that the ϖs are known in advance.

(C1) F: $\Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ is a Frechet differentiable operator and there exists $\mathbf{x}^* \in \Omega$ such that $F(\mathbf{x}^*) = 0$ and $F'(\mathbf{x}^*)^{-1} \in \mathfrak{L}(\mathbb{R}^n, \mathbb{R}^n)$.

(C2) There exists a continuous and non-decreasing function $\varpi_0 : R_0 \to R_0$ with $\varpi_0(0) = 0$ such that for each $x \in \Omega$:

$$\left\|F'(\mathbf{x}^*)^{-1}(F'(\mathbf{x}) - F'(\mathbf{x}^*))\right\| \le \varpi_0\|(\mathbf{x} - \mathbf{x}^*)\|. \tag{19}$$

We set $\Omega_0 = \Omega \cap B(x^*, \rho)$, where $\rho = \sup\{\gamma \in R_0 : \varpi_0(\gamma) < 1\}$.

(C3) There exist continuous and non-decreasing functions $\varpi_1, \varpi : [0, \rho) \to R_0$ with $\varpi(0) = \varpi_1(0) = 0$ such that for each $\mathbf{x}, \mathbf{x}_i^{[1]} \in \Omega_0$:

$$\begin{gathered}\left\|F'(\mathbf{x}^*)^{-1}\left(F'(\mathbf{x}) - F'\left(\mathbf{x}_i^{[1]}\right)\right)\right\| \le \varpi\left(\left\|\mathbf{x} - \mathbf{x}_i^{[1]}\right\|\right), \\ \left\|F'(\mathbf{x}^*)^{-1}F'(\mathbf{x})\right\| \le \varpi_1(\|\mathbf{x} - \mathbf{x}^*\|).\end{gathered} \tag{20}$$

(C4) For all $\mathbf{x} \in \Omega_0$

$$\left\|F'(\mathbf{x}^*)^{-1}F'(\mathbf{x})\right\| \le \varpi_1(\|\mathbf{x} - \mathbf{x}^*\|). \tag{21}$$

(C5) Define $B\left(x^*, d^{[1]}\right) \subseteq \Omega$, where the radius $d^{[1]}$ is defined by

$$d^{[1]} = \min\left\{d_1^{[1]}, d_2^{[1]}, d_3^{[1]}\right\}. \tag{22}$$

(C6) There exists $\mu > d^{[1]}$ such that

$$\int_0^1 \varpi\left(\mu d^{[1]}\right) d\mu < 1. \tag{23}$$

We show the main theorem of local convergence analysis using the assumptions and conditions stated in this section. The theorem computes the convergence radius and guides the selection of initial guess values to assure convergence to the exact root.

Theorem 2 Under the condition (C1)-(C6) for $d^{[1]} = r$, further suppose that $\mathbf{x}_0 \in B\left(\mathbf{x}^*, d^{[1]}\right) - \{\mathbf{x}^*\}$. The sequence $\{\mathbf{x}_i\}$ generated for point $\mathbf{x}_0$ by the iterative scheme $\text{MM}^{\supset^{[1]}}$ is well defined, remains in $B\left(\mathbf{x}^*, d^{[1]}\right)$ for each $i = 0,1,2, \ldots$ and converges to $\mathbf{x}^*$. Moreover the following assertions hold:

$$\left\|\mathbf{x}_i^{[1]} - \mathbf{x}^*\right\| \leq \S_1(\|\mathbf{x}_i - \mathbf{x}^*\|)\|\mathbf{x}_i - \mathbf{x}^*\| \leq \|\mathbf{x}_i - \mathbf{x}^*\| < d^{[1]}, \tag{24}$$

$$\left\|\mathbf{x}_i^{[2]} - \mathbf{x}^*\right\| \leq \S_2(\|\mathbf{x}_i - \mathbf{x}^*\|)\|\mathbf{x}_i - \mathbf{x}^*\| \leq \|\mathbf{x}_i - \mathbf{x}^*\| < d^{[1]}, \tag{25}$$

where the § functions are given previously, and $d^{[1]}$ is defined by

$$d^{[1]} = \min\left\{d_1^{[1]}, d_2^{[1]}\right\}. \tag{26}$$

Furthermore, $\mathbf{x}^*$ is the only solution of $F(\mathbf{x}) = 0$ given by (C6).

Proof:

Suppose that all conditions hold. Then for $\mathbf{x}_0 \in B\left(\mathbf{x}^*, d^{[1]}\right) - \{\mathbf{x}^*\}$, the sequence $\{\mathbf{x}_i\}$ generated by the $\text{MM}^{\supset^{[1]}}$ method is well defined in $B\left(\mathbf{x}^*, d^{[1]}\right)$, remains in $B\left(\mathbf{x}^*, d^{[1]}\right)$ for each $i = 0,1,2, \ldots$ and converges to $\mathbf{x}^*$. Let $\mathbf{x} \in B\left(\mathbf{x}^*, d^{[1]}\right)$. In view of the condition (C2),

$$\left\|F'(\mathbf{x}^*)^{-1}(F'(\mathbf{x}_i) - F'(\mathbf{x}^*))\right\| \leq \varpi_0(\|\mathbf{x}_i - \mathbf{x}^*\|) \leq \varpi_0\left(d^{[1]}\right) < 1. \tag{27}$$

The existence of an invertible operator in Banach space implies that $F'(\mathbf{x}_i)^{-1} = \left\|F'(\mathbf{x}_i)^{-1}F'(\mathbf{x}^*)\right\| \leq \frac{1}{\varpi(\|\mathbf{x}_i - \mathbf{x}^*\|)}$. So,

$$\begin{aligned}
\left\|F'(\mathbf{x}_i)^{-1}(F'(\mathbf{x}_i) - F'(\mathbf{x}^*))\right\| &\leq \varpi_0(\|\mathbf{x}_i - \mathbf{x}^*\|),\\
\left\|F'(\mathbf{x}_i)^{-1}F'(\mathbf{x}_i)\right\| + \left\|F'(\mathbf{x}_i)^{-1}F'(\mathbf{x}^*))\right\| &\leq \varpi_0(\|\mathbf{x}_i - \mathbf{x}^*\|),\\
\left\|F'(\mathbf{x}_i)^{-1}F'(\mathbf{x}^*))\right\| + 1 &\leq \varpi_0(\|\mathbf{x}_i - \mathbf{x}^*\|),\\
\left\|F'(\mathbf{x}_i)^{-1}F'(\mathbf{x}^*))\right\| &\leq \varpi_0(\|\mathbf{x}_i - \mathbf{x}^*\|) - 1,\\
\left\|F'(\mathbf{x}_i)F'(\mathbf{x}^*)^{-1}\right\| &\leq \frac{1}{1 - \varpi_0(\|\mathbf{x}_i - \mathbf{x}^*\|)},
\end{aligned}$$

and

$$\mathbf{x}_i^{[1]} = \mathbf{x}_i - \frac{F(\mathbf{x}_i)}{F'(\mathbf{x}_i)}, \tag{28}$$

$$\mathbf{x}_i^{[1]} = \mathbf{x}_i - F'(\mathbf{x}_i)^{-1}F(\mathbf{x}_i), \forall i = 1, 2, \ldots \tag{29}$$

Using the Mean Value Theorem (MVT) in integral from,

$$\frac{F(\mathbf{x}_i) - F(\mathbf{x}^*)}{\mathbf{x}_i - \mathbf{x}^*} = \int_0^1 F'\Big(\mathbf{x}^* + \theta^{1]}(\mathbf{x}_i - \mathbf{x}^*)\Big)d\theta^{1]}, \tag{30}$$

we can compute $F'(\mathbf{x}_i)^{-1}F(\mathbf{x}_i)$ by choosing $|F'(\mathbf{x})| \leq M$ for all $\mathbf{x} \in D$ and $\left\|F(\mathbf{x}) - F\left(\mathbf{x}_i^{[1]}\right)\right\| \leq c\left\|\mathbf{x} - \mathbf{x}_i^{[1]}\right\|$. Thus, we can write

$$\begin{aligned}
F(\mathbf{x}_i) &= \int_0^1 F'\Big(\mathbf{x}^* + \theta^{[1]}(\mathbf{x}_i - \mathbf{x}^*)\Big)d\theta^{[1]}(\mathbf{x}_i - \mathbf{x}^*), \\
F(\mathbf{x}_i) &= \begin{pmatrix} \int_0^1 [F'\big(\mathbf{x}^* + \theta^{[1]}(\mathbf{x}_i - \mathbf{x}^*)\big) \\ -F'(\mathbf{x}_0)]d\theta^{[1]} + \int_0^1 F'(\mathbf{x}_i)d\theta^{[1]} \end{pmatrix}(\mathbf{x}_i - \mathbf{x}^*), \\
F(\mathbf{x}_i)F'(\mathbf{x}_i)^{-1} &= \begin{pmatrix} \int_0^1 F'(\mathbf{x}_i)^{-1}[F'\big(\mathbf{x}^* + \theta^{[1]}(\mathbf{x}_i - \mathbf{x}^*)\big) \\ -F'(\mathbf{x}_i)]d\theta^{[1]}(\mathbf{x}_i - \mathbf{x}^*) + (\mathbf{x}_i - \mathbf{x}^*) \end{pmatrix}.
\end{aligned} \tag{31}$$

On the other hand, using $F(\mathbf{x}_i)F'(\mathbf{x}_i)^{-1}$ in (36) we have

$$\mathbf{x}_i^{[1]} - \mathbf{x}^* = \mathbf{x}_i - \mathbf{x}^* - (\mathbf{x}_i - \mathbf{x}^*) - \begin{pmatrix} \int_0^1 F'(\mathbf{x}_i)^{-1}[F'\big(\mathbf{x}^* + \theta^{[1]}(\mathbf{x}_i - \mathbf{x}^*)\big) \\ -F'(\mathbf{x}_i)]d\theta^{[1]}(\mathbf{x}_i - \mathbf{x}^*) \end{pmatrix}, \tag{32}$$

$$\begin{aligned}
\|\mathbf{x}_i - \mathbf{x}^*\| &\leq \left\|\int_0^1 F'(\mathbf{x}_i)^{-1}\Big[F'\Big(\mathbf{x}^* + \theta^{[1]}(\mathbf{x}_i - \mathbf{x}^*)\Big) - F'(\mathbf{x}_i)\Big]d\theta^{[1]}\right\| \times \|\mathbf{x}_i - \mathbf{x}^*\|, \\
\|\mathbf{x}_i - \mathbf{x}^*\| &\leq \begin{pmatrix} \left\|F'(\mathbf{x}_i)^{-1}F'(\mathbf{x}^*)\right\| \times \\ \left\|\int_0^1 \begin{bmatrix} F'(\mathbf{x}^*)^{-1}[F'(\mathbf{x}^* + \\ \theta^{[1]}(\mathbf{x}_i - \mathbf{x}^*)) - F'(\mathbf{x}_0)] \end{bmatrix} d\theta^{[1]}\right\| \end{pmatrix} \times \|\mathbf{x}_i - \mathbf{x}^*\|, \\
\|\mathbf{x}_i - \mathbf{x}^*\| &\leq \begin{pmatrix} \frac{\left\|\int_0^1 \varpi(\mathbf{x}^* + \theta^{[1]}(\mathbf{x}_i - \mathbf{x}^*) - \mathbf{x}_i)d\theta^{[1]}\right\|}{1 - \varpi_0(\|\mathbf{x}_i - \mathbf{x}^*\|)} \\ \times\|\mathbf{x}_i - \mathbf{x}^*\| \end{pmatrix}, \\
\|\mathbf{x}_i - \mathbf{x}^*\| &\leq \frac{\int_0^1 \varpi\big((1 - \theta^{[1]})\|\mathbf{x}_i - \mathbf{x}^*\|\big)d\theta^{[1]}}{1 - \varpi_0(\|\mathbf{x}_0 - \mathbf{x}^*\|)} \times \|\mathbf{x}_i - \mathbf{x}^*\|, \\
\|\mathbf{x}_i - \mathbf{x}^*\| &\leq \S_1(\|\mathbf{x}_i - \mathbf{x}^*\|)\|\mathbf{x}_0 - \mathbf{x}^*\| \leq \|\mathbf{x}_i - \mathbf{x}^*\| \leq d_1^{[1]}, \\
\mathbf{x}_i^{[1]} &\in B(\mathbf{x}^*, d) \text{and} \\
\S_1(\|\mathbf{x}_i - \mathbf{x}^*\|) &= \frac{\int_0^1 \varpi\big((1 - \theta^{[1]})\|\mathbf{x}_i - \mathbf{x}^*\|\big)d\theta^{[1]}}{1 - \varpi_0(\|\mathbf{x}_i - \mathbf{x}^*\|)} \leq d_1^{[1]}.
\end{aligned} \tag{33}$$

We now consider the second step of the numerical scheme $MM^{\circlearrowright^{[1]}}$. We have:

$$\mathbf{x}_i^{[2]} = \mathbf{x}_i^{[1]} - \left(\frac{F'(\mathbf{x}_i) - F'\left(\mathbf{x}_i^{[1]}\right)}{\alpha F'\left(\mathbf{x}_i^{[1]}\right) + (2-\alpha)F'(\mathbf{x}_i)}\right)\left(\frac{F(\mathbf{x}_i)}{F'(\mathbf{x}_i)}\right), \tag{34}$$

$$\mathbf{x}_i^{[2]} - \mathbf{x}^* = \mathbf{x}_i^{[1]} - \mathbf{x}^* - \left(\frac{F'(\mathbf{x}_i) - F'\left(\mathbf{x}_i^{[1]}\right)}{\alpha F'\left(\mathbf{x}_i^{[1]}\right) + (2-\alpha)F'(\mathbf{x}_i)}\right)\left(\frac{F(\mathbf{x}_i)}{F'(\mathbf{x}_i)}\right), \tag{35}$$

$$\left\|\mathbf{x}_i^{[2]} - \mathbf{x}^*\right\| = \left\|\mathbf{x}_i^{[1]} - \mathbf{x}^*\right\| - \left(\frac{F'(\mathbf{x}_i) - F'\left(\mathbf{x}_i^{[1]}\right)}{\alpha F'\left(\mathbf{x}_i^{[1]}\right) + (2-\alpha)F'(\mathbf{x}_i)}\right)\left(\frac{F(\mathbf{x}_i)}{F'(\mathbf{x}_i)}\right). \tag{36}$$

Let $A_0 = \alpha F'\left(\mathbf{x}_i^{[1]}\right) + (2-\alpha)F'(\mathbf{x}_i),$ Then, established the invertibility of $A_0,$ that is, we have

$$\left\|\mathbf{x}_i^{[2]} - \mathbf{x}^*\right\| \leq \left\|\mathbf{x}_i^{[1]} - \mathbf{x}^*\right\| - \frac{\int_0^1 \varpi_1\left(\theta^{[1]}(\|\mathbf{x}_i - \mathbf{x}^*\|)d\theta^{[1]}\|\mathbf{x}_i - \mathbf{x}^*\|\right)}{1 - \varpi_0(\|\mathbf{x}_0 - \mathbf{x}^*\|)}\left(\frac{\left(F'(\mathbf{x}_i) - F'\left(\mathbf{x}_i^{[1]}\right)\right)\left(F'(\mathbf{x}^*)^{-1}F'(\mathbf{x}^*)\right)}{A_0}\right) \tag{37}$$

$$A_0 = \alpha F'\left(\mathbf{x}_i^{[1]}\right) + (2-\alpha)F'(\mathbf{x}_i), \tag{38}$$

and we can write

$$\begin{aligned}
A_0F'(\mathbf{x}^*)^{-1} &= \left(\alpha F'\left(\mathbf{x}_i^{[1]}\right) + (2-\alpha)F'(\mathbf{x}_i)\right)F'(\mathbf{x}^*)^{-1} \\
&= \left(\alpha F'\left(\mathbf{x}_i^{[1]}\right) - \alpha F'(\mathbf{x}^*) + (2-\alpha)F'(\mathbf{x}_i) + \alpha F'(\mathbf{x}^*)\right)F'(\mathbf{x}^*)^{-1} \\
&= \left(\alpha\left(F'\left(\mathbf{x}_i^{[1]}\right) - F'(\mathbf{x}^*)\right) + 2F'(\mathbf{x}_i) - \alpha(F'(\mathbf{x}_i) + F'(\mathbf{x}^*))\right)F'(\mathbf{x}^*)^{-1} \\
&= \alpha F'(\mathbf{x}^*)^{-1}\left(\left(F'\left(\mathbf{x}_i^{[1]}\right) - F'(\mathbf{x}^*)\right) + 2F'(\mathbf{x}_i)F'(\mathbf{x}^*)^{-1} - \alpha F'(\mathbf{x}^*)^{-1}(F'(\mathbf{x}_i) + F'(\mathbf{x}^*))\right)F'(\mathbf{x}^*)^{-1} \\
A_0F'(\mathbf{x}^*)^{-1} &= \alpha\varpi_0\left(\left\|\mathbf{x}_i^{[1]} - \mathbf{x}^*\right\|\right) + 2\varpi_1(\|\mathbf{x}_i - \mathbf{x}^*\|) + \alpha\varpi_0(\|\mathbf{x}_i - \mathbf{x}^*\|).
\end{aligned} \tag{39}$$

It follows

$$\left\|\mathbf{x}_i^{[2]} - \mathbf{x}^*\right\| \leq \left\|\mathbf{x}_i^{[1]} - \mathbf{x}^*\right\| - \left(\frac{\int_0^1 \varpi_1\left(\theta^{[1]}(\|\mathbf{x}_i - \mathbf{x}^*\|)d\theta^{[1]}\|\mathbf{x}_i - \mathbf{x}^*\|\right)}{1 - \varpi_0(\|\mathbf{x}_0 - \mathbf{x}^*\|)}\right)\left(\frac{(\varpi_1(\|\mathbf{x}_i - \mathbf{x}^*\|))\left(\varpi_1\left(\left\|\mathbf{x}_i^{[1]} - \mathbf{x}^*\right\|\right)\right)}{A_0}\right), \tag{40}$$

$$\left\|\mathbf{x}_i^{[2]} - \mathbf{x}^*\right\| \leq \S_2(\|\mathbf{x}_i - \mathbf{x}^*\|)\|\mathbf{x}_i - \mathbf{x}^*\| \leq \|\mathbf{x}_i - \mathbf{x}^*\| \leq d_2^{[1]}, \tag{41}$$

$$\left\|\mathbf{x}_i^{[2]} - \mathbf{x}^*\right\| \leq d_2^{[1]}, \tag{42}$$

$$\mathbf{x}_i^{[2]} \in B\left(\mathbf{x}^*, d_2^{[1]}\right), \tag{43}$$

where

$$\S_2(t) = \S_1(t) - \frac{\int_0^1 \varpi_1\left(\left(\theta^{[1]}(t)\right) t d\theta^{[1]}\right)}{1 - \varpi_0(t)} \frac{\varpi_1(t) + (\varpi_1(\S_1(t)))t}{\alpha\varpi_1(\S_1(t)) + 2\varpi_1(t) + \alpha\varpi_0(t)} \tag{44}$$

and $t = |\mathbf{x}_i - \mathbf{x}^*|$. ■

The proved local convergence Theorem 2 provides important insights into the behavior of iterative schemes $\mathrm{MM}^{\circlearrowright[1]}$ for solving system of nonlinear equations. Under the assumption that the function is locally Lipschitz continuous and the initial guess is close enough to the solution, the iterations will converge to the solution within a specific neighborhood of the initial approximation. In practice, local convergence analysis in Banach space guides the choice of starting points and guides in the estimation of the rate at which classical and hybrid numerical optimization schemes will converge to exact solutions of (1), as explained in more detail in the subsequent sections.

3. Butterfly optimization scheme $\mathrm{OM}^{\circlearrowright[\mathbf{1}*]}$

The Butterfly optimization algorithm (BOA) is a powerful metaheuristic optimization technique that has demonstrated to be effective in tackling complex optimization problems across diverse domains. Inspired by the foraging behavior of butterflies in nature, BOA simulates the process of scent-based communication among butterflies to guide the search for optimal solutions. Central to the BOA is the concept of scent emission and detection, which serves as the primary mechanism for guiding the movement of butterflies within the solution space. The algorithm operates in three main stages: initialization, iteration, and optimization.

- **Initialization**. The initialization stage involves setting up algorithmic parameters and generating an initial population of candidate solutions. This population is typically randomly generated within the solution boundaries, ensuring diversity in the initial set of solutions.

- **Iteration**. During the iteration phase, butterflies explore the solution space by emitting and detecting scents. Each butterfly emits a scent that attracts other butterflies toward it, with the strength of the scent determined by the objective function value associated with the butterfly's current position. Butterflies may also engage in random exploration or move toward butterflies emitting stronger scents.

- **Optimization**. The optimization phase focuses on iteratively refining the solutions toward the global optimum. By continuously updating the scent intensities and adjusting the movements of butterflies based on the objective function evaluations, BOA attempts to converge toward high-quality solutions.

During the iteration phase, each butterfly's stimulus intensity ($\breve{I}$) is set based on the objective function assessment. The stimulus intensity, indicated as $\breve{I}$, for the i^{th} butterfly can be computed as:

$$\breve{I} = f\left(\mathbf{x}_i^{[\vartheta]}\right),$$

where $f(\cdot)$ represents the objective function, ϑ represents the iteration, and $\mathbf{x}_i^{[\vartheta]}$ is the i^{th} butterfly in the population. The intensity of the smell that each firefly detects—called σ_i—can be calculated as a function of $\breve{I}$ in the following way:

$$\sigma_i = \breve{C}\breve{I}^{\eth}.$$

The parameter $\breve{C}$ represents the sensory scent, and $\eth$ stands for the power exponent, so that $\breve{C} \in [0, 1]$. If $\eth = 0$, no other butterfly is able to detect the scent of another butterfly, and if $\eth = 1$, no fragrance is absorbed. Each butterfly in the population travels in a specific route; a global or local search is done in this respect. The global search phase is executed with the expression

$$\mathbf{x}_i^{[\vartheta+1]} = \mathbf{x}_i^{[\vartheta]} + \left(\delta^2 \times Q^* - \mathbf{x}_i^{[\vartheta]}\right)\sigma_i,$$

where Q^* signifies the butterfly that released the strongest scent, achieving the best objective function value, and $\delta \in [0, 1]$ is a uniformly generated random number. The local search is carried out as:

$$\mathbf{x}_i^{[\vartheta+1]} = \mathbf{x}_i^{[\vartheta]} + \left(\delta^2 \times \mathbf{x}_j^{[\vartheta]} - \mathbf{x}_k^{[\vartheta]}\right)\sigma_i.$$

In the solution space, the j^{th} and k^{th} butterflies are randomly selected as $\mathbf{x}_j^{[\vartheta]}$ and $\mathbf{x}_k^{[\vartheta]}$, respectively. Consequently, the i^{th} butterfly's motion is explained as:

$$\mathbf{x}_i^{[\vartheta+1]} = \begin{cases} \mathbf{x}_i^{[\vartheta]} + \left(\delta^2 \times Q^* - \mathbf{x}_i^{[\vartheta]}\right)\sigma_i, \text{for } \delta_p < p \\ \mathbf{x}_i^{[\vartheta]} + \left(\delta^2 \times \mathbf{x}_j^{[\vartheta]} - \mathbf{x}_k^{[\vartheta]}\right)\sigma_i, \text{otherwise}. \end{cases}$$

The BOA method applies a dynamic balance between extensive local search and common global search, guided by a probability parameter p typically ranging between 0 and 1. This parameter determines the likelihood of a butterfly engaging in either local or global exploration during each iteration of the algorithm. The iteration phase of BOA continues until one of the termination criteria is met. These criteria may include reaching a specified number of iterations, exceeding a predetermined error threshold, or utilizing a predetermined amount of CPU time. In the final stage of the algorithm, the solution with the highest fitness calculation, as evaluated by the objective function, is selected as the output solution.

Through the interplay of scent-based communication and stochastic movement, BOA facilitates the discovery of optimal or near-optimal solutions to challenging optimization problems. In practice, BOA has been successfully applied to a wide range of optimization tasks, including engineering design, scheduling, machine learning, and robotics. Its versatility, robustness, and computational efficiency make it a valuable tool for addressing real-world optimization problems. However, as the no-free-lunch theorem asserts [60], no single algorithm is universally optimal for solving all problems. While the BOA excels at efficiently exploring solution spaces and finding satisfactory solutions for nonlinear systems of equations, it does not guarantee globally optimal solutions. The effectiveness of BOA depends on several factors, including the selection of optimal parameters, initialization procedures, and the complexity of the equations being solved. The algorithm may encounter difficulties when confronted with highly nonlinear systems, complex optimization landscapes, or high-dimensional solution spaces. Additionally, the computational cost associated with BOA can be prohibitive, especially for large-scale problems.

To address these limitations, in the next section we propose the implementation of a hybrid numerical technique that combines $OM^{\circlearrowright[1*]}$ with our recently developed $MM^{\circlearrowright[1]}$. By integrating the characteristics of both approaches, our hybrid method aims to improve convergence rates, mitigate divergence issues, and increase the likelihood of finding globally optimal or near-optimal solutions.

4. Hybrid numerical butterfly optimization scheme ($MM^{\circlearrowright[2]}$)

The hybrid numerical Butterfly optimization strategy employs a two-step numerical iterative scheme $MM^{\circlearrowright[1]}$ in each iteration. First, the $OM^{\circlearrowright[1*]}$ method is used to determine the optimal butterfly position (BFP). Using this BFP as input in a two-step numerical iterative technique allows us to refine the starting position to the desired precision. The initial location is iteratively fed into the two-step process $MM^{\circlearrowright[1]}$, which results in significantly enhanced butterfly locations. The framework is shown in **Figure 1**. Combining $OM^{\circlearrowright[1*]}$ and $MM^{\circlearrowright[1]}$ improves the solution of the nonlinear system of equations under the conditions of the local convergence theorem in Banach space (LCT-IIB) derived in the previous section. Finally, the $MM^{\circlearrowright[1]}$ produces refined butterfly positions.

In developing our hybrid numerical optimization scheme for solving nonlinear systems of equations, we leverage the power of iterative refinement to enhance robustness and efficiency. Our approach is based on three key principles:

- **Iterative Refinement of Solutions:** We employ an iterative technique that iteratively refines and improves the solution obtained. By leveraging local convergence assumptions in Banach space, we ensure rapid progress toward optimal solutions, enhancing the overall efficiency of the optimization process.

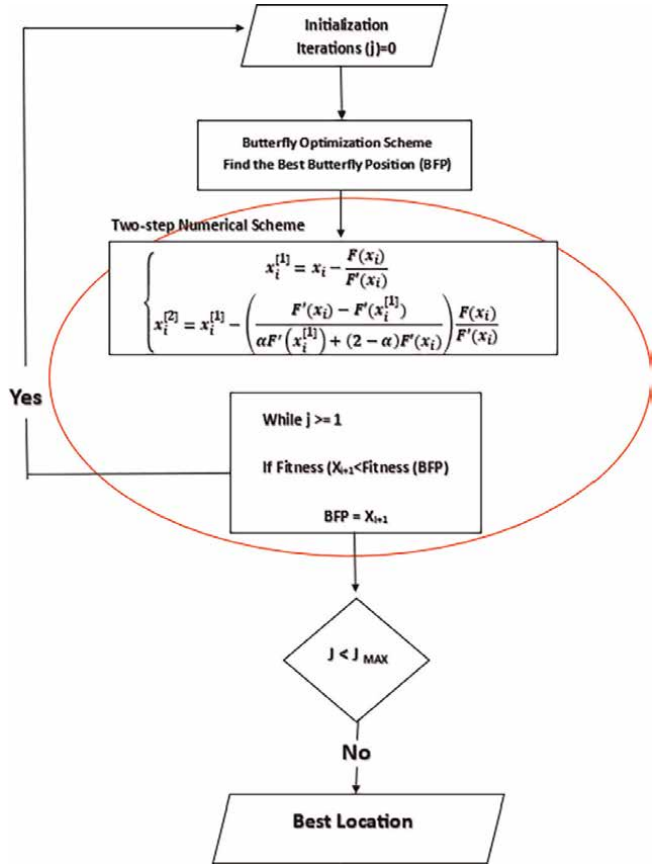


Figure 1.
The framework of the hybrid numerical optimization scheme $MM^{\circ[2]}$ for solving Eq. (1).

- **Adaptability to Complex Problem Landscapes:** Real-world optimization problems often feature complex and dynamic landscapes characterized by multiple local minima and maxima. Our iterative approach excels in navigating such intricate landscapes by adapting to changing conditions and exploring diverse regions of the solution space in each iteration.

- **Sensitivity to Initial Guesses:** Our iterative approach is carefully designed to mitigate sensitivity to initial guesses over successive iterations. By iteratively refining solutions and gradually reducing dependence on initial conditions, we enhance the robustness of the optimization process and ensure consistent performance across different starting points.

In summary, our hybrid numerical optimization scheme combines the power of iterative refinement with adaptability to complex landscapes and reduced sensitivity to initial guesses, aiming to deliver a robust and efficient approach for solving nonlinear systems of equations. Algorithm 1 describes the whole procedure.

Algorithm 1: Hybrid Numerical Butterfly Optimization Scheme ($MM^{\supset[2]}$) for solving of nonlinear system of equations.

Step 1: Inputs

•**"System of Equations"**: The system of nonliner equations

•**"System of Equations"**: Initial gusses for the solution vector

•**"Convergence Criteria"**: Threshold for convergence

•**"Maximum Iterationss"**: Maximum number of iterations

Step 2: Initialization

• Initialize the solution vector "**x**" with "**InilizalSolution**"

• Set iteration "**i**" to 0

Step 3: Main Loop

While "**i**" is less than " **MaxIterations**"

Perform Butterfly Optimazation Techanique

Call "x=ButterflyExplore(X)"

Find Best Butterfly Position (BFP)

Perform Two-Step Numerical Scheme using BFP as an input

Call "x=Two_Step_Numerical_Scheme(System _of_Nonlinear_Eq,x)"

For iteration=1:MazIteration

F(x)=Objective_function()%nonlinar system of equation

$F'(\mathbf{x})$=J=Jacobian(Solution)

$$\mathbf{x}_i^{[1]} = \mathbf{x}_i - \frac{F(\mathbf{x}_i)}{F'(\mathbf{x}_i)}$$

$$\mathbf{x}_i^{[2]} = \mathbf{x}_i^{[1]} - \left(\frac{F'(\mathbf{x}_i)-F'(\mathbf{x}_i^{[1]})}{\alpha F'(\mathbf{x}_i^{[1]})+(2-\alpha)F'(\mathbf{x}_i)}\right)\left(\frac{F(\mathbf{x}_i)}{F'(\mathbf{x}_i)}\right),$$

End do

Check Convergence

• If "Convergene (System_of_Eq,x,Convergence_Criteria)" hold

Exist Loop

Incriment the iterations counter i.e.,

• If "**i=i+1**"

End do

Step 4: Output

Return the final solution Vector "**X**"

End Computer Parogramme

5. Dynamical analysis of the $MM^{\supset[1]}$ and $MM^{\supset[2]}$ methods

In the context of nonlinear root-finding algorithms, the dynamical plane refers to a geometric representation used to study the behavior and convergence properties of iterative methods for finding roots of nonlinear equations. The dynamical plane typically consists of a coordinate system where each point represents an initial guess for the root, and the trajectory of each point under the iteration process illustrates how the algorithm converges (or diverges) toward the actual root. By studying the patterns of convergence or divergence in the dynamical plane, researchers can gain insights into the efficiency, stability, and robustness of the root-finding algorithm under consideration. On the other hand, the basin of attraction refers to the region of the dynamical plane where initial guesses converge to a specific root under the iteration process. Each root of the equation typically has its own basin of attraction, and points

within each basin converge to that particular root. The basin of attraction helps visualize which initial guesses will lead to convergence to a particular root, and which will not. In our dynamical analysis, we will analyze in particular the following aspects.

Convergence patterns. Basins of attraction illustrate how iteration can diverge or converge depending on the initial values. Observing how points in the dynamical plane approach roots (or fail to approach them) over iterations helps identify whether convergence is smooth, erratic, or exhibits oscillatory behavior. A wider basin indicates that the method is more likely to converge from a larger set of initial values.

Stability. The shape and size of the basins of attraction can indicate the stability of iterative solutions. Greater stability is often suggested by larger, more uniformly formed basins. Fractured or inconsistent basins, on the other hand, may indicate regions of instability or sensitivity to initial conditions. Stability refers to the behavior of the algorithm when subjected to small perturbations, such as rounding errors or noise in the system. Understanding the stability of the algorithm helps ensure reliable performance in practical applications.

Rate of convergence. A faster rate of convergence means fewer iterations are needed to obtain an accurate result. The convergence behavior varies across the solution space [61, 62]. Basins of attraction can provide insights into the speed of convergence of iterative methods. Methods with basins that quickly attract trajectories to their respective attractors indicate faster convergence rates, while methods with slower convergence may have smaller basins or regions of slower convergence. Based on this analysis, we may choose the best numerical method for solving a nonlinear problem.

Sensitivity on initial guess. Different initial guesses can lead to convergence to different roots or no convergence at all. Examining how sensitive the algorithm is to variations in initial guesses provides insight into its robustness and reliability in finding solutions. If, for a particular choice of starting values, the iterative method fails to generate basins of attraction after a predetermined number of iterations, it is sensitive to the initial values.

Percentage convergence or divergence. Computing the percentage convergence or divergence of an iterative method using basins of attraction involves analyzing the distribution of initial conditions within the dynamical plane, and determining the proportion of points that converge or diverge under iteration and multiply the result by 100.

We generate the basins of attraction using a 800×800 grid of squares $[-8, 8] \times [-8, 8]$ in the complex plane. If the iterative technique $\mathrm{MM}^{\circlearrowright[1]}$-$\mathrm{MM}^{\circlearrowright[2]}$ approximates the root within 20 iterations and satisfies the tolerance $\|x_{i+1} - x_i\| < 10^{-3}$, each root is assigned a color; otherwise, a dark red color is used. Using a larger radius for the numerical schemes generated from the preceding local convergence (see **Table 1**), basin color brightness, and smoothness results in fewer iterations and faster numerical scheme convergence. We develop an efficient numerical approach based on the percentage convergence divergence in the basins of attraction.

Below, we illustrate the basins of attractions for the nonlinear function F on $\Omega = x_i^{[1]} = \mathbb{R}$ and $\Omega\left[-\frac{1}{2}, \frac{3}{2}\right]$ defined as:

Method	$d_1^{[1]}$	$d_2^{[1]}$	$\min\left\{d_1^{[1]}, d_2^{[1]}\right\}$
$\mathrm{MM}^{\circlearrowright[1]}$	0.006896313920	0.00307650880	0.00307650880

Table 1.
Comparison of convergence radii for numerical scheme $\mathrm{MM}^{\circlearrowright[1]}$.

Method	Els-Time	Average-It	T-Points	D-Points	C-Points
Analysis of the dynamical planes using different convergent-divergent points					
$MM^{\supset[1]}$	0.0124010	19.0	640,000	1500.1256	638499.8744
$MM^{\supset[2]}$	2.2500124	17.0	640,000	650.0013	639349.9987
Analysis using condition of the LCT-IIB					
$MM^{\supset[1]}$	0.145213	19.0	640,000	703.51	639296.49
$MM^{\supset[2]}$	3.545213	16.0	640,000	132.5463	639867.4537

Table 2.
Dynamical study of numerical schemes $MM^{\supset[1]} - MM^{\supset[2]}$.

$$F(x) = \begin{cases} x \ln x^2 + x^5 - x^4, x \neq 0, \\ 0, x = 0, \end{cases} \tag{45}$$

for $x^* = 1$. We set $\varpi_0(\gamma) = 96.67\gamma$, $\varpi(\gamma) = 96.67\gamma$, $\varpi_1(\gamma) = 2$. **Table 1** summarizes our analysis of convergence radii.

In **Table 2**, D-points indicates the diverging points, C-Points denotes the converging points, Els-Time reports the elapsed time in seconds, and Average-It is the average number of iterations required to converge to the exact root. **Table 2** indicates that $MM^{\supset[2]}$ outperforms $MM^{\supset[1]}$ in terms of converging points utilizing the local convergence theorem's criteria and has a faster convergence rate. **Figure 2** depicts dynamic patterns corresponding to (45) utilizing $MM^{\supset[1]}$ and $MM^{\supset[2]}$. Brighter hues in basins of attraction imply fewer iterations needed to converge to exact roots. The percentage convergence of $MM^{\supset[1]}$ and $MM^{\supset[2]}$ is 99.7 and 99.89 percent, respectively, without utilizing LCT-IIB, and increases to 99.89 and 99.97 percent with LCT-IIB.

