KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Global Solution to Problems in System Identification and Model Order Reduction

Sibren LAGAUW

Dissertation presented in partial fulfillment
of the requirements for the degree of
Doctor of Engineering Science (PhD):
Electrical Engineering

Supervisor:
Prof. dr. ir. B. De Moor

January 2026

GLOBAL SOLUTION TO PROBLEMS IN SYSTEM IDENTIFICATION AND MODEL ORDER REDUCTION

Sibren LAGAUW

Supervisor:
Prof. dr. ir. B. De Moor

Members of the
Examination Committee:
Prof. dr. ir. L. Vandewalle, chair
Prof. dr. ir. K. Meerbergen
Dr. ir. K. De Cock
Prof. dr. ir. L. De Lathauwer
Dr. ir. S. Telen
(MPI MiS, Leipzig)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

January 2026

© 2026 Sibren LAGAUW
Uitgegeven in eigen beheer, Sibren LAGAUW, Kasteelpark Arenberg 10, 3001 Leuven (België) (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Preface

First and foremost, I would like to thank my supervisor, Prof. Bart De Moor, for his advice and guidance. For creating the "back-to-the-roots" program, which laid out the overarching research framework within which my fellow PhD students and I were able to grow, learn, and develop ourselves. The program sparked my interest in fundamental problems in systems and control, and guided me toward discovering new and fascinating technical topics. On a day-to-day basis, Bart expected me to take initiative and assume responsibility for my own research path. He believed in me and gave me the means to execute. This balance between independence and guidance inevitably led to moments of doubt and difficulty, but I am confident that the process has been formative.

The work presented in this thesis was supported by the Research Foundation Flanders (FWO) through a Fundamental Research PhD Fellowship (Grant No. 11K5625N), for which I am sincerely grateful. I would also like to thank the members of my supervisory committee for their guidance and constructive feedback throughout my PhD, as well as the members of my examination committee for their insightful comments and careful reading of the manuscript.

Every obstacle becomes more manageable when approached with a second pair of eyes. I am deeply grateful for the many collaborations I have been part of during this PhD. Collaboration is not without its challenges: it requires articulating ideas with clarity and allowing others into one's thought process—often to question assumptions or probe weak points. Yet I have learned that this process is immensely valuable. Engaging in collaborative work not only improves the quality of the research itself, but also enriches the experience on a human level, as collaborators motivate one another and complement each other's strengths.

I would therefore like to thank my direct collaborators at our research group—Mauricio, Lukas, Sem, and Christof—for their insightful interactions, intellectual rigor, attention, and patience. I am incredibly proud of our joint work and our

co-authored publications. Together with Katrien, Benoît, Hans, Sarthak, and Marie, these people form the "rooters" group, comprising members of our unit working within the context of the "back-to-the-roots" project. I would like to thank the rooters for being exceptional companions along the way, and for the many insightful discussions and engaging group meetings.

Thanks to the kind invitation of Simon, I was able to undertake a two-week research visit to the Numerical Algebraic Geometry group at the Max Planck Institute for Mathematics in the Sciences in Leipzig in October 2024. This visit had a substantial impact on my doctoral work and evolved into a long-term collaboration with Leonie, a PhD student in Simon's group. I have greatly enjoyed working with Leonie on our shared projects; her energy and curiosity were invaluable during the final stage of my PhD. I have learned a great deal from her, and I will be incredibly proud to see our joint research papers published. I am equally grateful to Bart and Simon for fostering an environment in which this collaboration could develop.

Ultimately, it was my master's thesis that convinced me to pursue this PhD. I would like to thank Christof, Katrien, and Philippe for their inspirational guidance during this project. You set an example that I decided I wanted to follow. During my PhD, I in turn had the opportunity to supervise several master's thesis projects. In this role, I was fortunate to work alongside excellent companions. Thank you, Lukas, Katrien, Lola, Vincent, and Hans; I greatly enjoyed our co-supervision.

Not all of my time at the offices near Arenberg Castle was spent on research. Beyond the names already mentioned, I would also like to thank Aldona, Alexandre, Axel, Bram, Elsy, Fernando, Giulia, Ida, Jaldert, John, Kostas, Melanie, Naomi, Oliver, Thomas, Yamini, Yoeri, and Zahra for the enjoyable lunch discussions, walks around the park, your help with administrative matters, and the many coffee chats that made daily life at the department all the more pleasant.

This PhD was not only an academic pursuit, but also a personal one, and I would like to thank those who were part of that journey. Thanks to Oscar, for keeping me up to date with what is new among the tech hippies of Silicon Valley; from podcasts to new songs. Thanks to Anton, Stijn, and Francis: our yearly ski trips were always something to look forward to. Thanks to Emile, Marie, and Mattias, for the good times we shared in the Kozijntjesstraat, and for the many occasions on which you chose not to say out loud, "what on earth are you working on?" when catching a glimpse of my computer screen. Thanks to Aline for the coffee breaks while working from home; to Alexander for the evening tennis and padel sessions; and to Tom for the occasional catch-up running sessions. Thanks to Dominiek, for accompanying me on escapes to the

sun somewhere around the Mediterranean. Thanks to Xanne, Edward, and Jolien, for occasionally reminding me how we behaved as students. Thanks to Jochem and Sem for the fun DIY weekends. Thanks to Colin, Yente, Thijs and the others for forming such a warm group of new friends in Ghent.

Above all, I owe my deepest gratitude to my family. To my parents, Martin and Oda, whose unwavering support has accompanied me throughout these past years. To Mieke, Denise, and Marcel: over the course of this PhD, I have come to see how much of who I am carries traces of each of you. To Jan, Martine, An and Stefaan for the warm welcome with which they embraced me into their family. To Lars: it has been some time since you were my little brother; you are now someone I genuinely look up to. To Uk, the most wonderful four-legged companion I could imagine. And finally, to Laura—my partner in crime. I am already incredibly proud of you, and we still have so many adventures ahead of us.

In a world where so much is changing, where technology evolves at an astonishing pace, and where the fundaments of our society are questioned and rewritten, the rigidity of the exact mathematics that I explored in this PhD often felt like a safe haven. Mathematics is a game with fixed and well-defined rules. Given a set of assumptions, something is either right or wrong. Although it may be difficult to obtain, questions posed in this setting admit decisive answers. Once the mathematics—and, where applicable, the numerics—work out, there is little room left for discussion.

Over the past four years of working on this thesis, one is inclined to grow accustomed to this mode of clear-cut reasoning. I realize, however, that the world in which we live behaves quite differently. Questions rarely admit a single answer, and different answers may be equally defensible depending on the perspective from which they are approached. Scientific methods are intended to allow us to draw objective conclusions from collected information, yet such information is not always readily available, and even when it is, its sources can—and should—be questioned. The current political and environmental climate additionally ensures that confidence intervals widen and long-tail events begin to dominate. Certainty is elusive, and, prediction machines as we are, we are forced to continually readjust our expectations.

After completing this PhD, I intend to step closer towards the above-mentioned, messy, true world. While it is chaotic, noisy, and therefore difficult to navigate, it also offers plenty opportunities to shape and improve it; to contribute positively to the place and the people around us. It will take some adjustment to re-

embrace uncertainty, nuance, and the coexistence of multiple, context-dependent truths. Yet I look forward to these challenges.

Beyond abstract thinking, independence, project management, and the many other skills one is expected to develop during a PhD, there are two lessons that resonated particularly deeply with me and have left a lasting impression. They are insights that, I believe, extend well beyond academia and are valuable to anyone.

The first is the value of structure. Through reading scientific literature and contributing to it myself, I have come to realize that the way knowledge is organized is at least as important as the knowledge itself. As Jeff Hawkins describes[1], learning is not merely the acquisition of new information; rather, it is the process of integrating that information into an existing knowledge base. That is, "learning" consists of placing a new knowledge node within a structured framework of previously established and interconnected concepts. The connections between these nodes encode relationships, dependencies, and hierarchies among ideas, while the strength of each connection encodes a notion of distance. The act of "thinking" can then be understood as movement within this framework: progressing from one concept to another through cause and effect, inference, and deduction.

Research, then, is fundamentally an exercise in structuring. New results are always relative to an existing representation of knowledge; they extend it, reshape it, or expose its blind spots. What distinguishes strong research is not the sheer volume of information, but the clarity with which ideas are organized and the connections that are drawn. This is why certain scientific books or papers read as if everything is obvious: their authors have constructed a framework in which each idea appears inevitable. That, to me, is true mastery. Similarly, once a new result is properly embedded within such a structure, it becomes difficult to "unsee" it. Much like the structure of this dissertation (cf. Figure 1.4), which may now appear inevitable, but was anything but at the outset.

I believe this is one of the most important skills we learn through our training as researchers: the ability to start from nothing and create a valuable, structured framework within which one can begin to reason and explore; and to have the courage and ability to adjust that framework when necessary. Carrying this lesson forward means, in everything that we do, remaining aware of the framework, or belief system, that shapes how we perceive and act.

In a natural way, this leads to the second insight, which is of a more reflective nature: the power of perception. Over the course of this PhD, I have come

[1] Jeff Hawkins, *"A Thousand Brains"*, 2022.

to realize that while we have little control over the physical world around us, we do have tremendous control over how we perceive it. Trying to obsessively control reality itself is, more often than not, a losing battle. What we can shape, however, is the lens through which that "reality" reaches us—how experiences are interpreted, integrated into our belief system, and given meaning. A small shift in that lens can be more powerful than any attempt to change the world directly. After all, the world does not exist independently of observation: by perceiving and experiencing it, we participate in its creation. One may therefore choose to construct a world that is loving rather than harsh, inspiring rather than draining and beautiful rather than broken. Let us create a world filled with wonder, admiration, appreciation and imagination.

Popularized Abstract

The world is full of systems that evolve and interact over time. *Systems theory* provides the mathematical tools to describe these systems, allowing us, among other uses, to simulate, predict, and control real-world processes. A *system* may be understood as an interconnected set of components acting together in time, or more abstractly, as the set of all signals it can generate. This abstraction makes systems theory applicable across a wide range of disciplines: systems may represent physical entities, such as industrial processes, biological networks, or mechanical structures, but equally abstract constructs in economics or sociology.

Within this framework, two key challenges arise: how to infer these mathematical models directly from measured data (system identification), and how to simplify complex models without losing the most important features (model reduction). Both tasks involve solving a non-convex optimization problem. The state-of-the-art techniques used in engineering practice are highly effective, but they provide only locally optimal solutions, which need not correspond to the true best model.

This dissertation takes a fundamentally different approach. For two specific problem instances, we develop a novel, *deterministic* method that characterizes the complete set of so-called *critical points* (the "valleys" and "hills" of the optimization landscape), thereby allowing us to identify the *globally optimal* solution(s). The results deepen our theoretical understanding while establishing useful benchmarks for future computational methods.

Gepopulariseerde Samenvatting

De wereld bestaat uit dynamische systemen: processen die zich in de tijd ontwikkelen en onderling interageren. *Systeemtheorie* biedt een wiskundig kader om dergelijke systemen te beschrijven, waardoor het mogelijk wordt om deze processen te simuleren, te voorspellen en te regelen. Een *dynamisch systeem* kan worden gezien als een geheel van verbonden componenten die samen doorheen de tijd evolueren, of, in meer abstracte zin, als de verzameling van alle signalen en configuraties die het kan genereren. Deze abstractie maakt systeemtheorie toepasbaar in zowat elke discipline: systemen kunnen fysisch zijn, zoals industriële processen, biologische netwerken of mechanische structuren, maar evenzeer abstracte constructen in de economie of sociologie.

Binnen dit systeemtheoretisch kader zijn er twee belangrijke uitdagingen: hoe kan men, gegeven een systeem, een passend wiskundig model construeren op basis van metingen (systeemidentificatie), en hoe kan men complexe modellen vereenvoudigen zonder de belangrijkste eigenschappen te verliezen (modelreductie). Beide taken leiden tot een niet-convex optimalisatieprobleem. De technieken die hiervoor in de ingenieurspraktijk worden gebruikt zijn effectief, maar leveren slechts lokaal optimale oplossingen, die niet noodzakelijk overeenkomen met het globaal optimale model.

Dit proefschrift hanteert een fundamenteel andere benadering. Voor twee specifieke probleemopstellingen werken we een nieuwe, *deterministisch* methode uit die de volledige verzameling van lokaal optimale punten (de “dalen” en “heuvels” van het glooiende optimalisatielandschap) karakteriseert. Dit stelt ons in staat om de *globaal optimale* oplossing(en) uit deze set te selecteren. De verkregen resultaten verdiepen ons theoretisch inzicht in de onderliggende structuur van de problemen, en leveren ons tegelijkertijd waardevolle ijkingspunten voor toekomstige computationele methodes.

Abstract

Systems theory provides a mathematical framework for describing dynamical phenomena through the abstract notion of a "system". Such a system is typically described by differential (continuous-time) or difference (discrete-time) equations that govern the evolution of internal states, and their relationship to inputs and outputs, over time. The abstract notion of a system makes it applicable across a wide range of disciplines: systems may represent physical entities, such as industrial processes, biological networks, or mechanical structures, but equally abstract constructs in economics or sociology. The obtained mathematical models form the basis for analysis, simulation, prediction, and control, the latter of which refers to the manipulation of the system's inputs to achieve certain objectives (e.g., stabilization, trajectory tracking or disturbance rejection).

This dissertation addresses two closely related topics: *system identification*, the task of inferring dynamical models from observed data, and *model reduction*, the approximation of a given high-order model by a model of lower complexity. Both problems naturally lead to non-convex optimization formulations, even when restricted to linear model classes. Consequently, the state of the art is dominated by iterative approaches that compute (at best) locally optimal solutions. While such methods are highly effective in engineering practice, they leave important questions unresolved: global optimality cannot be guaranteed, so one is never certain whether even better solutions exist, the outcome often depends strongly on the chosen initialization, and reproducibility is not always guaranteed. From a mathematical standpoint, these problems can therefore not yet be considered "solved."

In this dissertation, we develop a novel *deterministic* methodology that leads to *globally optimal* solutions. We specifically focus on the (weighted) least squares realization problem for identifying autonomous linear time-invariant models and the single-input single-output $\mathcal{H}_2$-norm model reduction problem. The key insight is that, for both problems, the first-order optimality conditions can be expressed in polynomial form, with their real common roots fully characterizing

the set of (local) minimizers. By exploiting the partially linear structure, we reformulate the polynomial system as a *multiparameter eigenvalue problem (MEP)*, which generalizes the standard eigenvalue problem by allowing more than one spectral parameter and removing the restriction of linearity in its parameter(s). The solutions of these MEPs, the so-called *eigentuples*, contain the globally minimizing solutions of the original problem.

The proposed framework is powerful in that it provides a unified methodology for two problems that are closely related yet fundamentally distinct. The resulting formulations also represent a substantial advancement over the state of the art in globally optimal approaches, both in terms of improved computational efficiency as theoretical simplicity. Finally, the contributions are non-trivial: in particular, the role of Walsh's Theorem as an important building block in the development of globally optimal solution methods had not previously been recognized.

The pursuit of global optimality, however, comes at a considerable computational cost. The efficiency of the proposed methodologies is governed by the difficulty of computing all common roots of the system of polynomial equations, or equivalently, the eigentuples of the resulting MEPs, the number of which grows rapidly with the problem size. For this reason, the proposed framework is best viewed as a theoretical benchmark, rather than a practical tool for large-scale applications.

To better understand these computational limitations, we present an algebro-geometric analysis of the solution sets of the obtained MEPs. This comprehensive analysis, which forms our second main contribution, leads to novel insights into the solution set to the critical equations. Most notably, we establish reducedness of the solutions of the MEPs (i.e., each solution occurs with multiplicity one) and, through the Thom–Porteous formula, we derive exact solution counts.

Beknopte samenvatting

Systeemtheorie biedt een wiskundig kader om dynamische fenomenen te beschrijven via het abstracte begrip van een "systeem". Deze systemen worden doorgaans voorgesteld door differentiaalvergelijkingen (continue-tijd) of differentievergelijkingen (discrete-tijd) die de evolutie van interne toestanden en hun relatie tot ingangs- en uitgangssignalen in de tijd beschrijven. Het brede begrip van een systeem maakt de theorie toepasbaar: systemen kunnen fysisch zijn, zoals industriële processen, biologische netwerken of mechanische structuren, maar evenzeer abstracte constructen in de economie of sociologie. De wiskundige modellen vormen de basis voor analyse, simulatie, voorspelling en regeling, waarbij dit laatste verwijst naar het manipuleren van de ingangen van het systeem om bepaalde doelstellingen te bereiken (bijvoorbeeld stabilisatie, het volgen van een traject of onderdrukking van verstoringen).

Dit proefschrift behandelt twee gerelateerde onderwerpen: *systeemidentificatie*, het opstellen van dynamische modellen op basis van waargenomen data, en *modelreductie*, de benadering van een gegeven model door een model van lagere complexiteit. Beide problemen leiden tot een niet-convex optimalisatieprobleem, zelfs wanneer enkel lineaire modelklassen worden beschouwd. Omwille van de niet-convexiteit wordt de *state of the art* gedomineerd door iteratieve methodes die (in het beste scenario) resulteren in een lokaal optimale oplossing. Hoewel deze methodes in de ingenieurspraktijk zeer effectief zijn, blijven belangrijke theoretische vragen onbeantwoord: globale optimaliteit kan niet worden gegarandeerd, waardoor men nooit zeker is of er betere oplossingen bestaan; het resultaat hangt vaak sterk af van de gekozen initialisatie; en reproduceerbaarheid van de resultaten is niet altijd gegarandeerd. Vanuit een wiskundig perspectief kunnen deze problemen daarom nog niet als "opgelost" worden beschouwd.

In dit proefschrift werken we een nieuwe, *deterministische* methode uit die leidt tot *globaal optimale* oplossingen. We richten ons specifiek op het (gewogen) kleinste-kwadraten-realisatieprobleem voor het identificeren van autonome

lineaire tijdsinvariante modellen en het enkele-ingang enkele-uitgang $\mathcal{H}_2$-norm modelreductieprobleem. Het belangrijkste inzicht is dat voor beide problemen de eerste-orde optimaliteitsvoorwaarden kunnen worden voorgesteld aan de hand van veeltermvergelijkingen, waarbij hun reële gemeenschappelijke nulpunten de volledige verzameling van lokaal optimale oplossingen karakteriseren. Door gebruik te maken van de gedeeltelijk lineaire structuur herformuleren we het bekomen veeltermstelsel als een *multiparameter-eigenwaardeprobleem (MEP)*, een veralgemening van het standaard eigenwaardeprobleem waarin meerdere spectrale parameters voorkomen en de beperking tot lineariteit in de parameter(s) wordt opgeheven. De oplossingen van deze MEPs, de zogenoemde *eigentuples*, bevatten de globaal optimale oplossingen van het oorspronkelijke probleem.

De voorgestelde methode is krachtig omdat het een uniforme aanpak biedt voor twee problemen die gerelateerd maar fundamenteel verschillend zijn. De resulterende methodes zijn bovendien aanzienlijk beter dan de bestaande globaal optimale methodes, zowel wat betreft verbeterde computationele eigenschappen als theoretische eenvoud. Tenslotte zijn de bijdragen niet triviaal: de rol van *Walsh's Theorem* als een belangrijke bouwblok in de ontwikkeling van globaal optimale oplossingsmethodes was tot dusver niet gekend.

De zoektocht naar globale optimaliteit komt echter met een aanzienlijke computationele kost. De effectiviteit van de uitgewerkte methodes wordt volledig bepaald door de moeilijkheid van het berekenen van alle gemeenschappelijke nulpunten van het stelsel veeltermvergelijkingen, of equivalent, de oplossingen van de resulterende MEPs, waarvan het aantal snel toeneemt met de probleemgrootte. De resultaten moeten daarom in de eerste plaats worden gezien als een theoretisch ijkpunt, eerder dan als een praktisch instrument dat bruikbaar is voor grootschalige toepassingen.

Om de computationele beperkingen beter in kaart te brengen, bestuderen we de oplossingsverzamelingen van de verkregen MEPs vanuit algebraisch-geometrisch standpunt. Deze uitgebreide analyse, die onze tweede fundamentele contributie vormt, leidt tot nieuwe inzichten in de oplossingsruimte van de zogenoemde kritieke vergelijkingen. We tonen aan dat de oplossingsverzamelingen van de MEPs *gereduceerd* zijn (d.w.z. dat elke oplossing met multipliciteit één voorkomt), en bekomen via de Thom–Porteous-formule een exacte uitdrukking voor het oplossingsaantal.

Contents

Preface i

Popularized Abstract vii

Gepopulariseerde Samenvatting ix

Abstract xi

Beknopte samenvatting xiii

Contents xv

List of Figures xxi

List of Tables xxv

List of Abbreviations xxvii

List of Symbols xxix

Introduction 1

1 Introduction 2

1.1 Motivational Prologue 2

1.1.1 System identification and model reduction 3

1.1.2 Non-convexity and engineering practice 4

1.1.3 Achieving global optimality 8

1.1.4 Research objectives 9

1.2 Overview of the Dissertation 10

1.2.1 Part I: Novel globally optimal solution approaches . . . 12

1.2.2 Part II: Solution set of the critical equations 16

1.2.3 Part III: Other contributions to system identification . . 21
1.2.4 General considerations 24

I Novel Globally Optimal Solution Approaches 25

2 Globally Optimal SISO $\mathcal{H}_2$-Norm Model Reduction Using Walsh's Theorem 26
2.1 Introduction 29
2.1.1 Problem formulation 29
2.1.2 Literature overview 30
2.1.3 Contributions and outline 31
2.2 Characterizing First-Order Optimality 31
2.2.1 Interpolatory conditions for optimality 32
2.2.2 Walsh's Theorem 33
2.3 Methodology 34
2.3.1 System of multivariate polynomial equations 34
2.3.2 Quadratic m-parameter eigenvalue problem 37
2.3.3 Computing the critical points 39
2.3.4 Discrete-time setting 41
2.3.5 Secular equations 41
2.3.6 Solution set at infinity 42
2.4 Numerical Examples 43
2.5 Conclusion and Future Work 45

3 Exact Solution to the Least Squares Realization Problem as a Multiparameter Eigenvalue Problem 47
3.1 Introduction 50
3.1.1 Problem formulation 50
3.1.2 Contributions and outline 52
3.1.3 Background and literature overview 53
3.2 Model Dynamics 55
3.2.1 Kernel representation 55
3.2.2 Image representation 56
3.3 Characterizing Optimality 57
3.3.1 Orthogonal decomposition of the ambient space 57
3.3.2 Optimal model parameters 60
3.4 Multiparameter Eigenvalue Problem 61
3.4.1 Eigentuples and critical points 61
3.4.2 Computation of the eigentuples 62
3.4.3 Comparison with alternatives 64
3.5 Model Complexity and Lower Order Models 64
3.5.1 Increasing the model complexity 65

3.5.2 Lower-order dynamics . . . 67
3.5.3 Computation of all critical points . . . 69
3.6 Numerical Examples . . . 70
3.7 Weighted Low-Rank Approximation of Hankel Matrices . . . 77
3.7.1 Problem formulation . . . 77
3.7.2 Literature overview . . . 81
3.7.3 Globally optimal weighted low-rank approximation . . . 86
3.7.4 Examples . . . 95
3.8 Sequences with Infinite Length . . . 107
3.9 Conclusions and Future Work . . . 109

4 Incorporating Fixed Pole Information in the Least Squares Realization Problem 111
4.1 Introduction . . . 114
4.2 Problem Formulation and Motivational Example . . . 115
4.3 Methodology . . . 119
4.3.1 Characterizing first-order optimality . . . 119
4.3.2 Solutions with lower-order dynamics . . . 120
4.3.3 Multiparameter eigenvalue problem . . . 121
4.3.4 How to solve the MEP in Theorem 4.5? . . . 122
4.4 Numerical Examples . . . 123
4.5 Conclusions and Future Work . . . 127

II Solution Set of the Critical Equations 129

5 On the Number of Critical Points for SISO $\mathcal{H}_2$-Norm Model Reduction 130
5.1 Introduction . . . 133
5.1.1 Problem formulation . . . 133
5.1.2 Scope and contributions . . . 134
5.1.3 Outline and notation . . . 136
5.2 The Critical Equations . . . 137
5.3 Relevant Background in Algebraic Geometry . . . 140
5.3.1 Affine varieties . . . 140
5.3.2 Dimension of fibers and intersections . . . 145
5.3.3 Projective space and determinantal varieties . . . 148
5.3.4 Degree and a special case of Thom–Porteous . . . 152
5.4 Finiteness of the Solution Set . . . 154
5.5 Multiparameter Eigenvalue Problem . . . 161
5.6 Degree of the Solution Set . . . 164
5.7 Conclusions and Future Work . . . 170

6 The Euclidean Distance Degree for Low-Rank Hankel Matrix Approximation 171
6.1 Introduction . . . 174
6.1.1 Problem formulation . . . 174
6.1.2 Contributions and outline . . . 175
6.2 Euclidean Distance Degree . . . 177
6.3 The Critical Equations . . . 180
6.4 Removal of the Singular Locus . . . 183
6.5 The Substituted ED Correspondence . . . 186
6.6 Multiparameter Eigenvalue Problem . . . 189
6.7 Degree of the Solution Set . . . 192
6.8 The Singular Locus . . . 198
6.9 Conclusions and Future Work . . . 202

III Other Contributions to System Identification 205

7 Least Squares Projection Onto the Behavior for SISO LTI Models 206
7.1 Introduction . . . 209
7.1.1 Problem formulation . . . 209
7.1.2 Literature overview . . . 210
7.1.3 Contributions and outline . . . 211
7.2 Projection onto the Behavior . . . 212
7.2.1 The behavior . . . 213
7.2.2 The misfit space . . . 214
7.2.3 Optimal projection . . . 216
7.3 The State-Space Dual . . . 216
7.3.1 Behavioral system theoretic models . . . 216
7.3.2 Structured misfits . . . 218
7.3.3 Optimal projection . . . 220
7.3.4 The z-domain . . . 221
7.4 Least Squares Realization for SISO LTI Models . . . 222
7.4.1 Notational reformulation . . . 222
7.4.2 First-order optimality . . . 223
7.4.3 Riemannian SVD . . . 227
7.4.4 Global optimality . . . 229
7.5 Conclusions and Future Work . . . 237

8 Least Squares Realization of Complex-Valued Multiple-Output LTI Systems via the Riemannian Singular Value Decomposition 239
8.1 Introduction . . . 242
8.2 Optimization over Complex Decision Variables . . . 244
8.3 Complex Riemannian SVD (RiSVD) . . . 245

8.3.1 Model-compliance 245
8.3.2 Data optimality 246
8.3.3 Model optimality 248
8.3.4 Complex Riemannian SVD (RiSVD) 248
8.4 An Inverse-Iteration Method 249
8.5 Numerical Experiments 251
8.5.1 Nuclear magnetic resonance (NMR) modeling 251
8.5.2 Direction of arrival (DOA) estimation 253
8.6 Conclusions and Future Work 256

Conclusion **259**

9 Contributions, Discussion and Perspectives for Future Work **260**
9.1 Research Contributions 260
9.1.1 SISO $\mathcal{H}_2$-norm model reduction 260
9.1.2 Least squares realization of autonomous LTI models . . 262
9.1.3 Other contributions 263
9.2 Discussion 264
9.2.1 Global optimality and computational complexity 265
9.2.2 Polynomial system solving vs. MEPs 265
9.2.3 Evaluating model fit 267
9.2.4 Hankel matrix approximation and Walsh's Theorem . . 268
9.3 Perspectives for Future Work 268

A Appendices **271**
A.1 Gentle Introduction to Optimization and Systems & Control . 271
A.1.1 Mathematical optimization 271
A.1.2 Systems theory and control 274
A.2 Polynomial Systems and MEPs 278
A.2.1 Polynomial system solving 279
A.2.2 From polynomial system to MEP 282
A.2.3 MEP solvers 284

Bibliography **287**

List of publications **305**

List of Figures

1.1 Comparison of convex and non-convex optimization problems. . 5
1.2 Contour plots of the objective function of the model reduction problem for varying reduced order models. 6
1.3 The system identification cycle. Practical methods are effective but heuristic, lacking guarantees of global optimality. 7
1.4 Overview of the dissertation structure. The dissertation is divided into three main parts. 11
1.5 Schematic overview of the strategy for deterministically computing globally optimal solutions to non-convex optimization problems in model reduction and system identification. 14
1.6 Relation between the polynomial system formulation and the corresponding multiparameter eigenvalue problem (MEP). . . 18
1.7 Schematic overview of the LS realization problem and its extensions considered in Part III. 22

3.1 Schematic representation of the autonomous LTI single-output model with data modified by the misfit. 51
3.2 Contour plot of the LS realization objective in (a_0, a_1), highlighting non-convexity and the distribution of critical points. 51
3.3 Graphical illustration of increasing the model order by adding a second pole to a first-order model. 67
3.4 Given data and optimal model-compliant data at the global minimizer. 71
3.5 Given data and model-compliant data corresponding to the globally optimal solution(s) for $n = 1, \ldots, 5$. 73
3.6 Given data and globally optimal model-compliant data in the (relatively) large-scale example. 74
3.7 Scatter plots of parameter values (a_0, a_1, a_2) for critical points in the (relatively) large-scale example. 75

3.8 Contour plot of the weighted LS realization objective, highlighting non-convexity and the distribution of critical points. 100
3.9 Globally optimal data in the case of unit weights. 103
3.10 Globally optimal data in the case of increasing weights. 104
3.11 Globally optimal data in the case of decreasing weights. 105
3.12 Globally optimal data in the case of cap-shaped weights. 106

4.1 Objective function of the least squares realization problem, illustrating the effect of a fixed pole constraint. 116
4.2 Box plots of realization experiments at different noise levels, comparing S-GOR and FP-GOR performance 126

5.1 Overview of the algebro-geometric notions used throughout the chapter. 141
5.2 Schematic overview of the logical structure and dependencies between the arguments developed in this chapter. 155

6.1 Schematic overview of the Euclidean distance minimization problem and related concepts. 178
6.2 Schematic overview of the logical structure and dependencies between the arguments developed in this chapter. 187
6.3 The generic and complete Euclidean distance degree for several problem sizes (N, r). 201

7.1 Structure of the SISO misfit modeling setup: observed signals are decomposed into model-compliant components and misfits. 210

8.1 Schematic overview of the autonomous LTI multiple-output model with model-compliant data and misfits. 243
8.2 Distribution of the squared 2-norm of the misfits for various noise levels and various number of sensors p. 253
8.3 Distribution of the relative errors on f_4, f_5 for various noise levels and various number of sensors p. 254
8.4 Distribution of the relative errors on d_4, and d_5 for various noise levels and various number of sensors p. 255
8.5 Linear sensor array with equally spaced sensors receiving a signal arriving at a common angle. 255
8.6 Distribution of misfits at varying noise levels. 256
8.7 Histograms of estimated DOA at varying noise levels. 257

A.1 Illustration of the LS fitting problem with data points, regression line, and residuals. 274

A.2 Measured data versus simulated output of a SISO system, with residuals indicated by vertical dotted lines. 277
A.3 Surface plots of a variety. 280
A.4 Contour plot of a function. 280

List of Tables

2.1 Properties of the systems of polynomial equations underlying the globally optimal techniques described in the literature. These equations determine the computational complexity of the proposed solution methods. 40
2.2 Critical points of the *four-disk* example. 44
2.3 The poles of the four critical points with the lowest value objective function value. 45

3.1 Affine common roots of the polynomial system with their corresponding cost values. 64
3.2 Comparison of coefficient matrix sizes of the proposed MEP and an alternative MEP from the literature, for various (N, n). . . . 65
3.3 Objective values, poles, and order for critical points obtained from the proposed MEP formulation. 72
3.4 Global minimum and number of real-valued affine eigentuples for $n = 1, \ldots, 5$. 73
3.5 Objective function values and dominant poles for selected critical points (a_0, a_1, a_2). 76
3.6 Real-valued solutions to the Riemannian SVD with unit weights and the corresponding singular values of the associated Hankel matrices. 90
3.7 Real-valued solutions to the Riemannian SVD with Frobenius weights and the corresponding singular values of the associated Hankel matrices. 90
3.8 Real-valued solutions to the Riemannian SVD and the corresponding singular values of the associated Hankel matrices. . . 97
3.9 Real-valued solutions to the Riemannian SVD and the corresponding singular values of the associated Hankel matrices. . . 98

3.10 Real-valued solutions of the weighted MEP together with the singular values of the corresponding Hankel matrices in the case of unit weights. 103
3.11 Real-valued solutions of the weighted MEP together with the singular values of the corresponding Hankel matrices in the case of increasing weights. 104
3.12 Real-valued solutions of the weighted MEP together with the singular values of the corresponding Hankel matrices in the case of decreasing weights. 105
3.13 Real-valued solutions of the weighted MEP together with the singular values of the corresponding Hankel matrices in the case of cap-shaped weights. 106

4.1 Comparison of realization approaches (NPF, advanced prefilter, FP-GOR, and S-GOR) for the motivational example. 124

5.1 Evaluation of the solution count formula for different problem sizes (N, n) under genericity assumptions. 166

6.1 Values of $\mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r})$ for unit weights, with bold entries matching the generic ED degree. 195
6.2 Values of $\mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r})$ for Frobenius weights, with bold entries matching the generic ED degree. 195
6.3 Values of $\mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r})$ for Bombieri weights, with bold entries matching the generic ED degree. 196
6.4 The common roots of the defining equations of $\mathcal{Y}^{\mathrm{h}}_{\boldsymbol{\Lambda}}$. 198

7.1 Real-valued solutions of the polynomial system with parameter values $\boldsymbol{a}$ and $\boldsymbol{b}$ and singular values of the associated block Hankel matrices. 230
7.2 Real-valued solutions of the polynomial system with parameter values $\boldsymbol{a}$ and $\boldsymbol{b}$ and singular values of the associated block Hankel matrices. 232

8.1 Exact parameter values of the simulated NMR signal used in the experiment. 251

List of Abbreviations

AR	autoregressive
ARMAX	autoregressive moving-average model with exogenous inputs
DOA	direction of arrival
ED	Euclidean distance
EIV	errors-in-variables
FIR	finite impulse response
FONC	first-order necessary conditions for optimality
FP-GOR	fixed pole globally optimal realization
GTLS	global total least squares
IRKA	Iterative rational Krylov approximation
KKT	Karush–Kuhn–Tucker
LS	least squares
LTI	linear time-invariant
NMR	nuclear magnetic resonance
MEP	multiparameter eigenvalue problem
MIMO	multiple-input multiple-output
MO	multiple-output
PEM	prediction-error method
PEP	polynomial eigenvalue problem
RSS	residual sum of squares
SCF	self-consistent field iteration
SISO	single-input single-output
SLRA	structured low-rank approximation
SNR	signal to noise ratio
SO	single-output
SVD	singular value decomposition
S-GOR	standard globally optimal realization
TLS	total least squares
TSD	time series deflation

List of Symbols

Scalars, vectors, matrices and polynomials

a	scalar
j	complex unit
$\boldsymbol{a}$	vector
$\boldsymbol{A}$	matrix
$\boldsymbol{I}_k$	identity matrix of size $k \times k$
$\boldsymbol{x}^{\boldsymbol{\alpha}}$	monomial $\prod_i x_i^{\alpha_i}$
$f(\boldsymbol{x})$	function/polynomial in $\boldsymbol{x}$
$\boldsymbol{M}(\boldsymbol{x})$	matrix polynomial in $\boldsymbol{x}$

Relations

$\coloneqq$	"defined as", used to introduce new notation
$\cong$	isomorphic
$\approx$	approximately equal
$\mapsto$	maps element to its image under a function/morphism

System-theoretic quantities (model reduction and identification)

N	order of given model / length of given data
n	order of to-be-estimated model
$\left\|\ldots\right\|_2$	ℓ_2-norm
$\left\|\ldots\right\|_{\mathcal{H}_2}$	$\mathcal{H}_2$-norm
$(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$	state-space representation
$H(s)$	transfer function of a continuous-time system
$h(t)$	impulse response of a continuous-time system
$H(z)$	transfer function of a discrete-time system
h_k	impulse response of a discrete-time system

Number systems, sets, and algebraic structures

$\emptyset$	empty set
$\mathbb{N}$	nonnegative integers
$\mathbb{Z}$	integers
$\mathbb{R}$	real numbers
$\mathbb{C}$	complex numbers
$\mathbb{P}$	projective space
$(x_0, \ldots, x_n)$	point in $\mathbb{C}^n$
$(x_0 : \cdots : x_n)$	point in $\mathbb{P}^n$
ℓ_2	space of square-summable sequences
$\mathcal{H}_2$	Hardy space on the right half-plane
$P_{\leq k}$	univariate polynomials of degree at most k
$\mathbb{C}[\boldsymbol{x}]$ or R_n	multivariate polynomial ring in n variables
$\mathbb{C}[\boldsymbol{x}]^\ell$	column vector of length ℓ with polynomial entries
X or $\mathcal{X}$	algebraic variety or set
$\overline{X}$	algebraic closure of X
$V(F)$	variety defined by set of polynomials F
$I(X)$	defining ideal of the variety X
$\langle F \rangle$	ideal generated by the set of polynomials F
$\sqrt{I}$	radical of the ideal I
X^{sm}	smooth locus of X
X^{sing}	singular locus of X
$X(\mathbb{R})$	real locus of X; set of points of X with real coordinates
$I_r(\boldsymbol{M})$	ideal generated by the (r)-minors of $\boldsymbol{M}$
$\Sigma_k(\boldsymbol{M})$	determinantal variety defined by $(k+1)$-minors of $\boldsymbol{M}$
$\text{Sym}_{\mathbb{R}}(N)$	set of real symmetric $N \times N$ matrices

Operators and transformations

∂	partial derivative operator
$\nabla(\cdot)$	gradient operator
$\nabla^2(\cdot)$	Hessian matrix
$.^{\mathrm{Re}}$	real part
$.^{\mathrm{Im}}$	imaginary part
$\binom{\cdot}{\cdot}$	binomial coefficient
$\mathrm{diag}(\cdot)$	diagonal matrix
$\mathrm{rev}(\cdot)$	vector reversal
$.^{\mathsf{T}}$	transpose
$.^{\mathsf{H}}$	Hermitian transpose
$.^{-1}$	inverse
$.^{\dagger}$	pseudo-inverse
$.^{\perp}$	orthogonal complement
$\mathrm{rank}(\cdot)$	rank
$\mathrm{null}\,(\cdot)$	null space
$\mathrm{range}\,(\cdot)$	range / column space
$\mathrm{row}\,(\cdot)$	row space
$\mathrm{nullity}\,(\cdot)$	nullity
$\det(\cdot)$	determinant
$\deg(\cdot)$	degree
$\gcd(\cdot)$	greatest common divisor
$\oplus$	direct sum of vector spaces
$\times$	product of vector spaces
$\otimes$	Kronecker product
$\setminus$	set difference
$\cup$	union of sets
$\cap$	intersection of sets
$\subseteq$	subset
$\subsetneq$	proper subset
$\not\subseteq$	not a subset
$\bar{\cdot}$	complex conjugation / closure (depending on context)
$\dim(\cdot)$	dimension
$\mathrm{codim}_X(\cdot)$	codimension wrt. X
$\#(\cdot)$	cardinality
$\mathbb{P}(\cdot)$	projectivization
$\mathrm{EDD}_{\mathbf{\Lambda}}(\cdot)$	Euclidean distance degree wrt. the weight matrix $\mathbf{\Lambda}$
$\mathrm{EDD}_{\mathrm{gen}}(\cdot)$	Euclidean distance degree wrt. generic weights
$\mathrm{EDDdef}(\cdot)$	Euclidean distance degree defect

Introduction

Chapter 1

Introduction

This introductory chapter is organized in two sections. In Section 1.1, we present a motivational prologue that informally introduces the central questions of this dissertation and situates them within the broader scientific and engineering context. Section 1.2 then provides a structured overview of the thesis, outlining its three main parts and summarizing the contents, research objectives and contributions of each.

1.1 Motivational Prologue

This dissertation addresses model reduction and system identification, two cornerstones of data-driven and model-based engineering. *Model reduction* seeks to approximate a given high-order model, often derived from first principles or large-scale simulations, by a model of lower complexity that preserves the most important dynamical features of the original model. *System identification*, in contrast, starts from measured data and aims to infer dynamical models that explain and predict the given data.

Despite their broad applicability and effective use, the current state-of-the-art approaches in both fields remain suboptimal from a mathematical point of view. In engineering practice, reduction and identification tasks are typically addressed using heuristic and/or iterative methods. Consequently, because these problems are inherently non-convex, several fundamental questions remain unresolved: In what sense is the computed solution optimal? Is it unique? Are the results reproducible?

To address this gap, this dissertation takes a fundamentally different perspective: one that, contrary to the conventional engineering practice, focuses on *global optimality* and employs *deterministic*, rather than heuristic, methods. Using tools from systems theory, (numerical) linear algebra and algebraic geometry, we obtain a rigorous and coherent framework for computing globally optimal solutions. The work also deepens our understanding of the mathematical structure of the problems under consideration, leading to novel theoretical insights into their global complexity.

The remainder of this section outlines the motivation behind this dissertation in more detail. It is intentionally brief and limited to essential references, serving as a conceptual prelude to the subsequent chapters, where more comprehensive and topic-specific literature reviews are provided.

1.1.1 System identification and model reduction

System identification In system identification, one builds mathematical models of dynamical systems from measured data and studies their properties [120]. The task is to determine, from given input-output data, a model structure and associated parameters so that the resulting model captures the dynamical behavior of the true system. The model is typically chosen from a user-specified model class, reflecting prior knowledge about the system's nature and complexity (e.g., linear vs. nonlinear, or time-variant vs. time-invariant).

In practice, however, the given data are almost never *model-compliant*, meaning that they do not satisfy the equations of the chosen model class exactly [86, 187]. These discrepancies between measured data and the model structure arise for many reasons: measurement inaccuracies, missing samples, observational errors, or a mismatch between the true system and the chosen model class. A common remedy is to introduce stochastic features, interpreting the discrepancy as random noise [120], but this requires explicit statistical assumptions based on a priori knowledge. An alternative, assumption-free approach is to minimally adjust the observed data by subtracting so-called *misfits*, which are small corrections that make the given data model-compliant, whose magnitude then reflects the "distance" between the obtained model and the given data [114].

Model order reduction Although closely related to system identification, model reduction addresses a different problem. Rather than starting from observed data, one is given a representation of a high-order dynamical system, and the task is to construct a lower-order model that approximates its input-output behavior as accurately as possible. By reducing the order while

maintaining accuracy in the relevant operating regimes, one obtains a more compact model that is significantly cheaper to simulate, analyze, and control. The main use cases of model reduction arise in engineering applications where first-principles modeling leads to very large state-space descriptions, such as fluid dynamics, circuit simulation, or structural mechanics, making direct computation too expensive [10].

These identification and model reduction problems clearly constitute an optimization task: the equations of the chosen model class serve as *constraints*; the *objective function* corresponds to the criterion assessing model quality; and the *decision variables* are the parameters of the selected model class.

1.1.2 Non-convexity and engineering practice

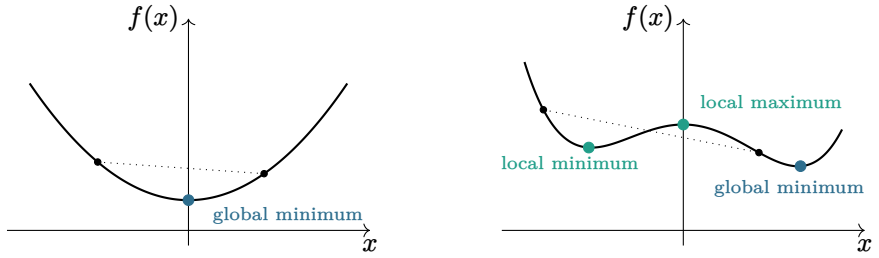
Convex[1] optimization problems have only one critical point[2] (cf. Figure 1.1) and there exist many methods that are able to globally solve large-scale problem instances in polynomial time [18]. Non-convex problems, by contrast, exhibit multiple local minima, local maxima and saddle points. In this case, the first-order necessary conditions for optimality (FONC) of the problem are not sufficient: they do not certify global optimality.

Unfortunately, both model reduction and system identification problems, even when restricted to linear model classes, are typically nonlinear and (heavily) non-convex, so that their solution space contains (many) distinct local optima. This non-convexity is evident even in very small problem instances, as illustrated by the contour plots in Figure 1.2.

Because of this underlying non-convexity, aiming for global optimality in this context is extremely challenging. Instead, the state-of-the-art techniques typically "solve" system identification and model reduction problems by iteratively computing a solution that satisfies the FONC, leading to, at best, locally optimal solutions. The term "iterative" reflects the fact that these methods iteratively update a current estimate $\boldsymbol{x}^{(k)}$, e.g., using information from its local neighborhood, producing a new estimate $\boldsymbol{x}^{(k+1)}$.

[1] A set $\mathcal{X} \subseteq \mathbb{R}^n$ is convex if it contains the line segment between any two of its points. Similarly, a differentiable function $f : \mathcal{X} \to \mathbb{R}$ is convex if $f(\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leq \theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ and $\theta \in [0, 1]$.

[2] In the unconstrained setting, a point is called *critical* if the gradient of the objective function vanishes. More generally, in constrained optimization, critical points are feasible points that satisfy the first-order necessary optimality conditions of the problem.



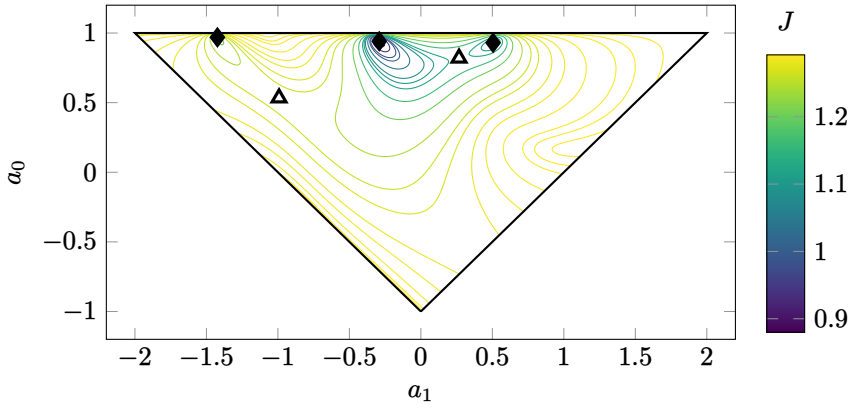
(a) Convex (b) Non-convex

Figure 1.1: Convexity versus non-convexity. In convex problems (a), there is only one critical point, corresponding to the global minimizer. In non-convex problems (b), multiple critical points arise; first-order optimality conditions alone do not identify the global minimizer.

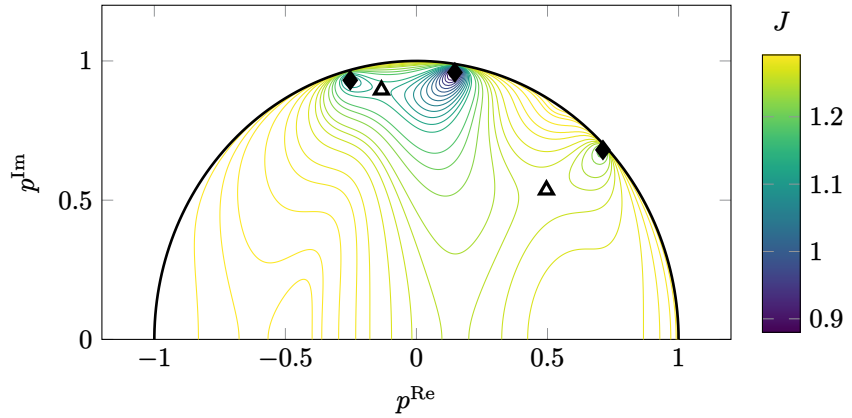
System identification techniques

Today's two most used identification techniques are prediction-error methods (PEMs) [120] and subspace identification methods [171]. From the mid-1970s, PEMs completely dominated the identification field, catalyzed by the book of Ljung [120] and accompanying software packages. Subspace identification emerged around the 1990s with help of multiple contributors (see [171] and references therein) and now both techniques constitute the main approaches for identifying linear dynamical systems from input-output data. Errors-in-variables (EIV) methods constitute a third class of approaches [154], where the noise on the observed data is projected out by so-called instrumental variables.

Despite their effectiveness in practice, these approaches remain suboptimal from a mathematical point of view (cf. Figure 1.3). Prediction-error methods minimize a least squares objective function, with a user-defined parametrization as constraints (e.g., AR, ARMAX), using numerical optimization algorithms. This approach does not guarantee convergence, let alone convergence to a global optimum. Subspace identification methods find a full state space model from the observed input-output data by using linear algebra tools like projectors and the singular value decomposition (SVD). In short, they first estimate the Kalman filter state sequence and afterwards construct the state space model matrices. Both steps are in a certain sense optimal, but the combination has not been proven to be optimal in any sense. Errors-in-variables methods assume noise on both input and output variables. Although several different approaches exist, EIV methods generally fall back to nonlinear optimization routines, resulting in the same suboptimality as encountered for PEMs.



(a) All stable models $a(z) = a_0 + a_1 z + z^2$.



(b) All models $a(z) = (z - p)(z - \overline{p})$ for $\|p\| < 1$ and $p^{\mathrm{Im}} \geq 0$.

Figure 1.2: Contour plots of the objective function value $J = ||H(z) - \widehat{H}(z)||_{\mathcal{H}_2}$ for varying reduced order models $\widehat{H}(z)$, illustrating the (strongly) non-convex nature of the problem. The problem setup corresponds to Example 2.6, in which a given 6th order model $H(z)$, with $\|H(z)\|_{\mathcal{H}_2} = 1.2919$, is approximated by a 2nd order model $\widehat{H}(z) = b(z)/a(z) = \frac{b_0 + b_1 z}{a_0 + a_1 z + z^2}$. For each set of parameters $\{a_0, a_1\}$, the optimal $\{b_0, b_1\}$ are obtained as the unique solution of a linear system. The local minimizers (♦) and saddle points (▵) in the displayed region are highlighted. This example is considered in more detail in Chapter 2.