6. Numerical outcomes

In order to estimate the computational order of convergence $\left(CO^{\vartheta[1]}\right)$, the approximate computational order of convergence $\left(CO^{\vartheta[2]}\right)$, and the local computational order of convergence $\left(CO^{\vartheta[3]}\right)$ of our method, we compute the following quantities:

$$CO^{\vartheta[1]} = \frac{\ln\left(\left\|\frac{\mathbf{x}_{i+1}-\mathbf{x}^*}{\mathbf{x}_i-\mathbf{x}^*}\right\|\right)}{\ln\left(\left\|\frac{\mathbf{x}_i-\mathbf{x}^*}{\mathbf{x}_{i-1}-\mathbf{x}^*}\right\|\right)}; CO^{\vartheta[12]} = \frac{\ln\left(\left\|\frac{\mathbf{x}_{i+1}-\mathbf{x}_i}{\mathbf{x}_i-\mathbf{x}_{i-1}}\right\|\right)}{\ln\left(\left\|\frac{\mathbf{x}_i-\mathbf{x}_{i-1}}{\mathbf{x}_{i-1}-\mathbf{x}_{i-2}}\right\|\right)}; CO^{\vartheta[3]} = \frac{\ln(\|\mathbf{x}_{i+1}-\mathbf{x}^*\|)}{\ln(\|\mathbf{x}_i-\mathbf{x}^*\|)}.$$

Note that we use $CO^{\vartheta[1]}$-$CO^{\vartheta[3]}$ to compute the convergence order without considering the function's higher-order derivative. **Tables 3–12** present estimates of the convergence radii for numerical examples 1–3 using local convergence theorem conditions. In all Tables F^{T^*} represents the fitness function:

$$F^{T*} = \sqrt{\left(f_1(x_1, \cdots, x_n)\right)^2 + \cdots + \left(f_n(x_1, \cdots, x_n)\right)^2}$$

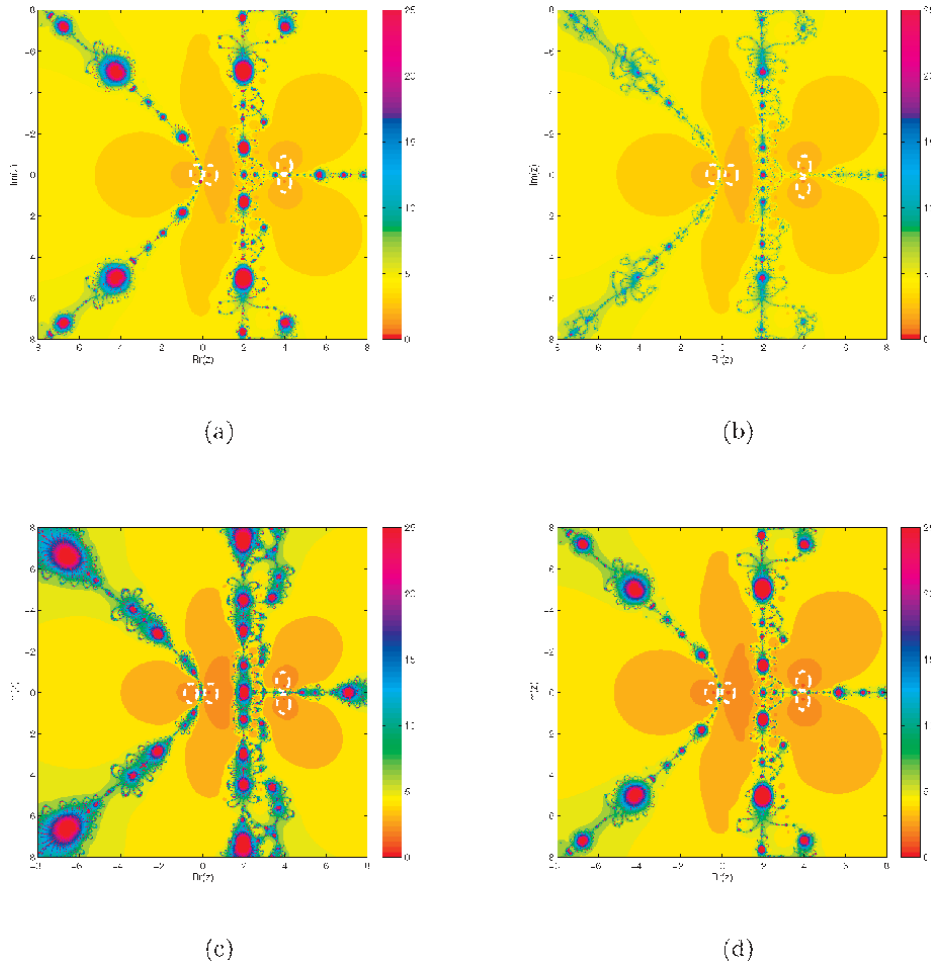


Figure 2.
(a–d) Basins of attraction of $MM^{\circlearrowleft[1]}$-$MM^{\circlearrowleft[2]}$ for (45). Figure (a,b)-shows the basins of attraction of $MM^{\circlearrowleft[1]}$-$MM^{\circlearrowleft[2]}$ without using the assumption LCT-IIB and Figure (c,d)-basins of attraction of $MM^{\circlearrowleft[1]}$-$MM^{\circlearrowleft[2]}$ using the assumption of LCT-IIB. In basins of attractions, the white circle represents the roots of (45). The color brightening and wide regular shape of (c,d) show the less number of iterations and are more stable than (a,b).

which measures the accuracy of the optimization scheme without using the assumptions of LCT-IIB to solve (46). The Euclidean norm (or norm-2) of $F(x) = f_1(x_1, \ldots, x_n) + \ldots + f_n(x_1, \ldots, x_n)$ was used to calculate the fitness function. To show the efficiency and stability of our hybrid scheme in comparison with existing classical iterative approaches, we combine the well-known Newton Trapezoidal methodology [57].

$$\begin{cases} \mathbf{x}_i^{[1]} = \mathbf{x}_i - \dfrac{F(\mathbf{x}_i)}{F'(\mathbf{x}_i)}, \\ \mathbf{x}_i^{[2]} = \mathbf{x}_i - 2\left(\dfrac{F(\mathbf{x}_i)}{F'\left(\mathbf{x}_i^{[1]}\right) + F'(\mathbf{x}_i)}\right), \end{cases}$$

Method	$MM^{\partial^{[1]}}$	$MM^{\partial^{[2]}}$	$MM^{\partial^{[3]}}$	$OM^{\partial^{[1*]}}$	$OM^{\partial^{[2*]}}$	$OM^{\partial^{[3*]}}$	$OM^{\partial^{[4*]}}$
Analysis of numerical approximations without applying LCT-IIB conditions							
x_1	1.340191844	1.340157654	1.340157654	1.3401913234	1.3401914634	1.3401921545	1.3401841
x_2	0.850232925	0.850223841	0.850223841	0.8502323415	0.8502301675	0.8502301735	0.85012423
F^{T^*}	2.01e-03	1.05e-05	1.01e-03	1.36e-03	4.51e-05	9.1e-0.3	6.5e-02
Analysis of numerical approximations without applying LCT-IIB conditions							
x_1	1.34019184	1.340157655	1.3401914634	1.3401913667	1.3401914630	1.3401921523	1.34018415
x_2	0.85023292	0.850223841	0.8502301675	0.8502323414	0.8502301679	0.8502301703	0.85012455
F^{T^*}	2.01e-15	1.05e-15	3.35e-11	1.36e-03	4.51e-07	9.1e-0.9	6.5e-06

Table 3.
Comparison of the hybrid numerical schemes with different optimization schemes for solving Example 1.

Method	CPU-Time	Maximum -Error	Minmum-Error	COC	$COC^{\ominus[i]}$
Analysis of $MM^{\ominus[1]}$, $MM^{\ominus[2]}$, $MM^{\ominus[3]}$ approximations without using the LCT-IIB conditions					
$MM^{\ominus[1]}$	2.12167	1.12e-02	2.15e-05	3.0124	3.02141
$MM^{\ominus[2]}$	1.13167	6.12e-05	1.7e-11	3.1421	3.31450
$MM^{\ominus[3]}$	1.13167	6.12e-05	1.7e-11	3.1421	3.31450
Analysis of $MM^{\ominus[1]}$, $MM^{\ominus[2]}$, $MM^{\ominus[3]}$ approximations using the LCT-IIB conditions					
$MM^{\ominus[1]}$	1.25401	4.1e-05	1.1e-14	3.21014	3.2145
$MM^{\ominus[2]}$	0.01234	7.7e-12	3.1e-18	3.7514	3.5146
$MM^{\ominus[3]}$	0.01234	7.7e-12	3.1e-18	3.7514	3.5146

Table 4.
Comparative analysis of the consistency of classical and hybrid numerical schemes utilized in the solution of nonlinear equations for solving Example 1.

Schemes	$x_0 = (1.5,1)$	$x_0 = (1.5,1)$	$x_0 = (1.5,1)$	$x_0 = (1.5,1)$	$x_0 = (1.5,1)$
$MM^{\ominus[1]}$	1.3e-03	2.1e-05	1.1e-02	1.1e-04	3.1e-02
$MM^{\ominus[2]}$	6.1e-05	1.7e-10	1.6e-08	6.1e-07	1.6e-06
Iterations	06	10	06	10	06
Comparison of errors in the approximation of the solution to Example 1 when local convergence conditions in Banach space (LCT-IIB) are applied					
$MM^{\ominus[1]}$	1.1e-14	6.1e-06	4.1e-05	5.2e-07	3.1e-06
$MM^{\ominus[2]}$	3.1e-18	7.1e-12	7.1e-14	6.4e-14	8.4e-13
Iterations	06	10	06	10	06

Table 5.
Comparison of $MM^{\ominus[z]} - MM^{\ominus[z]}$ errors for different initial guess values in order to solve the system of nonlinear equations in Example 1.

Initial guesses	Parameter	$MM^{\ominus[1]}$			$MM^{\ominus[2]}$		
$\left\{x_1^{(0)}, x_2^{(0)}\right\}$	α	$CO^{\vartheta[1]}$	$CO^{\vartheta[2]}$	$CO^{\vartheta[3]}$	$CO^{\vartheta[1]}$	$CO^{\vartheta[2]}$	$CO^{\vartheta[3]}$
{0.01,0.02}	0.1201	2.1562	2.9814	3.0124	3.0145	3.3115	3.4125
{0.25,0.081}	0.5201	3.1311	2.7165	3.0021	3.2151	3.0124	3.6714
{0.031,0.31}	5.1201	2.1142	2.7145	3.0121	3.5612	3.4210	3.3711
{0.011,0.014}	0.1201	2.9191	2.6984	2.9912	3.7815	3.1415	3.3431
{0.01,0.012}	2.1761	2.5122	2.8741	2.6113	3.7841	3.6951	2.9541

Table 6.
Comparison of the local computational order of convergence between standard and hybrid numerical schemes for solving Example 1 for different parameter values using the LCT-IIB conditions.

Method	$MM^{\partial[1]}$	$MM^{\partial[2]}$	$MM^{\partial[3]}$	$OM^{\partial[1*]}$	$OM^{\partial[2*]}$	$OM^{\partial[3*]}$	$OM^{\partial[4*]}$
Analysis of numerical approximations without applying LCT-IIB conditions							
x_1	0.682791454	0.6827914123	0.682791412	0.682791412	0.6827914	0.682791412	0.6827914
x_2	0.506474324	0.506474978	0.506474456	0.506474121	0.506474	0.506474456	0.5064742
x_3	−0.703061854	−0.70306187	−0.703061123	−0.70306112	−0.703061	−0.703061123	−0.7030611
x_4	−0.8622550452	−0.862255078	−0.8622550	−0.86225504	−0.8622550	−0.86225502	−0.8622550
x_5	3.8812e-07456	3.881287e-07	3.8812e-07	3.881214e-07	3.8812e-07	3.881254e-07	3.88102e-07
x_6	−3.01314e-13	−3.01312e-13	−3.01397e-13	−3.01398e-13	−3.01397e-13	−3.01378e-13	−3.0103e-13
F^{T*}	1.145e-03	4.12e-03	6.104e-01	3.514e-04	6.514e-05	3.014e-03	1.214e-02
Analysis of numerical approximations without applying LCT-IIB conditions							
x_1	0.6827914542	0.6827914	0.682791478	0.682791465	0.68279144	0.682791478	0.68279149
x_2	0.506474698	0.506474	0.506474897	0.506474454	0.50647465	0.506474897	0.50647487
x_3	−0.70306154	−0.703061	−0.70306179	−0.70306174	−0.70306145	−0.70306179	−0.7030617
x_4	−0.86225504	−0.8622550	−0.86225504	−0.862255054	−0.862255087	−0.86225504	−0.8622550
x_5	3.881298e-07	3.88127e-07	3.88192e-07	3.88126e-07	3.88129e-07	3.88192e-07	3.88120e-07
x_6	−3.01336e-13	−3.0135e-13	−3.0913e-13	−3.0133e-13	−3.0134e-13	−3.0913e-13	−3.0130e-13
F^{T*}	0.214e-07	3.0125e-06	1.104e-04	4.0251e-07	3.154e-05	6.124e-05	8.941e-06

Table 7.
Comparison of the hybrid numerical schemes with different optimization schemes for solving Example 2.

Method	CPU-Time	Maximum-Error	Minimum-Error	COC	$COC^{\partial[4]}$
Analysis of $MM^{\partial[1]}$, $MM^{\partial[2]}$, $MM^{\partial[3]}$ approximations without using the LCT-IIB conditions					
$MM^{\partial[1]}$	3.1211	7.71632e-01	3.15125e-06	2.801121129	2.91298554
$MM^{\partial[2]}$	2.3491	8.12451e-03	9.12654e-12	3.012321008	3.31223145
$MM^{\partial[3]}$	4.2391	7.10051e-01	8.65034e-05	2.095323423	2.90043318
Analysis of $MM^{\partial[1]}$, $MM^{\partial[2]}$, $MM^{\partial[3]}$ approximations using the LCT-IIB conditions					
$MM^{\partial[1]}$	2.3121	8.8124e-06	7.61254e-11	3.01221665	3.011351
$MM^{\partial[2]}$	0.0124	9.3635e-13	3.12351e-18	3.39121987	3.561364
$MM^{\partial[3]}$	3.6154	7.6535e-05	4.5561e-10	2.9900987	3.000375

Table 8.
Comparative analysis of the consistency of classical and hybrid numerical schemes utilized in the solution of the nonlinear system of equations used in Example 2.

Schemes	V_1	V_2	V_3	V_4	V_5
$MM^{\partial[1]}$	1.7145154e-04	3.1458741e-06	7.7154824e-01	6.954711e-01	6.145212e-02
$MM^{\partial[2]}$	4.2151254e-07	9.145217e-012	9.9854135e-04	3.654128e-03	6.325145e-03
Iterations	06	04	06	06	06
Comparison of errors in the approximation of the solution to Example 2 when local convergence conditions in Banach space (LCT-IIB) are applied					
$MM^{\partial[1]}$	9.2145128e-011	3.21456e-006	4.6215481e-006	6.352145e-006	3.201545e-006
$MM^{\partial[2]}$	1.001245e-018	6.21451e-014	9.3215425e-013	5.002145e-014	6.321543e-016
Iterations	06	04	06	06	06

We denote $V_1 = \{0.02,0.11,0.5,0.01,0.4,0.214\}, V_2 = \{0.0,0.3,0.4,0.7,0.8,0.21\}$, $V_3 = \{0.012,0.1,0.05,0.10,0.09,0.014\}, V_4 = \{0.141,0.08,0.14,0.2,0.0,0.021\}$, *and* $V_5 = \{0.0,0.1,0.06,0.01,0.05,0.0\}$..

Table 9.
Comparison of $MM^{\partial[4]} - MM^{\partial[2]}$ errors for different initial guess values theorem in Banach space.

Initial guesses	Parameter	$MM^{\partial[1]}$			$MM^{\partial[2]}$		
$\left\{x_1^{(0)}, \cdots, x_6^{(0)}\right\}$	α	$CO^{\vartheta[1]}$	$CO^{\vartheta[2]}$	$CO^{\vartheta[3]}$	$CO^{\vartheta[1]}$	$CO^{\vartheta[2]}$	$CO^{\vartheta[3]}$
V_1	3.21451	2.91452	2.3651	3.0215	3.0124	3.5412	3.5142
V_2	3.21451	2.98541	2.9845	3.2145	3.1124	3.0124	3.0061
V_3	3.21451	2.39654	2.9874	3.1241	3.2415	3.6541	3.7145
V_4	3.21451	3.02145	2.1451	3.0214	3.0124	2.9991	3.0014
V_5	3.21451	2.99841	2.6541	2.9991	3.3210	3.001	3.1748

Table 10.
Comparison of the local computational order of convergence between standard and hybrid numerical schemes for solving Example 2 for different parameter values using the LCT-IIB conditions.

Method	CPU-Time	Maximum-Error	Minimum-Error	COC	$COC^{\partial[\ell]}$
Analysis of $MM^{\partial[1]}$, $MM^{\partial[2]}$, $MM^{\partial[3]}$ approximations without using the LCT-IIB conditions					
$MM^{\partial[1]}$	6.12112	0.72132e-01	1.12125e-06	3.411121129	2.9529855
$MM^{\partial[2]}$	4.52191	3.10051e-03	3.11454e-12	3.0123001008	3.30023145
$MM^{\partial[3]}$	5.672153	0.47651e-02	7.14544e-10	2.986454375	1.30023145
Analysis of $MM^{\partial[1]}$, $MM^{\partial[2]}$, $MM^{\partial[3]}$ approximations using the LCT-IIB conditions					
$MM^{\partial[1]}$	3.31215	2.82240e-06	9.61254e-11	3.014451665	3.087451
$MM^{\partial[3]}$	1.01246	3.36005e-13	0.10051e-18	3.31121987	3.561365
$MM^{\partial[3]}$	2.01543	0.65991e-07	0.56436e-11	2.00543987	2.564755

Table 11.
Comparative analysis of the consistency of classical and hybrid numerical schemes utilized in the solution of the nonlinear system of equations used in Example 3.

Initial guesses	Parameter	$MM^{\partial[1]}$			$MM^{\partial[2]}$		
$\left\{x_1^{(0)}, \cdots, x_6^{(0)}\right\}$	α	$CO^{\vartheta[1]}$	$CO^{\vartheta[2]}$	$CO^{\vartheta[3]}$	$CO^{\vartheta[1]}$	$CO^{\vartheta[2]}$	$CO^{\vartheta[3]}$
V_1	3.23251	2.91452	2.3321	3.0215	3.0224	3.5412	3.51242
V_2	3.00151	2.98111	2.9845	3.2125	3.1100	3.0124	3.0161
V_3	3.5451	2.39224	2.9324	3.1541	3.2415	3.6651	3.7525
V_4	3.28451	3.02355	2.1111	3.0314	3.0124	2.9991	3.0014
V_5	3.20051	2.94241	2.6041	2.0091	3.3210	3.001	3.1658

We denote $V_1 = \{0.2,0.5,0.1,0.4,0.1,0.5,0.1,0.4,0.214,0.5,0.1,0.4,0.1,1.0\}$,
$V_2 = \{0.0,0.4,0.7,0.8,0.3,0.4,0.9,0.8,0.21,0.14,0.7,0.8,0.0,0.12\}$,
$V_3 = \{0.12,0.5,0.1,0.9,0.14,0.1,0.5,0.1,0.9,0.14,3.5,8.1,0.9,0.14\}$,
$V_4 = \{0.2,0.0,0.21,0.11,0.8,0.14,0.2,0.0,0.21,0.1,1.1,0.2,14.3,3.1\}$, *and*
$V_5 = \{0.0,3.1,0.6,0.1,0.5,0.0,3.2,0.0,0.21,0.2,1.3,2.5,0.6,6.3\}$..

Table 12.
Comparison of the local computational order of convergence between standard and hybrid numerical schemes for solving Example 3 for different parameter values using the LCT-IIB conditions.

with Algorithm 1 to construct a hybrid-optimized scheme abbreviated as abbreviated as $MM^{\partial[3]}$.

Example 1: Consider the nonlinear system of equations shown below [63].

$$F(\mathbf{x}) = \begin{cases} x_1 + e^{x_2} - \cos(x_2) = 0, \\ 3x_1 - \sin(x_1) - x_2 = 0, \end{cases} \tag{46}$$

which has the solution $\mathbf{x}^* = (0,0)^t$.

Based on the results presented in **Table 3** and **Figure 3(a,b)**, it is evident that the hybrid numerical optimization approach $MM^{\partial[2]}$ outperforms $MM^{\partial[1]}$, $MM^{\partial[3]}$, $OM^{\partial[1*]} - OM^{\partial[4*]}$ in both cases, regardless of whether the local convergence theorem assumptions in Banach space are implemented or not.

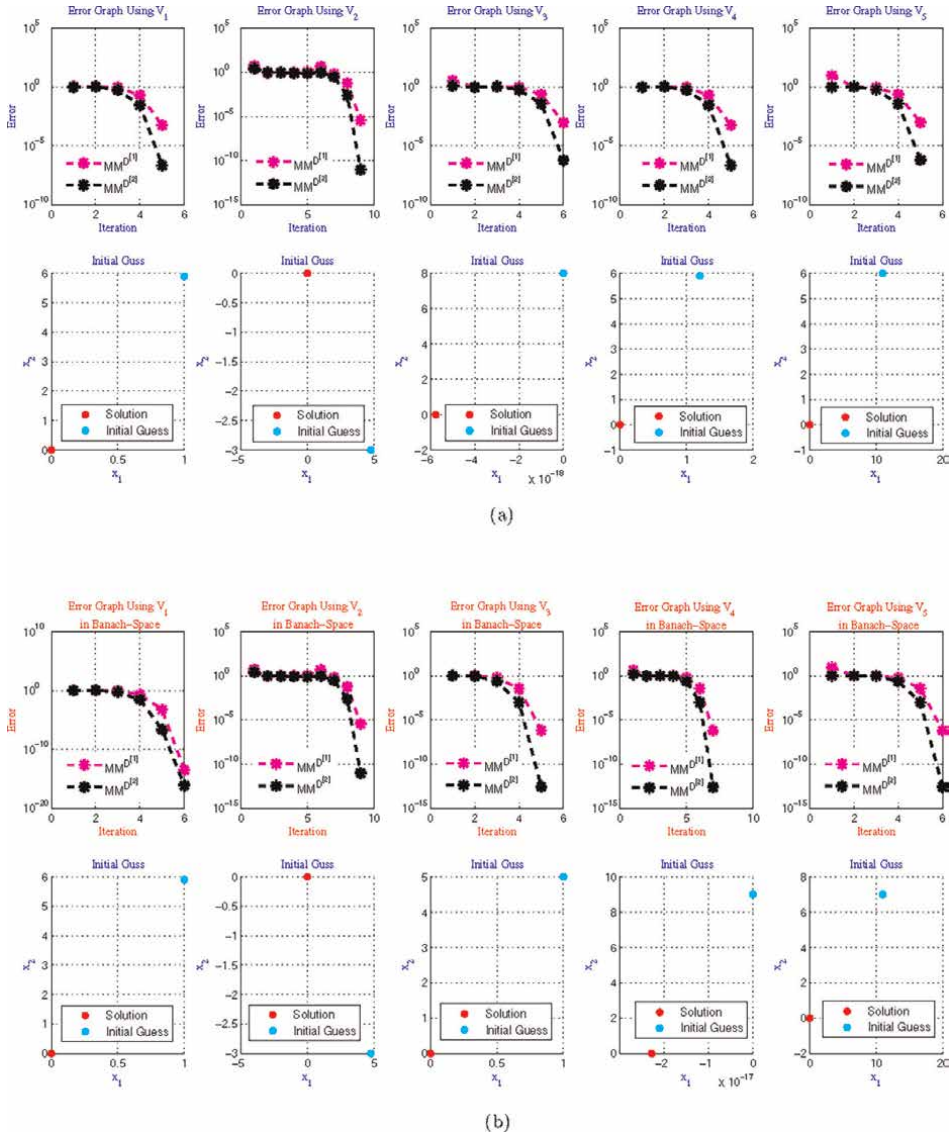


Figure 3.
(a,b) A comparison of error graphs between hybrid and classical numerical algorithms with different initial guess values for solving (46).

The results of the consistency and stability analyses for the numerical schemes $MM^{\supset[1]}$, $MM^{\supset[2]}$, and $MM^{\supset[3]}$ are presented in **Tables 4** and **5**, respectively. The consistency results of **Table 4** demonstrates unequivocally that the methods $MM^{\supset[1]}$, $MM^{\supset[2]}$, and $MM^{\supset[3]}$ are more consistent when the LCT-IIB conditions are used. Implementing these assumptions improves the convergence order, COC, and $COC^{\supset[i]}$. Regardless of whether the local convergence conditions are satisfied or not, **Table 5** demonstrates that the residual error of numerical scheme $MM^{\supset[2]}$ is considerably superior to that of $MM^{\supset[1]}$ for various initial vectors in both scenarios. A hybrid numerical optimization

technique exhibits a markedly superior performance in solving (46) compared to a classical iterative scheme. The computational order of convergence of the numerical schemes $MM^{\mathcal{D}^{[1]}} - MM^{\mathcal{D}^{[2]}}$ for a variety of initial vectors is illustrated in **Table 6**.

Example 2: Consider a nonlinear system of equations resulting from a neurophysiology problem describe below [64]:

$$F(\mathbf{x}) = \begin{cases} x_1^2 + x_3^2 - 1 = 0, \\ x_2^2 + x_4^2 - 1 = 0, \\ x_5 x_3^3 + x_6 x_4^3 = 0, \\ x_5 x_1^3 + x_6 x_2^3 = 0, \\ x_5 x_1 x_3^2 + x_6 x_2 x_4^2 = 0, \\ x_5 x_3 x_1^2 + x_6 x_4 x_2^2 = 0, \end{cases} \quad -10 \leq x_1, \ldots, x_6 \leq 10. \tag{47}$$

The above nonlinear equations have several solutions within the interval $[-10,10]$ and are difficult to solve.

Table 7, show that the hybrid numerical optimization approach $MM^{\mathcal{D}^{[2]}}$ outperforms $MM^{\mathcal{D}^{[1]}}$, $MM^{\mathcal{D}^{[3]}}$, $OM^{\mathcal{D}^{[1*]}} - OM^{\mathcal{D}^{[4*]}}$ in both cases, with and without LCT-IIB assumptions (**Figure 4**).

Tables 8 and **9** illustrate the results of the consistency and stability analyses performed on the numerical schemes $MM^{\mathcal{D}^{[1]}} - MM^{\mathcal{D}^{[3]}}$. Implementing the criteria of Theorem 2 in Banach space improves the consistency of methods $MM^{\mathcal{D}^{[1]}} - MM^{\mathcal{D}^{[2]}}$, as seen in **Table 8**. The maximum and minimum errors, which apply local convergence assumptions in Banach space, demonstrate the system's increased stability. Applying these assumptions improves convergence order, COC, and $COC^{\mathcal{D}^{[i]}}$. **Table 9** shows that $MM^{\mathcal{D}^{[2]}}$'s residual error outperforms $MM^{\mathcal{D}^{[1]}}$'s for various starting vectors, regardless of whether local convergence requirements are met. For solving (47), a hybrid numerical optimization technique $MM^{\mathcal{D}^{[2]}}$ outperforms $MM^{\mathcal{D}^{[1]}}$. **Table 10** shows the

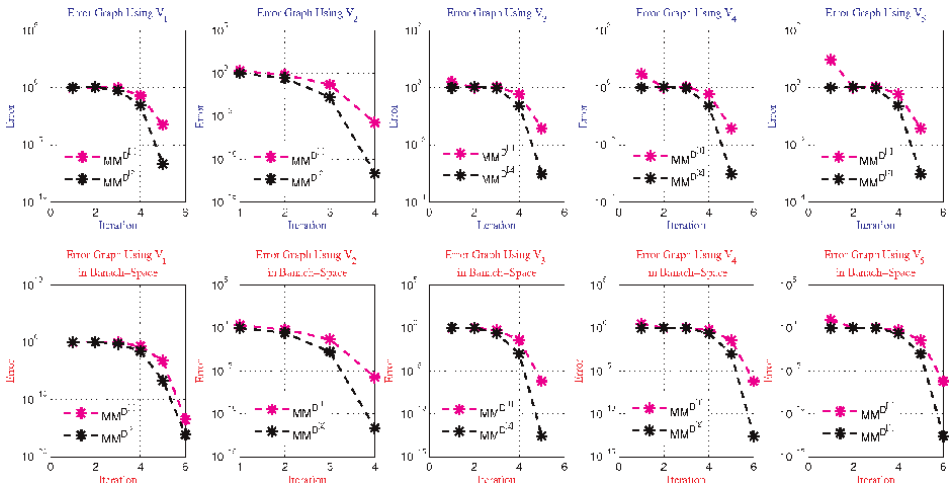


Figure 4.
A comparison of error graphs between hybrid and classical numerical algorithms with different initial guess values for solving (47).

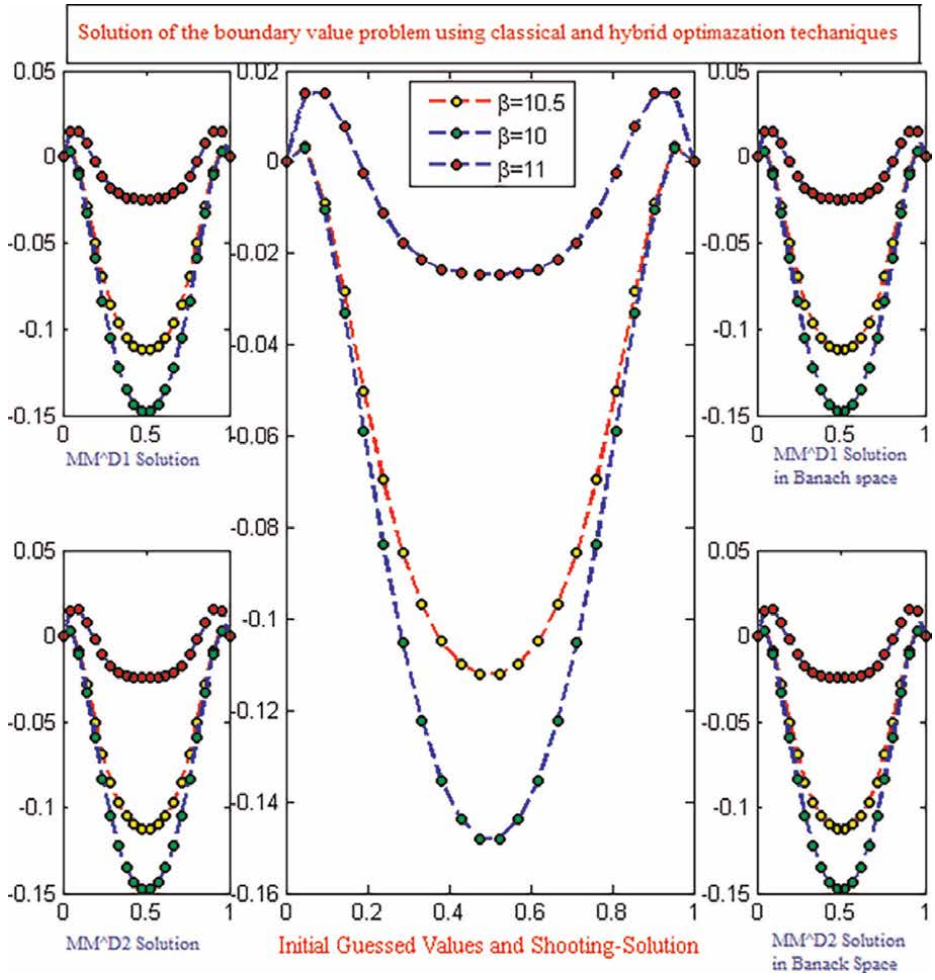


Figure 5.
Solution of the boundary value problem using classical and hybrid numerical schemes, with and without the use assumptions of LCT-IIB.

computational order of convergence of the numerical schemes $MM^{\ominus[1]} - MM^{\ominus[2]}$ for different initial vectors.

Example 3: We consider a real life application described by the nonlinear boundary value problem [65].

$$\begin{cases} \dfrac{d^2y}{dx^2} = -\beta e^y, & ; 0 \le x \le 1. \\ y(0) = 0; y(1) = 0, \end{cases} \tag{48}$$

The interval [0,1] is discretized as follows:

$$x_0 = 0 < x_1 < x_2 < \cdots < x_n < x_{n-1} = 1, x_{j-1} = x_j + h, \tag{49}$$

where $h = \frac{1}{m+1}$ is the step length and m is the size of the nonlinear system of equations which is obtained after discretizing the problem using finite difference approximation. Using central difference approximation of

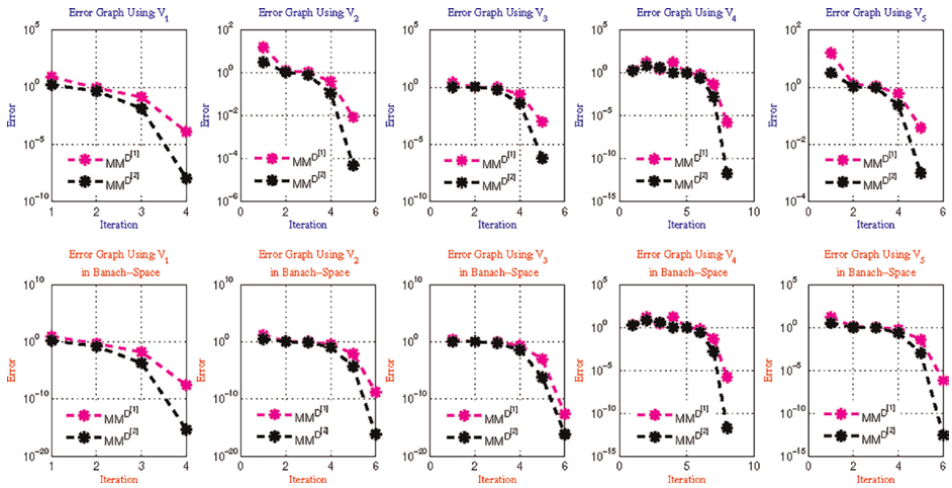


Figure 6.
A comparison of error graphs between hybrid and classical numerical algorithms with different initial guess values for solving (48).

$$y'' \approx \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2}, i = 1, 2, \ldots m. \tag{50}$$

in (48), we obtained the following nonlinear system of equations for m = 15.

$$F(\mathbf{y}) = y_{i+1} - 2y_i + y_{i-1} + h^2 \beta e^{y_i} = 0, i = 1,2,3, \ldots m. \tag{51}$$

Using MATLAB's built-in package BVP4C, we can solve the problem (48) to four decimal places.

$$\mathrm{V} = \begin{bmatrix} 16.86666667, 16.73333333, 16.6, 16.46666667, 16.33333333, 16.2, \\ 16.06666667, 15.9333333, 15.8, 15.66666667, 15.53333333, 15.4, 15.26666667 \\ 15.13333333, 15, 1.86666667, 14.733333333, 14.6, 14.46666667 \end{bmatrix}^T$$

Figure 5 and **Table 11** show the approximate solution to the boundary values problem using $\mathrm{MM}^{\circlearrowright[1]} - \mathrm{MM}^{\circlearrowright[2]}$, with and without the LCT-IIB condition (**Figure 6**).

Table 11 displays the consistency and stability analysis results for numerical schemes $\mathrm{MM}^{\circlearrowright[1]} - \mathrm{MM}^{\circlearrowright[3]}$. **Table 11** demonstrates that applying the criteria of Theorem 2 in Banach space improves the consistency of methods $\mathrm{MM}^{\circlearrowright[1]}$, $\mathrm{MM}^{\circlearrowright[2]}$, and $\mathrm{MM}^{\circlearrowright[3]}$. The residual error curve in **Figure 6** shows that $\mathrm{MM}^{\circlearrowright[2]}$ has a higher convergence rate than $\mathrm{MM}^{\circlearrowright[1]}$. **Table 12** depicts the computational order of convergence of the numerical schemes $\mathrm{MM}^{\circlearrowright[1]} - \mathrm{MM}^{\circlearrowright[2]}$ for various initial vectors.

7. Conclusion

In this research, we proposed a two-step numerical approach to solving nonlinear equations. We discussed the local convergence of the proposed scheme in Banach

space. Using the assumption of the convergence theorem in Banach space improves convergence rate, as demonstrated in the dynamical analysis. **Tables 1** and **2** and **Figure 2** show the convergence radii and results from the dynamical analysis. Furthermore, a hybrid numerical scheme is proposes that combines the two-step iterative method with the Butterfly optimization method to overcome the limitation of local minima, increase the convergence rate, and yields better approximations than $OM^{\mathfrak{I}[1*]} - OM^{\mathfrak{I}[4*]}$.

Our study compares the efficiency of the proposed technique for different engineering applications with and without the use of the local convergence theorem in Banach space. In all circumstances, the 2-norm fitness function is utilized. **Tables 3**, **7**, and **11** show that $MM^{\mathfrak{I}[2]}$ outperforms alternative techniques. **Tables 4**, **5**, **8**, **9**, and **11** demonstrate the consistency and stability of the two-step iterative process, as well as classical and hybrid optimization techniques. **Tables 1**–**12** demonstrate that the $MM^{\mathfrak{I}[2]}$ is significantly more reliable and consistent than the $MM^{\mathfrak{I}[1]}$, $MM^{\mathfrak{I}[3]}$ and $OM^{\mathfrak{I}[1*]} - OM^{\mathfrak{I}[4*]}$, respectively. Finally, **Tables 6**, **10**, and **12** show both the convergence rate and the local computational order of convergence. Using the assumptions of Theorem 2 improves both the rate and the local computational order of convergence.

In the future, we will investigate and analyze higher-order numerical iterative schemes and hybrid optimization schemes utilizing the methods described in this article to solve more complex engineering problems.

Acknowledgements

The work is supported by the Free University of Bozen-Bolzano (IN200Z SmartPrint). Bruno Carpentieri belongs to the Gruppo Nazionale per il Calcolo Scientifico (GNCS) of the Istituto Nazionale di Alta Matematica (INdAM) and this work was partially supported by INdAM-GNCS under Progetti di Ricerca 2024.

Author details

Mudassir Shams[1,2*] and Bruno Carpentieri[1*]

1 Faculty of Engineering, Free University of Bozen-Bolzano (BZ), Italy

2 Department of Mathematics and Statistics, Riphah International University, Islamabad, Pakistan

*Address all correspondence to: mudassir.shams@unibz.it
and bruno.carpentieri@unibz.it

IntechOpen

© 2024 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

References

[1] Su Y, Chen G. Iterative methods for solving linear matrix equation and linear matrix system. International Journal of Computer Mathematics. 2010;**87**(4): 763-774

[2] Elfving T. Block-iterative methods for consistent and inconsistent linear equations. Numerische Mathematik. 1980;**35**:1-12

[3] Sherman AH. On Newton-iterative methods for the solution of systems of nonlinear equations. SIAM Journal on Numerical Analysis. 1978;**15**(4):755-771

[4] Gale D. The Theory of Linear Economic Models. McGraw-Hill, University of Chicago Press; 1989

[5] Dymova L, Sevastjanov P, Pilarek M. A method for solving systems of linear interval equations applied to the Leontief input–output model of economics. Expert Systems with Applications. 2013; **40**(1):222-230

[6] Kardashov V, Einav S, Eppelbaum L, Ismail-Zadeh A. A novel approach to investigation and control of nonlinear nonstationary processes: Application to environments and biomedical engineering. Sci Israel. 1999;**3**:24-33

[7] Masri SF, Caffrey JP, Caughey TK, Smyth AW, Chassiakos AG. Identification of the state equation in complex non-linear systems. International Journal of Non-Linear Mechanics. 2004;**39**(7):1111-1127

[8] Verdon N, Allery C, Beghein C, Hamdouni A, Ryckelynck D. Reduced-order modelling for solving linear and non-linear equations. International Journal for Numerical Methods in Biomedical Engineering. 2011;**27**(1): 43-58

[9] Proinov PD, Vasileva MT. On the convergence of family of weierstrass-type root-finding methods. Comptes Rendus de l'Académie Bulgare des Sciences. 2015;**68**:697-704

[10] Mir NA, Muneer R, Jabeen I. Some families of two-step simultaneous methods for determining zeros of non-linear equations. ISRN Applied Mathematics. 2011;**1999**:1-11

[11] Shams M, Kausar N, Araci S, Oros GI. Numerical scheme for estimating all roots of non-linear equations with applications. AIMS Mathematics. 2023; **8**(10):23603-23620

[12] Carpentieri B, Duff IS, Giraud L, Sylvand G. Combining fast multipole techniques and an approximate inverse preconditioner for large electromagnetism calculations. SIAM Journal on Scientific Computing. 2005; **27**(3):774-792

[13] Rizzo F, Carpentieri B, Motta G, Storer JA. Low-complexity lossless compression of hyperspectral imagery via linear prediction. IEEE Signal Processing Letters. 2005;**12**(2):138-141

[14] Carpentieri B, Weinberger MJ, Seroussi G. Lossless compression of continuous-tone images. Proceedings of the IEEE. 2000;**88**(11):1797-1809

[15] Shams M, Carpentieri B. On highly efficient fractional numerical method for solving nonlinear engineering models. Mathematics. 2023;**11**(24):4914

[16] Thukral R. Introduction to a Newton-type method for solving nonlinear equations. Applied Mathematics and Computation. 2008; **195**(2):663-668

[17] Younes AB, Turner J. Generalized least squares and Newton's method algorithms for nonlinear root-solving applications. The Journal of the Astronautical Sciences. 2013;**60**:517-540

[18] Rahman NHA, Ibrahim A, Jayes MI. Newton homotopy solution for nonlinear equations using Maple14. Journal of Science and Technology. 2011; **3**(2):69-75

[19] Tatari M, Dehghan M. On the convergence of He's variational iteration method. Journal of Computational and Applied Mathematics. 2007;**207**(1): 121-128

[20] Abdirashid Ogli MM, Mashrabjon Ogli SS, Hakimjon Ogli HH. Determination of gas pressure distribution in a pipeline network using the Broyden method. Texas Journal of Engineering and Technology. 2023;**20**: 27-31

[21] Gdawiec K, Argyros IK, Qureshi S, Soomro A. An optimal homotopy continuation method: Convergence and visual analysis. Journal of Computational Science. 2023;**74**:102166

[22] Akram S, Shams M, Rafiq N, Mir NA. On the stability of Weierstrass type method with King's correction for finding all roots of non-linear function with engineering application. Applied Mathematical Sciences. 2020;**14**(10): 461-473

[23] Chen CY, Ghazali AH, Leong WJ. Scaled parallel iterative method for finding real roots of nonlinear equations. Optimization. 2022;**71**(9):2595-2611

[24] Özel M. A new decomposition method for solving system of nonlinear equations. Mathematical and Computational Applications. 2010;**15**(1): 89-95

[25] Kalaba R, Tesfatsion L. Solving nonlinear equations by adaptive homotopy continuation. Applied Mathematics and Computation. 1991;**41** (2):99-115

[26] Ali H, Datta T, Kamrujjaman M. Efficient family of iterative methods for solving nonlinear simultaneous equations: A comparative study. Journal of Applied Mathematics and Computation. 2021;**5**(4):331-337

[27] Saad Y. Iterative methods for linear systems of equations: A brief historical journey. In: Brenner S, Shparlinski I, Shu C-W, Szyld D, editors. Contemporary Mathematics. Vol. 754. Providence, Rhode Island: American Mathematical Society; 2020. pp. 197-215

[28] Shams M, Carpentieri B. Efficient inverse fractional neural network-based simultaneous schemes for nonlinear engineering applications. Fractal and Fractional. 2023;**7**(12):849

[29] Sihwail R, Solaiman OS, Ariffin KAZ. New robust hybrid Jarratt-butterfly optimization algorithm for nonlinear models. Journal of King Saud University-Computer and Information Sciences. 2022;**34**(10):8207-8220

[30] Sabir Z, Raja MAZ, Umar M, Shoaib M. Design of neuro-swarmingbased heuristics to solve the third-order nonlinear multi-singular Emden–Fowler equation. The European Physical Journal Plus. 2020;**135**(5):410

[31] Raouf OA, Hezam IM. Sperm motility algorithm: A novel metaheuristic approach for global optimisation. International Journal of Operational Research. 2017;**28**(2): 143-163

[32] Nobahari H, Nasrollahi S. A non-linear estimation and model predictive

control algorithm based on ant colony optimization. Transactions of the Institute of Measurement and Control. 2019;**41**(4):1123-1138

[33] Shehadeh HA, Shagari NM. A hybrid grey wolf optimizer and sperm swarm optimization for global optimization. In: Handbook of Intelligent Computing and Optimization for Sustainable Development. USA: Wiley-Scrivener Publishing; 2022. pp. 487-507

[34] Aberth O. Iteration methods for finding all zeros of a polynomial simultaneously. Mathematics of Computation. 1973;**27**:339-344

[35] Shehadeh HAT. Single-objective and multi-objective optimization algorithms based on sperm fertilization procedure [thesis]. Malaysia: University of Malaya; 2018

[36] Ramalingam SP, Shanmugam PK. Hardware implementation of a home energy management system using remodeled sperm swarm optimization (RMSSO) algorithm. Energies. 2022;**15** (14):5008

[37] Sundararaju N, Vinayagam A, Veerasamy V, Subramaniam G. A chaotic search-based hybrid optimization technique for automatic load frequency control of a renewable energy integrated power system. Sustainability. 2022;**14**(9):5668

[38] Rana N, Latiff MSA, Abdulhamid SIM, Chiroma H. Whale optimization algorithm: A systematic review of contemporary applications, modifications and developments. Neural Computing and Applications. 2020;**32**: 16245-16277

[39] Singh N, Son LH, Chiclana F, Magnot JP. A new fusion of salp swarm with sine cosine for optimization of non-linear functions. Engineering with Computers. 2020;**36**:185-212

[40] Martnez E, Singh S, Hueso JL, Gupta DK. Enlarging the convergence domain in local convergence studies for iterative methods in Banach spaces. Applied Mathematics and Computation. 2016; **281**:252-265

[41] Maroju P, Magreñán ÁA, Sarra Í, Kumar A. Local convergence of fourth and fifth order parametric family of iterative methods in Banach spaces. Journal of Mathematical Chemistry. 2020;**58**:686-705

[42] Argyros IK, Magreñán ÁA, Orcos L. Local convergence and a chemical application of derivative free root finding methods with one parameter based on interpolation. Journal of Mathematical Chemistry. 2016;**54**:1404-1416

[43] Shams M, Rafiq N, Kausar N, Agarwal P, Park C, Mir NA. On iterative techniques for estimating all roots of nonlinear equation and its system with application in differential equation. Advances in Difference Equations. 2021; **2021**(1):1-18

[44] Shams M, Kausar N, Samaniego C, Agarwal P, Ahmed SF, Momani S. On efficient fractional caputo-type simultaneous scheme for finding all roots of polynomial equations with biomedical engineering applications. Fractals. 2023; **31**:2340075

[45] Kumar S, Sharma JR, Bhagwan J, Jäntschi L. Numerical solution of nonlinear problems with multiple roots using derivative-free algorithms. Symmetry. 2023;**15**(6):1249

[46] Argyros IK. A unifying local–semilocal convergence analysis and applications for two-point Newton-like

methods in Banach space. Journal of Mathematical Analysis and Applications. 2004;**298**(2):374-397

[47] Amat S, Busquier S, editors. Advances in Iterative Methods for Nonlinear Equations (Vol. 10). Cham, Switzerland: Springer; 2016

[48] Maheshwari AK. A fourth order iterative method for solving nonlinear equations. Applied Mathematics and Computation. 2009;**211**(2):383-391

[49] Argyros IK. On the semilocal convergence of inexact Newton methods in Banach spaces. Journal of Computational and Applied Mathematics. 2009;**228**(1):434-443

[50] Sharma JR, Kumar S, Argyros IK. Local convergence of an efficient multipoint iterative method in Banach space. Algorithms. 2020;**13**(1):25

[51] Singh A, Jaiswal JP. Several new third-order and fourth-order iterative methods for solving nonlinear equations. International Journal of Engineering Mathematics. 2014;**2014**:828409

[52] Huen K. Neue methode zur approximativen integration der differentialge-ichungen einer unabhngigen variablen. Zeitschrift für Angewandte Mathematik und Physik. 1900;**45**:23-38

[53] Amat S, Busquier S, Guti Arrez JM. Third-order iterative methods with applications to Hammerstein equations: A unified approach. Journal of Computational and Applied Mathematics. 2011;**235**(9): 2936-2943

[54] Chun C, Kim Y-I. Several new third-order iterative methods for solving nonlinear equations. Acta Applicandae Mathematicae. 2010;**109**:1053-1063

[55] Kou J, Li Y, Wang X. A modification of Newton method with third-order convergence. Applied Mathematics and Computation. 2006;**181**:1106-1111

[56] Darvishi MT, Barati A. A third-order Newton-type method to solve systems of nonlinear equations. Applied Mathematics and Computation. 2007; **187**:630-635

[57] Cordero A, Torregrosa JR. Variants of Newton's method for functions of several variables. Applied Mathematics and Computation. 2006;**183**:199-208

[58] Khirallah MQ, Hafiz MA. Solving system of nonlinear equations using family of Jarratt methods. International Journal of Differential Equations and Applications. 2013;**12**(2):69-83

[59] Zhang X, Peng H, Hu G. A high order iteration formula for the simultaneous inclusion of polynomial zeros. Applied Mathematics and Computation. 2006;**179**:545-552

[60] Adam SP, Alexandropoulos SAN, Pardalos PM, Vrahatis MN. No Free Lunch Theorem: A Review. Approximation and optimization: Algorithms, complexity and applications; 2019. pp. 57-82

[61] Cordero A, Soleymani F, Torregrosa JR. Dynamical analysis of iterative methods for nonlinear systems or how to deal with the dimension? Applied Mathematics and Computation. 2014; **244**:398-412

[62] Ramm AG. Dynamical Systems Method for Solving Nonlinear Operator Equations. Oxford, UK: Elsevier; 2006

[63] Sharma E, Panday S, Dwivedi M. New optimal fourth order iterative method for solving nonlinear equations.

International Journal on Emerging Technologies. 2020;**11**(3):755-758

[64] Gupta RK. Numerical Methods: Fundamentals and Applications. Cambridge University Press; 2019

[65] Margrave GF, Lamoureux MP. Numerical Methods of Exploration Seismology: With Algorithms in MATLAB. Cambridge University Press; 2019

Chapter 11

Perspective Chapter: Enhancing Regression Analysis with Splines and Machine Learning – Evaluation of How to Capture Complex Non-Linear Multidimensional Variables

Alexander A. Huang and Samuel Y. Huang

Abstract

This chapter focuses upon the use of both splines and machine-learning in prediction and the methodology for constructing splines in a predictive context. In the realm of predictive modeling, machine learning and splines represent two pivotal approaches that address the complexity of capturing nonlinear relationships within data. Machine learning excels in identifying intricate patterns and relationships through algorithms that learn from data, making it a powerful tool for prediction across vast datasets. However, its often opaque nature can pose challenges for interpretability. In contrast, splines offer a bridge between the simplicity of linear regression and the complexity of machine learning. By introducing cutpoints in the data, splines allow for flexible modeling of nonlinear trends, providing a clearer interpretation of how independent variables influence the dependent variable across different segments. This makes splines particularly valuable in multivariable regression contexts, where understanding the nuanced effects of covariates is crucial. While machine learning may deliver superior predictive power in some cases, splines provide a compelling balance of predictability and interpretability, especially in scenarios where understanding the underlying model is as important as the accuracy of predictions.

Keywords: splines, regression analysis, nonlinear relationships, model interpretability, cutpoints, overfitting, multivariable regression, predictive modeling

1. Introduction

Splines play an essential role in enhancing predictive modeling by introducing flexibility that linear models lack, especially when dealing with complex, nonlinear relationships inherent in many real-world datasets. Their unique ability to segment data into intervals and fit different polynomial equations within these segments

allows for a nuanced capture of trends that might otherwise be missed by more straightforward approaches. This adaptability makes splines particularly useful in fields such as epidemiology, finance, and environmental science, where understanding the intricacies of variable interactions can lead to more accurate and meaningful predictions. Moreover, the strategic placement of knots or cut points in the data can fine-tune the model's sensitivity to changes, optimizing the balance between overfitting and underfitting. As a result, splines provide a powerful tool for predictive modeling, offering a blend of precision and interpretability that enhances the quality of insights derived from complex [1–5].

Machine learning has revolutionized predictive modeling by leveraging algorithms that learn from data, enabling the identification of complex patterns and relationships that are not readily apparent. Unlike traditional statistical methods, machine learning can handle vast amounts of unstructured data—ranging from images and text to intricate sensor data—facilitating the development of models that can predict outcomes with remarkable accuracy. Its applications span a wide array of fields, including finance, where it predicts market trends and credit risks; healthcare, where it forecasts disease progression and patient outcomes; and e-commerce, where it enhances customer experience through personalized recommendations. By employing techniques such as regression, classification, and neural networks, machine learning automates the model-building process, continuously improving its predictions as more data becomes available. This self-improving capability allows machine learning models to adapt over time, making them invaluable for predictive modeling in dynamic environments where patterns and relationships can change rapidly [2, 3, 6].

The benefits of splines and machine learning in predictive modeling, while distinct, collectively contribute to advancing the field of data analysis. Splines offer the advantage of modeling non-linear relationships with a high degree of interpretability, allowing researchers to understand and explain the effects of independent variables on the dependent variable across different data segments. This characteristic is particularly beneficial in fields requiring clear explanations of model behavior, such as epidemiology and economics, where understanding the nature of variable interactions is as important as the predictions themselves. On the other hand, machine learning excels in handling complex, high-dimensional datasets, offering superior predictive accuracy through its ability to learn from data patterns and adjust its algorithms accordingly. This makes it particularly useful in applications where the primary goal is prediction accuracy, such as in image recognition, natural language processing, and real-time decision-making systems. Together, splines and machine learning encompass a spectrum of benefits from interpretability and precision to adaptability and accuracy, catering to diverse needs within the predictive modeling landscape [7–11].

This chapter delves into the multifaceted role of splines, starting with their construction and moving through to their interpretation within multivariable models. It concludes by exploring their application in predictive contexts and drawing comparisons with machine learning techniques.

2. Constructing, visualizing and interpreting nonlinear relationships: the role of splines in regression analysis

Constructing a univariable model using splines involves a systematic approach to model nonlinear relationships within a dataset. The process begins with the selection of the variable of interest and the identification of its potential nonlinear relationship

with the outcome. Splines allow for this relationship to be modeled by breaking the data into segments and fitting a polynomial function to each segment. The first step in constructing such a model is to determine the appropriate type of spline—linear, quadratic, cubic, or of a higher degree—based on the nature of the data and the relationship under investigation. The choice of spline affects the flexibility and complexity of the model, with higher-degree splines offering greater flexibility at the cost of increased complexity and risk of overfitting [9, 12–14].

After selecting the spline type, the next critical step is the placement of knots. Knots are specific points in the range of the data where the spline's polynomial degree changes, allowing the model to fit more closely to the data in regions where the relationship between the variable and the outcome changes. The placement of these knots can be based on quantiles, domain knowledge, or through optimization techniques that seek to minimize prediction error. It's essential to strike a balance between having enough knots to accurately model the relationship and having too many, which could lead to overfitting. Typically, starting with a smaller number of knots and incrementally adding more based on model performance and validation metrics is a prudent approach [6, 15–18].

Finally, once the spline type and knots are determined, the model can be constructed using statistical software that supports spline modeling. The software will fit the specified spline model to the data, segmenting it at the chosen knots and applying the polynomial functions accordingly. The resulting model coefficients provide insights into the relationship between the variable and the outcome, adjusted for the nonlinearity captured by the splines. It's crucial to evaluate the model's fit and predictive accuracy through diagnostics and validation techniques, such as cross-validation, to ensure that the model adequately captures the underlying relationship without overfitting. By carefully following these steps, researchers can construct a robust univariable spline model that provides valuable insights into complex relationships within their data [2, 19–23].

Splines can effectively enhance the goodness of fit commonly utilized in linear regression. In linear regression, a common practice involves analyzing a set of data comprising two variables by regressing one against the other using least squares regression. This method is frequently employed in medical literature to discern the relationship between two variables, typically through a simple linear regression. This regression model aids in understanding the relationship's parameters, such as the slope and variability of the line concerning the variables, often summarized by the standard error of the intercept and slope, along with the overall goodness of fit, usually represented by the R-squared value. This analysis quickly determines the significance of the computed slopes by comparing them to 0, indicating whether the covariate is a significant predictor. However, the major weakness of linear regression lies in its inability to accurately capture nonlinear relationships, such as curvilinear or asymptotic relationships, requiring rigorous thinking to model these complexities accurately. This is where a simple spline can be instrumental.

One exercise can be to understand how two covariates in the medical literature relate to each other, for example, the effect that vitamin E has on depression screening scores. In this mathematics learning exercise, students will explore the concept of splines to understand how nonlinear relationships between two variables can be modeled and interpreted. The focus will be on the real-world application of how the intake of vitamin E influences various health outcomes like depression, sleep quality, and general health, noting that the beneficial effects tend to level off beyond a certain dosage, approximately at 15 units. This scenario illustrates the critical role of

nonlinear modeling in determining the optimal nutrient levels, avoiding both deficiency and excess. Using a spline, we can establish nonlinear relationships between two variables, where the association depends on the independent variables' values. For example, in medical literature, the effect of vitamin E varies across different dependent variables, such as depression, sleep, and overall general health, with the relationship often plateauing at around 15 units. This example underscores the importance of modeling nonlinear relationships. One application of this is in understanding nutritional covariance with other variables. For instance, having more of a particular nutrient does not necessarily equate to better outcomes, as insufficient levels can lead to nutritional deficiencies and poor health. Modeling nonlinear relationships enables us to determine the optimal dose of medication or nutrient, showing when the effect plateaus. This can be achieved through segmenting the data and modeling different areas, allowing us to understand the model's benefits. Cubic functions are commonly used for spline execution, fitting polynomial curves at different cut points. This approach improves goodness of fit by accurately modeling nonlinear functions and providing more information, leading to increased model accuracy throughout. A figure generated like below shows how splines can effectively evaluate the relationship between independent and dependent covariates (**Figure 1**) [24–26].

One significant advantage of univariate or simple linear regression is the ability to visualize the goodness of fit through plotting. This allows for easy evaluation of how the dependent and independent variables relate across the entire range of the independent variable. Consequently, it becomes apparent if there's a strong nonlinear relationship present in specific areas. Visualizing the spline results alongside the covariance output enables a robust assessment of the benefits and goodness of fit achieved through spline development. Comparing this with linear regression

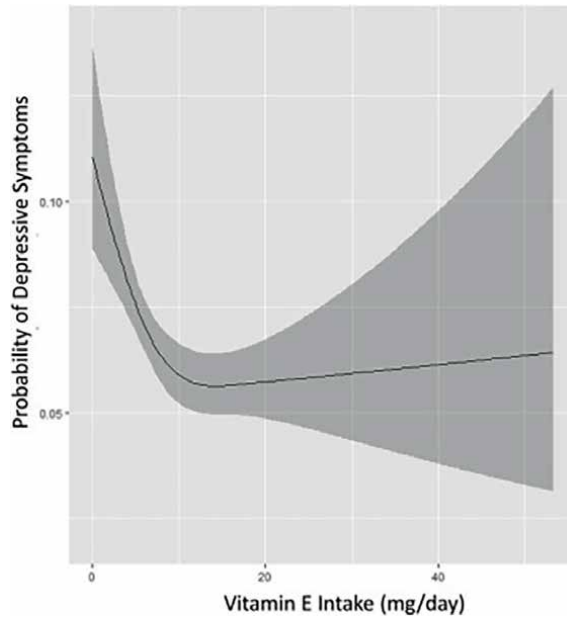


Figure 1.
The association between vitamin E intake and the likelihood of experiencing depressive symptoms, as assessed by the Patient Health Questionaire (PHQ)-9 scale. The connection between dietary vitamin E intake (represented on the x-axis, determined from dietary logs) and the likelihood of depressive symptoms (represented on the y-axis, evaluated through the PHQ-9 questionnaire).

plots facilitates a clear understanding of the advantages gained. Furthermore, this approach enhances interpretability by illustrating the spline's influence within the dataset context. Utilizing tools like bootstrap regression to address spline uncertainty facilitates a robust interaction between covariates, providing a visual summary and aiding interpretation [24–26].

3. Interpreting splines in multivariable regression: principles and considerations

Multivariate regression in the context of splines is very similar to the previously mentioned simple linear regression, wherein there is an ability to quantify the nonlinear relations present in ways that take into account how the independent and dependent relationships vary as the independent variable changes across its domain. Thus, it's crucial to extend this logic to multivariate regression, which is common in medical literature as a means to summarize the relationships between independent variables and the final dependent variable. However, the challenge lies in the development and evaluation of the spline. Additionally, interpretation of the spline itself differs. In evaluating multivariate linear regression, we often assume that all independent covariates are not only orthogonal to each other, meaning each covariate does not affect the others' relationship with the final product, but also that each covariate's interpretation remains stable throughout, implying the slope has no change over its domain. However, with the introduction of a spline, both of these assumptions become more challenging. First, the independence assumption is questioned. Secondly, in interpretation, two key considerations arise: whether the current independent variables being compared are independent of each other and whether the two splines created and the relationships built upon them remain independent [21, 23, 27, 28].

The first assumption is addressed similarly to how it's done in normal linear regression. We compare variables to determine if they are independent of each other. If they are, the same assumptions apply to that assembly, and the logic holds because functions of independent variables are independent of each other, regardless of the function, as long as the function does not contain the other covariate or covariates that are non-independent or dependent with the original covariate. Thus, it's essential that we still rely on the same assumptions as in normal multivariate regression, where we need to have independent dependent variables. The splines do not complicate these results; we are able to take into account the principle that functions of independent variables are not independent functions. On the flip side, it's important to understand the next component of the ration, which is how to interpret the splines in the context of other covariates. In normal linear regression, it does not matter what the other covariates are since we already found that they are independent, and thus holding one constant no longer changes the other. The question is, with the implementation of the spline, does this still hold? What is true is that they still hold based on the principle that functions of independent variables still hold this function, and thus the interpretation of a spline is the same, that we can [27, 29–32].

I interpret a slice of the data based on graphical principles. Each individual covariate is plotted separately on a suitable landscape, allowing us to interpret each slice independently, based on how the independent and dependent variables covary with each other. As long as we start with the initial assumption, demonstrated through normal linear regression techniques, of independence between all covariates, we can effectively interpret each spline as an independent entity. This understanding stems

from the fact that functions of independent covariates are functionally independent from each other. There exists a strong relationship between interpreting each covariate quantitatively [1, 4, 5, 33].

Another way to interpret these is as individual functions. Multivariate functions provide an output for the entire multivariate regression, from which we can identify significant components and assess their importance. A crucial principle in multivariable regression is recognizing that when one spline is overfitted, it may affect the significance of other covariates. For instance, if we overfit one spline with $n - 1$ degrees of freedom, resulting in a perfect fit for all data points in our dataset, the significance of other variables diminishes. Even if they are individually independent, there are no degrees of freedom left to fit them. Thus, it's essential to understand how different degrees of freedom play a role in constructing the model [2].

I analyze a slice of the data based on graphical principles. Each individual covariate is printed and plotted separately on a suitable landscape, enabling separate interpretation of each based on how the independent and dependent variables covary. As long as we maintain the initial assumption, demonstrated through normal linear regression techniques, of independence between all covariates, we can effectively interpret each spline as an independent entity. This understanding arises from the fact that functions of independent covariates are functionally independent from each other, establishing a strong relationship between the interpretation of each covariance. Another method of interpretation is viewing them as individual functions. Multivariate functions provide an output for the entire multivariate regression, allowing us to assess which splines have significant components.

4. Balancing complexity and interpretability: the comparative use of splines and machine learning in data analysis

In this section, we explore the utilization of Shapley additive explanations to illuminate the predictive mechanisms of machine learning models, based on varying feature inputs. We provide a potential exercise analyzing sample medical data. This methodology enables an analysis of how changes in feature values can influence the model's predictions, either by escalating risks or diminishing adverse outcomes. Such an approach underscores the models' proficiency in navigating complex, interactive, and nonlinear relationships without depending on conventional assumptions about normality or the independence of covariates.

The challenge, however, lies in the interpretability of these machine learning models to human users. The detailed components of these models, including elements like gradient boosting trees, neural network nodes, and the intricate computational steps, are largely inscrutable to non-expert users. This complexity contrasts sharply with the transparency offered by traditional regression techniques, such as logistic or linear regression. These techniques elucidate model dynamics through easily understandable metrics like slopes and odds ratios, enabling clear comprehension of covariate impacts. Unlike machine learning models, which may produce variable outcomes due to their stochastic nature, regression techniques offer consistent results across repeated applications, thanks to their deterministic framework [1, 3, 5, 34].

Our investigation revealed that, often, interpretations derived from Shapley additive explanations align with those obtained from restricted cubic spline analysis for several covariates, suggesting a level of consistency in their predictive insights. However, discrepancies did emerge, potentially rooted in the distinct methods these

approaches use to incorporate interactive effects. While restricted cubic splines adjust for variables upfront and apply corrections based on least squares error, Shapley additive explanations assess each covariate's contribution in a more sequential manner. Such differences may lead to variances in accounting for interactive effects. This study leverages the complementary strengths of both methods, using the transparency of restricted cubic splines to validate the insights gained from Shapley Additive Values (SHAP) values. This dual approach confirmed that the machine

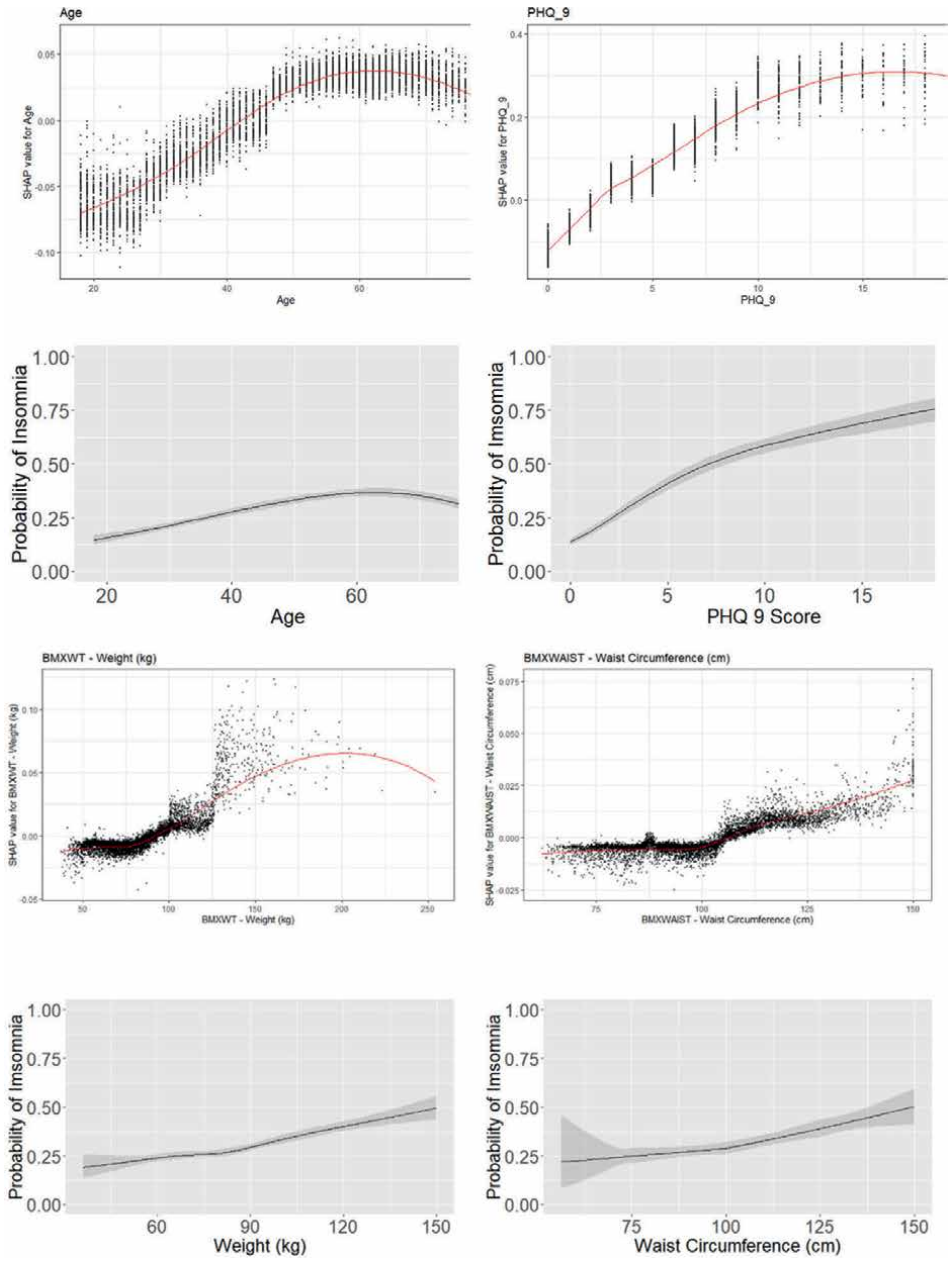


Figure 2.
Visualization of Shapley additive values with restricted cubic splines below.

learning models' utilization of covariates mirrors clinical realities, as evidenced by the consistent link between conditions like depression and insomnia identified by both methods. Consequently, our research supports a combined application of SHAP values and machine learning techniques, validated through the interpretability afforded by restricted cubic splines, to create a predictive modeling approach that is both accurate and accessible (**Figure 2**) [26, 35–37].

We can visualize the differences in how splines model data through direct comparison of model output. Comparative analysis of SHAP explanation contours (presented above) against the restricted cubic spline profiles for each covariate (displayed below).

Machine learning is a widely used tool that accounts for the complex, non-literal connections between variables in datasets, regardless of their completeness, independence, or the understanding of the variables themselves. Its application across literature, engineering, medicine, and mathematics demonstrates the promise of these models in prediction, capable of generating complex decision trees and identifying long-range covariate connections that are beyond human analytical capabilities. This complexity arises from the intricate covariance relationships and the sheer volume of data samples, often requiring computational power beyond human calculation [38–40].

Splines, in a similar vein, offer a complexity that surpasses multilinear regression models, providing greater accuracy. The choice between using splines and machine learning techniques depends on the objectives of the study. For those seeking explicit, interpretable models, simple splines or multilinear regression might be preferable over machine learning, as each variable can be precisely interpreted, almost independently from one another. Splines are noted for their strong interpretability, a contrast to the often opaque nature of machine learning models. If splines and machine learning models show equivalent predictive power, splines can be chosen to clearly demonstrate covariate relationships. However, if machine learning models significantly outperform splines in prediction accuracy, they should be utilized to leverage their superior predictive capabilities [31, 41–44].

Despite the advantages, splines enable the visualization of relationships between variables, including curvilinear connections, in a straightforward manner by accounting for various covariates. This positions splines as a powerful tool, offering both better predictability than linear models and easier interpretability compared to machine learning models. Yet, they inhabit a "middle ground," not as easily interpretable as linear models nor as predictively powerful as machine learning models. Therefore, splines represent a balanced option that requires careful evaluation to determine their optimal use in research, capturing the benefits and drawbacks of both worlds [45].

5. Optimizing model accuracy and interpretability: strategies for handling dependent variables and preventing overfitting

A critical principle in multivariable regression is recognizing that when one spline is overfitted, it may affect the significance of other covariates. For instance, if we overfit one spline with $n - 1$ degrees of freedom, resulting in a perfect fit for all data points in our dataset, the significance of other variables diminishes. Even if they are individually independent, there are no degrees of freedom left to fit the others. Thus, it's crucial to understand how different degrees of freedom play a role in construction [22, 23, 46].

Thus, we were able to carry out the spline analysis with accurate selection, building an entire model that can be interpreted as previously stated. The next step involves recognizing that independence in these models is not always present. This issue can be most accurately addressed by combining variables, specifically by including only one of two dependent variables. This approach helps in understanding potential interaction terms and necessitates the use of simplification for accurate interpretation. In situations where interpretability is not a concern for multivariate regression, an alternative method involves placing all variables in the model and performing an accurate fit. This can be a potent method if compatibility is not an issue, as it is purely a predictive exercise where overfitting might not be critically detrimental [21].

To prevent overfitting, which typically involves ensuring all variants are significant and there is an increase in predictive value with a decrease in degrees of freedom, another strategy involves using a train-test set. By allocating 80% of the data for training and 20% for testing and ensuring the predictive variability does not significantly differ between these two sets, we can propose a model with strong predictive power both internally and externally. Therefore, we have confidence that with additional new data, this model will continue to predict accurately moving forward [40, 47, 48].

6. Balancing predictive power and complexity: evaluating splines for optimal goodness of fit

Evaluating splines for goodness of fit presents a significant challenge due to the need to balance the amount of variability they explain and the direction of this variability. For instance, some splines may be highly predictive at extreme values of the independent variable, demonstrating substantial predictive capacity in certain ranges but not in others. Additionally, as data varies along the independent variable, the sample size within each range—despite uniform absolute sizes—may differ, possibly requiring various transformations to account for this variability. Therefore, it's crucial to consider both the predictive value of each spline and where they are most predictive [18, 39].

One method to evaluate the goodness of fit of a spline is through comparison to its performance across different ranges of its domain, identifying where it is most predictive. This approach is especially useful in contexts such as pharmaceutical trials, where understanding the effect size within specific ranges can be critical.

Another approach involves comparing the spline's fit to its performance across varying degrees of freedom. This comparison can help determine whether adding more points to the spline significantly improves the model's explanatory power compared to simpler models. Conducting an Analysis of Variance (ANOVA) test is one method to assess this, allowing us to see if an additional predictor improves variability explanation beyond what would be expected by chance. Principles such as the Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), and adjusted R-squared can also be employed to evaluate the inclusion of extra variables that are considered beyond the reduced degrees of freedom provided by the spline, effectively treating these additions as independent covariates [1, 4, 34].

Lastly, a straightforward evaluation involves assessing the spline's significance directly through its degrees of freedom and explained variability. If significant, this may justify either further development or maintaining the model as constructed. This

approach offers a simplified and direct method for interpreting the spline's effectiveness without the need for additional adjustments [1, 4, 34].

7. Adaptive complexity: the advantage of splines over polynomial regression in data modeling

One analogy for the use of splines is their similarity to polynomial regression. In polynomial regression, different powers of the covariate being analyzed are used as distinct variables. These variables are treated as independent, despite being dependent on the original covariate, essentially acting as indicator variables within the regression. One of the key advantages of this approach is the ease of interpreting the model, where summarizing the covariates' slopes or coefficients offers a clear depiction of the relationship. By constructing a polynomial function, where the sum of the powers combined with their respective coefficients provides a comprehensive description of the dataset, we achieve an intuitive understanding of the data relationships [4, 38, 49].

However, the challenge with polynomial regression lies in its least squares nature, which requires the data's curvature to be modeled under uniform constraints. This becomes problematic when different segments of the dataset exhibit varying behaviors, such as some areas being linear and others polynomial, making it difficult for polynomial regression to fully capture the dataset's intricacies.

Splines address this limitation by employing local polynomial regressions that can be adjusted based on the specific segment of the dataset being examined. This flexibility allows for different levels of complexity in the model across various parts of the data, leading to a more accurate and locally tailored representation. The ability of splines to adapt their complexity to match the dataset's local characteristics is one

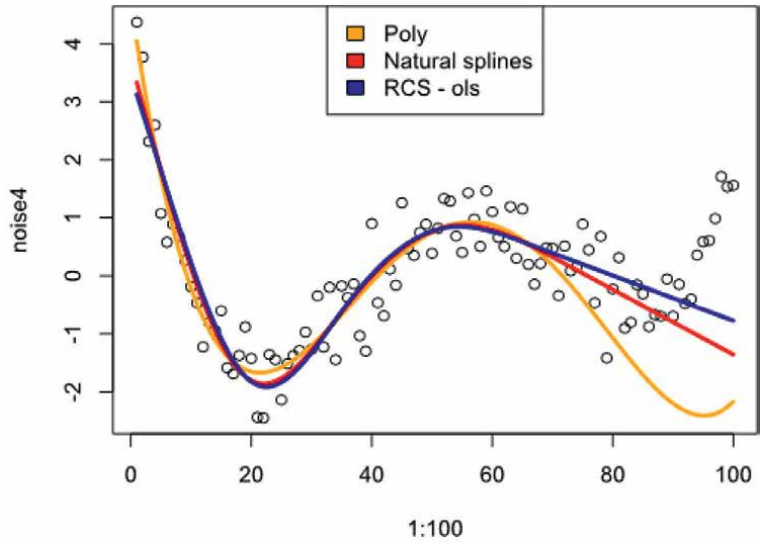


Figure 3.
Visualization of differences in fit between polynomial fitting, natural splines, and restricted cubic splines.

of their significant benefits, enabling the creation of simpler models where appropriate (**Figure 3**) [4, 38, 49].

This plot shows there are slight differences between comparing splines to polynomial regression.

8. Striking the balance: the art and science of setting cutpoints in restricted cubic spline models

In the realm of statistical modeling, particularly when employing restricted cubic splines, the selection of appropriate cutpoints stands as a pivotal step. These splines, designed to capture nonlinear relationships without pre-specifying the relationship's form, require cutpoints—specific locations on the predictor variable axis where the spline function alters its polynomial degree. This chapter is dedicated to navigating the intricacies of choosing these cutpoints, which is a delicate balance aimed at achieving a model that is both robust and interpretable.

Restricted cubic splines, an extension of cubic splines, are particularly noteworthy for their ability to reduce the risk of overfitting at the data's boundaries by enforcing linearity beyond the outermost knots. The cutpoints imbue the spline with its flexibility, enabling it to model complex relationships. However, this flexibility is a double-edged sword, as too many cutpoints can lead to overfitting, capturing noise rather than the underlying relationship, while too few can oversimplify the model, obscuring important data characteristics. Thus, selecting the optimal number and placement of cutpoints is crucial, necessitating a careful consideration of flexibility against parsimony.

Several strategies exist for setting cutpoints, ranging from default methods provided by statistical software, which often place cutpoints at quantiles of the predictor distribution, to more data-driven approaches. These approaches might involve exploratory data analysis (EDA) to visually identify changes in the relationship or derivative-based techniques to find significant variations in the rate of change. Additionally, domain knowledge can inform cutpoint placement, particularly in fields like biomedicine where established clinical thresholds can serve as natural cutpoints. Moreover, cross-validation offers a rigorous method to ascertain the best configuration of cutpoints by minimizing prediction error across different data splits.

Implementing restricted cubic splines and deciding on cutpoints involves practical considerations, including the choice of software and the balance between model complexity and interpretability. Most statistical packages provide tools for spline implementation, allowing for flexibility in cutpoint specification. However, it's important to remember that increased cutpoint numbers, while potentially improving model fit, also make the model more complex and harder to interpret for those without statistical expertise.

To illustrate the application of these concepts, consider a study exploring the relationship between body mass index (BMI) and diabetes risk. By employing restricted cubic splines and setting cutpoints at clinically significant BMI thresholds, researchers can flexibly model the potentially nonlinear increase in diabetes risk with BMI. Such an approach not only enhances the model's accuracy but also aids in interpreting complex relationships, ultimately guiding interventions and patient counseling strategies based on nuanced understanding of the data.

In summary, the process of setting cutpoints for restricted cubic splines is a nuanced endeavor that blends statistical techniques with domain-specific knowledge and practical considerations. This careful balancing act ensures the development of models that not only accurately reflect complex relationships within the data but are also accessible and interpretable to a broader audience, thereby maximizing their utility and impact.

9. Applications in the literature: author's own use of linear-models, splines, and machine-learning in a variety of contexts with explanation

Linear models have the benefit of being easy to compute and having the easiest interpretation with a slope, and can be easily applied in a variety of contexts, from mass evaluation of covariates to understanding how different variables change overtime [25, 35, 50–54]. Linear model type extensions such as survival analysis make use of similar interpretation for slopes. Machine-learning has benefits of being able to generate strong predictions [24, 26, 37, 55–59]. Additionally, with the addition of methods such as shapely additive explanations, machine-learning has been able to have easily to understand explanations to understand covariates. When combining both of these methods together and balancing their benefit, we are able to achieve this through splines [25, 36, 52, 53, 57–61]. This is present through application of sections 1–7 to generate accurate models, as can be shown when evaluation of splines through rigorous analysis to understand how to set cutoffs in a practical application such as with vitamin E cutoffs [24, 54–56].

10. Conclusion

Splines significantly enhance the predictive modeling landscape by providing the flexibility to model nonlinear relationships, a feature often missing in linear models. This flexibility is crucial for accurately capturing the nuances of real-world data across various domains. By dividing data into segments and applying unique polynomial equations to each, splines adeptly reveal hidden patterns within the data. This characteristic is invaluable across numerous fields, such as epidemiology, financial analysis, and environmental studies, where a deep understanding of variable interactions is key to making precise predictions. Additionally, the judicious placement of knots within the data allows for refined adjustments to the model, striking an optimal balance between avoiding overfitting and underfitting, thus bolstering the model's predictive power while maintaining clarity and interpretability.