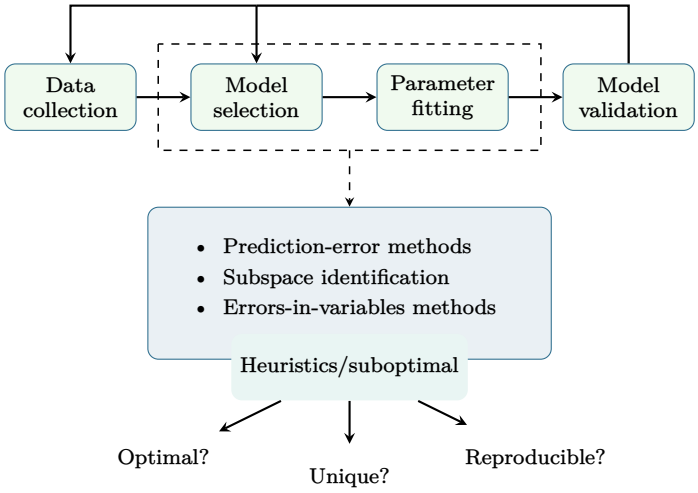


Figure 1.3: The system identification cycle. While state-of-the-art methods used in engineering practice are effective for large-scale problems, they remain heuristic in nature: global optimality cannot be guaranteed, and the resulting solutions often depend strongly on the initialization of the algorithm.

Model reduction techniques

Similarly, in the context of model reduction, the standard engineering practice does not aim for global optimality but instead relies on FONC to obtain locally minimizing solutions. Under the $\mathcal{H}_2$-norm approximation criterion, the first-order optimality conditions have been formulated both through Lyapunov equations [94, 188] and through interpolatory conditions [77, 128, 168]. Accordingly, existing algorithms are generally classified as Lyapunov based (e.g., [156]) or interpolation based methods (e.g., [77]), although the two formulations have been shown to be mathematically equivalent [77, Lemma 3.7]. Another prominent approach is $\mathcal{H}_\infty$-norm reduction, which aims to minimize the worst-case error. However, this problem setup is generally regarded as even more challenging than $\mathcal{H}_2$-norm reduction. In balanced truncation, a reduced model is obtained by truncating some of the states of the given model. Even though this approach provides a priori error bounds on the $\mathcal{H}_\infty$-norm, it is not globally optimal with respect to either the $\mathcal{H}_2$ or $\mathcal{H}_\infty$ criteria, see e.g., [10, 141].

We draw a similar conclusion as before: even though the state of the art is effective in practice, the same fundamental questions regarding global optimality,

dependence on initialization, uniqueness and reproducibility remain unresolved.[3]

The premise of this dissertation is the view that, despite their widespread use and practical effectiveness, the aforementioned iterative and/or heuristic techniques do not "solve" the identification and model reduction problems from a mathematical point of view. Indeed, the performance of these approaches depends heavily on the initialization strategy; stochastic elements compromise reproducibility; and at most local optimality is guaranteed, so that one never knows whether better solutions exist.

By contrast, we deem the identification/reduction task to be mathematically "solved" if and only if the *globally optimal* solutions have been identified by means of a *deterministic* procedure, for instance, by solving a system of linear equations, an eigenvalue problem, or a convex optimization problem.

1.1.3 Achieving global optimality

In this dissertation, we thoroughly study specific model reduction and system identification problems, and exploit the obtained insights in order to develop novel, deterministic methods that provide globally optimal solutions. Our results are inspired by, and build upon, several foundational contributions in the literature, which we briefly review below.

After it was established that globally optimal polynomial optimization is essentially equivalent to solving a large eigenvalue problem (see e.g., [52]), a number of system identification and model reduction problems have been addressed in a globally optimal manner. The strategy of the few existing globally optimal methods is, in broad terms, to compute all solutions that satisfy the FONC—the so-called *critical points*—thereby ensuring that the globally optimal solution(s) can be selected from the resulting set of solutions.

In the context of system identification, a globally optimal approach for the prediction-error method formulation was proposed in [15]; the globally optimal identification of ARMA models was studied in [174]; and the globally optimal least squares realization of autonomous linear time-invariant models was considered in [45]. Each of these approaches exploit the polynomial nature of the identification problem to reduce it to a large-scale eigenvalue problem.

[3] The Hankel-norm approximation problem, which minimizes the error in terms of the Hankel operator, constitutes an exception: Glover's seminal result establishes that this problem admits a closed-form, globally optimal solution via Riccati equations [72].

Preliminary results have also been obtained in the context of model reduction, with a specific focus on single-input single-output (SISO) models and the $\mathcal{H}_2$-norm criterion. In [81], the globally optimal solution for order-one reductions is identified by computing the common roots of a system of quadratic polynomial equations. Along the same lines, [7] proposes a method to determine all optimal interpolation points, which implicitly characterize the full set of critical points. This work generalizes the earlier results of [5], where only first- and second-order reduced models were addressed. Alternatively, [4] exploits the Lyapunov-based optimality conditions to characterize the globally optimal solutions of the SISO $\mathcal{H}_2$-norm model reduction problem.

Despite these promising advances, the main bottleneck of all globally optimal approaches lies in their computational complexity. Even for small sized problem instances, the task of computing all critical points quickly becomes computationally intractable. This makes it important to exploit as much of the specific problem structure as possible in order to obtain more efficient characterizations of the set of (local) minimizers.

Addressing this challenge is one of the main contributions of this dissertation: using problem-specific insights (e.g., Lemma 2.4 and Theorem 3.3) we develop novel characterizations that significantly reduce the resulting computational cost, thereby making it possible to solve (relatively) larger problem instances in a globally optimal and deterministic manner.

1.1.4 Research objectives

The overarching aim of this dissertation is to develop a rigorous mathematical framework for globally optimal methods in system identification and model reduction, and to use this framework to obtain both theoretical insights and practical tools. The largest part of the work focuses on two problems: $\mathcal{H}_2$-norm model reduction in the linear time-invariant (LTI) single-input single-output (SISO) setting and the least squares (LS) realization problem for autonomous single-output LTI models.

Research objective I Our first objective is to develop globally optimal methodologies for the SISO $\mathcal{H}_2$-norm model reduction problem and the LS realization problem for autonomous single-output LTI models. In both settings, we want to exploit the first-order optimality conditions and use problem-specific insights to obtain an efficient (polynomial) characterization of the set of (local) minimizers. These formulations enable a deterministic approach that identifies all local and global minimizers. The obtained complete "mathematical solution"

to the identification and reduction problems (e.g., Theorems 2.5 and 3.6) enables the identification of the globally optimal solution(s) and makes it possible to benchmark heuristic methods used in practice.

Research objective II The second objective is to study the critical point equations over the complex numbers, as opposed to only considering real solutions, since the most important numerical tools used in the context of the first objective, polynomial and MEP solvers, generally operate over $\mathbb{C}$. We aim to develop a rigorous mathematical understanding of the complex solution sets associated with the globally optimal formulations derived earlier. By applying tools from algebraic geometry, we study the geometric structure of these solution sets and quantify their size (e.g., Theorems 5.27 and 6.15). In doing so, we provide a theoretical framework that gives insight into the computational complexity of the globally optimal methodologies.

Research objective III The third objective focuses on the LS realization problem. We want to broaden the scope of the framework for characterizing all locally optimal models derived earlier to more general model classes. We consider the case where an input is added to the single-output models (cf. the SISO setting), where our analysis lays the foundation for the eventual development of globally optimal methodologies. In addition, we investigate multiple-output autonomous models for which the coefficients are allowed to be complex-valued, which extends the applicability of LS realization techniques to areas such as nuclear magnetic resonance spectroscopy.

1.2 Overview of the Dissertation

The text is organized into three main parts, summarized in Figure 1.4. Each part addresses one of the research objectives discussed before. **Part I** develops novel globally optimal methodologies for several problems in model reduction and system identification by reformulating the first-order optimality conditions of the corresponding optimization problems as a multiparameter eigenvalue problem (MEP). **Part II** shifts to an algebro-geometric perspective, analyzing the solution sets of the resulting MEPs over the complex numbers. This leads to new fundamental insights into the underlying mathematical structure and provides exact counts of the solutions. **Part III** is more exploratory in scope, gathering additional contributions on broader model classes and focusing on the characterization and computation of locally optimal models.

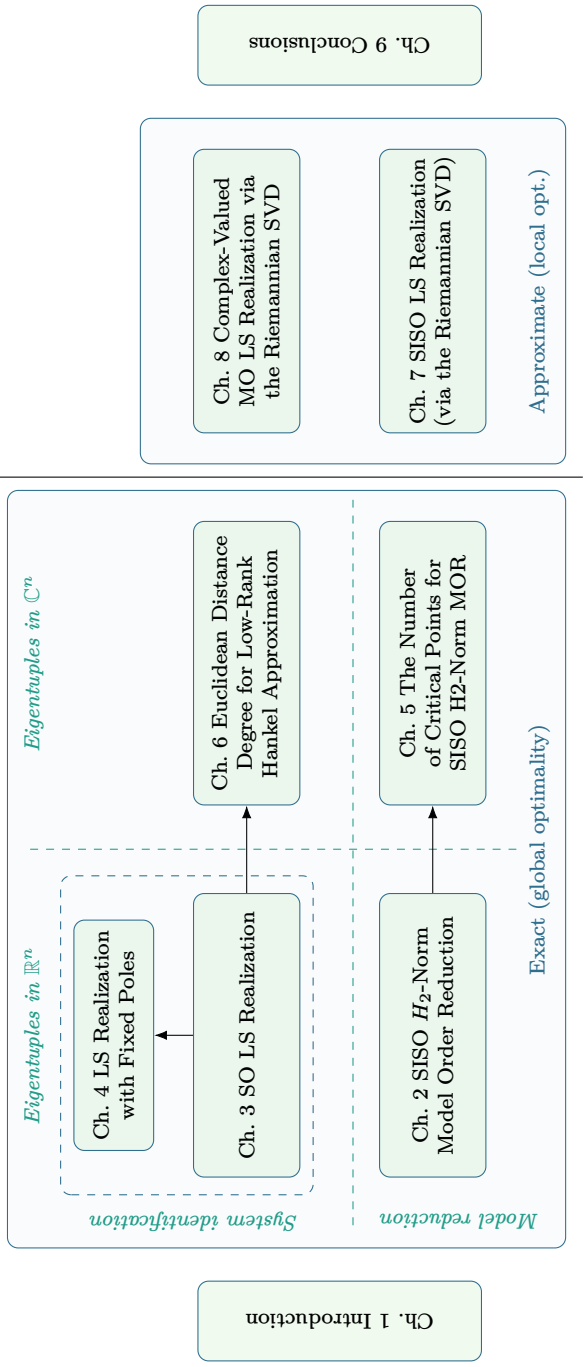
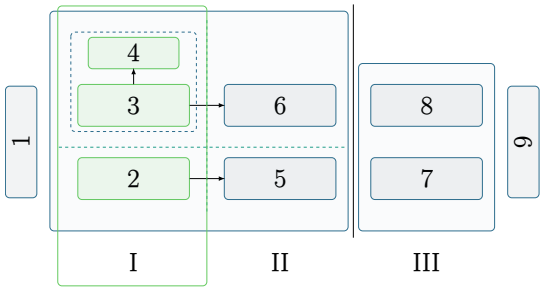


Figure 1.4: Structure of the dissertation. The text is organized into three main parts, framed by an introduction and a conclusion. Part I develops MEP-based, globally optimal solution approaches for problems in model reduction and system identification. Part II shifts the perspective and analyzes the solution sets of these MEPs over the complex numbers. Part III collects additional contributions on system identification, considering more general model classes.

1.2.1 Part I: Novel globally optimal solution approaches



The first part of this dissertation considers the development of *globally optimal* solution methods for problems in model reduction and system identification using linear time-invariant (LTI) systems. These problems are typically formulated as non-convex optimization tasks and are, in engineering practice, most often addressed using local or heuristic algorithms. Such approaches, while effective in practice, provide no guarantees of global optimality and may return different solutions depending on initialization.

The main contribution of Part I is to illustrate how the first-order optimality conditions can be exploited in order to characterize all the critical points of interest as the real-valued common roots of a system of polynomial equations. The obtained polynomial systems are reformulated as a *multiparameter eigenvalue problem (MEP)*, making it possible to deterministically compute all local and global minimizers of the optimization problems. This approach is visualized in Figure 1.5.

Chapter-wise overview of the contributions

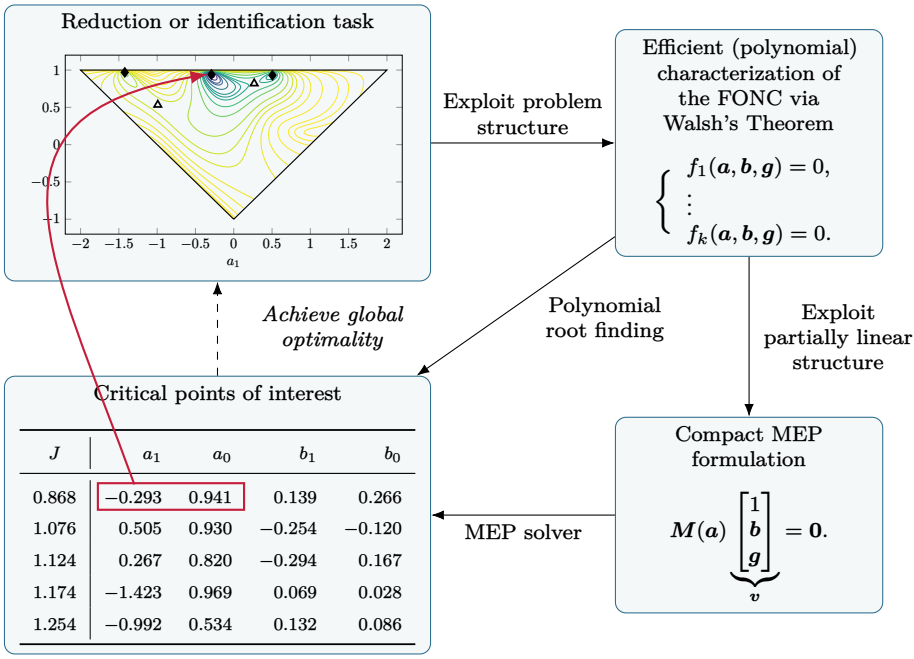
Chapter 2 develops a new methodology for $\mathcal{H}_2$-norm model reduction in the single-input single-output (SISO) setting. By exploiting Walsh's Theorem, which is an elegant formulation of the first-order necessary conditions for optimality, we show that the problem can be reformulated as an MEP, whose real-valued eigentuples describe all local and global minimizers of the model reduction problem. Contrary to the earlier globally optimal approaches available in the literature, the degree of the proposed MEP is independent of the order of the original and reduced systems and the size of its coefficient matrices grows less rapidly, leading to a significant improvement in the involved computational cost. The resulting approach can therefore be applied to (relatively) larger problem sizes, thereby broadening the scope of globally optimal $\mathcal{H}_2$-norm reduction.

In **Chapter 3**, we shift our attention from model reduction to system identification, specifically the least squares (LS) realization problem for autonomous single-output LTI systems. The LS realization problem is fundamental in data-driven modeling: from a finite time series of observed output data, it aims to estimate the autonomous single-output LTI system that generated the data. The problem arises in a broad spectrum of applications, including high-resolution frequency estimation and harmonic retrieval, direction-of-arrival estimation in array processing and shape-from-moments problems. Much like the model reduction problem considered in Chapter 2, the LS realization problem is inherently non-convex and is typically approached in engineering practice through heuristic or iterative algorithms that provide no guarantees of global optimality.

Building on the insights from Chapter 2, we investigate the appearance of a finite-dimensional counterpart of Walsh's Theorem in the context of this LS realization problem. Specifically, the so-called misfit associated with each optimal realization can be interpreted as the result of a double filtering operation through a finite-impulse response filter. Motivated by the analogy with Chapter 2, we exploit this novel characterization of the optimal misfit to reformulate the LS realization problem as an MEP, whose real-valued eigentuples describe all local and global minimizers, thereby establishing a novel, globally optimal solution method. Our approach provides a better understanding of the structure of the LS realization problem and, moreover, proves to be more efficient than the other globally optimal method currently available in the literature.

We also extend the novel globally optimal methodology to the setting of weighted structured low-rank approximation (SLRA) of Hankel matrices, which arises when the samples of the observed data come with differing levels of confidence or when certain samples are required to be fitted with higher accuracy than others. By introducing non-uniform weights into the residual norm to be minimized, we obtain a reformulation as a weighted MEP, of which the unweighted variant studied before appears as a special case. To the best of the author's knowledge, this constitutes the first general globally optimal approach for weighted low-rank Hankel matrix approximation.

Chapter 4 proposes an extension of the LS realization method developed in the previous chapter. It addresses scenarios where some of the poles of the autonomous system are known beforehand. Such prior knowledge is not uncommon in practice: in process control and econometrics, certain modes in the data can often be reliably derived from first principles. Likewise, when the LS realization problem is used for parameter estimation of (real) exponential signals, partial information about the poles may be available from earlier experiments, knowledge of the excitation source, intrinsic physical properties of the setup, or



J	a_1	a_0	b_1	b_0
0.868	−0.293	0.941	0.139	0.266
1.076	0.505	0.930	−0.254	−0.120
1.124	0.267	0.820	−0.294	0.167
1.174	−1.423	0.969	0.069	0.028
1.254	−0.992	0.534	0.132	0.086

Figure 1.5: Schematic overview of the strategy used in Part I of the dissertation for *deterministically* computing *globally optimal* solutions to non-convex optimization problems stemming from model reduction and system identification. By exploiting Walsh's Theorem (or variants thereof), the first-order necessary conditions for optimality are reformulated as a system of multivariate polynomial equations. From this system, the nonlinear part of the problem is isolated by removing the variables appearing only linearly in the equations. This leads to a multiparameter eigenvalue problem (MEP), the eigentuples of which characterize all (local and global) minimizers of the problem. Once these eigentuples have been computed, the globally optimal solution(s) can be selected by comparing the respective objective function values of the obtained points.

even from dynamics introduced to capture noise disturbances. We show how this information can be incorporated into the globally optimal framework by reformulating the problem as yet an alternative MEP.

Significance, scope and target audience

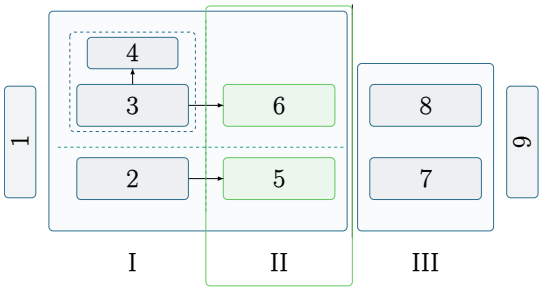
The significance of these contributions is threefold. First, the proposed framework is powerful in that it provides a unified methodology for two problems that are closely related yet fundamentally distinct: the LS realization problem for autonomous LTI models and the SISO $\mathcal{H}_2$ model reduction problem. Second, the resulting formulations represent a substantial advancement over the state of the art, both in terms of improved computational efficiency as theoretical elegance and simplicity. Finally, the contributions are non-trivial: in particular, the role of Walsh's Theorem as an important building block in the development of globally optimal solution methods had not previously been recognized.

The pursuit of global optimality in model reduction and system identification comes at a significant computational cost. The efficiency of the proposed methodologies is governed almost entirely by the difficulty of computing all eigentuples of the resulting MEPs, whose size and degree grow rapidly with the problem size. In essence, an MEP is a system of multivariate polynomial equations for which a subset of the variables appears linearly, and it is therefore similarly difficult to solve[4]. As such, MEP-based approaches are best viewed as theoretical benchmarks rather than as practical tools for large-scale engineering applications.

The primary target audience of Part I consists of engineers who seek insight into the full landscape of local and global minimizers: on the one hand, to validate and benchmark the heuristic or iterative algorithms commonly used in practice, and on the other hand, to inform the design of new solvers. A secondary target audience consists of applied mathematicians working in numerical algebraic geometry or related areas, for whom the presented formulations offer concrete, application-driven instances of polynomial system solving and MEPs. These applications may in turn motivate the development of more efficient numerical algebraic geometry solvers.

[4]The conversion from a polynomial system to an MEP is described in Appendix A.2.2.

1.2.2 Part II: Solution set of the critical equations



The second part of this dissertation revisits the systems of polynomial equations—referred to as the *critical equations*—and the corresponding MEP formulations introduced in Part I, but shifts the perspective from engineering applications to a more rigorous mathematical analysis. In Part I, the attention was given to the real-valued solutions of these critical equations and corresponding MEPs, since only those correspond to meaningful critical points of the model reduction and system identification problems under consideration.

In contrast, Part II considers the entire set of complex solutions, which is motivated by the following two considerations: First, numerical solvers for polynomial systems and MEPs inherently operate over the complex numbers, so the total number of complex solutions is a natural measure of the computational complexity involved in solving these problems; Second, by passing to the algebraically closed field of complex numbers, we can exploit powerful tools from algebraic geometry to study the geometry and structural properties of the solution sets.

The main contribution of Part II is a comprehensive analysis of the complex solution sets arising from the polynomial and MEP formulations derived in Part I. We first establish that, for generic problem instances, the set of complex solutions to the polynomial formulations is finite. Then, by applying the Thom–Porteous formula to the MEP formulation, we obtain an exact count of its complex solutions. This approach is visualized in Figure 1.6.

In particular, we show that, when considered over the complex numbers, the reformulation of the system of polynomial equations as an MEP can give rise to additional, spurious solutions, which we denote by the *bad locus*. Importantly, we prove that in most cases the bad locus is in fact empty, explaining why these spurious solutions did not appear in the numerical experiments of Part I. We also establish that the solutions to the formulations from Part I are generally reduced,

i.e., multiplicity-free, which is highly advantageous from a computational point of view.

Perhaps most significantly, the MEP formulations from Part I allow us to derive an exact count of the complex solutions, under the mild assumption that the bad locus is finite, which fails only in rare, exceptional cases. In the context of the LS realization problem, our count agrees with earlier results in the literature, but the derivation presented here is more accessible and provides deeper insight into why the count holds. For the SISO $\mathcal{H}_2$ model reduction problem, by contrast, our work establishes the first count applicable to all problem sizes.

Chapter-wise overview of the contributions

Chapters 5 and 6 revisit the MEPs introduced in Chapters 2 and 3, respectively, but now analyzed from an algebro-geometric perspective. Throughout both chapters, the *fiber dimension theorem* (see e.g., [152, Theorem 1.25]) and the *Thom–Porteous formula* (cf. [56, Theorem 12.4]) are the most important tools driving the analysis.

While it was known that the critical equations, which characterize all first-order optimal solutions associated with the $\mathcal{H}_2$-norm model reduction problem, admit finitely many complex solutions, no general formula upper-bounding their total number had previously been established. **Chapter 5** addresses this gap by deriving, for the first time, an explicit count of the number of complex solutions to the associated MEP. We begin with a comprehensive algebro-geometric analysis of the solution set of the critical equations, stratify its components, derive its dimension and establish reducedness of the points of interest.

Next, we consider the MEP formulation introduced in Chapter 2 and show that it corresponds to a particular homogenization of the critical equations, which introduces a spurious component: the bad locus $\mathcal{B}$. We prove that this bad locus is empty whenever the higher-order model has simple poles only. Finally, an exact count of the complex eigentuples is obtained by applying the Thom–Porteous formula for the corank one case to the determinantal variety arising from the MEP formulation.

For weighted LS realization, or equivalently, SLRA of Hankel matrices, the number of complex solutions to the critical equations was already (indirectly) accessible through the literature on the Euclidean Distance (ED) degree. The MEP-based analysis presented in **Chapter 6** not only recovers these results but also provides new geometric insights: in particular, it explains why the number of solutions may change under special choices of weights.

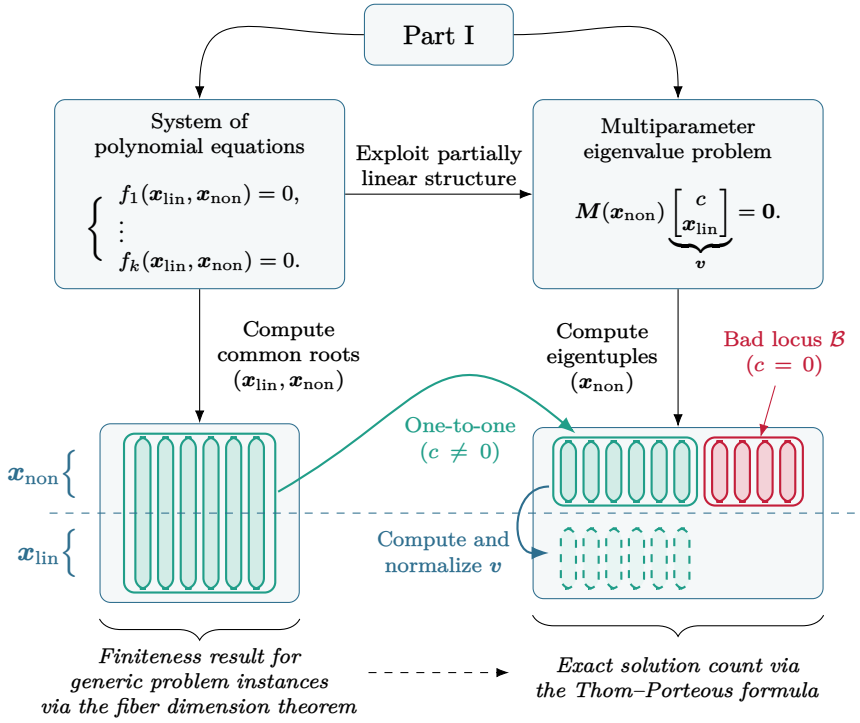


Figure 1.6: Part II of this dissertation analyzes the complex solution set of the systems of polynomial equations and corresponding MEP formulations obtained in Part I. The MEP is obtained by eliminating the variables that appear only linearly ($\boldsymbol{x}_{\text{lin}}$) from the system of equations, resulting in a nonlinear eigenvalue problem in the remaining variables ($\boldsymbol{x}_{\text{non}}$). We illustrate that, over the complex numbers, this reformulation as an MEP may introduce spurious additional components, referred to as the *bad locus* $\mathcal{B}$. In the conversion from the system to the MEP one entry of $\boldsymbol{v}$ is fixed as a nonzero constant c. The bad locus corresponds to eigentuples $\boldsymbol{x}_{\text{non}}$ for which this entry in the corresponding eigenvector $\boldsymbol{v}$ is zero, such that $\boldsymbol{v}$ cannot be appropriately normalized to retrieve the associated values for $\boldsymbol{x}_{\text{lin}}$. We prove that for most problem instances of interest the bad locus is generically empty and formally establish finiteness of solution set of the polynomial systems for generic problem instances, regardless of the problem size, thereby confirming our numerical observations from Part I. By applying the *Thom–Porteous formula*, we then obtain an exact count of the complex eigentuples of the MEP under the assumption that the bad locus is finite, an assumption that fails only in very exceptional cases.

The size of the bad locus is shown to depend on the selected weights, and we demonstrate that, for certain problem sizes and weight configurations, the bad locus can become positive-dimensional. Since the derivation in Chapter 6 relies on basic tools from intersection theory, it offers a more transparent and accessible perspective than the existing treatments of the ED degree available in the literature.

Significance, scope and target audience

The obtained results mark a clear advance over the state of the art: whereas previous results on the complex solution sets were scattered across different communities and relied on diverse techniques, the application of the fiber dimension theorem and the Thom–Porteous formula provides a unified approach to proving finiteness and deriving exact solution counts to the critical points arising from engineering applications. In particular, the notion of the bad locus, which arises naturally from the MEP formulations obtained in Part II, is novel and leads to important new insights.

The formulae for the solution counts obtained in Chapters 5 and 6 are general, in the sense that they are applicable for all problem sizes, and are independent from the specific problem instance (i.e., the observed data or given higher-order model). Because the critical points of the optimization problems form a subset[5] of the complex solutions to the critical equations, these counts provide novel upper bounds on the number of genuine critical points for both the $\mathcal{H}_2$-norm model reduction problem and the weighted LS realization problem.

From a practical perspective, these results allow engineers to assess the fundamental complexity involved in achieving global optimality and possibly, to evaluate the scalability of iterative algorithms commonly used in practice. In addition, the geometric analysis of the solution sets shows that, in most cases, the MEP solutions are reduced (i.e., they occur with multiplicity one). This property is computationally favorable: reduced solutions can be approximated reliably by numerical methods, whereas multiplicities typically give rise to ill-conditioning and increased computational difficulty.

Also note that the results of Part II build directly on the MEP-based reformulation of the critical equations developed in Part I. The MEP description characterizes the solution set as a determinantal variety, which in turn makes it possible to apply the Thom–Porteous formula leading to the solution

[5] Only real-valued (stable) solutions correspond to genuine critical points (e.g., local minima), and in the context of model reduction, they must additionally represent a stable model.

count. Consequently, Parts I-II collectively establish multiparameter eigenvalue formulations as a powerful computational and theoretical framework.

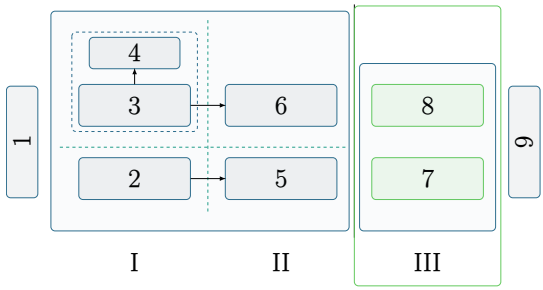
The results in Part II are directed both at engineers and applied mathematicians. Although the analysis relies on tools from algebraic geometry, which can be difficult to access without prior exposure, the presentation emphasizes intuition and applications to make the results as accessible as possible to a broader engineering audience. Examples are used throughout to illustrate important concepts, and Chapter 5 opens with an elementary introduction to the relevant concepts from algebraic geometry (cf. Section 5.3).

Collaborative efforts

Part II is the outcome of a close collaboration with *Leonie Kayser*, a PhD researcher in the Numerical Algebraic Geometry group at the Max Planck Institute (MPI) for Mathematics in the Sciences, Leipzig. This collaboration was initiated during a two-week research visit of the author to the MPI in October 2024, at the invitation of Dr. Telen, S., leader of the Numerical Algebraic Geometry group at the institute.

The primary objective of the visit was to deepen the algebro-geometric understanding of the polynomial systems and MEPs encountered in Part I. While earlier work had already suggested, through extensive numerical experiments, that the obtained MEPs generally admit finitely many complex solutions, a formal proof of finiteness and, ideally, a precise count of the solutions was still lacking. In addition, the broader theoretical consequences of reformulating polynomial systems as MEPs had not yet been completely understood. Since the initial research visit, the author and Leonie have worked together intensively, combining perspectives from systems theory and (numerical) algebraic geometry to establish the theoretical foundations for the polynomial formulations, and corresponding MEPs, studied in Part I.

1.2.3 Part III: Other contributions to system identification



Part III is less tightly connected to the main theme of this dissertation and instead collects additional contributions developed by the author during the course of the PhD.

The main contribution of Part III is to consider extensions of the LS realization problem, which was studied in Chapter 3, to more complex model classes. As in Parts I and II, the analysis is guided by the first-order necessary conditions, however, the results are of a more exploratory nature, focusing on the properties of locally optimal models and on heuristic methods for their computation. Although the pursuit of globally optimal methods is briefly touched upon in the second half of Chapter 7, it is mostly left as an open direction for future work. An overview of the tackled problem formulations is given in Figure 1.7.

Chapter-wise overview of the contributions

Chapter 7 considers the LS realization problem for SISO LTI models. The analysis begins with LS projection onto the behavior of a known model, where observed input-output data are modified in a LS sense so that the corrected data satisfy a prescribed difference equation. We show that this orthogonal projection induces an orthogonal decomposition of the data space. This generalizes earlier results for autonomous systems (cf. Section 3.3.1) to the more general SISO setting and makes the connection to insights from behavioral systems theory. Furthermore, we introduce a novel image representation for the behavior of SISO models.

We then turn to the full LS realization problem, in which the model parameters themselves must be estimated from observed sequences of input-output data. This problem admits a natural interpretation as an SLRA task, motivating the

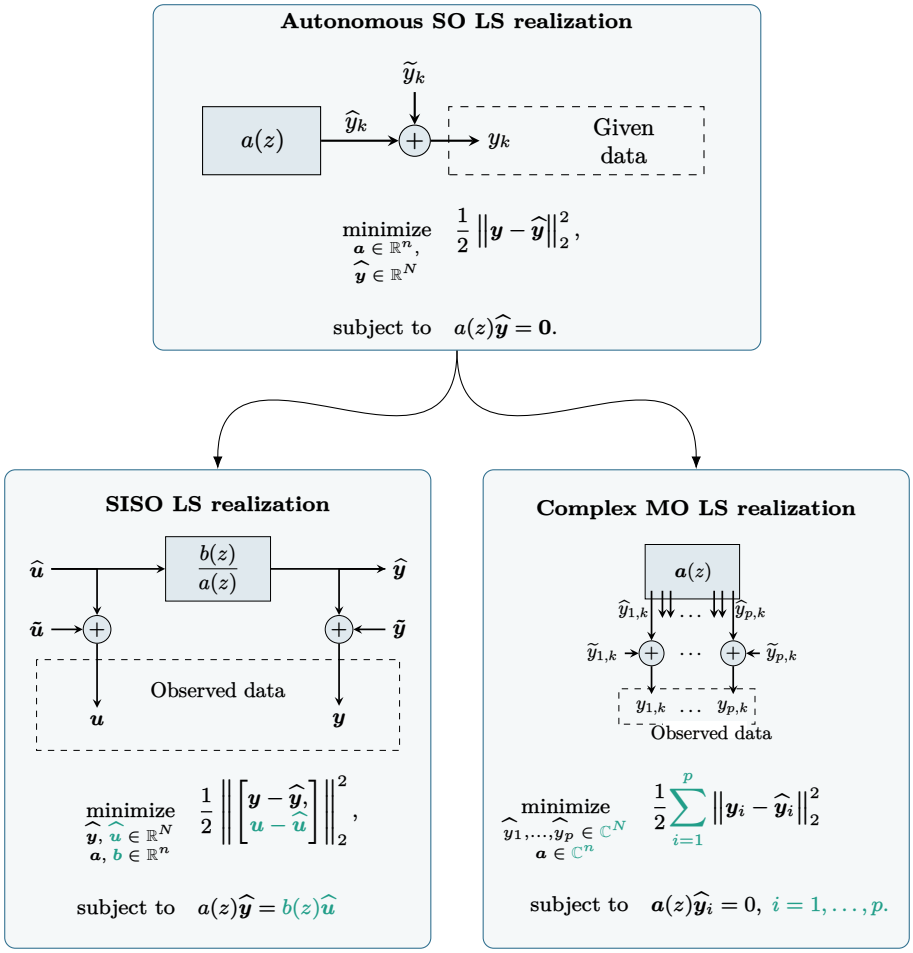


Figure 1.7: Schematic overview of the extensions of the LS realization problem considered in Part III. Building on the autonomous single-output (SO) formulation of Chapter 3, Chapter 7 addresses the general SISO LS realization problem, while Chapter 8 extends the framework to complex-valued, multiple-output (MO) systems. As in Parts I and II, the analysis is guided by first-order necessary conditions for optimality. The focus, however, shifts: whereas the earlier parts pursued global optimality via MEPs, Part III remains more exploratory, characterizing properties of locally optimal models and investigating heuristic approaches for their computation. The figure highlights the fundamental differences between the model structures and the corresponding optimization problems.

use of the Riemannian SVD and associated heuristic methods for computing locally optimal realizations. We revisit the kernel-based Riemannian SVD described in the literature, and propose an alternative version that builds on the novel image representation. Although several promising directions are identified, the development and comprehensive study of a framework similarly to the globally optimal methodologies proposed in Parts I and II remains an open challenge for future research.

Chapter 8 considers the extension of the LS realization problem to complex-valued, autonomous multiple-output LTI systems. This extension from the classical real-valued systems and signals to the complex case broadens the scope of LS realization methods to application domains such as array signal processing, nuclear magnetic resonance spectroscopy, and multi-channel system identification, where complex-valued data and multiple outputs occur. The chapter formulates the problem as a structured total least squares problem and derives a novel, complex-valued Riemannian SVD, characterizing all locally optimal solutions. An existing heuristic algorithm for the real case is adapted to the complex case, and the methodology is validated on illustrative examples drawn from a nuclear magnetic resonance modeling problem and a direction of arrival estimation problem.

Significance, scope and target audience

Although a comprehensive framework for the computation of globally optimal solutions remains a topic for future work, the analysis in Chapter 7 identifies several promising directions and highlights important challenges that need to be addressed. The introduction of a novel image representation, which had not previously been described in the literature, enriches the framework available for describing the behavior of SISO systems. Equally important is the developed unification of the behavioral and state-space-based characterizations with the input-output difference equation framework. Taken together, these contributions advance the understanding of the properties of the SISO LS realization problem and, in particular, of its set of locally optimal solutions.

Chapter 8 moves, beyond real-valued, autonomous single-output models to complex-valued, multiple-output systems. This generalization broadens the applicability of LS realization methods to domains where complex data naturally arise. This constitutes, to the best of our knowledge, the first treatment of LS realization in the complex, multiple-output setting. The chapter derives a novel complex-valued Riemannian SVD and adapts an existing algorithm from the real to the complex setting, complementing the theoretical characterization with a method that is directly applicable in practice. More broadly, the developments

in this chapter provide a template for extending other system identification approaches to the complex numbers. The results will be of particular interest to engineers and applied scientists working with complex-valued and multi-channel data.

1.2.4 General considerations

We outline two general considerations regarding the structure of the chapters in this dissertation and the computational tools employed throughout.

Individual chapters

Each chapter is written to be largely self-contained, with most chapters adapted—sometimes with minor modifications—from prior individual publications. To some extent, this allows the reader to consult the chapters independently, without requiring a linear reading of the dissertation. A detailed introduction and literature overview is provided at the beginning of every chapter. In addition, we systematically position the contents of each chapter within the broader context of this dissertation. Wherever possible, didactic running examples are included to guide the reader through the novel (theoretical) insights, while concluding numerical experiments validate the proposed methodologies in the context of (relatively) larger-scale problems.

Computational tools

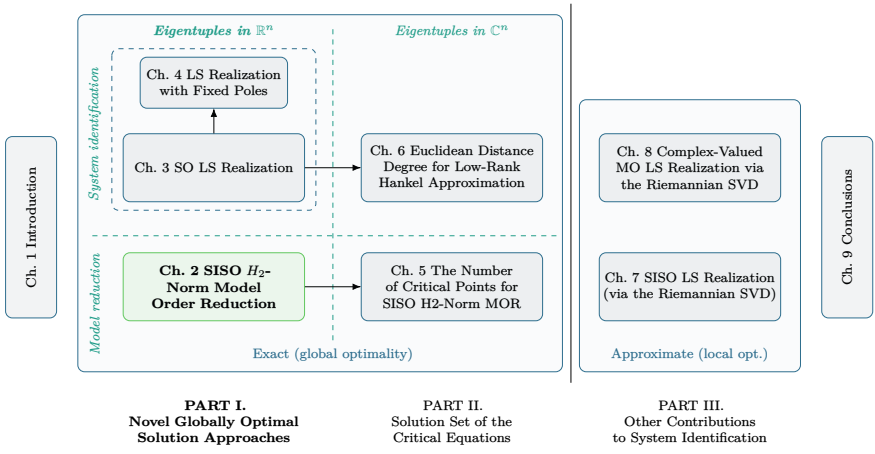
While solving systems of polynomial equations or MEPs is essential to numerically validate our results, the focus of this dissertation lies upstream: on deriving the algebraic formulations and characterizations rather than on the design or analysis of numerical solvers. Our contributions are thus not in the development of specific numerical methods, but in providing problem formulations that make such methods applicable in the first place. For the computational experiments, we use the HomotopyContinuation.jl package [20] to solve systems of polynomial equations, and the block Macaulay matrix framework made available in MacaulayLab [177] to solve MEPs. These tools are chosen for their effectiveness and ease of use, but alternative solvers could equally be applied.

Part I

Novel Globally Optimal Solution Approaches

Chapter 2

Globally Optimal SISO $\mathcal{H}_2$-Norm Model Reduction Using Walsh's Theorem



Chapter abstract

We present a novel methodology for single-input single-output (SISO) $\mathcal{H}_2$-norm model reduction that guarantees global optimality of the obtained solution(s). By exploiting Walsh's Theorem, which is an elegant formulation of the first-order necessary conditions for optimality, we reformulate the model reduction problem as a multiparameter eigenvalue problem (MEP), the real-valued eigentuples of which characterize the globally optimal solution(s) of the model reduction problem. While aiming for global optimality comes at the cost of a combinatorial growth of the problem complexity for increasing model orders, the novel methodology allows us to tackle larger problems compared to the few other globally optimal approaches in the literature. In particular, the degree of the obtained MEP is independent of the order of the original higher order and obtained reduced-order model, a property that is favorable from a computational point of view. We perform three numerical experiments to illustrate the effectiveness of the methodology.

Candidate's contributions and relation to published work

This work has been previously published as:

- **Lagauw, S.**, Agudelo, O. M., and De Moor, B. (2023). "Globally Optimal SISO H2-Norm Model Reduction Using Walsh's Theorem". *IEEE Control Systems Letters* 7, pp. 1670–1675.

The candidate led the theoretical development, implemented the code, conducted the experiments, and authored the text.

This chapter differs slightly from the original article:

- The notation and terminology was adapted to ensure consistency with the conventions used throughout the remainder of this dissertation.
- The formulation of [109, Theorem 3] was slightly adjusted for improved clarity and a minor error in the original statement of [109, Proposition 2] is corrected.
- A compact expression, based on banded Toeplitz matrices, was introduced for the system of equations from [109, Theorem 3].
- Remark 2.7 is new, more information is given on the solution set at infinity and a contour plot is added to Example 2.12.

2.1 Introduction

2.1.1 Problem formulation

We consider the model reduction problem for systems in $\mathcal{M}$, the class of minimal, stable, causal single-input single-output (SISO) linear time-invariant (LTI) continuous-time systems with real-valued impulse response, a problem that already has been studied extensively, see e.g., [10, 141] and references therein. The model reduction problem can be summarized as follows: given a strictly proper transfer function $H(s) \in \mathcal{M}$ of order N,

$$H(s) = \frac{d(s)}{c(s)} = \frac{d_{N-1}s^{N-1} + \cdots + d_1 s + d_0}{c_N s^N + c_{N-1}s^{N-1} + \cdots + c_1 s + c_0}, \tag{2.1}$$

we search for a model $\widehat{H}(s) \in \mathcal{M}$ of order $n < N$,

$$\widehat{H}(s) = \frac{b(s)}{a(s)} = \frac{b_{n-1}s^{n-1} + \cdots + b_1 s + b_0}{a_n s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0}, \tag{2.2}$$

so that $\widehat{H}(s)$ is a 'good approximation' of $H(s)$. Define the coefficient vectors $\boldsymbol{c} \in \mathbb{R}^{N+1}$ and $\boldsymbol{d} \in \mathbb{R}^N$ as $(c_0, \ldots, c_N)^\mathsf{T}$ and $(d_0, \ldots, d_{N-1})^\mathsf{T}$ respectively, and define $\boldsymbol{a} \in \mathbb{R}^{n+1}$ and $\boldsymbol{b} \in \mathbb{R}^n$ similarly using the coefficients of respectively $a(s)$, $b(s)$.

To ensure uniqueness of the parametrization and without loss of generality, we fix $c_N = a_n = 1$ for the remainder of this chapter, so that the models $H(s), \widehat{H}(s)$ are completely determined by $2N$ and $2n$ parameters, respectively. We will therefore denote the vector of unknown parameters as $\boldsymbol{a}_{\text{tr}} = [a_0 \; a_1 \; \cdots \; a_{n-1}]^\mathsf{T} \in \mathbb{R}^n$.

In the model reduction problem, we seek for an approximant $\widehat{H}(s)$ that minimizes the $\mathcal{H}_2$-norm of the approximation error $E(s) = H(s) - \widehat{H}(s)$. The problem under consideration becomes:

$$\widehat{H}(s) \in \underset{\widehat{H}(s) \in \mathcal{M}}{\operatorname{argmin}} \; J^2, \tag{2.3}$$

where

$$\begin{aligned} J^2 = \|E(s)\|_{\mathcal{H}_2}^2 &= \frac{1}{2\pi}\int_{-\infty}^{\infty} |H(j\omega) - \widehat{H}(j\omega)|^2 \, \mathrm{d}\omega \\ &= \int_0^\infty \Big(h(t) - \widehat{h}(t)\Big)^2 \, \mathrm{d}t, \end{aligned} \tag{2.4}$$

with $h(t)$ and $\widehat{h}(t)$ the impulse responses of H, $\widehat{H}$, respectively. Assuming that H and $\widehat{H}$ are stable implies that the roots of $c(s)$ and $a(s)$ lie in the left half of the complex s-plane, guaranteeing that the $\mathcal{H}_2$-norm is well-defined [10].

2.1.2 Literature overview

The optimization problem in (2.3) over the decision variables $\{a_i, b_i\}_{i=1,\ldots,n}$ is generally non-convex, implying that the solution space can contain (many) local optima. This non-convexity makes it very challenging to aim for global optimality. Instead, state-of-the-art techniques 'solve' the problem by iteratively computing a solution that satisfies the first-order necessary conditions for optimality (FONC), i.e., a *critical point* of the optimization problem in (2.3). These FONC have been formulated both in terms of Lyapunov equations [188, 94] and interpolatory conditions [128, 77], dividing the landscape of high-performance algorithms into interpolation-based methods (e.g., [77, 168]) and Lyapunov-based methods (e.g., [156]). Note that both frameworks for the FONC have been shown to be equivalent [77, Lemma 3.7].

Although the above-mentioned iterative techniques are very useful in practice, we consider them suboptimal from a mathematical point of view: their performance depends heavily on the heuristic choice of a 'good' initial point and at most local optimality with respect to (2.3) can be guaranteed. By contrast, we deem the model reduction problem mathematically 'solved' if and only if the globally optimal reduced model(s) have been identified up to machine precision by means of a deterministic procedure, e.g., by solving a system of linear equations or an eigenvalue problem.

The approach of the few existing globally optimal methods [5, 81, 7, 4] is, roughly speaking, to compute all reduced models that satisfy the FONC of (2.3), so that this set of critical points is guaranteed to contain the globally optimal reduced model(s). Inspired by iterative rational Krylov algorithms (e.g., [77]), the author of [7] shows that all optimal sets of interpolation points, which implicitly define the critical points $\widehat{H}(s)$, can be retrieved via a multiparameter eigenvalue problem (MEP) [35]. In particular, each eigentuple of the MEP is used to construct a univariate polynomial, the roots of which correspond to a set of interpolation points. The coefficient matrices of the MEP are composed using a state-space model of the higher order system. The methodology is a generalization of the results in [5], where only first- and second-order reduced models are considered. Alternatively, the authors of [4] exploit the Lyapunov equation to rewrite the objective function (2.4) in terms of a controllability Grammian, where they use the control canonical form to describe the reduced model in terms of the model parameters $\{a_i, b_i\}_{i=0,\ldots,m-1}$. In turn, the authors reformulate the FONC of this obtained objective function as an MEP, the eigentuples/-vectors of which contain the optimal model parameters. For the case of order-one reductions, the authors of [81] retrieve the globally optimal solution(s) via the common roots of a system of quadratic polynomial equations.

Aiming for global optimality comes at the cost of an (enormous) increase in computational burden. The efficiency of the globally optimal methods [7, 4] is completely determined by the difficulty involved with the computation of all the eigentuples of the obtained MEP. So, it is clear that the viability of these approaches depends on the degree and the size of the to-be-solved MEPs.

2.1.3 Contributions and outline

In this chapter, we propose a novel methodology for globally optimal SISO $\mathcal{H}_2$-norm model reduction (Section 2.3). The methodology combines properties of the two globally optimal approaches in the literature: we use the interpolatory framework of the FONC (Walsh's Theorem) to reformulate the problem as an MEP, similarly as in [7], and the optimal model parameters can be retrieved directly from the eigentuples and eigenvectors of the obtained MEP, similarly as in [4]. However, contrary to the existing approaches [7, 4], the proposed methodology solely uses the frequency-domain representation (transfer function) of the models, eliminating the use of state-space descriptions. Furthermore, we argue that the MEPs obtained with the proposed methodology have favorable properties from a computational point of view, especially noticeable for growing model orders. Lastly, we provide a concise summary of the literature related to globally optimal $\mathcal{H}_2$ model reduction in Section 2.2.

Remark 2.1. *Note that we do not insinuate that the above-mentioned 'approximate' methods [77, 168, 156] are not useful in practice. The globally optimal approaches should not be considered as meant to be competitive with respect to these established techniques in the context of large-scale models. Rather, we address a theoretical problem, the solution of which could potentially be used for benchmarking purposes.*

2.2 Characterizing First-Order Optimality

In this Section, we briefly describe the interpolatory formulation of the FONC of the model reduction problem in (2.3). In the context of systems theory and control, these optimality conditions date back to the 1940s, where they arose from a network-synthesis problem considering real-valued and simple poles [6]. Later, the framework has been generalized to multiple poles that are allowed to be complex-valued [128]. Model reduction for linear systems interacts in many ways with the mathematical discipline of approximation on the complex domain, a connection investigated in [23]. In particular, the model reduction problem in (2.3) and its equivalent for discrete-time systems can be considered

as rational l_2-approximation problems, see e.g., [141]. So, it comes as no surprise that the interpolatory conditions derived in [6, 128] were already known in the field of classical analysis in the 1920s, e.g., by Walsh [184].