Simultaneously, machine learning has transformed predictive modeling through its sophisticated algorithms that glean insights from data, uncovering intricate patterns and relationships. Capable of processing and learning from vast datasets, including unstructured data like images, text, and sensor data, machine learning has broadened the horizons of predictive modeling. Its application across various sectors—from financial trend prediction and healthcare outcome forecasting to personalized e-commerce experiences—demonstrates its versatility and capability to deliver highly accurate predictions. The iterative improvement of predictions with new data, facilitated by techniques like regression, classification, and neural networks, allows machine learning models to remain relevant in ever-changing environments.

The integration of splines and machine learning into predictive modeling offers a comprehensive toolkit for data analysts. While splines provide a detailed and interpretable model of nonlinear relationships, machine learning offers unparalleled predictive accuracy across complex datasets. This synergy between the two approaches furnishes the predictive modeling field with tools that are not only adaptable and precise but also interpretable, meeting a wide array of analytical needs. Consequently, the combination of splines and machine learning propels the advancement of data analysis, catering to both the requirement for deep understanding and the demand for predictive accuracy across various disciplines.

In conclusion, the journey from the foundational principles of splines through to the practical considerations of their implementation encapsulates a broader narrative on the progression of statistical methods in research. It underscores the ongoing need for methods that can adapt to the complexity of data while providing insights that are both accurate and interpretable. As we continue to push the boundaries of what is possible with data analysis, the thoughtful application of splines and similar techniques will remain a crucial part of the statistical toolkit, bridging the gap between theoretical models and the multifaceted reality they seek to explain. In this way, the exploration of splines not only enriches our understanding of data but also empowers researchers to uncover deeper truths within their fields of inquiry, demonstrating the enduring value of marrying complex mathematical models with the nuanced intricacies of the natural world.

Author details

Alexander A. Huang[1,2*] and Samuel Y. Huang[1,3]

1 Cornell University, USA

2 Northwestern University Feinberg School of Medicine, USA

3 Icahn School of Medicine at Mount Sinai South Nassau, USA

*Address all correspondence to: alexander.huang@northwestern.edu

IntechOpen

© 2024 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

[1] Agyar O, Tirink C, Onder H, Sen U, Piwczynski D, Yavuz E. Use of multivariate adaptive regression splines algorithm to predict body weight from body measurements of Anatolian buffaloes in Turkiye. Animals (Basel). 2022;**12**(21):50-70. DOI: 10.3390/ani12212923

[2] Alavi J, Aminikhah H. Orthogonal cubic splines for the numerical solution of nonlinear parabolic partial differential equations. MethodsX. 2023;**10**:102190. DOI: 10.1016/j.mex.2023.102190

[3] Athey TL, Teneggi J, Vogelstein JT, Tward DJ, Mueller U, Miller MI. Fitting splines to axonal arbors quantifies relationship between branch order and geometry. Frontiers in Neuroinformatics. 2021;**15**:704627. DOI: 10.3389/fninf.2021.704627

[4] Austin PC, Fang J, Lee DS. Using fractional polynomials and restricted cubic splines to model non-proportional hazards or time-varying covariate effects in the cox regression model. Statistics in Medicine. 2022;**41**(3):612-624. DOI: 10.1002/sim.9259

[5] Bach NH, Vu LH, Nguyen VD, Pham DP. Classifying marine mammals signal using cubic splines interpolation combining with triple loss variational auto-encoder. Scientific Reports. 2023;**13**(1):19984. DOI: 10.1038/s41598-023-47320-4

[6] Azzolina D, Berchialla P, Bressan S, Da Dalt L, Gregori D, Baldi I. A Bayesian sample size estimation procedure based on a B-splines semiparametric elicitation method. International Journal of Environmental Research and Public Health. 2022;**19**(21):100-120. DOI: 10.3390/ijerph192114245

[7] Dyrting S, Taylor A. Estimating age-specific mortality using calibrated splines. Population Studies. 2023;**1**:1-18. DOI: 10.1080/00324728.2023.2228297

[8] Ebrahimi-Khusfi Z, Nafarzadegan AR, Khosroshahi M. Using multivariate adaptive regression splines and extremely randomized trees algorithms to predict dust events frequency around an international wetland and prioritize its drivers. Environmental Monitoring and Assessment. 2021;**193**(7):437. DOI: 10.1007/s10661-021-09198-5

[9] Elhakeem A et al. Using linear and natural cubic splines, SITAR, and latent trajectory models to characterise nonlinear longitudinal growth trajectories in cohort studies. BMC Medical Research Methodology. 2022;**22**(1):68. DOI: 10.1186/s12874-022-01542-8

[10] Feng X et al. Relationship between body mass index and kidney stones based on dose-response analyses using restricted cubic splines applied to NHANES 2011-2016 data. Journal of Renal Nutrition. 2021;**31**(3):263-269. DOI: 10.1053/j.jrn.2020.05.003

[11] Gascoigne C, Smith T. Penalized smoothing splines resolve the curvature identifiability problem in age-period-cohort models with unequal intervals. Statistics in Medicine. 2023;**42**(12):1888-1908. DOI: 10.1002/sim.9703

[12] Chen L, de Borst R. Analysis of progressive fracture in fluid-saturated porous medium using splines. International Journal for Numerical Methods in Engineering. 2023;**124**(1):264-281. DOI: 10.1002/nme.7120

[13] D'Urso P, De Giovanni L, Vitale V. Spatial robust fuzzy clustering of COVID 19 time series based on B-splines. Spatial Statistics. 2022;**49**:100518. DOI: 10.1016/j.spasta.2021.100518

[14] Dantony E et al. Multidimensional penalized splines for survival models: Illustration for net survival trend analyses. International Journal of Epidemiology. 2024;**53**(2):132-153. DOI: 10.1093/ije/dyae033

[15] Bantis LE, Tsimikas JV, Georgiou SD. Survival estimation through the cumulative hazard with monotone natural cubic splines using convex optimization-the HCNS approach. Computer Methods and Programs in Biomedicine. 2020;**190**:105357. DOI: 10.1016/j.cmpb.2020.105357

[16] Bekar Adiguzel M, Cengiz MA. Model selection in multivariate adaptive regressions splines (MARS) using alternative information criteria. Heliyon. 2023;**9**(9):e19964. DOI: 10.1016/j.heliyon.2023.e19964

[17] Belias M, Rovers MM, Hoogland J, Reitsma JB, Debray TPA, IntHout J. Predicting personalised absolute treatment effects in individual participant data meta-analysis: An introduction to splines. Research Synthesis Methods. 2022;**13**(2):255-283. DOI: 10.1002/jrsm.1546

[18] Celma A, Bade R, Sancho JV, Hernandez F, Humphries M, Bijlsma L. Prediction of retention time and collision cross section (CCS(H+), CCS(H-), and CCS(Na+)) of emerging contaminants using multiple adaptive regression splines. Journal of Chemical Information and Modeling. 2022;**62**(22):5425-5434. DOI: 10.1021/acs.jcim.2c00847

[19] Uhry Z et al. Multidimensional penalized splines for incidence and mortality-trend analyses and validation of national cancer-incidence estimates. International Journal of Epidemiology. 2020;**49**(4):1294-1306. DOI: 10.1093/ije/dyaa078

[20] Wang J et al. Stitching locally fitted T-splines for fast fitting of large-scale freeform point clouds. Sensors (Basel). 2023;**23**(24):52-53. DOI: 10.3390/s23249816

[21] Whetten AB. Smoothing splines of apex predator movement: Functional modeling strategies for exploring animal behavior and social interactions. Ecology and Evolution. 2021;**11**(24):17786-17800. DOI: 10.1002/ece3.8294

[22] Xu Y et al. Using restricted cubic splines to study the duration of antibiotic use in the prognosis of ventilator-associated pneumonia. Frontiers in Pharmacology. 2022;**13**:898630. DOI: 10.3389/fphar.2022.898630

[23] Zheng S et al. Using restricted cubic splines to study the trajectory of systolic blood pressure in the prognosis of acute myocardial infarction. Frontiers in Cardiovascular Medicine. 2021;**8**:740580. DOI: 10.3389/fcvm.2021.740580

[24] Huang AA, Huang SY. Quantification of the relationship of pyridoxine and spirometry measurements in the United States population. Current Developments in Nutrition. 2023;**7**(8):100078. DOI: 10.1016/j.cdnut.2023.100078

[25] Huang AA, Huang SY. Quantification of the effect of vitamin E intake on depressive symptoms in United States adults using restricted cubic splines. Current Developments in Nutrition. 2023;**7**(2):100038. DOI: 10.1016/j.cdnut.2023.100038

[26] Huang AA, Huang SY. Use of machine learning to identify risk factors for coronary artery disease. PLoS ONE. 2023;**18**(4):e0284103. DOI: 10.1371/journal.pone.0284103

[27] Wu TE, Chen JW, Liu TC, Yu CH, Jhou MJ, Lu CJ. Identifying and exploring the impact factors for intraocular pressure prediction in myopic children with atropine control utilizing multivariate adaptive regression splines. Journal of Personalized Medicine. 2024;**14**(1):11-24. DOI: 10.3390/jpm14010125

[28] Xu J, Hou Q, Qu K, Sun Y, Meng X. A fast weighted fuzzy C-medoids clustering for time series data based on P-splines. Sensors (Basel). 2022;**22**(16):34-52. DOI: 10.3390/s22166163

[29] Hamza T, Furukawa TA, Orsini N, Cipriani A, Iglesias CP, Salanti G. A dose-effect network meta-analysis model with application in antidepressants using restricted cubic splines. Statistical Methods in Medical Research. 2022;**1**:9622802211070256. DOI: 10.1177/09622802211070256

[30] Noakes L. Planar interpolation by second order spiral splines. MethodsX. 2020;**7**:100776. DOI: 10.1016/j.mex.2019.100776

[31] Pandey R, Tolani H. Penalized splines model to estimate time-varying reproduction number for Covid-19 in India: A Bayesian semi-parametric approach. Clinical Epidemiology and Global Health. 2022;**18**:101176. DOI: 10.1016/j.cegh.2022.101176

[32] Piatek M, Lisowski A, Dabrowska M. The effects of solid lignin on the anaerobic digestion of microcrystalline cellulose and application of smoothing splines for extended data analysis of its inhibitory effects. Bioresource Technology. 2021;**320**(Pt A):124262. DOI: 10.1016/j.biortech.2020.124262

[33] Gogel B, Welham S, Cullis B. Empirical comparison of time series models and tensor product penalised splines for modelling spatial dependence in plant breeding field trials. Frontiers in Plant Science. 2022;**13**:1021143. DOI: 10.3389/fpls.2022.1021143

[34] Ammothum Kandy AK, Wadbro E, Aradi B, Broqvist P, Kullgren J. Curvature constrained splines for DFTB repulsive potential parametrization. Journal of Chemical Theory and Computation. 2021;**17**(3):1771-1781. DOI: 10.1021/acs.jctc.0c01156

[35] Huang A et al. Lack of compensation of energy intake explains the success of alternate day feeding to produce weight loss. Physiology & Behavior. 2023;**263**:114128. DOI: 10.1016/j.physbeh.2023.114128

[36] Huang AA, Huang SY. Use of feature importance statistics to accurately predict asthma attacks using machine learning: A cross-sectional cohort study of the US population. PLoS ONE. 2023;**18**(11):e0288903. DOI: 10.1371/journal.pone.0288903

[37] Huang AA, Huang SY. Use of machine learning to identify risk factors for insomnia. PLoS ONE. 2023;**18**(4):e0282622. DOI: 10.1371/journal.pone.0282622

[38] Heinecke A, Tallarita M, De Iorio M. Bayesian splines versus fractional polynomials in network meta-analysis. BMC Medical Research Methodology. 2020;**20**(1):261. DOI: 10.1186/s12874-020-01113-9

[39] Jover IL, Debarre T, Aziznejad S, Unser M. Coupled splines for sparse curve fitting. IEEE Transactions on

Image Processing. 2022;**31**:4707-4718. DOI: 10.1109/TIP.2022.3187286

[40] Karciauskas K, Peters J. Low degree splines for locally quad-dominant meshes. Computer Aided Geometric Design. 2020;**83**:32-53. DOI: 10.1016/j.cagd.2020.101934

[41] Mubarik S, Hu Y, Yu C. A multi-country comparison of stochastic models of breast cancer mortality with P-splines smoothing approach. BMC Medical Research Methodology. 2020;**20**(1):299. DOI: 10.1186/s12874-020-01187-5

[42] Munoz-Osorio GA et al. Using fat thickness and longissimus thoracis traits real-time ultrasound measurements in Black Belly ewe lambs to predict carcass tissue composition through multiresponse multivariate adaptive regression splines algorithm. Meat Science. 2024;**207**:109369. DOI: 10.1016/j.meatsci.2023.109369

[43] Mushtaq K et al. Multivariate wind power curve modeling using multivariate adaptive regression splines and regression trees. PLoS ONE. 2023;**18**(8):e0290316. DOI: 10.1371/journal.pone.0290316

[44] Nacar S, Mete B, Bayram A. Estimation of daily dissolved oxygen concentration for river water quality using conventional regression analysis, multivariate adaptive regression splines, and TreeNet techniques. Environmental Monitoring and Assessment. 2020;**192**(12):752. DOI: 10.1007/s10661-020-08649-9

[45] Tirink C et al. Use of multivariate adaptive regression splines for prediction of body weight from body measurements in Marecha (*Camelus dromedaries*) camels in Pakistan. Tropical Animal Health and Production. 2021;**53**(3):339. DOI: 10.1007/s11250-021-02788-y

[46] Zhou Z, Zhang R, Zhu Z. Retraction notice to "Uncalibrated dynamic visual servoing via multivariate adaptive regression splines and improved incremental extreme learning machine" [ISA Transactions 92 (2019) 298-314]. ISA Transactions. 2020;**98**:505. DOI: 10.1016/j.isatra.2020.02.021

[47] Lamichhane BP. A mixed finite element discretisation of linear and nonlinear multivariate splines using the Laplacian penalty based on biorthogonal systems. MethodsX. 2023;**10**:101962. DOI: 10.1016/j.mex.2022.101962

[48] Yang Y et al. The relationship between ventilatory ratio (VR) and 28-day hospital mortality by restricted cubic splines (RCS) in 14,328 mechanically ventilated ICU patients. BMC Pulmonary Medicine. 2022;**22**(1):229. DOI: 10.1186/s12890-022-02019-6

[49] Momen M, Campbell MT, Walia H, Morota G. Predicting longitudinal traits derived from high-throughput phenomics in contrasting environments using genomic Legendre polynomials and B-splines. G3 : Genes - Genomes - Genetics. 2019;**9**(10):3369-3380. DOI: 10.1534/g3.119.400346

[50] Huang SY, Johnathan R, Shah N, Srivastava P, Huang AA, Gress F. Technical report: Protocol for characterizing phenotype variants using phenome-wide association study (PheWAS) utilizing the nationwide inpatient sample 2020 in individuals with pancreatic cysts and lung cancer. Cureus. 2023;**15**(12):e50982. DOI: 10.7759/cureus.50982

[51] Huang AA, Huang SY. Increasing transparency in machine learning through bootstrap simulation and shapely additive explanations.

PLoS ONE. 2023;**18**(2):e0281922. DOI: 10.1371/journal.pone.0281922

[52] Huang AA, Huang SY. Hospitalized COVID-19 patients with diabetes have an increased risk for pneumonia, intensive care unit requirement, intubation, and death: A cross-sectional cohort study in Mexico in 2020. Health Science Reports. 2023;**6**(4):e1222. DOI: 10.1002/hsr2.1222

[53] Huang AA, Huang SY. Computation of the distribution of model accuracy statistics in machine learning: Comparison between analytically derived distributions and simulation-based methods. Health Science Reports. 2023;**6**(4):e1214. DOI: 10.1002/hsr2.1214

[54] Huang AA, Huang SY. Diabetes is associated with increased risk of death in COVID-19 hospitalizations in Mexico 2020: A retrospective cohort study. Health Science Reports. 2023;**6**(7):e1416. DOI: 10.1002/hsr2.1416

[55] Huang AA, Huang SY. Dendrogram of transparent feature importance machine learning statistics to classify associations for heart failure: A reanalysis of a retrospective cohort study of the Medical Information Mart for Intensive Care III (MIMIC-III) database. PLoS ONE. 2023;**18**(7):e0288819. DOI: 10.1371/journal.pone.0288819

[56] Huang AA, Huang SY. Increased vigorous exercise and decreased sedentary activities are associated with decreased depressive symptoms in United States adults: Analysis of The National Health and Nutrition Examination Survey (NHANES) 2017-2020. Health Science Reports. 2023;**6**(8):e1473. DOI: 10.1002/hsr2.1473

[57] Huang AA, Huang SY. Exploring depression and nutritional covariates amongst US adults using shapely additive explanations. Health Science Reports. 2023;**6**(10):e1635. DOI: 10.1002/hsr2.1635

[58] Huang AA, Huang SY. Technical report: Machine-learning pipeline for medical research and quality-improvement initiatives. Cureus. 2023;**15**(10):e46549. DOI: 10.7759/cureus.46549

[59] Huang AA, Huang SY. Shapely additive values can effectively visualize pertinent covariates in machine learning when predicting hypertension. Journal of Clinical Hypertension (Greenwich, Conn.). 2023;**25**(12):1135-1144. DOI: 10.1111/jch.14745

[60] Huang AA, Huang SY. Covariate dependent Markov chains constructed with gradient boost modeling can effectively generate long-term predictions of obesity trends. BMC Research Notes. 2023;**16**(1):346. DOI: 10.1186/s13104-023-06610-w

[61] Huang AA, Huang SY. Stochastic modeling of obesity status in United States adults using Markov chains: A nationally representative analysis of population health data from 2017-2020. Obesity Science and Practice. 2023;**9**(6):653-660. DOI: 10.1002/osp4.697

Section 2

Matrix Analysis

Chapter 12

Introductory Chapter: The Matrices, Their History, Importance and Applications

Victor Martinez-Luaces

1. Introduction and history

Matrices historically began as rectangular arrangements of numbers, although currently their elements are not necessarily numbers and they are not necessarily rectangular either. Indeed, in Analysis, there are matrices such as the Jacobian or the Hessian whose entries are not numbers but functions [1]. Regarding its rectangular shape, this is not always true, for instance, there exist triangular-shaped matrices (also called tables), which have their theory and close relationships with other areas of mathematics [2].

Since ancient times, these mathematical objects have been fundamentally linked to the study and resolution of linear systems of equations. They appeared in ancient Chinese writings [3] dating back more than 2200 years. Moreover, some magic squares have been known for over 25 centuries [4].

That is to say, it seems that its origin is closely linked either to the simultaneous resolution of linear equations or to certain mathematical curiosities, mainly with recreational interests.

2. Importance and applications of matrices

As usually happens, mathematical objects widely overtook those that could have been their initial purposes, and the matrices are precisely one of the most paradigmatic examples of this phenomenon. In fact, matrices have invaded practically all areas of mathematics.

In Linear Algebra, several of the most important methods for solving linear systems of equations, such as the Gauss method, Gauss-Jordan, or using the inverse matrix, are ultimately matrix methods. In addition to the above, the matrices have an obvious connection with linear transformations, quadratic forms, bilinear applications, and determinants, and they play a fundamental role in some processes such as diagonalization, Cramer's method, orthogonal diagonalization, and change of basis in finite-dimensional vector spaces [5]. Besides, some algebraic objects such as permanents, minors, characteristic polynomials, and the trace of a matrix, among others, are also defined from matrices.

In mathematical analysis, matrices also play an essential role, and it is enough to mention a couple of well-known examples—the Jacobian $\mathbb{J}_f$ and the Hessian $\mathbb{H}_f$ —and

their importance in the search and classification of relative extremes of a given function of several variables, the chain rule and the Taylor expansion, among many other applications [6].

In several branches of Geometry, matrices are also very important. For instance, in Geometry in $\mathbb{R}^2$ and $\mathbb{R}^3$, matrices are used to represent rigid movements (rotations, translations, symmetries, etc.); also they are relevant in the study of conics and quadrics. Matrices also provide simple methods to perform some operations between vectors in the space $\mathbb{R}^3$, as is the case of the vector product and the mixed product, which also have intrinsic meanings and geometric applications. In Differential Geometry, the vector product—along with the scalar product—has a primary role in the definition of the of the Frenet Trihedron's vectors and in the Fundamental Theorem of the Theory of Curves [7]. It should be noted that in the proof of that theorem, the involved matrix relates the derivatives of the tangent, normal and binormal vectors with those same vectors of the aforementioned trihedron.

In the theory of Ordinary Differential Equations (ODE), matrices naturally appear in first-order linear ODE systems. If the given ODE system is homogeneous, it is enough to put it in its matrix form, that is $\mathbb{X}' = \mathbb{A}.\mathbb{X}$ and its solution will be of the form $\mathbb{X}(t) = \exp(t\,\mathbb{A}).\mathbb{X}_0$ being the vector $\mathbb{X}_0$ corresponding to the initial condition. In other words, any homogeneous linear ODE system is solved by using the exponential of matrices [8]. It is important to note that they also could be solved by decoupling the system through a diagonalization process or semi-decoupling it using Jordan's Canonical Form (in case the system's matrix is not diagonalizable).

If a non-homogeneous linear ODE system is considered, one of the most typical methods is the constants variation method, or indeterminate coefficients method, which again uses matrices, particularly the Wronskian matrix $\mathbb{W}(t)$ and its determinant $W(t)$.

Additionally, another alternative approach for homogeneous and non-homogeneous linear ODE systems is to solve it by using Laplace Transform [9], which converts the ODE system into a linear system of algebraic equations, so once again, it can be solved by matrix methods.

If we consider the Numerical Methods, matrices are also very relevant. For instance, in Numerical Linear Algebra, the Jacobi and Gauss-Seidel methods are in fact matrix methods [10]. If the given system of equations is non-linear, one of the fundamental methods is Newton-Raphson, which works with matrices, in particular with the already mentioned Jacobian Matrix. In a different area of Numerical Methods, when solving second-order EDPs, both parabolic and hyperbolic, it usually appears tridiagonal matrices both in the finite difference methods (Euler explicit and implicit, Crank-Nicolson and others) as in finite element methods based on piecewise linear functions with a compact support. On the other hand, when discretizing elliptic equations—specifically when solving the Poisson equation by a method of finite differences of second order on an equally spaced mesh—the resulting matrix is much more complicated and is made up of tridiagonal submatrices surrounded by copies of the identity matrix. In short, beyond the complexity of the matrices, all these methods lead to linear systems of algebraic equations or linear ODE systems, depending on the methodology followed.

As it can be expected, also in Probability, Statistics, Experimental Design and Stochastic Methods, the matrices are fundamental. Some examples of the previous statement are the Normal Equations in Linear Regression [11], the Latin and Greco-Latin Squares in Experimental Design [12] and the Stochastic Matrices in Markov Chains [13], among many others that could be mentioned.

In Discrete Mathematics, matrices also have a lot to say. As an example, in graph theory, matrices are used to represent a given graph through its adjacency matrix. Moreover, the positive powers of that matrix provide the number of paths of a certain length between two given vertices [14]. Another example in Discrete Mathematics is related to the study of functions and relations. Indeed, the transitive closure of a relation can be carried out by means of the Warshall algorithm [15], which is nothing more than a matrix method. A concrete application of the latter involves obtaining the minimum equivalence relation that contains the given relation, that is, the so-called equivalence closure.

All of the above shows only some of the best-known examples of matrices' applications in different branches of mathematics.

Taking into account that Matrix Theory transversally crosses almost all other branches of mathematics, it is not surprising that matrices also make their appearance in nearly all scientific disciplines.

Hence, for example, in experimental sciences such as Chemistry, Physics and Biology, matrices have a relevant role. In Chemical Kinetics, for example, certain special matrices appear associated with the chemical mechanisms consisting of first-order reactions (so-called FOCKM-matrices) [16]. Something similar happens with mixture problems, which are modeled by first-order ODE linear systems, giving rise to other types of matrices (so-called MP-matrices) [17].

Similarly, it happens in the social sciences. For example, in Economics, the so-called payment matrices are used that report profits or losses associated with certain situations [18]. Matrices are also utilized to optimize economic variables by linear and non-linear programming methods [19]. Another example takes place in Sociology where directed graphs are used to study relationships among large groups of individuals which in turn are represented by adjacency matrices [20]. In Education, when applying didactic analysis, the different variables can be encoded in Boolean form (by using zeros and ones, indicating the presence or absence of a certain characteristic) which can then be used as the feed for cluster analysis [21].

It is worth mentioning that various disciplines apply the design of experiments and the subsequent treatment of data, to different experimental sciences like Chemistry or Biology, and Economics, Sociology and Psychology among the social sciences. These mixed disciplines are called Chemometrics, Biometrics, Econometrics, Sociometrics and Psychometrics, respectively. In a first approach, it might seem that all of them are basically the same applications of Statistics and Experimental Design to different areas of knowledge; however, the shape of the data matrices in each case is decisive regarding the methods that are preferably used. For example, in Chemistry, usually, there are just a few samples, or runs, but sometimes in each of them, the absorbances are measured between 200 and 700 nm, that is, the data matrix has, for example, a number of rows of one or at most two digits, whereas the number of columns is around several hundred. So, this gives a data matrix having a rectangular shape with a wide base and low height, whereas in the social sciences, usually, exactly the opposite happens. For instance, in order to predict the vote of the population in a presidential election, several thousand citizens should be interviewed, and at the same time, the variables usually considered are very few: gender, age, education level, income level, place where the person lives and perhaps one or two more. Consequently, the data matrix will have hundreds or even thousands of rows, but just a few columns.

Due to these facts, the shape of the data matrix—as a rectangle with a wide base and low height or a narrow base and an important height—will determine the

methodological choice. For instance, if the researcher has a small number of experimental runs, the asymptotic properties of the estimators are almost irrelevant, making it more desirable to use unbiased or minimum variance estimators. In conclusion, these disciplines—for example, Chemometrics and Sociometry—both use statistical and experimental design methods, although they differ in the selection of the methods to be applied and this is mainly due to the different shapes of their data matrices.

Thus, if matrices play an important role in almost all branches of mathematics and scientific disciplines, it is not surprising to observe that they are also extremely relevant in engineering and technology due to their applications. Just to mention a few examples, the matrices have been used to optimize a drug manufacturing process [22], the treatment of atmospheric corrosion data [23], image processing [24] and to study and improve the production of a protein for immunodiagnostics [25], just to mention a few previous experiences where matrices have been successfully utilized. Matrices also have applications in graphic design, route planning, computer animation, data analysis and machine learning, facial recognition, cryptography and cybersecurity, financial analysis and risk management and portfolio evaluation, among many other applications [26].

In a few words, matrices are among the most useful existing mathematical tools. As an example, it is enough to mention that there are estimates that suggest that more than 75% of scientific, industrial or engineering problems involve, at some stage, the resolution of a linear system of equations [27], which in turn is just one of the many matrix applications.

Consequently, due to the importance of matrices and their enormous applicability in science, engineering, economics and industry, among other areas, a permanent update on this topic is essential. This update needs to reflect the advances in Matrix Theory and its applications in the most diverse disciplines.

These have been the main ideas when proposing this volume: updating the subject theoretical aspects and at the same time, providing new research in different areas that complement our knowledge about the immense range of applications of Matrix Theory.

Author details

Victor Martinez-Luaces
University of the Republic of Uruguay, Montevideo, Uruguay

*Address all correspondence to: victorml@fing.edu.uy

IntechOpen

© 2024 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

[1] Polanco C. Advanced Calculus: Fundamentals of Mathematics. Singapore: Bentham Science Publishers; 2019. 215 p

[2] Zatorsky R. Introduction to the theory of triangular matrices (tables). In: Kyrchei I, editor. Advances in Linear Algebra Research. New York: Nova Science Publishers; 2015. pp. 185-238

[3] Shen K, Crossley JN, Lun AW-C. Nine Chapters of the Mathematical Art, Companion and Commentary. 2nd ed. Beijing: Oxford University Press; 1999. 596 p

[4] Swaney M. History of Magic Squares. [Internet] 2000Available from: https://arthurmag.com/2004/02/16/na-84/ [Accessed: March 06, 2024]

[5] Anton H, Rorres C, Kaul A. Elementary Linear Algebra. 12th ed. Hoboken NJ: John Wiley & Sons; 2018. 800 p

[6] Lovric M. Vector Calculus. Hoboken NJ: John Wiley & Sons; 2007. 623 p

[7] Kühnel W. Differential Geometry. 2nd ed. Providence RI: American Mathematical Society; 2006. 380 p

[8] Chicone C. Ordinary Differential Equations with Applications. 2nd ed. New York: Springer Science & Business Media; 2008. 636 p

[9] Adkins WA, Davidson MG. Ordinary Differential Equations. New York: Springer Science & Business Media; 2012. 799 p

[10] Epperson JF. An Introduction to Numerical Methods and Analysis. 2nd ed. Hoboken NJ: John Wiley & Sons; 2013. 614 p

[11] Monahan JF. A Primer on Linear Models. Boca Raton FL: CRC Press; 2008. 304 p

[12] Montgomery DC. Design and Analysis of Experiments. 9th ed. Hoboken NJ: John Wiley & Sons; 2017. 734 p

[13] Neuts MF. Structured Stochastic Matrices of M/G/1 Type and their Applications. Boca Raton FL: CRC Press; 2021. 512 p

[14] Aldous JM, Wilson RJ. Graphs and Applications: An Introductory Approach. London: Springer; 2000. 444 p

[15] Sridharan S, Balakrishnan R. Discrete Mathematics: Graph Algorithms, Algebraic Structures, Coding Theory, and Cryptography. Boca Raton FL: CRC Press; 2019. 340 p

[16] Martinez-Luaces V. Concentration curves in first order chemical kinetics: All the possible cases. In: Martinez-Luaces V, editor. A Closer Look at Chemical Kinetics. New York: Nova Science Publishers; 2023. pp. 13-28

[17] Martinez-Luaces V. Square matrices associated to mixing problems ODE Systems. In: Yasser HA, editor. Matrix Theory. London: InTechOpen Science; 2018. pp. 41-58

[18] Ekel P, Pedrycz W, Pereira J. Multicriteria Decision-Making under Conditions of Uncertainty: A Fuzzy Set Perspective. Hoboken NJ: John Wiley & Sons; 2019. 368 p

[19] Franklin JN. Methods of Mathematical Economics: Linear and Nonlinear Programming, Fixed-Point

Theorems. Berlin Heidelberg: Springer; 2013. 299 p

[20] Vasantha Kandasamy WB, Ilanthenral K, Smarandache F. Subset Vertex Graphs for Social Networks. Bruxelles: EuropaNova ASBL; 2018. 290 p

[21] Martinez-Luaces V. Posing Inverse Modeling Problems for Task Enrichment in a Secondary Mathematics Teachers Training Program. Granada: University of Granada; 2021. 269 p

[22] Martinez-Luaces V, Ohanian M. Drug production optimization using experimental design methods. In: Mellal MA, editor. Manufacturing Systems. New York: Nova Science Publishers; 2020. pp. 365-380

[23] Díaz V, Martinez-Luaces V, Guineo-Cobs G. Corrosión Atmosférica: validación de modelos empleando técnicas estadísticas. Meta. 2003;**39**(4):243-251

[24] Martinez-Luaces M, Martinez-Luaces V. Scene analysis as a source of statistical and numerical problems: Experiences in informatics' engineering courses. In: D'Arcy A, Martinez-Luaces V, Oates G, Varsavsky C, editors. Vision and Change for a New Century. Montevideo: ISC-Delta; 2007. pp. 105-110

[25] Martinez-Luaces V, Guineo-Cobs G, Velazquez B, Chabalgoity A, Massaldi H. Bifactorial design applied to recombinant protein expressions. Journal of Data Science. 2006;**4**(2):247-255

[26] Matematizame.com. Ejemplos prácticos de aplicaciones de matrices en la vida real [Internet]. 2023. Available from: https://matematizame.com/ejemplos-practicos-de-aplicaciones-de-matrices-en-la-vida-real/ [Accessed: March 06, 2024]

[27] Butt R. An Introduction to Numerical Methods and Analysis Using MATLAB [Internet]. 2021. Available from: https://faculty.ksu.edu.sa/sites/default/files/LL-M254.pdf [Accessed: March 06, 2024]

Chapter 13

Eigenvalues of Matrices in Chemical Kinetics and Their Algebraic and Geometric Multiplicities

Victor Martinez-Luaces

Abstract

Every mechanism or system of chemical reactions gives rise to a system of ordinary differential equations when the variation of concentrations with respect to time is studied. Furthermore, if such reactions are all first-order kinetic reactions, then a linear system of differential equations is obtained and its associated matrix has special properties. In particular, the matrix eigenvalues and their algebraic and geometric multiplicities determine the form of the solutions as well as their qualitative behavior. In this chapter, the theoretical nine possible cases are analyzed, and it is proved that all but one can occur experimentally, and examples are provided. For those eight cases that can take place, the stability and asymptotic stability of the solutions are studied.

Keywords: chemical kinetics matrices, eigenvalues, algebraic multiplicity, geometric multiplicity, stability of ODE systems

1. Introduction

There is a very important relationship between mathematics and chemistry [1, 2], particularly when modeling chemical problems using ordinary differential equations (ODE) and ODE systems [3, 4]. When the kinetics of any mechanism formed by chemical reactions is studied, the corresponding mathematical model gives rise to a differential equation or a system of differential equations, depending on the number of species involved [5–7]. The simplest example occurs when a chemical species S_1 gives several products P_1, P_2, and so on, in such a way that the reaction rate is directly proportional to the remaining concentration of the reactant S_1. This situation is illustrated in **Figure 1**.

The mathematical model corresponding to this first-order chemical kinetics problem is

$$\frac{d[S_1]}{dt} = -k[S_1] \tag{1}$$

IntechOpen

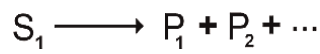


Figure 1.
First-order chemical reaction.

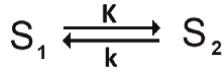


Figure 2.
Opposed reactions.

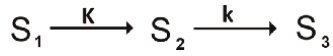


Figure 3.
Consecutive reactions.

where $[S_1]$ is the concentration of the species S_1, t is time, and k is the kinetic constant of the reaction.

A more interesting situation occurs if two chemical species S_1 and S_2 are present, such that S_1 is transformed into S_2 by means of a first-order chemical reaction (FOCR), with kinetic constant K and, at the same time, S_2 is transformed into S_1 by another FOCR with kinetic constant k, as is shown in **Figure 2**.

In this case, we have a first-order chemical kinetic mechanism (FOCKM), which consists of a set of reactions, all of them being FOCRs. This new problem gives rise to the following ODE system:

$$\begin{cases} \dfrac{d[S_1]}{dt} = -K[S_1] + k[S_2] \\ \dfrac{d[S_2]}{dt} = K[S_1] - k[S_2] \end{cases} \text{ or } \frac{d}{dt}\begin{pmatrix} [S_1] \\ [S_2] \end{pmatrix} = \begin{pmatrix} -K & k \\ K & -k \end{pmatrix}\begin{pmatrix} [S_1] \\ [S_2] \end{pmatrix} \tag{2}$$

We will call FOCKM matrix

$$\mathbb{A} = \begin{pmatrix} -K & k \\ K & -k \end{pmatrix} \tag{3}$$

to the ODE system associated matrix, whose eigenvalues and eigenvectors can be easily computed as $\lambda_1 = 0$, $\lambda_2 = -K - k$ and $\vec{v_1} = (k \;\; K)$, $\vec{v_2} = (-1 \;\; 1)$.

In other words, this case—known as opposed reactions (see [7])—gives rise to a couple of different eigenvalues (each with algebraic multiplicity one) and two one-dimensional eigenspaces (i.e., each with geometric multiplicity one).

Now suppose that a chemical species S_1 produces another chemical species S_2 and this in turn gives S_3. In this case, we have consecutive reactions (e.g., [7]), which is illustrated in **Figure 3**.

Hence, the corresponding mathematical model is

$$\begin{cases} \dfrac{d[S_1]}{dt} = -K[S_1] \\ \dfrac{d[S_2]}{dt} = K[S_1] - k[S_2] \\ \dfrac{d[S_3]}{dt} = k[S_2] \end{cases} \text{ or } \frac{d}{dt}\begin{pmatrix} [S_1] \\ [S_2] \\ [S_3] \end{pmatrix} = \begin{pmatrix} -K & 0 & 0 \\ K & -k & 0 \\ 0 & k & 0 \end{pmatrix}\begin{pmatrix} [S_1] \\ [S_2] \\ [S_3] \end{pmatrix} \tag{4}$$

Figure 4.
Competitive reactions.

In this case, the FOCKM matrix is

$$\mathbb{A} = \begin{pmatrix} -K & 0 & 0 \\ K & -k & 0 \\ 0 & k & 0 \end{pmatrix}, \tag{5}$$

and since it is a lower triangular matrix, its eigenvalues are $\lambda_1 = -K$, $\lambda_2 = -k$, and $\lambda_3 = 0$. Since all are simple eigenvalues, it results in $AM_{\lambda_i} = 1, i = 1, 2, 3$, and $GM_{\lambda_i} = 1, i = 1, 2, 3$; that is, all the algebraic and geometric multiplicities are one. Moreover, it is easy to obtain $\vec{v_1} = (\, k - K \quad K \quad -k \,)$, $\vec{v_2} = (\, 0 \quad -1 \quad 1 \,)$ and $\vec{v_3} = (\, 0 \quad 0 \quad 1 \,)$, the corresponding eigenvectors.

The last case to be analyzed in this introduction is the one that occurs when a chemical species S_1 gives S_2 and at the same time, in parallel, S_1 also gives S_3. These are called competitive reactions (see [7]) and can be illustrated as in **Figure 4**.

In this case, the ODE system is the following:

$$\begin{cases} \dfrac{d[S_1]}{dt} = -(K+k)[S_1] \\ \dfrac{d[S_2]}{dt} = K[S_1] \\ \dfrac{d[S_3]}{dt} = k[S_1] \end{cases} \quad \text{or} \quad \frac{d}{dt}\begin{pmatrix} [S_1] \\ [S_2] \\ [S_3] \end{pmatrix} = \begin{pmatrix} -(K+k) & 0 & 0 \\ K & 0 & 0 \\ k & 0 & 0 \end{pmatrix}\begin{pmatrix} [S_1] \\ [S_2] \\ [S_3] \end{pmatrix} \tag{6}$$

In this case, the FOCKM matrix is

$$\mathbb{A} = \begin{pmatrix} -K-k & 0 & 0 \\ K & 0 & 0 \\ k & 0 & 0 \end{pmatrix}. \tag{7}$$

Once again, it is a lower triangular matrix, so its eigenvalues are $\lambda_1 = -K - k$, $\lambda_2 = 0$, and $\lambda_3 = 0$, with $AM_{\lambda_1} = 1$ and $AM_{\lambda=0} = 2$. The corresponding eigenvectors are in this case $\vec{v_1} = (\, -K - k \quad K \quad k \,)$, $\vec{v_2} = (\, 0 \quad 1 \quad 0 \,)$ and $\vec{v_3} = (\, 0 \quad 0 \quad 1 \,)$, so once again, the FOCKM matrix is diagonalizable.

In all the cases seen so far, all the eigenvalues are nonpositive real numbers; they can be single or double, and the FOCKM matrix is always diagonalizable. We will see which of these properties can be proven in general and which cannot, beginning by revisiting some results from previous works [8–10].