2.2.1 Interpolatory conditions for optimality

Theorem 2.2 (Meier and Luenberger [128])**.** *Given a stable SISO model $H(s) \in \mathcal{M}$ of order N, let $\widehat{H}(s)$ be a reduced model of order n ($n < N$). The model $\widehat{H}(s)$ is a critical point of the $\mathcal{H}_2$-norm model reduction problem in* (2.3) *if and only if for all poles p_i of $\widehat{H}(s)$ the following interpolatory conditions hold:*

$$H(-p_i)^{(j)} = \widehat{H}(-p_i)^{(j)}, \quad j = 0, \ldots, d_i, \tag{2.5}$$

where d_i is the order of the pole p_i and the superscript j denotes the jth derivative with respect to s, i.e.,

$$F(\xi)^{(j)} = \left.\frac{\mathrm{d}^j F(s)}{\mathrm{d}s^j}\right|_{s=\xi},$$

for any function $F(s)$ and $\xi \in \mathbb{C}$.

We give a high-level view on how the conditions in (2.5) follow naturally from the model reduction problem (2.3), assuming, for the ease of notation, that $\widehat{H}(s)$ has simple poles only. By expressing $\widehat{H}(s)$ in a partial fraction expansion:

$$\widehat{H}(s) = \frac{r_1}{s - p_1} + \frac{r_2}{s - p_2} + \cdots + \frac{r_n}{s - p_n}, \tag{2.6}$$

the model reduction problem (2.3) becomes an optimization problem over the $2n$ decision variables r_i and p_i. The FONC could now be constructed by equating all partial derivatives of J^2 with respect to the real and imaginary parts of these decision variables to zero. However, Wirtinger derivatives[1] can be used to deal with the complex decision variables, resulting in an equivalent but more elegant formulation of the FONC. The Wirtinger derivative of J^2 with respect to $\overline{r_i}$, where $\widehat{H}(s)$ is parametrized as in (2.6) and $\overline{r_i}$ denotes the complex conjugate of r_i, gives:

$$\begin{aligned}\frac{\partial J^2}{\partial \overline{r_i}} &= \frac{1}{\pi j} \int_{-j\infty}^{j\infty} \left(H(s) - \widehat{H}(s)\right) \frac{\partial}{\partial \overline{r_i}} \overline{\left(H(s) - \widehat{H}(s)\right)}\, \mathrm{d}s \\ &= \frac{-1}{\pi j} \int_{-j\infty}^{j\infty} \left(H(s) - \widehat{H}(s)\right) \left(\frac{1}{\overline{s} - \overline{p_i}}\right) \mathrm{d}s, \end{aligned} \tag{2.7}$$

[1] The Wirtinger derivative of $f(s, \overline{s})$ wrt. s (resp. $\overline{s}$) is calculated using the standard rules of differentiation, while considering the variable $\overline{s}$ (resp. s) as a constant, i.e., s and $\overline{s}$ are treated as two independent variables for the purpose of differentiation. This is also called $\mathbb{CR}$-*calculus* [104].

and similarly:

$$\frac{\partial J^2}{\partial \overline{p_i}} = \frac{-1}{\pi j} \int_{-j\infty}^{j\infty} \Big(H(s) - \widehat{H}(s)\Big) \left(\frac{\overline{r_i}}{(\overline{s} - \overline{p_i})^2}\right) \mathrm{d}s. \tag{2.8}$$

Because the objective function J^2 is real-valued, we know that the partial derivatives with respect to r_i and p_i are equal to the complex conjugates of (2.7) and (2.8), respectively:

$$\overline{\left(\frac{\partial J^2}{\partial r_i}\right)} = \frac{\partial J^2}{\partial \overline{r_i}} \quad \text{and} \quad \overline{\left(\frac{\partial J^2}{\partial p_i}\right)} = \frac{\partial J^2}{\partial \overline{p_i}}.$$

It therefore suffices to equate the expressions in (2.7)–(2.8) to zero for all $i = 1, \ldots, n$, to compose the FONC. Using properties of the Laplace transform, these $2n$ equations can be shown to be equivalent to the interpolatory conditions described in Theorem 2.2. See, e.g., [128], for more details.

2.2.2 Walsh's Theorem

The conditions in Theorem 2.2 can be expressed compactly using the polynomial $a(-s)$, which has the mirror images of the poles p_i of the critical point $\widehat{H}(s)$ as its roots.

Theorem 2.3 (Regalia [141])**.** *Given a stable SISO model $H(s) \in \mathcal{M}$ of order N, let $\widehat{H}(s)$ of order n with $n < N$ be a critical point of the model reduction problem* (2.3)*. Then for all $s \in \mathbb{C}$:*

$$H(s) - \widehat{H}(s) = \frac{d(s)}{c(s)} - \frac{b(s)}{a(s)} = \left[\frac{a(-s)}{a(s)}\right]^2 G(s), \tag{2.9}$$

with $G(s)$ the Laplace-transform of some real-valued, stable and causal signal, and where $c(s), d(s), a(s)$ and $b(s)$ are defined as in (2.1)–(2.2)*.*

The denominator of the right-hand side of (2.9) contains the factor $[a(s)]^2$ to illustrate that the residual contains an 'all-pass' filter defined by the poles of $\widehat{H}(s)$, hinting to the connection with the operator-theoretic result of Beurling–Lax[2]. Because the roots of $a(s)$ lie in the complex left half-plane, the factor could equally well be 'absorbed' into $G(s)$.

[2]The transfer function of systems in $\mathcal{M}$ can be considered as an element of the Hardy space $\mathcal{H}_2$ of functions analytic on the complex right half-plane (for discrete-time systems this becomes the Hardy space of functions analytic on the exterior of the unit-disc). Beurling–Lax's theorem describes how an *inner function* (in this case the all-pass filter) induces an orthogonal decomposition of this space into two shift-invariant subspaces [141].

The formulation in (2.9) has been encountered in [117] and the equivalent relation phrased for discrete-time models, which is similar to the continuous-time case upon changing the setting from the Laplace- to the z-domain, has been derived as a limiting case of an iterative model reduction algorithm [47], based on polynomial algebra [64] and using the result of Beurling–Lax [141]. The author of [141] labeled the result as *Walsh's Theorem*, recognizing its origins within approximation theory. The relation has been extended to multiple-input multiple-output (MIMO) models in [103] and [24], in the continuous- and discrete-time setting respectively.

2.3 Methodology

In this Section, we show how Theorem 2.3 can be exploited to retrieve all nth order critical points $\widehat{H}(s)$ of the model reduction problem in (2.3) via the eigentuples of an MEP.

2.3.1 System of multivariate polynomial equations

Lemma 2.4. *For any given stable SISO model $H(s)$ of order N and nth order stable approximant $\widehat{H}(s)$ with $n < N$ as defined in (2.1)–(2.2), define the polynomial,*

$$f(s) = d(s)a(s) - c(s)b(s) - [a(-s)]^2 g(s), \tag{2.10}$$

where $g(s)$ is a degree $N-n-1$ polynomial, parametrized in the coefficients $\boldsymbol{g} = (g_0, \ldots, g_{N-n-1})^{\mathsf{T}} \in \mathbb{R}^{N-n}$:

$$g(s) = g_0 + g_1 s + \cdots + g_{N-n-1} s^{N-n-1}.$$

Then, $\widehat{H}(s)$ is a critical point of (2.3) if and only if,

$$\exists\, \boldsymbol{g} \quad s.t. \quad f(s) = 0,\ \forall s \in \mathbb{C}. \tag{2.11}$$

Proof. Define $g(s) = G(s)[c(s)/a(s)]$ so that we can rewrite (2.9) from Theorem 2.3 as,

$$d(s)a(s) - c(s)b(s) = [a(-s)]^2 g(s). \tag{2.12}$$

The left-hand side of (2.12) is a polynomial in s of degree at most $N+n-1$, so that $g(s)$ must be a polynomial of degree $N-n-1$. It follows that, given a stable SISO model $H(s)$ of order N and a critical point $\widehat{H}(s)$ of order n of the model reduction problem (2.3), there exists a polynomial $g(s)$ such that (2.11) holds. This proves the $\Rightarrow$ direction. Conversely, if, for a given approximant $\widehat{H}(s)$

of order n, there exists a polynomial $g(s)$ such that (2.11) is true, then (2.9) holds as well. The latter implies that $\widehat{H}(s)$ satisfies the interpolatory conditions described in Theorem 2.2, indicating that $\widehat{H}(s)$ is a critical point of the model reduction problem in (2.3). □

Theorem 2.5. *Let f_k be the coefficient corresponding to s^k in the polynomial from* (2.10) *and consider the parameter vectors $\boldsymbol{a}_{\text{tr}}, \boldsymbol{b}$ containing the coefficients of respectively $a(s)$ and $b(s)$ as in* (2.2)*. Define the algebraic variety,*

$$\mathcal{V}_{\mathbb{R}} = \big\{(\boldsymbol{a}_{\text{tr}}, \boldsymbol{b}, \boldsymbol{g}) \in \mathbb{R}^{N+n} \ : \ f_k(\boldsymbol{a}_{\text{tr}}, \boldsymbol{b}, \boldsymbol{g}) = 0, \quad \forall k = 0, \ldots, N+n-1\big\}, \tag{2.13}$$

and consider the semialgebraic set $\mathcal{V}'_{\mathbb{R}} \subseteq \mathcal{V}_{\mathbb{R}}$ that contains the tuples $(\boldsymbol{a}_{\text{tr}}, \boldsymbol{b}, \boldsymbol{g}) \in \mathcal{V}_{\mathbb{R}}$ for which all roots of $a(s)$ lie in the complex left half-plane. Then, in the context of the model reduction problem (2.3)*, we know by Lemma 2.4 that the tuples in $\mathcal{V}'_{\mathbb{R}}$ describe all critical points $\widehat{H}(s)$ of order n.*

Critical points for which $\widehat{H}(s)$ has one or more pole-zero cancellations–the order of $\widehat{H}(s)$ is strictly smaller than n–are generally not described by the relation in (2.9). As such, they cannot be retrieved via $\mathcal{V}'_{\mathbb{R}}$. However, these critical points have been shown to correspond to saddle points of the $\mathcal{H}_2$-norm model reduction problem [141, Theorem 5.3] and, by consequence, they are of no interest to us in our search for the globally optimal minimizer of (2.3).

Consider the coefficient vectors $\boldsymbol{a}, \boldsymbol{b}$ corresponding to $a(s)$ and $b(s)$ as in (2.2). Let $\boldsymbol{T}_N^{\boldsymbol{a}} \in \mathbb{R}^{N \times (N+n)}$ be the banded Toeplitz matrix[3]generated by the coefficient vector $\boldsymbol{a}$:

$$\boldsymbol{T}_N^{\boldsymbol{a}} = \begin{bmatrix} a_0 & \cdots & \cdots & a_{n-1} & 1 & 0 & \cdots & 0 \\ 0 & a_0 & \cdots & \cdots & a_{n-1} & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_0 & \cdots & \cdots & a_{n-1} & 1 \end{bmatrix},$$

and define the banded Toeplitz matrix $\boldsymbol{T}_n^{\boldsymbol{c}} \in \mathbb{R}^{n \times (N+n)}$, which is constructed from the known coefficient vector $\boldsymbol{c}$, similarly. Denote the vector of coefficients corresponding to the polynomial $a(-s)$ by $\widetilde{\boldsymbol{a}} = [a_0, -a_1, a_2, \ldots, (-1)^{n-1}a_{n-1}, (-1)^n]^\mathsf{T} \in \mathbb{R}^{n+1}$, such that $\boldsymbol{T}_N^{\widetilde{\boldsymbol{a}}} \in \mathbb{R}^{N \times (N+n)}$

[3]We use $\boldsymbol{T}_k^{\boldsymbol{x}}$ to denote the banded Toeplitz matrix constructed from k forward shifts of the coefficient vector $\boldsymbol{x}$.

corresponds to:

$$\boldsymbol{T}_N^{\widetilde{\boldsymbol{a}}} = \begin{bmatrix} a_0 & -a_1 & a_2 & \cdots & (-1)^n & 0 & \cdots & 0 \\ 0 & a_0 & -a_1 & a_2 & \cdots & (-1)^n & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \\ 0 & \cdots & 0 & a_0 & -a_1 & a_2 & \cdots & (-1)^n \end{bmatrix}. \tag{2.14}$$

We now exploit the correspondence between polynomial multiplication and the convolution of coefficient vectors, noting that such convolutions can be represented as matrix-vector products using banded Toeplitz matrices. The coefficients $\boldsymbol{f} = [f_0, f_1, \ldots, f_{N-n-1}]^\mathsf{T} \in \mathbb{R}^{N-n}$ of the polynomial from (2.10) can be expressed as follows:

$$\boldsymbol{f} = (\boldsymbol{T}_N^{\boldsymbol{a}})^\mathsf{T}\boldsymbol{d} - (\boldsymbol{T}_n^{\boldsymbol{c}})^\mathsf{T}\boldsymbol{b} - (\boldsymbol{T}_N^{\widetilde{\boldsymbol{a}}})^\mathsf{T}(\boldsymbol{T}_{N-n}^{\widetilde{\boldsymbol{a}}})^\mathsf{T}\boldsymbol{g}. \tag{2.15}$$

Example 2.6. *Consider the third order model ($N = 3$) used in [4]:*

$$H(s) = \frac{s^2 + 9s - 10}{s^3 + 12s^2 + 49s + 78},$$

such that for this example $\boldsymbol{c} = [78, 49, 12, 1]^\mathsf{T}$ and $\boldsymbol{d} = [-10, 9, 1]^\mathsf{T}$. We search the optimal first-order ($n = 1$) approximation. The polynomial $f(s)$ from (2.10) *is equal to,*

$$\begin{aligned} f(s) = \underbrace{(1 - g_1 - b_0)}_{f_3} s^3 + \underbrace{(a_0 - 12b_0 - g_0 + 2a_0g_1 + 9)}_{f_2} s^2 \\ + \underbrace{(9a_0 - 49b_0 + 2a_0g_0 - {a_0}^2g_1 - 10)}_{f_1} s \\ + \underbrace{(-g_0a_0^2 - 10a_0 - 78b_0)}_{f_0}. \end{aligned}$$

The expression in (2.15) *elegantly expresses the coefficients $\boldsymbol{f} = [f_0, \ldots, f_3]^\mathsf{T}$ as a function of the model parameters $\{a_0, b_0\}$ and auxiliary variables $\{g_0, g_1\}$:*

$$\begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \end{bmatrix} = \begin{bmatrix} a_0 & 0 & 0 \\ 1 & a_0 & 0 \\ 0 & 1 & a_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -10 \\ 9 \\ 1 \end{bmatrix} - \begin{bmatrix} 78 \\ 49 \\ 12 \\ 1 \end{bmatrix} b_0 - \begin{bmatrix} a_0 & 0 & 0 \\ -1 & a_0 & 0 \\ 0 & -1 & a_0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} a_0 & 0 \\ -1 & a_0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} g_0 \\ g_1 \end{bmatrix}.$$

Consequently, the variety $\mathcal{V}_\mathbb{R}$ defined in (2.13) *corresponds to the set of real-valued common roots of the following square system of polynomial equations in*

the model parameters $\{a_0, b_0\}$ *and auxiliary variables* $\{g_0, g_1\}$,

$$\begin{aligned} 0 &= f_0 = -g_0 {a_0}^2 - 10a_0 - 78b_0, \\ 0 &= f_1 = 9a_0 - 49b_0 + 2a_0 g_0 - {a_0}^2 g_1 - 10, \\ 0 &= f_2 = a_0 - 12b_0 - g_0 + 2a_0 g_1 + 9, \\ 0 &= f_3 = 1 - g_1 - b_0. \end{aligned} \tag{2.16}$$

The semialgebraic set $\mathcal{V}'_{\mathbb{R}}$ *consists of the common roots for which* $a_0 > 0$.

Remark 2.7 (Optimal zeros for a given set of poles)**.** *It is well known that the nonlinearity inherent to the* $\mathcal{H}_2$*-norm model reduction problem is contained in the search for the optimal poles of the reduced model. Once the denominator* $a(s)$ *is fixed, the optimal numerator* $b(s)$ *can be found via an orthogonal projection, where orthogonality has to be understood wrt. the inner-product defined in the Hardy space* $\mathcal{H}_{22}$ *[77].*

By repeating the derivation in Sections 2.2 and 2.3.1 under the assumption that the parameters $\boldsymbol{a}$ *are fixed, one observes that the optimal parameters* $\boldsymbol{b}$ *arise from the solution of a linear system. Thus, the problem becomes indeed linear in* $\boldsymbol{b}$ *when* $\boldsymbol{a}$ *is prescribed.*

More precisely, if the poles of the reduced model are fixed beforehand, the optimality conditions involve only the partial derivatives in (2.7)*, which leads to the following characterization of the optimal models (alternative to* (2.10)*)*

$$d(s)a(s) - c(s)b(s) = a(-s)g(s), \tag{2.17}$$

where $g(s)$ *is a polynomial of degree at most* $(N-1)$*. In this case also the polynomial* $a(s)$ *is known, such that* $b(s)$ *and* $g(s)$ *satisfying* (2.17) *can be obtained as the unique solution of a system of* $N+n$ *linear equations:*

$$\begin{bmatrix} (\boldsymbol{T}_n^{\boldsymbol{c}})^{\mathsf{T}} & (\widetilde{\boldsymbol{T}_N^{\boldsymbol{a}}})^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{g} \end{bmatrix} = (\boldsymbol{T}_{n+1}^{\boldsymbol{d}})^{\mathsf{T}} \boldsymbol{a}.$$

2.3.2 Quadratic m-parameter eigenvalue problem

Closer observation of the system of equations in (2.16) shows that only the parameter a_0 appears nonlinearly, justifiably indicating that the nonlinearity of the model reduction problem is solely caused by the search for the optimal poles, i.e., the optimal values for the parameters in $\boldsymbol{a}_{\mathrm{tr}}$. We can make this (partially) linear structure of the problem explicit by 'extracting' the variables that appear

only linearly (the $\{b_i\}_{i=1,\dots,n}$ and $\{g_i\}_{i=0,\dots,N-n-1}$) in the system of polynomial equations from Theorem 2.5, and organize them into an eigenvector, giving a quadratic n-parameter eigenvalue problem (MEP):

$$\boldsymbol{M}(\boldsymbol{a}_{\text{tr}})\boldsymbol{v} = \left[(\boldsymbol{T}_N^{\boldsymbol{a}})^{\mathsf{T}}\boldsymbol{d}, \quad (\boldsymbol{T}_n^{\boldsymbol{c}})^{\mathsf{T}}, \quad (\widetilde{\boldsymbol{T}_N^{\boldsymbol{a}}})^{\mathsf{T}}(\widetilde{\boldsymbol{T}_{N-n}^{\boldsymbol{a}}})^{\mathsf{T}}\right]\boldsymbol{v} = \boldsymbol{0}, \tag{2.18}$$

where $\boldsymbol{M}(\boldsymbol{a}_{\text{tr}}) = \sum_{\{\boldsymbol{\alpha}\}} \boldsymbol{M}_{\boldsymbol{\alpha}} {\boldsymbol{a}_{\text{tr}}}^{\boldsymbol{\alpha}} \in \mathbb{R}^{(N+n)\times(N+1)}[\boldsymbol{a}_{\text{tr}}]$ is a matrix polynomial[4] in the monomials ${\boldsymbol{a}_{\text{tr}}}^{\boldsymbol{\alpha}} = a_0^{\alpha_0}\dots a_{n-1}^{\alpha_{n-1}}$ and $\|\boldsymbol{v}\| \neq 0$. Indeed, multiplication of the matrix polynomial from (2.18) from the right with the vector $\boldsymbol{v} = [-1, \boldsymbol{b}, \boldsymbol{g}]^{\mathsf{T}}$, writing out the matrix vector product and changing the sign of the obtained equations leads to the right-hand side of (2.15). The affine values of the variables $\boldsymbol{a}_{\text{tr}}$ for which the matrix in (2.18) drops rank are the so-called *affine eigentuples* of the MEP, i.e., the algebraic variety:

$$\mathcal{W}_{\mathbb{R}} = \left\{\boldsymbol{a}_{\text{tr}} \in \mathbb{R}^n \ : \ \operatorname{rank}\boldsymbol{M}(\boldsymbol{a}_{\text{tr}}) \leq N\right\}.$$

Let $\mathcal{W}'_{\mathbb{R}} \subseteq \mathcal{W}_{\mathbb{R}}$ be the semialgebraic set that contains the tuples $\boldsymbol{a}_{\text{tr}} \in \mathcal{W}_{\mathbb{R}}$ for which all roots of $a(s)$ lie in the complex left half-plane.

Proposition 2.8. *Given the square system of $N+n$ equations $f_0 = \dots = f_{N+n-1} = 0$ and the variety $\mathcal{V}_{\mathbb{R}}$, both defined as in Theorem 2.5, consider the corresponding quadratic n-parameter eigenvalue problem as defined in* (2.18) *and its set of real-valued eigentuples $\mathcal{W}_{\mathbb{R}}$. Then, for generic $H(s)$, $\mathcal{W}_{\mathbb{R}}$ is equal to the projection[5] of $\mathcal{V}_{\mathbb{R}}$ onto the variables $\boldsymbol{a}_{\text{tr}}$.*

Remark 2.9. *The original statement in [109] claimed that $\mathcal{W}_{\mathbb{R}}$ is equal to the projection of $\mathcal{V}_{\mathbb{R}}$ onto the variables $\boldsymbol{a}_{\text{tr}}$, for all models $H(s)$. We will formally show in Chapter 5 that this is true for generic $H(s)$, but that there exist special cases for which $\mathcal{W}_{\mathbb{R}}$ contains additional points.*

Example 2.10 (Example 2.6 continued)**.** *Extracting the variables that appear only linearly out of the system in* (2.16) *leads to the polynomial matrix $\boldsymbol{M}(a_0)$:*

$$\underbrace{\begin{bmatrix} -10a_0 & 78 & a_0^2 & 0 \\ -10+9a_0 & 49 & -2a_0 & a_0^2 \\ 9+a_0 & 12 & 1 & -2a_0 \\ 1 & 1 & 0 & 1 \end{bmatrix}}_{\boldsymbol{M}(a_0)} \begin{bmatrix} -1 \\ b_0 \\ g_0 \\ g_1 \end{bmatrix} = \boldsymbol{0},$$

[4]Each term in the system of equations from Theorem 2.5 is mapped to a specific position in one of the coefficient matrices of the MEP. The term has a monomial in $\boldsymbol{a}_{\text{tr}}$ as a factor that determines in which coefficient matrix the coefficient of the term has to be included. The position within that matrix depends on the equation to which the term belongs (the row) and the presence of a variable from the eigenvector of the MEP in the term (the column).

[5]Here, "projecting onto the variables $\boldsymbol{a}_{\text{tr}}$" simply means disregarding all other variables, thereby retaining only the components associated with $\boldsymbol{a}_{\text{tr}}$.

which can be rewritten as a quadratic polynomial eigenvalue problem (PEP) in the parameter a_0:

$$\left(\underbrace{\begin{bmatrix} 0 & 78 & 0 & 0 \\ -10 & 49 & 0 & 0 \\ 9 & 12 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}}_{\boldsymbol{M}_0} + \underbrace{\begin{bmatrix} -10 & 0 & 0 & 0 \\ 9 & 0 & -2 & 0 \\ 1 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{\boldsymbol{M}_1} a_0 + \underbrace{\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{\boldsymbol{M}_2} a_0^2 \right) \begin{bmatrix} -1 \\ b_0 \\ g_0 \\ g_1 \end{bmatrix} = \mathbf{0}. \quad (2.19)$$

2.3.3 Computing the critical points

From Lemma 2.4, we know that all critical points $\widehat{H}(s)$ of order n can be computed via a polynomial root finding problem, for which advanced solvers exists (e.g., [20]). However, as these root finding techniques generally work over the field of complex numbers, one has to prune away the complex-valued common roots to obtain $\mathcal{V}_{\mathbb{R}}$. A second pruning step is then needed to isolate the critical points $\mathcal{V}'_{\mathbb{R}}$, that is, removing the solutions for which the corresponding model is unstable. Alternatively, we can exploit the partially linear structure of the problem and calculate $\mathcal{V}_{\mathbb{R}}$ via $\mathcal{W}_{\mathbb{R}}$, the real-valued affine eigentuples of the MEP, as described in the aforementioned proposition. By definition, the matrix $\boldsymbol{M}(\boldsymbol{a}_{\text{tr}})$ is rank deficient when evaluated in one of its eigentuples, allowing to calculate for each eigentuple $\boldsymbol{a}_{\text{tr}} \in \mathcal{W}_{\mathbb{R}}$ the corresponding values for the parameters $\boldsymbol{b}$ via the null space of $\boldsymbol{M}(\boldsymbol{a}_{\text{tr}})$. Specialized solvers (e.g., [175, 138]) can be used to compute the eigentuples of the MEP.

There is no standard means to quantify the complexity of a system of polynomial equations before it has been solved, but the size of the system, measured as the number of variables (#vars) and the number of equations (#eqs), and the highest polynomial degree present in the system ($d_{\max}$) serve as good indicators.

Bézout's number n_{B}, equal to the product of the degree of each of the polynomials in a square system of equations, scales with the above-mentioned indicators and could therefore be used as an elementary[6] proxy for its complexity.

[6]There exist more sophisticated bounds that, often based on *mixed-volume* computations, additionally consider the sparsity of the polynomials [119].

Table 2.1: Properties of the systems of polynomial equations underlying the globally optimal techniques described in the literature. These equations determine the computational complexity of the proposed solution methods.

(N, n)	Approach	#eqs	#vars	$d_{\max}$	n_{B}
$(3, 1)$	Agudelo's approach [4]	2	2	8	32
	Alsubaie's approach* [7]	7	7	2	128
	Novel methodology	4	4	3	18
$(6, 2)$	Agudelo's approach [4]	4	4	15	10290
	Alsubaie's approach* [7]	14	14	2	16384
	Novel methodology	8	8	3	1458
$(17, 4)$	Agudelo's approach [4]	8	8	41	$\mathcal{O}(10^{11})$
	Alsubaie's approach* [7]	38	38	2	$\mathcal{O}(10^{11})$
	Novel methodology	21	21	3	$\mathcal{O}(10^{9})$

*for this methodology additional post-processing is required.

We compare the alternative approaches for globally optimal SISO $\mathcal{H}_2$-norm model reduction [7, 4] in Table 2.1, which depicts the properties of the to-be-solved system of polynomial equations[7] for three scenarios of (N, n) that correspond to the examples tackled in Section 2.4.

For growing (N, n) the novel approach comes with the lowest complexity. Contrary to the approach proposed in [4], each polynomial in our system of equations is at most *cubic*, which makes the corresponding MEP *quadratic*, independent of the model orders (N, n). This low degree is advantageous, e.g., in the context of homotopy continuation methods the step size of the continuation procedure typically has to be reduced when operating on systems of higher degree to avoid *path jumping* [165], thereby increasing the overall computation time. While Alsubaie's approach [7] leads to linear MEPs, it results in relatively larger systems, with more equations and variables (see Table 2.1). The increased system size—in particular, the larger number of variables—leads to a higher computational cost compared to the novel approach.

[7] Comparing the complexity of MEPs is even less straightforward. So, to be able to include the methods of [4, 7], which lead to MEPs, we consider the properties of the system of polynomial equations obtained by writing out the matrix-vector product of the to-be-solved MEP.

2.3.4 Discrete-time setting

The results described in Theorem 2.2, Theorem 2.3, Lemma 2.4 and Theorem 2.5 admit a straightforward counterpart in the discrete-time setting. In this case, the forward shift s (the variable of the Laplace transform) has to be substituted by the discrete-time forward shift z (the variable corresponding to the $\mathcal{Z}$-transform). Along similar lines, the factor $[a(-s)]^2$ in (2.10) has to be replaced by $[z^n a(z^{-1})]^2$, adjusting for the change in the definition of the 'mirror images' of the poles (i.e., reflection over the imaginary axis in a continuous-time setting becomes reflection over the unit circle in the discrete-time setting). I.e., the vector $\widetilde{\boldsymbol{a}}$ used in (2.14) has to be replaced by the vector $\boldsymbol{a_r} = \mathrm{rev}(\boldsymbol{a}) = [1, a_{n-1}, \ldots, a_0]^\mathsf{T}$ such that $\boldsymbol{T}_N^{\boldsymbol{a_r}} \in \mathbb{R}^{N \times (N+n)}$ corresponds to:

$$\boldsymbol{T}_N^{\boldsymbol{a_r}} = \begin{bmatrix} 1 & a_{n-1} & \cdots & \cdots & a_0 & 0 & \cdots & 0 \\ 0 & 1 & a_{n-1} & \cdots & \cdots & a_0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & a_{n-1} & \cdots & \cdots & a_0 \end{bmatrix},$$

then we can define the MEP for the discrete-time case as follows. This leads to the following alternative MEP:

$$\boldsymbol{M}(\boldsymbol{a}_{\mathrm{tr}})_{\mathrm{d}} = \left[(\boldsymbol{T}_N^{\boldsymbol{a}})^\mathsf{T} \boldsymbol{d}, \quad (\boldsymbol{T}_n^{\boldsymbol{c}})^\mathsf{T}, \quad (\boldsymbol{T}_N^{\boldsymbol{a_r}})^\mathsf{T} (\boldsymbol{T}_{N-n}^{\boldsymbol{a_r}})^\mathsf{T} \right]. \tag{2.20}$$

A discrete-time analogue of Proposition 2.8 follows naturally.

2.3.5 Secular equations

When $n = 1$, the MEP from (2.18) becomes a PEP with square coefficient matrices $\{\boldsymbol{M}_\alpha\}_{\alpha=0,1,2} \in \mathbb{R}^{N+1 \times N+1}$. In this case, we can alternatively calculate the eigenvalues of the PEP by rooting a univariate polynomial: the square matrix $\boldsymbol{M}(a_0)$ is singular if its determinant is equal to zero. To give an example, the determinant of the PEP in (2.19) gives:

$$\det(\boldsymbol{M}(a_0)) = a_0^5 + 15a_0^4 - 89a_0^3 - 1191a_0^2 - 2596a_0 + 780. \tag{2.21}$$

This reformulation of the system of equations (2.13), eliminating the N 'linear' variables $(\boldsymbol{b}, \boldsymbol{g})$ at the cost of higher polynomial degrees, is possible for arbitrary model orders (N, n). Indeed, all minors[8] of the rectangular matrix $\boldsymbol{M}(\boldsymbol{a}_{\mathrm{tr}})$ from (2.18) should vanish when evaluated in an eigentuple of the MEP. So, by

[8]The minors are the $\binom{N+n}{n-1}$ determinants of the $(N+1)\times(N+1)$ submatrices, obtained by omitting $n-1$ rows, of the polynomial matrix $\boldsymbol{M}(\boldsymbol{a}_{\mathrm{tr}})$.

equating these minors to zero we obtain an overdetermined system of equations in the n variables from $\boldsymbol{a}_{\text{tr}}$, the common roots of which correspond to the eigentuples of the MEP. However, the number of all possible minors grows quickly with (N, n). We observed cases where not all minors are required to be able to retrieve all the eigentuples of the MEP, i.e., the full collection of minors does not form a so-called minimal generating set. However, it is not immediately clear how to determine such a minimal set, and doing so is generally computationally expensive.

2.3.6 Solution set at infinity

The nullity of block Macaulay matrices for growing degree d and the rank structure of the rows in their null space, can be used to analyse both the affine solution set of the MEP in (2.18) as the solution set at infinity.

We say that an eigentuple $\boldsymbol{a}_{\text{tr}}$ lies "at infinity" when at least one of its components a_i tends to infinity. These solutions at infinity arise because the entries of the MEP matrix are not homogeneous in $\boldsymbol{a}_{\text{tr}}$, that is, the entries do not all share the same total degree. In the context of (2.3), however, only the affine solutions are of interest. Within the Macaulay matrix framework, the effect of solutions at infinity can be removed by means of a *column compression*. See, e.g., [35, 175].

Although for all of our numerical experiments the affine solution set of the MEP has been zero-dimensional, the solution set at infinity seems to be $(n-1)$-dimensional: for $n = 2$ the number of solutions at infinity found via the block Macaulay matrix of degree d grows linearly with d (constant increases of 5), for $n = 3$ the growth of the number of solutions at infinity grows linearly with d (constant increases of 7), and so on.

Furthermore, we observed that the system of equations (2.13) has an n-dimensional solution set at infinity. Note that this notion of "infinity", which includes points where $\boldsymbol{a}_{\text{tr}}$, $\boldsymbol{b}$ and/or $\boldsymbol{g}$ tend to infinity, differs from the "infinity" referred to in the context of the MEP. Finally, the set of secular equations of the MEP in (2.18), i.e., the minors of the matrix $\boldsymbol{M}(\boldsymbol{a}_{\text{tr}})$, has no solutions at infinity. This is a direct consequence of their homogeneity.

2.4 Numerical Examples

To validate the proposed methodology we compute[9] the numerical solutions of Example 2.6 and tackle two other examples from the literature. We only consider the affine common roots (eigentuples) of the obtained systems of polynomial equations (respectively the MEPs).

Example 2.11 (Example 2.10 continued)**.** *The univariate polynomial in* (2.21) *has five roots:*

$$\mathcal{V} = \{9.679, -16.619, 0.267, -4.164 \pm 0.903j\},$$

which coincide with the eigenvalues of the PEP in (2.19)*. The nullity of the block Macaulay matrix [35] composed from* (2.19) *stabilizes at* $d = 4$*, allowing to retrieve the eigenvalues via the shift-invariance property of its null space. The system of equations in* (2.16) *has 5 common roots, the values for the variable* a_0 *of which correspond to the above-mentioned eigenvalues of the PEP. Only two solutions* a_0 *remain after pruning for the critical points:* $\mathcal{V}'_{\mathbb{R}} = \{9.679, 0.267\}$*. The corresponding values for* b_0 *are* 1.279 *and* -0.0437*, respectively. We can select the globally optimal minimizer using their respective objective function values:* $J = \{0.278, 0.398\}$*.*

Example 2.12. *Consider the four disk system (discrete-time,* $N = 6$*) described in [47] for which we want to calculate the globally optimal second-order approximant (*$n = 2$*):*

$$H(z) = \frac{0.0448z^5 + 0.2368z^4 + 0.0013z^3 + 0.0211z^2 + 0.2250z + 0.0219}{z^6 - 1.2024z^5 + 2.3675z^4 - 2.0039z^3 + 2.2337z^2 - 1.0420z + 0.8513}.$$

Equating the coefficients with respect to the powers of z *in the discrete-time equivalent of* (2.10) *to zero, gives a square system of 8 polynomial equations. The polynomial matrix* $\boldsymbol{M}(\boldsymbol{a}_{\mathrm{tr}})_{\mathrm{d}}$ *from the MEP formulation in* (2.20) *is equal to:*

$$\left[\begin{array}{c:c:c:c:c:c:c}
0.045 & 0 & -1 & 0 & 0 & 0 & -a_0^2 \\ \hdashline
0.045a_1{+}0.24 & -1 & 1.20 & 0 & 0 & -a_0^2 & -2a_0a_1 \\ \hdashline
0.045a_0{+}0.24a_1{+}0.001 & 1.20 & -2.37 & 0 & -a_0^2 & -2a_0a_1 & -a_1^2{-}2a_0 \\ \hdashline
0.24a_0{+}0.001a_1{+}0.02 & -2.37 & 2.01 & -a_0^2 & -2a_0a_1 & -a_1^2{-}2a_0 & -2a_1 \\ \hdashline
0.001a_0{+}0.02a_1{+}0.22 & 2.01 & -2.23 & -2a_0a_1 & -a_1^2 - 2a_0 & -2a_1 & -1 \\ \hdashline
0.02a_0{+}0.22a_1{+}0.02 & -2.23 & 1.04 & -a_1^2 - 2a_0 & -2a_1 & -1 & 0 \\ \hdashline
0.22a_0{+}0.02a_1 & 1.04 & -0.85 & -2a_1 & -1 & 0 & 0 \\ \hdashline
0.02a_0 & -0.85 & 0 & -1 & 0 & 0 & 0
\end{array}\right]$$

[9] We used a MacBook Pro with a 6-core Intel i7 CPU (2019) working at 2.6 GHz with access to 32 GB RAM. All *timings* are averaged over 5 consecutive runs and the numerical results are rounded for displaying purposes. A MATLAB implementation of the (block) Macaulay method [175] is available online at `www.macaulaylab.net`.

Table 2.2: The objective function value, model parameters and corresponding poles for the critical points encountered in Example 2.12.

$J^\dagger$	a_1	a_0	b_1	b_0	$p_{1,2}$
0.868	-0.293	0.941	0.139	0.266	$0.1465 \pm 0.9589j$
1.076	0.505	0.930	-0.254	-0.120	$-0.2525 \pm 0.9307j$
1.124	0.267	0.820	-0.294	0.167	$-0.1335 \pm 0.8956j$
1.174	-1.423	0.969	0.069	0.028	$0.7115 \pm 0.6803j$
1.254	-0.992	0.534	0.132	0.086	$0.4960 \pm 0.5366j$

†the discrete-time equivalent of (2.4).

which leads to a quadratic 2-parameter eigenvalue problem:

$$\Big(\boldsymbol{M}_0 + \boldsymbol{M}_{10}a_0 + \boldsymbol{M}_{01}a_1 + \boldsymbol{M}_{20}a_0^2 + \boldsymbol{M}_{11}a_0a_1 + \boldsymbol{M}_{02}a_1^2\Big)\begin{bmatrix}1 & b_0 & b_1 & g_0 & g_1 & g_2 & g_3\end{bmatrix}^{\mathsf{T}} = \boldsymbol{0},$$

with coefficient matrices $\boldsymbol{M}_{\boldsymbol{\alpha}} \in \mathbb{R}^{8\times 7}$. The block Macaulay matrix of degree 10 suffices to retrieve the 49 eigentuples, 11 of which real-valued.

The solution set $\mathcal{W}'_{\mathbb{R}}$ contains 5 tuples $\boldsymbol{a}_{\mathrm{tr}}$. The values of the corresponding critical points are listed in Table 2.2 and a contour plot of the objective function value for various second-order reduced models $\widehat{H}(z)$ has been given in Figure 1.2. To obtain the objective function value for each set of parameter values $\{a_0, a_1\}$ when constructing the contour plot, the corresponding optimal parameters $\{b_0, b_1\}$ have been computed via the procedure described in Remark 2.7. Note that Figure 1.2b depicts only a subset of the feasible reduced models $\widehat{H}(z)$, excluding reduced models with two (differing) real-valued poles.

Example 2.13. *We search for the globally optimal 4th order reduced model ($n = 4$) of the continuous-time state-space model ($N = 17$) considered in Example 3 of [181]. The model describes the interaction between a torque activator and an approximately collocated torsional rate sensor for the ACES structure [31]. Because of the high complexity (see Table 2.1) we used the package of [20] to calculate the common roots of* (2.13)*, which took* 3 h 24 min 2 s*. Contrary to the block Macaulay method [175], homotopy continuation methods can exploit parallelism in the computations, allowing to tackle larger problems. The poles p_i of the four critical points with smallest value for J are depicted in Table 2.3, the first of which corresponds to the globally optimal solution. For comparison, we applied the iterative method described in [77] using random sets of initial interpolation points to this example: the iteration generally converged*

Table 2.3: The poles p_i of the four critical points with the lowest value of J, corresponding to Example 2.13.

J	$p_{1,2}$	$p_{3,4}$
9.14×10^{-3}	$-0.032 \pm 78.54j$	$-0.111 \pm 15.43j$
9.22×10^{-3}	$-0.032 \pm 78.54j$	$-5.713 \pm 52.57j$
1.03×10^{-2}	$-0.032 \pm 78.54j$	$-0.023 \pm 3.842j$
1.09×10^{-2}	$-0.032 \pm 78.54j$	$-4.663 \pm 15.88j$

to locally optimal reduced models. Only in (very) rare cases the globally optimal solution was obtained.

2.5 Conclusion and Future Work

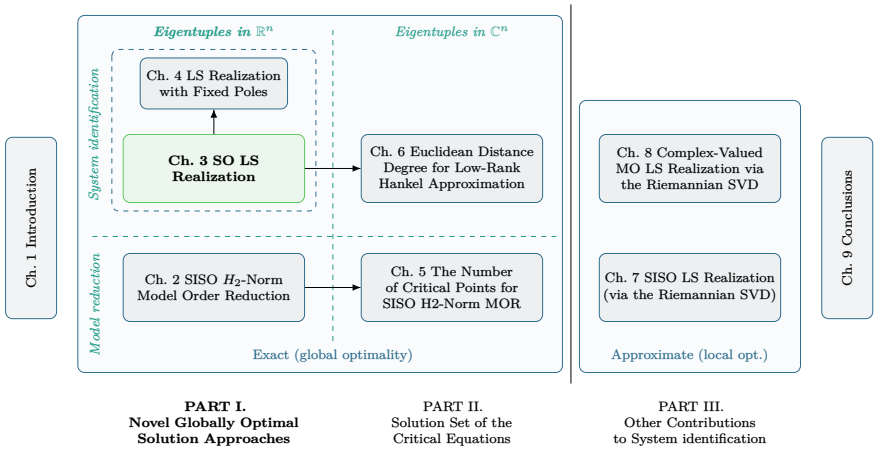
We derived a novel methodology for globally optimal SISO $\mathcal{H}_2$-norm model reduction. The parameters of the optimal model(s) are retrieved via the eigentuples of a multiparameter eigenvalue problem (MEP). The MEP is quadratic, independent of the model orders, which is favorable from a computational point of view. We compared the complexity of the novel technique with the few alternative globally optimal approaches in the literature, and solved three numerical examples to validate the proposed methodology.

In future work, we want to develop specialized solvers that compute only the real-valued eigentuples of the MEP, since these are the only ones of interest. Ideally, we would solve for the global optimal minimizer(s) only. The observation that also for MIMO models (tangential) interpolatory conditions for first-order optimality exist, motivates the search for a generalization of the presented methodology.

Another direction concerns the analysis of existing iterative methods, such as Iterative Rational Krylov Approximation (IRKA). Although IRKA performs remarkably well in practice, particularly when supplied with good initial interpolation data, no general proof of convergence is known, except in the special case of symmetric models [65, 77]. By comparing IRKA's iterates with the globally optimal solutions obtained using the proposed framework, one may gain insights into when IRKA converges, why it fails to do so in certain settings, and how it might be modified to improve robustness.

Chapter 3

Exact Solution to the Least Squares Realization Problem as a Multiparameter Eigenvalue Problem



Chapter abstract

We consider the least squares (LS) realization of an autonomous single-output linear time-invariant dynamical model for a given sequence of output data. As opposed to standard system identification practices, which often rely on (heuristic) iterative optimization techniques, we propose a global solution to this non-convex optimization problem: the *globally optimal* solution(s) are identified by means of a *deterministic* eigenvalue procedure. In particular, we illustrate that for all (local) minimizers, the corresponding misfit can be characterized as the result of filtering an unknown signal twice through a finite-impulse response filter. Exploiting this insight, we propose a novel (rectangular) multiparameter eigenvalue problem (MEP), the eigentuples of which allow to retrieve all local and global minimizers of the identification problem. The proposed MEP is of great theoretical interest and offers new insights into the structure of the LS realization problem, which we explore in detail. We then extend the novel framework by incorporating weights into the objective function, thereby making the connection to weighted low-rank Hankel matrix approximation. Several numerical examples are worked out to illustrate our findings.

Candidate's contributions and relation to published work

This work has been previously published as:

- **Lagauw, S.**, Vanpoucke, L., and De Moor, B. (2026). "Exact Solution to the Least Squares Realization Problem as a Multiparameter Eigenvalue Problem". *Automatica* 185, pp. 112792.
- **Lagauw, S.**, Vanpoucke, L., and De Moor, B. (2024). "Exact Characterization of the Global Optima of Least Squares Realization of Autonomous LTI Models as a Multiparameter Eigenvalue Problem". In: *Proceedings of the 22nd European Control Conference (ECC)*. KTH Royal Institute of Technology, Stockholm, Sweden, pp. 3439– 3444.

The candidate led the theoretical development, implemented the code, conducted the experiments, and authored the text.

This chapter differs slightly from the *Automatica* article:

- The notation and terminology was adapted to ensure consistency with the conventions used throughout the remainder of this dissertation.
- Section 3.7, which was not part of the original publication [110], extends the globally optimal methodology for LS realization to the setting of weighted structured low-rank approximation for Hankel matrices.
- Section 3.8, which was not part of the original publication [110], offers additional observations that complement the material presented in this chapter. Rather than introducing new technical developments, the emphasis is placed on conceptual insights.

3.1 Introduction

This chapter presents novel mathematical insights into the *least squares (LS) realization problem*. While this problem has been extensively studied, the results obtained here were, to the best of our knowledge, previously unknown and have not been documented in the literature.

3.1.1 Problem formulation

We consider *autonomous single-output linear time-invariant (LTI) causal dynamical models in discrete-time*, with a pre-specified model order n, which corresponds to the number of states. For this model class, output data $\widehat{\boldsymbol{y}} = [\widehat{y}_0 \ \cdots \ \widehat{y}_{N-1}]^\mathsf{T} \in \mathbb{R}^N$ (with $N > 2n$) satisfy a difference equation of the form:

$$a_0\widehat{y}_k + \cdots + a_{n-1}\widehat{y}_{k+n-1} + a_n\widehat{y}_{k+n} = a(z)\widehat{y}_k = 0, \tag{3.1}$$

where $a(z) = a_0 + \cdots + a_{n-1}z^{n-1} + a_n z^n$ is a degree n polynomial in the forward-shift operator z, i.e., $z\widehat{y}_k = \widehat{y}_{k+1}$. The coefficients of $a(z)$ correspond to the parameters $\boldsymbol{a} = [a_0 \ a_1 \ \cdots \ a_n]^\mathsf{T} \in \mathbb{R}^{n+1}$ of the autonomous LTI model. Data $\widehat{\boldsymbol{y}}$ that satisfies (3.1) for all $k = 0, \ldots, N{-}n{-}1$ will be called *model-compliant* with respect to the model parameters.

To ensure uniqueness of the parametrization and without loss of generality, we fix $a_n = 1$ for the remainder of this chapter, so that the model $a(z)$ is completely determined by n parameters. We will therefore denote the vector of unknown parameters as $\boldsymbol{a}_{\text{tr}} = [a_0 \ a_1 \ \cdots \ a_{n-1}]^\mathsf{T} \in \mathbb{R}^n$.

The LS realization problem is formulated as follows: given data $\boldsymbol{y} = [y_0 \ \cdots \ y_{N-1}]^\mathsf{T} \in \mathbb{R}^N$, we want to identify the model parameters $\boldsymbol{a}_{\text{tr}}$ that minimize the 2-norm of the so-called misfit $\widetilde{\boldsymbol{y}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$, which is the modification needed to make the given data $\boldsymbol{y}$ model-compliant:

$$\underset{\boldsymbol{a}_{\text{tr}} \in \mathbb{R}^n, \, \widehat{\boldsymbol{y}} \in \mathbb{R}^N}{\text{minimize}} \quad \frac{1}{2}\left\|\widetilde{\boldsymbol{y}}\right\|_2^2 = \frac{1}{2}\left\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\right\|_2^2, \tag{3.2a}$$

$$\text{subject to} \quad a(z)\widehat{y}_k = 0, \ \ \forall k = 0, \ldots, N{-}n{-}1. \tag{3.2b}$$

This modeling setup is depicted in Figure 3.1.

Remark that even though the considered model class is linear in the parameters $\boldsymbol{a}$, it is nonlinear in the unknowns $\boldsymbol{a}_{\text{tr}}$ and $\widehat{\boldsymbol{y}}$ simultaneously. Consequently, the LS realization problem constitutes a *non-convex optimization problem*, implying that (many) local optima can exist (see e.g., Figure 3.2).

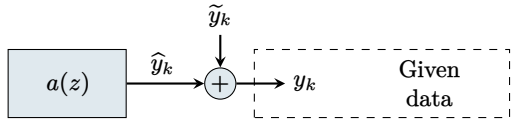


Figure 3.1: Schematic representation of the autonomous LTI single-output model. Given data $\boldsymbol{y} \in \mathbb{R}^N$ are to be modified using the so-called misfit $\widetilde{\boldsymbol{y}} \in \mathbb{R}^N$, so that the modified data $\widehat{\boldsymbol{y}} = \boldsymbol{y} - \widetilde{\boldsymbol{y}}$ are model-compliant.

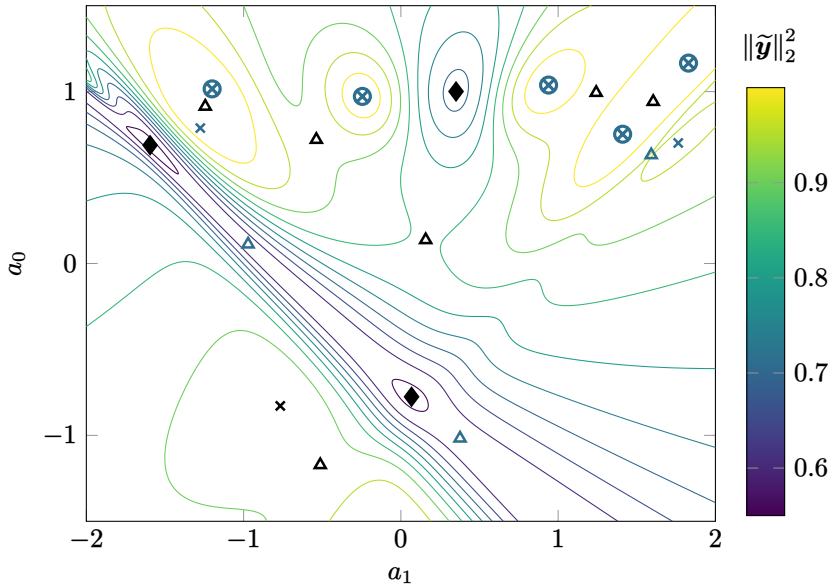


Figure 3.2: Contour plot of the objective function of the LS realization problem as a function of the parameters a_0, a_1, illustrating the non-convex nature of the problem. For all values (a_0, a_1), the optimal $\widehat{\boldsymbol{y}}$ was determined via (3.13) to compute the corresponding objective function value (cf. the formulation in (3.14)). The problem setup corresponds to Example 3.13 described in Section 3.6. Critical points contained in the displayed region are highlighted: local minimizers (♦), saddle points (△) and local maximizers (×), where critical points appearing as eigentuples of the proposed MEP (3.19) are depicted in black. Local maximizers for which the cost function reaches the global maximum $\|\boldsymbol{y}\|_2^2$ are encircled (⊗). The global minimizer lies outside the displayed region.

3.1.2 Contributions and outline

In this chapter, we propose a *global solution* to the LS realization problem: contrary to standard engineering practices (see Section 3.1.3), we deem the non-convex optimization problem (3.2) to be mathematically 'solved' if and only if the *globally optimal* model(s) have been identified by means of a *deterministic* procedure.

The main claim and results of this chapter are:

1. We formalize a finite-dimensional version of what is called 'Walsh's Theorem: the optimal misfit is the result of filtering an unknown signal twice by the same finite impulse response (FIR) filter (cf. Theorem 3.3).
2. We claim that the global solution to the LS realization problem is an eigenvalue problem: exploiting the characterization of Theorem 3.3, we propose a novel, *rectangular multiparameter eigenvalue problem (MEP)*, the set of eigentuples[1] of which contains all local and global minimizers of the problem (cf. Theorem 3.6). Besides its computational advantages, the novel MEP is also of great theoretical interest, leading new insights into the structure of the LS realization problem (see e.g., Chapter 6).
3. We demonstrate that, for a subset of the critical points of the problem, the corresponding model-compliant data $\widehat{\boldsymbol{y}}$ have so-called *lower-order dynamics*, and we show that these critical points must be either saddle points or local maximizers of the problem. This eventually leads to the proof of Theorem 3.6.

The remainder of this chapter is organized as follows. We conclude Section 3.1 by presenting an overview of the background and the key literature relevant to our work. Section 3.2 recapitulates the two alternative representations for the model dynamics described in [45]. The necessary conditions for first-order optimality are discussed in Section 3.3, leading to a finite-dimensional version of Walsh's Theorem. Exploiting the results from Sections 3.2-3.3, we propose the novel MEP in Section 3.4 and characterize the so-called 'lower-order' critical points of the LS realization problem in Section 3.5. Several numerical examples to illustrate the obtained results are provided in Section 3.6. We then extend the novel framework by incorporating weights into the objective function (Section 3.7) and in Section 3.8, make the connection to the SISO $\mathcal{H}_2$-norm model reduction problem considered in the previous chapter. We conclude the chapter in Section 3.9.

[1] A k-parameter eigenvalue problem has k-tuples of eigenvalues, which we refer to as *eigentuples*.

3.1.3 Background and literature overview

The literature about the difference relation (3.1) and methods to estimate the coefficients from data, from an applied engineering point of view, is vast. The first idea goes back to the French engineer Prony, in 1795 [48], see e.g., [85, 88, 91] for more information. From a contemporary systems theory point of view, (3.1) is the characteristic (or minimal) polynomial of a linear system, and hence relates subsequent samples of model-compliant output data from an autonomous LTI system or an LTI impulse response. Therefore, it should come as no surprise that the LS realization problem occurs in a wide variety of data driven engineering fields, such as (we limit ourselves to the most basic references as the literature here is enormous): high-resolution frequency estimation and harmonic retrieval [1, 21, 25, 40, 108, 166] (see [160] for an extensive overview), direction-of-arrival estimation in array processing [78, 134, 147, 161], shape-from-moments problems [57], and realization algorithms for converting impulse response samples into state-space models [89, 107, 191].

Engineers typically resort to various 'approximate' approaches to efficiently solve the non-convex optimization problem (3.2). Examples are the iterative quadratic maximum likelihood [116], Steiglitz-McBride [59, 141], constrained total least squares [1], the Riemannian SVD [38, 40, 42, 43], the canonical polyadic decomposition [61, 62, 63], and the variable projection method proposed in [74]. In addition to these specialized approaches, plain numerical optimization techniques, such as those often used in prediction error methods [120] and errors-in-variables methods [154], provide yet another alternative means of solving the problem.

Despite their effectiveness in practice, most of these methods are heuristic in the sense that, due to the non-convex nature of the problem, their performance is often highly sensitive to the initialization, convergence issues can arise, reproducibility of the obtained results is not always guaranteed, and even if some form of optimality is obtained, certification of *global* optimality is generally impossible[2]. Additionally, in a maximum-likelihood setting, a priori probabilistic assumptions are often made on the inaccuracies perturbing the model-compliant data, for instance that they follow a multivariate Gaussian distribution and/or that they are 'white'. While these assumptions can be justified asymptotically for large N, they are heuristic in a finite-sample setting.

The aforementioned approaches stand in marked contrast to our objective, which is to obtain a *global* solution: a deterministic approach that achieves global

[2]For completeness, we must note that, under a suitable bound on the perturbation of the given data, the canonical polyadic decomposition does admit a guarantee that the problem is well-posed and has a unique solution. We refer the interested reader to [60].

optimality. In prior work, various system identification problems have been shown to be essentially an eigenvalue problem [15, 174]. These findings are based on the combined insights that the problems in question are polynomial in nature and that global polynomial optimization is equivalent to solving a multiparameter eigenvalue problem (MEP) [52, 53].

The MEP generalizes the standard eigenvalue problem by allowing more than one eigenvalue and removing the restriction of linearity in its parameter(s). Consequently, each MEP is inherently connected to a system of multivariate polynomial equations. As a comprehensive treatment of MEPs lies beyond the scope of this text, we refer the interested reader to recent literature [35, 90, 173]. Efficient numerical algorithms to find all the (real-valued) affine eigentuples of an MEP are available, e.g., the (block) Macaulay framework described in [173, 175, 176] or the methods from [90, 138]. However, the computational complexity involved with solving an MEP grows exponentially with its degree and the dimensions of its coefficient matrices. Accordingly, MEP-based global solution approaches serve primarily as theoretical benchmarks rather than as practical tools suitable for large-scale problems.

The LS realization problem under consideration was surprisingly only recognized to be an MEP in [45]. Despite the notable theoretical importance of these findings, the obtained MEPs exhibit certain less advantageous characteristics: the size of the coefficient matrices and/or the degree of the MEPs grow fast for increasing problem sizes. By contrast, the MEP that we derive is *cubic* and has significantly more compact coefficient matrices, so that the novel methodology allows us to tackle (relatively) larger examples. The solution set of our MEP differs from that of the MEPs described in [45], as it excludes the so-called lower-order critical points (see Section 3.5.3), thereby reducing the problem complexity.

The 'double-FIR-filter' property of the optimal misfits, which was noted as a secondary observation in [45], serves as the foundation for the proposed MEP. The property is reminiscent of the interpolatory conditions described in Walsh's Theorem [141, Theorem 3.4], which originates from the field of rational approximation [184]. Our approach is further inspired by our previous work [109], where we leveraged Walsh's Theorem in an infinite-dimensional setting to derive a new methodology for globally optimal single-input single-output H_2-norm model reduction.

3.2 Model Dynamics

We briefly recapitulate the two alternative representations of the dynamics of a single-output autonomous LTI model described in [45] (Sections 3.2.1-3.2.2). Both representations reveal the structure of the model-compliant data $\widehat{\boldsymbol{y}}$, which will be instrumental in Sections 3.3 and 3.5.

3.2.1 Kernel representation

The set of all sequences $\widehat{\boldsymbol{y}} \in \mathbb{R}^N$ that are model-compliant with respect to given model parameters $\boldsymbol{a} \in \mathbb{R}^{n+1}$, is referred to as the *behavior* of that model. Writing out (3.1) for all $k = 0, \dots, N-n-1$, gives

$$\begin{bmatrix} a_0 & \cdots & \cdots & a_{n-1} & 1 & 0 & \cdots & 0 \\ 0 & a_0 & \cdots & \cdots & a_{n-1} & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_0 & \cdots & \cdots & a_{n-1} & 1 \end{bmatrix} \widehat{\boldsymbol{y}} = \boldsymbol{T}_{N-n}^{\boldsymbol{a}} \widehat{\boldsymbol{y}} = \boldsymbol{0}, \tag{3.3}$$

indicating that the behavior of the model can be characterized as the n-dimensional kernel (i.e., right null space) of the banded-Toeplitz[3] matrix $\boldsymbol{T}_{N-n}^{\boldsymbol{a}} \in \mathbb{R}^{(N-n)\times N}$. Note that one can replace the constraints (3.2b) in the LS realization problem by (3.3).

The matrix-vector product in (3.3) can be rewritten as

$$\boldsymbol{T}_{N-n}^{\boldsymbol{a}} \widehat{\boldsymbol{y}} = \widehat{\boldsymbol{Y}} \boldsymbol{a} = \boldsymbol{0}, \tag{3.4}$$

where $\widehat{\boldsymbol{Y}} \in \mathbb{R}^{(N-n)\times(n+1)}$ is the Hankel matrix constructed from the elements of $\widehat{\boldsymbol{y}}$,

$$\widehat{\boldsymbol{Y}} = \begin{bmatrix} \widehat{y}_0 & \widehat{y}_1 & \widehat{y}_2 & \cdots & \widehat{y}_n \\ \widehat{y}_1 & \widehat{y}_2 & \widehat{y}_3 & \cdots & \widehat{y}_{n+1} \\ \vdots & \vdots & \vdots & & \vdots \\ \widehat{y}_{N-n-1} & \widehat{y}_{N-n} & \widehat{y}_{N-n+1} & \cdots & \widehat{y}_{N-1} \end{bmatrix}. \tag{3.5}$$

Hence, given a (sufficiently large) sequence of model-compliant data $\widehat{\boldsymbol{y}}$, one can retrieve the corresponding model parameters $\boldsymbol{a}$ via the kernel of $\widehat{\boldsymbol{Y}}$.

For the optimization problem (3.2) to be well-defined, the given data must be incompatible with the model class under consideration. That is, the Hankel matrix $\boldsymbol{Y} \in \mathbb{R}^{(N-n)\times(n+1)}$, constructed from the given data $\boldsymbol{y}$ and defined

[3] The super- and subscript denote the generating vector and number of rows of the Toeplitz matrix $\boldsymbol{T}_{N-n}^{\boldsymbol{a}}$, respectively.

similarly as in (3.5), is assumed to be of full column rank. Indeed, when this is not the case, one can always find model parameters $\boldsymbol{a}$ for which the given data $\boldsymbol{y}$ lie in the behavior of the model: $\boldsymbol{Y}\boldsymbol{a} = \boldsymbol{T}_{N-n}^{\boldsymbol{a}}\boldsymbol{y} = \boldsymbol{0}$, such that $\widetilde{\boldsymbol{y}} = \boldsymbol{0}$. Since $\boldsymbol{Y}$ is, by construction, not of full column rank when $n < N \leq 2n$, we require $N > 2n$.

3.2.2 Image representation

Alternatively, the behavior of the autonomous model can be expressed via a Vandermonde matrix, the image of which equals the kernel of $\boldsymbol{T}_{N-n}^{\boldsymbol{a}}$, leading to the so-called *image representation* of the model dynamics. In this case, the behavior is parametrized by the n poles $\{p_i \in \mathbb{C} \,|\, i = 1, \ldots, n\}$ of the model, i.e., the roots of the monic characteristic polynomial $a(z)$:

$$a(z) = \prod_{i=1}^{n} (z - p_i).$$

For real-valued models, complex-valued poles appear in complex-conjugate pairs $p_{1,2} = \alpha \pm \beta j$ with $\alpha, \beta \in \mathbb{R}$. The model can therefore be parametrized using n real-valued coefficients, as is the case with the kernel representation. For ease of notation, we will assume that the model has simple poles only, i.e., all roots of $a(z)$ have multiplicity one. In this case, each pole p_i determines a mode $\boldsymbol{p_i} \in \mathbb{C}^N$ that contributes to the behavior of the model:

$$\boldsymbol{p_i} = \begin{bmatrix} 1 & p_i & p_i^2 & \cdots & p_i^{N-1} \end{bmatrix}^{\mathsf{T}}. \tag{3.6}$$

This allows us to represent the model-compliant output data $\widehat{\boldsymbol{y}}$ as the following weighted sum,

$$\widehat{\boldsymbol{y}} = \boldsymbol{V}\widehat{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{p_1} & \boldsymbol{p_2} & \cdots & \boldsymbol{p_n} \end{bmatrix} \widehat{\boldsymbol{x}}, \tag{3.7}$$

where $\boldsymbol{V} \in \mathbb{C}^{N \times n}$ is a Vandermonde matrix and $\widehat{\boldsymbol{x}} \in \mathbb{C}^n$.

The Vandermonde matrix $\boldsymbol{V}$ is guaranteed to be of full column rank since Vandermonde vectors $\boldsymbol{p_i}$ and $\boldsymbol{p_j}$ generated by different points $p_i \neq p_j$ are linearly independent by construction. If one were to consider poles with multiplicity greater than one, the matrix $\boldsymbol{V}$ has to be composed using a combination of standard and confluent Vandermonde vectors (see [45] for more details). Also in this case, the matrix $\boldsymbol{V}$ is guaranteed to be of full column rank.

Notice that, in contrast to the image representation, the kernel representation conceals the complexity arising from multiple and complex-conjugate poles. Indeed, this parametrization accommodates all cases without requiring modification, which is advantageous from an optimization standpoint and motivates our choice to employ the kernel representation in formulating our MEP.

3.3 Characterizing Optimality

Consider the Lagrangian of (3.2),

$$\mathcal{L}(\boldsymbol{a}, \widehat{\boldsymbol{y}}, \boldsymbol{\ell}) = \frac{1}{2} \left\| \boldsymbol{y} - \widehat{\boldsymbol{y}} \right\|_2^2 + \boldsymbol{\ell}^{\mathsf{T}} \boldsymbol{T}_{N-n}^{\boldsymbol{a}} \widehat{\boldsymbol{y}},$$

where the variables $\boldsymbol{\ell} \in \mathbb{R}^{N-n}$ are Lagrange multipliers. The first-order necessary conditions for optimality of (3.2) can now be obtained as,

$$\partial \mathcal{L} / \partial \boldsymbol{\ell} = \boldsymbol{T}_{N-n}^{\boldsymbol{a}} \widehat{\boldsymbol{y}} = \boldsymbol{0}, \tag{3.8a}$$

$$\partial \mathcal{L} / \partial \widehat{\boldsymbol{y}} = \widehat{\boldsymbol{y}} - \boldsymbol{y} + (\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{T}} \boldsymbol{\ell} = \boldsymbol{0}, \tag{3.8b}$$

$$\partial \mathcal{L} / \partial \boldsymbol{a}_{\mathrm{tr}} = (\widehat{\boldsymbol{Y}}|)^{\mathsf{T}} \boldsymbol{\ell} = \boldsymbol{0}, \tag{3.8c}$$

where $\widehat{\boldsymbol{Y}}| \in \mathbb{R}^{(N-n) \times n}$ denotes the matrix obtained by removing the rightmost column from $\widehat{\boldsymbol{Y}}$. The solutions $(\boldsymbol{a}, \widehat{\boldsymbol{y}}, \boldsymbol{\ell})$ to the system of equations (3.8) are the *critical points* of the LS realization problem (3.2), of which only the real-valued affine ones are relevant for our purposes.

The LS realization problem (3.2) has a nested structure:

1. for given data $\boldsymbol{y}$ and an instance of model parameters $\boldsymbol{a}_{\mathrm{tr}}$, find the minimal 2-norm misfit that makes the given data model-compliant with respect to the given model $\boldsymbol{a}_{\mathrm{tr}}$ (inner-minimization);
2. for given data $\boldsymbol{y}$, the model parameters $\boldsymbol{a}_{\mathrm{tr}}$ are to be optimized so that the optimal misfit obtained from the inner-minimization has a minimal 2-norm.

In what follows, we will illustrate that the optimality conditions (3.8a)-(3.8b) lead to a closed-form solution for the inner-minimization, i.e., the optimal data sequences $\widehat{\boldsymbol{y}}$ and $\widetilde{\boldsymbol{y}}$ can be expressed as a function of the given data $\boldsymbol{y}$ and the model parameters $\boldsymbol{a}_{\mathrm{tr}}$ (Section 3.3.1). The implications of optimality of the model parameters $\boldsymbol{a}_{\mathrm{tr}}$, via (3.8c), are discussed in Section 3.3.2.

3.3.1 Orthogonal decomposition of the ambient space

The relation in (3.8b) gives

$$\widetilde{\boldsymbol{y}} = (\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{T}} \boldsymbol{\ell}. \tag{3.9}$$

The matrix-vector product corresponds to a convolution operation, indicating that the misfit is the result of filtering the signal $\boldsymbol{\ell}$ through the FIR filter defined by $a(z^{-1})$. Combined with (3.8a), this illustrates that a particular model, defined by its parameters $\boldsymbol{a}$, induces an orthogonal decomposition of the ambient data space $\mathbb{R}^N$: the optimal misfit $\widetilde{\boldsymbol{y}}$ resides in the $(N-n)$-dimensional row space of $\boldsymbol{T}_{N-n}^{\boldsymbol{a}}$, whereas the optimal model-compliant data $\widehat{\boldsymbol{y}}$ lie in the n-dimensional kernel of $\boldsymbol{T}_{N-n}^{\boldsymbol{a}}$. It is a natural consequence that the optimal model-compliant data and misfit are orthogonal: $\widehat{\boldsymbol{y}}^{\mathsf{T}}\widetilde{\boldsymbol{y}} = 0$.

Combining (3.8a) and (3.9) leads to

$$\begin{aligned} \partial\mathcal{L}/\partial\boldsymbol{\ell} = \boldsymbol{T}_{N-n}^{\boldsymbol{a}}\widehat{\boldsymbol{y}} = \boldsymbol{T}_{N-n}^{\boldsymbol{a}}(\boldsymbol{y}-\widetilde{\boldsymbol{y}}) = \boldsymbol{0} \\ \iff \boldsymbol{T}_{N-n}^{\boldsymbol{a}}\boldsymbol{y} = \boldsymbol{T}_{N-n}^{\boldsymbol{a}}(\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{T}}\boldsymbol{\ell}, \qquad (3.10) \\ \iff \boldsymbol{\ell} = (\boldsymbol{D}_{N-n}^{\boldsymbol{a}})^{-1}\boldsymbol{T}_{N-n}^{\boldsymbol{a}}\boldsymbol{y}, \qquad (3.11) \end{aligned}$$

where the matrix $\boldsymbol{D}_{N-n}^{\boldsymbol{a}} \coloneqq \boldsymbol{T}_{N-n}^{\boldsymbol{a}}(\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{T}} \in \mathbb{R}^{(N-n)\times(N-n)}$ is invertible for any nonzero choice of $\boldsymbol{a} \in \mathbb{R}^{n+1}$. Subsequently, substitution of (3.11) into (3.9) gives

$$\widetilde{\boldsymbol{y}} = (\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{T}}(\boldsymbol{D}_{N-n}^{\boldsymbol{a}})^{-1}\boldsymbol{T}_{N-n}^{\boldsymbol{a}}\boldsymbol{y}, \qquad (3.12)$$

which, using $\widehat{\boldsymbol{y}} = \boldsymbol{y} - \widetilde{\boldsymbol{y}}$, implies that

$$\widehat{\boldsymbol{y}} = \left(\boldsymbol{I} - (\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{T}}(\boldsymbol{D}_{N-n}^{\boldsymbol{a}})^{-1}\boldsymbol{T}_{N-n}^{\boldsymbol{a}}\right)\boldsymbol{y}. \qquad (3.13)$$

This indicates that, if the model parameters $\boldsymbol{a}$ were known (cf. the inner-minimization), they would uniquely define the corresponding optimal $\boldsymbol{\ell}$, $\widetilde{\boldsymbol{y}}$ and $\widehat{\boldsymbol{y}}$ via (3.11), (3.12) and (3.13), respectively. In particular, the optimal data sequences $\widetilde{\boldsymbol{y}}$ and $\widehat{\boldsymbol{y}}$ are obtained via the orthogonal projection of the given data $\boldsymbol{y}$ onto the linear subspaces defined by the matrix $\boldsymbol{T}_{N-n}^{\boldsymbol{a}}$. The existence of this closed-form solution for the inner-minimization implies that the LS realization problem is an instance of a *separable nonlinear LS problem*[4] [73].

Consequently, for given data $\boldsymbol{y}$ and model order n, a critical point $(\boldsymbol{a}, \widehat{\boldsymbol{y}}, \boldsymbol{\ell})$ of (3.2) is completely determined by its parameter values $\boldsymbol{a}$. From this point onward, whenever referring to specific model parameters $\boldsymbol{a}$, we assume that $\boldsymbol{\ell}$, $\widetilde{\boldsymbol{y}}$ and $\widehat{\boldsymbol{y}}$ are defined as in (3.11), (3.12) and (3.13).

While tangential to our main discussion, observe that substitution of the closed-form solution (3.12) into the outer problem (3.2) eliminates the N decision

[4]The 'separability' of the nested optimization problem is a natural property of the LS misfit modeling setup and arises also when more complex model classes are considered, such as multiple-input multiple-output (MIMO) models [144]. It has been exploited by iterative solution approaches, e.g., the alternating LS iterations in [74] and [144, Algorithm 4].

variables $\widehat{\boldsymbol{y}}$, leading to an equivalent, nonlinear, unconstrained optimization problem over the n model parameters $\boldsymbol{a}$ only

$$\underset{\boldsymbol{a}\in\mathbb{R}^n}{\text{minimize}}\ \frac{1}{2}\left\|\widetilde{\boldsymbol{y}}\right\|_2^2 = \frac{1}{2}\boldsymbol{y}^{\mathsf{T}}(\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{T}}(\boldsymbol{D}_{N-n}^{\boldsymbol{a}})^{-1}\boldsymbol{T}_{N-n}^{\boldsymbol{a}}\boldsymbol{y}. \tag{3.14}$$

This formulation has been used in Figure 3.2.

Example 3.1. *Consider the numerical example described in [45, Section 8.2], where the globally optimal first-order autonomous LTI LS realization is computed for the sequence of given output data*

$$\boldsymbol{y} = \begin{bmatrix}4 & 3 & 2 & 1\end{bmatrix}^{\mathsf{T}}.$$

For $n = 1$, the polynomial $a(z)$ from (3.1) corresponds to $z + a_0$. By (3.12), the minimal norm misfit associated to the globally optimal solution $a_0 = -0.6764$ (this value will be computed in Section 3.4) can be obtained via the following orthogonal projection:

$$\widetilde{\boldsymbol{y}} = \begin{pmatrix}-0.1155\\ 0.2165\\ 0.1173\\ -0.2734\end{pmatrix} = \underbrace{\begin{pmatrix}0.4326 & -0.3838 & -0.2596 & -0.1756\\ -0.3838 & 0.7404 & -0.1756 & -0.1187\\ -0.2596 & -0.1756 & 0.8813 & -0.0803\\ -0.1756 & -0.1187 & -0.0803 & 0.9457\end{pmatrix}}_{\boldsymbol{Q}}\begin{pmatrix}4\\3\\2\\1\end{pmatrix},$$

where

$$\boldsymbol{Q} = \left[(\boldsymbol{T}_3^{\boldsymbol{a}})^{\mathsf{T}}(\boldsymbol{D}_3^{\boldsymbol{a}})^{-1}\boldsymbol{T}_3^{\boldsymbol{a}}\right]_{a_0=-0.6764}. \tag{3.15}$$

The image representation (3.7) provides an alternative expression for (3.13). It follows from $\operatorname{null}\left(\boldsymbol{T}_{N-n}^{\boldsymbol{a}}\right) = \operatorname{range}(\boldsymbol{V})$ that if the poles $\{p_i \in \mathbb{C} \,|\, i = 1, \ldots, n\}$ of the model are known, the optimal model-compliant data $\widehat{\boldsymbol{y}}$ can be obtained via the orthogonal projection of $\boldsymbol{y}$ onto the image of $\boldsymbol{V}$:

$$\widehat{\boldsymbol{y}} = \boldsymbol{V}[(\boldsymbol{V})^{\mathsf{H}}\boldsymbol{V}]^{-1}(\boldsymbol{V})^{\mathsf{H}}\boldsymbol{y}, \tag{3.16}$$

where $(\boldsymbol{V})^{\mathsf{H}}$ denotes the Hermitian conjugate of $\boldsymbol{V}$.

Example 3.2 (Example 3.1 continued)**.** *The model described by the parameter $a_0 = -0.6764$ has one pole: $p_1 = 0.6764$. It can be verified numerically that*

$$\boldsymbol{Q} = \left[\boldsymbol{I} - \boldsymbol{p_1}[\boldsymbol{p_1}^{\mathsf{T}}\boldsymbol{p_1}]^{-1}\boldsymbol{p_1}^{\mathsf{T}}\right]_{p_1=0.6764},$$

where $\boldsymbol{p_1} = [1\ p_1\ p_1^2\ p_1^3]^{\mathsf{T}}$ and $\boldsymbol{Q}$ is the matrix from (3.15).

3.3.2 Optimal model parameters

Taking into account the optimality of the model parameters (3.8c) illustrates that in a subset of the critical points of (3.2), the misfit is the result of filtering an unknown signal twice through the FIR filter defined by $a(z^{-1})$.

Theorem 3.3 (Double FIR filter)**.** *Given a sequence of data $\boldsymbol{y}$, a model order n and a critical point $\boldsymbol{a}_{\text{tr}}$ of* (3.2) *for which* $\operatorname{rank} \widehat{\boldsymbol{Y}} = n$*, then the optimal misfit $\widetilde{\boldsymbol{y}}$ can be expressed as,*

$$\widetilde{\boldsymbol{y}} = (\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{T}} (\boldsymbol{T}_{N-2n}^{\boldsymbol{a}})^{\mathsf{T}} \boldsymbol{g}, \tag{3.17}$$

for some signal $\boldsymbol{g} \in \mathbb{R}^{N-2n}$, where the matrix $\boldsymbol{T}_{N-2n}^{\boldsymbol{a}} \in \mathbb{R}^{(N-2n)\times(N-n)}$ is a banded Toeplitz matrix defined similarly to the matrix $\boldsymbol{T}_{N-n}^{\boldsymbol{a}}$ from (3.3)*.*

Proof. Consider a critical point $\boldsymbol{a}_{\text{tr}}$ and corresponding model-compliant data $\widehat{\boldsymbol{y}}$ obtained via the orthogonal projection (3.13). Model-compliance of the data ensures that

$$\boldsymbol{T}_{N-2n}^{\boldsymbol{a}} \widehat{\boldsymbol{Y}}| = \boldsymbol{0}.$$

It follows from (3.4) that $\operatorname{rank} \widehat{\boldsymbol{Y}} = n$ implies $\operatorname{rank} \widehat{\boldsymbol{Y}}| = n$, such that in this case the rows of the banded Toeplitz matrix $\boldsymbol{T}_{N-2n}^{\boldsymbol{a}}$, which are linearly independent by construction, span the entire $(N-2n)$-dimensional left null space of $\widehat{\boldsymbol{Y}}|$. Combining this insight with (3.8c) illustrates that there must exist a vector $\boldsymbol{g} \in \mathbb{R}^{N-2n}$ such that,

$$\boldsymbol{\ell} = (\boldsymbol{T}_{N-2n}^{\boldsymbol{a}})^{\mathsf{T}} \boldsymbol{g}. \tag{3.18}$$

Substituting (3.18) into (3.8b) gives the required result. □

Example 3.4 (Example 3.2 continued)**.** *The misfit associated to the global minimizer $a_0 = -0.6764$ can be characterized as*

$$\widetilde{\boldsymbol{y}} = \begin{pmatrix} -0.1155 \\ 0.2165 \\ 0.1173 \\ -0.2734 \end{pmatrix} = \left[(\boldsymbol{T}_3^{\boldsymbol{a}})^{\mathsf{T}} (\boldsymbol{T}_2^{\boldsymbol{a}})^{\mathsf{T}} \right]_{a_0 = -0.6764} \underbrace{\begin{bmatrix} -0.2525 \\ -0.2734 \end{bmatrix}}_{\boldsymbol{g}}.$$

Remark that (3.18) is generally not applicable to critical points $\boldsymbol{a}$ for which $\operatorname{rank} \widehat{\boldsymbol{Y}} = m < n$. Indeed, in this case row $\left(\boldsymbol{T}_{N-2n}^{\boldsymbol{a}}\right)$ is only a subspace of the $(N-n-m)$-dimensional left null space of $\widehat{\boldsymbol{Y}}|$. This phenomenon is considered in more detail in Section 3.5.2.

3.4 Multiparameter Eigenvalue Problem

The characterization of the optimal misfit in Theorem 3.3 allows us to formulate a *cubic rectangular n-parameter eigenvalue problem (MEP)* in the parameters $\boldsymbol{a}_{\text{tr}}$:

$$\boldsymbol{M}(\boldsymbol{a}_{\text{tr}})\boldsymbol{v} = \begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-n}\boldsymbol{y} & \boldsymbol{T}^{\boldsymbol{a}}_{N-n}(\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2n})^{\mathsf{T}} \end{bmatrix} \boldsymbol{v} = \boldsymbol{0}, \tag{3.19}$$

with $\|\boldsymbol{v}\| \neq 0$. The derivation of this MEP will be detailed in Section 3.4.1.

The matrix $\boldsymbol{M}(\boldsymbol{a}_{\text{tr}}) = \sum_{\{\boldsymbol{\alpha}\}} \boldsymbol{M}_{\boldsymbol{\alpha}} \boldsymbol{a}^{\boldsymbol{\alpha}}$, is a third-degree matrix polynomial in the monomials $\boldsymbol{a}_{\text{tr}}{}^{\boldsymbol{\alpha}} = a_0^{\alpha_0} \cdots a_{n-1}^{\alpha_{n-1}}$, with coefficient matrices $\boldsymbol{M}_{\boldsymbol{\alpha}} \in \mathbb{R}^{(N-n)\times(N-2n+1)}$, the size of which indicates that (3.19) is overdetermined when $n > 1$. The values $\boldsymbol{a}_{\text{tr}} \in \mathbb{C}^n$ for which the matrix polynomial $\boldsymbol{M}(\boldsymbol{a}_{\text{tr}})$ becomes rank-deficient, so that there exists a nonzero vector $\boldsymbol{v} \in \mathbb{C}^{N-2n+1}$ for which these equations are satisfied, are the *affine eigentuples* of this MEP.

3.4.1 Eigentuples and critical points

The following two results relate the real-valued eigentuples $\boldsymbol{a}$ of the proposed MEP (3.19) to the critical points of (3.2). Theorem 3.6, the proof of which is given in Section 3.5.2, constitutes our *main contribution.*

Lemma 3.5. *For given data $\boldsymbol{y}$ and model order n:*

1. *each critical point $\boldsymbol{a}_{\text{tr}}$ for which* $\operatorname{rank} \widehat{\boldsymbol{Y}} = n$ *maps to an affine, real-valued eigentuple of the MEP* (3.19)*;*
2. *each affine, real-valued eigentuple $\boldsymbol{a}_{\text{tr}}$ of the MEP* (3.19) *constitutes a critical point of* (3.2)*.*

Proof. Substitution of (3.18) into (3.10) illustrates that for a critical point $\boldsymbol{a}_{\text{tr}}$ of (3.2) for which the corresponding model-compliant data $\widehat{\boldsymbol{y}}$ satisfy $\operatorname{rank} \widehat{\boldsymbol{Y}} = n$, there must exist a vector $\boldsymbol{g} = [g_0, g_1, \ldots, g_{N-2n-1}]^{\mathsf{T}} \in \mathbb{R}^{N-2n}$ such that

$$\boldsymbol{T}^{\boldsymbol{a}}_{N-n}\boldsymbol{y} = \boldsymbol{T}^{\boldsymbol{a}}_{N-n}(\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2n})^{\mathsf{T}}\boldsymbol{g}. \tag{3.20}$$

Bringing everything to the right-hand side and separating out the variables $\boldsymbol{g}$ from (3.20) into the eigenvector gives (3.19), with $\boldsymbol{v} = [-1,\, \boldsymbol{g}^{\mathsf{T}}]^{\mathsf{T}}$. This proves the first claim.

Secondly, for all $\boldsymbol{a}_{\text{tr}} \in \mathbb{R}^n$ the product $\boldsymbol{T}^{\boldsymbol{a}}_{N-n}(\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}}$ is symmetric positive definite since $\boldsymbol{T}^{\boldsymbol{a}}_{N-n}$ is full row rank, such that the matrix $\boldsymbol{T}^{\boldsymbol{a}}_{N-n}(\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2n})^{\mathsf{T}}$

has full column rank for all $\boldsymbol{a}_{\mathrm{tr}} \in \mathbb{R}^n$. By consequence, there is a one-to-one correspondence between the real-valued affine eigentuples $\boldsymbol{a}_{\mathrm{tr}}$ of the MEP (3.19) and the real-valued affine common roots $(\boldsymbol{a}_{\mathrm{tr}}, \boldsymbol{g})$ of the system of polynomial equations (3.20). Indeed, the first element of the eigenvector $\boldsymbol{v}$ of (3.19) is always nonzero[5], such that the eigenvector $\boldsymbol{v}$ can be appropriately normalized to retrieve $\boldsymbol{g}$. In addition, the matrix obtained after substitution of $\boldsymbol{a}_{\mathrm{tr}}$ into $\boldsymbol{M}(\boldsymbol{a}_{\mathrm{tr}})$ can drop rank by at most one, such that each eigentuple $\boldsymbol{a}_{\mathrm{tr}}$ has a unique eigenvector $\boldsymbol{v}$ (up to normalization).

It now suffices to show that each tuple $(\boldsymbol{a}_{\mathrm{tr}}, \boldsymbol{g})$ satisfying (3.20) also satisfies (3.8). If, given such a tuple $(\boldsymbol{a}_{\mathrm{tr}}, \boldsymbol{g})$, $\boldsymbol{\ell}$ is defined as in (3.11) and $\widehat{\boldsymbol{y}}$ via the orthogonal projection (3.13), then (3.8a) and (3.8b) are satisfied by construction. Multiplication of (3.20) from the left with $(\boldsymbol{D}^{\boldsymbol{a}}_{N-n})^{-1}$ gives $(\boldsymbol{D}^{\boldsymbol{a}}_{N-n})^{-1}\boldsymbol{T}^{\boldsymbol{a}}_{N-n}\boldsymbol{y} = (\boldsymbol{T}^{\boldsymbol{a}}_{N-2n})^{\mathsf{T}}\boldsymbol{g}$, which, using (3.11), implies that (3.8c) is also satisfied. □

Theorem 3.6. *For given data $\boldsymbol{y}$ and model order n, consider the MEP defined in* (3.19)*. The set of affine, real-valued eigentuples of this MEP contains the model parameters $\boldsymbol{a}_{\mathrm{tr}}$ of all (locally) minimizing critical points, and is therefore guaranteed to include the* global minimizer(s) *of the LS realization problem* (3.2)*.*

See Section 3.5.2 for the proof of Theorem 3.6, which builds on the analysis of lower-order dynamics presented in Lemma 3.10.

3.4.2 Computation of the eigentuples

The MEP is polynomial in nature. In the case $n = 1$, the MEP (3.19) is a polynomial eigenvalue problem (PEP) in the variable a_0 with square coefficient matrices. As such, its eigenvalues correspond to the zeros of the *secular equation*: $\det(\boldsymbol{M}(\boldsymbol{a}_{\mathrm{tr}})) = 0$, which is a univariate polynomial of degree $3N-5$ in the variable a_0. For $n = 2$, the matrix polynomial $\boldsymbol{M}(a_0, a_1)$ has dimensions $(N-2)\times(N-3)$, such that its eigentuples could be computed as the common roots of the system of bivariate polynomial equations obtained by equating all $(N-3)\times(N-3)$ minors[6] of that matrix to zero. However, there exist better approaches for obtaining the eigentuples of an MEP. Indeed, it is well established that working with determinants is numerically unfavorable when using finite-precision arithmetic.

[5] If the first element of $\boldsymbol{v}$ equaled zero, then $\boldsymbol{g}$ must be such that $\boldsymbol{T}^{\boldsymbol{a}}_{N-n}(\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2n})^{\mathsf{T}}\boldsymbol{g} = \boldsymbol{0}$. This could only be satisfied for $\boldsymbol{g} = \boldsymbol{0}$, which contradicts $\|\boldsymbol{v}\|_2 \neq 0$.

[6] The minors of an $(N-n) \times (N-2n+1)$ matrix are the $\binom{N-n}{n-1}$ determinants of the submatrices obtained by omitting $n-1$ rows of the polynomial matrix of the MEP (3.19).

The set of (real-valued) affine eigentuples $\boldsymbol{a}_{\mathrm{tr}}$ of the MEP (3.19) can be computed numerically using the (block) Macaulay framework described in [173, 175, 176] or the methods from [90, 138]. Alternatively, one can obtain the eigentuples of the MEP by computing the set of common roots $(\boldsymbol{a}_{\mathrm{tr}}, \boldsymbol{g})$ of the system of multivariate polynomial equations (3.20) using off-the-shelf polynomial root-finding techniques, e.g., [20]. Then, the globally optimal solution(s) of the LS realization problem (3.2) can be selected by evaluating the objective function for each obtained critical point. Note that because (multiparameter) eigenvalue-solvers and polynomial root-finding techniques generally work over the field of complex numbers, the complex-valued eigentuples or common roots have to be pruned away since they have no valid interpretation in the context of the LS realization problem.

Example 3.7 (Example 3.4 continued)**.** *For this example, the system of quartic polynomial equations from* (3.20) *in the variables* $\{a_1, g_0, g_1\}$ *corresponds to*

$$\begin{cases} 0 = 2a_0g_0 + a_0^2g_1 + a_0^3g_0 - 4a_0 - 3, \\ 0 = g_0 + 2a_0g_1 + 2a_0^2g_0 + a_0^3g_1 - 3a_0 - 2, \\ 0 = g_1 + a_0g_0 + 2a_0^2g_1 - 2a_0 - 1. \end{cases} \tag{3.21}$$

Observe that the variables $\{g_0, g_1\}$ *appear only linearly. Reformulating this system of multivariate polynomial equations gives the following PEP in the parameter* a_0*:*

$$\Bigg(\underbrace{\begin{pmatrix} 3 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}}_{\boldsymbol{M}_0} + \underbrace{\begin{pmatrix} 4 & 2 & 0 \\ 3 & 0 & 2 \\ 2 & 1 & 0 \end{pmatrix}}_{\boldsymbol{M}_1} a_0 + \underbrace{\begin{pmatrix} 0 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}}_{\boldsymbol{M}_2} a_0^2 + \underbrace{\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}}_{\boldsymbol{M}_3} a_0^3 \Bigg) \begin{pmatrix} -1 \\ g_0 \\ g_1 \end{pmatrix} = \mathbf{0}. \tag{3.22}$$

The secular equation of this PEP is given as,

$$\det(\boldsymbol{M}(a_0)) = 2a_0^7 - 5a_0^6 + 12a_0^5 - a_0^4 + 6a_0^3 + 3a_0^2 + 3 = 0. \tag{3.23}$$

The seven affine common roots of the system of polynomial equations in (3.21) *are depicted in Table 3.1. In addition, the system has 41 solutions at infinity, so that the total number of solutions equals Bézout's number:* $4 \cdot 4 \cdot 3 = 48$*.*

Computing the affine eigenvalues a_0 *of the PEP from* (3.22) *or computing the roots of the univariate polynomial in* (3.23) *gives the same results. The set of*

Table 3.1: The affine common roots of the system of polynomial equations in (3.21) and corresponding cost function values $\|\widetilde{\boldsymbol{y}}\|_2^2$.

$\|\widetilde{\boldsymbol{y}}\|_2^2$	a_0	g_0	g_1
0.1486	-0.6764	-0.2525	-0.2734
/	$-0.1589 \mp 0.808j$	$1.3577 \pm 3.8194j$	$1.8359 \pm 3.3491j$
/	$0.4209 \pm 0.6233j$	$3.0425 \mp 2.9959j$	$-0.0785 \pm 1.2013j$
/	$1.3261 \pm 2.0058j$	$-0.2739 \mp 0.6279j$	$0.3793 \mp 0.3849j$

real-valued eigenvalues corresponds to a singleton: $a_0 = -0.6764$*, which is the global minimizer of the LS realization problem. It can be verified numerically that the* 3×3 *matrix obtained after substitution of* $a_0 = -0.6764$ *into the matrix polynomial* $\boldsymbol{M}(a_0)$ *in* (3.22) *has rank two.*

3.4.3 Comparison with alternatives

The observation that globally optimal LS realization of autonomous LTI models is an MEP was first shown in [45]. The author starts from the formulation in (3.14) and illustrates that the first-order necessary conditions for optimality of this unconstrained optimization problem over the model parameters $\boldsymbol{a}$ can be used to construct an MEP [45, Equation (13)]. We will denote this alternative MEP by $\boldsymbol{M}_{\text{alt}}$. The disadvantage of this approach is that many auxiliary variables have to be introduced to cope with the inverse of the matrix $\boldsymbol{D}_{N-n}^{\boldsymbol{a}}$. As a result, the coefficient matrices of $\boldsymbol{M}_{\text{alt}}$ tend to grow very large: $((N-n)(n+1)+n) \times ((N-n)(n+1)+1)$, which is approximately $n+1$ times larger than the coefficient matrices of (3.19). This becomes especially noticeable for increasing problem sizes (N, n), see, e.g., Table 3.2. As the computational complexity involved with solving an MEP depends on the interplay of its degree, the number of parameters, and the size of the coefficient matrices, the smaller coefficient matrices of our novel methodology are favorable from a computational point of view.

3.5 Model Complexity and Lower Order Models

Theorem 3.3, which is foundational for the proposed MEP, is contingent on the assumption $\operatorname{rank} \widehat{\boldsymbol{Y}} = n$. In this section, we investigate the implications of this assumption. To gain insight into the structure of the problem (3.2), we start with a formal study of the effects of increasing the model order n. We then

Table 3.2: Comparison of the sizes of the coefficient matrices of $\boldsymbol{M}_{\text{alt}}$, the MEP obtained from [45, Equation (13)], and $\boldsymbol{M}$, the MEP described in (3.19), for several combinations of (N, n).

(N, n)	$\text{size}(\boldsymbol{M}_{\text{alt}})$	$\text{size}(\boldsymbol{M})$
$(4, 1)$	7×7	3×3
$(16, 6)$	76×71	10×5
$(50, 8)$	386×379	42×35
$(200, 15)$	2975×2961	185×171

consider a special type of models that have so-called *lower order dynamics*, i.e., models for which the optimal model-compliant data is such that $\operatorname{rank} \widehat{\boldsymbol{Y}} < n$. The obtained results eventually allow us to prove Theorem 3.6.

Recall that, for a given set of model parameters $\boldsymbol{a}$, we assume the misfit and model-compliant data to be defined via the orthogonal projections in (3.12) and (3.13).

3.5.1 Increasing the model complexity

Although intuitively clear, we formalize the insight that if the globally optimal solution of (3.2) has a nonzero misfit $\|\widetilde{\boldsymbol{y}}\|_2 > 0$, increasing the complexity of the to-be-realized model (e.g., $n \to n+1$) results in a strict decrease of the global minimum.

Lemma 3.8. *For given data $\boldsymbol{y}$ and poles $\{p_i\}_{i=1,\dots,n}$ of an nth order model, we have that*

$$\|\widetilde{\boldsymbol{y}}_{\boldsymbol{n+1}}\|_2 \leq \|\widetilde{\boldsymbol{y}}_{\boldsymbol{n}}\|_2 \neq 0, \tag{3.24}$$

where $\widetilde{\boldsymbol{y}}_{\boldsymbol{n}}$ is the optimal misfit corresponding to the given nth order model and $\widetilde{\boldsymbol{y}}_{\boldsymbol{n+1}}$ is the optimal misfit corresponding to the $(n+1)$th-order model obtained by adding an arbitrary pole $p_{n+1} \in \mathbb{R}$ to the set of poles[7] of the given nth-order model. Moreover, equality holds in (3.24) *if and only if p_{n+1} is such that*

$$\boldsymbol{y} \perp \left[\boldsymbol{I} - \boldsymbol{V}[(\boldsymbol{V})^{\mathsf{H}}\boldsymbol{V}]^{-1}(\boldsymbol{V})^{\mathsf{H}}\right] \boldsymbol{p}_{\boldsymbol{n+1}}, \tag{3.25}$$

where the image of $\boldsymbol{V} \in \mathbb{C}^{N \times n}$ is the behavior of the nth order model, as defined in (3.7)*, and $\boldsymbol{p}_{\boldsymbol{n+1}}$ is the Vandermonde vector* (3.6) *defined by p_{n+1}.*

Proof. Denote by $\widehat{\boldsymbol{y}}_{\boldsymbol{n}}$ the optimal model-compliant data associated to the given nth-order model and data $\boldsymbol{y}$, and define $\boldsymbol{u}_{\boldsymbol{n+1}} \in \mathbb{R}^N$ as the component of $\boldsymbol{p}_{\boldsymbol{n+1}}$

[7] If p_{n+1} is already a pole of the nth-order model, then $\boldsymbol{p}_{\boldsymbol{n+1}}$ must be taken as the appropriate confluent Vandermonde vector, such that also in this case $\boldsymbol{p}_{\boldsymbol{n+1}} \notin \operatorname{range}(\boldsymbol{V})$.

that lies orthogonal to range $(\boldsymbol{V})$, i.e.,

$$\boldsymbol{u_{n+1}} = \left[\boldsymbol{I} - \boldsymbol{V}[(\boldsymbol{V})^{\mathsf{H}}\boldsymbol{V}]^{-1}(\boldsymbol{V})^{\mathsf{H}}\right]\boldsymbol{p_{n+1}}.$$

Then, it follows from the orthogonal projection in (3.16) that if the model order is increased by adding a new pole p_{n+1} to the model, the new model-compliant data $\widehat{\boldsymbol{y}}_{\boldsymbol{n+1}}$ can be expressed as

$$\widehat{\boldsymbol{y}}_{\boldsymbol{n+1}} = \widehat{\boldsymbol{y}}_{\boldsymbol{n}} + \frac{(\boldsymbol{y} - \widehat{\boldsymbol{y}}_{\boldsymbol{n}})^{\mathsf{H}}\boldsymbol{u_{n+1}}}{(\boldsymbol{u_{n+1}})^{\mathsf{H}}\boldsymbol{u_{n+1}}}\boldsymbol{u_{n+1}}. \tag{3.26}$$

Since $\boldsymbol{u_{n+1}} \perp \widehat{\boldsymbol{y}}_{\boldsymbol{n}} \in \text{range}\,(\boldsymbol{V})$, (3.26) implies

$$\|\widehat{\boldsymbol{y}}_{\boldsymbol{n+1}}\|_2^2 = \|\widehat{\boldsymbol{y}}_{\boldsymbol{n}}\|_2^2 + \left\|\frac{\boldsymbol{y}^{\mathsf{H}}\boldsymbol{u_{n+1}}}{(\boldsymbol{u_{n+1}})^{\mathsf{H}}\boldsymbol{u_{n+1}}}\boldsymbol{u_{n+1}}\right\|_2^2,$$

which, knowing that $\|\boldsymbol{y}\|_2^2 = \|\widehat{\boldsymbol{y}}_{\boldsymbol{n+1}}\|_2^2 + \|\widetilde{\boldsymbol{y}}_{\boldsymbol{n+1}}\|_2^2$, establishes the first part of the proof. Secondly, since Vandermonde vectors $\boldsymbol{p_i}, \boldsymbol{p_j}$ generated by different points $p_i \neq p_j$ are linearly independent, $\boldsymbol{p_{n+1}} \notin \text{range}\,(\boldsymbol{V})$. Consequently, the vector $\boldsymbol{u_{n+1}}$ is guaranteed to be nonzero, such that $\widehat{\boldsymbol{y}}_{\boldsymbol{n+1}} = \widehat{\boldsymbol{y}}_{\boldsymbol{n}}$ arises only if (3.25) is satisfied. □

A graphical illustration of the effects of the condition (3.25) in Lemma 3.8 is given in Figure 3.3.