2. Some previous results revisited

If we consider again the previous FOCKM matrices,

$$\mathbb{A} = \begin{pmatrix} -K & k \\ K & -k \end{pmatrix}, \tag{8}$$

$$\mathbb{A}' = \begin{pmatrix} -K & 0 & 0 \\ K & -k & 0 \\ 0 & k & 0 \end{pmatrix}, \tag{9}$$

and

$$\mathbb{A}'' = \begin{pmatrix} -K-k & 0 & 0 \\ K & 0 & 0 \\ k & 0 & 0 \end{pmatrix}, \tag{10}$$

it is easy to observe that all non-diagonal entries of a given column are nonnegative (i.e., $a_{i,j} \geq 0, \forall j \neq i$) and the diagonal entry is the sum of all these elements, multiplied by a factor (-1) (i.e., $a_{1,1} = -\sum_{i\neq 1} a_{i,1}, a_{2,2} = -\sum_{i\neq 2} a_{i,2}, \ldots, a_{n,n} = -\sum_{i\neq n} a_{i,n}$).

The following theorem formalizes and generalizes this result:

Theorem 1.

The general form of a FOCKM matrix is as follows:

$$\mathbb{A} = \begin{pmatrix} -\sigma_1 & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & -\sigma_n \end{pmatrix}, \tag{11}$$

with

$$\sigma_1 = \sum_{i\neq 1} a_{i,1}, \ \ldots, \sigma_n = \sum_{i\neq n} a_{i,n}. \tag{12}$$

Proof:

If we consider n chemical species $S_1, S_2, \ldots, S_n$ such that S_i is transformed into S_j by an FOCR $S_i \xrightarrow{k_{ij}} S_j$, k_{ij} being the kinetic constant of this reaction, then all the possible reactions can be schematized as shown in **Figure 5**.

In this diagram, if there is no reaction $S_i \xrightarrow{k_{ij}} S_j$, then we consider $k_{ij} = 0$.

The ODE corresponding to the concentration of the species S_1 can be written as:

$$\frac{d[S_1]}{dt} = -k_{12}[S_1] - k_{13}[S_1] - \ \ldots\ - k_{1n}[S_1] + k_{21}[S_2] + k_{31}[S_3] + \ \ldots\ + k_{n1}[S_n] \tag{13}$$

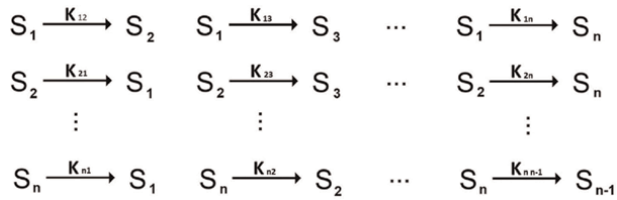


Figure 5.
All the possible reactions between n chemical species.

This ODE can be written as:

$$\frac{d[S_1]}{dt} = -(k_{12} + k_{13} + \ ... \ + k_{1n})[S_1] + k_{21}[S_2] + k_{31}[S_3] + \ ... \ + k_{n1}[S_n] \tag{14}$$

Following the same procedure, we obtain:

$$\frac{d[S_2]}{dt} = k_{12}[S_1] - (k_{21} + k_{23} + \ ... \ + k_{2n})[S_2] + k_{32}[S_3] + \ ... \ + k_{n2}[S_n] \tag{15}$$

$$\frac{d[S_n]}{dt} = k_{1n}[S_1] + k_{2n}[S_2] + \ ... \ + k_{n-1n}[S_{n-1}] - (k_{n1} + k_{n2} + \ ... \ + k_{nn-1})[S_n] \tag{16}$$

The ODE system can be written as:

$$\frac{d}{dt}\begin{pmatrix} [S_1] \\ \vdots \\ [S_n] \end{pmatrix} = \begin{pmatrix} -(k_{12} + k_{13} + \ ... \ + k_{1n}) & \cdots & k_{n1} \\ \vdots & \ddots & \vdots \\ k_{1n} & ... & -(k_{n1} + k_{n2} + \ ... \ + k_{nn-1}) \end{pmatrix}\begin{pmatrix} [S_1] \\ \vdots \\ [S_n] \end{pmatrix} \tag{17}$$

And it is easy to observe that formula (17) corresponds to the form given by matrix $\mathbb{A}$ in Eq. (11) with a different notation. Then, the theorem is proved.

Corollary 1.

The determinant of an FOCKM matrix is zero.

Proof:

Since the general form of an FOCKM matrix is

$$\mathbb{A} = \begin{pmatrix} -\sigma_1 & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & -\sigma_n \end{pmatrix} \tag{18}$$

and

$$\sigma_1 = \sum_{i \neq 1} a_{i,1}, \ ..., \sigma_n = \sum_{i \neq n} a_{i,n}, \tag{19}$$

then $row\ 1 + row\ 2 + \ ... \ + row\ n = \vec{0}$, which can be written as $row\ n = -row\ 1 - row\ 2 - \ ... \ - row(n-1)$. Then, the corollary is proved.

Corollary 2.

For any FOCKM $\mathbb{A}$, $\lambda = 0$ is one of its eigenvalues.

Proof:

It is obvious, since $det(\mathbb{A}) = 0$.

For the proof of the following result, we will use the Gershgorin's circle theorem [11], which establishes that the eigenvalues λ_i of a matrix $\mathbb{A}$ lie within the union of disks $\bigcup_{i=1}^{n} \mathbb{D}_i$, $\mathbb{D}_i$ being the disk of center a_{ii} and radius $R_i = \sum_{j \neq i} |a_{ji}|$, that is, the sum of the modules of the non-diagonal elements. The theorem can be applied to $\mathbb{A}^T$, and in that case, the Gershgorin's disks correspond to the columns of matrix $\mathbb{A}$ [11]. We will use this last version of the theorem.

Theorem 2.

Let $\mathbb{A} = \begin{pmatrix} -\sigma_1 & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & -\sigma_n \end{pmatrix}$ be an FOCKM matrix and let λ_k be an eigenvalue of $\mathbb{A}$. Then, $\lambda_k \in \bigcup_{i=1}^{n} \mathbb{D}_i$, $\mathbb{D}_i$ being the disk of center $-\sigma_i$ and radius σ_i.

Proof.

Since $\mathbb{A} = \begin{pmatrix} -\sigma_1 & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & -\sigma_n \end{pmatrix}$, the Gershgorin's disk corresponding to the first column is $\mathbb{D}_1$ $\mathbb{D}_1$ being the disk of center $-\sigma_1$ and radius $|a_{21}| + |a_{31}| + \ldots |a_{n1}| = \sum_{j\neq 1} |a_{j1}| = \sum_{j\neq 1} a_{j1} = \sigma_1$. The same happens with all the other disks, and the theorem is proved.

Figure 6a shows one of these Gershgorin's circles, and **Figure 6b** shows a diagram of the union of all those disks.

Taking into account the previous results, the following corollary results.

Corollary 3.

If λ_k is an eigenvalue of $\mathbb{A}$, then $Re(\lambda_k) \leq 0$ and $Re(\lambda_k) = 0$ if and only if $\lambda_k = 0$.

Proof.

Immediately results from looking at **Figure 6**, since the union of disks is contained in the negative complex half-plane and is tangent to the imaginary axis at $z = 0$.

3. The possible cases for the eigenvalues and their multiplicities

In principle, there would be three possible cases for the eigenvalues of an FOCKM matrix: (1) $\lambda = 0$, which is always one of the FOCKM matrix eigenvalues; (2) $\lambda = a < 0$, which appeared in several examples analyzed in the introduction section; and (3) $\lambda = a + ib$, with $a < 0$ and $b \neq 0$, which we will see later that it is also possible.

Concerning the multiplicities (algebraic and geometric), we have seen examples of single and double eigenvalues, and in the latter case, $AM_\lambda = GM_\lambda$, at least in those introductory examples.

Taking into account all the previous results, there would be—at least potentially—nine cases to study, as can be seen in **Figure 7**.

We will study all these nine cases, which implies posing a double inverse problem:

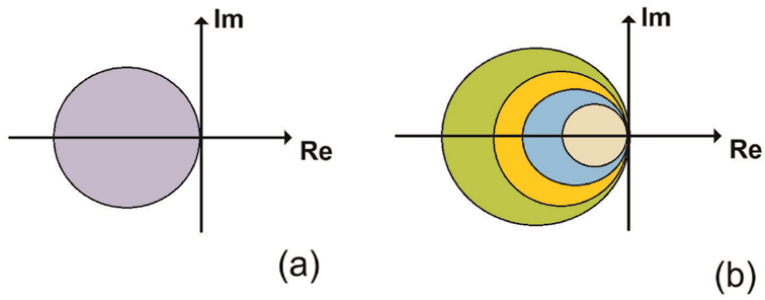


Figure 6.
a. One Gershgorin's circle. b. Diagram of the union of all the Gershgorin's disks.

		$\lambda = 0$	$\lambda = a < 0$	$\lambda = a \pm ib$
Simple		i	ii	iii
Multiple	MA = MG	iv	v	vi
	MA ≠ MG	vii	viii	ix

Figure 7.
The nine cases to study in FOCKM problems.

A. Given an eigenvalue λ, with algebraic multiplicity AM_λ , and geometric multiplicity GM_λ , is it possible to find an FOCKM matrix $\mathbb{A}$, which has this eigenvalue with the given algebraic and geometric multiplicities?

B. Once the matrix $\mathbb{A}$ is found, if it exists, is it possible to find a set of chemical reactions that correspond to it?

We will study both inverse problems for the nine cases described in **Figure 7**.

Case i. Simple eigenvalue $\lambda = 0$.

This case occurs, for example, when there are opposed reactions (see **Figure 2**). Specifically, in the example mentioned in the introduction section, the FOCKM matrix was

$$\mathbb{A} = \begin{pmatrix} -K & k \\ K & -k \end{pmatrix}, \tag{20}$$

whose eigenvalues are $\lambda_1 = 0$ and $\lambda_2 = -K - k$, both being simple eigenvalues.

Case ii. Simple eigenvalue $\lambda = a < 0$

Once again, the example of the opposite reactions serves to illustrate this situation, since the FOCKM matrix eigenvalues are $\lambda_1 = 0$ and $\lambda_2 = -K - k$, the second one being a negative simple eigenvalue.

Case iii. Simple eigenvalues $\lambda = a \pm ib, a < 0, b \neq 0$

In **Figure 8**, an FOCKM involving five chemical species is depicted by a directed graph. The numerical values of the kinetic constants are also included.

The corresponding ODE system is:

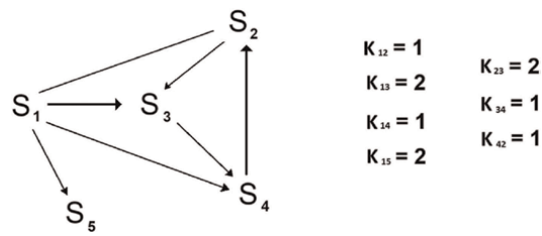


Figure 8.
Reactions involving chemical species S_1, S_2, S_3, S_4, and S_5.

$$\begin{cases} \frac{d[S_1]}{dt} = -(1+2+1+2)[S_1] \\ \frac{d[S_2]}{dt} = 1[S_1] - 2[S_2] + 1[S_4] \\ \frac{d[S_3]}{dt} = 2[S_1] + 2[S_2] - 1[S_3] \\ \frac{d[S_4]}{dt} = 1[S_1] + 1[S_3] - 1[S_4] \\ \frac{d[S_5]}{dt} = 2[S_1] \end{cases} \tag{21}$$

And then, the FOCKM matrix is:

$$\mathbb{A} = \begin{pmatrix} -6 & 0 & 0 & 0 & 0 \\ 1 & -2 & 0 & 1 & 0 \\ 2 & 2 & -1 & 0 & 0 \\ 1 & 0 & 1 & -1 & 0 \\ 2 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{22}$$

The eigenvalues are $\lambda_1 = -6$, $\lambda_2 = 0$, $\lambda_3 = 0$, $\lambda_4 = -2 + i$, and $\lambda_5 = -2 - i$, so simple complex eigenvalues $\lambda = a \pm ib, a < 0, b \neq 0$ can occur, at least theoretically.

Case iv. Multiple eigenvalue $\lambda = 0$ **with** $AM_\lambda = GM_\lambda$.

The example of competitive reactions serves to illustrate this situation (**Figure 4**). As it was observed, the FOCKM matrix is

$$\mathbb{A} = \begin{pmatrix} -K-k & 0 & 0 \\ K & 0 & 0 \\ k & 0 & 0 \end{pmatrix} \tag{23}$$

and its eigenvalues are $\lambda_1 = -K - k$, $\lambda_2 = 0$, and $\lambda_3 = 0$. If we consider the double eigenvalue $\lambda_{2,3} = 0$, we can obtain two independent associated eigenvectors $\vec{v_2} = (0 \quad 1 \quad 0)$ and $\vec{v_3} = (0 \quad 0 \quad 1)$, and so, $AM_{\lambda=0} = 2 = GM_{\lambda=0}$. Then, this case in theory can occur.

Case v. Multiple eigenvalue $\lambda = a < 0$ **with** $AM_\lambda = GM_\lambda$.

Let us consider a set of reactions involving chemical species S_1, S_2, S_3, S_4, and S_5, illustrated in **Figure 9**. In **Figure 9** presents a directed graph that schematizes all the FOCRs involved.

If $k_{12} = 1$, $k_{14} = 1$, $k_{23} = 1$, and $k_{45} = 1$, the corresponding ODE system is:

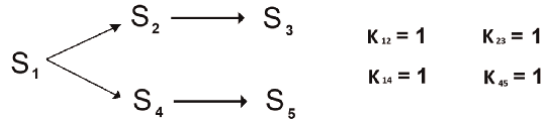


Figure 9.
Other reactions involving chemical species S_1, S_2, S_3, S_4, and S_5.

$$\begin{cases} \dfrac{d[S_1]}{dt} = -(1+1)[S_1] \\ \dfrac{d[S_2]}{dt} = 1[S_1] - 1[S_2] \\ \dfrac{d[S_3]}{dt} = 1[S_2] \\ \dfrac{d[S_4]}{dt} = 1[S_1] - 1[S_4] \\ \dfrac{d[S_5]}{dt} = 1[S_4] \end{cases} \tag{24}$$

And then, the FOCKM matrix is:

$$\mathbb{A} = \begin{pmatrix} -2 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{25}$$

The eigenvalues are: $\lambda_1 = 0$, $\lambda_2 = 0$, $\lambda_3 = -1$, $\lambda_4 = -1$, and $\lambda_5 = -2$. Then, there is a double negative eigenvalue $\lambda_{3,4} = -1$, and its associated eigenvectors are $\vec{v_3} = (\,0 \quad 0 \quad 0 \quad 1 \quad -1\,)$ and $\vec{v_4} = (\,0 \quad 1 \quad -1 \quad 0 \quad 0\,)$. Then, this case with a multiple eigenvalue $\lambda = -1$ and $AM_\lambda = GM_\lambda = 2$ occurs.

Case vi. Multiple eigenvalues $\lambda = a \pm ib, a < 0, b \neq 0$ **with** $AM_\lambda = GM_\lambda$.

In order to exemplify this case, we can adapt the diagram of **Figure 8** by "duplicating" the subgraph that corresponds to the chemical species S_1, S_2, S_3, and S_4. The result can be observed in **Figure 10**, which presents a directed graph that schematizes all the FOCRs involved.

If $k_{1j} = 1 \forall j = 2, 3, \ldots, 8$, $k_{23} = 2$, $k_{34} = 1$, $k_{42} = 1$, $k_{56} = 2$, $k_{67} = 1$, and $k_{75} = 1$, the corresponding ODE system is:

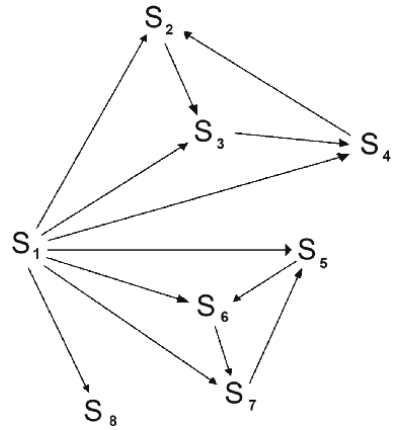


Figure 10.
Reactions involving eight chemical species.

$$\begin{cases} \dfrac{d[S_1]}{dt} = -(1+1+1+1+1+1+1)[S_1] \\ \dfrac{d[S_2]}{dt} = 1[S_1] - 2[S_2] + 1[S_4] \\ \dfrac{d[S_3]}{dt} = 1[S_1] + 2[S_2] - 1[S_4] \\ \dfrac{d[S_4]}{dt} = 1[S_1] + 1[S_3] - 1[S_4] \\ \dfrac{d[S_5]}{dt} = 1[S_1] - 2[S_5] + 1[S_7] \\ \dfrac{d[S_6]}{dt} = 1[S_1] + 2[S_5] - 1[S_7] \\ \dfrac{d[S_7]}{dt} = 1[S_1] + 1[S_6] - 1[S_7] \\ \dfrac{d[S_8]}{dt} = 1[S_1] \end{cases} \tag{26}$$

And then, the FOCKM matrix is:

$$\mathbb{A} = \begin{pmatrix} -7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -2 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -2 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 2 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{27}$$

The eigenvalues are $\lambda_1 = -7$, $\lambda_2 = -2+i$, $\lambda_3 = -2+i$, $\lambda_4 = -2-i$, $\lambda_5 = -2-i$, $\lambda_6 = 0$, $\lambda_7 = 0$, and $\lambda_8 = 0$. Then, $\lambda_{2,3} = -2+i$ and $\lambda_{4,5} = -2-i$ are both double complex eigenvalues, and their associated eigenvectors are

$$\begin{aligned} \vec{v_2} &= (\,0 \quad 0 \;\; 0 \quad 0 \quad -i \;\; -1+i \quad 1 \quad 0\,), \\ \vec{v_3} &= (\,0 \quad -i \;\; -1+i \quad 1 \quad 0 \;\; 0 \quad 0 \quad 0\,), \\ \vec{v_4} &= (\,0 \quad 0 \;\; 0 \quad 0 \quad -i \;\; -1-i \quad 1 \quad 0\,), \\ \vec{v_5} &= (\,0 \quad i \;\; -1-i \quad 1 \quad 0 \;\; 0 \quad 0 \quad 0\,). \end{aligned} \tag{28}$$

So, this example shows that multiple complex eigenvalues $\lambda = -2 \pm i$ with $AM_\lambda = GM_\lambda$ are possible in FOCKM matrices.

Case vii. Multiple eigenvalue $\lambda = 0$ with $AM_\lambda \neq GM_\lambda$.

In a previous work [10], it was shown analytically that this case is not possible. The proof is based on the study of the solutions of the ODE system corresponding to the FOCKM matrix, which are unbounded if $AM_{\lambda=0} \neq GM_{\lambda=0}$, and this result is nonsensical if we are considering concentrations of chemical species.

In the next section, we will give a pure algebraic proof of such an impossibility, which is independent of the interpretation of the solutions $[S_i](t)$ as concentrations of chemical substances.

Then, this case cannot occur.

Case viii. Multiple eigenvalue $\lambda = a < 0$ **with** $AM_\lambda \neq GM_\lambda$.

Let us consider again the case where a chemical species S_1 produces another chemical species S_2 and this in turn gives S_3 (consecutive reactions), illustrated in **Figure 3**. As a particular case, if the two kinetic constants were equal, that is, $K = k$, then the corresponding mathematical model is

$$\begin{cases} \dfrac{d[S_1]}{dt} = -K[S_1] \\ \dfrac{d[S_2]}{dt} = K[S_1] - K[S_2], \\ \dfrac{d[S_3]}{dt} = K[S_2] \end{cases} \tag{29}$$

The FOCKM matrix is

$$\mathbb{A} = \begin{pmatrix} -K & 0 & 0 \\ K & -K & 0 \\ 0 & K & 0 \end{pmatrix}, \tag{30}$$

and since it is a lower triangular matrix, its eigenvalues are $\lambda_1 = -K$, $\lambda_2 = -K$, and $\lambda_3 = 0$. So $\lambda_{1,2} = -K$ is a negative double eigenvalue, and there exists only one independent eigenvector $\vec{v_{1,2}} = (\,0 \quad 1 \quad -1\,)$, associated to $\lambda_{1,2} = -K$. Then, this case with a negative eigenvalue λ being $AM_\lambda \neq GM_\lambda$ occurs.

Case ix. Multiple eigenvalues $\lambda = a \pm ib, a < 0, b \neq 0$ **with** $AM_\lambda \neq GM_\lambda$.

For this last case, we consider six chemical species S_1, S_2, S_3, S_4, S_5, and S_6, and the reactions among them can be observed in **Figure 11**, which presents a directed multigraph that schematizes all the FOCRs involved.

If $k_{12} = 5/3$, $k_{23} = 5/3$, $k_{31} = 13/15$, $k_{32} = 4/5$, $k_{41} = 1$, $k_{45} = 1$, $k_{52} = 1$, $k_{56} = 1$, $k_{63} = 1$, and $k_{64} = 1$, the ODE system is

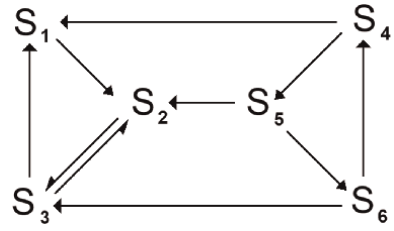


Figure 11.
Reactions involving six chemical species.

$$\begin{cases} \dfrac{d[S_1]}{dt} = -\dfrac{5}{3}[S_1] + \dfrac{13}{15}[S_3] + 1[S_4] \\ \dfrac{d[S_2]}{dt} = \dfrac{5}{3}[S_1] - \dfrac{5}{3}[S_2] + \dfrac{4}{5}[S_3] + 1[S_5] \\ \dfrac{d[S_3]}{dt} = \dfrac{5}{3}[S_2] - \dfrac{5}{3}[S_3] + 1[S_6] \\ \dfrac{d[S_4]}{dt} = -2[S_4] + 1[S_6] \\ \dfrac{d[S_5]}{dt} = 1[S_4] - 2[S_5] \\ \dfrac{d[S_6]}{dt} = 1[S_5] - 2[S_6] \end{cases} \tag{31}$$

And then, the associated FOCKM matrix is:

$$\mathbb{A} = \begin{pmatrix} -5/3 & 0 & 13/15 & 1 & 0 & 0 \\ 5/3 & -5/3 & 4/5 & 0 & 1 & 0 \\ 0 & 5/3 & -5/3 & 0 & 0 & 1 \\ 0 & 0 & 0 & -2 & 0 & 1 \\ 0 & 0 & 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 & 1 & -2 \end{pmatrix} \tag{32}$$

The characteristic equation is $\lambda(\lambda+1)\left[\lambda^2+5\lambda+7\right]^2 = 0$, and so, the eigenvalues are $\lambda_1 = 0$, $\lambda_2 = -1$, $\lambda_3 = \frac{1}{2}\left(-5+i\sqrt{3}\right)$, $\lambda_4 = \frac{1}{2}\left(-5+i\sqrt{3}\right)$, $\lambda_5 = \frac{1}{2}\left(-5-i\sqrt{3}\right)$, and $\lambda_6 = \frac{1}{2}\left(-5-i\sqrt{3}\right)$. Then, there is a couple of double complex eigenvalues $\lambda_{3,4} = \frac{1}{2}\left(-5+i\sqrt{3}\right)$ and, $\lambda_{5,6} = \frac{1}{2}\left(-5-i\sqrt{3}\right)$, and their associated eigenvectors are $\vec{v} = \left(\frac{1}{10}\left(-5-i3\sqrt{3}\right) \quad \frac{1}{10}\left(-5+i3\sqrt{3}\right) \quad 1 \quad 0 \quad 0 \quad 0\right)$ and $\vec{w} = \left(\frac{1}{10}\left(-5+i3\sqrt{3}\right) \quad \frac{1}{10}\left(-5-i3\sqrt{3}\right) \quad 1 \quad 0 \quad 0 \quad 0\right)$. Then, in this case, we have complex eigenvalues $\lambda_{3,4,5,6} = \frac{1}{2}\left(-5 \pm i\sqrt{3}\right)$ with $AM_\lambda = 2 \neq GM_\lambda = 1$.

4. An algebraic proof of the impossibility of the case vii

The main idea of this section is to show that for any FOCKM, the null eigenvalue $\lambda = 0$ always has the same algebraic and geometric multiplicity (i.e., $AM_{\lambda=0} = GM_{\lambda=0}$), which makes case vii of the previous section impossible to take place.

For this purpose, we will analyze an example that is an adaptation of the one considered in **Figure 9**. The new version of this example includes opposed reactions, and it is shown in **Figure 12**. All the reactions involved are schematized through a directed multigraph.

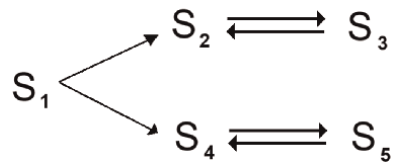


Figure 12.
Reactions involving five chemical species.

The corresponding ODE system is:

$$\begin{cases} \dfrac{d[S_1]}{dt} = -(k_{12} + k_{14})[S_1] \\ \dfrac{d[S_2]}{dt} = k_{12}[S_1] - k_{23}[S_2] + k_{32}[S_3] \\ \dfrac{d[S_3]}{dt} = k_{23}[S_2] - k_{32}[S_3] \\ \dfrac{d[S_4]}{dt} = k_{14}[S_1] - k_{45}[S_4] + k_{54}[S_5] \\ \dfrac{d[S_5]}{dt} = k_{45}[S_4] - k_{54}[S_5] \end{cases} \tag{33}$$

And then, the FOCKM matrix is:

$$\mathbb{A} = \begin{pmatrix} -k_{12} - k_{14} & 0 & 0 & 0 & 0 \\ k_{12} & -k_{23} & k_{32} & 0 & 0 \\ 0 & k_{23} & -k_{32} & 0 & 0 \\ k_{14} & 0 & 0 & -k_{45} & k_{54} \\ 0 & 0 & 0 & k_{45} & -k_{54} \end{pmatrix} \tag{34}$$

It is easy to observe that this matrix has rank $r = 3$, so there must be two rows that can be written as linear combinations of the others. In fact, it is possible to write:

$$row_1 = -row_2 - row_3 - row_4 - row_5 \tag{35}$$

and

$$row_2 = -row_3 + \frac{k_{12}}{k_{14}} row_4 + \frac{k_{12}}{k_{14}} row_5. \tag{36}$$

It should be noted that if the Eq. (36) is replaced in Eq. (35), then both row_1 and row_2 can be expressed as linear combinations of the other rows (i.e., row_3, row_4, and row_5). For this example, we do not need to perform these operations.

Utilizing the dimension theorem [12], we can write:

$n = \dim(Ker\mathbb{A}) + rank(\mathbb{A})$, so we have $\dim(Ker\mathbb{A}) = 5 - 3 = 2$ and with $Ker[\mathbb{A} - 0\mathbb{I}] = 2$, it results in $GM_{\lambda=0} = 2$.

Now, to analyze the algebraic multiplicity, we consider the characteristic equation:

$$0 = det(\mathbb{A} - \lambda\mathbb{I}) = \begin{vmatrix} -k_{12} - k_{14} - \lambda & 0 & 0 & 0 & 0 \\ k_{12} & -k_{23} - \lambda & k_{32} & 0 & 0 \\ 0 & k_{23} & -k_{32} - \lambda & 0 & 0 \\ k_{14} & 0 & 0 & -k_{45} - \lambda & k_{54} \\ 0 & 0 & 0 & k_{45} & -k_{54} - \lambda \end{vmatrix} \tag{37}$$

The determinant does not change if we add to a certain row a linear combination of the other rows. Then, we can replace the first row by

$$row_1 + row_2 + row_3 + row_4 + row_5 \tag{38}$$

and the second row by

$$row_2 + row_3 - \frac{k_{12}}{k_{14}} row_4 - \frac{k_{12}}{k_{14}} row_5, \tag{39}$$

without changing the determinant.

It is important to note that these linear combinations (38) and (39) correspond to formulas (35) and (36), previously obtained.

The result is the following:

$$0 = \begin{vmatrix} -\lambda & -\lambda & -\lambda & -\lambda & -\lambda \\ 0 & -\lambda & -\lambda & \frac{k_{12}}{k_{14}}\lambda & \frac{k_{12}}{k_{14}}\lambda \\ 0 & k_{23} & -k_{32} - \lambda & 0 & 0 \\ k_{14} & 0 & 0 & -k_{45} - \lambda & k_{54} \\ 0 & 0 & 0 & k_{45} & -k_{54} - \lambda \end{vmatrix} \tag{40}$$

One λ can be extracted from each of the first two rows, obtaining:

$$0 = \lambda^2 \begin{vmatrix} -1 & -1 & -1 & -1 & -1 \\ 0 & -1 & -1 & \frac{k_{12}}{k_{14}} & \frac{k_{12}}{k_{14}} \\ 0 & k_{23} & -k_{32} - \lambda & 0 & 0 \\ k_{14} & 0 & 0 & -k_{45} - \lambda & k_{54} \\ 0 & 0 & 0 & k_{45} & -k_{54} - \lambda \end{vmatrix} \tag{41}$$

Thus, this new version of the characteristic equation will have $\lambda = 0$ as a double root, and then, $AM_{\lambda=0} = 2$.

These same ideas can be generalized, obtaining the following theorem:

Theorem 3.

In any FOCKM matrix, the null eigenvalue ($\lambda = 0$) has the same algebraic and geometric multiplicity (i.e., $AM_{\lambda=0} = GM_{\lambda=0}$).

Proof.

Let us consider an FOCKM matrix $\mathbb{A} = \begin{pmatrix} -\sigma_1 & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & -\sigma_n \end{pmatrix}$, with $rank(\mathbb{A}) = r$.

By using the dimensions theorem, we can write $n = \dim(Ker\mathbb{A}) + rank(\mathbb{A})$ [12], so we have $\dim(Ker\mathbb{A}) = n - r$ and then $GM_{\lambda=0} = n - r$.

Now, let us consider the characteristic equation:

$$0 = det(\mathbb{A} - \lambda\mathbb{I}) = det\begin{pmatrix} -\sigma_1 - \lambda & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & -\sigma_n - \lambda \end{pmatrix} \tag{42}$$

Since $rank(\mathbb{A}) = r$, there are then $n - r$ rows that can be written as a linear combination of the remaining r rows. Without loss of generality, we can assume that the last $n - r$ rows can be written as linear combinations of the first r rows, which are linearly independent. For instance, if $j > r$, then we have:

$$row_j = \alpha_1 row_1 + \alpha_2 row_1 + \ ... \ + \alpha_r row_r, \tag{43}$$

which can be expanded as:

$$\begin{aligned}\left(a_{j,1}\, a_{j,2} \ldots a_{j,r} \ldots - \sigma_j \ldots a_{j,n}\right) = \ & \alpha_1\left(-\sigma_1\, a_{1,2} \ldots a_{1,r} \ldots a_{1,j} \ldots a_{1,n}\right) + \ ... \\ & + \alpha_r\left(a_{r,1}\, a_{r,2} \ldots - \sigma_r \ldots a_{r,j} \ldots a_{r,n}\right)\end{aligned} \tag{44}$$

Then, we can write the following equalities:

$$\begin{cases} a_{j,1} = -\alpha_1\sigma_1 + \ ... \ + \alpha_r a_{r,1} \\ a_{j,2} = \alpha_1 a_{1,2} + \ ... \ + \alpha_r a_{r,2} \\ \vdots \\ a_{j,r} = \alpha_1 a_{1,r} + \ ... \ - \alpha_r\sigma_r \\ \vdots \\ -\sigma_j = \alpha_1 a_{1,j} + \ ... \ + \alpha_r a_{r,j} \\ \vdots \\ a_{j,n} = \alpha_1 a_{1,n} + \ ... \ + \alpha_r a_{r,n} \end{cases} \tag{45}$$

Now, returning to the characteristic Eq. (42), this equality does not change if a linear combination is added to the j-th row as follows:

$$row_j - \alpha_1 row_1 - \alpha_2 row_2 - \ ... \ - \alpha_r row_r. \tag{46}$$

The new $j-$th row of the determinant is:

$$\begin{aligned}&\left(a_{j,1}\, a_{j,2} \ldots a_{j,r} \ldots - \sigma_j - \lambda \ldots a_{j,n}\right) - \alpha_1\left(-\sigma_1 - \lambda\, a_{1,2} \ldots a_{1,r} \ldots a_{1,j} \ldots a_{1,n}\right) \\ &\ldots - \alpha_r\left(a_{r,1}\, a_{r,2} \ldots - \sigma_r - \lambda \ldots a_{r,j} \ldots a_{r,n}\right)\end{aligned} \tag{47}$$

So, the first entry of this row is

$$a_{j,1} + \alpha_1\sigma_1 + \alpha_1\lambda \ldots - \alpha_r a_{r,1} = \alpha_1\lambda, \tag{48}$$

being the last equality, a direct consequence of the first equation of the system (45).

The same procedure can be applied to the second, third ... , and the $r-$th entry, which can be written as

$$a_{j,r} - \alpha_1 a_{1,r} \ldots + \alpha_r \sigma_r + \alpha_r \lambda = \alpha_r \lambda, \tag{49}$$

where the last equality is a consequence of the $r-$th equation of the system (45).

If we consider now the $j-$th entry, we have

$$-\sigma_j - \lambda - \alpha_1 a_{1,j} \ldots - \alpha_r a_{r,j} = -\lambda, \tag{50}$$

being the last equality, an obvious consequence of the $r-$th equation of the system (45).

Finally, the $n-$th entry is

$$a_{j,n} - \alpha_1 a_{1,n} \ldots - \alpha_r a_{r,n} = 0, \tag{51}$$

and the last equality is due to the last equation of the system (45).

Taking into account (48), (49), (50), and (51), the new $j-$th row of the determinant is

$$(\alpha_1 \lambda \; \alpha_2 \lambda \ldots \alpha_r \lambda \ldots - \lambda \; 0 \ldots 0) = \lambda.(\alpha_1 \; \alpha_2 \ldots \alpha_r \ldots - 1 \; 0 \ldots 0). \tag{52}$$

In other words, a λ can be extracted for each row, from row_{r+1} to row_n, and it is impossible to extract more λs, since the first r rows are linearly independent.

Then, $\lambda = 0$ is a multiple eigenvalue, with $AM_{\lambda=0} = n - r$, and this equality ends the proof of the theorem, since $AM_{\lambda=0} = GM_{\lambda=0} = n - r$.

The previous theorem—in addition to proving the impossibility of case vii (multiple eigenvalue $\lambda = 0$ with $AM_\lambda \neq GM_\lambda$)—also has important consequences on the stability of the solutions of the FOCKM problem, as will be seen in next section.

5. About the stability of the solutions to the FOCKM problem

In the previous analysis, we observed that in the FOCKM problem, there are three types of eigenvalues:

The null eigenvalue $\lambda = 0$, simple or multiple, with $AM_{\lambda=0} = GM_{\lambda=0}$.

Negative eigenvalues $\lambda = a < 0$, simple or multiple, with $AM_\lambda = GM_\lambda$ or $AM_\lambda \neq GM_\lambda$.

Complex eigenvalues $\lambda = a + ib$, with $a < 0$ and $b \neq 0$, simple or multiple, with $AM_\lambda = GM_\lambda$ or $AM_\lambda \neq GM_\lambda$.

We start by analyzing the second case, where the solutions corresponding to an eigenvalue $\lambda = a < 0$ are linear combinations of $\left\{e^{at}, te^{at}, t^2 e^{at}, \ \ldots, t^p e^{at}\right\}$, where the exponent p depends on whether the eigenvalue is simple or multiple and whether $AM_\lambda = GM_\lambda$ or not [13, 14]. In any case, the solution will be of the form $\left(\alpha_0 + \alpha_1 t + \alpha_2 t^2 + \ \ldots \ + \alpha_p t^p\right) e^{at}$, which, regardless of the degree of the polynomial, satisfies

$$\lim_{n \to \infty} \left(\alpha_0 + \alpha_1 t + \alpha_2 t^2 + \ \ldots \ + \alpha_p t^p\right) e^{at} = 0. \tag{53}$$

In the third case, the solutions corresponding to complex eigenvalues $\lambda = a + ib$, with $a < 0$ and $b \neq 0$ are linear combinations of $\{e^{at}\cos bt, e^{at}\sin bt, t\,e^{at}\cos bt, t\,e^{at}\sin bt, \ \ldots, t^q e^{at}\cos bt, t^q e^{at}\sin bt\}$, where the exponent q depends on whether the eigenvalue is simple or multiple and whether $AM_\lambda = GM_\lambda$ or not [13, 14]. In any case, the solution will be of the form $\left(\alpha_0 + \alpha_1 t + \alpha_2 t^2 + \ldots + \alpha_q t^q\right) e^{at}\cos bt + \left(\beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_q t^q\right) e^{at}\sin bt$, which, regardless of the degree of the polynomial, satisfies

$$\lim_{n\to\infty}\left(\alpha_0 + \alpha_1 t + \alpha_2 t^2 + \ldots + \alpha_q t^q\right) e^{at}\cos bt + \left(\beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_q t^q\right) e^{at}\sin bt = 0. \tag{54}$$

Now we are going to consider the null eigenvalue, $\lambda = 0$, which can be simple or multiple, always satisfying $AM_{\lambda=0} = GM_{\lambda=0}$, due to theorem 3. As a consequence, only linear combinations of e^{0t} can appear, and the general solution for this case is

$$\alpha_0\, e^{0t} + \alpha_1\, e^{0t} + \ldots + \alpha_n\, e^{0t} = \alpha_0 + \alpha_1 + \ldots + \alpha_n = C. \tag{55}$$

Taking into account the previous results (53), (54), and (55), it is easy to conclude that the solutions of the ODE system corresponding to any FOCKM problem will be stable, but not asymptotically stable (because of the null eigenvalue).

In other words, slight perturbations in the initial conditions of the problem will cause only slight differences in the solution, but these perturbations do not tend to disappear over time.

6. Conclusion

In this work, a special type of matrices has been analyzed, those that come from a linear ODE system, which represents the mathematical model associated with a set of chemical reactions, all following first-order kinetics law.

These matrices (FOCKM matrices) have a very special format, since all their non-diagonal entries are nonnegative, and the diagonal elements are equal to the sum of the other elements in the same column, with the sign changed.

This very particular format gives them certain special properties, such as that all their eigenvalues have a nonpositive real part and that the null eigenvalue is always present in the spectrum of the matrix. Moreover, the structure of these FOCKM matrices determines their algebraic and geometric multiplicity, at least in the case of the null eigenvalue, where both multiplicities must be the same.

In this chapter, the nine possible combinations of eigenvalues and multiplicities were analyzed, leading to the main conclusion that all of them are possible except one.

As a consequence of the above, the solutions to any FOCKM problem will be stable. This fact has an important consequence from the point of view of the original chemical kinetics problem. Indeed, any small error that may occur when weighing the reagents and/or diluting them in their respective solvents remains bounded, although —as a general rule—it does not tend to disappear.

Therefore, small errors in the determination of the initial conditions do not dramatically affect reactions with first-order kinetics, as if it could occur—at least in the first approach—in other types of reactions with nonlinear kinetics, which are not analyzed in this work.