Corollary 3.9. *Let*

$$s_n = \min_{\boldsymbol{a}\in\mathbb{R}^n} \boldsymbol{y}^{\mathsf{T}}(\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{T}}(\boldsymbol{D}_{N-n}^{\boldsymbol{a}})^{-1}\boldsymbol{T}_{N-n}^{\boldsymbol{a}}\boldsymbol{y}$$

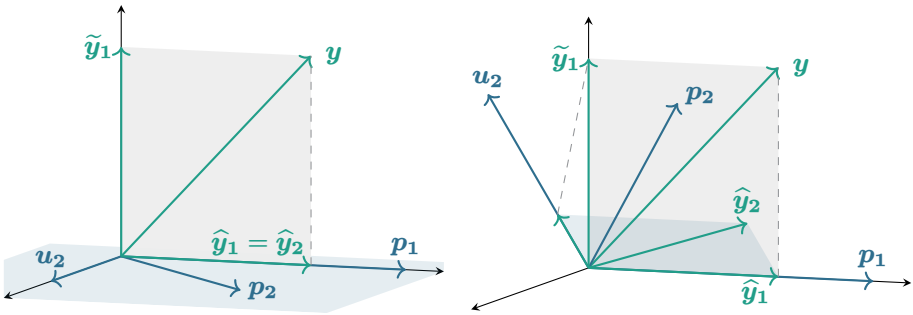
correspond to the global minimum for a model order n. Then,

$$s_n \neq 0 \iff s_{n+1} < s_n.$$

Proof. It follows from Lemma 3.8 that the set of poles p_{n+1} for which equality holds in (3.24), consist of finitely many points. Indeed, by the geometry of the problem ($\boldsymbol{y} = \widehat{\boldsymbol{y}}_{\boldsymbol{n}} + \widetilde{\boldsymbol{y}}_{\boldsymbol{n}}$ with $\widehat{\boldsymbol{y}}_{\boldsymbol{n}} \perp \widetilde{\boldsymbol{y}}_{\boldsymbol{n}}$ and $\widehat{\boldsymbol{y}}_{\boldsymbol{n}} \in \text{range}\,(\boldsymbol{V})$), (3.25) can be shown to be equivalent to

$$\begin{aligned} &\widetilde{\boldsymbol{y}}_{\boldsymbol{n}} \perp \left[\boldsymbol{I} - \boldsymbol{V}[(\boldsymbol{V})^{\mathsf{H}}\boldsymbol{V}]^{-1}(\boldsymbol{V})^{\mathsf{H}}\right]\boldsymbol{p_{n+1}}, \\ \iff &\widetilde{\boldsymbol{y}}_{\boldsymbol{n}} \perp \boldsymbol{p_{n+1}}. \end{aligned} \tag{3.27}$$

Define $\widetilde{y}_n(z)$ as the $(N-1)$th degree polynomial generated by the nonzero misfit $\widetilde{\boldsymbol{y}}_{\boldsymbol{n}}$: $\widetilde{y}_n(z) = \widetilde{y}_0 + \widetilde{y}_1 z + \widetilde{y}_2 z^2 + \cdots + \widetilde{y}_{N-1}z^{N-1}$, then (3.27) is satisfied



(a) Non-contributing $\boldsymbol{p_2}$ (b) Contributing $\boldsymbol{p_2}$

Figure 3.3: Graphical illustration of the two scenarios described in Lemma 3.8 in the case where a new pole p_2 is added to an existing first-order model with pole p_1, with $N = 3$. Within the orthogonal coordinate system (black), the plane depicted in blue corresponds to the behavior of the obtained 2nd order model. The vector $\boldsymbol{u_2}$ is the residual of the orthogonal projection of $\boldsymbol{p_2}$ onto $\boldsymbol{p_1}$, such that $\boldsymbol{u_2} \perp \boldsymbol{p_1}$. The model-compliant data $\widehat{\boldsymbol{y}}_\mathbf{2}$ correspond to the sum of the projection of $\boldsymbol{y}$ onto $\boldsymbol{p_1}$ and the projection of $\widetilde{\boldsymbol{y}}_\mathbf{1}$ onto $\boldsymbol{u_2}$.

exclusively by Vandermonde vectors constructed from the roots of this univariate polynomial: for each root p_i of $\widetilde{y}_n(z)$, $\widetilde{y}_n(p_i) = \boldsymbol{p}_i^{\mathsf{H}} \widetilde{\boldsymbol{y}}_\boldsymbol{n} = 0$ by definition, where we used that complex roots appear in complex-conjugate pairs. Consequently, as long as the misfit $\widetilde{\boldsymbol{y}}_\boldsymbol{n}$ is nonzero, one can always select a value for p_{n+1} so that (3.25) does not hold. □

Note that when p_{n+1} is such that (3.25) is satisfied, the mode $\boldsymbol{p_{n+1}}$ does not contribute to the new model-compliant data. As such, we can express $\widehat{\boldsymbol{y}}_{\boldsymbol{n+1}}$ in this case as

$$\widehat{\boldsymbol{y}}_{\boldsymbol{n+1}} = \begin{bmatrix} \boldsymbol{p_1} & \cdots & \boldsymbol{p_n} & \boldsymbol{p_{n+1}} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{x}} \\ 0 \end{bmatrix} = \widehat{\boldsymbol{y}}_\boldsymbol{n}, \tag{3.28}$$

where we used the image representation from (3.7). By consequence, the Hankel matrix $\widehat{\boldsymbol{Y}}' \in \mathbb{R}^{N-(n+1)\times n+2}$ constructed from $\widehat{\boldsymbol{y}}_{\boldsymbol{n+1}}$ is at least rank-two deficient.

3.5.2 Lower-order dynamics

The derivation leading to (3.28) provides insight into the structure of models with lower-order dynamics, that is, models for which the optimal model-compliant data is such that rank $\widehat{\boldsymbol{Y}} < n$. Consider such an nth-order model, then it follows

from Lemma 3.8 that

$$\operatorname{rank}\widehat{\boldsymbol{Y}} < n \iff \exists j \in \{1,\dots,n\} \mid \\ \boldsymbol{y}^{\mathsf{H}}\left[\boldsymbol{I} - \boldsymbol{V}'[(\boldsymbol{V}')^{\mathsf{H}}\boldsymbol{\Lambda}'_{\boldsymbol{N}}]^{-1}(\boldsymbol{\Lambda}'_{\boldsymbol{N}})^{\mathsf{H}}\right]\boldsymbol{p_j} = 0, \quad (3.29)$$

where $\boldsymbol{\Lambda}'_{\boldsymbol{N}} \in \mathbb{C}^{N\times(n-1)}$ is the Vandermonde matrix constructed from all but one of the poles $\{p_i\}_{i=1,\dots,n|i\neq j}$.

Confining our attention to the critical points of the LS realization problem (3.2), we can use (3.29) to show that a critical point with lower-order dynamics, such that Theorem 3.3 might not be applicable for this model, cannot be a locally minimizing solution of (3.2).

Lemma 3.10. *Given the parameters $\boldsymbol{a}_{\mathrm{tr}}$ of an nth order model that is a (locally) minimizing critical point of* (3.2)*, then the corresponding optimal model-compliant data $\widehat{\boldsymbol{y}}$ have exactly nth order LTI dynamics, i.e.,* $\operatorname{rank}\widehat{\boldsymbol{Y}} = n$.

Proof. The proof is by contradiction. Suppose that the model $\boldsymbol{a}$ is a (locally) minimizing critical point of (3.2), for which the corresponding model-compliant data $\widehat{\boldsymbol{y}}$ have $\operatorname{rank}\widehat{\boldsymbol{Y}} < n$. Then, as in (3.29), the latter implies that at least one of the modes[8], take $\boldsymbol{p_j} \in \mathbb{R}^N$, of the model $\boldsymbol{a}$ does not contribute to the model-compliant data $\widehat{\boldsymbol{y}}$. If we now locally perturb the coefficients of $\boldsymbol{a}$ so that this pole $p_j \to p_{j,\epsilon} = p_j + \epsilon$, with $\epsilon \in \mathbb{R}$ arbitrarily small and so that

$$\boldsymbol{y}^{\mathsf{H}}\left[\boldsymbol{I} - \boldsymbol{V}'[(\boldsymbol{V}')^{\mathsf{H}}\boldsymbol{V}']^{-1}(\boldsymbol{V}')^{\mathsf{H}}\right]\boldsymbol{p_{j,\epsilon}} \neq 0,$$

while keeping the other poles of the model unchanged, then, the pole $p_{j,\epsilon}$ will contribute to the obtained model-compliant data $\widehat{\boldsymbol{y}}_{\boldsymbol{\epsilon}}$.

By Lemma 3.8, the misfit $\widetilde{\boldsymbol{y}}_{\boldsymbol{\epsilon}}$ of this novel model, which is arbitrarily close to the given lower-order model, is guaranteed to have a smaller 2-norm than the misfit $\widetilde{\boldsymbol{y}}$ associated to the parameters $\boldsymbol{a}_{\mathrm{tr}}$. In other words, it is possible to locally perturb the model parameters $\boldsymbol{a}_{\mathrm{tr}}$ so that the objective function value decreases, which contradicts the assumption that the lower-order model defined by $\boldsymbol{a}_{\mathrm{tr}}$ is a (local) minimizer. This completes the proof. □

Lemma 3.10 implies that critical points of the LS realization problem (3.2) with lower-order dynamics need to correspond to saddle points or (local) maximizers.

[8]We assume for simplicity that the pole p_j is real-valued. The proof, however, can easily be generalized to the complex-valued case: if p_j is complex-valued, then also the mode corresponding to the complex conjugate $\overline{p_j}$ does not contribute to the model-compliant data $\widehat{\boldsymbol{y}}$. In this case, both modes $\boldsymbol{p_j}$ and $\overline{\boldsymbol{p_j}}$ have to be perturbed simultaneously.

Extreme examples of these lower-order critical points are the *global maximizers* of the LS realization problem (3.14). For these model parameters $\boldsymbol{a}$, the given data $\boldsymbol{y}$ lie completely in the row space of $\boldsymbol{T}_{N-n}^{\boldsymbol{a}}$, such that the orthogonal projections from (3.12) and (3.13) result in $\widetilde{\boldsymbol{y}} = \boldsymbol{y}$ and $\widehat{\boldsymbol{y}} = \boldsymbol{0}$, respectively.

***PROOF of* Theorem 3.6.** *The claim is a direct consequence of Lemma 3.5 and Lemma 3.10.*

Note that (3.18) and consequently Theorem 3.3 are generically[9] not satisfied for these lower-order critical points. Consider for example the above-mentioned global maximizers: in this case $\operatorname{rank} \widehat{\boldsymbol{Y}} = 0$, such that the optimality constraint (3.8c) does not put any restriction on the vector $\boldsymbol{\ell}$. As such, these lower order critical points are generally not included in the set of eigentuples of the proposed MEP (3.19), since the derivation of the MEP is contingent on the characterization given in Theorem 3.3.

Example 3.11 (Example 3.7 continued)**.** *Consider the model* $a_0 = 1.6506$*, which is a global maximizer of the LS realization problem* (3.14)*. For this parameter value,*

$$\boldsymbol{y} = \begin{bmatrix} 4 & 3 & 2 & 1 \end{bmatrix}^{\mathsf{T}} \in \operatorname{row}\left(\boldsymbol{T}_3^{\boldsymbol{a}}|_{a_0=1.6506}\right),$$

such that the misfit $\widetilde{\boldsymbol{y}}$ *obtained with the projection* (3.12) *is equal to the given data:* $\widetilde{\boldsymbol{y}} = \boldsymbol{y}$*. One can verify numerically that the characterization in* (3.17) *from Theorem 3.3 does not hold for this global maximizer:*

$$\widetilde{\boldsymbol{y}} = \boldsymbol{y} = \begin{bmatrix} 4 & 3 & 2 & 1 \end{bmatrix}^{\mathsf{T}} \notin \operatorname{range}\left(\left[(\boldsymbol{T}_3^{\boldsymbol{a}})^{\mathsf{T}}(\boldsymbol{T}_2^{\boldsymbol{a}})^{\mathsf{T}}\right]_{a_0=1.6506}\right).$$

3.5.3 Computation of all critical points

As the MEP $\boldsymbol{M}_{\text{alt}}$ from [45] does not exploit Theorem 3.3–it follows directly from the first-order necessary conditions for optimality (3.8)–its real-valued eigentuples are in one-to-one correspondence to the set of all critical points of (3.2). This is in contrast to the proposed MEP (3.19), for which we know from Section 3.5.2 that some saddle points and/or maximizers of the problem do not appear as eigentuples. As such, $\boldsymbol{M}_{\text{alt}}$ could be advantageous if one is interested in obtaining a complete characterization of the search space. However, this also

[9] Even though these lower-order critical points do not satisfy (3.17) when generic data $\boldsymbol{y}$ are used, nothing forbids them to 'coincidentally' do so. One can easily construct such an example: take an arbitrary set of model parameters $\boldsymbol{a}$ and construct the data $\boldsymbol{y}$ as $\boldsymbol{y} = (\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{T}}(\boldsymbol{T}_{N-2n}^{\boldsymbol{a}})^{\mathsf{T}}\boldsymbol{f}$ for some $\boldsymbol{f} \in \mathbb{R}^{2N-n}$. Then, for this sequence $\boldsymbol{y}$, $\boldsymbol{a}$ is a global maximizer of (3.14), corresponding to the 0th order solution $\widehat{\boldsymbol{y}} = \boldsymbol{0}$, which nevertheless satisfies (3.18).

implies that one can expect the number of (real-valued) eigentuples of $\boldsymbol{M}_{\text{alt}}$ to be significantly larger than for the proposed MEP. To substantiate this claim, consider the following example. Define $y(z)$ as the polynomial generated by the given data $\boldsymbol{y}$: $y(z) = y_0 + y_1 z + y_2 z^2 + \cdots + y_{N-1} z^{N-1}$. To each root p_i of $y(z)$, corresponds a Vandermonde vector $\boldsymbol{p_i}$ such that $\boldsymbol{y}^{\mathsf{T}} \boldsymbol{p_i} = 0$. As such, by selection of n roots from $y(z)$ as the poles of a model, one can construct $\binom{N-1}{n}$ models for which $\widehat{\boldsymbol{y}}$ obtained via (3.13) is the zero vector. The parameters of each of these models will appear as a (possibly complex-valued[10]) eigentuple of the MEP $\boldsymbol{M}_{\text{alt}}$. A similar argument can be made to characterize the number of lower-order critical points with rank $\widehat{\boldsymbol{Y}} = m$, where $0 < m < n$.

Example 3.12 (Example 3.11 continued)**.** *When we use the MEP $\boldsymbol{M}_{alt}$ from [45, Equation (13)] for $\boldsymbol{y} = [4, 3, 2, 1]^{\mathsf{T}}$ and $n = 1$, 10 affine eigenvalues a_0 are retrieved, 7 of which correspond to the affine eigenvalues of the PEP* (3.22) *depicted in Table 3.1. The three additional eigenvalues* $\{1.6506, 0.1747 \pm 1.5469j\}$ *correspond to the roots of the polynomial: $y(z) = 4 + 3z + 2z^2 + z^3$. The model with $a_0 = 1.6506$ is a global maximizer of* (3.14)*.*

3.6 Numerical Examples

We consider several numerical examples to illustrate the results of Section 3.4. Numerical values are rounded for displaying purposes. A MATLAB implementation of the (block) Macaulay method [173, 175, 176], available online at `www.macaulaylab.net`, has been used to compute the eigentuples of the obtained MEPs.

Example 3.13. *In this example we fit a second-order model ($n = 2$) to given data $\boldsymbol{y}$ ($N = 16$) with $\|\boldsymbol{y}\|_2^2 = 1.0323$:*

$$\begin{aligned}\boldsymbol{y} = [&-0.3845, -0.2211, -0.1000, -0.1976, -0.6810,\\ &-0.1417, -0.2469, -0.0766, -0.0595, 0.2074, 0.0506,\\ &-0.1412, 0.0690, 0.2549, -0.0975, -0.2959]^{\mathsf{T}}.\end{aligned}$$

The MEP (3.19) *has 739 affine eigentuples, 15 of which are real-valued (see Table 3.3). The globally optimal solution $(a_0, a_1) = (9.7392, -12.692)$ has an objective function value approximately equal to* 0.4985*. Notice from Table 3.3 that*

[10] When $y(z)$ has complex-valued roots, a subset of the $\binom{N-1}{n}$ models will have complex-valued parameters $\boldsymbol{a}$. Such a 'complex-valued' model has no valid interpretation in the context of the identification problem (3.2).

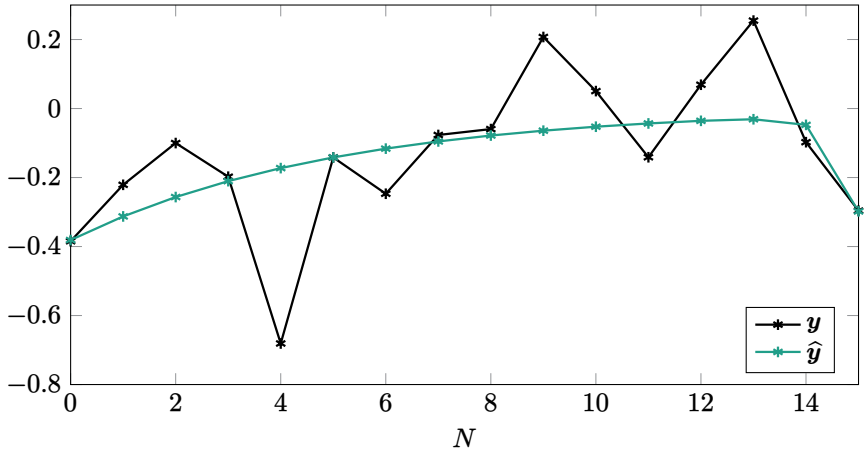


Figure 3.4: The given data $\boldsymbol{y}$ and optimal model-compliant data $\widehat{\boldsymbol{y}}$ corresponding to the global minimizer of Example 3.13.

the matrix $\widehat{\boldsymbol{Y}}$ corresponding to each critical point is of rank n, indicating that no lower-order critical points are retrieved (cf. the discussion in Section 3.5.2). The globally optimal model-compliant data $\widehat{\boldsymbol{y}}$ are depicted in Figure 3.4. The objective function value for model parameters (a_0, a_1) in the interval $[-1.5, 1.5] \times [-2, 2]$ and corresponding critical points are depicted in Figure 3.2. Notice that only a subset (including all minimizers) of the critical points appear as eigentuples of the proposed MEP (3.19).

Example 3.14. *We fit models for $n = 1, 2, 3, 4, 5$ to a sequence of given data $\boldsymbol{y}$ ($N = 10$) that is generated by a fifth-order autonomous LTI model with poles $(0.5, 0.25 \pm 0.75j, -0.3 \pm 0.5j)$:*

$$\begin{aligned}\boldsymbol{y} = [&-0.0001, -0.0963, 0.8035,\\ &2.1025, -1.0864, -1.2530, 0.1556,\\ &0.6388, 0.2978, -0.2089]^{\mathsf{T}}.\end{aligned}$$

The objective function value of the globally optimal model and the number of real-valued affine eigentuples of the MEP are depicted in Table 3.4. For $n = 5$, the MEP has one real-valued eigentuple. It corresponds to the model parameters

Table 3.3: The objective function value and poles p_i for the critical points (a_0, a_1) of Example 3.13 obtained using the MEP from (3.19). The rightmost column depicts the next to smallest singular value of the model-compliant data Hankel matrix $\widehat{\boldsymbol{Y}}$. The nonzero values indicate that rank $\widehat{\boldsymbol{Y}} = n$ for all the obtained critical points, illustrating that none of the retrieved models have lower-order dynamics (cf. Section 3.5.2).

$\|\widetilde{\boldsymbol{y}}\|_2^2$	a_0	a_1	$\{p_1, p_2\}$	$\sigma_n(\widehat{\boldsymbol{Y}})$
0.4985	9.7392	−12.692	11.871, 0.8204	0.2361
0.5328	0.6882	−1.5946	$0.7973 + 0.2292j$	0.1605
0.5381	−0.7755	0.0679	−0.9153, 0.8473	0.2786
0.6846	1.0007	0.3508	$-0.1754 \pm 0.9849j$	0.5499
0.8478	−11.401	−11.632	12.541, -0.9091	0.2809
0.8479	0.1371	0.1578	$-0.0789 \pm 0.3618j$	0.1563
0.8697	−5.1738	−7.9710	8.5744, -0.6034	0.2718
0.9118	0.7206	−0.5376	$0.2688 \pm 0.8052j$	0.2696
0.9134	3.4371	−2.7291	$1.3646 \pm 1.2550j$	0.1560
0.9139	−1.1719	−0.5126	1.3688, -0.8561	0.2378
0.9147	23.256	16.548	−14.998, −1.5506	0.2023
0.9271	−0.8285	−0.7668	1.3711, -0.6043	0.2278
0.9829	0.9939	1.2420	$-0.6210 \pm 0.7799j$	0.2388
1.0293	0.9413	1.6045	$-0.8023 \pm 0.5456j$	0.0428
1.0308	0.9175	−1.2470	$0.6235 \pm 0.7271j$	0.0396

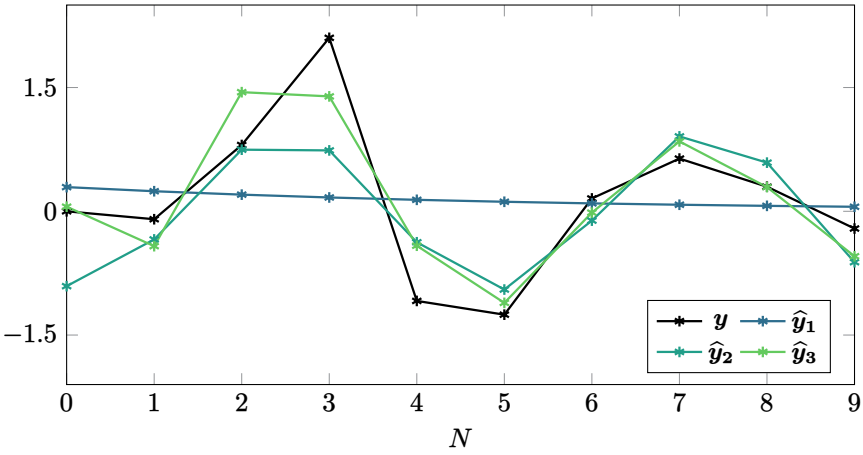


Figure 3.5: The given data $\boldsymbol{y}$ and the globally optimal model-compliant data $\widehat{\boldsymbol{y}}$ corresponding to the globally optimal solution(s) of Example 3.14. Note that $\boldsymbol{y} = \widehat{\boldsymbol{y}}_{\boldsymbol{5}} \approx \widehat{\boldsymbol{y}}_{\boldsymbol{4}}$.

Table 3.4: The global minimum and the number of real-valued affine eigentuples n_a corresponding to $n = 1, 2, \ldots, 5$ in Example 3.14.

n	1	2	3	4	5
$\|\widetilde{\boldsymbol{y}}\|_2^2$	8.12	3.74	1.68	6.06×10^{-4}	$\mathcal{O}(10^{-31})$
n_a	5	10	17	13	1

that were used to generate $\boldsymbol{y}$*, i.e., the coefficients of the fifth-degree polynomial:*

$$a(z) = (z - 0.5)(z - 0.25 + 0.75j)$$
$$(z - 0.25 - 0.75j)(z + 0.3 + 0.5j)(z + 0.3 - 0.5j).$$

The number of real-valued affine eigentuples n_a *peaks at* $n = 3$*, after which it decreases as* n *approaches the model order with which the given data has been generated. The model-compliant data corresponding to the globally optimal solution(s)* $\{\widehat{\boldsymbol{y}}_{\boldsymbol{n}}\}_{n=1,\ldots,5}$ *are depicted in Figure 3.5.*

Example 3.15. *We fit a third-order model (*$n = 3$*) to given data* $\boldsymbol{y}$ *(*$N = 30$*), which is constructed as:*

$$\boldsymbol{y} = \boldsymbol{y}_{3rd} + \alpha \cdot \texttt{randn(N,1)}, \tag{3.30}$$

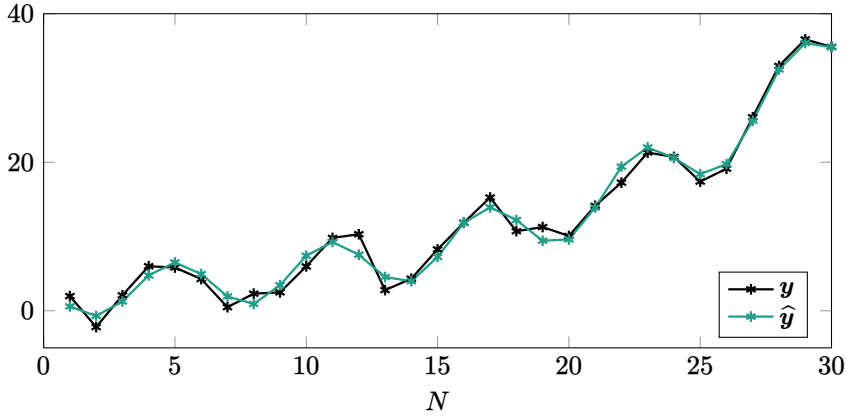


Figure 3.6: The given data $\boldsymbol{y}$ and the globally optimal model-compliant data $\widehat{\boldsymbol{y}}$ corresponding to the globally optimal solution of Example 3.15.

where $\alpha = 1.5$ *and* $\boldsymbol{y}_{3rd}$ *is generated by a third-order model with poles* $\{1.1, 0.5 \pm 0.9j\}$ *and initial conditions:* $y_{3rd,0} = 1$, $y_{3rd,1} = 0$, $y_{3rd,2} = 1.5$. *To make the results reproducible, the random number generator used by MATLAB's* `randn()` *function was seeded via* `rng(10)`. *Solving the system of equations* (3.20) *using the software package*[11] *from [20] allows us to retrieve 45 critical points. The parameter values of (some of) these critical points are depicted in Table 3.5. The global minimizer has an objective function value of* $\|\widetilde{\boldsymbol{y}}\|_2^2 = 41.3091$ *(compared to* $\|\boldsymbol{y}\|_2^2 = 7574.04$*). The obtained data sequence is visualized in Figure 3.6.*

Based on the objective function values, the set of critical points can be divided into three groups: 1) the global minimizer, 2) 24 points for which $225 < \|\widetilde{\boldsymbol{y}}\|_2^2 < 255$ *and 3) 20 critical points with objective function value close to the global maximum* $\|\boldsymbol{y}\|_2^2$. *Interestingly, Figure 3.7 shows that the critical points of the first and second group seem to lie on a plane in the ambient space* $\mathbb{R}^3$. *Closer observation learns that this plane spans all the coefficients of third-order monic polynomials* $a(z)$ *for which (at least) one of the roots lies close to 1.1, which is the dominant pole*[12] *of the model used to generate* $\boldsymbol{y}_{3rd}$. *This 'clustering effect' of the critical points arises due to the high signal-to-noise level present in the given data* $\boldsymbol{y}$. *The effect becomes less outspoken when the example is repeated for larger values of* α *in* (3.30).

[11] Contrary to the block Macaulay method [173], homotopy continuation methods can exploit *parallelism* in the computations, allowing to tackle larger problems.

[12] The dominant pole (pair) of a model $\boldsymbol{a}$ corresponds to the root(s) of $a(z)$ with the largest magnitude.

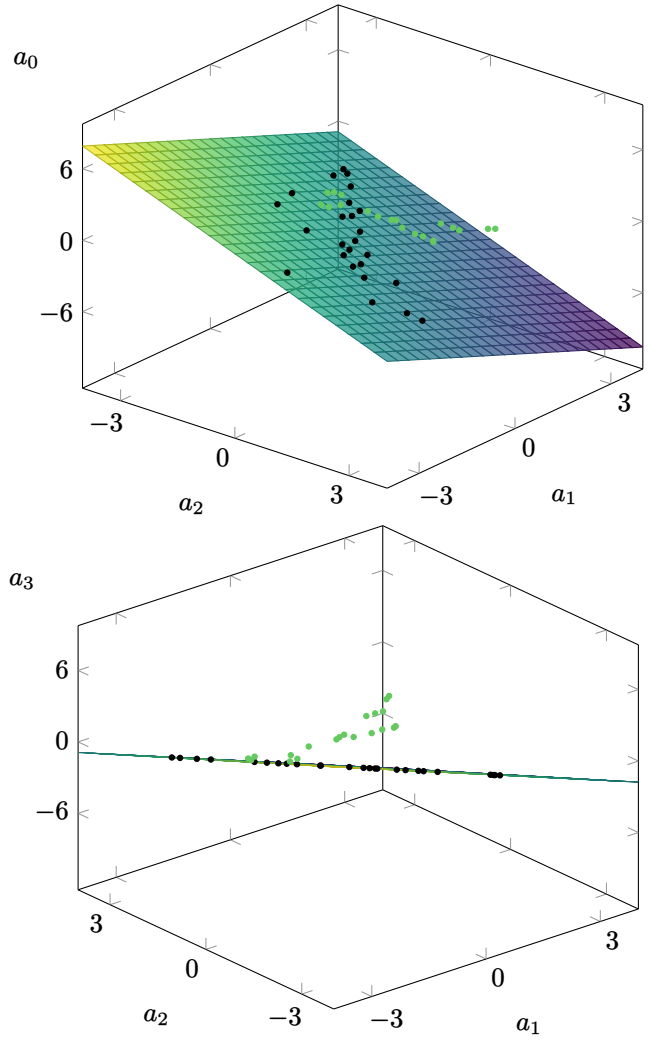


Figure 3.7: Scatter plots of the parameters values (a_0, a_1, a_2) of the critical points of Example 3.15. The critical points depicted in black (blue) have an objective function value for which $\|\widetilde{\boldsymbol{y}}\|_2^2 < 255$ (resp. $\|\widetilde{\boldsymbol{y}}\|_2^2 > 7500$). The plane depicts all parameter values for which (at least) one of the roots of the polynomial $a(z) = a_0 + a_1 z + a_2 z^2 + z^3$ is equal to 1.1. The plots illustrate that the best-performing critical points have one pole approximating the dominant pole of the model used to generate the data $\boldsymbol{y}$.

Table 3.5: The objective function value and the dominant pole (pair) p_{d} for (some of) the critical points (a_0, a_1, a_2) of Example 3.15, obtained using the system of polynomial equations from (3.20).

$\|\widetilde{\boldsymbol{y}}\|_2^2$	a_0	a_1	a_2	p_{d}	
41.309	-1.132	2.120	-2.091	1.100	25 critical points
225.73	-1.084	0.950	-1.079	1.106	
$\vdots$					
251.56	-1.307	-1.217	1.063	1.106	
253.94	-1.113	0.023	-0.215	1.105	
7504.5	0.147	0.420	-1.443	$0.820 \pm 0.267j$	20 critical points
7507.0	-0.172	1.124	-1.870	$0.8189 \pm 0.271j$	
$\vdots$					
7566.3	-0.128	0.691	0.458	$-0.310 \pm 0.834j$	
7567.4	-0.150	0.546	1.395	$-0.788 \pm 0.458j$	

3.7 Weighted Low-Rank Approximation of Hankel Matrices

This section extends the globally optimal framework presented before to the setting of weighted structured low-rank approximation (SLRA) for Hankel matrices. In particular, we incorporate arbitrary positive weights into the objective function, thereby generalizing the previously unweighted formulation. From an engineering perspective, this extension is particularly relevant in scenarios where certain data samples carry different levels of confidence or importance.

The section is organized as follows. We formulate the weighted Hankel low-rank approximation problem (Section 3.7.1) and review the relevant literature (Section 3.7.2). The generalized methodology is then described (Section 3.7.3), where we focus on the main difference introduced by the incorporation of weights. Several illustrative numerical examples highlighting the effect of non-uniform weighting are presented in Section 3.7.4.

3.7.1 Problem formulation

Structured low-rank matrix approximation

Given a sequence of data $\boldsymbol{y} \in \mathbb{R}^N$, the *weighted structured low-rank matrix approximation problem* refers to the following optimization problem:

$$\underset{\widehat{\boldsymbol{y}} \in \mathbb{R}^N}{\text{minimize}} \quad \|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|_{\boldsymbol{\Lambda}}, \tag{3.31a}$$

$$\text{subject to} \quad \operatorname{rank} \mathcal{S}(\widehat{\boldsymbol{y}}) \leq r, \tag{3.31b}$$

where r is a positive integer with $r < \min(L, K)$, the positive definite matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{N \times N}$ determines the weighted norm, and

$$\mathcal{S}(\widehat{\boldsymbol{y}}) \coloneqq \sum_{i=0}^{N-1} \widehat{\boldsymbol{y}}_i \boldsymbol{S_i} + \boldsymbol{S_N},$$

which is defined by the linearly independent basis matrices $\boldsymbol{S_i} \in \mathbb{R}^{L \times K}$ for $i \in \{0, \ldots, N\}$, organizes the elements of the vector $\widehat{\boldsymbol{y}} = [\widehat{\boldsymbol{y}}_0, \ldots, \widehat{\boldsymbol{y}}_{N-1}]^{\mathsf{T}} \in \mathbb{R}^N$ into an (affinely) structured matrix. Several well-known examples of affine matrix structures include Hankel matrices, circulant matrices, Toeplitz matrices, centrosymmetric matrices, matrices with a certain sparsity pattern and combinations of the former, such as block-Toeplitz matrices.

The goal of the optimization problem is to compute the vector $\widehat{\boldsymbol{y}} \in \mathbb{R}^N$, approximating the given data $\boldsymbol{y}$, for which the affinely structured matrix $\mathcal{S}(\widehat{\boldsymbol{y}})$ is low-rank while minimizing a predefined cost function that quantifies the "distance" between the given data and its approximation. Although the usual 2-norm is often sufficient, several applications require weighted variants:

$$\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|_{\boldsymbol{\Lambda}}^2 = (\boldsymbol{y} - \widehat{\boldsymbol{y}})^{\mathsf{T}} \boldsymbol{\Lambda} (\boldsymbol{y} - \widehat{\boldsymbol{y}}),$$

where $\boldsymbol{\Lambda} \in \mathbb{R}^{N \times N}$ is symmetric, positive definite. In practice, the weights are often determined based on prior knowledge or application-specific considerations. For example, the matrix $\boldsymbol{\Lambda}$ can be chosen as the inverse covariance matrix of the noise that is expected to have perturbed the measurements $\boldsymbol{y}$. When $\boldsymbol{\Lambda}$ is diagonal, we get the element-wise weighted norm:

$$\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|_{\boldsymbol{\lambda}}^2 = \sum_{i=0}^{N-1} \lambda_i (y_i - \widehat{y}_i)^2, \tag{3.32}$$

with $\boldsymbol{\lambda} = \operatorname{diag} \boldsymbol{\Lambda} = [\lambda_0, \ldots, \lambda_{N-1}]^{\mathsf{T}} \in \mathbb{R}^N$.

One could alternatively define the problem (3.31) using a matrix expression in the objective function:

$$\|\mathcal{S}(\boldsymbol{y}) - \mathcal{S}(\widehat{\boldsymbol{y}})\|_{\boldsymbol{W}}^2 = \sum_{i=1}^{L} \sum_{j=1}^{K} \boldsymbol{W}_{i,j} \left(\mathcal{S}(\boldsymbol{y})_{i,j} - \mathcal{S}(\widehat{\boldsymbol{y}})_{i,j}\right)^2 \tag{3.33}$$

where $\boldsymbol{W} \in \mathbb{R}^{L \times K}$ has positive entries only. When $\boldsymbol{W}$ is a matrix consisting of all ones, we recover the Frobenius matrix norm

$$\|\mathcal{S}(\boldsymbol{y}) - \mathcal{S}(\widehat{\boldsymbol{y}})\|_F^2 = \sum_{i=1}^{L} \sum_{j=1}^{K} \left(\mathcal{S}(\boldsymbol{y})_{i,j} - \mathcal{S}(\widehat{\boldsymbol{y}})_{i,j}\right)^2.$$

Hankel matrices

We restrict our attention to SLRA problems using *Hankel matrices*, which have identical elements along each anti-diagonal. The Hankel matrix generated by given data $\boldsymbol{y} \in \mathbb{R}^N$ is obtained by the following map $\boldsymbol{H}_L : \mathbb{R}^N \mapsto \mathbb{R}^{L \times K}$ with $L + K - 1 = N$:

$$\boldsymbol{H}_L(\boldsymbol{y}) = \begin{bmatrix} y_0 & y_1 & y_2 & \cdots & y_{K-1} \\ y_1 & y_2 & y_3 & \cdots & y_K \\ \vdots & \vdots & \vdots & & \vdots \\ y_{L-1} & y_L & y_{L+1} & \cdots & y_{N-1} \end{bmatrix} \in \mathbb{R}^{L \times K}, \tag{3.34}$$

The parameter L, which is often referred to as the *window size*, determines the size of the Hankel matrix and has to be specified by the user. One could set the parameter so that the resulting Hankel matrix is (close to) square [123, 132] or so that the matrix is as tall as possible (choose L large) [67]. It is important to note that, except for the boundary cases, the particular size of the Hankel matrix does not influence its rank, as shown by the following result.

Lemma 3.16 (Corollary 5.2 in [87])**.** *For any sequence $\widehat{\boldsymbol{y}} \in \mathbb{C}^N$ there exists a number $d \leq \left\lfloor \frac{N+1}{2} \right\rfloor$ such that*

$$\operatorname{rank} \boldsymbol{H}_L(\widehat{\boldsymbol{y}}) = \min\left(L, K, d\right)$$

where $\boldsymbol{H}_L(\widehat{\boldsymbol{y}})$ is defined as in (3.34) *and $L + K - 1 = N$ as before.*

If $d < (N+1)/2$, the rank of the Hankel matrices $\boldsymbol{H}_L(\widehat{\boldsymbol{y}})$ reaches a plateau at d before decreasing back to one as the window length L is varied from 1 to N. In this scenario, the data sequence $\widehat{\boldsymbol{y}}$ can be represented as the impulse response of a minimal LTI system of order d (cf. Section 3.2). Consequently, such data $\widehat{\boldsymbol{y}}$ is commonly referred to as a *time series of finite rank* in the signal spectrum analysis (SSA) literature [76]. In contrast, when $d = \frac{N+1}{2}$ and N odd, the resulting Hankel matrix $\boldsymbol{H}_L(\widehat{\boldsymbol{y}})$ is full-rank for every L. The latter is always the case when the data generating the Hankel matrix are sufficiently generic.

Because of the rank-structure of Hankel matrices described by Lemma 3.16, it suffices to consider the SLRA problem (3.31) with $K = r + 1$ (i.e., $L = N - r$) to cover all other cases for which $K > r + 1$ as long as $r < \min\left(L, K\right)$ (see e.g., [68, Corollary 3.1]). As such, we will only consider problems of the form $L = N - r$. To ease the notation, we denote the $(N-r) \times (r+1)$-Hankel matrix generated by the data $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}}$ by the capital letters:

$$\boldsymbol{Y} = \boldsymbol{H}_{N-r}(\boldsymbol{y}) \quad \text{and} \quad \widehat{\boldsymbol{Y}} = \boldsymbol{H}_{N-r}(\widehat{\boldsymbol{y}}).$$

Definition 3.17 (Weighted SLRA for Hankel matrices)**.** *Given a symmetric, positive definite matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{N \times N}$, data $\boldsymbol{y} \in \mathbb{R}^N$ and a positive integer r for which the matrix $\boldsymbol{Y} \in \mathbb{R}^{(N-r)\times(r+1)}$ has full column-rank, consider the weighted SLRA problem for Hankel matrices:*

$$\underset{\widehat{\boldsymbol{y}} \in \mathbb{R}^N}{\text{minimize}} \quad \frac{1}{2} \left\| (\boldsymbol{y} - \widehat{\boldsymbol{y}}) \right\|_{\boldsymbol{\Lambda}}^2 = \frac{1}{2} (\boldsymbol{y} - \widehat{\boldsymbol{y}})^{\mathsf{T}} \boldsymbol{\Lambda} (\boldsymbol{y} - \widehat{\boldsymbol{y}}), \tag{3.35a}$$

$$\text{subject to} \quad \operatorname{rank} \widehat{\boldsymbol{Y}} \leq r. \tag{3.35b}$$

We need $N > 2r$ for (3.35) *to be well-defined, since otherwise $\boldsymbol{Y}$ has at least one linearly dependent column by construction.*

Remark 3.18. *Note that the weighted SLRA problem for Hankel matrices* (3.35) *generalizes the LS realization problem that we considered before. Indeed, when unit weights are used and the rank r corresponds to the given order n of the to be estimated LTI model,* (3.35) *is equivalent to* (3.2)*.*

Note that in the case of Hankel matrices, the 'vector-norm' (3.32) and the 'matrix-norm' (3.33) are equivalent upon the substitution:

$$\lambda_k = \sum_{i=\max(1,k+2-(r+1))}^{\min(N-r,k+1)} w_{i,k+2-i}, \quad \forall k \in \{0,\ldots,N-1\}, \tag{3.36}$$

in which we assumed the case $L = N-r$, i.e., $K = r+1$. As such, each problem with a cost function of the form (3.33) can be recast into a problem of the form (3.35). See e.g., [96] for more details.

Choice of weights

We describe several specific types of weights that will be used extensively in the remainder of this section. First, we denote *unit weights* by $\boldsymbol{\lambda_I} = [1,\ldots,1]^{\mathsf{T}} \in \mathbb{R}^N$, corresponding to the standard unweighted 2-norm. Next, we consider *generic weights*, denoted by $\boldsymbol{\lambda_G} = \texttt{rand}(N,1) \in \mathbb{R}^N$, which are randomly sampled from a uniform distribution. We also define two weight vectors that correspond to Frobenius norms derived from structured Hankel matrices. The *Frobenius weights associated with a tall Hankel structure* are given by

$$(\boldsymbol{\lambda_F})_i = \min(i+1,\, d,\, N-i), \quad \forall\, i \in \{0,\ldots,N-1\},$$

where $d = \min(N-r,\, r+1)$. Under the assumption $N > 2r$, this simplifies to $d = r+1$. The term "tall" refers to the fact that these weights arise from the Frobenius norm of the approximation error using the tall Hankel matrix structure $\boldsymbol{H}_{N-r}$ as defined in equation (3.34):

$$\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2_{\boldsymbol{\lambda_F}} = \|\boldsymbol{H}_{N-r}(\boldsymbol{y}) - \boldsymbol{H}_{N-r}(\widehat{\boldsymbol{y}})\|^2_F = \left\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\right\|^2_F.$$

In a similar fashion, the *Frobenius weights corresponding to a square (or nearly square) Hankel structure* are given by

$$(\boldsymbol{\lambda_S})_i = \min(i+1, N-i), \quad \forall\, i \in \{0,\ldots,N-1\}.$$

These weights correspond to the Frobenius norm when the Hankel structure $\boldsymbol{H}_d$ has approximately equal dimensions, with $d = N/2$ when N is even and $d = (N+1)/2$ when N is odd:

$$\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2_{\boldsymbol{\lambda_S}} = \|\boldsymbol{H}_d(\boldsymbol{y}) - \boldsymbol{H}_d(\widehat{\boldsymbol{y}})\|^2_F.$$

Lastly, we consider a class of weights introduced in [132, Section 4], which we shall refer to as the *Bombieri weights*. They are defined by

$$(\boldsymbol{\lambda_T})_i = \binom{N-1}{i}, \quad \forall\, i \in \{0, \ldots, N-1\}.$$

3.7.2 Literature overview

We provide a brief overview of the most important developments related to structured low-rank approximation, with a particular focus on the case of Hankel matrices and the incorporation of weighted norms. We begin by situating the problem within the broader context of low-rank matrix approximation and the use of Hankel matrices, before reviewing several classes of solution approaches. We conclude by discussing globally optimal methods, which aim to characterize and compute the entire set of critical points of the problem.

The literature review is partially based on the comprehensive survey provided in [68], but the discussion has been tailored to emphasize those aspects most relevant to the methodology presented in this here. Another insightful review of SLRA for Hankel matrices is provided in [123].

Low-rank matrix approximation

Low-rank matrix approximation is a fundamental problem in applied mathematics and engineering, connected to principal component analysis [153], factor analysis [157] and *total least squares* (TLS) [75, 124, 170]. In the classical TLS setting, which is also referred to as *orthogonal regression* or *errors-in-variables*, the objective is to approximate a given full-rank matrix by one of lower rank, while accounting for perturbations in all entries of the data matrix. This traces back to the the 19th century, with early contributions by Adcock and Pearson [2, 3, 135]. This unstructured and unweighted low-rank approximation problem admits a closed-form, globally optimal solution via the singular value decomposition (SVD) of the given full-rank data matrix. In particular, the optimal rank-r approximation is obtained by retaining the top r singular values and associated singular vectors, as described in the classical results of Eckart and Young [54, 92, 189], cf. the *Eckart–Young–Mirsky Theorem* [75]. This result lies at the basis of a wide range of data-driven modeling techniques in statistics and signal processing.

In recent decades, there has been considerable interest in the use of low-rank approximation techniques for problems arising in time series analysis, system identification, and forecasting. In such applications, it is often insufficient to

merely approximate the given data matrix by one of lower rank. Instead, the approximating matrix is required to have (or preserve) a prescribed algebraic structure, such as Hankel, Toeplitz, or block-symmetric. This leads to the formulation of the SLRA, as described in (3.31). The problem is also referred to as the structured or constrained TLS problem, see e.g., [1, 40, 44, 115] and references therein.

The SLRA problem is fundamentally different from the classical, unstructured low-rank approximation setup. Although the introduction of non-uniform weights in the cost function in the unstructured setting is conceptually straightforward, as first proposed by Young in the 1940s [190], it nevertheless makes the traditional SVD based algorithms inapplicable. Additionally, the presence of structural constraints significantly increases the mathematical complexity of the problem and the computational difficulty involved with obtaining accurate solutions. Unlike the unstructured and unweighted case, which admits a globally optimal closed-form solution, no such explicit solution exists for the general SLRA problem. Except for specific cases, for instance, when the structure is circulant or when certain classes of square matrices are considered [149], the SLRA problem is generally known to be NP-hard in the global optimization sense [71]. As a result, one typically resorts to heuristic methods or iterative numerical optimization techniques to compute approximate (local) minimizers.

Hankel matrices

The connection between SLRA and linear system theory naturally leads to the study of SLRA problems for *Hankel matrices*, which play a central role in system identification, signal processing, and time-series analysis. In practical scenarios, the available data are typically contaminated by measurement noise, so that the associated Hankel matrix has full rank [40]. In such cases, we want to approximate the given Hankel matrix with a rank-deficient matrix, while preserving its Hankel structure, so that we can capture the underlying low-dimensional dynamics of the given time series. Beyond measurement noise, other factors such as missing data, model mismatch (e.g., under-modeling), and nonlinearities in the signal can further contribute to the degradation of the low-rank structure.

Hankel matrices have a long and rich history in both mathematics and engineering. Early foundational work by Prony [48] and Sylvester [163] explored these matrices in the context of polynomial and binary form decompositions. Their significance was later established in linear system theory. A fundamental result in this area, which traces back to Kronecker's work on rational

approximation [105], is the fact that the rank deficiency of a Hankel matrix formed from a given time series, is equivalent to the existence of a linear time-invariant (LTI) system whose impulse response matches the given data (except for the first sample) [98]. The minimal order of such a system is exactly equal to the rank of the corresponding Hankel matrix.

Beyond their role in system theory, Hankel matrices also appear in the context of symmetric tensor decompositions, particularly in relation to moment problems and symmetric multilinear forms [101]. We refer the interested reader to [87] for a more detailed overview of the theory on Hankel matrices.

Solution approaches

Classical algorithms for solving the SLRA problem for the unit weight case have been discussed at the start of this chapter. The most prominent methods that arose from this system-theoretic perspective are iterative quadratic maximum likelihood (IQML)[21] and Steiglitz-McBride [158]. The latter, which is not restricted to Hankel matrices, was shown to be equivalent to IQML in the context of (3.35) [127]. Its convergence behavior is discussed in [162]. Related methods that predate IQML but are algorithmically equivalent include those proposed in [58], [99] and [150].

Cadzow's iteration [25] is a heuristic approach which alternates between two steps: a singular value decomposition (SVD) to compute a rank-deficient approximation of the data matrix, and a least squares fit to restore the desired matrix structure. In the case of Hankel matrices, the latter step is performed by averaging over the anti-diagonals. This alternating procedure is repeated until convergence. Although straightforward and computationally efficient, Cadzow's method, like the previously mentioned approaches, does not guarantee convergence to a locally optimal solution. This lack of optimality was explicitly demonstrated using a numerical example in [40].

Another family of methods rooted in systems theory includes SVD-based realization algorithms, which aim to reconstruct a state-space model from impulse response data. Classical examples of such methods include those by Kung [107] and by Zeiger and McEwen [191]. While primarily developed for state-space identification, these algorithms implicitly perform SLRA.

There is an extensive body of literature on the structured total least squares (STLS) problem, which in the case of Hankel matrices is equivalent to (3.35). The so-called *kernel representation* methods exploit the observation that the rank deficiency constraint (3.35b) of the Hankel matrix generated by the modified data $\widehat{\boldsymbol{y}}$ can be equivalently enforced by requiring that the data $\widehat{\boldsymbol{y}}$ lie in the null

space of a banded Toeplitz matrix, generated from the vector in the null space of the rank deficient Hankel matrix $\widehat{\boldsymbol{Y}}$. This characterization led to a nonlinear generalized singular value decomposition, the so-called *Riemannian SVD*. The solutions to this Riemannian SVD, which can be solved via an algorithm that is inspired by inverse iteration [40, 44], are the critical points of the SLRA problem. The equivalence between this nonlinear SVD-based approach and the constrained TLS solution, proposed earlier by Abatzoglou [1], was established in [115].

A continuous-time dynamical system formulation of the Riemannian SVD and its convergence behavior has also been investigated [41]. Other solution approaches in this category include [146], who proposed a linear approximation to the Toeplitz-based rank-deficiency constraint and continues by using Newton's method to obtain a solution, and [192], who proposes an iterative reweighting scheme. A modification of the STLS approach from [40] for rank reduction by more than one was described in [169], but this approach involves Kronecker products which dramatically inflate the dimension of the involved matrices.

Several approaches explicitly exploit the nested, or separable, structure of the optimization problem: the SLRA problem under consideration is a so-called *separable nonlinear LS problems*, for which part of the optimization problem can be shown to admit a closed form solution, the "projection" [74]. Substituting this projection into the original optimization problem allows to analytically eliminate a subset of the decision variables (cf Section 3.3.1). The result is an equivalent, often unconstrained, nonlinear optimization problem which can be solved using standard local optimization techniques such as Gauss-Newton and Levenberg-Marquardt [125, 167, 193]. These *variable projection* methods reduce the dimensionality of the search space and explicitly account for the problem's inherent structure.

The SLRA problem has also been formulated as a rank minimization problem, see e.g., [123] for a detailed treatment. While this alternative formulation is theoretically equivalent, it may lead to non-uniqueness of solutions.

Remark 3.19. *The globally optimal LS realization framework developed earlier in this chapter builds upon several of the aforementioned characterizations. Exploiting the dynamics of an autonomous linear time-invariant (LTI) system, the model-compliant data were shown to lie in the null space of a structured (banded Toeplitz) matrix (cf. the kernel representation). The framework also conceptually aligns with variable projection methods: the optimal misfit was obtained as the orthogonal projection of the given data onto the row space of the banded Toeplitz matrix, thereby eliminating the model-compliant data from the optimization problem.*

Global optimality

A final class of approaches attempts to address the fundamental challenge of global optimality in SLRA. For the case of the weighted Frobenius norm, it has been shown in [132, Corollary 3.9] that the number of complex solutions to the first-order necessary conditions for optimality of the low-rank approximation problem for Hankel matrices grows polynomially with the data length N and exponentially with the rank r. Furthermore, the objective function often exhibits a highly non-convex landscape with narrow basins of attraction, particularly in the presence of noise [69]. This behavior, which is evident from the numerical examples in Section 3.7.4, raises concerns regarding the performance of heuristic methods, which can guarantee local optimality at best. Even if such a heuristic method converges to a solution, it is very likely that many other locally optimal solutions exist that offer a (potentially significantly) better fit.

To mitigate these issues, several heuristic strategies have been proposed to improve the likelihood of finding better, if not globally optimal, solutions. One such method is the alternating projections with backtracking and randomization (APBR) algorithm [70], a refinement of Cadzow's iteration. This approach incorporates a backtracking mechanism to maintain proximity to the observed data, while random perturbations are used to escape poor local minima. Another direction involves convex relaxation techniques based on kernel representations, such as the approach proposed in [28], which can provably recover globally optimal solutions when the given data lies sufficiently close to a structured low-rank model. Despite the efforts, these approaches remain heuristic, as they do not provide absolute guarantees of global convergence.

Notably, the engineering literature has yet to propose a methodology for low-rank approximation of Hankel matrices that consistently achieves global optimality. The few examples in the literature that do guarantee global optimality originate from the (numerical) algebraic geometry community, where the focus lies on the so-called *Euclidean distance (ED) degree* of the SLRA problems under consideration. The ED-degree quantifies the number of complex solutions to the system of equations defined by the first-order optimality conditions of the optimization problem. For instance, [51] characterizes the ED-degree for the unweighted case using the standard Euclidean distance, while [132] extends the analysis to the weighted Frobenius norm. In both cases, the authors reparameterize a particular instance of the problem, derive a system of polynomial equations describing the critical solutions and solve the system using Gröbner-basis methods. Although the primary goal of these studies is to characterize the algebraic complexity of the problem, the computed set of critical points include the global minimizer(s), thereby establishing a globally optimal solution in principle. A detailed discussion of the ED-degree in the case

of SLRA for Hankel matrices will be presented in Chapter 6.

3.7.3 Globally optimal weighted low-rank approximation

The approach presented below builds upon the framework introduced for the unweighted globally optimal LS realization problem. As many of the underlying derivations remain structurally similar, we focus on the key differences introduced by the incorporation of weights into the cost function.

Rank deficiency via the kernel representation

A common algebraic characterization of the candidate space of rank-deficient Hankel matrices for (3.35) involves enforcing the vanishing of all minors of size $r+1$, which ensures that the matrix has rank at most r. While this determinantal formulation is theoretically correct, it is numerically unfavorable. For tall or fat Hankel matrices, the number of such minors grows combinatorially with the matrix dimensions and each minor is a polynomial of degree $r+1$ in the variables $\widehat{\boldsymbol{y}}$. As a result, the associated system of equations is typically large and contains many algebraically dependent equations, leading to ill-conditioned solution sets that are extremely sensitive to numerical perturbations. From an algebraic-geometry perspective, the solution set defined by the determinantal equations is typically highly singular, particularly near points where the rank of the Hankel matrix drops further than strictly required. This singular structure poses computational challenges for (numerical) solvers.

A numerically better alternative is to explicitly parametrize the kernel vector of the Hankel matrix by introducing auxiliary variables $\boldsymbol{a} = [a_0\ a_1\ \cdots\ a_{r-1}\ a_r]^\mathsf{T} \in \mathbb{R}^{n+1}$, which define a nonzero vector in the null space of the matrix. The rank-deficiency condition (3.35b) is then reformulated as a linear constraint:

$$\operatorname{rank} \widehat{\boldsymbol{Y}} \leq r \iff \exists\, \boldsymbol{a} \ : \ \widehat{\boldsymbol{Y}}\boldsymbol{a} = \boldsymbol{0} \text{ and } \boldsymbol{a}^\mathsf{T}\boldsymbol{e} = 1, \tag{3.37}$$

where $\boldsymbol{e} \in \mathbb{R}^{n+1}$ is some nonzero vector that determines the particular choice of normalization of the parameter vector $\boldsymbol{a}$. Typical choices include random $\boldsymbol{e}$, $\boldsymbol{e} = [0, \ldots, 0, 1]^\mathsf{T}$) or $\boldsymbol{e} = \boldsymbol{a}$ leading to the quadratic constraint $\boldsymbol{a}^\mathsf{T}\boldsymbol{a} = 1$.

The kernel representation avoids the combinatorial explosion of the determinantal formulation by replacing those many nonlinear constraints stemming from the $(r+1)$ minors with the much smaller system of (at most) quadratic equations (3.37). Moreover, it is not affected by the singularities that come with the determinantal description (see Chapter 6 for a detailed geometric analysis).

However, note that this "kernel representation" is straightforward only in the case where the Hankel matrix is rank deficient by one (i.e., the corank-1 case). For general $r > 1$, one must parametrize an r-dimensional subspace, which is substantially more involved.

The Hankel structure of $\widehat{\boldsymbol{Y}}$ together with the rank-deficiency constraint implies that the approximating data $\widehat{\boldsymbol{y}}$ is a rank-r time series, i.e., it satisfies an rth order difference equation. Although the low-rank approximation problem does not need to originate from a system theoretic context, the coefficients $\boldsymbol{a}$ can be considered as the model parameters of an autonomous LTI model that generates the data $\widehat{\boldsymbol{y}}$. Replacing the Hankel-vector product in (3.37) using the banded-Toeplitz matrix $\boldsymbol{T}_{N-r}^{\boldsymbol{a}} \in \mathbb{R}^{(N-r)\times N}$ leads to

$$\widehat{\boldsymbol{Y}}\boldsymbol{a} = \begin{bmatrix} a_r & \cdots & \cdots & a_1 & a_0 & 0 & \cdots & 0 \\ 0 & a_r & \cdots & \cdots & a_1 & a_0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_r & \cdots & \cdots & a_1 & a_0 \end{bmatrix} \widehat{\boldsymbol{y}} = \boldsymbol{T}_{N-r}^{\boldsymbol{a}}\widehat{\boldsymbol{y}} = \boldsymbol{0}. \tag{3.38}$$

The image representation remains a valid basis for the r-dimensional null space of $\boldsymbol{T}_{N-r}^{\boldsymbol{a}}$, in which $\widehat{\boldsymbol{y}}$ resides. Indeed, each root of the characteristic polynomial

$$a(z) = a_0 + a_1 z + \cdots + a_r z^r = c\prod_{i=1}^{n}(z - p_i)$$

generates a Vandermonde vector $\boldsymbol{p_i} = \begin{bmatrix} 1 & p_i & p_i^2 & \cdots & p_i^{N-1} \end{bmatrix}^{\mathsf{T}}$ such that for some $\widehat{\boldsymbol{x}} \in \mathbb{C}^n$,

$$\widehat{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{p_1} & \boldsymbol{p_2} & \cdots & \boldsymbol{p_r} \end{bmatrix} \widehat{\boldsymbol{x}}.$$

The coefficient $c \in \mathbb{R}$ in $a(z) = c\prod_{i=1}^{n}(z - p_i)$ is determined by the particular choice of normalization vector $\boldsymbol{e}$.

Using this kernel representation, the Lagrangian of the weighted low-rank Hankel matrix approximation problem (3.35) corresponds to

$$\mathcal{L}(\boldsymbol{a}, \widehat{\boldsymbol{y}}, \boldsymbol{\ell}) = \frac{1}{2}\left\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\right\|_{\boldsymbol{\Lambda}}^2 + \boldsymbol{\ell}^{\mathsf{T}}\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\widehat{\boldsymbol{y}} + \mu(\boldsymbol{a}^{\mathsf{T}}\boldsymbol{e} - 1),$$

where the variables $\boldsymbol{\ell} = [\ell_0, \ldots, \ell_{N-r-1}]^{\mathsf{T}} \in \mathbb{R}^{N-r}$ and $\mu \in \mathbb{R}$ are Lagrange multipliers. The first-order necessary conditions for optimality (FONC) follow

naturally:

$$\partial\mathcal{L}/\partial\widehat{\boldsymbol{y}} = \boldsymbol{\Lambda}(\widehat{\boldsymbol{y}} - \boldsymbol{y}) + (\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}\boldsymbol{\ell} = \boldsymbol{0}, \tag{3.39a}$$

$$\partial\mathcal{L}/\partial\boldsymbol{\ell} = \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{Y}}\boldsymbol{a} = \boldsymbol{0}, \tag{3.39b}$$

$$\partial\mathcal{L}/\partial\boldsymbol{a} = (\widehat{\boldsymbol{Y}})^{\mathsf{T}}\boldsymbol{\ell} + \mu\boldsymbol{e} = \boldsymbol{0}, \tag{3.39c}$$

$$\partial\mathcal{L}/\partial\mu = \boldsymbol{a}^{\mathsf{T}}\boldsymbol{e} - 1 = 0. \tag{3.39d}$$

The *critical points* of the approximation problem (3.35) are the real-valued solutions $(\boldsymbol{a}, \widehat{\boldsymbol{y}}, \boldsymbol{\ell}, \mu)$ of these equations.

Multiplication of (3.39c) from the left with $\boldsymbol{a}^{\mathsf{T}}$ gives

$$\boldsymbol{a}^{\mathsf{T}}(\widehat{\boldsymbol{Y}})^{\mathsf{T}}\boldsymbol{\ell} + \mu\boldsymbol{a}^{\mathsf{T}}\boldsymbol{e} = 0$$

$$\iff \mu\boldsymbol{a}^{\mathsf{T}}\boldsymbol{e} = \mu = 0,$$

where we used $\widehat{\boldsymbol{Y}}\boldsymbol{a} = \boldsymbol{0}$ which follows from (3.39b), and $\boldsymbol{a}^{\mathsf{T}}\boldsymbol{e} = 1$ from (3.39d). As such, independent of our choice for the normalization $\boldsymbol{e}$, we can substitute $\mu = 0$ in the FONC (3.39). After this substitution, (3.39a)-(3.39d) constitute a system of $2N + 1$ quadratic polynomial equations in $2N + 1$ unknowns $(\widehat{\boldsymbol{y}}, \boldsymbol{a}, \boldsymbol{\ell})$.

The Riemannian SVD

Elimination of the approximating data $\widehat{\boldsymbol{y}}$ leads to the *Riemannian singular value decomposition* (SVD) formulation of (3.35).

From $\partial\mathcal{L}/\partial\widehat{\boldsymbol{y}} = \boldsymbol{0}$ we get

$$\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}\boldsymbol{\ell} = \boldsymbol{y} - \widehat{\boldsymbol{y}}.$$

Substituting this expression for $\widehat{\boldsymbol{y}}$ into the condition $\partial\mathcal{L}/\partial\boldsymbol{\ell} = \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\widehat{\boldsymbol{y}} = \boldsymbol{0}$ yields

$$\boldsymbol{T}^{\boldsymbol{a}}_{N-r}(\boldsymbol{y} - \boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}\boldsymbol{\ell}) = \boldsymbol{0}$$

$$\iff \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}\boldsymbol{\ell} = \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{y} \tag{3.40}$$

$$\iff \boldsymbol{\ell} = (\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}})^{-1}\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{y}. \tag{3.41}$$

The inverse $\boldsymbol{\Lambda}^{-1}$ exists because $\boldsymbol{\Lambda}$ is positive definite.

Similarly to (3.38), we can replace the Hankel-vector product in (3.39c) by a Toeplitz-vector product

$$(\widehat{\boldsymbol{Y}})^{\mathsf{T}}\boldsymbol{\ell} = \boldsymbol{T}^{\boldsymbol{\ell}}_{r+1}\widehat{\boldsymbol{y}} = \boldsymbol{0}, \tag{3.42}$$

with $\boldsymbol{T}_{r+1}^{\boldsymbol{\ell}} \in \mathbb{R}^{(r+1)\times N}$ the banded Toeplitz matrix constructed from $\boldsymbol{\ell}$. Consequently, from $\partial\mathcal{L}/\partial\boldsymbol{a} = \boldsymbol{0}$:

$$\begin{aligned} & \boldsymbol{T}_{r+1}^{\boldsymbol{\ell}}\widehat{\boldsymbol{y}} = \boldsymbol{0}, \\ \iff & \boldsymbol{T}_{r+1}^{\boldsymbol{\ell}}(\boldsymbol{y} - \boldsymbol{\Lambda}^{-1}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}}\boldsymbol{\ell}) = \boldsymbol{0} \\ \iff & \boldsymbol{T}_{r+1}^{\boldsymbol{\ell}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}_{r+1}^{\boldsymbol{\ell}})^{\mathsf{T}}\boldsymbol{a} = \boldsymbol{T}_{r+1}^{\boldsymbol{\ell}}\boldsymbol{y}. \end{aligned} \tag{3.43}$$

Combining (3.40) and (3.43) gives:

$$\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}}\boldsymbol{\ell} = \boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{y}, \tag{3.44a}$$

$$\boldsymbol{T}_{r+1}^{\boldsymbol{\ell}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}_{r+1}^{\boldsymbol{\ell}})^{\mathsf{T}}\boldsymbol{a} = \boldsymbol{T}_{r+1}^{\boldsymbol{\ell}}\boldsymbol{y}, \tag{3.44b}$$

which together with the normalization (3.39d) gives a system of $N+2$ (at most cubic) polynomial equations in the $N+1$ variables $(\boldsymbol{a}, \boldsymbol{\ell})$.

This set of equations corresponds to the Riemannian SVD formulation for the weighted low-rank Hankel matrix approximation problem. It is an alternative, yet equivalent, formulation of the FONC (3.39). Indeed, the bijection between the solutions $(\widehat{\boldsymbol{y}}, \boldsymbol{a}, \boldsymbol{\ell})$ of (3.39) and the solutions $(\boldsymbol{a}, \boldsymbol{\ell})$ of (3.44) is given by

$$\widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{\Lambda}^{-1}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}}\boldsymbol{\ell}.$$

The real-valued elements of the solution set of the Riemannian SVD correspond to *all critical points* of the low-rank approximation problem (3.35).

Example 3.20 (Running example – FONC)**.** *Consider the Hankel matrix $\boldsymbol{Y}$ ($N = 6$):*

$$\boldsymbol{Y} = \begin{bmatrix} 6 & 5 & 4 & 3 & 2 \\ 5 & 4 & 3 & 2 & 1 \end{bmatrix}^{\mathsf{T}},$$

We are interested in the optimal ($r = 1$) low-rank approximation $\widehat{\boldsymbol{Y}}$. We will consider both unit weights $\boldsymbol{\lambda_I}$ and Frobenius weights $\boldsymbol{\lambda_F} = [1, 2, 2, 2, 2, 1]^{\mathsf{T}}$, and set $\boldsymbol{e}$ so that we can substitute $a_r = 1$. This problem setup corresponds to the example from [40, Section VI.A].

For both $\boldsymbol{\lambda_I}$ and $\boldsymbol{\lambda_F}$, (3.44) *corresponds to a system of 7 polynomial equations in 6 variables. They were solved using homotopy continuation [20]. Both systems have 18 common roots, 2 of which real-valued. They are displayed in Tables 3.6-3.7. In both cases, the global minimizer of the problem corresponds to the values obtained in [40, Tables I-II].*

Table 3.6: The real-valued solutions to the Riemannian SVD for Example 3.20 in the case of unit weights $\boldsymbol{\lambda_I}$. For each solution, σ_i denotes the singular values of the corresponding Hankel matrix $\widehat{\boldsymbol{Y}}$.

i	$\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|^2_{\boldsymbol{\Lambda}}$	a_0	ℓ_0	ℓ_1	ℓ_2	ℓ_3	ℓ_4	σ_i
1	0.6875	-0.7602	0.3844	0.2209	-0.1874	-0.5562	-0.5978	$\{11.77,\ 7.928\times10^{-16}\}$
2	91	1.492	4.0220	0.6556	2.2420	0.5082	1	$\{4.87\times10^{-15},\ 3.603\times10^{-15}\ \}$

Table 3.7: The real-valued solutions to the Riemannian SVD for Example 3.20 in the case of Frobenius weights $\boldsymbol{\lambda_F}$. For each solution, σ_i denotes the singular values of the corresponding Hankel matrix $\widehat{\boldsymbol{Y}}$.

i	$\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|^2_{\boldsymbol{\Lambda}}$	a_0	ℓ_0	ℓ_1	ℓ_2	ℓ_3	ℓ_4	σ_i
1	0.8726	-0.7629	0.5081	0.3337	-0.3021	-0.8247	-0.651	{12.01, 1.769e-16}
2	145	2.292	2.6180	3.2210	2.0850	1.7080	1	$\{1.966\times10^{-15},\ 1.883\times10^{-15}\}$

Decomposition of the ambient space

Similarly as in Section 3.3.1, the LS-criterion leads to an orthogonal decomposition of the ambient space. However, in the context of non-uniform weights, orthogonality must be reinterpreted with respect to the inner product defined by the weighting matrix $\boldsymbol{\Lambda}$. In what follows, we demonstrate that once this generalized notion of orthogonality is adopted, the insights and results established before generalize straightforwardly to the weighted setting.

Combining (3.41) with (3.39a) gives

$$\begin{aligned}
\boldsymbol{\Lambda}(\boldsymbol{y}-\widehat{\boldsymbol{y}}) &= (\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}})^{-1}\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{y} \\
\iff \widetilde{\boldsymbol{y}} &= \boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}})^{-1}\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{y}, \\
&= (\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1})^{\mathsf{T}})^{-1}\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\boldsymbol{y}, \\
&= (\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1})^{\mathsf{T}})^{-1}\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{y}, \qquad (3.45)
\end{aligned}$$

which implies that

$$\widehat{\boldsymbol{y}} = \left(\boldsymbol{I} - (\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1})^{\mathsf{T}})^{-1}\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\right)\boldsymbol{y}. \qquad (3.46)$$

It follows from (3.45) that, if the parameters $\boldsymbol{a}$ were given, the optimal misfit $\widetilde{\boldsymbol{y}}$ corresponds to the *oblique projection* of the given data $\boldsymbol{y}$ onto the row-space of $\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}$, along the direction of null $\left(\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\right)$. This oblique projection is defined

by the projection matrix[13]:

$$\boldsymbol{Q}_{\boldsymbol{\Lambda}} = (\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1})^{\mathsf{T}}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1})^{\mathsf{T}})^{-1}\boldsymbol{T}_{N-r}^{\boldsymbol{a}}.$$

Similarly, (3.46) expresses the optimal $\widehat{\boldsymbol{y}}$ as the projection of $\boldsymbol{y}$ onto $\operatorname{null}\left(\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\right)$, thereby guaranteeing that $\widehat{\boldsymbol{Y}}$ will be rank-deficient, along the direction of $\operatorname{row}\left(\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1}\right)$, so that the $\boldsymbol{\Lambda}$-weighted norm of $\widetilde{\boldsymbol{y}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ is minimized.

As such, the ambient space $\mathbb{R}^N$ is decomposed into two subspaces:

$$\operatorname{null}\left(\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\right) \oplus \operatorname{row}\left(\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1}\right) = \mathbb{R}^N.$$

This decomposition is orthogonal with respect to $\boldsymbol{\Lambda}$-weighted Euclidean space, i.e., the Euclidean space determined by the inner product

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\boldsymbol{\Lambda}} = \boldsymbol{u}^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{v}. \tag{3.47}$$

Example 3.21 (Running example – oblique projection)**.** *The projection matrix from* (3.46) *for the global minimizer ($a_0 = -0.7629$) of the setup from Example 3.20 with $\boldsymbol{\lambda}_{\boldsymbol{F}}$ corresponds to:*

$$\boldsymbol{P} = (\boldsymbol{I} - \boldsymbol{Q}_{\boldsymbol{\lambda}_{\boldsymbol{F}}}) = \begin{bmatrix} 0.3558 & 0.3839 & 0.2929 & 0.2234 & 0.1704 & 0.0919 \\ 0.2714 & 0.2929 & 0.2234 & 0.1704 & 0.1300 & 0.0701 \\ 0.2071 & 0.2234 & 0.1704 & 0.1300 & 0.0992 & 0.0535 \\ 0.1580 & 0.1704 & 0.1300 & 0.0992 & 0.0757 & 0.0408 \\ 0.1205 & 0.1300 & 0.0992 & 0.0757 & 0.0577 & 0.0311 \\ 0.0919 & 0.0992 & 0.0757 & 0.0577 & 0.0440 & 0.0237 \end{bmatrix}.$$

One can verify that $\boldsymbol{P}^2 = \boldsymbol{P}$, which is what we expect from a projection. The projection invoked by multiplication with $\boldsymbol{P}$ is clearly not orthogonal with respect to the standard Euclidean space, since $\boldsymbol{P}$ is not symmetric. The matrix has one eigenvalue at 1 and five eigenvalues at 0, indicating that the range of the projection is a one-dimensional subspace. This subspace corresponds to $\operatorname{null}\left(\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\right)$*, which is indeed spanned by a single vector in the case $r = 1$.*

The (local) minimizers of the problem

The optimization problem (3.35) over the variables $\boldsymbol{a}$ and $\widehat{\boldsymbol{y}}$ has no (local) maximizers: in each critical point $\boldsymbol{a}$, the optimal minimizing $\widehat{\boldsymbol{y}}$ is uniquely defined via (3.46), so that a critical point $(\boldsymbol{a}, \widehat{\boldsymbol{y}})$ of (3.35) must be either a

[13] Let $\boldsymbol{A} \in \mathbb{R}^{p\times k}$ and $\boldsymbol{B} \in \mathbb{R}^{k\times p}$ with $\boldsymbol{B}^{\mathsf{T}}\boldsymbol{A}$ invertible. The matrix $\boldsymbol{Q} = \boldsymbol{A}(\boldsymbol{B}^{\mathsf{T}}\boldsymbol{A})^{-1}\boldsymbol{B}^{\top}$ is then an *oblique projection matrix* with $\operatorname{range}(\boldsymbol{Q}) = \operatorname{range}(\boldsymbol{A})$ and $\operatorname{null}(\boldsymbol{Q}) = \operatorname{null}\left(\boldsymbol{B}^{\top}\right)$ (i.e., $\boldsymbol{Q}$ projects onto $\operatorname{range}(\boldsymbol{A})$ along $\operatorname{null}\left(\boldsymbol{B}^{\mathsf{T}}\right)$).

saddle point (it is maximizing over $\boldsymbol{a}$ and minimizing over $\widehat{\boldsymbol{y}}$) or a (local) minimum.

The following reformulation of (3.35), which exploits the separable structure of the problem, serves to be useful for developing intuition about the underlying structure of the problem. By replacing $\widehat{\boldsymbol{y}}$ with its optimal value obtained via the projection (3.46) determined by $\boldsymbol{a}$, we can consider the low-rank approximation problem (3.35) as an optimization problem over the decision variables $\boldsymbol{a}$ only:

$$\begin{aligned} \underset{\boldsymbol{a}\in\mathbb{R}^{r+1}}{\text{minimize}} \quad & \boldsymbol{y}^{\mathsf{T}}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}}(\boldsymbol{\Lambda})^{-1}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}})\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{y} \\ \text{subject to} \quad & \boldsymbol{e}^{\mathsf{T}}\boldsymbol{a} = 1. \end{aligned}$$

As observed in Tables 3.6-3.7, the system of equations (3.44) has solutions for which $\widehat{\boldsymbol{y}} = \boldsymbol{0}$ and thus $\operatorname{rank}\widehat{\boldsymbol{Y}} = 0 < r$. These special critical points correspond to values $\boldsymbol{a}$ for which the given data $\boldsymbol{y}$ lies completely in row $\left(\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1}\right)$. Indeed, for this type of critical points the cost function value equals $\|\boldsymbol{y}\|_{\boldsymbol{\Lambda}}^2$, which is the maximal value that is reachable if one obtains $\widetilde{\boldsymbol{y}}$ and $\widehat{\boldsymbol{y}}$ via the projections (3.45) and (3.46). In this sense, these critical points are global maximizers of the objective function in the optimization problem (3.48).

Lemma 3.22. *Let the given data $\boldsymbol{y} \in \mathbb{R}^N$ be such that $\operatorname{rank}\boldsymbol{Y} = r+1$, i.e., $\boldsymbol{y}$ is sufficiently generic, and consider a positive definite $\boldsymbol{\Lambda}$. Then any local minimizer $\widehat{\boldsymbol{y}} \in \mathbb{R}^N$ of the problem (3.35) is such that $\operatorname{rank}\widehat{\boldsymbol{Y}} = r$.*

Proof. The proof of Lemma 3.10, originally formulated in the standard Euclidean space, extends naturally to the weighted Euclidean space defined by the inner product in (3.47). The argument remains valid under this generalized inner product, requiring only a reinterpretation of orthogonality with respect to the weighting matrix $\boldsymbol{\Lambda}$.

□

We can once again exploit Lemma 3.22 to further adapt the system of equations (3.44). When $\operatorname{rank}\widehat{\boldsymbol{Y}} = r$, which is guaranteed for all local minimizers of (3.35), the rows of the matrix $\boldsymbol{T}_{N-2r}^{\boldsymbol{a}} \in \mathbb{R}^{(N-2r)\times(N-r)}$ span a basis for the $(N-2r)$-dimensional left null space of $\widehat{\boldsymbol{Y}}$. As such, it follows from (3.42) that

$$\boldsymbol{\ell} = (\boldsymbol{T}_{N-2r}^{\boldsymbol{a}})^{\mathsf{T}}\boldsymbol{g}, \tag{3.49}$$

for some $\boldsymbol{g} \in \mathbb{R}^{N-2r}$. When combined with (3.44), this leads to the following result.

Theorem 3.23 (Weighted double FIR filter)**.** *For generic given data $\boldsymbol{y}$ and positive definite $\boldsymbol{\Lambda}$, the critical points $\widehat{\boldsymbol{y}}$ of* (3.35) *for which* $\operatorname{rank}\widehat{\boldsymbol{Y}} = r$*, are such that:*

$$\boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2r})^{\mathsf{T}}\boldsymbol{g}.$$

The vector $\boldsymbol{a}$, which generates the banded-Toeplitz matrices $\boldsymbol{T}^{\boldsymbol{a}}_{N-r}$ and $\boldsymbol{T}^{\boldsymbol{a}}_{N-2r}$, spans the one-dimensional null space of $\widehat{\boldsymbol{Y}}$.

Remark 3.24. *After obtaining the results presented in this section, we found that Theorem 3.23, which is the generalization of the double FIR filter property from Theorem 3.3 to non-unit weights, also appeared in [193], in which a heuristic variable projection approach for solving* (3.35) *using the Gauss–Newton method is proposed. The authors do not mention its connection to Walsh's Theorem, nor acknowledge its effectiveness in formulating a globally optimal methodology.*

Substitution of (3.49) into (3.44a) leads to:

$$\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{y} - \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2r})^{\mathsf{T}}\boldsymbol{g} = \boldsymbol{0}, \tag{3.50}$$

which, together with the normalization (3.39d), encompasses a system of $N-r+1$ quartic polynomial equations in $N - r + 1$ variables.

We will formally show in Chapter 6 that, for generic data $\boldsymbol{y}$, none of the solutions for which $\operatorname{rank}\widehat{\boldsymbol{Y}} < n$ satisfy (3.50). As such, for generic data $\boldsymbol{y}$, the formulation (3.50) characterizes only a subset of the critical points.

Example 3.25 (Running example – Double FIR filter)**.** *The system* (3.50) *corresponding to the problem setup ($r = 1$) from Example 3.20 with $\boldsymbol{\lambda_F}$ has 5 polynomial equations of degree at most four. It has one real-valued solution:*

$$(a_0, g_0, g_1, g_2, g_3) = (-0.7629, -0.6659, -1.310, -1.321, -0.651),$$

which, via the transformation (3.49)*, corresponds to the solution shown in the first row of Table 3.7. Indeed, we have*

$$\begin{pmatrix} 0.5081 \\ 0.3337 \\ -0.3021 \\ -0.8247 \\ -0.651 \end{pmatrix} = \begin{pmatrix} -0.7629 & 0 & 0 & 0 \\ 1 & -0.7629 & 0 & 0 \\ 0 & 1 & -0.7629 & 0 \\ 0 & 0 & 1 & -0.7629 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -0.6659 \\ -1.310 \\ -1.321 \\ -0.6510 \end{pmatrix}.$$

The critical point corresponding to the second row of Table 3.7, for which $\operatorname{rank}\widehat{\boldsymbol{Y}} = 0 < r = 1$, does not appear as one of the solutions of (3.50)*.*

The multiparameter eigenvalue problem

We can restructure the system of polynomial equations of (3.50) into a (rectangular) multiparameter eigenvalue problem (MEP). To ease the notation, we will assume from this point on that $\boldsymbol{e}$ is chosen so that $a_r = 1$. Consequently, we introduce the vector $\boldsymbol{a}_{\mathrm{tr}} = [a_0, \ldots, a_{r-1}]^{\mathsf{T}} \in \mathbb{R}^r$ to denote the remaining unknowns.

Consider the third degree matrix polynomial in the monomials $\boldsymbol{a}_{\mathrm{tr}}^{\boldsymbol{\alpha}} = a_0^{\alpha_0} \cdots a_{r-1}^{\alpha_{r-1}}$:

$$\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{a}_{\mathrm{tr}}) = \begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{y} & \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2r})^{\mathsf{T}} \end{bmatrix}, \tag{3.51}$$

where $\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{a}_{\mathrm{tr}}) = \sum_{\{\boldsymbol{\alpha}\}} \boldsymbol{M}_{\boldsymbol{\alpha}} \boldsymbol{a}^{\boldsymbol{\alpha}}$ with coefficient matrices $\boldsymbol{M}_{\boldsymbol{\alpha}} \in \mathbb{R}^{(N-r)\times(N-2r+1)}$. The values $\boldsymbol{a}_{\mathrm{tr}} \in \mathbb{C}^r$ for which $\operatorname{rank} \boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{a}_{\mathrm{tr}}) \leq N - 2r$, so that there exists a nonzero vector $\boldsymbol{z} \in \mathbb{C}^{N-2r+1}$ for which $\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{a}_{\mathrm{tr}})\boldsymbol{z} = \boldsymbol{0}$, are the affine eigentuples of this MEP.

Lemma 3.26. *For given data $\boldsymbol{y} \in \mathbb{R}^N$, rank r and positive definite $\boldsymbol{\Lambda} \in \mathbb{R}^{N\times N}$,*

- *each real-valued critical point $\widehat{\boldsymbol{y}}$ of* (3.35) *for which* $\operatorname{rank} \widehat{\boldsymbol{Y}} = r$ *corresponds to an affine, real-valued eigentuple $\boldsymbol{a}_{\mathrm{tr}}$ of the MEP* (3.51)*,*
- *each affine, real-valued eigentuple of $\boldsymbol{a}_{\mathrm{tr}}$ of the MEP* (3.51) *corresponds to a critical point of* (3.35)*.*

Proof. It suffices to show that the matrix $\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2r})^{\mathsf{T}}$ has full column rank for all nonzero $\boldsymbol{a} \in \mathbb{R}^{r+1}$, as this guarantees that the corresponding vector $\boldsymbol{z}$ can always be normalized so that its first entry equals -1.

Let $\boldsymbol{u} \in \mathbb{R}^{N-r}$, with $\boldsymbol{u} \neq \boldsymbol{0}$, and define $\boldsymbol{x} = (\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}\boldsymbol{u}$. Since $\boldsymbol{T}^{\boldsymbol{a}}_{N-r}$ has full row rank, it follows that $\boldsymbol{x} \neq \boldsymbol{0}$. Consider the quadratic form $\boldsymbol{u}^{\mathsf{T}}\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}\boldsymbol{u} = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{x}$. Because $\boldsymbol{\Lambda}^{-1}$ is positive definite and $\boldsymbol{x}$ real, the form is strictly positive. This implies that, for all $\boldsymbol{a} \in \mathbb{R}^{r+1}$, the matrix $\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}$ is invertible. Since $\boldsymbol{T}^{\boldsymbol{a}}_{N-2r}$ also has full row rank, it follows immediately that the product $\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2r})^{\mathsf{T}}$ has the full column rank. □

Remark 3.27. *It should be noted that Lemma 3.26 considers only real-valued eigentuples. In particular, the matrix $\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}$ may become singular if the variables $\boldsymbol{a}$ are allowed to take complex values. As such, the MEP formulation* (3.51) *can admit additional complex solutions beyond those arising from the system of equations* (3.50)*. This observation has several important computational implications, particularly since MEP solvers typically operate*

over the complex field $\mathbb{C}$ rather than the reals $\mathbb{R}$. A detailed examination of these implications is deferred to Part II of this dissertation.

Remark 3.28. *Note that Lemma 3.26 does not guarantee that every affine, real-valued eigentuple $\boldsymbol{a}_{\text{tr}}$ of the MEP corresponds to a critical point with $\operatorname{rank} \widehat{\boldsymbol{Y}} = n$. Although this is almost always the case, there exist special $\boldsymbol{y}$ for which this is no longer guaranteed. We will formalize this phenomenon in Chapter 6.*

Example 3.29 (Running example – MEP)**.** *Consider the MEP formulation of the system of equations from Example 3.25:*

$$\left[\begin{pmatrix} 5 & 0 & 0 & 0 & 0 \\ 4 & 0.5 & 0 & 0 & 0 \\ 3 & 0 & 0.5 & 0 & 0 \\ 2 & 0 & 0 & 0.5 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 6 & 1 & 0 & 0 & 0 \\ 5 & 0 & 1 & 0 & 0 \\ 4 & 0.5 & 0 & 1 & 0 \\ 3 & 0 & 0.5 & 0 & 1 \\ 2 & 0 & 0 & 0.5 & 0 \end{pmatrix} a_0 \right.$$
$$\left. + \begin{pmatrix} 0 & 0 & 0.5 & 0 & 0 \\ 0 & 1 & 0 & 0.5 & 0 \\ 0 & 0 & 1 & 0 & 0.5 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} a_0^2 + \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} a_0^3 \right] \begin{pmatrix} -1 \\ g_0 \\ g_1 \\ g_2 \\ g_3 \end{pmatrix} = \mathbf{0},$$

which can be written compactly via

$$\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{a}_{\text{tr}}) = \begin{pmatrix} -5-6a_0 & a_0+a_0^3 & 0.5a_0^2 & 0 & 0 \\ -4-5a_0 & 0.5+a_0^2 & a_0+0.5a_0^3 & 0.5a_0^2 & 0 \\ -3-4a_0 & 0.5a_0 & 0.5+a_0^2 & a_0+0.5a_0^3 & 0.5a_0^2 \\ -2-3a_0 & 0 & 0.5a_0 & 0.5+a_0^2 & a_0+0.5a_0^3 \\ -1-2a_0 & 0 & 0 & 0.5a_0 & 1+a_0^2 \end{pmatrix}.$$

The matrix polynomial $\boldsymbol{M}_{\boldsymbol{\Lambda}}(a_0)$ is square such that its eigentuples can be obtained via its determinant, which is a degree 13 univariate polynomial in a_0. Only one out of the 13 roots is real-valued: $a_0 = -0.7629$. The corresponding values for $\boldsymbol{g}$ can be found via the null space of $\boldsymbol{M}_{\boldsymbol{\Lambda}}(a_0)$:

$$\begin{pmatrix} 0.4224 & -1.206 & 0.2910 & 0 & 0 \\ 0.1853 & 1.0820 & -0.9849 & 0.2910 & 0 \\ -0.0516 & -0.3814 & 1.0820 & -0.984953 & 0.2910 \\ -0.2887 & 0 & -0.3814 & 1.08205 & -0.9849 \\ -0.5258 & 0 & 0 & -0.3814 & 1.5820 \end{pmatrix} \begin{pmatrix} -1 \\ -0.6659 \\ -1.3102 \\ -1.3213 \\ -0.6509 \end{pmatrix} = \mathbf{0}.$$

3.7.4 Examples

Several numerical examples are worked out to illustrate the methodology.

Example 3.30. *We consider the problem used in [132, Example 2], in which the optimal, unstructured rank-one ($r = 1$) reduction of a given circulant matrix is computed:*

$$\boldsymbol{U} = \begin{pmatrix} -59 & 11 & 59 \\ 11 & 59 & -59 \\ 59 & -59 & 11 \end{pmatrix}, \tag{3.52}$$

with respect to the element-wise weighted Frobenius norm (3.33) *determined by the weights:*

$$\boldsymbol{W} = \begin{pmatrix} 9 & 6 & 1 \\ 6 & 1 & 9 \\ 1 & 9 & 6 \end{pmatrix}. \tag{3.53}$$

The authors of [132] report 39 solutions to the first-order necessary conditions for optimality of the unstructured, weighted low-rank approximation problem, 19 of which are real-valued. Following the results in [132], the 7 (local) minimizers of the unstructured problem correspond to

$$\widehat{\boldsymbol{U}}_{\mathbf{1}} = \begin{pmatrix} 0.0826 & 2.7921 & -1.5452 \\ 2.7921 & 94.3235 & -52.2007 \\ -1.5452 & -52.2007 & 28.8890 \end{pmatrix}, \quad \widehat{\boldsymbol{U}}_{\mathbf{2}} = \begin{pmatrix} -52.2007 & 28.8890 & -1.5452 \\ 2.7921 & -1.5452 & 0.0826 \\ 94.3235 & -52.2007 & 2.7921 \end{pmatrix},$$

$$\widehat{\boldsymbol{U}}_{\mathbf{3}} = \begin{pmatrix} -52.2007 & 2.7921 & 94.3235 \\ 28.8890 & -1.5452 & -52.2007 \\ -1.5452 & 0.0826 & 2.7921 \end{pmatrix}, \quad \widehat{\boldsymbol{U}}_{\mathbf{4}} = \begin{pmatrix} -29.8794 & 36.2165 & -27.2599 \\ -32.7508 & 39.6968 & -29.8794 \\ 39.6968 & -48.1160 & 36.2165 \end{pmatrix},$$

$$\widehat{\boldsymbol{U}}_{\mathbf{5}} = \begin{pmatrix} -48.1160 & 36.2165 & 39.6968 \\ 36.2165 & -27.2599 & -29.8794 \\ 39.6968 & -29.8794 & -32.7508 \end{pmatrix}, \quad \widehat{\boldsymbol{U}}_{\mathbf{6}} = \begin{pmatrix} -29.8794 & -32.7508 & 39.6968 \\ 36.2165 & 39.6968 & -48.1160 \\ -27.2599 & -29.8794 & 36.2165 \end{pmatrix},$$

$$\widehat{\boldsymbol{U}}_{\mathbf{7}} = \begin{pmatrix} -25.375 & -25.375 & -25.375 \\ -25.375 & -25.375 & -25.375 \\ -25.375 & -25.375 & -25.375 \end{pmatrix}, \tag{3.54}$$

the first three of which are the minimizers attaining the global minimum:

$$\left\| \boldsymbol{U} - \widehat{\boldsymbol{U}}_{\boldsymbol{i}} \right\|_{\boldsymbol{W}}^2 = 4.3556 \times 10^4, \quad \forall\, i \in \{1, 2, 3\}.$$

In this example, we will compute the optimal rank-1 Hankel matrix approximation of the given matrix (3.52) *with respect to the weights determined by* (3.53)*. The given matrix $\boldsymbol{U}$ corresponds to the following data vector ($N = 5$):*

$$\boldsymbol{U} = \boldsymbol{H}_3(\boldsymbol{y}), \quad \textit{where} \quad \boldsymbol{y} = [-59, 11, 59, -59, 11]^{\mathsf{T}}.$$

Table 3.8: The set of all critical points, obtained as the real-valued solution set of the Riemannian SVD (3.44), corresponding to the low-rank Hankel matrix approximation problem ($r = 1$) described in Example 3.30. For each solution, σ_i denotes the singular values of the corresponding Hankel matrix $\widehat{\boldsymbol{Y}}$.

i	$\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|_{\boldsymbol{\Lambda}}^2$	a_0	l_0	l_1	l_2	l_3	σ_i
1	7.345×10^4	1.744	-332.9	332.6	-118	-239.7	$\{71.27,\ 1.061 \times 10^{-14}\}$
2	7.447×10^4	0.0958	36.8	281.1	-1069	66.03	$\{59.94,\ 1.139 \times 10^{-14}\}$
3	7.57×10^4	-1	302.6	-133.9	-387	218.3	$\{71.77,\ 8.753 \times 10^{-15}\}$
4	8.265×10^4	-0.4407	276.1	-218.9	-958.3	76.3	$\{55.33,\ 1.407 \times 10^{-14}\}$
5	1.066×10^5	-15.92	33.36	-6.197	-11.51	66	$\{2.003 \times 10^{-14},\ 1.544 \times 10^{-14}\}$
6	1.066×10^5	0.779	-681.6	1044	-1113	66	$\{2.319 \times 10^{-14},\ 1.701 \times 10^{-14}\}$

We can use our methodology to compute the optimal rank $r = 1$ reduction of the $(N - r) \times (r + 1)$ Hankel matrix

$$\boldsymbol{Y} = \begin{pmatrix} -59 & 11 \\ 11 & 59 \\ 59 & -59 \\ -59 & 11 \end{pmatrix}.$$

By Lemma 3.16, the (3×3)-Hankel matrix composed from the elements of the obtained approximations $\widehat{\boldsymbol{y}}$ will also be of rank one. Furthermore, we use the conversion from (3.36) *to transform the element-wise weights from* (3.53) *to our problem setup, we obtain:*

$$\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{\lambda}), \quad \textit{with} \quad \boldsymbol{\lambda} = [9, 12, 3, 18, 6]^{\mathsf{T}}.$$

The set of all critical points, obtained by solving (3.44) *using [20], is depicted in Table 3.8.*

The affine real-valued solutions of the system of equations (3.50)*, which exploits the double FIR filter property, are depicted in Table 3.9. Notice that this characterization does not describe the subset of the critical points for which* $\operatorname{rank} \widehat{\boldsymbol{Y}} = 0 < r = 1$.

Given the values for $\boldsymbol{a}$ from Table 3.9, we can compute the corresponding data $\widehat{\boldsymbol{y}}$ via the oblique projection in (3.46)*. Alternatively, we can use the obtained tuples $(\boldsymbol{a}, \boldsymbol{g})$ and compute $\widehat{\boldsymbol{y}}$ by exploiting Theorem 3.23:*

$$\widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{\Lambda}^{-1} (\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}} (\boldsymbol{T}_{N-2r}^{\boldsymbol{a}})^{\mathsf{T}} \boldsymbol{g}.$$

Table 3.9: The set of real-valued solutions to the system of equations from (3.50), corresponding to the low-rank Hankel matrix approximation problem ($r = 1$) described in Example 3.30. For each solution, σ_i denotes the singular values of the corresponding Hankel matrix $\widehat{\boldsymbol{Y}}$.

i	$\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2_{\boldsymbol{\Lambda}}$	a_0	g_0	g_1	g_2	σ_i
1	7.345×10^4	1.744	-190.8	300.1	-239.7	$\{71.27,\ 1.788 \times 10^{-14}\}$
2	7.447×10^4	0.0958	384.1	-1076	66.03	$\{59.94,\ 5.175 \times 10^{-15}\}$
3	7.57×10^4	-1	-302.6	-168.8	218.2	$\{71.77,\ 2.427 \times 10^{-14}\}$
4	8.265×10^4	-0.4407	-626.4	-924.7	76.3	$\{55.33,\ 3.897 \times 10^{-14}\}$

The (3×3)*-Hankel matrices* $\boldsymbol{H}_3(\widehat{\boldsymbol{y}}_i)$ *corresponding to the critical points from Table 3.9 are given as:*

$$
\begin{aligned}
\boldsymbol{H}_3(\widehat{\boldsymbol{y}}_1) &= \begin{pmatrix} 5.506 & -9.6034 & 16.75 \\ -9.6034 & 16.75 & -29.214 \\ 16.75 & -29.214 & 50.955 \end{pmatrix}, \\
\boldsymbol{H}_3(\widehat{\boldsymbol{y}}_2) &= \begin{pmatrix} -59.392 & 5.6897 & -0.54507 \\ 5.6897 & -0.5451 & 0.0522 \\ -0.5451 & 0.0522 & -0.0050 \end{pmatrix}, \\
\boldsymbol{H}_3(\widehat{\boldsymbol{y}}_3) &= \begin{pmatrix} -25.375 & -25.375 & -25.375 \\ -25.375 & -25.375 & -25.375 \\ -25.375 & -25.375 & -25.375 \end{pmatrix}, \\
\boldsymbol{H}_3(\widehat{\boldsymbol{y}}_4) &= \begin{pmatrix} -45.48 & -20.045 & -8.8348 \\ -20.045 & -8.8348 & -3.8939 \\ -8.8348 & -3.8939 & -1.7162 \end{pmatrix}.
\end{aligned}
\tag{3.55}
$$

Numerical computation confirms that these matrices are rank one.

The number of critical points obtained in Table 3.8 does not correspond the number of solutions of the unstructured problem described in [132], nor does one of the minimizers from (3.54) *appear as one of the locally optimal solutions of the structured problem* (3.55)*. It is clear that the introduction of the additional structural constraint (i.e., the Hankel structure) significantly alters the optimization problem. The global minima of the unstructured problem (*4.3556×10^4*) gives a lower-bound for the global minima of the structured problem (*7.345×10^4*). This is as expected, since the set of feasible solutions of the structured formulation is a subset of the feasible set of the unstructured formulation.*

Example 3.31. *We fit a second-order model ($n = 2$) to given data $\boldsymbol{y}$ of length $N = 10$, using the Bombieri weights $\boldsymbol{\lambda_T}$.*

$$\boldsymbol{y} = [0.6, -0.5, -0.1, 0.6, -1.0, -0.3, -1.1, -0.6, 0.9, -1.4]^{\mathsf{T}}.$$

Solving the MEP (3.51) *using* MACAULAYLAB *[178] indicates that the MEP has 208 affine eigentuples, of which 12 are real-valued. The globally optimal solution is $\boldsymbol{a} = [-1.0952, -0.3295, 1]^{\mathsf{T}}$, with an objective function value $\|\widetilde{\boldsymbol{y}}\|_{\boldsymbol{\Lambda}} = 9.6981$, compared to $\|\boldsymbol{y}\|_{\boldsymbol{\Lambda}} = 17.1581$.*

Although 208 homotopy paths are tracked, the polynomial system (3.50) *has only 82 affine solutions, 12 of which are real. Each of these 82 solutions coincide with a solution to the MEP. The remaining 126 paths diverge to infinity in the direction of $\boldsymbol{g}$. This discrepancy between the affine solutions of the MEP* (3.51) *and those of the system* (3.50) *will be examined further in Chapter 6.*

Solving the system of equations associated with the Riemannian SVD formulation (3.44) *using homotopy continuation [20] illustrates that it has 181 solutions, 29 of which real-valued. Note that only a subset (including all minimizers) of the real-valued solutions to* (3.44) *appears as solutions to the proposed MEP* (3.51) *(or equivalently, the system of equations* (3.50)*).*

The objective function values over (a_0, a_1) and the obtained critical points in the domains $[-10, 5] \times [-10, 6]$ and $[-2, 2] \times [-2.5, 2.75]$ are shown in Figures 3.8a and 3.8b, respectively.

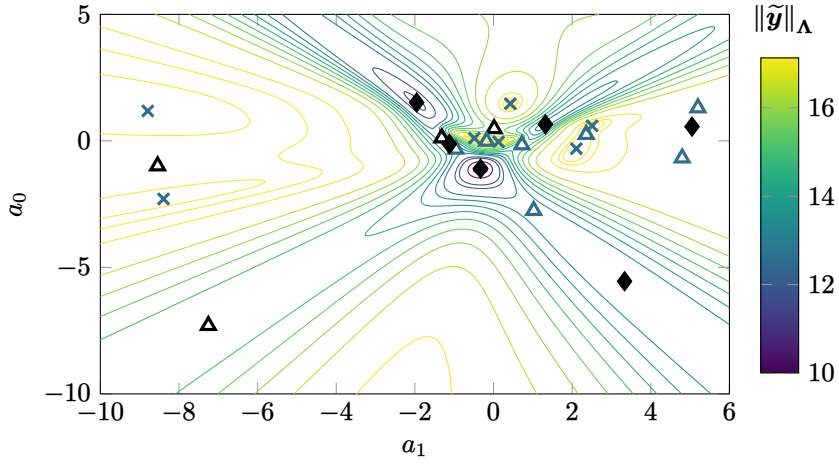
Example 3.32 (Walsh's citations)**.** *In this example we consider the number of yearly citations*[14] *of the seminal work [183] for the period ranging from 2000 to 2024 ($N = 25$):*

$$\boldsymbol{y} = [28, 26, 16, 24, 23, 39, 23, 30, 30, 26, 28, 38, 50,$$
$$32, 45, 43, 42, 36, 42, 27, 30, 16, 17, 12, 17]^{\mathsf{T}}.$$

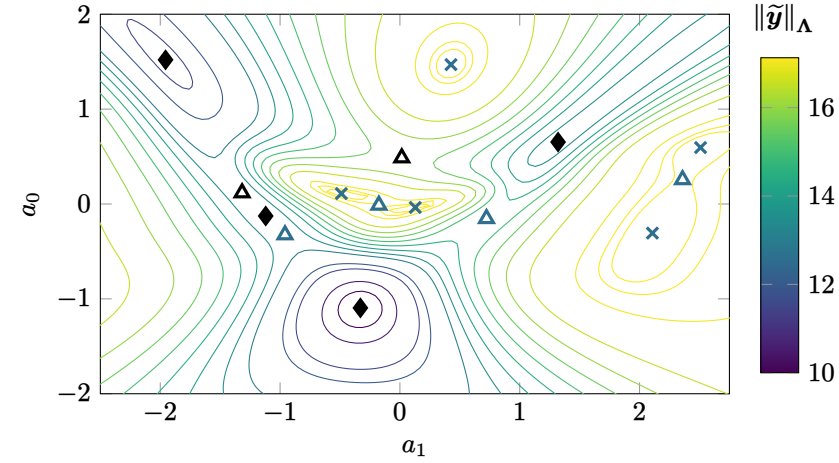
The (13×13)-Hankel matrix $\boldsymbol{H}_{13}(\boldsymbol{y})$ is invertible, implying that the sequence $\boldsymbol{y}$ can not be realized as the output of an autonomous single-output LTI model of order 13 or lower.

In this example, we will approximate the given time series $\boldsymbol{y}$ by the output of a first-, second- and third-order autonomous LTI model. In addition to the standard LS realization setting with uniform weights ($\boldsymbol{\lambda_I}$), we also consider

[14] We obtained the yearly number of citations of [183] via its corresponding entry on the website `https://www.semanticscholar.org`, accessed on 07/04/2025.



(a) The domain $[-10, 5] \times [-10, 6]$.



(b) Close-up to $[-2, 2] \times [-2.5, 2.75]$.