Author details

Victor Martinez-Luaces
University of the Republic of Uruguay, Montevideo, Uruguay

*Address all correspondence to: victorml@fing.edu.uy

IntechOpen

© 2024 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

References

[1] Emanuel NM, Kritsman VA, Zaikov GE, Markovi N. Chemical Kinetics and Chain Reactions: Historical Aspects. New York; NY: Nova Science Publishers; 1995

[2] Doggett G, Cockett M, Davies AG, Phillips D, Derek-Woollins J. Maths for Chemists. 2nd ed. London: Royal Society of Chemistry; 2015

[3] Martinez-Luaces V. Square matrices associated to mixing problems ODE Systems. In: Matrix Theory-Applications and Theorems. London: InTech; 2018. DOI: 10.5772/intechopen.74437

[4] Martinez-Luaces V. Matrices in chemical problems modeled using directed graphs and multigraphs. In: Kyrchei I, editor. Hot Topics in Linear Algebra. New York: Nova Science Publishers; 2020. pp. 233-266

[5] Martinez-Luaces V. Chemical kinetics and inverse modelling problems. In: Chemical Kinetics. London: InTech; 2012. DOI: 10.5772/37376

[6] Martinez-Luaces V. Qualitative behavior of concentration curves in first-order chemical kinetics mechanisms. In: Taylor JC, editor. Advances in Chemistry Research. Vol. 34. New York: Nova Science Publishers; 2017. pp. 139-169

[7] Martinez-Luaces V. Concentration curves in first-order chemical kinetics: All possible cases. In: A Closer Look at Chemical Kinetics. New York: Nova Science Publishers; 2023. pp. 13-28

[8] Martinez-Luaces V. First-order chemical kinetics matrices and stability of O.D.E. Systems. In: Kyrchei I, editor. Advances in Linear Algebra Research. New York: Nova Science Publishers; 2015. pp. 325-343

[9] Martinez-Luaces V. Stability of ODE systems associated with first-order chemical kinetics mechanisms with and without final products. Konuralp Journal of Mathematics. 2016;**4**(1):80-87

[10] Martinez-Luaces V. Matrices in chemical problems: Characterization, properties and consequences about the stability of ODE Systems. In: Baswell AR, editor. Advances in Mathematics Research. New York: Nova Science Publishers; 2017. pp. 1-33

[11] Varga RS. Geršgorin and his Circles. Berlin Heildelberg: Springer-Verlag; 2011

[12] Roman S. Advanced Linear Algebra. Berlin Heildelberg: Springer-Verlag; 2013

[13] Ricardo HJ. A Modern Introduction to Differential Equations. Amsterdam: Elsevier Science; 2020

[14] Zaitsev VF, Polyanin AD. Handbook of Exact Solutions for Ordinary Differential Equations. Boca Raton, FL: CRC Press; 2002

Chapter 14

Matrices with a Diagonal Commutator

Armando Martínez-Pérez and Gabino Torres-Vega

Abstract

It is well known that there are no two matrices with a diagonal commutator. However, the commutator can behave as if it is diagonal when acting on a particular vector. We discuss pairs of matrices that give rise to a diagonal commutator when applied to a given arbitrary vector. Some properties of these matrices are discussed. These matrices have additional, continuous eigenvalues and eigenvectors than the dimension of the matrix, and their inverse also has this property. Some of these matrices are discrete approximations of the derivative and integration of a function and are exact for the exponential function. We also determine the adjoint of the obtained discrete derivative.

Keywords: commutator between matrices, pair of matrices with diagonal commutator, exact finite differences derivative, exact finite differences integration, matrices as discrete operators

1. Introduction

Let us consider the matrices that shift the entries of a vector. The usual matrix that cyclically shifts the entries of a vector to the left is

$$\mathbf{R}_1 = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & 0 & \dots & 0 & 0 \end{pmatrix}. \tag{1}$$

Given an arbitrary vector $\mathbf{h} = (h_1, h_2, \dots, h_N)^T \in \mathbb{C}$, $h_j \neq 0$, the action of $\mathbf{R}_1$ on this vector results in

$$\mathbf{R}_1\mathbf{h} = (h_2, h_3, \dots, h_N, h_1)^T. \tag{2}$$

There is also a matrix that shifts the entries of a vector to the right, the matrix with the lower diagonal entries different from zero.

But, we can also use a diagonal matrix to rotate to the left the entries of the vector **h**, in particular. The action of the diagonal matrix

$$\mathbf{R}_2 = \begin{pmatrix} \frac{h_2}{h_1} & 0 & 0 & 0 & \dots & 0 \\ 0 & \frac{h_3}{h_2} & 0 & 0 & \dots & 0 \\ 0 & 0 & \frac{h_4}{h_3} & 0 & \dots & 0 \\ 0 & 0 & 0 & \frac{h_5}{h_4} & \dots & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & 0 & \dots & \frac{h_1}{h_N} \end{pmatrix} \tag{3}$$

on **h** is $\mathbf{R}_2\mathbf{h} = (h_2, h_3, \dots, h_N, h_1)^T$. It is also possible to perform a cyclic rotation to the right. The action of this matrix on another vector is only to rescale its entries.

Combining the above ideas gives rise to a third type of shifting matrix. There is a non-diagonal matrix that acts like an identity matrix for the particular vector **h**:

$$\mathbf{R}_3 = \begin{pmatrix} 0 & \frac{h_1}{h_2} & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \frac{h_2}{h_3} & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \frac{h_3}{h_4} & \dots & 0 & 0 \\ \vdots & & & & & & \\ 0 & 0 & 0 & 0 & \dots & 0 & \frac{h_{N-1}}{h_N} \\ \frac{h_N}{h_1} & 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix}. \tag{4}$$

We have that $\mathbf{R}_3\mathbf{h} = \mathbf{h}$. The eigenvalues of this matrix are the same as those of $\mathbf{R}_1$, the N roots of unity: $\lambda_n = e^{i2\pi n/N}$, $n = 0,1,2, \dots, N-1$, and **h** is the eigenvector that corresponds to the eigenvalue $\lambda_0 = 1$.

Other special matrices are the matrices that admit a continuous eigenvalue, besides the usual constant eigenvalues. For instance, the matrix

$$\mathbf{Q}_1 = \begin{pmatrix} q_{11} & q_{12} & 0 \\ 0 & q_{22} & q_{23} \\ 0 & q_{32} & q_{33} \end{pmatrix} \tag{5}$$

has eigenvalues q_{11}, $\frac{1}{2}(q_{22} + q_{33} - z)$, $\frac{1}{2}(q_{22} + q_{33} + z)$, and the corresponding eigenvectors are

$$(1,0,0)^T, \tag{6}$$

$$\left(\frac{q_{12}(-q_{22}+q_{33}+z)}{q_{32}(2q_{11}-q_{22}-q_{33}+z)}, -\frac{-q_{22}+q_{33}+z}{2q_{32}}, 1\right)^T, \tag{7}$$

$$\left(\frac{q_{12}(q_{22}-q_{33}+z)}{q_{32}(-2q_{11}+q_{22}+q_{33}+z)}, \frac{q_{22}-q_{33}+z}{2q_{32}}, 1\right)^T, \tag{8}$$

where, $z=\sqrt{4q_{23}q_{32}+(q_{22}-q_{33})^2}$. These are the only eigenvalues and eigenvectors. The eigenvalues are fixed quantities and depend on the fixed entries of the matrix. However, the matrix

$$\begin{pmatrix} \frac{v_1\tau}{v_1-v_2} & -\frac{v_1\tau}{v_1-v_2} & 0 \\ 0 & \frac{v_2\tau}{v_2-v_3} & -\frac{v_2\tau}{v_2-v_3} \\ 0 & -\frac{v_3\tau}{v_3-v_2} & \frac{v_3\tau}{v_3-v_2} \end{pmatrix}, \tag{9}$$

has eigenvalues 0, τ, $\frac{v_1\tau}{v_1-v_2}$ with corresponding eigenvectors

$$(1,1,1)^T, \quad (v_1,v_2,v_3)^T, \quad (1,0,0)^T, \tag{10}$$

where the eigenvalue τ is independent of v_j, and viceversa, and it is a continuous variable.

There is the usual procedure to obtain a set of eigenvalues and eigenvectors [1]. But if the entries of a matrix contain variables, we can solve the eigenvalue set of simultaneous equations now for the entries of the matrix, obtaining additional continuous eigenvalues and eigenvectors.

Now, in general, there are no two matrices that have a diagonal commutator [2], that is, proportional to the identity matrix. We use the above facts about matrices to define two matrices with a diagonal commutator when applied to an arbitrary vector **h**. One of the obtained matrices corresponds to a finite differences derivation of a function, exact for the exponential function.

2. The matrices

There are many matrices with a diagonal commutator along a given direction, but we consider a simple set for simplicity.

We consider the pair of $N\times N$, $N\in\mathbb{N}$, matrices **A**, with non-zero elements around and on the diagonal, and **B**, diagonal, of the form

$$\mathbf{A}=\begin{pmatrix} a_{11} & a_{12} & 0 & \dots & 0 & 0 \\ 0 & a_{22} & a_{23} & \dots & 0 & 0 \\ 0 & 0 & a_{33} & \dots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \dots & a_{N-1,N-1} & a_{N-1,N} \\ 0 & 0 & 0 & \dots & a_{N,N-1} & a_{NN} \end{pmatrix}, \tag{11}$$

$$\mathbf{B} = \begin{pmatrix} b_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & b_2 & 0 & 0 & \dots & 0 \\ 0 & 0 & b_3 & 0 & \dots & 0 \\ 0 & 0 & 0 & b_4 & \dots & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & 0 & \dots & b_N \end{pmatrix}, \tag{12}$$

both matrices with complex entries.

A straightforward calculation yields the characteristic polynomial of the matrix $\mathbf{A}$,

$$\begin{aligned} &\det(\mathbf{A} - \lambda \mathbf{I}_N) = (a_{11} - \lambda)\cdots(a_{N-2,N-2} - \lambda) \\ &\times \left[\lambda^2 - \lambda(a_{NN} + a_{N-1,N-1}) - a_{N-1,N}a_{N,N-1} + a_{N-1,N-1}a_{NN}\right], \end{aligned} \tag{13}$$

whose solutions are

$$\lambda = a_{11}, a_{22}, \dots, a_{N-2,N-2}, a^+, a^-, \tag{14}$$

where

$$a^{\pm} = \frac{1}{2}\left(a_{N-1,N-1} + a_{NN} \pm \sqrt{(a_{N-1,N-1} - a_{NN})^2 + 4a_{N-1,N}a_{N,N-1}}\right). \tag{15}$$

The corresponding eigenvectors are a bit complicated to write them down here, but, for instance, for dimension four, the eigenvalues are $a_{11}, (a_{22} + a_{33} - z)/2, (a_{22} + a_{33} + z)/2$, and the corresponding eigenvectors are

$$(1,0,0)^T, \tag{16}$$

$$\left(\frac{a_{12}(-a_{22} + a_{33} + z)}{a_{32}(2a_{11} - a_{22} - a_{33} + z)}, -\frac{-a_{22} + a_{33} + z}{2a_{32}}, 1\right)^T, \tag{17}$$

$$\left(\frac{a_{12}(a_{22} - a_{33} + z)}{a_{32}(-2a_{11} + a_{22} + a_{33} + z)}, \frac{a_{22} - a_{33} + z}{2a_{32}}, 1\right)^T, \tag{18}$$

where

$$z = \sqrt{4a_{23}a_{32} + (a_{22} - a_{33})^2}. \tag{19}$$

These eigenvalues are discrete and fixed and the eigenvectors are different between themselves.

Still, we can generate additional eigenvalues and eigenvectors by choosing the values of a_{jj}. The eigenvalue equation $\mathbf{Au} = \gamma \mathbf{u}$, $\gamma \in \mathbb{C}$, $\mathbf{u} = (u_1, \dots, u_N)^T \in \mathbb{C}^N$, leads to a system of simultaneous equations which is now solved for the a_{jj}'s with solutions

$$a_{11} = \gamma - \frac{a_{12}u_2}{u_1}, \dots, a_{N-1,N-1} = \gamma - \frac{a_{N-1,N}u_N}{u_{N-1}}, a_{NN} = \gamma - \frac{a_{N,N-1}u_{N-1}}{u_N}. \tag{20}$$

This solution set gives rise to a matrix that results in weighted differences between the entries of a vector, resembling a finite-differences derivative [3–15].

Now, the determinant of the matrix **A** is not zero; it is equal to

$$\det(\mathbf{A}) = (a_{N-1,N-1}a_{NN} - a_{N-1,N}a_{N,N-1}) \prod_{j=1}^{N-2} a_{jj}, \tag{21}$$

and then, we can compute its inverse in the usual way. We could find it with $\mathbf{A}^{-1} = [\det(\mathbf{A})]^{-1} \operatorname{adj} \mathbf{A}$, where the adj **A** is the adjugate of **A**, whose entries are given by $(-1)^{i+j}\det(\mathbf{A}(j,i))$, where $\mathbf{A}(j,i)$ is the minor obtained by deleting the jth row and ith column from **A** [1]. The inverse matrix **A** becomes

$$\mathbf{A}^{-1} = \begin{pmatrix} \frac{1}{a_{11}} & -\frac{a_{12}}{a_{11}a_{22}} & \frac{a_{12}a_{23}}{a_{11}a_{22}a_{33}} & \cdots & -\frac{a_{NN}\prod_{k=1}^{N-2} a_{k,k+1}}{w\prod_{k=i}^{N-2} a_{kk}} & \frac{\prod_{k=1}^{N-1} a_{k,k+1}}{w\prod_{k=1}^{N-2} a_{kk}} \\ 0 & \frac{1}{a_{22}} & -\frac{a_{23}}{a_{22}a_{33}} & \cdots & \frac{a_{NN}\prod_{k=2}^{N-2} a_{k,k+1}}{w\prod_{k=2}^{N-2} a_{kk}} & -\frac{\prod_{k=2}^{N-1} a_{k,k+1}}{w\prod_{k=2}^{N-2} a_{kk}} \\ \vdots & & & & & \\ 0 & 0 & 0 & \cdots & \frac{a_{NN}}{w} & -\frac{a_{N-1,N}}{w} \\ 0 & 0 & 0 & \cdots & -\frac{a_{N,N-1}}{w} & \frac{a_{N-1,N-1}}{w} \end{pmatrix}, \tag{22}$$

where $w = a_{N-1,N-1}a_{NN} - a_{N-1,N}a_{N,N-1}$. When w vanishes, there is no inverse of the matrix **A**. This matrix resembles a discrete integration approximation over subintervals, with weight factors given by the matrix elements [3].

The usual eigenvectors are too complicated to write them here, but the complement eigenvector $\mathbf{v} = (v_1, v_2, \ldots, v_N)^T$, with eigenvalue τ, is obtained when the diagonal matrix elements are given as

$$a_{jj} = \frac{1}{\tau} - \frac{v_{j+1}}{v_j} a_{j,j+1}, 1 \leq j < N, \tag{23}$$

$$a_{NN} = \frac{1}{\tau} - \frac{v_{N-1}}{v_N} a_{N,N-1}. \tag{24}$$

We saw some properties of the general matrix **A**. Next, we analyze particular versions of this matrix.

3. The commutator between matrices A and B

The commutator between the matrices **A** and **B** results in

$$\mathbf{A}, \mathbf{B} = \begin{pmatrix} 0 & a_{12}\Delta_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & a_{23}\Delta_2 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \cdots & 0 & a_{N-1,N}\Delta_{N-1} \\ 0 & 0 & 0 & \cdots & -a_{N,N-1}\Delta_{N-1} & 0 \end{pmatrix}, \tag{25}$$

where $\Delta_j = b_{j+1} - b_j$ is the difference between b_{j+1} and b_j. This matrix is not diagonal; it does not depend on the diagonal elements of the matrix **A**, and it introduces the differences between adjacent diagonal elements of the matrix **B**. We can choose the values of $a_{j,j+1}$ and $a_{N,N-1}$ to obtain the effect of a diagonal commutator when acting on the particular arbitrary vector **h**. This procedure can be applied to more general matrices **A** and **B**. Still, our choice of the form of matrices **A** and **B** is for the sake of simplicity and to define a two-point derivative matrix that satisfies the commutator relationship and has as eigenvector the exponential function.

We ask matrices **A** and **B** to comply with the requirement that

$$[\mathbf{A}, \mathbf{B}]\mathbf{h} = \alpha \mathbf{h}, \quad \alpha \in \mathbb{C}, \tag{26}$$

where $\mathbf{h} = (h_1, h_2, \ldots, h_N)^T \in \mathbb{C}^N$ is an arbitrary complex vector of length N but with non-zero entries $h_j \neq 0$. The requirement Eq. (26) leads to a system of simultaneous equations that are solved not for the vector **h** but for the *entries of the matrix* **A**. The resulting matrix is

$$\mathbf{A}_c = \begin{pmatrix} a_{11} & \frac{h_1\alpha}{h_2\Delta_1} & 0 & \ldots & 0 & 0 \\ 0 & a_{22} & \frac{h_2\alpha}{h_3\Delta_2} & \ldots & 0 & 0 \\ 0 & 0 & a_{33} & \ldots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \ldots & a_{N-1,N-1} & \frac{h_{N-1}\alpha}{h_N\Delta_{N-1}} \\ 0 & 0 & 0 & \ldots & -\frac{h_N\alpha}{h_{N-1}\Delta_{N-1}} & a_{N,N} \end{pmatrix}, \tag{27}$$

which depends on **h** and α. With these matrices, the commutator becomes a permutation matrix with weights $\alpha h_j/h_{j+1}$ or $\alpha h_N/h_{N-1}$,

$$\mathbf{A}_c, \mathbf{B} = \begin{pmatrix} 0 & \alpha\frac{h_1}{h_2} & 0 & \ldots & 0 & 0 \\ 0 & 0 & \alpha\frac{h_2}{h_3} & \ldots & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \ldots & 0 & \alpha\frac{h_{N-1}}{h_N} \\ 0 & 0 & 0 & \ldots & \alpha\frac{h_N}{h_{N-1}} & 0 \end{pmatrix}. \tag{28}$$

This matrix performs two shifts when acting on the vector **h**. Since the matrix has only upper off-diagonal elements different from zero, there is a shift to the left. But the ratios h_j/h_{j+1} shift back the vector. Then, the commutator is diagonal along the direction of **h**.

Now, we use another condition on $\mathbf{A}_c$ to determine its diagonal entries. The version of the matrix $\mathbf{A}_c$ that complies with the eigenvalue equation $\tilde{\mathbf{A}}_c\mathbf{v} = \gamma\mathbf{v}$ is

$$\tilde{\mathbf{A}} = \begin{pmatrix} \gamma - \frac{h_1 v_2 \alpha}{h_2 v_1 \Delta_1} & \frac{h_1 \alpha}{h_2 \Delta_1} & 0 & \dots & 0 & 0 \\ 0 & \gamma - \frac{h_2 v_3 \alpha}{h_3 v_2 \Delta_2} & \frac{h_2 \alpha}{h_3 \Delta_2} & \dots & 0 & 0 \\ 0 & 0 & \gamma - \frac{h_3 v_4 \alpha}{h_4 v_3 \Delta_3} & \dots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \dots & \gamma - \frac{h_{N-1} v_N \alpha}{h_N v_{N-1} \Delta_{N-1}} & \frac{h_{N-1} \alpha}{h_N \Delta_{N-1}} \\ 0 & 0 & 0 & \dots & -\frac{h_N \alpha}{h_{N-1} \Delta_{N-1}} & \frac{h_N v_{N-1} \alpha}{h_{N-1} v_N \Delta_{N-1}} + \gamma \end{pmatrix}. \tag{29}$$

The action of this matrix on a vector $\mathbf{g} = (g_1, g_2, \dots, g_N)^T \in \mathbb{C}^N$ is

$$(\tilde{\mathbf{A}}\mathbf{g})_j = \frac{h_j \alpha}{h_{j+1} \Delta_j} g_{j+1} + \left(\gamma - \frac{h_j v_{j+1} \alpha}{h_{j+1} v_j \Delta_j} \right) g_j, \quad 1 \le j < N, \tag{30}$$

$$(\tilde{\mathbf{A}}\mathbf{g})_N = \left(\gamma + \frac{h_N v_{N-1} \alpha}{h_{N-1} v_N \Delta_{N-1}} \right) g_N - \frac{h_N \alpha}{h_{N-1} \Delta_{N-1}} g_{N-1}. \tag{31}$$

If we use power series expansions for the quotients $h(b)/h(b+\Delta)$ and $v(b+\Delta)/v(b)$, at mesh points, we obtain

$$\frac{h_j}{h_{j+1} \Delta_j} = \frac{1}{\Delta_j} - \frac{h_j'}{h_j} + O(\Delta_j), \quad \text{and} \quad \frac{v_{j+1}}{v_j \Delta_j} = \frac{1}{\Delta_j} + \frac{v_j'}{v_j} + O(\Delta_j). \tag{32}$$

With these expansions, we obtain the small Δ_j expressions

$$(\tilde{\mathbf{A}}\mathbf{g})_j \underset{\Delta_j \to 0}{\to} \alpha \frac{g_{j+1} - g_j}{\Delta_j} + g_j \left(\gamma - \alpha \frac{v_j'}{v_j} \right) - \alpha \frac{h_j'}{h_j} \left(g_{j+1} - g_j \right) + O(\Delta_j), \quad 1 \le j < N, \tag{33}$$

$$(\tilde{\mathbf{A}}\mathbf{g})_N \underset{\Delta_j \to 0}{\to} \alpha \frac{h_N}{h_{N-1} \Delta_{N-1}} \left(\frac{v_{N-1}}{v_N} g_N - g_{N-1} \right) + \gamma g_N + O(\Delta_j). \tag{34}$$

Note that, since $g_{j+1} - g_j \to 0$ and $\left(g_{j+1} - g_j \right)/\Delta_j \to g'(b_j)$ as $\Delta_j \to 0$, if we ask that $\gamma - \alpha v_j'/v_j \to 0$, we would get $\alpha g'(b_j)$ from Eqs. (32) and (33) in the limit $\Delta_j \to 0$. Thus, if we make the choice $v(b) = e^{\gamma b/\alpha}$, Eqs. (30) and (31) lead to finite difference approximations to the derivative when $\mathbf{A}$ acts on $\mathbf{g}$. The right hand side of Eq. (33) becomes the finite differences derivative of $g(b)$ at the boundary plus the boundary term γg_N. Hereafter, we will use $v(b) = e^{\gamma b/\alpha}$, and finite Δ_j.

4. The derivative matrix

With the choices $h_j = v_j$ and $v_j = v(b_j) = e^{\gamma b_j/\alpha}$, the matrix $\alpha^{-1}\tilde{\mathbf{A}}$ becomes the derivative matrix

$$\mathbf{D} = \begin{pmatrix} \frac{\gamma}{\alpha} - \frac{1}{\Delta_1} & \frac{e^{-\frac{\gamma\Delta_1}{\alpha}}}{\Delta_1} & 0 & \dots & 0 & 0 \\ 0 & \frac{\gamma}{\alpha} - \frac{1}{\Delta_2} & \frac{e^{-\frac{\gamma\Delta_2}{\alpha}}}{\Delta_2} & \dots & 0 & 0 \\ 0 & 0 & \frac{\gamma}{\alpha} - \frac{1}{\Delta_3} & \dots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \dots & \frac{e^{-\frac{\gamma\Delta_{N-2}}{\alpha}}}{\Delta_{N-2}} & 0 \\ 0 & 0 & 0 & \dots & \frac{\gamma}{\alpha} - \frac{1}{\Delta_{N-1}} & \frac{e^{-\frac{\gamma\Delta_{N-1}}{\alpha}}}{\Delta_{N-1}} \\ 0 & 0 & 0 & \dots & -\frac{e^{\frac{\gamma\Delta_{N-1}}{\alpha}}}{\Delta_{N-1}} & \frac{\gamma}{\alpha} + \frac{1}{\Delta_{N-1}} \end{pmatrix}. \tag{35}$$

When the matrix $\mathbf{D}$ acts to the left on a vector $\mathbf{f}^T = (f_1, f_2, \dots, f_N)^T \in \mathbb{C}^N$, we obtain

$$\mathbf{f}^T\mathbf{D} = -\big((\mathcal{B}\mathbf{f})_1, (\mathcal{D}\mathbf{f})_2, \dots, (\mathcal{D}\mathbf{f})_{N-2}, (\mathcal{D}\mathbf{f})_{N-1} + (\mathcal{B}\mathbf{f})_{N-1}, , (\mathcal{D}\mathbf{f})_N + (\mathcal{B}\mathbf{f})_N\big)^T, \tag{36}$$

where

$$(\mathcal{D}\mathbf{f})_j = \left(\frac{\gamma}{\alpha} - \frac{1}{\Delta_j}\right) f_j + e^{-\gamma\Delta_{j-1}/\alpha}\frac{f_{j-1}}{\Delta_{j-1}}, \quad 1 < j \le N, \tag{37}$$

$$(\mathcal{B}\mathbf{f})_1 = \left(\frac{1}{\Delta_1} - \frac{\gamma}{\alpha}\right) f_1, \tag{38}$$

$$(\mathcal{B}\mathbf{f})_{N-1} = e^{\frac{\gamma}{\alpha}\Delta_{N-1}}\frac{f_N}{\Delta_{N-1}}, \tag{39}$$

$$(\mathcal{B}\mathbf{f})_N = -2\frac{f_N}{\Delta_{N-1}}, \tag{40}$$

which is the negative of a finite-differences approximation to the derivative of $\mathbf{f}$. When Δ_j is small and using the exponential expansions $e^{\mp\gamma\Delta_j/\alpha} = 1 \mp \gamma\frac{\Delta_j}{\alpha} + \cdots$, we find that

$$(\mathcal{D}\mathbf{f})_j \underset{\Delta_j \to 0}{\to} \frac{f_j}{\Delta_j} - \frac{f_{j-1}}{\Delta_{j-1}} - \frac{\gamma}{\alpha}\left(f_j - f_{j-1}\right) + O(\Delta_j), \quad 1 < j < N. \tag{41}$$

When $\mathbf{D}$ is applied to the right to the vector $\mathbf{g} = (g_1, \dots, g_N)^T$, we obtain

$$\mathbf{Dg} = \left((\mathrm{D}\mathbf{g})_1, (\mathrm{D}\mathbf{g})_2, \ldots, (\mathrm{D}\mathbf{g})_{N-1}, (\mathrm{D}\mathbf{g})_N\right)^T, \tag{42}$$

where

$$(\mathrm{D}\mathbf{g})_j = \left(\frac{\gamma}{\alpha} - \frac{1}{\Delta_j}\right) g_j + e^{-\frac{\gamma}{\alpha}\Delta_j} \frac{g_{j+1}}{\Delta_j}, \quad 1 \le j \le N-1, \tag{43}$$

$$(\mathrm{D}\mathbf{g})_N = \left(\frac{\gamma}{\alpha} + \frac{1}{\Delta_{N-1}}\right) g_N - e^{\frac{\gamma}{\alpha}\Delta_{N-1}} \frac{g_{N-1}}{\Delta_{N-1}}. \tag{44}$$

For small Δ_j, these equalities become

$$(\mathrm{D}\mathbf{g})_j \underset{\Delta_j \to 0}{\to} \frac{g_{j+1} - g_j}{\Delta_j} - \frac{\gamma}{\alpha}\left(g_{j+1} - g_j\right) + O(\Delta_j), \quad 1 \le j \le N-1, \tag{45}$$

$$(\mathrm{D}\mathbf{g})_N \underset{\Delta_j \to 0}{\to} \frac{g_N - g_{N-1}}{\Delta_{N-1}} + \frac{\gamma}{\alpha}\left(g_N - g_{N-1}\right) + O(\Delta_j). \tag{46}$$

These approximations show that $\mathbf{D}$ is a derivation matrix, exact for the exponential function $e^{\gamma b/\alpha}$ for any value of Δ_j.

For instance, for dimension five, the eigenvalues of the derivative matrix $\mathbf{D}$ are $\left\{\frac{\gamma}{\alpha}, \frac{\gamma}{\alpha}, \frac{\gamma}{\alpha} - \frac{1}{\Delta_1}, \frac{\gamma}{\alpha} - \frac{1}{\Delta_2}, \frac{\gamma}{\alpha} - \frac{1}{\Delta_{N-2}}\right\}$, and the corresponding eigenvectors are

$$\left(e^{\frac{\gamma}{\alpha}b_1}, e^{\frac{\gamma}{\alpha}b_2}, \ldots, e^{\frac{\gamma}{\alpha}b_4}, e^{\frac{\gamma}{\alpha}b_5}\right)^T, \tag{47}$$

$$(0,0,0,0,0)^T, \tag{48}$$

$$(1,0,0,0,0)^T, \tag{49}$$

$$\left(-\frac{e^{\frac{\gamma}{\alpha}b_1}\Delta_2}{\Delta_1 - \Delta_2}, e^{\frac{\gamma}{\alpha}b_2}, 0, 0, 0\right)^T, \tag{50}$$

$$\left(-\frac{e^{\frac{\gamma}{\alpha}b_1}\Delta_3^2}{(\Delta_1 - \Delta_3)(-\Delta_2 + \Delta_3)}, -\frac{e^{\frac{\gamma}{\alpha}b_2}\Delta_3}{\Delta_2 - \Delta_3}, e^{\frac{\gamma}{\alpha}b_3}, 0, 0\right)^T, \tag{51}$$

and the exponential function is an eigenfunction of the derivative matrix with eigenvalue γ/α.

The determinant of the derivative matrix $\mathbf{D}$ is

$$|\mathbf{D}| = \frac{\gamma^2}{\alpha^2} \prod_{k=1}^{N-2} \left(\frac{1}{\Delta_k} - \frac{\gamma}{\alpha}\right). \tag{52}$$

Therefore, as long as $\alpha, \Delta_k, \gamma \neq 0$, and $\alpha - \gamma\Delta_k \neq 0$, $k = 1, 2, \ldots, N-2$, we can compute the inverse of $\mathbf{D}$. For dimension 4×4, the inverse matrix is

$$\mathbf{D}^{-1} = \begin{pmatrix} -\frac{\alpha\Delta_1}{\alpha-\gamma\Delta_1} & -\frac{e^{-\frac{\gamma}{\alpha}\Delta_1}\alpha^2\Delta_2}{(\alpha-\gamma\Delta_1)(\alpha-\gamma\Delta_2)} & \frac{e^{-\frac{\gamma}{\alpha}(\Delta_1+\Delta_2)}\alpha^3(\alpha+\gamma\Delta_4)}{\gamma^2(\alpha-\gamma\Delta_1)(\alpha-\gamma\Delta_2)\Delta_4} & -\frac{e^{-\frac{\gamma}{\alpha}(\Delta_1+\Delta_2+\Delta_3)}\alpha^4}{\gamma^2(\alpha-\gamma\Delta_1)(\alpha-\gamma\Delta_2)\Delta_4} \\ 0 & -\frac{e^{-\frac{\gamma}{\alpha}\Delta_2}\alpha\Delta_2}{\alpha-\gamma\Delta_2} & \frac{e^{-\frac{\gamma}{\alpha}(\Delta_2+\Delta_3)}\alpha^2(\alpha+\gamma\Delta_4)}{\gamma^2(\alpha-\gamma\Delta_2)\Delta_4} & -\frac{e^{-\frac{\gamma(\Delta_2+\Delta_3)}{\alpha}}\alpha^3}{\gamma^2(\alpha-\gamma\Delta_2)\Delta_4} \\ 0 & 0 & \frac{\alpha(\alpha+\gamma\Delta_4)}{\gamma^2\Delta_4} & -\frac{e^{-\frac{\gamma\Delta_3}{\alpha}}\alpha^2}{\gamma^2\Delta_4} \\ 0 & 0 & \frac{e^{\frac{\gamma\Delta_3}{\alpha}}\alpha^2}{\gamma^2\Delta_3} & -\frac{\alpha(\alpha-\gamma\Delta_3)}{\gamma^2\Delta_3} \end{pmatrix} \tag{53}$$

When this matrix acts on a vector, the result is a collection of partial summations with weights given by its entries, an integration matrix. The eigenvalues of the inverse matrix are $\frac{\alpha}{\gamma}$, $\frac{\alpha}{\gamma}$, $-\frac{\alpha\Delta_1}{\alpha-\gamma\Delta_1}$, $-\frac{\alpha\Delta_2}{\alpha-\gamma\Delta_2}$ and the corresponding eigenvectors are

$$\left(e^{\frac{\gamma}{\alpha}b_1}, e^{\frac{\gamma}{\alpha}b_2}, e^{\frac{\gamma}{\alpha}b_3}, e^{\frac{\gamma}{\alpha}b_4}\right)^T, \tag{54}$$

$$(0,0,0,0)^T, \tag{55}$$

$$\left(e^{\frac{\gamma}{\alpha}b_1},0,0,0\right)^T, \tag{56}$$

$$\left(-\frac{e^{\frac{\gamma}{\alpha}b_1}\Delta_2}{\Delta_1-\Delta_2}, e^{\frac{\gamma}{\alpha}b_2},0,0\right)^T. \tag{57}$$

The exponential function $e^{\gamma b/\alpha}$ is an eigenvector of the inverse matrix (a finite differences integration matrix) with eigenvalue α/γ.

4.1 Summation by parts

A practical result is the summation by parts theorem (the discrete version of the continuous variable integration by parts theorem), the subject of this section.

We start by defining the summation matrix

$$\mathbf{S} = \begin{pmatrix} \Delta_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \Delta_2 & 0 & \dots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \dots & \Delta_{N-1} & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix}. \tag{58}$$

The summation matrix $\mathbf{S}$ allows for determining the equality

$$\mathbf{f}^T\mathbf{S}\mathbf{D}\mathbf{g} = \mathbf{g}^T\left(\tilde{\mathbf{B}} - \mathbf{S}\mathbf{D}^\dagger\right)\mathbf{f}, \tag{59}$$

where

$$\tilde{\mathbf{B}} = \begin{pmatrix} \frac{\gamma}{\alpha}\Delta_1 - 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \\ 0 & 0 & \dots & 0 & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & e^{-\frac{\gamma\Delta_{N-1}}{\alpha}} & 0 \end{pmatrix} \tag{60}$$

is the boundary matrix, and

$$\mathbf{D}^\dagger = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ -\frac{e^{-\frac{\gamma\Delta_1}{\alpha}}}{\Delta_2} & \frac{1}{\Delta_2} - \frac{\gamma}{\alpha} & 0 & \dots & 0 & 0 & 0 \\ 0 & -\frac{e^{-\frac{\gamma\Delta_2}{\alpha}}}{\Delta_3} & \frac{1}{\Delta_3} - \frac{\gamma}{\alpha} & \dots & 0 & 0 & 0 \\ \vdots & & & & & & \\ 0 & 0 & 0 & 0 & \dots & -\frac{e^{-\frac{\gamma\Delta_{N-2}}{\alpha}}}{\Delta_{N-1}} & \frac{1}{\Delta_{N-1}} - \frac{\gamma}{\alpha} \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \end{pmatrix} \tag{61}$$

is the with continuous entries, adjoint of the discrete derivative matrix $\mathbf{D}$. When this matrix acts on a vector $\mathbf{f}$, we obtain a vector with entries

$$\left(\mathbf{D}^\dagger \mathbf{f}\right)_j = \left(\frac{1}{\Delta_j} - \frac{\gamma}{\alpha}\right) f_j - \frac{1}{\Delta_j} e^{-\frac{\gamma\Delta_{j-1}}{\alpha}} f_{j-1}, \quad 1 < j < N. \tag{62}$$

The small Δ_j limit of the boundary terms is

$$\left(\mathbf{g}\tilde{\mathbf{B}}\mathbf{f}\right)_1 \underset{\Delta_j \to 0}{\to} -f_1 g_1 + O(\Delta_j), \tag{63}$$

$$\left(\mathbf{g}\tilde{\mathbf{B}}\mathbf{f}\right)_N \underset{\Delta_j \to 0}{\to} f_{N-1} g_N + O(\Delta_j), \tag{64}$$

and, for the adjoint matrix $\mathbf{D}^\dagger$, we obtain

$$\left(\mathbf{D}^\dagger \mathbf{f}\right)_j \underset{\Delta_j \to 0}{\to} \frac{f_j - f_{j-1}}{\Delta_j} - \frac{\gamma}{\alpha} f_j + O(\Delta_j). \tag{65}$$

Thus, the adjoint matrix is a derivation minus a constant term.

For a matrix with dimension 4×4, for instance, the eigenvalues are 0 of multiplicity two, $\frac{1}{\Delta_2} - \frac{\gamma}{\alpha}$, and $\frac{1}{\Delta_3} - \frac{\gamma}{\alpha}$, and the eigenvectors are

$$(0,0,0,1)^T, \tag{66}$$

$$\left(e^{-\frac{b_1\gamma}{\alpha}}\left(1 - \Delta_2 \frac{\gamma}{\alpha}\right)\left(1 - \Delta_3 \frac{\gamma}{\alpha}\right), e^{-\frac{b_2\gamma}{\alpha}}\left(1 - \Delta_3 \frac{\gamma}{\alpha}\right), e^{-\frac{b_3\gamma}{\alpha}}, 0\right)^T, \tag{67}$$

$$\left(0, -\frac{(\Delta_3 - \Delta_2)e^{-\frac{b_2\gamma}{\alpha}}}{\Delta_2}, e^{-\frac{b_3\gamma}{\alpha}}, 0\right)^T, \tag{68}$$

$$(0,0,1,0)^T, \tag{69}$$

and the complement, for arbitrary dimension, eigenvector is

$$\left(0, \frac{e^{-\gamma b_2/\alpha}}{\alpha - (\alpha\tau + \gamma)\Delta_2}, \ldots, \frac{e^{-\gamma b_{N-1}/\alpha}}{\prod_{k=2}^{N-2}(\alpha - (\alpha\tau + \gamma)\Delta_k)}, 0\right)^T, \tag{70}$$

with eigenvalue $\tau \in \mathbb{C}$, and $\tau \neq \frac{1}{\Delta_k} - \frac{\gamma}{\alpha}$, $k = 2, 3, \ldots, N-1$, the complement of the regular eigenvalues.

The determinant of the matrix $\mathbf{D}^\dagger$ vanishes, and then, there is no inverse matrix at all.

We arrived at a finite-differences derivative of a function that complies with the discrete versions of the properties that the continuous derivative has [16].

4.2 The upper diagonal matrix

The simplest case of matrices with diagonal commutator along the direction of $\mathbf{h}$ is obtained when the diagonal elements of the matrix $\mathbf{A}$ vanish, $a_{jj} = 0$. That matrix is

$$\mathbf{A}_2 = \begin{pmatrix} 0 & \frac{h_1\alpha}{h_2\Delta_1} & 0 & \ldots & 0 & 0 \\ 0 & 0 & \frac{h_2\alpha}{h_3\Delta_2} & \ldots & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \ldots & 0 & \frac{h_{N-1}\alpha}{h_N\Delta_{N-1}} \\ 0 & 0 & 0 & \ldots & -\frac{h_N\alpha}{h_{N-1}\Delta_{N-1}} & 0 \end{pmatrix}. \tag{71}$$

This matrix is also a cyclic shifting matrix to the left, with rescaling, in general, and only a rescaling matrix when acting on the vector $\mathbf{h}$ as if it were the matrix $\alpha\ \mathrm{diag}\left(\Delta_j^{-1}\right)$, with $\Delta_N = \Delta_{N-1}$.