Figure 3.8: Contour plot of the objective function of the weighted LS realization problem as a function of the parameters (a_0, a_1), for the Bombieri weights $\boldsymbol{\lambda}_{\boldsymbol{T}}$. For all values (a_0, a_1), the optimal $\widehat{\boldsymbol{y}}$ was determined via (3.46) to compute the corresponding objective function value (cf. the formulation in (3.48)). The problem setup corresponds to Example 3.31. Critical points contained in the displayed region are highlighted: local minimizers (♦), saddle points (△) and local maximizers (×), where critical points appearing as eigentuples of the proposed MEP (3.51) (or equivalently, as solutions to the system of equations (3.50)) are depicted in black.

three weighted scenarios that assign greater emphasis to specific segments of the time series. A time series $\widehat{\boldsymbol{y}}$ is the output of an r-th order LTI autonomous single-output model if the Hankel matrix $\boldsymbol{H}_{N-r}(\widehat{\boldsymbol{y}})$ is rank-deficient. As such, we can apply our methodology for the (weighted) low-rank approximation problem for Hankel matrices, using $r = 1, 2, 3$.

Standard case *We begin with the unit weight $\boldsymbol{\lambda_I}$ scenario. The real-valued solutions of* (3.50) *are depicted in Table 3.10. Figure 3.9 displays the original data sequence $\boldsymbol{y}$, alongside the globally optimal approximations $\widehat{\boldsymbol{y}}_r$ for $r = 1, 2, 3$. The applied weights λ_i are visualized in the background using purple bars, scaled on the secondary vertical axis to the right. The flat weight profile suggests that the optimization gives uniform importance across the time interval. The third-order approximation $\widehat{\boldsymbol{y}}_3$ closely matches the true signal, effectively capturing both the rise in citations around 2012–2016 and the subsequent decline. The second-order model $\widehat{\boldsymbol{y}}_2$ reproduces the general trend but deviates more noticeably near the peak and tail regions. In contrast, the first-order model $\widehat{\boldsymbol{y}}_1$ exhibits only a coarse approximation of the data, with no ability to reflect the signal's curvature.*

Increasing importance *Suppose greater confidence is placed in the more recent measurements than in the older ones. To incorporate this preference, the following set of weights can be applied:*

$$\lambda_i = 1 - (0.9)^{(i+1)}, \quad i = 0, \ldots, N-1. \tag{3.56}$$

The real-valued solutions of (3.50) *are depicted in Table 3.11 and the data $\widehat{\boldsymbol{y}}_r$ corresponding to the global minimizers are depicted in Figure 3.10.*

Compared to the results in Figure 3.9, the errors are larger in the early years of this century, while the approximation accuracy improves significantly for more recent years. This reflects the effect of the weighting scheme, which causes the optimization to prioritize fitting the recent data more closely. The effect is particularly evident in the first- and second-order approximations.

Decreasing importance *Suppose now that we reverse our objective and aim for a more accurate approximation of the first decade of this century. To reflect this preference, we introduce the following set of exponentially decaying weights:*

$$\lambda_i = 1 - (0.9)^{(N-i)}, \quad i = 0, \ldots, N-1. \tag{3.57}$$

The real-valued solutions of (3.50) *are depicted in Table 3.12 and the data $\widehat{\boldsymbol{y}}_r$ corresponding to the global minimizers are depicted in Figure 3.11.*

As expected, these weights give the opposite effect compared to those in Figure 3.10: the approximation becomes more accurate in the early years, while the errors increase in the more recent period. This behavior is again particularly pronounced for the first- and second-order approximations.

Interestingly, the globally optimal third-order approximation ($r = 3$) appears largely unaffected by the choice of weights, whether they emphasize early or recent data. For the weights in (3.56)*, the poles of the model are* $\{1.1126 \pm 0.4121j, 1.02744\}$*, and for those in* (3.57)*, they are* $\{1.1021 \pm 0.3998j, 1.02568\}$*. Both are only slight deviations from the poles of the third-order solution obtained using uniform weights* $\boldsymbol{\lambda_I}$*, which are* $\{1.1103 \pm +0.4106j, 1.02712\}$*.*

A plausible explanation for this observation is that the third-order model fitted using uniform weights already captures the dynamics of $\boldsymbol{y}$ *in both temporal extremes, namely 2000–2008 and 2014–2024, sufficiently well, so that altering the weighting has little additional effect. Notably, all third-order models remain unstable, as each has at least one pole located outside the unit circle.*

Tolerate errors on boundaries *Finally, consider the case in which we are willing to tolerate larger errors at the beginning and end of the time interval. This objective is reflected by the following set of weights:*

$$\lambda_i = 2(1 - (0.9)^{(i+1)} - (0.9)^{(N-i)}), \quad i = 0, \ldots, N-1. \tag{3.58}$$

The real-valued solutions of (3.50) *are depicted in Table 3.13 and the data* $\widehat{\boldsymbol{y}}_r$ *corresponding to the global minimizers are depicted in Figure 3.12.*

It is immediately evident from Figure 3.12 that the real pole in the globally optimal third-order model is negative. This results in sign alternations of the form $(-1)^i$ *in one of the components of the reconstructed signal. The globally optimal third-order solution yields the pole set* $\{1.1252 \pm 0.0497j, -1.1439\}$*.*

Table 3.10: Real-valued solutions to (3.50) corresponding to Example 3.32 with unit weights $\boldsymbol{\lambda_I}$, where σ_i denotes the singular values of $\widehat{\boldsymbol{Y}}$.

i	$\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|^2_{\boldsymbol{\Lambda}}$	a_0	σ_i
1	2493	-0.9977	$\{205.1,\ 9.159\times10^{-14}\}$
2	2.422×10^4	1.172	$\{17.75,\ 1.249\times10^{-13}\}$
3	2.425×10^4	1.633	$\{14.73,\ 5.023\times10^{-14}\}$
4	2.44×10^4	0.8696	$\{3.302,\ 1.48\times10^{-13}\ \}$

(a) r = 1

i	$\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|^2_{\boldsymbol{\Lambda}}$	a_1	a_0	σ_i
1	1012	-2.15	1.161	$\{261.3,\ 16.39,\ 3.26\times10^{-12}\ \}$
2	2449	-0.1949	-0.8	$\{246.5,\ 8.985,\ 4.668\times10^{-14}\ \}$
3	2456	0.1032	-1.098	$\{245.6,\ 9.092,\ 5.083\times10^{-14}\ \}$
4	2488	-0.8095	-0.1872	$\{246.3,\ 1.97,\ 5.652\times10^{-14}\}$
5	2.355×10^4	-1.833	0.9933	$\{45.95,\ 15.41,\ 1.032\times10^{-11}\}$
6	2.42×10^4	2.084	1.121	$\{23.35,\ 2.457,\ 1.476\times10^{-11}\}$
⋮	⋮	⋮	⋮	⋮
18	2.438×10^4	-1.389	1.07	$\{6.079,\ 4.430,\ 7.876\times10^{-13}\}$
19	2.44×10^4	1.546	1.154	$\{3.512,\ 2.665,\ 1.241\times10^{-12}\}$

(b) r = 2

i	$\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|^2_{\boldsymbol{\Lambda}}$	a_2	a_1	a_0	σ_i
1	674.8	-3.248	3.682	-1.439	$\{298.6,\ 27.68,\ 5.439,\ 2.535\times10^{-11}\ \}$
2	859.8	-1.001	-1.357	1.383	$\{298.7,\ 26.19,\ 17.59,\ 3.099\times10^{-13}\ \}$
3	899.8	-0.2848	-2.963	2.282	$\{298.8,\ 25.42,\ 11.31,\ 5.397\times10^{-13}\}$
4	1003	-1.323	-0.6243	0.9669	$\{298.8,\ 23.88,\ 4.274,\ 3.933\times10^{-13}\}$
5	2228	0.9848	-0.9359	-1.04	$\{278.5,\ 30.51,\ 7.272,\ 4.379\times10^{-13}\ \}$
6	2243	-2.052	1.723	-0.6683	$\{280.4,\ 18.91,\ 10.32,\ 2.762\times10^{-12}\ \}$
⋮	⋮	⋮	⋮	⋮	⋮
50	2.435×10^4	2.916	3.146	1.19	$\{10.05,\ 4.96,\ 0.649,\ 5.269\times10^{-11}\}$
51	2.438×10^4	-0.5167	-0.1363	0.9278	$\{6.466,\ 5.564,\ 3.446,\ 1.308\times10^{-13}\}$

(c) r = 3

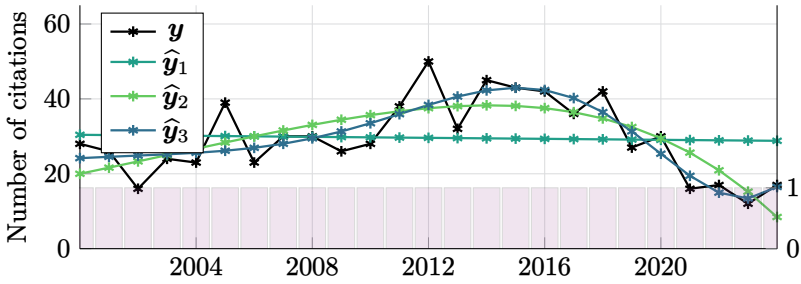


Figure 3.9: The globally optimal data sequences $\widehat{\boldsymbol{y}}$ corresponding to Example 3.32. The weights $\boldsymbol{\lambda_I}$ are depicted using purple bars.

Table 3.11: Real-valued solutions to (3.50) corresponding to Example 3.32 with weights (3.56), where σ_i denotes the singular values of $\widehat{\boldsymbol{Y}}$.

i	$\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|_{\boldsymbol{\Lambda}}^2$	a_0	σ_i
1	1313	−0.9789	$\{\ 229.4,\ 6.688\times10^{-14}\ \}$
2	1.264×10^4	1.139	$\{\ 18.32,\ 1.499\times10^{-13}\ \}$
3	1.266×10^4	1.559	$\{14.89,\ 4.203\times10^{-14}\ \}$
4	1.279×10^4	0.6996	$\{16.83,\ 6.49\times10^{-14}\ \}$

(a) r = 1

i	$\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|_{\boldsymbol{\Lambda}}^2$	a_1	a_0	σ_i
1	457.1	−2.051	1.064	$\{256.5,\ 18.59,\ 2.65\times10^{-12}\ \}$
2	1260	0.06327	−1.019	$\{\ 274.9,\ 14.05,\ 3.409\times10^{-14}\ \}$
3	1304	−0.2852	−0.679	$\{274,\ 17.92,3.42\times10^{-14}\ \}$
4	1312	−0.7685	−0.2059	$\{\ 274,\ 7.048,\ 4.295\times10^{-14}\ \}$
5	1.262×10^4	1.89	0.887	$\{\ 31.86,\ 4.641,\ 1.144\times10^{-11}\ \}$
6	1.265×10^4	2.272	1.106	$\{\ 22.95,\ 7.171,1.259\times10^{-12}\ \}$
⋮	⋮	⋮	⋮	⋮
18	1.279×10^4	0.97	0.564	$\{7.935,\ 3.588,\ 1.249\times10^{-13}\ \}$
19	1.279×10^4	−0.4221	0.7671	$\{4.024,\ 3.407,\ 1.039\times10^{-13}\ \}$

(b) r = 2

i	$\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|_{\boldsymbol{\Lambda}}^2$	a_2	a_1	a_0	σ_i
1	283.9	−3.253	3.694	−1.447	$\{\ 298.6,\ 27.73,\ 5.501,\ 6.545\times10^{-12}\ \}$
2	338.2	−0.9624	−1.209	1.199	$\{\ 294,\ 29.74,\ 18.71,\ 3.189\times10^{-13}\ \}$
3	380	−0.06485	−3.129	2.236	$\{\ 294.2,\ 28.38,\ 9.966,\ 3.783\times10^{-13}\ \}$
4	449	−1.352	−0.3702	0.7439	$\{293.9,\ 30.67,\ 14.73,2.684\times10^{-13}\ \}$
5	1185	0.9743	−0.9124	−0.979	$\{\ 309.7,\ 31.36,\ 6.322,\ 5.066\times10^{-13}\ \}$
6	1192	1.292	−0.8806	−1.317	$\{\ 309.7,\ 24.38,\ 6.247,\ 1.476\times10^{-13}\ \}$
⋮	⋮	⋮	⋮	⋮	⋮
49	1.278×10^4	2.027	1.832	0.6306	$\{19.71,\ 4.068,\ 1.712,\ 6.866\times10^{-12}\ \}$
50	1.278×10^4	0.2743	0.4675	0.5373	$\{19.15,\ 3.97,\ 2.965,\ 7.117\times10^{-14}\ \}$

(c) r = 3

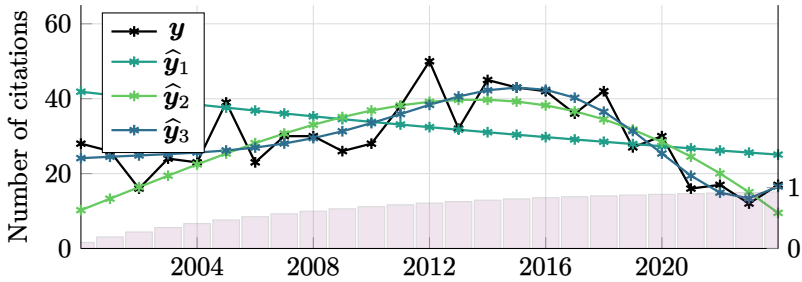


Figure 3.10: The globally optimal data sequences $\widehat{\boldsymbol{y}}_r'$ corresponding to the results in Table 3.11. The weights (3.56) are depicted using purple bars.

Table 3.12: Real-valued solutions to (3.50) corresponding to Example 3.32 with weights (3.57), where σ_i denotes the singular values of $\widehat{\boldsymbol{Y}}$.

i	$\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|_{\boldsymbol{\Lambda}}^2$	a_0	σ_i
1	734.3	-1.022	$\{233.8,\ 5.61\times10^{-14}\}$
2	1.256×10^4	1.197	$\{21.01,\ 1.517\times10^{-13}\}$
3	1.258×10^4	3.783	$\{10.61,\ 4.711\times10^{-14}\}$
4	1.258×10^4	0.9341	$\{1.528,\ 1.625\times10^{-13}\}$

(a) r = 1

i	$\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|_{\boldsymbol{\Lambda}}^2$	a_1	a_0	σ_i
1	478.7	-2.341	1.357	$\{261.5,\ 28.54,\ 9.649\times10^{-13}\}$
2	710.5	0.1952	-1.245	$\{280.7,\ 22.15,\ 2.893\times10^{-14}\}$
3	711.2	-0.1628	-0.8774	$\{279.7,\ 7.672,\ 4.7\times10^{-14}\}$
4	726.8	-1.051	0.02761	$\{280.3,\ 2.588,\ 8.672\times10^{-14}\}$
5	1.211×10^4	-1.919	1.139	$\{67.17,\ 22.24,\ 6.882\times10^{-12}\}$
6	1.256×10^4	0.9374	1.258	$\{15.59,\ 11.72,\ 1.031\times10^{-13}\}$
⋮	⋮	⋮	⋮	⋮
13	1.258×10^4	1.688	3.752	$\{4.404,\ 3.123,\ 5.153\times10^{-14}\}$
14	1.258×10^4	-1.064	2.221	$\{3.049,\ 1.577,\ 5.915\times10^{-14}\}$

(b) r = 2

i	$\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|_{\boldsymbol{\Lambda}}^2$	a_2	a_1	a_0	σ_i
1	408.8	-3.23	3.635	-1.41	$\{298.4,\ 27.67,\ 4.977,\ 2.969\times10^{-11}\}$
2	416	-2.596	2.057	-0.4498	$\{297.7,\ 36.72,\ 4.144,\ 5.812\times10^{-12}\}$
3	420	-2.901	2.781	-0.8747	$\{297.7,\ 34.66,\ 2.33,\ 1.157\times10^{-11}\}$
4	453.2	-1.122	-1.501	1.659	$\{298.9,\ 38.42,\ 22.86,\ 1.835\times10^{-13}\}$
5	464.6	-1.477	-0.6843	1.191	$\{299.1,\ 39.49,\ 6.322,\ 1.977\times10^{-13}\}$
6	474.6	0.3817	-5.055	3.733	$\{299,\ 36.16,\ 12.55,\ 5.089\times10^{-13}\}$
⋮	⋮	⋮	⋮	⋮	⋮
49	1.258×10^4	2.63	5.344	3.51	$\{4.489,\ 3.373,\ 1.397,\ 6.408\times10^{-13}\}$
50	1.258×10^4	-0.1302	1.226	2.074	$\{3.253,\ 1.971,\ 1.513,\ 5.218\times10^{-14}\}$

(c) r = 3

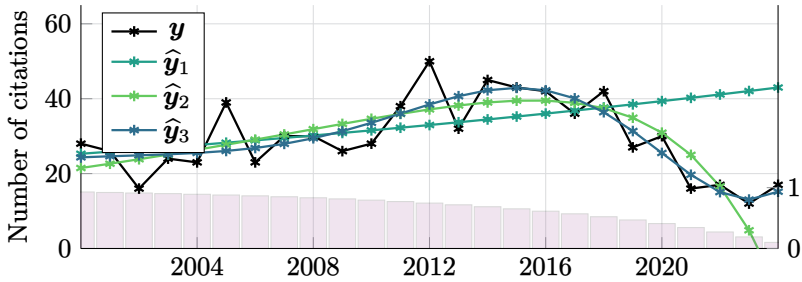


Figure 3.11: The globally optimal data sequences $\widehat{\boldsymbol{y}}_r'$ corresponding to the results in Table 3.12. The weights (3.57) are depicted using purple bars.

Table 3.13: Real-valued solutions to (3.50) corresponding to Example 3.32 with weights (3.58), where σ_i denotes the singular values of $\widehat{\boldsymbol{Y}}$.

i	$\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|_{\boldsymbol{\Lambda}}^2$	a_0	σ_i
1	954.3	-1.013	$\{238.9,\ 4.228\times10^{-14}\ \}$
2	1.644×10^{4}	1.146	$\{21.47,\ 9.242\times10^{-14}\ \}$
3	1.648×10^{4}	0.6885	$\{\ 25.05,\ 5.81\times10^{-14}\ \}$
4	1.651×10^{4}	5.973	$\{\ 1.993,\ 3.217\times10^{-14}\ \}$

(a) r = 1

i	$\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|_{\boldsymbol{\Lambda}}^2$	a_1	a_0	σ_i
1	477.6	-2.253	1.272	$\{256.9,\ 30.18,\ 3.517\times10^{-13}\ \}$
2	883.7	0.1389	-1.166	$\{\ 287,\ 23.96,\ 9.302\times10^{-14}\ \}$
3	929.8	-0.3126	-0.7093	$\{\ 285.7,\ 22.82,\ 2.267\times10^{-14}\ \}$
4	954.3	-0.9615	-0.05195	$\{\ 286.3,\ 0.05492,\ 3.776\times10^{-14}\ \}$
5	1.636×10^{4}	1.889	0.9325	$\{31.5,\ 4.727,\ 4.486\times10^{-12}\ \}$
6	1.636×10^{4}	-1.538	0.8848	$\{\ 25.27,\ 14.63,\ 1.196\times10^{-12}\ \}$
⋮	⋮	⋮	⋮	
18	1.651×10^{4}	0.3272	3.084	$\{8.251,\ 2.739,\ 3.424\times10^{-14}\ \}$
19	1.651×10^{4}	-0.8133	1.999	$\{0.7373,\ 0.6818,\ 4.904\times10^{-14}\ \}$

(b) r = 2

i	$\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|_{\boldsymbol{\Lambda}}^2$	a_2	a_1	a_0	σ_i
1	411.1	-1.106	-1.306	1.451	$\{294.1,\ 41.72,\ 23.24,\ 1.204\times10^{-13}\ \}$
2	420.2	-3.188	3.556	-1.372	$\{299.6,\ 27.18,\ 5.344,\ 8.428\times10^{-12}\ \}$
3	451.5	-1.557	-0.2959	0.8836	$\{293.8,\ 45.03,\ 21.87,\ 1.703\times10^{-13}\ \}$
4	477.5	1.959	-8.219	5.355	$\{294.2,\ 40.91,\ 2.671,\ 2.952\times10^{-13}\ \}$
5	805.1	0.8795	-0.9807	-0.9479	$\{323.1,\ 34.26,\ 6.462,\ 1.336\times10^{-13}\ \}$
6	812.3	-0.4518	0.3616	-0.9417	$\{322.9,\ 24.71,\ 18.97,\ 3.093\times10^{-14}\ \}$
⋮	⋮	⋮	⋮		
57	1.651×10^{4}	5.622	-1.612	2.887	$\{6.274,\ 2.979,\ 1.788,\ 4.497\times10^{-14}\ \}$
58	1.651×10^{4}	5.13	-2.849	11.88	$\{1.711,\ 0.5861,\ 0.1655,\ 4.381\times10^{-14}\ \}$

(c) r = 3

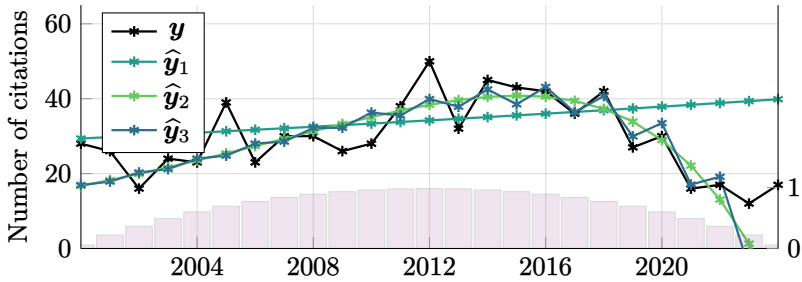


Figure 3.12: The globally optimal data sequences $\widehat{\boldsymbol{y}}_r'$ corresponding to the results in Table 3.13. The weights (3.58) are depicted using purple bars.

3.8 Sequences with Infinite Length

The constraint $\boldsymbol{T}^{\boldsymbol{a}}_{N-n}\widehat{\boldsymbol{y}} = \boldsymbol{0}$ in the LS realization problem enforces $\operatorname{rank}(\widehat{\boldsymbol{Y}}) < n$, with $\widehat{\boldsymbol{Y}} \in \mathbb{R}^{N\times(n+1)}$ the Hankel matrix built from $\widehat{\boldsymbol{y}}$. If $N \to \infty$, then for any finite n the LS realization problem becomes

$$\underset{\widehat{\boldsymbol{y}}}{\text{minimize}} \quad \sum_{k=0}^{\infty}(\boldsymbol{y} - \widehat{\boldsymbol{y}})^2 \tag{3.59a}$$

$$\text{subject to} \quad \operatorname{rank}(\widehat{\boldsymbol{Y}}) \leq n. \tag{3.59b}$$

The rank-deficiency constraint is equivalent to the existence of an nth order linear recurrence relation satisfied by the infinite-length sequence $\widehat{\boldsymbol{y}} \in \mathbb{R}^{\infty}$:

$$a_0\widehat{y}_k + \cdots + a_{n-1}\widehat{y}_{k+n-1} + a_n\widehat{y}_{k+n} = 0.$$

In the context of strictly causal SISO systems—that is, SISO systems for which the transfer function is a strictly proper rational function—this recurrence property holds naturally for the impulse response (excluding its first entry h_0). Indeed, consider a minimal, strictly causal, time-invariant system of McMillan degree n with rational transfer function $H(z) = b(z)/a(z)$, for which the time-domain input-output relation corresponds to

$$a_0\widehat{y}_k + \cdots + a_{n-1}\widehat{y}_{k+n-1} + a_n\widehat{y}_{k+n} = b_0\widehat{u}_k + \cdots + b_{n-1}\widehat{u}_{k+n-1}.$$

If we give an impulse to the system, i.e., choose $u_k = \delta_k$ with

$$\delta_k = \begin{cases} 1 & k = 0, \\ 0 & k > 0, \end{cases}$$

and consider zero initial state, then the output y_k is the impulse response h_k of the system. In this case, for $k > 0$, all terms on the right-hand side vanish, leaving the n-th order homogeneous linear recurrence relation

$$a_0h_k + \cdots + a_{n-1}h_{k+n-1} + a_nh_{k+n} = 0$$

whose coefficients are determined by the system's poles. The first n samples of h_k can be interpreted as initial conditions, subsequent samples ($k \geq n$) are generated by the linear recurrence relation. Any infinite sequence satisfying such a recurrence is therefore the impulse response of some n-th order SISO system (cf. Kronecker's Theorem [105]).

As such, when the infinite sequence $\boldsymbol{y}$ is chosen as the impulse response $\{h_k\}_{k=1,\ldots,\infty}$ of a stable m-th order SISO system (with $m > n$), (3.59) corresponds to the H_2-norm SISO model reduction problem:

- The optimal approximant $\widehat{\boldsymbol{y}}$ represents the impulse response $\{\widehat{h}_k\}_{k=1,\ldots,\infty}$ of a strictly causal SISO system with order at most n.
- Using Parseval's identity, one can show that

$$\sum_{k=0}^{\infty}\left(h_k-\widehat{h}_k\right)^2=\frac{1}{2\pi}\int_0^{2\pi}\left\|H(e^{j\omega})-\widehat{H}(e^{j\omega})\right\|^2\,\mathrm{d}\omega=\left\|H(z)-\widehat{H}(z)\right\|_{H_2}^2,$$

  where $H(z)$ and $\widehat{H}(z)$ are the stable, strictly causal and finite-order systems corresponding to the impulse responses h_k and $\widehat{h}_k$, respectively. I.e., the time-domain cost (sum of squared impulse-response errors) is exactly the H_2-norm of the transfer-function error.

We can conclude that autonomous LS realization (3.2) on an infinite horizon becomes equivalent to the H_2-norm SISO model reduction problem (cf. Chapter 2). Because of this relation between both problems, many of the results obtained in this chapter have a functional counterpart, some of which we encountered in Chapter 2.

Beurling–Lax–Halmos

The orthogonal subspaces defined by the Toeplitz matrix $\boldsymbol{T}_{N-n}^{\boldsymbol{a}}$ are reminiscent of the operator-theoretic result of Beurling–Lax–Halmos [17, 79, 111], which describes how each inner function (the mathematical counterpart of a causal all-pass transfer function[15]) in the Hardy space $\mathcal{H}_2$ of functions analytic on the exterior of the unit-disc, induces an orthogonal decomposition of that space, see, e.g., [141, Chapter 3]. Indeed, for a square summable sequence $\boldsymbol{y}\in\ell_2$, the row space of the doubly-infinite matrix $\boldsymbol{T}_{N-n}^{\boldsymbol{a}}\in\mathbb{R}^{\infty\times\infty}$, in which the approximation error $\widetilde{\boldsymbol{y}}$ resides, constitutes an infinite-dimensional forward shift-invariant subspace, where each row is a forward-shift of the vector $[a_0\ \cdots\ a_n\ 0\ 0\ \cdots]^{\mathsf{T}}\in\mathbb{R}^{\infty}$, corresponding to the Taylor coefficients of the functions $\{z^{n+k}a(z)\}_{k=0,1,\ldots}$. The n-dimensional orthogonal complementary subspace containing the model-compliant sequence $\widehat{\boldsymbol{y}}$, the so-called *model space*, is backward shift-invariant and is completely determined by the roots of $a(z)$: it corresponds to the range of the infinite Vandermonde matrix $\boldsymbol{V}_{\infty}\in\mathbb{R}^{\infty\times n}$, the truncated version of which has been defined in (3.7).

[15] It can be shown that the optimal misfit $\widetilde{\boldsymbol{y}}$ of (3.2) satisfies nth order SISO LTI dynamics: $\widetilde{\boldsymbol{y}}$ corresponds to a particular zero state response of the so-called *misfit model*, which has the *all-pass* completion $a(z)/z^n a(z^{-1})$ as its transfer function. This misfit model is also encountered in the context of SISO and MIMO misfit modeling problems, see, e.g., [144]. The misfit model will be studied in Chapter 7.

Lower-order models

The lower-order models (i.e., the models for which rank $\widehat{\boldsymbol{Y}} < n$) encountered in Section 3.5.2 can be seen as the finite-dimensional equivalent of SISO models with pole-zero cancellations in the context of H_2-norm model reduction. The insights that we obtained in Section 3.5.2 are therefore resemblant to results in the literature on SISO H_2-norm model reduction, e.g., Theorem 3.8 can be seen as the finite-dimensional equivalent of [13, Proposition 1] and [141, Lemma 5.2], whereas Theorem 3.10 is the counterpart of [141, Theorem 5.3].

Walsh's Theorem

Theorem 3.3 is a finite-dimensional version of [141, Theorem 3.14]. The result shows that the optimal misfit $\widetilde{\boldsymbol{y}}$, i.e., the approximation error, in (some of) the critical points of the LS realization problem is the result of filtering an unknown signal $\boldsymbol{g} \in \mathbb{R}^{N-2n}$ twice by the FIR filter $a(z)$, which determines the LTI dynamics of the optimal model. The original result [141, Theorem 3.14], which we encountered in Chapter 2 and named after Walsh to recognize its origins in rational approximation theory [184], states that the approximation error function in the critical points of the SISO H_2-norm model reduction problem can always be expressed in the form $[z^n a(z^{-1})]^2 g(z)$ for some $g(z) \in \mathcal{H}_2$, where the nth order monic polynomial $a(z)$ is the denominator of the transfer function of the optimal, lower-order approximant. Notice that, contrary to the result in the context of model reduction, Theorem 3.3 does not assume any form of stability of the estimated model.

3.9 Conclusions and Future Work

Least squares (LS) realization of autonomous single-output linear time-invariant dynamical models, despite being a well-established topic in the statistical and system identification literature, remained partially unresolved in terms of its *global* solution. We define a global solution as one that identifies the *global minimizer(s)* of the (non-convex) optimization problem by means of a *deterministic* procedure. This approach is fundamentally different from the methods commonly used in engineering practice, which often rely on (heuristic) iterative optimization techniques and invoke asymptotic statistical assumptions about the given data.

As demonstrated in this chapter, the global solution of the (weighted) LS realization problem can be formulated as a rectangular multiparameter

eigenvalue problem (MEP). The MEP is a generalization of the standard eigenvalue problem, where more than one eigenvalue is allowed and the restriction of linearity in its parameter(s) is removed.

We demonstrated how many of the insights from the unweighted setting, such as the orthogonal decomposition of the data space, kernel-based representations, and projections, can be naturally extended to the weighted case through a reinterpretation of the inner product. Several numerical examples illustrated the effects of weighting, particularly in terms of improved modeling flexibility.

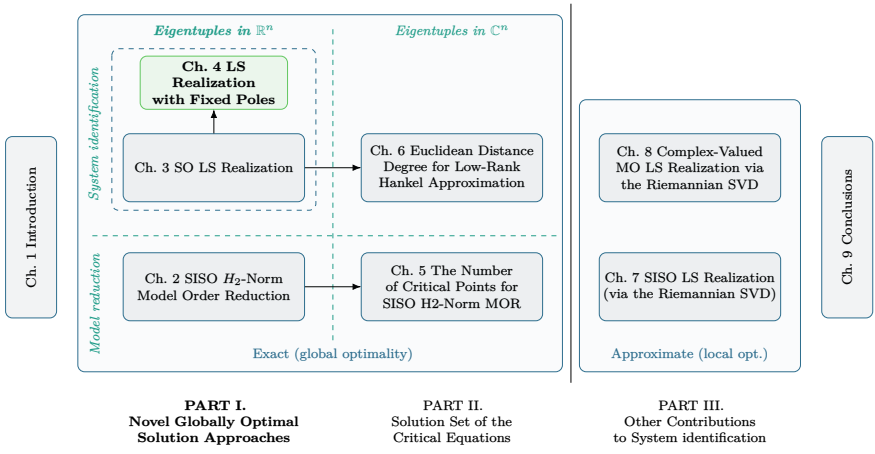
Ultimately, we combined general principles from optimization theory, such as first-order optimality conditions, with problem-specific insights from system theory (e.g., the "double-FIR-filter" property) to obtain our results. This paradigm is undoubtedly relevant beyond the immediate context of the (weighted) LS realization problem. More broadly, we demonstrated that MEPs constitute an effective framework for characterizing the critical points of a (highly) non-convex estimation problem.

As briefly noted, the real-valuedness of the eigentuples is important in the validity of the results presented in this chapter. This raises concerns when solving the resulting MEPs using numerical solvers, which typically operate over the complex field. Another challenge considers the number of critical points. In the special case $n = 1$, finiteness of the affine solution set of the proposed MEP is easily proved via its secular equation. However, even though in all of our numerical examples the set of affine eigentuples of the MEP has been zero-dimensional–it consists of a finite number of isolated elements–finiteness of the affine solution set for all (N, n) remains to be proven. These fundamental questions will be answered in Part II of this dissertation, where we will conduct a detailed algebraic and geometric analysis of the solution set to the obtained MEP.

Recognizing that the computational difficulty involved with solving the obtained MEP grows exponentially with the problem size (N, n), more research is needed to make our theoretical findings applicable in practice. In future work, we will try to exploit the fact that we are only interested in the real-valued eigentuples of the MEP. We also want to perform more numerical examples to get better insights into the nature of the local/global minimizer(s) of the LS realization problem, pushing the problem size (N and/or n) to be as large as possible, e.g., using a supercomputer.

Chapter 4

Incorporating Fixed Pole Information in the Least Squares Realization Problem



Chapter abstract

In practical least squares realization problems, partial information about the poles of the dynamical model may be known *a priori*. Existing techniques for incorporating prior knowledge, such as prefiltering the given data, are typically heuristic and lack theoretical guarantees. We extend our previously developed globally optimal approach for the least squares realization problem to accommodate fixed poles. In particular, we reformulate the problem as a (rectangular) multiparameter eigenvalue problem, the eigenvalues of which characterize all local and global minimizers of the constrained estimation problem. We present numerical examples to demonstrate the effectiveness of the proposed method and experimentally validate this letter's central hypothesis: incorporating *a priori* information on the poles enhances the estimation results.

Candidate's contributions and relation to published work

This work has been previously published as:

- Vermeersch, C., **Lagauw, S.**, and De Moor, B. (2025). "Incorporating Fixed Pole Information in the Data-Driven Least Squares Realization Problem". A preprint of the manuscript is available online at: `https://arxiv.org/abs/2509.09394`.

The candidate was a co-author of the original article, contributing equally with Christof Vermeersch to the research, numerical examples and writing.

The contents differ slightly from the original article:

- The notation and terminology was adapted to ensure consistency with the conventions used throughout the remainder of this dissertation.

4.1 Introduction

Given a sequence of (discrete-time) output data, the *standard least squares realization problem* addresses the question "how can we modify the given data points in a least squares sense so that they become the output of an autonomous, single-output, linear time-invariant dynamical model, and what are the corresponding model parameters". For a predetermined model order, the realized model is completely characterized by the coefficients or roots of its characteristic equation. Although the model class is linear in the parameters, the least squares realization problem of finding the smallest adaptation of the observed output data is nonlinear in both the parameters and misfit, resulting in a non-convex optimization problem with potentially many local optima (see Figure 4.1).

Many approaches for determining the coefficients or poles of the model exist, with first occurrences in the literature dating back to Prony in the 18th century [85]. Recently, globally optimal approaches that minimize the misfit between the observed and modified output data have been proposed by some of the authors [45, 110].

Most methods for solving the least squares realization problem, including the globally optimal approaches in [45, 110], only consider the observed output data and a predetermined model order. However, some of the model's poles may be known ("fixed") in practical settings. Such situations commonly arise in process control, structural dynamics, and econometrics, where certain modes are either predetermined or estimated from first principles. Similarly, in the context of parameter estimation for exponential signals, information about the signal properties may be available from previous experiments, inherent (physical) properties of the experimental setup, or dynamics introduced to model noise disturbances [27, 26]. The claim that using this information improves the estimates of the remaining unknown poles in a certain sense is this letter's central hypothesis. Therefore, we extend in this chapter one of our previously developed globally optimal methods to incorporate fixed pole information.

The idea that using information about the fixed pole locations may improve the estimates of the remaining unknown poles is not new. Three alternative versions of Prony's method were already described in [12], each constraining the original method in a different way to handle the fixed poles. The linearly constrained total least squares (TLS) method employs linear constraints to incorporate *a priori* knowledge about the poles of the model [50]. And, in the context of nuclear magnetic resonance spectroscopy data, the Hankel total least squares (HTLS) method has been extensively used for spectral estimation [93], with several variants, the so-called HTLS-PK methods [26, 27], incorporating *a*

priori knowledge on the signal properties. Alternatively, rather than adapting parameter estimation techniques to deal with the fixed poles, one could also prefilter the observed output data in an attempt to remove the effects of the fixed poles before estimation. A naive approach is to start the realization procedure from the optimal misfit associated to the lower-order model defined by the given fixed poles. The more advanced time series deflation (TSD) technique [122] applies a moving-average prefilter to the observed output data to remove the known components.

Although these heuristic methods provide practical solutions, a globally optimal approach for dealing with fixed poles does not exist. The approach that we take in this chapter starts from our globally optimal methodology in [110], the so-called *standard globally optimal realization technique*. If one or more poles of the model are known beforehand, then the new *fixed pole globally optimal realization technique* finds the globally optimal approximation of the remaining, unknown poles. We end up with a (rectangular) multiparameter eigenvalue problem (MEP), which generalizes the polynomial eigenvalue problem by allowing multiple spectral parameters. It is the first globally optimal approach that takes the fixed poles into account. It is fundamentally different than the above-mentioned heuristics, replacing engineering practice by exactness, but at the cost of a higher computational complexity (similar as in [110]). The statistical experiment in the last numerical example validates the central hypothesis that *a priori* information improves the estimation in the presence of noise.

The remainder of the chapter continues as follows: We give a mathematical description of the problem and a motivational example in Section 4.2. Next, in Section 4.3, we explain how we obtain, via the first-order optimality conditions, an MEP. *The MEP reformulation of the fixed poles least squares realization problem is our main contribution.* We solve the motivational example numerically and perform a statistical experiment in Section 4.4. Finally, we conclude this chapter and look at opportunities for future work in Section 4.5.

4.2 Problem Formulation and Motivational Example

We consider an autonomous, single-output, linear time-invariant dynamical model in discrete time with a predefined model order n that corresponds to the number of states. Model-compliant output data $\widehat{\boldsymbol{y}} \in \mathbb{R}^N$, with the $N > 2n$ data points y_k, satisfy a difference equation of the form

$$a_0\widehat{y}_k + a_1\widehat{y}_{k+1} + + \cdots a_n\widehat{y}_{k+n} = a(z)\widehat{y}_k = 0, \tag{4.1}$$

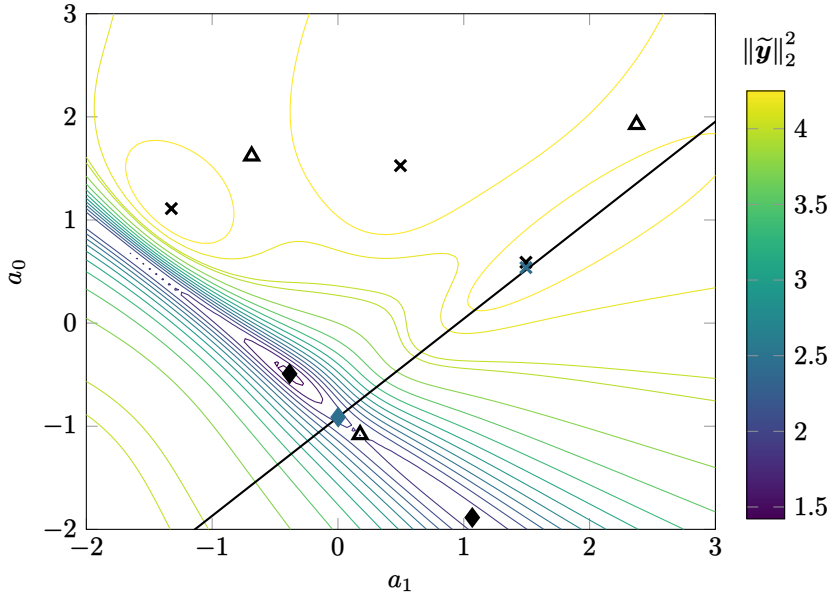


Figure 4.1: Objective function $\|\widetilde{\boldsymbol{y}}\|_2^2$ of the least squares realization problem in Example 4.3, plotted against the model parameters a_0 and a_1, normalized by setting $a_2 = 1$. The surface illustrates the non-convex nature of the underlying optimization problem, with critical points of the standard least squares objective function including minimizers (♦), saddle points (△), and maximizers (×). Given a fixed pole, the fixed pole least squares realization problem is subject to an additional constraint (–) leading to different optima. A different minimizer (♦) and maximizer (×) are obtained in this example.

for all $k = 0, 1, 2, \ldots, N-n-1$, where $a(z) = a_0 + a_1 z + \ldots + a_n z^n$ is a degree n polynomial in the forward-shift operator z (i.e., $z\widehat{y}_k = \widehat{y}_{k+1}$). Writing out (4.1) for every k,

$$\underbrace{\begin{bmatrix} a_0 & \cdots & a_{n-1} & a_n & 0 & \cdots & 0 \\ 0 & a_0 & \cdots & a_{n-1} & a_n & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_0 & \cdots & a_{n-1} & a_n \end{bmatrix}}_{\boldsymbol{T}_{N-n}^{\boldsymbol{a}}} \widehat{\boldsymbol{y}} = \boldsymbol{0} \tag{4.2}$$

expresses the model via the (banded) Toeplitz matrix $\boldsymbol{T}_{N-n}^{\boldsymbol{a}} \in \mathbb{R}^{(N-n)\times N}$ for the coefficient vector $\boldsymbol{a} = \begin{bmatrix} a_0 & a_1 & \cdots & a_n \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{n+1}$. The null space of $\boldsymbol{T}_{N-n}^{\boldsymbol{a}}$ is

spanned by Vandermonde vectors generated by the roots of $a(z)$. Consequently, the model-compliant output data $\widehat{\boldsymbol{y}}$ can be expressed as a weighted sum (with coefficients x_i) of exactly n linearly independent (confluent[1]) Vandermonde vectors $\boldsymbol{p_i} = \begin{bmatrix} 1 & p_i & \cdots & p_i^{N-1} \end{bmatrix}^\mathsf{T} \in \mathbb{C}^N$ for each pole p_i:

$$\widehat{\boldsymbol{y}} = \sum_{i=1}^{n} \boldsymbol{p_i} x_i. \tag{4.3}$$

Definition 4.1. *Given observed output data $\boldsymbol{y} \in \mathbb{R}^N$, the* standard least squares realization problem *seeks to determine the model parameters $\boldsymbol{a} \in \mathbb{R}^{n+1}$ of the nth order model that minimizes the l_2 norm of the misfit $\widetilde{\boldsymbol{y}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$. It corresponds to the following optimization problem:*

$$\begin{aligned} \underset{\boldsymbol{a}, \widehat{\boldsymbol{y}}}{\text{minimize}}\ & \frac{1}{2} \left\| \widetilde{\boldsymbol{y}} \right\|_2^2 = \frac{1}{2} \left\| \boldsymbol{y} - \widehat{\boldsymbol{y}} \right\|_2^2, \\ \textit{subject to}\ & \boldsymbol{T}_{N-n}^{\boldsymbol{a}} \widehat{\boldsymbol{y}} = \boldsymbol{0}, \\ & \langle \boldsymbol{a}, \boldsymbol{e} \rangle = 1, \end{aligned} \tag{4.4}$$

where the last constraint avoids the trivial solution $\boldsymbol{a} = \boldsymbol{0}$.

The *normalization vector* $\boldsymbol{e} \in \mathbb{R}^{n+1}$ in (4.4) must have at least one nonzero element. A commonly used choice is $\boldsymbol{e} = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}^\mathsf{T}$, which boils down to setting $a_n = 1$.

Suppose that $m < n$ poles, $\{p_i\}_{i=1,2,\ldots,m}$, of the model are fixed, such that the model to be estimated can be factorized as

$$a(z) = b(z) \prod_{i=1}^{m} (z - p_i) = b(z) c(z), \tag{4.5}$$

where $b(z)$ is a qth order polynomial in the forward-shift z with unknown coefficients $\boldsymbol{b} = \begin{bmatrix} b_0 & \cdots & b_{q-1} & b_q \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{q+1}$, in which we introduced $q = n - m$ for notational convenience. The coefficients $\boldsymbol{c} = \begin{bmatrix} c_0 & \cdots & c_{m-1} & c_m \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{m+1}$ of the mth order polynomial $c(z) = \prod_{i=1}^{m} (z - p_i)$ follow directly from the fixed poles. After defining (banded) Toeplitz matrices $\boldsymbol{T}_{N-q}^{\boldsymbol{b}}$ and $\boldsymbol{T}_{N-n}^{\boldsymbol{c}}$ from $\boldsymbol{b}$ and $\boldsymbol{c}$, the factorization in (4.5) can be used to rewrite (4.2) as

$$\boldsymbol{T}_{N-n}^{\boldsymbol{a}} \widehat{\boldsymbol{y}} = \boldsymbol{T}_{N-n}^{\boldsymbol{c}} \boldsymbol{T}_{N-q}^{\boldsymbol{b}} \widehat{\boldsymbol{y}} = \boldsymbol{0}. \tag{4.6}$$

[1] Since our methodology uses the coefficients of $a(z)$ to parameterize the model, it holds for models with simple and multiple poles. More information about the case with multiple poles can be found in [45].

The order of $\boldsymbol{T}^{\boldsymbol{b}}_{N-q}$ and $\boldsymbol{T}^{\boldsymbol{c}}_{N-n}$ can be changed as long as the dimensions are adjusted appropriately. Indeed,

$$\boldsymbol{T}^{\boldsymbol{c}}_{N-n}\boldsymbol{T}^{\boldsymbol{b}}_{N-q} = \boldsymbol{T}^{\boldsymbol{b}}_{N-n}\boldsymbol{T}^{\boldsymbol{c}}_{N-m}. \tag{4.7}$$

Requiring the model $a(z)$ to have certain poles in (4.3), fixes m of the Vandermonde vectors beforehand:

$$\widehat{\boldsymbol{y}} = \sum_{i=1}^{m} \boldsymbol{p_i} x_i + \sum_{i=m+1}^{n} \boldsymbol{p_i} x_i = \sum_{i=1}^{m} \boldsymbol{p_i} x_i + \widehat{\boldsymbol{y}}'. \tag{4.8}$$

Definition 4.2. *Given observed output data $\boldsymbol{y} \in \mathbb{R}^N$ and m fixed poles $\{p_i\}_{i=1,2,\ldots,m}$, the* fixed pole least squares realization problem *seeks to determine the model parameters $\boldsymbol{b} \in \mathbb{R}^{q+1}$ of the unknown factor of the nth order model that minimizes the ℓ_2-norm of the misfit $\widetilde{\boldsymbol{y}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$. It corresponds to the following optimization problem:*

$$\begin{aligned} &\underset{\boldsymbol{b},\widehat{\boldsymbol{y}}}{\text{minimize}}\, \frac{1}{2}\left\|\widetilde{\boldsymbol{y}}\right\|_2^2 = \frac{1}{2}\left\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\right\|_2^2, \\ &\text{subject to}\, \boldsymbol{T}^{\boldsymbol{c}}_{N-n}\boldsymbol{T}^{\boldsymbol{b}}_{N-q}\widehat{\boldsymbol{y}} = \boldsymbol{0}, \\ &\qquad\qquad \langle \boldsymbol{b}, \boldsymbol{e} \rangle = 1, \end{aligned} \tag{4.9}$$

where the last constraint avoids the trivial solution $\boldsymbol{b} = \boldsymbol{0}$.

One could wonder whether an adaptation of the least squares methodology is truly necessary to be able to cope with fixed poles. Would it not suffice to remove the effects of the given poles from the observed output data, i.e., to prefilter the data and use standard least squares realization techniques to estimate the remaining poles? We motivate our methodology by considering two (heuristic) prefiltering approaches.

Example 4.3 (Motivational example)**.** *Given the output data*

$$\boldsymbol{y} = \begin{bmatrix} 3.0000 \\ 5.0000 \\ 2.0000 \\ 3.0000 \\ 4.0000 \\ 2.0000 \\ 3.0000 \end{bmatrix},$$

the sole pole of the globally optimal least squares $n = 1$ estimation is[2]

$$p_1 = -0.9557. \tag{4.10}$$

[2]We adopt a four-digit precision convention for all floating-point numbers.

Suppose that we want to estimate a second-order ($n = 2$) model, for which one of its poles is fixed by p_1. We will compute the globally optimal solution for the second pole with the methodology of this chapter in Example 4.8, resulting in $p_2 = 0.9538$ with $\|\widetilde{\boldsymbol{y}}\|_2^2 = 5.9112$. The objective functions $\|\widetilde{\boldsymbol{y}}\|_2^2$ for the standard (4.4) and fixed pole optimization problem (4.9) are visualized in Figure 4.1.

By using an orthogonal projection (see (12) in [110]), we can compute the misfit $\widetilde{\boldsymbol{y}}'$ associated with the first-order model defined by p_1. We estimate the second pole by performing another $n = 1$ estimation, this time using the misfit $\widetilde{\boldsymbol{y}}'$ as the 'observed' output data, leading to an estimate of the most dominant pole, $p_2 = 0.8630$, present in the first-order misfit $\widetilde{\boldsymbol{y}}'$. This naive prefiltering approach results in a misfit $\widetilde{\boldsymbol{y}}$ with $\|\widetilde{\boldsymbol{y}}\|_2^2 = 8.4181$, which is not globally optimal; $p_2 = 0.9538$ in Example 4.8 provides a smaller misfit. Alternatively, we can prefilter the observed output data $\boldsymbol{y}$ via the TSD technique described in [122]:

$$\boldsymbol{y}' = \boldsymbol{T}_{N-m}^{\boldsymbol{c}} \boldsymbol{y}.$$

After solving the standard least squares realization problem (4.4) for a first-order model with $\boldsymbol{y}'$ as the 'observed' output data, we find a second $p_2 = 0.9361$, yielding also a larger misfit than with $p_2 = 0.9538$, namely $\|\widetilde{\boldsymbol{y}}\|_2^2 = 6.0070$. The observation that both prefiltering approaches are suboptimal w.r.t. (4.9) remains valid for other problem setups (i.e., different N, n, or m).