The eigenvalues of the matrix $\mathbf{A}_2$ are

$$\lambda = 0\ (\text{with multiplicity } N-2), \textit{and}\ \ \lambda_\pm = \pm\frac{i\alpha}{\Delta_{N-1}}, \tag{72}$$

and the corresponding eigenvectors are

$$\mathbf{x}_\lambda = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\Delta_{N-1}^{N-2} h_1}{\prod_{j=1}^{N-2} \Delta_j} \\ i \frac{\Delta_{N-1}^{N-3} h_2}{\prod_{j=2}^{N-2} \Delta_j} \\ -\frac{\Delta_{N-1}^{N-4} h_3}{\prod_{j=3}^{N-2} \Delta_j} \\ \vdots \\ i^{N-2} h_{N-1} \\ i^{N-1} h_N \end{pmatrix}, \begin{pmatrix} -\frac{\Delta_{N-1}^{N-2} h_1}{\prod_{j=1}^{N-2} \Delta_j} \\ -i \frac{\Delta_{N-1}^{N-3} h_2}{\prod_{j=2}^{N-2} \Delta_j} \\ \frac{\Delta_{N-1}^{N-4} h_3}{\prod_{j=3}^{N-2} \Delta_j} \\ \vdots \\ -i^{N-2} h_{N-1} \\ -i^{N-1} h_N \end{pmatrix}. \tag{73}$$

Additionally, there are $N-3$, not well-defined vectors corresponding to the null eigenvalue with degeneracy $N-3$. These eigenvectors give rise to a vector space of dimension three. However, if we just look for the solution of the simultaneous linear equations, there is another set of eigenvectors, the complement eigenvalue β, and the eigenvector $\mathbf{v}$ with entries—for dimension five—given by

$$\frac{v_5}{v_4} = -i\frac{h_5}{h_4}, \quad \frac{v_4}{v_3} = \frac{h_4 \beta \Delta_3}{h_3 \alpha}, \quad \frac{v_3}{v_2} = \frac{h_3 \beta \Delta_2}{h_2}, \quad \frac{v_2}{v_1} = \frac{h_2 \beta \Delta_1}{h_1 \alpha}, \quad v_1 \in \mathbb{C}. \tag{74}$$

with eigenvalue $\beta = \pm i\alpha/\Delta_4$.

The determinant of the matrix $\mathbf{A}_2$ vanishes, and then, there is no inverse matrix at all.

5. Conclusion

We discussed several properties of matrices, namely, pairs of matrices with a diagonal commutator when applied to a given vector, exact finite-differences derivation and integration, and complement eigenvalues and eigenvectors.

These results are relevant in quantum mechanics theory, in which some operators have a discrete spectrum. Our scheme might be of interest in quantum gravity theory, too, because the space is quantized, and then a discrete derivative with respect to the length variable is needed [17, 18].

Acknowledgements

A. Martínez-Pérez would like to acknowledge the support from the UNAM Postdoctoral Program (POSDOC).

Author details

Armando Martínez-Pérez[1†] and Gabino Torres-Vega[2*†]

1 IIMAS, UNAM, México City, Mexico

2 Departamento de F'isica, Cinvestav, México City, Mexico

*Address all correspondence to: gabino.torres@cinvestav.mx

† These authors contributed equally.

IntechOpen

© 2024 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

References

[1] Piziak R, Odell PL. Matrix Theory: From Generalized Inverses to Jordan Form. Boca Raton: Champan & Hall/ CRC; 2007

[2] Putnam CR. Commutation Properties of Hilbert Space Operators and Related Topics. Berlin: Springer-Verlag; 1967

[3] Martínez-Pérez A, Torres-Vega G. The Inverse of the Discrete Momentum Operator, cap.10, Schrödinger Equation - Fundamentals Aspects and Potential Applications. Rijeka: IntechOpen; 2023. DOI: 10.5772/intechopen.112376

[4] Mickens RE. Nonstandard Finite Difference Models of Differential Equations. Singapore: World Scientific; 1994

[5] Mickens M. Discretizations of nonlinear differential equations using explicit nonstandard methods. Journal of Computational and Applied Mathematics. 1999;**110**:181

[6] Mickens RE. Nonstandard finite difference schemes for differential equations. Journal of Difference Equations and Applications. 2010;**8**:823

[7] Mickens RE. Applications of Nonstandard Finite Difference Schemes. Singapore: World Scientific; 2000

[8] Mickens RE. Calculation of denominator functions for nonstandard finite difference schemes for differential equations satisfying a positivity condition. Numerical Methods for Partial Differential Equations. 2006;**23**:672

[9] Potts RB. Differential and difference equations. The American Mathematical Monthly. 1982;**89**:402-407

[10] Potts RB. Ordinary and partial differences equations. The Journal of the Australian Mathematical Society. Series B. 1986;**27**:488

[11] Tarasov VE. Exact discretization by Fourier transforms. Communications in Nonlinear Science and Numerical Simulation. 2016;**37**:31

[12] Tarasov VE. Exact discrete Analogs of derivatives of integer orders: Differences as infinite series. Journal of Mathematics. 2015;**2015**:134842. DOI: 10.1155/2015/134842

[13] Tarasov VE. Exact discretization of Schrödinger equation. Physics Letters A. 2016;**380**:68. DOI: 10.1016/j.physleta. 2015.10.039

[14] Martínez Pérez A, Torres-Vega G. Exact finite differences. The derivative on non uniformly spaced partitions. Symmetry. 2017;**9**:217. DOI: 10.3390/ sym9100217

[15] Martínez-Pérez A, Torres-Vega G. Discrete self-adjointness and quantum dynamics. Travel times. Journal of Mathematical Physics. 2021;**62**:012013. DOI: 10.1063/5.0021565

[16] Gitman DM, Tyutin IV, Voronov BL. Self-Adjoint Extensions in Quantum Mechanics. General Theory and Applications to Schrödinger and Dirac Equations with Singular Potentials. New York: Birkhäuser; 2012

[17] Bishop M, Contreras J, Singleton D. "The more things change the more they stay the same" Minimum lengths with unmodified uncertainty principle and dispersion relation. International Journal of Modern Physics D. 2022;**31**:2241002

[18] Bishop M, Contreras J, Singleton D. Reconciling a quantum gravity minimal length with lack of photon dispersion. Physics Letters B. 2021;**816**:136265

Chapter 15

On the Universal Realizability Problem: New Results

Ana I. Julio and Ricardo L. Soto

Abstract

Let $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ be a list of complex numbers. Λ is said to be realizable if there is a nonnegative matrix with spectrum Λ. The list Λ is said to be universally realizable ($\mathcal{UR}$) if it is realizable for each possible Jordan canonical form (*JCF*) allowed by Λ. The problem of determining the universal realizability of Λ is called universal realizability problem (*URP*). The first results concerning *URP* (formerly called nonnegative inverse elementary divisors problem) are due to H. Minc and they establish that if Λ is the spectrum of a diagonalizable positive matrix, then Λ is $\mathcal{UR}$. In this chapter, we introduce new results that contain extensions of Minc's results and that allow us to show the universal realizability of lists of complex numbers not positively realizable. We also prove new universal realizability criteria and structured universal realizability criteria.

Keywords: nonnegative matrices, universal realizability problem, nonnegative inverse eigenvalue problem, Jordan canonical form, realizability of spectra

1. Introduction

A list $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ of complex numbers (with repeats allowed) is said to be realizable if there is an $n \times n$ nonnegative matrix A whose spectrum is Λ. In this case, A is said to be a realizing matrix. It is well known that if A is a nonnegative matrix, then the spectral radius of A, $\rho(A) = \max\{|\lambda_i|,\ \lambda_i \text{ eigenvalue of } A\}$, is an eigenvalue of A called its Perron eigenvalue. Throughout this chapter, if $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ is realizable, λ_1 will be the Perron eigenvalue of the corresponding realizing matrix. The problem of determining the realizability of Λ is called Nonnegative Inverse Eigenvalue Problem (*NIEP*, see Ref. [1]). If the realizing matrix A is diagonalizable, we say that Λ is diagonalizably realizable (*DR*). A list $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ is universally realizable ($\mathcal{UR}$) if it is realizable for every possible Jordan canonical form (*JCF*) allowed by Λ. The problem of determining the universal realizability of a list Λ of complex numbers is called universal realizability problem (*URP*) (formerly called Nonnegative Inverse Elementary Divisors Problem (*NIEDP*, see Ref. [2])). Both problems, the *NIEP* and the *URP*, have been completely solved only for lists of $n \leq 4$ complex numbers, which show the difficulty of both problems.

A matrix $A = [a_{ij}]$ of order n is said to have constant row sums if all its rows sum up to the same constant α. We denote by $\mathcal{CS}_\alpha$ the set of all $n \times n$ real matrices with

constant row sums equal to $\alpha \in \mathbb{R}$. It is clear that any matrix in $\mathcal{CS}_\alpha$ has an eigenvector $\mathbf{e} = [1, \ldots, 1]^T$ corresponding to the eigenvalue α. The relevance of the real matrices with constant row sums is due to the well-known fact that the problem of finding a nonnegative matrix with spectrum $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$, λ_1 being the Perron eigenvalue, is equivalent to the problem of finding a nonnegative matrix in $\mathcal{CS}_{\lambda_1}$ with spectrum Λ (see Ref. [3]). We define the kth moment of the list Λ to be $s_k(\Lambda) := \sum_{j=1}^n \lambda_j^k$, $k = 1, 2, \ldots$. It is clear that if Λ is the spectrum of a nonnegative matrix A, then $s_k(\Lambda) = trace(A^k) \geq 0$, for $k = 1, 2, \ldots$. Moreover, we denote by $\mathbf{e}_k$ the vector with one in the kth position and zeros elsewhere. Finally, we denote by $E_{i,j}$ the matrix with 1 in position (i, j) and zero elsewhere, and we define the matrix $E = \sum_{i \in K} E_{i,i+1}, \subset \{2, \ldots, n-1\}$.

The *URP* includes the *NIEP* and both problems are equivalent if the elements of $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ are distinct. Since 1949, many works on the *NIEP* have been published. In contrast, few works are known about *URP*. As far as we know, the first results concerning *URP* (formerly *NIEDP*) are due to H. Minc. In Refs. [4, 5], Minc studied the problem for nonnegative and doubly stochastic matrices. In particular, he proved the following theorem, which we write in terms of the *URP* as:

Theorem 1.1 Let $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ be a list of complex numbers. If Λ is diagonalizably positively realizable, then Λ is universally realizable.

It is clear that the diagonalizability condition is necessary, while the positivity condition is essential for the proof of Minc's result. Minc set the question whether his result holds for nonnegative realizations. This question was open for almost 40 years. Recently, two extensions of Minc's result have been obtained, which we will discuss in Section 2.

In what follows we will use the following results, some of which have been employed with success to obtain sufficient conditions for the *NIEP* and the *URP* to have a solution. The first two are perturbation results: the first one, due to Brauer [6], shows how to change one single eigenvalue of an $n \times n$ matrix without changing any of the remaining $n-1$ eigenvalues. The second result, due to R. Rado and published by Perfect in Ref. [7], is a generalization of Brauer's result. It shows how to change r eigenvalues of an $n \times n$ matrix without changing any of the remaining $n-r$ eigenvalues. The third result, by Soto and Ccapa [8], shows how is the *JCF* of the Brauer perturbation $A + \mathbf{v}\mathbf{q}^T$. Next two results are a symmetric version of Rado's result [9] and a result, due to Laffey and Šmigoc [10], that solves the *NIEP* for left half-plane lists of complex numbers, that is, lists $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ with $\lambda_1 > 0$, $\text{Re}\, \lambda_i \leq 0,\ i = 2, \ldots, n$.

Theorem 1.2 [6] Let A be an $n \times n$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. Let $\mathbf{v} = [v_1, \ldots, v_n]^T$ be an eigenvector of A associated with the eigenvalue λ_k and let $\mathbf{q}$ be any n-dimensional vector. Then $A + \mathbf{v}\mathbf{q}^T$ has eigenvalues $\lambda_1, \ldots, \lambda_{k-1}, \lambda_k + \mathbf{v}^T\mathbf{q}, \lambda_{k+1}, \ldots, \lambda_n$.

Theorem 1.3 [7] Let A be an $n \times n$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. Let $X = [\mathbf{x}_1 | \cdots | \mathbf{x}_r]$ be such that $rank(X) = r$ and $A\mathbf{x}_i = \lambda_i \mathbf{x}_i,\ i = 1, \ldots, r,\ r \leq n$. Let C be an $r \times n$ matrix. Then $A + XC$ has eigenvalues $\mu_1, \ldots, \mu_r, \lambda_{r+1}, \ldots, \lambda_n$, where $\mu_1, \ldots, \mu_r$ are eigenvalues of the matrix $\Omega + CX$ with $\Omega = diag\{\lambda_1, \ldots, \lambda_r\}$.

Theorem 1.4 [8] Let $\mathbf{q} = [q_1, \ldots, q_n]^T$ be an arbitrary n-dimensional vector and let E_{ij} be an $n \times n$ matrix with 1 in position (i, j) and zeros elsewhere. Let $A \in \mathcal{CS}_{\lambda_1}$ with *JCF* $J(A) = S^{-1}AS = diag(J_1(\lambda_1), J_{n_2}(\lambda_2), \ldots, J_{n_k}(\lambda_k))$. If $\lambda_1 + \sum_{i=1}^n q_i \neq \lambda_i,\ i = 2, \ldots, n$, then $A + \mathbf{e}\mathbf{q}^T$ has Jordan canonical form $J(A + \mathbf{e}\mathbf{q}^T) = J(A) + (\sum_{i=1}^n q_i) E_{11}$. In particular, if $\sum_{i=1}^n q_i = 0$, then A and $A + \mathbf{e}\mathbf{q}^T$ are similar.

Theorem 1.5 [9] Let A be an $n\times n$ symmetric matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$ be an orthonormal set of eigenvectors of A such that $AX = X\Omega$, where $X = [\mathbf{x}_1|\cdots|\mathbf{x}_r]$ and $\Omega = diag\{\lambda_1, \ldots, \lambda_r\}$. Let C be any $r\times r$ symmetric matrix. Then the symmetric matrix $A + XCX^T$ has eigenvalues $\mu_1, \ldots, \mu_r, \lambda_{r+1}, \ldots, \lambda_n$, where $\mu_1, \ldots, \mu_r$ are eigenvalues of the matrix $\Omega + C$.

Theorem 1.6 [10] Let $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ be a list of complex numbers such that $\lambda_1 > 0$ and $\text{Re}\,\lambda_i \le 0$, $i = 2, \ldots, n$. Then Λ is realizable if and only if the following conditions are satisfied:

$$s_1 = \sum_{i=1}^{n} \lambda_i \ge 0, \quad s_2 = \sum_{i=1}^{n} \lambda_i^2 \ge 0, \quad s_1^2 \le n s_2.$$

Since a list of complex numbers $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ is always the spectrum of some $n\times n$ matrix A (for instance, a diagonal matrix), from now on we will use, interchangeably, the term list or spectrum. Regarding Minc's result, the following question arises: Are there spectra no positively realizable that are $\mathcal{UR}$? This question has a positive answer. The following results have been progressively obtained:

1. In Soto and Ccapa [8], it was proved that spectra of real Suleĭmanova type [11], that is,

$$\Lambda = \{\lambda_1, \ldots, \lambda_n\} \quad \text{with} \quad \lambda_1 > 0 \ge \lambda_2 \ge \cdots \ge \lambda_n, \quad \text{are} \quad \mathcal{UR}. \tag{1}$$

2. In Soto et al. [12], it was proved that spectra of complex Suleĭmanova type, that is,

$$\lambda_1 > 0, \quad \text{Re}\,\lambda_i \le 0, |\,\text{Re}\,\lambda_i| \ge |\text{Im}\,\lambda_i|, \quad i = 2, \ldots, n \quad \text{are} \quad \mathcal{UR}, \tag{2}$$

3. In Diaz and Soto [13], it was proved that spectra of Šmigoc type, that is,

$$\lambda_1 > 0, \quad \text{Re}\,\lambda_i \le 0, \quad |\sqrt{3}\,\text{Re}\,\lambda_i| \ge |\text{Im}\,\lambda_i|, \quad i = 2, \ldots, n \quad \text{are also} \quad \mathcal{UR}. \tag{3}$$

It is interesting to note that in all these three cases, Λ is $\mathcal{UR}$ if and only if Λ is realizable if and only if $\sum_{i=1}^{n}\lambda_i \ge 0$. Moreover, these three kinds of spectra are left half-plane spectra, and the good behavior of them led us to think, in a first moment, that any left half-plane spectra were $\mathcal{UR}$. In Julio et al. [14], the authors showed that this is not true. In fact, the spectra

$$\Lambda = \left\{a, -\frac{a}{4} \pm \frac{\sqrt{5}a}{4}i, -\frac{a}{4} \pm \frac{\sqrt{5}a}{4}i\right\}, \quad a > 0$$

are not DR, and therefore not $\mathcal{UR}$.

2. Extensions for Minc's result

In Ref. [15], the authors Borobia and Moro prove the following result:

Theorem 2.1 [15] Let A be a nonnegative irreducible matrix with spectrum $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ and a positive row or column. Then A is similar to a positive.

Recently, two extensions of Minc's result have been obtained in Collao et al. and Johnson et al. [16, 17]. In Ref. [16], Collao et al. showed that a nonnegative matrix $A \in \mathcal{CS}_{\lambda_1}$, with a positive column, is similar to a positive matrix (the irreducibility condition is not necessary if $A \in \mathcal{CS}_{\lambda_1}$). Note that if A is nonnegative with a positive row and A^T has a positive eigenvector, then A^T is also similar to a positive matrix. As a consequence we have:

Corollary 2.1 [16] If Λ is the spectrum of a diagonalizable nonnegative matrix $A \in \mathcal{CS}_{\lambda_1}$ having a positive column, then Λ is $\mathcal{UR}$. If A is diagonalizable nonnegative with a positive row and A^T has a positive eigenvector, then Λ is also $\mathcal{UR}$.

Regarding the second extension for Minc's result, we have:

Definition 2.1 We call a realization $A = [a_{ij}]$ off-diagonally positive (*ODP*), if $a_{ij} > 0$ whenever $i \neq j$, and on the diagonal, zero entries are allowed. Furthermore, a realization is quasi-*ODP*, if all off-diagonal entries are positive, except for one that is zero.

In Ref. [17], Johnson et al. introduced the concept of *ODP* matrices and proved that if Λ is diagonalizably *ODP* realizable, then Λ is $\mathcal{UR}$. Note that both extensions contain, as a particular case, Minc's result in Ref. [4]. Both extensions allow us to significantly increase the set of spectra that can be proved to be $\mathcal{UR}$. The extension in Johnson et al. [17], for instance, allows us to prove that certain spectra $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ with $s_1(\Lambda) = 0$, are $\mathcal{UR}$, which is not possible from Minc's result.

There are a number of spectra that are *ODP* realizable, as for instance, spectra of real numbers of Suleĭmanova type, which are realizable if and only if $\sum_{i=1}^{n} \lambda_i \geq 0$ (see [7, 18]). Moreover, it was proved in Johnson et al. [17] that real spectra of Suleĭmanova type are diagonalizably *ODP* realizable, and therefore $\mathcal{UR}$. In fact, it is clear that Λ is the spectrum of

$$A = \begin{bmatrix} \lambda_1 & & & \\ \lambda_1 - \lambda_2 & \lambda_2 & & \\ & & \ddots & \\ \lambda_1 - \lambda_n & & & \lambda_n \end{bmatrix} \in \mathcal{CS}_{\lambda_1}.$$

Then since $A\mathbf{e} = \lambda_1 \mathbf{e}$, for $\mathbf{q}^T = \left[\sum_{i=2}^{n} \lambda_i, -\lambda_2, \ldots, -\lambda_n\right]$, $A + \mathbf{e}\mathbf{q}^T$ is a diagonalizable *ODP* matrix with spectrum Λ. Thus, Λ is $\mathcal{UR}$.

Theorem 2.2 [17] Suppose that a spectrum $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ is realizable, with a certain *JCF*, by a matrix that is either *ODP* or quasi-*ODP*. Then, Λ is realizable, by a matrix with any coarser *JCF*, by a matrix that is *ODP* or quasi-*ODP*.

Proof: Suppose that Λ is *ODP* realizable, with a realizing matrix $A \in \mathcal{CS}_{\lambda_1}$, with *JCF* $J(A) = S^{-1}AS$. Let $B = SES^{-1}$, where $E = \sum_{i \in K} E_{i,i+1}, \quad K \subset \{2, \ldots, n-1\}$. Since B is such that $tr(B) = 0$ and $B \in \mathcal{CS}_0$, then from Theorem 1.2 and Theorem 1.4 with $\mathbf{q} = -\mathbf{b}$ ($\mathbf{b}$ is the vector of entries diagonal of B), the matrix $B + \mathbf{e}\mathbf{q}^T$ has all its diagonal entries zero. Therefore, by picking ε small enough, we have that $A + \varepsilon\left(B + \mathbf{e}\mathbf{q}^T\right)$ is nonnegative with the same eigenvalues as A, but the coarsened *JCF*. If A is quasi-*ODP* with a zero entry in position (j,k), $j \neq k$ and $B + \mathbf{e}\mathbf{q}^T$ has zero diagonal entries with a negative entry in position (j,k), $j \neq k$, then by choosing $\varepsilon < 0$ small enough, the (j,k), $j \neq k$ entry in $A + \varepsilon\left(B + \mathbf{e}\mathbf{q}^T\right)$ will be positive. Since $\varepsilon < 0$ is small enough, it does not matter if the other possible positive entries in $B + \mathbf{e}\mathbf{q}^T$

become negative. Anyway, $A+\varepsilon\left(B+\mathbf{e}\mathbf{q}^T\right)$ will be nonnegative, cospectral with A, and with the coarsened *JCF*.

In the event that the original realization is diagonalizable, we have the following important special case, which generalizes the classic result of Minc in [4].

Corollary 2.2 [17] If a spectrum Λ is diagonalizably *ODP* or quasi-*ODP* realizable, then Λ is $\mathcal{UR}$.

Proof: Apply Theorem 2.2 with the original blocks all 1×1 in the real eigenvalue case, and all blocks 2×2 (real blocks) in the complex conjugate eigenvalue case, any sequence of merged blocks may be achieved, resulting in any *JCF* allowed by the spectrum.

Example 2.1 Consider the list $\Lambda=\{6,1,1,-4,-4\}$ which is diagonalizably *ODP* realizable with realizing matrix

$$A=\begin{bmatrix} 0 & \frac{3+\sqrt{5}}{2} & \frac{3-\sqrt{5}}{2} & \frac{3-\sqrt{5}}{2} & \frac{3+\sqrt{5}}{2} \\ \frac{3+\sqrt{5}}{2} & 0 & \frac{3+\sqrt{5}}{2} & \frac{3-\sqrt{5}}{2} & \frac{3-\sqrt{5}}{2} \\ \frac{3-\sqrt{5}}{2} & \frac{3+\sqrt{5}}{2} & 0 & \frac{3+\sqrt{5}}{2} & \frac{3-\sqrt{5}}{2} \\ \frac{3-\sqrt{5}}{2} & \frac{3-\sqrt{5}}{2} & \frac{3+\sqrt{5}}{2} & 0 & \frac{3+\sqrt{5}}{2} \\ \frac{3+\sqrt{5}}{2} & \frac{3-\sqrt{5}}{2} & \frac{3-\sqrt{5}}{2} & \frac{3+\sqrt{5}}{2} & 0 \end{bmatrix}.$$

Let $S=[\mathbf{s}_1,\mathbf{s}_2,\ldots,\mathbf{s}_n]$ be the matrix of eigenvectors of A, where $\mathbf{s}_1=\mathbf{e}$, such that $S^{-1}AS=J(A)=diag\{6,1,1,-4,-4\}$. In particular, for $J(A)=diag\{J_1(6),J_2(1),J_2(-4)\}$, where $J_k(\lambda)$ denotes a $k\times k$ Jordan block corresponding to the eigenvalue λ, we have, for $\mathbf{q}=\left(-\frac{\sqrt{5}-1}{10},-\frac{2}{5},\frac{\sqrt{5}+1}{10},\frac{\sqrt{5}+1}{10},-\frac{\sqrt{5}-1}{10}\right)^T$, the matrix

$$SES^{-1}+\mathbf{e}\mathbf{q}^T=\begin{bmatrix} 0 & -\frac{\sqrt{5}}{10}-\frac{1}{2} & \frac{1}{2}-\frac{\sqrt{5}}{10} & \frac{2\sqrt{5}}{5} & -\frac{\sqrt{5}}{5} \\ -\frac{\sqrt{5}}{5} & 0 & \frac{\sqrt{5}}{5} & -\frac{\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \\ \frac{\sqrt{5}}{10}+\frac{1}{2} & \frac{\sqrt{5}}{10}-\frac{1}{2} & 0 & \frac{\sqrt{5}}{5} & -\frac{2\sqrt{5}}{5} \\ -\frac{\sqrt{5}}{5} & \frac{\sqrt{5}}{10}-\frac{1}{2} & \frac{\sqrt{5}}{5} & 0 & \frac{1}{2}-\frac{\sqrt{5}}{10} \\ -\frac{\sqrt{5}}{5} & -\frac{\sqrt{5}}{10}-\frac{1}{2} & \frac{\sqrt{5}}{5} & \frac{\sqrt{5}}{10}+\frac{1}{2} & 0 \end{bmatrix},$$

where $E=E_{23}+E_{45}$, having all its diagonal entries zero. Thus, we chose $\varepsilon\neq 0$ small enough, to obtain $A+\varepsilon\left(SES^{-1}+\mathbf{e}\mathbf{q}^T\right)$, which is a nonnegative *ODP* matrix with *JCF* having Jordan blocks $J_1(6),J_2(1),J_2(-4)$. In a similar way, we may obtain a nonnegative matrix with spectrum Λ, for each one of the other *JCF* allowed by Λ. Thus, Λ is $\mathcal{UR}$.

3. On universal realizability criteria

It has been commented in the Introduction that real and complex spectra of Suleĭmanova type are $\mathcal{UR}$ [8, 12]. In Collao et al. [19], the authors study lists with two positive eigenvalues and they prove that, under certain conditions, these types of lists are not only realizable, but also $\mathcal{UR}$.

Theorem 3.1 [19] Let $\Lambda = \{p, q, -r_1, -r_2, \ldots, -r_{n-2}\}$ be a list of n real numbers with

$$p + q - \sum_{j=1}^{n-2} r_j = 0,$$

in which $p, q, r_j > 0$; $p > q, r_j$, $j = 1, 2, \ldots, n-2$; $-r_j \geq -r_{j+1}, j = 1, 2, \ldots, n-3$. If there is a decomposition

$$R = \{r_1, r_2, \ldots, r_{n-2}\} = R_1 \cup R_2 \text{with}$$
$$R_1 = \{\alpha_1, \ldots, \alpha_s\}, R_2 = \{\beta_1, \ldots, \beta_{n-2-s}\}, \alpha_i, \beta_i \in R,$$

in such a way that

$$p \geq \sum_{i=1}^{s} \alpha_i \geq \sum_{i=1}^{n-2-s} \beta_i, \text{ then } \Lambda \text{ is } \mathcal{UR}.$$

Proof: Take

$$\Lambda = \Lambda_0 \cup \Lambda_1 \cup \Lambda_2 \text{ with } \Lambda_0 = \{p, q\},$$
$$\Lambda_1 = \{-\alpha_1, \ldots, -\alpha_s\}, \Lambda_2 = \{-\beta_1, \ldots, -\beta_t\}, \ t = n - 2 - s$$
$$\alpha_i, \beta_i > 0; \quad -\alpha_i, -\beta_i \in \{-r_1, -r_2, \ldots, -r_{n-2}\},$$

with the associated lists

$$\Gamma_1 = \{p - t, -\alpha_1, \ldots, -\alpha_s\}, \Gamma_2 = \{q + t, -\beta_1, \ldots, -\beta_t\},$$

with $t \geq 0$ and $p - t = \sum_{i=1}^{s} \alpha_i$, $q + t = \sum_{i=1}^{t} \beta_i$. Note that the lists Γ_1, Γ_2 are of real Suleĭmanova type, then from Soto and Ccapa [8], they are $\mathcal{UR}$. Let A_1 and A_2 be the realizing matrices of Γ_1 and Γ_2, respectively. Since $p - q - t \geq t \geq 0$,

$$B = \begin{bmatrix} p - t & t \\ p - q - t & q + t \end{bmatrix} \text{is nonnegative with spectrum } \Lambda_0$$

and the required diagonal entries. Thus, from Theorem 1.3 with X the $n \times 2$ nonnegative matrix of eigenvectors of $\begin{bmatrix} A_1 & \\ & A_2 \end{bmatrix}$ and C the $2 \times n$ nonnegative matrix, such that $CX = B - \Omega$ with $\Omega = diag\{p - t, q + t\}$,

$$\begin{bmatrix} A_1 & \\ & A_2 \end{bmatrix} + XC \text{ is nonnegative with spectrum, } \Lambda$$

and with the required *JCF* allowed by Λ. Then, Λ is $\mathcal{UR}$.

Note that if $\Lambda = \{p, q, -r, -r, \ldots, -r\}$ with $p + q - (n-2)r = 0,$ where $p, q, r > 0,\ p > q, r;\ q > r,$ then if n is even, Λ is $\mathcal{UR}$. If n is odd with $p \geq \frac{(n-1)}{2} r,$ then Λ is also $\mathcal{UR}$.

Example 3.1 Let $\Lambda = \{19, 1, -2, -2, -3, -3, -5, -5\}$ (with)

$$\Lambda_0 = \{19, 1\}, \Lambda_1 = \Lambda_2 = \{-2, -3, -5\}$$
$$\Gamma_1 = \Gamma_2 = \{10, -2, -3, -5\}.$$

We want to construct a nonnegative matrix A with *JCF*

$$J = diag\{J_1(19), J_1(1), J_2(-2), J_2(-3), J_1(-5), J_1(-5)\}.$$

Then we compute

$$A_1 = \begin{bmatrix} 10 & 0 & 0 & 0 \\ 15 & -5 & 0 & 0 \\ 13 & 0 & -2 & -1 \\ 13 & 0 & 0 & -3 \end{bmatrix} + \mathbf{e}\mathbf{q}^T = \begin{bmatrix} 0 & 5 & 2 & 3 \\ 5 & 0 & 2 & 3 \\ 3 & 5 & 0 & 2 \\ 3 & 5 & 2 & 0 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 10 & 0 & 0 & 0 \\ 13 & -2 & -1 & 0 \\ 13 & 0 & -3 & 0 \\ 15 & 0 & 0 & -5 \end{bmatrix} + \mathbf{e}\mathbf{q}^T = \begin{bmatrix} 0 & 2 & 3 & 5 \\ 3 & 0 & 2 & 5 \\ 3 & 2 & 0 & 5 \\ 5 & 2 & 3 & 0 \end{bmatrix} \text{ and } B = \begin{bmatrix} 10 & 9 \\ 9 & 10 \end{bmatrix}.$$

Then

$$A = \begin{bmatrix} 0 & 5 & 2 & 3 & 0 & 0 & 0 & 0 \\ 5 & 0 & 2 & 3 & 0 & 0 & 0 & 0 \\ 3 & 5 & 0 & 2 & 0 & 0 & 0 & 0 \\ 3 & 5 & 2 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 3 & 5 \\ 0 & 0 & 0 & 0 & 3 & 0 & 2 & 5 \\ 0 & 0 & 0 & 0 & 3 & 2 & 0 & 5 \\ 0 & 0 & 0 & 0 & 5 & 2 & 3 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 & 9 & 0 & 0 & 0 \\ 9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

is the desired matrix. In the same way, we may construct a nonnegative matrix for each one of the remaining *JCF*.

Here, we consider more general lists of complex numbers and we give new sufficient conditions for the *URP* to have a solution. We know that diagonalizability is a necessary condition. Since normal matrices are diagonalizable, we set the following result [20], in which we use normal *ODP* matrices:

Theorem 3.2 [20] Let $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ be a list of complex numbers with $\Lambda = \overline{\Lambda}$, $\lambda_1 \geq \max_i |\lambda_i|,\ \sum_{i=1}^{n} \lambda_i \geq 0$ and let

$$\begin{aligned}\Lambda &= \Lambda_0 \cup \Lambda_1 \cup \cdots \cup \Lambda_{p_0}\\ \Lambda_0 &= \{\lambda_{01}, \lambda_{02}, \ldots, \lambda_{0p_0}\},\ \ \lambda_{01} = \lambda_1\\ \Lambda_k &= \left\{\lambda_{k1}, \lambda_{k2}, \ldots, \lambda_{kp_k}\right\},\ \ k = 1, \ldots, p_0,\end{aligned}$$

where some lists $\Lambda_k,\ k = 1, \ldots, p_0,$ can be empty. Suppose that the following conditions are satisfied:

i. For each $k = 1, \ldots, p_0,$ there exists a normal *ODP* matrix with spectrum

$$\Gamma_k = \left\{\omega_k, \lambda_{k1}, \ldots, \lambda_{kp_k}\right\}, 0 \le \omega_k < \lambda_1, \text{where } \omega_k \text{ is the Perron eigenvalue.}$$

ii. There exists a $p_0 \times p_0$ normal *ODP* matrix with spectrum Λ_0 and diagonal entries $\omega_1 \ge \omega_2 \ge \cdots \ge \omega_{p_0}$.

Then Λ is $\mathcal{UR}$.

Proof: From *i*) let A_k be a $(p_k + 1) \times (p_k + 1)$ normal *ODP* matrix with spectrum

$$\Gamma_k = \left\{\omega_k, \lambda_{k1}, \ldots, \lambda_{kp_k}\right\}, k = 1, \ldots, p_0, 0 \le \omega_k < \lambda_1.$$

Then $A = diag\{A_1, A_2, \ldots, A_{p_0}\}$ is an $n \times n$ normal nonnegative matrix realizing $\Gamma = \Gamma_1 \cup \cdots \cup \Gamma_{p_0}$. Let

$$X = \begin{bmatrix} \mathbf{x}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_{p_0} \end{bmatrix}$$

the matrix of normalized eigenvectors of A, where $\mathbf{x}_k = \left[x_{k1}, \ldots, x_{kp_k}\right]^T$ is the Perron eigenvector of $A_k,\ A_k\mathbf{x}_k = \omega_k\mathbf{x}_k$ with $\|\mathbf{x}_k\| = 1$. Since A_k is an *ODP* matrix, it is nonnegative irreducible, and then $\mathbf{x}_k$ is a positive eigenvector.

From *ii*) let B be a $p_0 \times p_0$ normal *ODP* matrix with spectrum Λ_0 and diagonal entries $\omega_1, \ldots, \omega_{p_0}$. Let $\Omega = diag\{\omega_1, \ldots, \omega_{p_0}\},$ and $C = B - \Omega$. Then, since A, X and C are nonnegative with A and C normal, $M = A + XCX^T$ is normal nonnegative with spectrum Λ. Moreover, since $A_k,\ k = 1, \ldots, p_0,$ and C are *ODP* matrices, M is also an *ODP* matrix. Thus, Λ is realizable by a normal *ODP* matrix, and from the extension in Johnson et al. [17] Λ is $\mathcal{UR}$.

Many of the known sufficient conditions in the literature about both problems, *NIEP* and *URP*, have been obtained from Theorem 1.3 (Rado's Theorem). A number of distinct versions of Rado's result have also been obtained. In Arrieta et al. [21], it has been proved as regards a Rado diagonalizable version. It will be useful for constructing diagonalizable nonnegative matrices, with prescribed spectrum, and for deciding about the universal realizability of spectra.

Theorem 3.3 [21] Let A be an $n \times n$ diagonalizable matrix with spectrum $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$. Let $X = [\mathbf{x}_1|\mathbf{x}_2|\cdots|\mathbf{x}_r]$ be an $n \times r$ matrix, with $rank(X) = r,\ r \le n,$ such that $A\mathbf{x}_i = \lambda_i\mathbf{x}_i,\ i = 1, \ldots, r$. Let

$\Omega = diag\{\lambda_1, \ldots, \lambda_r\}$ and let $C = [c_{ij}]$ be an $r \times r$ matrix such that $B = \Omega + C$ is a diagonalizable matrix. Let $J(A) = S^{-1}AS$ be the *JCF* of A, with $S = [X|Y]$, $S^{-1} = \begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix}$. Then, the matrix $A + XC\tilde{X}$ is diagonalizable with spectrum $\mu_1, \ldots, \mu_r, \lambda_{r+1}, \ldots, \lambda_n$, where $\mu_1, \ldots, \mu_r$ are eigenvalues of B.

Proof: Since $S^{-1}S = I_n$, then $\tilde{X}X = I_r$, $\tilde{X}Y = 0$, $\tilde{Y}X = 0$, $\tilde{Y}Y = I_{n-r}$. Since $J(A) = \Omega \oplus D$, with $D = diag\{\lambda_{r+1}, \ldots, \lambda_n\}$, then,

$$\begin{aligned} S^{-1}(A + XC\tilde{X})S &= \Omega \oplus D + S^{-1}XC\tilde{X}S \\ &= \Omega \oplus D + \begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix} XC\tilde{X}[X \;\; Y] \\ &= \Omega \oplus D + \begin{bmatrix} I_r \\ 0 \end{bmatrix} C[I_r \quad 0] \\ &= \Omega \oplus D + \begin{bmatrix} C & 0 \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} B & \mathbf{0} \\ \mathbf{0} & D \end{bmatrix}. \end{aligned}$$

Then $A + XC\tilde{X}$ is diagonalizable, and from Theorem 1.3, it has the spectrum $\mu_1, \ldots, \mu_r, \lambda_{r+1}, \ldots, \lambda_n$, where $\mu_1, \ldots, \mu_r$ are eigenvalues of B.

Corollary 3.1 [21] If in Theorem 3.3, the matrices A and B are nonnegative diagonalizable and, $X, \tilde{X}$ are nonnegative, then $A + XC\tilde{X}$ is nonnegative diagonalizable with spectrum $\mu_1, \ldots, \mu_r, \lambda_{r+1}, \ldots, \lambda_n$.

If in Theorem 3.3, the matrix A is block diagonal, with diagonalizable *ODP* blocks, and the matrix B is diagonalizable *ODP*, then Λ is $\mathcal{UR}$.

Example 3.2 Consider the spectrum

$$\Lambda = \{7,2,0,-1,-2+i,-2-i,-2+i,-2-i\},$$

with $\Lambda_0 = \{7,2,0\}$, $\Gamma_1 = \Gamma_2 = \{4,-2+i,-2+i\}$, $\Gamma_3 = \{1,-1\}$.The matrices

$$B = \begin{bmatrix} 4 & \sqrt{\frac{24}{5}} & \sqrt{\frac{6}{5}} \\ \sqrt{\frac{24}{5}} & 4 & 2 \\ \sqrt{\frac{6}{5}} & 2 & 1 \end{bmatrix},$$

$$A_1 = A_2 = \begin{bmatrix} 0 & 2 & 2 \\ 3 & 0 & 1 \\ 1 & 3 & 0 \end{bmatrix}, \; A_3 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

are diagonalizable with spectrum Λ_0, Γ_1, Γ_2, and Γ_3, , respectively. Then,

$$A = (A_1 \oplus A_2 \oplus A_3) + XC\tilde{X}$$

is diagonalizable *ODP* with spectrum Λ. Hence, from the extension in Johnson et al. [17], Corollary 4.1, Λ is $\mathcal{UR}$.

The following result from Ref. [21] gives a diagonalizable version of a Lemma by Fiedler [22] that may be applied to decide the universal realizability of some lists of complex numbers. Although the result is more general, we establish it here, without proof, for diagonalizable nonnegative matrices.

Corollary 3.2 [21] Let A and B be $n\times n$ and $m\times m$ diagonalizable nonnegative matrices with spectrum $\Gamma_1 = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ and $\Gamma_2 = \{\beta_1, \beta_2, \ldots, \beta_m\}$, respectively. Let $\mathbf{u}$ and $\mathbf{v}$, with $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$, be the Perron eigenvectors of A and B, associated to α_1 and β_1, respectively. Let $M = A \oplus B$ and let $J(M) = S^{-1}MS$ be the *JCF* of M, with $S = [X|Y]$, $S^{-1} = \begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix}$. Then for $\rho > 0$, the matrix $F = \begin{bmatrix} A & \rho\mathbf{u}\mathbf{v}^* \\ \rho\mathbf{v}\mathbf{u}^* & B \end{bmatrix}$ is diagonalizable with spectrum $\Lambda = \{\gamma_1, \gamma_2, \alpha_2, \ldots, \alpha_n, \beta_2, \ldots, \beta_m\}$, where γ_1, γ_2 are eigenvalues of $B = \begin{bmatrix} \alpha_1 & \rho \\ \rho & \beta_1 \end{bmatrix}$.

Corollary 3.3 [21] If in Corollary 3.2, A and B are normal nonnegative *ODP* matrices, then the matrix

$$F = \begin{bmatrix} A & \rho\mathbf{u}\mathbf{v}^* \\ \rho\mathbf{v}\mathbf{u}^* & B \end{bmatrix},$$

is normal nonnegative *ODP* with the prescribed spectrum Λ. Hence, Λ is $\mathcal{UR}$.

Proof: If A and B are *ODP* matrices, they are irreducible. Therefore, the Perron eigenvectors $\mathbf{u}$ and $\mathbf{v}$ are positive. Thus, F is normal nonnegative *ODP*, and Λ is $\mathcal{UR}$.

The following result also establishes a universal realizability criterion.

Theorem 3.4 [21] Let $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ be a realizable list of complex numbers, with λ_1, λ_2 being real numbers. Suppose that the lists

$$\Lambda_1 = \{\alpha_1, \alpha_2, \ldots, \alpha_p\}, \quad \Lambda_2 = \{\beta_1, \beta_2, \ldots, \beta_q\},$$

with $p + q = n$, $\alpha_i \in \Lambda$, $i = 2, 3, \ldots, p$, $\beta_j \in \Lambda$, $j = 2, 3, \ldots, q$, and $\alpha_1 + \beta_1 = \lambda_1 + \lambda_2$, are diagonizably *ODP* realizable. Then, Λ is $\mathcal{UR}$.

Proof: Let A_1 and A_2 be $p\times p$ and $q\times q$ diagonalizable *ODP* matrices, with spectrum Λ_1 and Λ_2, respectively. Let $\mathbf{u}$ and $\mathbf{v}$, $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$, be the Perron eigenvectors of A_1 and A_2, associated to the Perron eigenvalues α_1 and β_1, respectively. Since A_1 and A_2 are irreducible, then $\mathbf{u}$ and $\mathbf{v}$ are positive. As $\alpha_1 + \beta_1 = \lambda_1 + \lambda_2$, there is a real number $\rho > 0$, such that the 2×2 matrix $\begin{bmatrix} \alpha_1 & \rho \\ \rho & \beta_1 \end{bmatrix}$ has eigenvalues λ_1, λ_2. Therefore, from Corollary 3.2,

$$A = \begin{bmatrix} A_1 & \rho\mathbf{u}\mathbf{v}^T \\ \rho\mathbf{v}\mathbf{u}^T & A_2 \end{bmatrix} \text{is diagonalizable ODP with spectrum } \Lambda$$

Hence, from the extension in Johnson et al. [17] Λ is $\mathcal{UR}$.

Finally, we apply Theorem 3.3 to more general lists of complex numbers (not in the left half-plane). For instance,

i) $\Lambda = \{8,6,3,3,-5,-5,-5,-5\}$, with $\Lambda_0 = \{8,6\},\ \Lambda_1 = \Lambda_2 = \{3,-5,-5\}$ and $\Gamma_1 = \Gamma_2 = \{7,3,-5,-5\},$ is diagonalizably *ODP* realizable and therefore it is $\mathcal{UR}$. ii) $\Lambda = \{7,5,1,1,-4,-4,-6\}$ with $\Lambda_0 = \{7,5\},\ \Gamma_1 = \{6,1,1,-4,-4\},\ \Gamma_2 = \{6,-6\}$ is also diagonalizably *ODP* realizable and therefore it is $\mathcal{UR}$. iii) $\Lambda = \{10,3,2,-1,-1,3i,-3i,-1\pm 2i,-1\pm 2i\}$ with $\Lambda_0 = \{10,3,2\},\ \Gamma_1 = \{6,3i,-3i\},\ \Gamma_2 = \Gamma_3 = \left\{\frac{9}{2},-1,-1\pm 2i\right\}$ is the spectrum of a diagonalizable nonnegative matrix $A \in \mathcal{CS}_{10}$ with a positive column. Hence, it is also $\mathcal{UR}$. iv) $\Lambda = \{13,3,1+4i,1-4i,1+4i,1-4i\}$ with $\Lambda_0 = \{13,3\},\ \Gamma_1 = \Gamma_2 = \{8,1+4i,1-4i\}$ is the spectrum of a diagonalizable positive matrix. Hence, it is also $\mathcal{UR}$. One of the advantages of applying this procedure is that to decide whether $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ is $\mathcal{UR}$ we do not need to compute a nonnegative matrix for each *JCF* allowed by Λ. It is enough to show that Λ is diagonalizably *ODP* realizable or Λ is the spectrum of a diagonalizable nonnegative matrix with constant row sums and a positive column.

4. The *URP* for structured matrices

4.1 The *URP* for permutative matrices

An $n\times n$ permutative matrix is a matrix in which every row is a permutation of its first row, that is,

Definition 4.1 Let $\mathbf{x} \in \mathbb{R}^n$ and let $P_2, P_3, \ldots, P_n$ be $n\times n$ permutation matrices. A permutative matrix is any matrix of the form

$$P = \begin{bmatrix} \mathbf{x}^T \\ (P_2\mathbf{x})^T \\ \vdots \\ (P_n\mathbf{x})^T \end{bmatrix}.$$

It is clear that $P \in \mathcal{CS}_S,$ where S is the sum of the entries of the vector $\mathbf{x}$. Permutative matrices were introduced and first studied in Hu et al. [23]. In this section, we study the permutative universal realizability problem, that is, the problem of determining the existence and construction of a nonnegative permutative matrix, with prescribed complex spectrum $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$, for each possible *JCF* allowed by Λ.

The following result gives a sufficient condition for that a list $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ of real numbers, with $\lambda_1 > 0 > \lambda_2 > \lambda_3 \geq \lambda_4 \geq \cdots \geq \lambda_n,$ to be permutatively universally realizable.

Theorem 4.1 [24] Let $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ be a list of real numbers, with $\lambda_1 > 0 > \lambda_2 > \lambda_3 \geq \lambda_4 \geq \cdots \geq \lambda_n$. Then the following statements are equivalent:

1. $\sum_{i=1}^{n} \lambda_i \geq 0$,

2. Λ is permutatively realizable,

3. Λ is permutatively $\mathcal{UR}$.

The following example illustrates Theorem 4.1. It shows how we may obtain, a permutative realization, for each *JCF* allowed by a given real list $\lambda_1 > 0 > \lambda_2 > \lambda_3 \geq \cdots \geq \lambda_n$.

Example 4.1 Let us consider the list

$$\Lambda = \{30, -1, -5, -5, -5, -7, -7\}.$$

We start with

$$B = \begin{bmatrix} 30 & 0 & 0 & 0 & 0 & 0 & 0 \\ 31 & -1 & 0 & 0 & 0 & 0 & 0 \\ 35 & 4 & -5 & -4 & 0 & 0 & 0 \\ 35 & 4 & 0 & -5 & -4 & 0 & 0 \\ 35 & 0 & 0 & 0 & -5 & 0 & 0 \\ 31 & 6 & 2 & 0 & 0 & -7 & -2 \\ 31 & 6 & 0 & 0 & 0 & 0 & -7 \end{bmatrix}.$$

Then, for $\mathbf{q}^T = [\,-30 \quad 1 \quad 5 \quad 5 \quad 5 \quad 7 \quad 7\,]$, we have that

$$B + \mathbf{e}\mathbf{q}^T = \begin{bmatrix} 0 & 1 & 5 & 5 & 5 & 7 & 7 \\ 1 & 0 & 5 & 5 & 5 & 7 & 7 \\ 5 & 5 & 0 & 1 & 5 & 7 & 7 \\ 5 & 5 & 5 & 0 & 1 & 7 & 7 \\ 5 & 1 & 5 & 5 & 0 & 7 & 7 \\ 1 & 7 & 7 & 5 & 5 & 0 & 5 \\ 1 & 7 & 5 & 5 & 5 & 7 & 0 \end{bmatrix}$$

is permutative with $JCF\, J\left(B + \mathbf{e}\mathbf{q}^T\right) = diag\{J_1(30), J_1(-1), J_3(-5), J_2(-7)\}$. For

$$B + \mathbf{e}\mathbf{q}^T = \begin{bmatrix} 30 & 0 & 0 & 0 & 0 & 0 & 0 \\ 31 & -1 & 0 & 0 & 0 & 0 & 0 \\ 35 & 4 & -5 & -4 & 0 & 0 & 0 \\ 35 & 0 & 0 & -5 & 0 & 0 & 0 \\ 35 & 0 & 0 & 0 & -5 & 0 & 0 \\ 31 & 6 & 2 & 0 & 0 & -7 & -2 \\ 31 & 6 & 0 & 0 & 0 & 0 & -7 \end{bmatrix} + \mathbf{e}\mathbf{q}^T,$$

we obtain the $JCF\, J\left(B + \mathbf{e}\mathbf{q}^T\right) = diag\{J_1(30), J_1(-1), J_2(-5), J_1(-5), J_2(-7)\}$, and so on.

From Theorem 1.3, we have the following Rado permutative version:

Theorem 4.2 [24] Let $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ be a realizable list of real numbers, where $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0 > \lambda_{p+1} \geq \lambda_{p+2} \geq \cdots \geq \lambda_n$, with $-\lambda_n \geq \lambda_2$, $n \geq 2p$ for n even, and $n \geq 2p + 1$ for n odd n, $p \geq 2$. Suppose that:

i. Λ admits a decomposition $\Lambda = \Lambda_0 \cup \underbrace{\Lambda_1 \cup \cdots \cup \Lambda_1}_{p\text{times}}$, where

$$\Lambda_0 = \{\lambda_1, \lambda_2, \ldots, \lambda_p\}, \quad \Lambda_1 = \{\lambda_{11}, \lambda_{12}, \ldots, \lambda_{1r}\},$$
$$\lambda_{1k} \in \{\lambda_{p+1}, \lambda_{p+2}, \ldots, \lambda_n\}, \quad k = 1, 2, \ldots, r,$$

such that $\Gamma_1 = \{\lambda\} \cup \Lambda_1, \ 0 \leq \lambda \leq \lambda_1,$ is permutatively (circulant) realizable.

ii. There exists a $p \times p$ permutative (circulant) nonnegative matrix with spectrum Λ_0 and diagonal entries $\lambda, \lambda, \ldots, \lambda$ (p times).

Then, Λ is permutatively $\mathcal{UR}$.

Example 4.2 Consider the spectrum

$$\Lambda = \{4,1,1,-2,-2,-2\}, \text{with}$$
$$\Lambda_0 = \{4,1,1\}, \Gamma_1 = \{2,-2\}.$$

Then,

$$B = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}, \text{and } A_1' = \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix},$$

are permutative, realizing Λ_0 and $\Gamma_1,$ respectively. Then,

$$A_1 = A_1' \oplus A_1' \oplus A_1' + XC = \begin{bmatrix} 0 & 2 & 1 & 0 & 1 & 0 \\ 2 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 2 & 1 & 0 \\ 1 & 0 & 2 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 2 \\ 1 & 0 & 1 & 0 & 2 & 0 \end{bmatrix}$$

is nonnegative permutative with diagonal *JCF*. Next, for

$$A'' = \begin{bmatrix} 0 & 2 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \\ -1 & 1 & 0 & 0 & 2 & 0 \end{bmatrix},$$

we obtain $A_2 = A'' + XC,$ nonnegative permutative, with *JCF* having one 2×2 Jordan block $J_2(-2)$. Next, for

$$A''' = \begin{bmatrix} 0 & 2 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 2 \\ 0 & 0 & 0 & 0 & 2 & 0 \end{bmatrix},$$

we obtain $A_3 = A''' + XC$, nonnegative permutative, with *JCF* having a 3×3 Jordan block $J_3(-2)$. Thus, $\Lambda = \{4,1,1,-2,-2,-2\}$ is permutatively $\mathcal{UR}$.

4.2 The *URP* for centrosymmetric matrices

Centrosymmetric matrices have applications in many fields, such as physics, communication theory, differential equations, numerical analysis, engineering and statistics. An $n\times n$ real matrix $C = [c_{ij}]$ is said to be centrosymmetric, if its entries satisfy the relation $c_{ij} = c_{n-i+1,n-j+1}$, or equivalently if $J_nCJ_n = C$, where $J_n = [\mathbf{e}_n, \mathbf{e}_{n-1}, \ldots, \mathbf{e}_1]$. In this section, we study the centrosymmetric universal realizability problem, that is, the problem of determining conditions for the existence and construction of a centrosymmetric realizing matrix, for each *JCF* allowed by a given list Λ of complex numbers.

The following two results are of constructive nature, in the sense that if they are satisfied, then a centrosymmetric realizing matrix with spectrum Λ can be constructed for each *JCF* associated to Λ. We introduce them here without proof.

Theorem 4.3 [25] Let $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ be a realizable list of complex numbers with λ_1 simple, $n = 2m$. Suppose $\Lambda = \Lambda_1 \cup \Lambda_2$ with $\Lambda_1 \cap \Lambda_2 = \emptyset$, where Λ_1 is diagonalizably realizable by an $m\times m$ matrix W_1 and Λ_2 is the spectrum of an $m\times m$ real diagonalizable matrix W_2 (not necessarily nonnegative). If $W_1 + W_2$ is *ODP* and $W_1 - W_2$ is positive, then Λ is centrosymmetrically $\mathcal{UR}$.

Theorem 4.4 [25] Let $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ be a realizable list of complex numbers with λ_1 simple, $n = 2m + 1$. Suppose $\Lambda = \Lambda_1 \cup \Lambda_2$ with $\Lambda_1 \cap \Lambda_2 = \emptyset$, where Λ_1 is diagonalizably realizable by the $(m + 1) \times (m + 1)$ *ODP* matrix

$$\begin{bmatrix} W_1 & \mathbf{a} \\ \mathbf{b}^T & c \end{bmatrix}$$

and Λ_2 is the spectrum of an $m\times m$ real diagonalizable matrix W_2. If $W_1 + W_2$ is *ODP* and $W_1 - W_2$ is positive, then Λ is centrosymmetrically $\mathcal{UR}$.

Example 4.3 Consider the list

$$\Lambda = \{10, -2, -2, -2, -1 + 2i, -1 - 2i, -1 + 2i, -1 - 2i\}.$$

We apply Theorem 4.3 to show that Λ is centrosymmetrically $\mathcal{UR}$. We take

$$\Lambda_1 = \{10, -2, -2, -2\}, \Lambda_2 = \{-1 + 2i, -1 - 2i, -1 + 2i, -1 - 2i\}$$

which are the spectrum of

$$W_1 = \begin{bmatrix} 1 & 3 & 3 & 3 \\ 3 & 1 & 3 & 3 \\ 3 & 3 & 1 & 3 \\ 3 & 3 & 3 & 1 \end{bmatrix} \text{ and } \quad W_2 = \begin{bmatrix} -1 & -2 & 0 & 0 \\ 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 2 & -1 \end{bmatrix}$$

respectively. Next, we compute the centrosymmetric *ODP* matrix

$$C_1 = \frac{1}{2}\begin{bmatrix} W_1 + W_2 & (W_1 - W_2)J_4 \\ J_4(W_1 - W_2) & J_4(W_1 + W_2)J_4 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 0 & 1 & 3 & 3 & 3 & 3 & 5 & 2 \\ 5 & 0 & 3 & 3 & 3 & 3 & 2 & 1 \\ 3 & 3 & 0 & 1 & 5 & 2 & 3 & 3 \\ 3 & 3 & 5 & 0 & 2 & 1 & 3 & 3 \\ 3 & 3 & 1 & 2 & 0 & 5 & 3 & 3 \\ 3 & 3 & 2 & 5 & 1 & 0 & 3 & 3 \\ 1 & 2 & 3 & 3 & 3 & 3 & 0 & 5 \\ 2 & 5 & 3 & 3 & 3 & 3 & 1 & 0 \end{bmatrix}$$

with diagonal *JCF*. Now, we compute a centrosymmetric *ODP* matrix C_2 with *JCF*

$$J(C_2) = diag\{J_1(10), J_2(-2), J_1(-2), J_2(-1+2i), J_2(-1-2i)\}.$$

To do this, we first compute the matrix of eigenvectors of C_1 :

$$S = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & -1 & 0 & -1 \\ 1 & 0 & 1 & 0 & 0 & -i & 0 & i \\ 1 & 1 & 0 & 0 & -1 & 0 & -1 & 0 \\ 1 & -1 & -1 & -1 & -i & 0 & i & 0 \\ 1 & -1 & -1 & -1 & i & 0 & -i & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & i & 0 & -i \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

Then for $E = E_{2,3} + E_{5,6} + E_{7,8}$, we have

$$SES^{-1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{3}{8} & \frac{3}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & \frac{3}{8} & -\frac{5}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & -\frac{7}{8} & \frac{1}{8} \\ \frac{1}{8} & -\frac{7}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ -\frac{5}{8} & \frac{3}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & \frac{3}{8} & \frac{3}{8} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and taking $\mathbf{q} = -\mathbf{d}$, where $\mathbf{d}$ is the vector of the diagonal entries of SES^{-1} we obtain that $SES^{-1} + \mathbf{e}\mathbf{q}^T$ has all its diagonal entries being zero. Thus,

$$C_2 = C_1 + \left(SES^{-1} + \mathbf{eq}^T\right) = \begin{bmatrix} 0 & \frac{1}{2} & \frac{13}{8} & \frac{11}{8} & \frac{11}{8} & \frac{13}{8} & \frac{5}{2} & 1 \\ \frac{5}{2} & 0 & \frac{13}{8} & \frac{11}{8} & \frac{11}{8} & \frac{13}{8} & 1 & \frac{1}{2} \\ \frac{15}{8} & \frac{15}{8} & 0 & \frac{1}{4} & \frac{9}{4} & 1 & \frac{15}{8} & \frac{7}{8} \\ \frac{13}{8} & \frac{13}{8} & \frac{11}{4} & 0 & 1 & \frac{3}{4} & \frac{5}{8} & \frac{13}{8} \\ \frac{13}{8} & \frac{5}{8} & \frac{3}{4} & 1 & 0 & \frac{11}{4} & \frac{13}{8} & \frac{13}{8} \\ \frac{7}{8} & \frac{15}{8} & 1 & \frac{9}{4} & \frac{1}{4} & 0 & \frac{15}{8} & \frac{15}{8} \\ \frac{1}{2} & 1 & \frac{13}{8} & \frac{11}{8} & \frac{11}{8} & \frac{13}{8} & 0 & \frac{5}{2} \\ 1 & \frac{5}{2} & \frac{13}{8} & \frac{11}{8} & \frac{11}{8} & \frac{13}{8} & \frac{1}{2} & 0 \end{bmatrix}$$

is nonnegative centrosymmetric with spectrum Λ and with the desired *JCF* $J(C_2)$. Applying the same procedure, changing the matrix E, say by, $E_1 = E_{2,3}$, $E_2 = E_{2,3} + E_{3,4}$, $E_3 = E_{2,3} + E_{3,4} + E_{5,6} + E_{7,8}$, $E_4 = E_{5,6} + E_{7,8}$, we may construct a nonnegative centrosymmetric matrix with spectrum Λ, for each one of the other four *JCF* allowed by Λ.

Theorems 4.3 and 4.4 allow us to show that certain real spectra of nonnegative numbers and of the Suleĭmanova type are centrosymmetrically $\mathcal{URC}$ (Corollaries 4.1 and 4.2 below).

In Ref. [7], Perfect introduces the $n \times n$ (matrix)

$$P = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & \cdots & 1 & -1 \\ 1 & 1 & \cdots & -1 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & -1 & \cdots & 0 & 0 \end{bmatrix} \tag{4}$$

and she proves that if $D = diag\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ with $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n \geq 0,$ then PDP^{-1} is a positive matrix in $\mathcal{CS}_{\lambda_1}$.

Corollary 4.1 [25] Let $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ be a list of nonnegative real numbers with $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. If $\lambda_m > \lambda_{m+1}$ when $n = 2m$ and $\lambda_{m+1} > \lambda_{m+2}$ when $n = 2m + 1$, then Λ is centrosymmetrically UR.

Proof: For $n = 2m$, we define the $m \times m$ diagonalizable positive matrix with spectrum $\Lambda_1 = \{\lambda_1, \ldots, \lambda_m\}$ as $W_1 = PDP^{-1}$, where $D = diag\{\lambda_1, \ldots, \lambda_m\}$ and P is the matrix in (4). Let $W_2 = diag\{\lambda_{m+1}, \ldots, \lambda_n\}$ with spectrum Λ_2. Note that $\Lambda = \Lambda_1 \cup \Lambda_2$ and since $\lambda_m > \lambda_{m+1}$, $\Lambda_1 \cap \Lambda_2 = \emptyset$. Moreover, it is clear that $W_1 + W_2$ is ODP. On the other hand, in [Julio et al. [26], Theorem 3.2], it was proved that if $d_{jj}, j = 1, 2, \ldots, m$, are the diagonal entries of $PDP^{-1},$ then $d_{jj} > \lambda_{m+j}$, for all $j = 1, 2, \ldots, m$. Thus, $W_1 - W_2$ is positive. Then, from Theorem 4.3, Λ is centrosymmetrically $\mathcal{URC}$.

For $n = 2m + 1$, we define the $(m + 1) \times (m + 1)$ diagonalizable positive matrix with spectrum $\Lambda_1 = \{\lambda_1, \ldots, \lambda_{m+1}\}$ as $PDP^{-1} = \begin{bmatrix} W_1 & \mathbf{a} \\ \mathbf{b}^T & c \end{bmatrix}$, where $D = diag\{\lambda_1, \ldots, \lambda_{m+1}\}$ and P is the $(m + 1) \times (m + 1)$ matrix in (4). Let $W_2 = diag\{\lambda_{m+2}, \ldots, \lambda_n\}$ with spectrum Λ_2. Again $\Lambda = \Lambda_1 \cup \Lambda_2$, $\Lambda_1 \cap \Lambda_2 = \emptyset$, $W_1 + W_2$ is *ODP* and $W_1 - W_2$ is positive. Then, from Theorem 4.4, Λ is centrosymmetrically $\mathcal{UR}$.

Corollary 4.2 [25] Let $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ be a realizable list of real numbers with $\lambda_1 > 0 > \lambda_2 \geq \cdots \geq \lambda_n$. If $\lambda_m > \lambda_{m+1}$ when $n = 2m$ and $\lambda_{m+1} > \lambda_{m+2}$ when $n = 2m + 1$, then Λ is centrosymmetrically $\mathcal{UR}$.

Proof: We assume without loss of generality that $\sum_{i=1}^{n} \lambda_i = 0$. For $n = 2m$ we define

$$W_1 = \begin{bmatrix} \lambda_1 & & & \\ \lambda_1 - \lambda_2 & \lambda_2 & & \\ \vdots & \vdots & \ddots & \\ \lambda_1 - \lambda_m & 0 & \cdots & \lambda_m \end{bmatrix} + \mathbf{e}\mathbf{q}^T$$

$$= \begin{bmatrix} -\lambda_{m+1} & -\lambda_{m+2} - \lambda_2 & \cdots & -\lambda_n - \lambda_m \\ -\lambda_{m+1} - \lambda_2 & -\lambda_{m+2} & \cdots & -\lambda_n - \lambda_m \\ \vdots & \vdots & \ddots & \vdots \\ -\lambda_{m+1} - \lambda_m & -\lambda_{m+2} - \lambda_2 & \cdots & -\lambda_n \end{bmatrix},$$

where $\mathbf{q}^T = [-\lambda_{m+1} - \lambda_1, -\lambda_{m+2} - \lambda_2, \cdots, -\lambda_n - \lambda_m]$. Note that W_1 is an $m \times m$ diagonalizable positive matrix with spectrum $\Lambda_1 = \{\lambda_1, \ldots, \lambda_m\}$. Let $W_2 = diag\{\lambda_{m+1}, \ldots, \lambda_n\}$ with spectrum Λ_2. It is clear that $\Lambda = \Lambda_1 \cup \Lambda_2$ and since $\lambda_m > \lambda_{m+1}$, $\Lambda_1 \cap \Lambda_2 = \emptyset$. Moreover,

$$W_1 + W_2 = \begin{bmatrix} 0 & -\lambda_{m+2} - \lambda_2 & \cdots & -\lambda_n - \lambda_m \\ -\lambda_{m+1} - \lambda_2 & 0 & \cdots & -\lambda_n - \lambda_m \\ \vdots & \vdots & \ddots & \vdots \\ -\lambda_{m+1} - \lambda_m & -\lambda_{m+2} - \lambda_2 & \cdots & 0 \end{bmatrix}$$

is *ODP* and

$$W_1 - W_2 = \begin{bmatrix} -2\lambda_{m+1} & -\lambda_{m+2} - \lambda_2 & \cdots & -\lambda_n - \lambda_m \\ -\lambda_{m+1} - \lambda_2 & -2\lambda_{m+2} & \cdots & -\lambda_n - \lambda_m \\ \vdots & \vdots & \ddots & \vdots \\ -\lambda_{m+1} - \lambda_m & -\lambda_{m+2} - \lambda_2 & \cdots & -2\lambda_n \end{bmatrix}$$

is positive. Then, from Theorem 4.3, Λ is centrosymmetrically $\mathcal{UR}$.

For $n = 2m + 1$ we define the $(m + 1) \times (m + 1)$ diagonalizable nonnegative matrix with spectrum $\Lambda_1 = \{\lambda_1, \ldots, \lambda_{m+1}\}$ as

$$\begin{bmatrix} W_1 & \mathbf{a} \\ \mathbf{b}^T & c \end{bmatrix} = \begin{bmatrix} \lambda_1 & & & & \\ \lambda_1-\lambda_2 & \lambda_2 & & & \\ \vdots & \vdots & \ddots & & \\ \lambda_1-\lambda_m & 0 & \cdots & \lambda_m & \\ \lambda_1-\lambda_{m+1} & 0 & \cdots & 0 & \lambda_{m+1} \end{bmatrix} + \mathbf{e}\mathbf{q}^T$$

$$= \begin{bmatrix} -\lambda_{m+2} & -\lambda_{m+3}-\lambda_2 & \cdots & -\lambda_n-\lambda_m & -\lambda_{m+1} \\ -\lambda_{m+2}-\lambda_2 & -\lambda_{m+3} & \cdots & -\lambda_n-\lambda_m & -\lambda_{m+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\lambda_{m+2}-\lambda_m & -\lambda_{m+3}-\lambda_2 & \cdots & -\lambda_n & -\lambda_{m+1} \\ -\lambda_{m+2}-\lambda_{m+1} & -\lambda_{m+3}-\lambda_2 & \cdots & -\lambda_n-\lambda_m & 0 \end{bmatrix},$$

where $\mathbf{q}^T = [-\lambda_{m+2}-\lambda_1, -\lambda_{m+3}-\lambda_2, \cdots, -\lambda_n-\lambda_m, -\lambda_{m+1}]$.

Let $W_2 = diag\{\lambda_{m+2}, \ldots, \lambda_n\}$ with spectrum Λ_2. Then from Theorem 4.4, the result follows.

4.3 The *URP* for M-matrices

M-matrices appear in many applications in the physical, biological and social sciences. A real matrix A is said to be an M-matrix if it is of the form $A = \alpha I - B$, where B is a nonnegative matrix and Although M-matrices are not nonnegative, they are related to nonnegative matrices. For instance, it is well known that the inverse of a nonsingular M-matrix is nonnegative. Moreover, the problem of finding an M-matrix $A = \alpha I - B$ with prescribed complex spectrum $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ can be seen as the problem of finding a nonnegative matrix B with eigenvalues $\alpha-\lambda_1, \alpha-\lambda_2, \ldots, \alpha-\lambda_n$. In Soto et al. [27], the authors study the *URP* for M-matrices. More precisely, they give sufficient conditions for the existence of M-matrices with prescribed elementary divisors. In particular, they solve the *URP* for certain lists of real numbers and for lists of complex numbers of the form $\Lambda = \{\lambda_1, a \pm bi, \ldots, a \pm bi\}$. In Soto et al. [27], the inverse M-matrix problem for symmetric generalized doubly stochastic M-matrices is also considered.

Proposition 4.1 Let $A = \alpha I - B$ be an $n \times n$ M-matrix. Then.

i. $B \in \mathcal{CS}_{\lambda_1}$ if and only if $A \in \mathcal{CS}_{\alpha-\lambda_1}$.

ii. $B, B^T \in \mathcal{CS}_{\lambda_1}$ if and only if $A, A^T \in \mathcal{CS}_{\alpha-\lambda_1}$.

iii. B is normal if and only if A is normal.

iv. B is symmetric if and only if A is symmetrix.

v. B is circulant if and only if A is circulant.

vi. A and B have the same eigenvectors.

Next result is an M-matrix version of Brauer Theorem.

Theorem 4.5 Let $A = (\alpha I - B) \in \mathcal{CS}_{\lambda_1}$ be an M-matrix with spectrum $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, $\lambda_1 > |\lambda_i|$, $i = 2, \ldots, n$. Let B be with a positive column. Then $A + \mathbf{eq}^T$ is also an M-matrix with the same spectrum and with elementary divisors as A.

Proof: $A + \mathbf{eq}^T = \alpha I - B + \mathbf{eq}^T = \alpha I - \left(B - \mathbf{eq}^T\right)$. As B is nonnegative with its *kth* column being positive, then by taking $\mathbf{q} = [q_i]$, $q_i \leq 0$, $i \neq k$, $q_k =$

$-\sum_{i \neq k} q_i \leq \min_{1 \leq i \leq n} \{b_{ik}\}$, we have that $B - \mathbf{eq}^T$ is nonnegative and $\sum_{i=1}^{n} q_i = 0$. Thus, from Theorem 1.4, $A + \mathbf{eq}^T$ is also an M-matrix with the same spectrum and with elementary divisors as A.

Theorem 4.6 Let $A = \alpha I - B$ be a diagonalizable M-matrix with spectrum $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, where B is positive. Then there is an M-matrix with spectrum Λ for each *JCF* allowed by Λ.

Proof: Since A is diagonalizable then $B = \left(b_{ij}\right)$ is diagonalizable and positive. Then there is a positive matrix B' with same spectrum as B for each *JCF* allowed by the spectrum of B. Hence, $A' = \alpha I - B'$ is an M-matrix with same spectrum as A for each *JCF* allowed by Λ.

Theorem 4.7 Let $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ be with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n-1} > \lambda_n \geq 0$. Then there exists a generalized stochastic M-matrix $A \in \mathcal{CS}_{\lambda_n}$ with spectrum Λ for each *JCF* allowed by Λ.

Proof: Let $\alpha \geq \lambda_1$. Consider the list

$$\Lambda' = \{\alpha - \lambda_n, \alpha - \lambda_{n-1}, \ldots, \alpha - \lambda_2, \alpha - \lambda_1\}.$$

Let $D = diag\{\alpha - \lambda_n, \alpha - \lambda_{n-1}, \ldots, \alpha - \lambda_1\}$ and let P the Perfect matrix in (4). Then $B = PDP^{-1} \in \mathcal{CS}_{\alpha - \lambda_n}$ is positive with spectrum Λ' and diagonal *JCF*. Let $D + E$, with E the matrix defined in the Introduction, the desired *JCF*. Then we have

$$D + E = P^{-1}BP + E = P^{-1}\left(B + PEP^{-1}\right)P.$$

It is clear that for $\varepsilon > 0$ small enough, $B + \varepsilon PEP^{-1}$ is positive with *JCF D+E* Therefore, since $D + \varepsilon E$ and *D+E* are similar (with the matrix $M = diag\left\{1, \varepsilon, \varepsilon^2 . \ldots, \varepsilon^{n-1}\right\}$), $A = \alpha I - \left(B + \varepsilon PEP^{-1}\right)$ is an M-matrix in $\mathcal{CS}_{\lambda_1}$ with the prescribed elementary divisors.

Theorem 4.8 Let $\Lambda = \{\lambda_1, a \pm bi, \ldots, a \pm bi\}$ be a list of n complex numbers with $\lambda_1 \geq 0$, $a \geq 0$, $b > 0$. If

$$\lambda_1 \leq a - \frac{n+1}{2}b, \tag{5}$$

then there exists an $n \times n$ M-matrix with spectrum Λ for each *JCF* allowed by Λ.

Proof: Let $\alpha \geq a$. Consider the list

$$\Lambda' = \{\alpha - \lambda_1, (\alpha - a) \pm bi, \ldots, (\alpha - a) \pm bi\}$$

Since $\lambda_1 \leq a - \frac{n+1}{2}b$, then $\alpha - \lambda_1 \geq \alpha - a + \frac{n+1}{2}b$. From [Arrieta et al. (21), Theorem 2.1], if (5) holds then there exists a nonnegative matrix $B \in \mathcal{CS}_{\alpha - \lambda_1}$ with spectrum Λ' and the prescribed elementary divisors. Then $A = \alpha I - B$ is an M-matrix with spectrum Λ for each *JCF* allowed by Λ.

5. Conclusions

In this chapter, we revisit the problem of universal realizability of spectra [2], with new advances and results, which contain two important extensions of the Minc result [4] and new universal realizability criteria for general and structured matrices. Several open questions have been answered, while others remain open, such as under what conditions diagonalizably realizable implies universally realizable? Recent developments on this question are introduced in Soto and Marijuán [28].

Author details

Ana I. Julio* and Ricardo L. Soto
Department of Mathematics, Universidad Católica del Norte, Antofagasta, Chile

*Address all correspondence to: ajulio@ucn.cl

IntechOpen

© 2024 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

References

[1] Soto RL. Nonnegative inverse eigenvalue problem, Capítulo 5 in linear algebra. In: Yasser HA, editor. Theorems and Applications. London, UK: IntechOpen; 2012. pp. 99-116

[2] Soto RL. In: Katsikis VN, editor. Nonnegative Inverse Elementary Divisors Problem, Capítulo 4 in Applied Linear Algebra in Action. London, UK: IntechOpen; 2016. pp. 85-114

[3] Johnson CR. Row stochastic matrices similar to doubly stochastic matrices. Linear and Multilinear Algebra. 1981;**10**: 113-130

[4] Minc H. Inverse elementary divisor problem for nonnegative matrices. Proceedings of the American Mathematical Society. 1981;**83**:665-669

[5] Minc H. Inverse elementary divisor problem for doubly stochastic matrices. Linear and Multilinear Algebra. 1982;**11**: 121-131

[6] Brauer A. Limits for the characteristic roots of a matrix IV. Applications to stochastic matrices. Duke Mathematical Journal. 1952;**19**:75-91

[7] Perfect H. Methods of constructing certain stochastic matrices II. Duke Mathematical Journal. 1955;**22**: 305-311

[8] Soto RL, Ccapa J. Nonnegative matrices with prescribed elementary divisors. Electronic Journal of Linear Algebra. 2008;**17**:287-303

[9] Soto RL, Rojo O, Moro J, Borobia A. Symmetric nonnegative realization of spectra. Electronic Journal of Linear Algebra. 2007;**16**:1-18

[10] Laffey TJ, Šmigoc H. Nonnegative realization of spectra having negative real parts. Linear Algebra and its Applications. 2006;**416**:148-159

[11] Suleĭmanova HR. Stochastic matrices with real characteristic values. Doklady Akademii Nauk SSSR. 1949;**66**:343-345

[12] Soto RL, Díaz RC, Nina H, Salas M. Nonnegative matrices with prescribed spectrum and elementary divisors. Linear Algebra and its Applications. 2013;**439**:3591-3604

[13] Díaz RC, Soto RL. Nonnegative inverse elementary divisors problem in the left half plane. Linear and Multilinear Algebra. 2016;**64**:258-268

[14] Julio AI, Marijuán C, Pisonero M, Soto RL. Universal realizability in low dimension. Linear Algebra and its Applications. 2021;**619**:107-136

[15] Borobia A, Moro J. On nonnegative matrices similar to positive matrices. Linear Algebra and its Applications. 1997;**266**:365-379

[16] Collao M, Salas M, Soto RL. Spectra universally realizable by doubly stochastic matrices. Special Matrices. 2018;**6**:301-309

[17] Johnson CR, Julio AI, Soto RL. Nonnegative realizability with Jordan structure. Linear Algebra and its Applications. 2020;**587**:302-313

[18] Soto RL. Existence and construction of nonnegative matrices with prescribed spectrum. Linear Algebra and its Applications. 2003;**369**:169-184

[19] Collao M, Johnson CR, Soto RL. Universal realizability of spectra with two positive eigenvalues. Linear Algebra and its Applications. 2018;**545**:226-239

[20] Julio AI, Soto RL. On the universal realizability problem. Linear Algebra and its Applications. 2020;**597**:170-186

[21] Arrieta LE, Millano AD, Soto RL. On spectra realizable and diagonalizably realizable. Linear Algebra and its Applications. 2021;**612**:273-288

[22] Fiedler M. Eigenvalues of nonnegative symmetric matrices. Linear Algebra and its Applications. 1974;**9**: 119-142

[23] Hu X, Johnson CR, Davis C, Zhang EY. Ranks of permutative matrices. Special Matrices. 2016;**4**: 233-246

[24] Soto RL, Julio AI, Alfaro JH. Permutative universal realizability. Special Matrices. 2021;**9**:66-77

[25] Julio AI, Linares YR, Soto RL. Centrosymmetric universal realizability. Electronic Journal of Linear Algebra. 2021;**37**:680-691

[26] Julio AI, Rojo O, Soto RL. Centrosymmetric nonnegative realization of spectra. Linear Algebra and its Applications. 2019;**581**:260-284

[27] Soto RL, Díaz RC, Salas M, Rojo O. M-matrices with prescribed elementary divisors. Inverse Problems. 2017;**33**: 095009

[28] Soto, R. L., Marijuán C., How Diagonalizably Realizable Implies Universally Realizable, Pre-print.

Edited by Peter Y.P. Chen and Victor Martinez-Luaces

Nonlinear system analysis is of interest to engineers, sociologists, physicists, mathematicians, and many other scientists since most systems are inherently nonlinear in nature. In mathematics, a nonlinear system does not satisfy the superposition principle such as in a linear system. Therefore, the theories underlining nonlinear analysis and their applications need to be developed on their own merit. The first section of this book is a collection of examples reporting recent advances in both theory and applications of nonlinear system analysis. The contents of each chapter will provide in-depth foresight to interested readers. As numerical linearization to a set of matrix equations is still the principal method used to solve a nonlinear system, matrix analysis is the topic of the second section of this book. The matrices have invaded practically all areas of mathematics, the experimental and social sciences, engineering, and technology. This volume updates purely mathematical theoretical aspects, and it also presents concrete examples of the wide range of applications of matrix theory in other disciplines.

Published in London, UK
© 2024 IntechOpen
© StudioM1 / iStock

IntechOpen

ISBN 978-1-83769-450-1