4.3 Methodology

Below, we characterize all minimizers of the fixed pole optimization problem (4.9) by adapting the standard globally optimal realization methodology from [110] and establishing a globally optimal realization approach for the fixed pole setting. We start by reformulating the first-order necessary conditions for optimality as matrix rank problem (Section 4.3.1). After showing that the minimizers result in a given rank (Section 4.3.2), we solve the fixed pole problem via an MEP (Section 4.3.3). Finally, we discuss how the obtained MEP can be solved to find the globally optimal model subject to the fixed poles (Section 4.3.4).

4.3.1 Characterizing first-order optimality

With the introduction of the Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^{N-n}$ and $\mu \in \mathbb{R}$, the Lagrangian of (4.9) is equal to

$$\mathcal{L}(\boldsymbol{b}, \widehat{\boldsymbol{y}}, \boldsymbol{\lambda}, \mu) = \frac{1}{2} \|\widetilde{\boldsymbol{y}}\|_2^2 + \boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{T}_{N-n}^{\boldsymbol{c}} \boldsymbol{T}_{N-q}^{\boldsymbol{b}} \widehat{\boldsymbol{y}} + \mu \left(\boldsymbol{b}^{\mathsf{T}} \boldsymbol{e} - 1 \right).$$

The first-order necessary conditions for optimality can be obtained via the partial derivatives as

$$\partial \mathcal{L}/\partial \boldsymbol{b} = \widehat{\boldsymbol{Y}}^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{c}}_{N-n})^{\mathsf{T}}\boldsymbol{\lambda} + \mu \boldsymbol{e} = \boldsymbol{0}, \tag{4.11}$$

$$\partial \mathcal{L}/\partial \widehat{\boldsymbol{y}} = \widehat{\boldsymbol{y}} - \boldsymbol{y} + (\boldsymbol{T}^{\boldsymbol{b}}_{N-q})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{c}}_{N-n})^{\mathsf{T}}\boldsymbol{\lambda} = \boldsymbol{0}, \tag{4.12}$$

$$\partial \mathcal{L}/\partial \boldsymbol{\lambda} = \boldsymbol{T}^{\boldsymbol{c}}_{N-n}\boldsymbol{T}^{\boldsymbol{b}}_{N-q}\widehat{\boldsymbol{y}} = \boldsymbol{T}^{\boldsymbol{c}}_{N-n}\widehat{\boldsymbol{Y}}\boldsymbol{b} = \boldsymbol{0}, \tag{4.13}$$

$$\partial \mathcal{L}/\partial \mu = \boldsymbol{b}^{\mathsf{T}}\boldsymbol{e} - 1 = 0, \tag{4.14}$$

where $\widehat{\boldsymbol{Y}} \in \mathbb{R}^{(N-q)\times(q+1)}$ denotes the Hankel matrix constructed from the model-compliant output data $\widehat{\boldsymbol{y}}$. The real solutions that satisfy (4.11)–(4.14) are the *critical points* of the problem. Multiplying (4.11) from the left by $\boldsymbol{b}^{\mathsf{T}}$ yields

$$\boldsymbol{b}^{\mathsf{T}}\widehat{\boldsymbol{Y}}^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{c}}_{N-n})^{\mathsf{T}}\boldsymbol{\lambda} + \mu \boldsymbol{b}^{\mathsf{T}}\boldsymbol{e} = \boldsymbol{0} \iff \mu = 0,$$

by using that $\boldsymbol{T}^{\boldsymbol{c}}_{N-n}\widehat{\boldsymbol{Y}}\boldsymbol{b} = \boldsymbol{0}$, from (4.13), and $\boldsymbol{b}^{\mathsf{T}}\boldsymbol{e} = 1$, from (4.14). Therefore, we can set $\mu = 0$ from this point on. As such, (4.11) implies that

$$\widehat{\boldsymbol{Y}}^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{c}}_{N-n})^{\mathsf{T}}\boldsymbol{\lambda} = \boldsymbol{0} \iff \boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{T}^{\boldsymbol{c}}_{N-n}\widehat{\boldsymbol{Y}} = \boldsymbol{0}, \tag{4.15}$$

meaning that the vector of Lagrange multipliers $\boldsymbol{\lambda}$ lies in the left null space of the filtered Hankel matrix $\widehat{\boldsymbol{Y}}' = \boldsymbol{T}^{\boldsymbol{c}}_{N-n}\widehat{\boldsymbol{Y}} \in \mathbb{R}^{(N-n)\times(q+1)}$. We know from (4.13) that the filtered Hankel matrix has to be rank-deficient, such that

$$\operatorname{rank} \widehat{\boldsymbol{Y}}' \leq q. \tag{4.16}$$

Furthermore, (4.13) shows that the output data $\widehat{\boldsymbol{y}}$ are model-compliant w.r.t. $a(z)$, while the filtered output data $\widehat{\boldsymbol{y}}' = \boldsymbol{T}^{\boldsymbol{c}}_{N-n}\widehat{\boldsymbol{y}}$ are model-compliant w.r.t. $b(z)$.

4.3.2 Solutions with lower-order dynamics

Similar as in the standard case without fixed poles [110], we show that for all (local) minimizers of the fixed pole optimization problem (4.9) equality holds in (4.16); a solution with lower-order dynamics can not be a (local) minimizer of (4.9). Strict inequality in (4.16) implies that the model-compliant output data $\widehat{\boldsymbol{y}}$ do not employ all degrees of freedom allowed by $a(z)$. At least one of the poles of $a(z)$ can be removed without changing the optimal $\widehat{\boldsymbol{y}}$.

Lemma 4.4. *Given output data $\widehat{\boldsymbol{y}}$ that comply with an nth order model $b(z)\prod_{i=1}^{m}(z-p_i)$, of which the parameters $\boldsymbol{b}$ correspond to a (local) minimizer of* (4.9)*. The corresponding filtered model-compliant output data $\widehat{\boldsymbol{y}}' = \boldsymbol{T}^{\boldsymbol{c}}_{N-n}\widehat{\boldsymbol{y}}$ have exactly qth order dynamics, i.e.,* $\operatorname{rank} \widehat{\boldsymbol{Y}}' = q$.

Proof. A formal proof of Lemma 4.4 is omitted, but the necessary steps follow closely the reasoning in [110, Section 5]. In particular, the multiplication of $\widehat{\boldsymbol{y}}$ with $\boldsymbol{T}^{\boldsymbol{c}}_{N-n}$ annihilates the first m terms in (4.8), so that $\widehat{\boldsymbol{y}}' = \boldsymbol{T}^{\boldsymbol{c}}_{N-n}\widehat{\boldsymbol{y}} = \sum_{i=m+1}^{n} \boldsymbol{p_i} x_i$ is determined solely by the zeros of $b(z)$. If $\operatorname{rank} \widehat{\boldsymbol{Y}}' < q$ with $\widehat{\boldsymbol{Y}}' = \boldsymbol{T}^{\boldsymbol{c}}_{N-n}\widehat{\boldsymbol{Y}}$, then the model $b(z)$ contains at least one non-contributing pole, i.e., a pole that does not influence $\widehat{\boldsymbol{y}}'$, and it can be removed without altering $\widehat{\boldsymbol{y}}$ or $\widehat{\boldsymbol{y}}'$. The orthogonality induced by the least squares measure implies that a small perturbation of the coefficients of $b(z)$ (which only modifies the non-contributing pole) leads to a strict reduction in the l_2 norm of the misfit. □

4.3.3 Multiparameter eigenvalue problem

If equality holds in (4.16), then the filtered Hankel matrix $\widehat{\boldsymbol{Y}}'$ has a left null space of dimension $N - 2n + m$. From (4.13) and (4.7), it follows that

$$\boldsymbol{T}^{\boldsymbol{b}}_{N-2n+m}\boldsymbol{T}^{\boldsymbol{c}}_{N-n}\widehat{\boldsymbol{Y}} = \boldsymbol{0},$$

which, via (4.15), implies that for all critical points for which equality holds in (4.16), we can express $\boldsymbol{\lambda}$ as

$$\boldsymbol{\lambda} = (\boldsymbol{T}^{\boldsymbol{b}}_{N-2n+m})^{\mathsf{T}}\boldsymbol{g}, \tag{4.17}$$

for some $\boldsymbol{g} \in \mathbb{R}^{N-2n+m}$. Substitution of (4.17) into (4.12) and multiplication from the left by $\boldsymbol{T}^{\boldsymbol{a}}_{N-n} = \boldsymbol{T}^{\boldsymbol{c}}_{N-n}\boldsymbol{T}^{\boldsymbol{b}}_{N-q}$ leads to a characterization for this subset of critical points:

$$-\boldsymbol{T}^{\boldsymbol{a}}_{N-n}\boldsymbol{y} + \boldsymbol{T}^{\boldsymbol{a}}_{N-n}(\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{b}}_{N-2n+m})^{\mathsf{T}}\boldsymbol{g} = \boldsymbol{0}. \tag{4.18}$$

Together with the normalization (4.14), (4.18) is a system of $N-n+1$ (at most quartic) polynomial equations in the $N-n+1$ variables $\boldsymbol{b}$ and $\boldsymbol{g}$.

Theorem 4.5. *Consider the MEP, $\boldsymbol{M}(\boldsymbol{b})\boldsymbol{z} = \boldsymbol{0}$, defined by the $(N-n) \times (N - 2n + m + 1)$ cubic matrix polynomial*

$$\boldsymbol{M}(\boldsymbol{b}) = \begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-n}\boldsymbol{y} & \boldsymbol{T}^{\boldsymbol{a}}_{N-n}(\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{b}}_{N-2n+m})^{\mathsf{T}} \end{bmatrix}, \tag{4.19}$$

where $\boldsymbol{T}^{\boldsymbol{a}}_{N-n}$ depends on $\boldsymbol{b}$ through $\boldsymbol{T}^{\boldsymbol{a}}_{N-n} = \boldsymbol{T}^{\boldsymbol{c}}_{N-n}\boldsymbol{T}^{\boldsymbol{b}}_{N-q}$. The eigenvalues are values of $\boldsymbol{b} \in \mathbb{R}^{q+1}$ for which the matrix polynomial $\boldsymbol{M}(\boldsymbol{b})$ drops rank so that there exists a nonzero vector $\boldsymbol{z} \in \mathbb{C}^{N-2n+m+1}$. Each real-valued, affine eigenvalue of (4.19) *corresponds to a critical point of* (4.9) *and the (local) minimizers are a subset of these real affine eigenvalues.*

Proof. The multiparameter formulation in Theorem 4.5 follows from Lemma 4.4 and the opening paragraph. As long as $\boldsymbol{a}$ is real, the product of the three Toeplitz matrices in (4.19) is of full column rank, such that the first entry of the eigenvector $\boldsymbol{z}$ can always be normalized. □

When $\boldsymbol{e} = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}^{\mathsf{T}}$ is used as the non-triviality constraint, the normalization equation (4.14) can be eliminated from the system by substituting $b_q = 1$ in (4.18). A similar adaptation of the MEP in (4.19) is possible.

Remark 4.6. *Our previously developed methodology in [110] and the methodology that follows from Theorem 4.5 are both globally optimal, but w.r.t. a different optimization problem. To avoid confusion, we call the approach from [110] the standard globally optimal realization (S-GOR) technique and the approach that follows from Theorem 4.5 the fixed pole globally optimal realization (FP-GOR) technique.*

Remark 4.7. *Recall that in the TSD technique [122] the observed output data $\boldsymbol{y}$ are prefiltered to remove the effects of the fixed poles and standard least squares realization techniques are used on the obtained $\boldsymbol{y}'$. As became evident in Example 4.3, this approach is heuristic since the prefilter is applied to the observed output data $\boldsymbol{y}$ and not to the model-compliant output data $\widehat{\boldsymbol{y}}$, so that also the misfit is prefiltered. The FP-GOR technique essentially applies the same prefilter, but it does so on the to-be-estimated model-compliant output data $\widehat{\boldsymbol{y}}$. By changing the order of the Toeplitz matrices in* (4.6)*, as shown in* (4.7)*, we get that $\boldsymbol{T}_{N-n}^{\boldsymbol{b}}\boldsymbol{T}_{N-m}^{\boldsymbol{c}}\widehat{\boldsymbol{y}} = \boldsymbol{T}_{N-n}^{\boldsymbol{b}}\widehat{\boldsymbol{y}}' = \boldsymbol{0}$, illustrating that the dynamics of the prefiltered model-compliant output data $\widehat{\boldsymbol{y}}' = \boldsymbol{T}_{N-m}^{\boldsymbol{c}}\widehat{\boldsymbol{y}}$ are completely determined by the unknown poles $\{p_i\}_{i=m+1}^{n}$.*

4.3.4 How to solve the MEP in Theorem 4.5?

In this letter, we refrain ourselves of diving too deep into possible solution algorithms for MEPs[3]. Two software toolboxes exist to solve these problems via structured matrices, namely MacaulayLab [177] and MultiParEig [138]. The former uses the block Macaulay matrix to construct a joint generalized eigenvalue problem that has the affine eigenvalues as its solutions [175], while the latter uses operator determinants or randomized sketching to reach that same goal [90]. More information can be found in [178].

While solving the MEP in Theorem 4.5, it becomes apparent that the more poles are known beforehand, the fewer eigenvalue parameters are contained in the

[3]Note that the multivariate polynomial system in (4.18) could also be tackled directly by polynomial root-finding algorithms (such as homotopy continuation methods and resultant-based approaches).

MEP, such that the computational complexity of determining the eigenvalues decreases for increasing m. There are two interesting limit cases:

- Theorem 4.5 results in the MEP from the S-GOR technique when $m = 0$.
- For $q = 1$ and $b_1 = 1$, the matrix polynomial in (4.19) is a square, univariate matrix polynomial in b_0.

4.4 Numerical Examples

Several numerical examples are worked out to illustrate the obtained methodology. The non-triviality constraint $\boldsymbol{e} = \begin{bmatrix} 0 & \dots & 0 & 1 \end{bmatrix}^{\mathsf{T}}$ is used in each of them, so that the normalization equation (4.14) and the variable b_q can be eliminated.

Example 4.8 (Example 4.3 continued)**.** *Consider the same observed output data $\boldsymbol{y}$ as in Example 4.3. Suppose, similarly as in Example 4.3, that we want to estimate the globally optimal second-order ($n = 2$) model, for which one of the poles ($m = 1$) is fixed by* (4.10)*. The resulting multivariate polynomial system* (4.18) *has* 5 *equations in the variables b_0 and $\boldsymbol{g}$, corresponding to a cubic polynomial eigenvalue problem* (4.19) *with 5×5 coefficient matrices. Out of the* 13 *affine eigenvalues, only one is real-valued: $b_0 = -0.9538$. The globally optimal solution has $p_2 = 0.9538$ as second pole and misfit $\|\widetilde{\boldsymbol{y}}\|_2^2 = 5.9112$. The associated model is*

$$\begin{aligned} a(z) = (z - p_1)b(z) &= (z + 0.9557)(z - 0.9538) \\ &= z^2 + 0.0019z - 0.9116. \end{aligned}$$

Now, consider the standard least squares problem where both poles can be chosen freely, i.e., we do not fix the pole p_1 beforehand. The globally optimal solution, $p_1 = -0.5351$ and $p_2 = 0.9194$, is found by solving a cubic two-parameter eigenvalue problem (via the S-GOR technique). The global minimum of this $n = 2$ fit, $\|\widetilde{\boldsymbol{y}}\|_2^2 = 3.8836$, is smaller than the global minimum of the $n = 1$ fit with one given pole.

The best model for both techniques (and the prefiltering approaches) is given in Table 4.1. Although (global) optimality is achieved in each intermediate step (the first $n = 1$ fit and the second $n = 2, m = 1$ fit), we observe that the nth order model obtained by concatenating the retrieved poles in FP-GOR differs from the globally optimal solution to the nth order S-GOR problem.

Table 4.1: Results for solving the motivational example via different realization approaches. The naive prefilter (NPF) approach uses the optimal misfit associated to the first-order model, while the advanced prefilter applies the TSD technique from [122]. The FP-GOR technique finds the optimal solution subject to one given fixed pole, but the S-GOR technique is not restricted by any *a priori* information.

approach	(n,m)	$\|\widetilde{\boldsymbol{y}}\|_2^2$	p_1	p_2
NPF (heuristic)	$(2,1)$	8.4181	$-0.9557^{\S}$	0.8630
TSD (heuristic)	$(2,1)$	6.0070	$-0.9557^{\S}$	0.9361
FP-GOR (optimal)	$(2,1)$	5.9112	$-0.9557^{\S}$	0.9538
S-GOR (optimal)	$(2,0)$	3.8836	-0.5351	0.9194

§This pole was fixed in the motivational example.

Notice that an optimal $n = 1$ fit as fixed pole does not automatically lead to an optimal higher-order fit. A similar conclusion could be seen when we start from an optimal $n = 1$ fit and recursively increase the degree by one. Recursively applying FPR will not result in as good results as directly using GOR. Example 4.8 illustrates a broader principle: the fact that each intermediate step in a methodology is executed optimally does not necessarily imply that the overall sequence of steps constitutes an optimal procedure.

Example 4.9. *As a sanity check, we now perform an $(n,m) = (2,1)$ estimate with one of the poles obtained form the S-GOR technique, i.e., $(n,m) = (2,0)$. We take $p_1 = -0.5351$ from Table 4.1 and use the FP-GOR technique to determine a globally optimal estimate for p_2. The result of solving* (4.18) *or* (4.19) *is $p_2 = 0.9194$, which is the same as the globally optimal result obtained via the S-GOR technique.*

During the next example, we verify whether the motivation for the adapted methodology was correct: "Does incorporating *a priori* information about a subset of the poles lead to improved statistical accuracy w.r.t. the estimates of the model-compliant data and/or the remaining poles?"

Example 4.10. *Consider the third-order $n = 3$ autonomous dynamical model with poles $p_{1,2} = e^{\pm 0.8j}$ and $p_3 = -0.75$ defined by the state-space model $(\boldsymbol{A}, \boldsymbol{C})$:*

$$\boldsymbol{A} = \boldsymbol{T}^{-1} \begin{bmatrix} \mathrm{Re}(p_1) & -\mathrm{Im}(p_1) & 0 \\ \mathrm{Im}(p_1) & \mathrm{Re}(p_1) & 0 \\ 0 & 0 & p_3 \end{bmatrix} \boldsymbol{T}, \qquad \boldsymbol{C} = \begin{bmatrix} 2 & 2 & 2 \end{bmatrix},$$

where $\boldsymbol{T} \in \mathbb{R}^{3\times 3}$ *is a non-singular transformation matrix. We can use this model to generate the true output data* $\boldsymbol{x} \in \mathbb{R}^N$*:*

$$x_k = \boldsymbol{C}\boldsymbol{A}^k\boldsymbol{x_0}, \quad k = 0, \dots, N-1,$$

with initial state $\boldsymbol{x}_0 = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\mathsf{T}$*. The observed output data* $\boldsymbol{y} \in \mathbb{R}^N$ *are obtained from the following noisy setup:*

$$y_k = x_k + \sigma\epsilon_k, \quad k = 0, \dots, N-1, \tag{4.20}$$

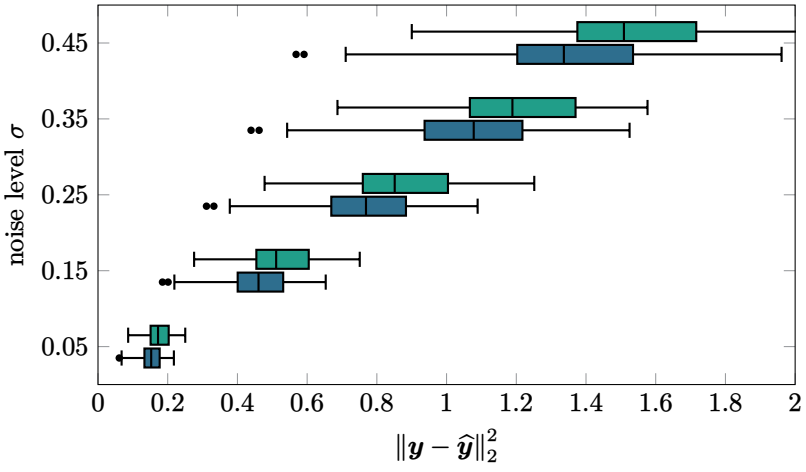
where $\epsilon_k \sim \mathcal{N}(0,1)$ *such that* $\sigma\epsilon_k \sim \mathcal{N}(0, \sigma^2)$*.*

Given sequences $\boldsymbol{y}$ *of length* $N = 16$*, we consider two alternative approaches to obtain an estimate for* p_3*: by using the S-GOR technique to find all three poles without* a priori *information* $(n, m) = (3, 0)$ *and by using the FP-GOR technique, in which we assume the complex pole pair* $p_{1,2} = e^{\pm 0.8j}$ *to be fixed while computing the remaining pole* p_3 $(n, m) = (3, 2)$*. We compute estimates of* p_3 *for* 50 *different realizations of the random noise sequence* ϵ_k *and repeat this experiment for several noise levels* $\sigma \in \{0.05, 0.15, \dots, 0.45\}$*. The results are summarized in Figure 4.2.*

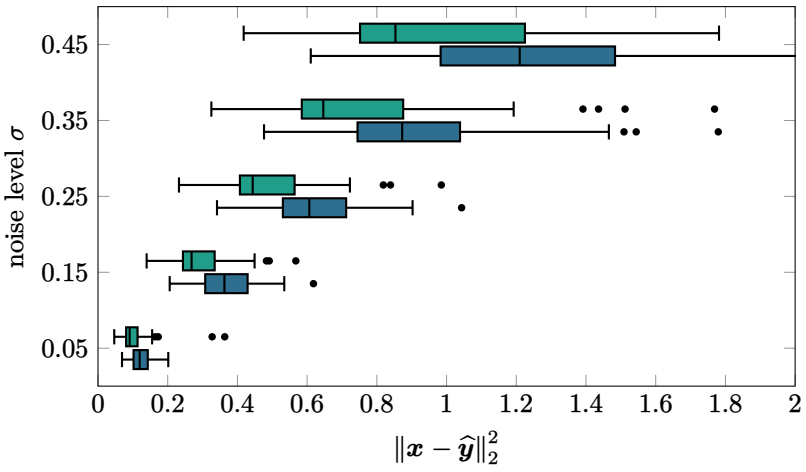
For the S-GOR technique, the MEP has 6466 *solutions, out of which only a small subset is real-valued, while the FP-GOR technique requires solving a univariate problem that has only* 37 *solutions. The standard approach is computationally much more expensive than the fixed pole approach.*

It is clear from Figure 4.2a that the misfits of the S-GOR approach are consistently lower than the ones obtained using the FP-GOR approach. This is as expected since the former is guaranteed to reach the global minimum of the standard least squares realization problem (4.4)*, which is a lower-bound for the fixed pole least squares realization problem* (4.9)*. Indeed, fixing the poles* $p_{1,2}$ *reduces the degrees of freedom.*

In practice, however, we are interested more in minimizing the difference to the true output data $\|\boldsymbol{x} - \widehat{\boldsymbol{y}}\|_2^2$ *than the misfit* $\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|_2^2$*. Figure 4.2b indicates that the model-compliant output data* $\widehat{\boldsymbol{y}}$ *obtained using FP-GOR are significantly closer to the noise-free output data* $\boldsymbol{x}$ *than the model-compliant output data obtained using S-GOR. Incorporating* a priori *information about the pole pair* $p_{1,2}$ *enhances the model's ability to distinguish signal from noise, thereby improving estimation accuracy.*



(a) comparison to given data



(b) comparison to true data

Figure 4.2: Box plots of the 50 realization experiments in Example 4.10. For different noise levels $\sigma \in \{0.05, 0.15, \ldots, 0.45\}$, a third-order model with two fixed poles is estimated by the standard (S-GOR) and fixed pole (FP-GOR) realization approach from given data $\boldsymbol{y} \in \mathbb{R}^{16}$ generated by a model as explained in (4.20). It is visible in Figure 4.2a that the S-GOR technique results in a smaller misfit than the FP-GOR technique. The latter obtains better results when we compare the model-compliant data to the true data.

4.5 Conclusions and Future Work

We presented an adaptation of the standard least squares realization problem that exploits fixed pole information in the identification of autonomous, single-output, linear time-invariant dynamical models. By embedding *a priori* information into the globally optimal formulation of [110], we adapted the (rectangular) multiparameter formulation to the fixed pole setting. This formulation allows computing all local and global minimizers of the constrained problem, establishing a globally optimal methodology for the identification of the unknown poles. Moreover, as some of the poles of the model are fixed, the number of unknowns decreases, which in turn leads to a reduction in computational complexity for obtaining the globally optimal solution(s) compared to [110].

We validated the proposed approach through a set of numerical examples. Additionally, we tested whether incorporating *a priori* information about a subset of the model's poles improves the statistical performance of the resulting estimator. The results indicated that the model-compliant output sequences obtained when a priori information is used in the estimation more accurately approximate the true, noise-free signal than those produced by the standard formulation.

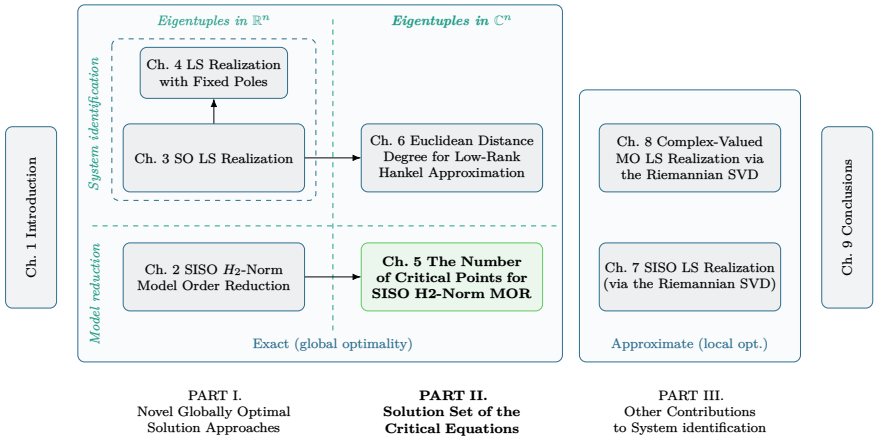
This investigation gives rise to several opportunities for future work. Firstly, our experimental analysis was limited in scope. A more extensive statistical evaluation of the estimator under a broader range of noise conditions and models will be of significant interest. Secondly, the proposed globally optimal framework provides a benchmarking tool that can be used to better understand the heuristic methods. Finally, the computational properties of the multiparameter formulations need further investigation, particularly in terms of scalability and stability for large, realistic problems.

Part II

Solution Set of the Critical Equations

Chapter 5

On the Number of Critical Points for SISO $\mathcal{H}_2$-Norm Model Reduction



Chapter abstract

We perform a comprehensive algebro-geometric analysis of the variety defined by the critical equations of the continuous-time, single-input single-output $\mathcal{H}_2$-norm model reduction problem. We stratify its components, derive its dimension and establish reducedness of the points of interest. Next, we consider the corresponding multiparameter eigenvalue problem (MEP) formulation and show that it corresponds to a particular homogenization of the critical equations, which introduces a spurious component, referred to as the bad locus. We prove that this bad locus is empty whenever the higher-order model has simple poles only. Moreover, we derive an exact count of the complex eigentuples of the MEP by applying the Thom–Porteous formula to the determinantal variety arising from the MEP formulation.

Our main contribution is an explicit formula, applicable to all problem sizes, that determines the number of complex solutions to the critical equations—the critical points. This provides the first uniform upper bound on the number of genuine critical points (e.g., local minima) of the model reduction problem, which correspond to real-valued solutions for which the reduced model is stable. While it was known that the critical equations associated with the model reduction problem admit finitely many complex solutions, no general formula upper-bounding their total number had previously been established. The count is of both theoretical and practical significance: it reveals the intrinsic algebraic complexity of the problem and provides a benchmark for assessing global solvers and the iterative algorithms used in engineering practice.

Candidate's contributions and relation to published work

The material presented in this chapter has not been published elsewhere. It is the result of joint work by Leonie Kayser, associated with the Numerical Algebraic Geometry group at the Max Planck Institute for Mathematics in the Sciences (MPI-MiS), Leipzig, and the candidate, with both authors contributing equally. This collaboration originated during a two-week research visit of the candidate to MPI-MiS in October 2024, at the invitation of Dr. Telen, leader of the Numerical Algebraic Geometry group.

The candidate's contributions to this work include formulating the research questions, contributing to the derivation of the theoretical results (e.g., the analysis of the bad locus), and developing illustrative numerical experiments. While Leonie Kayser provided the algebro-geometric expertise, the author contributed his perspective on the underlying structure of the systems-theoretic problem from which the critical equations arise. In addition to validating the results through numerical computations, the candidate played an important role in interpreting the theoretical findings from a systems-theoretic perspective and situating them within the broader engineering literature. By carefully adapting the presentation of the results given in this chapter, he ensured that the results became accessible to a wider engineering audience and established clear connections to the earlier developments in Part I of this dissertation. An alternative exposition of these results can be found in Leonie Kayser's PhD dissertation [100].

5.1 Introduction

We consider transfer functions of real, continuous-time, strictly causal, stable, single-input single-output (SISO) linear time-invariant (LTI) models of finite order. For a fixed integer $N \geq 1$, let $\mathcal{M}_N$ denote the set of all such transfer functions with McMillan degree exactly N. An element $H(s) \in \mathcal{M}_N$ is written as

$$H(s) = \frac{d(s)}{c(s)} = \frac{d_{N-1}s^{N-1} + \cdots + d_1 s + d_0}{c_N s^N + c_{N-1}s^{N-1} + \cdots + c_1 s + c_0},$$

where $c(s)$ and $d(s)$ are real polynomials with $c_N \neq 0$ and $\gcd(c, d) = 1$ (i.e., $c(s)$ and $d(s)$ are coprime). Stability means that all poles of H, i.e., all roots of $c(s)$, lie in the open left half-plane; equivalently, $c(s)$ is Hurwitz. Strict causality means that the degree of the denominator strictly exceeds that of the numerator, so that the impulse response of H vanishes for negative time and is integrable for $t \geq 0$. Transfer functions in $\mathcal{M}_N$ are therefore real rational functions belonging to the Hardy space $\mathcal{H}_2$, consisting of functions analytic in the open right half-plane and square-integrable along the imaginary axis.

5.1.1 Problem formulation

Given a target reduced order $n < N$, we seek an approximant $\widehat{H}(s)$ whose McMillan degree is at most n, that is $\widehat{H}(s) \in \mathcal{M}_{\leq n}$. We parameterize $\widehat{H}$ as

$$\widehat{H}(s) = \frac{b(s)}{a(s)} = \frac{b_{n-1}s^{n-1} + \cdots + b_1 s + b_0}{a_n s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0},$$

while allowing for the possibility that the actual McMillan degree of $\widehat{H}$ is strictly smaller than n, for instance when pole-zero cancellations occur between $a(s)$ and $b(s)$. The approximant is chosen to minimize the $\mathcal{H}_2$-norm of the error $E(s) = H(s) - \widehat{H}(s)$. Formally, the model reduction problem is

$$\widehat{H}(s) \in \underset{\widehat{H}(s) \in \mathcal{M}_{\leq n}}{\arg\min} \quad J^2, \tag{5.1}$$

where

$$J^2 = \|E(s)\|_{\mathcal{H}_2}^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} |H(j\omega) - \widehat{H}(j\omega)|^2 \, \mathrm{d}\omega = \int_0^{\infty} (h(t) - \widehat{h}(t))^2 \, \mathrm{d}t,$$

and $h(t)$ and $\widehat{h}(t)$ denote the impulse responses of $H(s)$ and $\widehat{H}(s)$, respectively, obtained via the inverse Laplace transform.

The $\mathcal{H}_2$-norm model reduction problem (5.1) is generally non-convex, implying that the solution space can contain (many) local optima. The first-order optimality conditions, characterizing these locally optimal solutions, admit an algebraic formulation: They can be expressed as a system of polynomial equations in the parameters of the reduced model, to which we will refer as the *critical equations*. Accordingly, any (complex) solution of the critical equations is called a *critical point*; however, only real critical points for which the reduced model is stable have a genuine interpretation as a critical point (e.g., local minimum or saddle point) of the original model reduction problem.

5.1.2 Scope and contributions

Although $\mathcal{H}_2$-norm model reduction is by now an established topic in the literature (see, e.g., [10] for a comprehensive treatment), the focus there is typically on practical reduction algorithms. In particular, the field is dominated by approximate iterative methods (e.g., [77]) which, given an initial estimate, converge to a single critical point. By contrast, the present work is of a more fundamental nature: our aim is not to propose yet another reduction algorithm, but to undertake a detailed mathematical study of the critical equations arising from the model reduction problem and their (complex) solution set.

Until now, it was only known that the set of critical points of the model reduction problem (5.1) is finite and consists of isolated points (see, e.g., [13, Theorem 3] for a differential-topological proof and [82, Theorem 3.2] for an algebro-geometric proof via the closure theorem under the additional assumption that H has only simple poles). No general upper bound on the number of such points, valid for arbitrary problem sizes (N, n), was available in the literature.

Our contributions are threefold:

- We provide a comprehensive proof of finiteness of the number of (complex) critical points and show that, generically, the solutions of the critical equations occur without multiplicity, which is computationally favorable[1] (Corollary 5.21).
- Under the assumption that H has simple poles only, we derive an explicit formula for the number of complex critical points, i.e., the total number of (complex) common roots of the critical equations as a function of the problem size (N, n) (Theorem 5.27).

[1] Multiplicities typically lead to ill-conditioning and increased computational difficulty when approximating solutions by numerical methods.

- We develop a presentation that makes powerful algebro-geometric tools accessible to the broader engineering community. In particular, to establish finiteness and obtain an exact solution count, we demonstrate how two fundamental results—the fiber dimension theorem and the Thom–Porteous formula—can be concretely applied to systems of polynomial equations and multiparameter eigenvalue problems[2] (MEPs), respectively. These techniques are not specific to the model reduction problem under consideration, but are equally applicable in other engineering problems where such algebraic formulations arise (see e.g., Chapter 6 for another use case).

The deeper algebro-geometric understanding of the solution set of the critical equations is not only of intrinsic theoretical interest. The obtained upper bound for the number of critical points serves as an indicator of the computational complexity of globally optimal model reduction approaches[3], which certify global optimality by computing all locally optimal solutions [4, 7, 80, 109, 136]. This can be done by directly computing the complex solution set of the critical equations using polynomial system solvers (such as homotopy continuation [20]) and/or specialized MEP solvers (e.g., [90, 176, 178]) and subsequently extract the real, stable solutions. For example, [80] and [136] proposed methodologies based on Gröbner basis computations for order-one and order-three reductions, respectively, while [109] introduces an MEP formulation whose eigentuples allow to retrieve all critical points. In the context of these methods, the expression for the number of critical points as a function of the problem size (N, n) derived here, provides a precise proxy for the computational cost—the difficulty of solving a polynomial system or an MEP is closely tied to the number of its solutions—while simultaneously offering a rigorous benchmark to verify that all critical points have been found.

Algebraic geometry is the study of systems of polynomial equations and the geometric structures (algebraic varieties) that arise as their solution sets, using algebraic means [32, 83]. Our analysis relies heavily on two classical, yet powerful, algebro-geometric results: the fiber dimension theorem (see e.g., [152, Theorem 1.25]) and the Thom–Porteous formula (see e.g., [56, Theorem 12.4]), both of which will be formally introduced later. In our presentation of these results, however, we will avoid as much as possible the full abstraction customary in algebraic geometry, in an effort to make the theory more accessible to the

[2]The multiparameter eigenvalue problem generalizes the standard eigenvalue problem by allowing multiple eigenvalues and by dropping the restriction of linearity in the parameters: a k-parameter eigenvalue problem has k-tuples of eigenvalues, which we refer to as *eigentuples*.

[3]These globally optimal methods are computationally feasible only for (relatively) small problems and are therefore best regarded as tools for theoretical investigation and benchmarking rather than routine engineering practice.

engineering community. We are not the first to draw on algebraic geometry in an engineering context, see e.g., [49] and [82].

The fiber dimension theorem is a foundational result describing the dimension of algebraic varieties. It arose in the context of the broader development of "dimension theory" in commutative algebra and algebraic geometry, which sought to formalize intuitive notions of dimension in the context of geometric objects. The theorem underpins many dimension-counting arguments and is an everyday tool for algebraic geometers [142].

The Thom–Porteous formula, also referred to as the Giambelli-Thom-Porteous formula, originally appeared in [139]. It gives an expression for the fundamental class of a degeneracy locus of a morphism of vector bundles in terms of Chern classes—we will clarify each part of this statement. For our purposes, the vector bundle morphism can be viewed concretely as a parameter-dependent matrix (cf. an MEP formulation): for each parameter value, the matrix defines a linear map between vector spaces of fixed dimension. In this setting, the degeneracy locus of the morphism corresponds to the determinantal variety associated with this matrix, that is, the set of parameter values for which the matrix drops rank (cf. the eigentuples of the MEP). Such determinantal varieties are often highly singular, and their geometric structure can be difficult to study directly from their defining equations. Instead, the Thom–Porteous formula uses only structural information about the parameter-dependent matrix to compute the characteristic class of the corresponding determinantal variety, thereby providing a quantitative measure of "how large" and "how complicated" the determinantal variety is. In the special case where the determinantal variety is known to be zero-dimensional (i.e., it consists of finitely many points), the formula provides a way to count those points.

Remark 5.1. *Note that the results of this chapter build directly on the MEP-based reformulation of the critical equations developed in Chapter 2. The MEP description characterizes the solution set of the critical equations as a determinantal variety, which in turn makes it possible to apply the Thom–Porteous formula leading to the solution count. Consequently, Chapters 2 and 5 collectively establish MEP formulations as a powerful computational and theoretical framework.*

5.1.3 Outline and notation

We conclude the introduction with a brief outline of the chapter. In Section 5.2, we review the interpolatory first-order optimality conditions and summarize the relevant literature. Since our analysis relies heavily on tools from algebraic geometry, Section 5.3 introduces the essential concepts needed in the remainder

of the chapter. In Section 5.4, we consider the (incidence) variety associated with the critical equations. We stratify its components, derive its dimension using the fiber dimension theorem and establish reducedness of the points of interest. As a corollary, this gives a comprehensive proof of finiteness of the number of critical points. Section 5.5 transitions to the multiparameter eigenvalue formulation (MEP) reformulation of the critical equations introduced in [109] and shows that it can be interpreted as a particular homogenization of the critical equations. This homogenization introduces a spurious component, referred to as the bad locus. We prove that this bad locus is empty whenever the higher-order model H has simple poles only, such that in this case, the complex solutions of the MEP are in one-to-one correspondence with those of the critical equations. Then, in Section 5.6, we derive an exact count of the complex eigentuples of the MEP by applying the Thom–Porteous formula to the determinantal variety arising from the MEP formulation. Finally, Section 5.7 discusses the implications of our results and outlines directions for future research.

Throughout, scalars are denoted by lowercase letters (e.g., $s \in \mathbb{C}$), vectors by boldface lowercase letters (e.g., $\boldsymbol{a} \in \mathbb{C}^k$), and matrices by boldface uppercase letters (e.g., $\boldsymbol{M} \in \mathbb{C}^{k \times k}$). Let $P_{\leq k} \coloneqq \{f(s) = \sum_{i=0}^{k} f_i s^i \mid f_i \in \mathbb{C}\} = \mathbb{C}[s]_{\leq k}$ denote the $(k+1)$-dimensional space of complex polynomials of degree at most k. For each polynomial $f \in P_{\leq k}$, we write its coefficient vector in boldface $\boldsymbol{f} = (f_0, \ldots, f_k)^{\mathsf{T}} \in \mathbb{C}^{k+1}$, and conversely, any $\boldsymbol{f} = [f_0, \ldots, f_k] \in \mathbb{C}^{k+1}$ defines a unique polynomial $f(s) = f_0 + f_1 s + \cdots + f_k s^k \in P_{\leq k}$. Varieties and algebraic sets are denoted by uppercase letters, with calligraphic font occasionally used for emphasis (e.g., Y or $\mathcal{Y}$). When required, additional notation will be introduced at the point of use.

5.2 The Critical Equations

The first-order optimality conditions for the $\mathcal{H}_2$-norm model reduction problem (5.1) can be derived in two complementary ways: analytically, by partial differentiation of the $\mathcal{H}_2$-norm error, and geometrically, by exploiting orthogonality in the Hardy space $\mathcal{H}_2$ (see e.g., [77, 128]). In both cases, one obtains a set of Hermite interpolation constraints at the reflected poles of the reduced model: H and $\widehat{H}$ interpolate not only in function value but also in the first derivatives at these points. The foundational result, originally due to Aigrain and Williams [6], was established for models with distinct poles and later extended by Meier and Luenberger [128] to cover higher-order poles.

Theorem 5.2 (Meier and Luenberger [128])**.** *Consider, for a given H, the model reduction problem as formulated in* (5.1)*. A necessary condition for a*

reduced model $\widehat{H} \in \mathcal{M}_{\leq n}$ to be a (local) minimizer is that, for each pole $p_i \in \mathbb{C}$ of $\widehat{H}$, the following interpolation conditions hold:

$$H(-p_i)^{(j)} = \widehat{H}(-p_i)^{(j)}, \quad j = 0, \dots, d_i, \tag{5.2}$$

where d_i is the multiplicity of the pole p_i and the superscript denotes derivation

$$F(a)^{(j)} = \frac{\mathrm{d}^j F(s)}{\mathrm{d}s^j}|_{s=a}.$$

When we restrict our attention to models $\widehat{H}$ with distinct poles only, the optimality conditions in Theorem 5.2 simplify to the following characterization.

Corollary 5.3. *Let $\widehat{H}(s) = b(s)/a(s) \in \mathcal{M}_{\leq n}$ be a candidate approximant with distinct poles only. Then, $\widehat{H}$ satisfies the interpolatory conditions* (5.2) *if and only if there exists a real polynomial $g(s) \in P_{\leq N-n-1}$ such that*

$$d(s)a(s) - c(s)b(s) = a(-s)^2 g(s). \tag{5.3}$$

Proof. Any model $\widehat{H}(s)$ with distinct poles only that satisfies the interpolatory conditions (5.2) must be such that the numerator of the residual $H(s) - \widehat{H}(s) = \frac{d(s)}{c(s)} - \frac{b(s)}{a(s)}$ is divisible by $a(-s)^2$. Expanding this expression, bringing terms to a common denominator, and multiplying out gives the required result. □

Since (5.3) must hold identically as a polynomial identity in $s \in \mathbb{C}$, it is equivalent to the vanishing of each coefficient in the expansion of

$$a(s)d(s) - b(s)c(s) - \widetilde{a}(s)^2 g(s), \tag{5.4}$$

where $\widetilde{a}(s) \coloneqq a(-s) \in P_{\leq n}$. These conditions define a system of $N + n$ polynomial equations in the unknown coefficients of the polynomials a, b, and g, referred to as the *critical equations*, whose common roots are called the *critical points*. Although only real critical points for which the reduced model $\widehat{H}$ is stable correspond to genuine critical points of the model reduction problem (5.1), performing the analysis of the solution set of the critical equations over the complex numbers allows the application of tools from algebraic geometry.

Let us derive a coefficient-based representation of these critical equations. For a polynomial $f \in P_{\leq k}$, let $\boldsymbol{T}_\ell^f \in \mathbb{C}^{\ell \times (\ell+k)}$ denote the banded Toeplitz matrix generated by ℓ forward-shifts of the coefficient vector $\boldsymbol{f} \in \mathbb{C}^{k+1}$:

$$\boldsymbol{T}_\ell^f \coloneqq \begin{bmatrix} f_0 & \cdots & f_{k-1} & f_k & 0 & \cdots & 0 \\ 0 & f_0 & \cdots & f_{k-1} & f_k & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & f_0 & \cdots & f_{k-1} & f_k \end{bmatrix}.$$

Toeplitz matrices constructed in this was allow us to represent polynomial multiplication in terms of coefficient vectors; for instance, if $h \in P_{\leq m}$ for some $m \geq 0$, then the product $h(s)f(s)$ has coefficient vector $(\boldsymbol{T}_{k+1}^{h})^{\mathsf{T}}\boldsymbol{f}$. The critical equations (5.4) can therefore be expressed compactly as:

$$(\boldsymbol{T}_{N}^{a})^{\mathsf{T}}\boldsymbol{d} - (\boldsymbol{T}_{n}^{c})^{\mathsf{T}}\boldsymbol{b} - (\widetilde{\boldsymbol{T}_{N}^{a}})^{\mathsf{T}}(\widetilde{\boldsymbol{T}_{N-n}^{a}})^{\mathsf{T}}\boldsymbol{g} = \boldsymbol{0}.$$

Higher-order poles Note that models $\widehat{H}$ satisfying the interpolatory conditions of Theorem 5.2 but having higher-order poles do not necessarily satisfy (5.3). Indeed, in this case the conditions (5.3) are more restrictive than the interpolatory conditions in Theorem 5.2 (see below for an example). Hence, strictly speaking, the algebraic characterization in (5.3) may exclude those models with higher-order poles. However, the case of higher-order poles has been investigated from a state-space perspective in [168], where the authors argue that, for generic H, all reduced models $\widehat{H}$ satisfying the interpolatory conditions of Theorem 5.2 have simple poles (cf. [168, Theorem 6.3 and subsequent discussion]). We therefore restrict our attention to the simple-pole case, while leaving a treatment of multiple poles as a direction for future work.

Example 5.4. *Take $n = 2$ and consider a model $\widehat{H}(s)$ with a double pole at $p \in \mathbb{R}$, so that $a(s) = (s-p)^2$. The interpolatory conditions of Theorem 5.2 correspond to (5.3) if its right-hand side, $a(-s)^2 g(s) = (s+p)^4 g(s)$, is replaced by $(s+p)^3 g'(s)$ with $g'(s) \in P_{\leq N-n}$.*

Discrete-time models All results presented in this chapter extend without difficulty to the setting of discrete-time models. Only the notation requires an adjustment; all derivations, formulations, and results remain valid. In the discrete-time case, the transfer functions under consideration are strictly proper rational functions of the complex variable z, analytic outside the closed unit disk. Given such an Nth order discrete-time model $H(z) \in \mathcal{M}_{N,\mathrm{d}}$ and a desired reduced order $n < N$, the $\mathcal{H}_2$-norm model reduction problem searches for a solution

$$\widehat{H}(z) \in \operatorname*{argmin}_{\widehat{H}(z) \in \mathcal{M}_{\leq n,\mathrm{d}}} J_{\mathrm{d}}^2,$$

where the subscript d indicates the discrete-time domain. For example,

$$J_{\mathrm{d}}^2 = \left\| H(z) - \widehat{H}(z) \right\|_{\mathcal{H}_2}^2 = \frac{1}{2\pi} \int_0^{2\pi} \left| H\left(e^{j\omega}\right) - \widehat{H}\left(e^{j\omega}\right) \right|^2 \mathrm{d}\omega = \sum_{k=1}^{\infty} \left(h_k - \widehat{h}_k \right)^2.$$

The samples $\{h_k\}_{k=1,\ldots,\infty}$ and $\{\widehat{h}_k\}_{k=1,\ldots,\infty}$ are the impulse responses of H and $\widehat{H}$, respectively.

In this setting, the interpolatory conditions in Theorem 5.2 change: instead of reflecting the poles across the imaginary axis, one now inverts them with respect to the unit circle, so that the interpolation points $-p_i$ are replaced by $1/p_i$. Accordingly, our notation $\widetilde{a}(s) = a(-s)$ used in Corollary 5.3 must be adapted to the discrete-time setting and becomes $\widetilde{a}(z) = z^n a(z^{-1})$.

5.3 Relevant Background in Algebraic Geometry

This section reviews the algebro-geometric concepts that lie at the basis of our analysis. The exposition is intentionally limited in scope and presents only the essential ideas in a accessible form (an overview of the discussed topics is given in Figure 5.1). For readers interested in applications, [155] provides an excellent introduction to polynomial system solving in the sciences and engineering. A complete reference for basic results in algebraic geometry are the introductory texts [83, 152, 33], while a more advanced treatment can be found in [56]. Readers already familiar with these notions may skip this section without loss of continuity. Similarly, those primarily interested in the final results rather than the underlying arguments can also proceed directly to Section 5.4.

5.3.1 Affine varieties

Let $R_n \coloneqq \mathbb{C}[x_1, \ldots, x_n]$ denote the ring of (multivariate) polynomials in the variables $x_1, \ldots, x_n$. An *affine variety* $X \subseteq \mathbb{C}^n$ is the set of all points in $\mathbb{C}^n$ at which a given collection of polynomials $F = \{f_1, \ldots, f_s\} \subseteq R_n$ vanishes simultaneously (the common zero locus of these polynomials)

$$X = V(F) \coloneqq \{\, \boldsymbol{x} \in \mathbb{C}^n \mid f(\boldsymbol{x}) = 0 \text{ for all } f \in F \,\}, \qquad F \subseteq R_n.$$

Closed subsets of $\mathbb{C}^n$ are precisely the sets of the form $V(F)$ for some $F \subseteq \mathbb{C}[x_1, \ldots, x_n]$; their complements are called *open sets*.

A *subvariety* or *closed subset* Y of X is an algebraic subset $Y \subseteq X$, meaning that $X \setminus Y$ is open. The closure $\overline{Y}$ of a subset $Y \subseteq X$ is the smallest closed subset $\overline{Y} \subseteq X$ containing Y. The set Y is *dense* in X if $\overline{Y} = X$. A point $\boldsymbol{x} \in X$ is called *generic* for a given property if it lies in some dense open subset of X, so that "almost all" points of X share that property. The variety X is *irreducible* if $X \neq \emptyset$ and whenever $X = Y_1 \cup Y_2$ with subvarieties $Y_1, Y_2 \subseteq \mathbb{C}^n$, then $X = Y_1$ or $X = Y_2$. Equivalently, every two non-empty open subsets of X intersect.

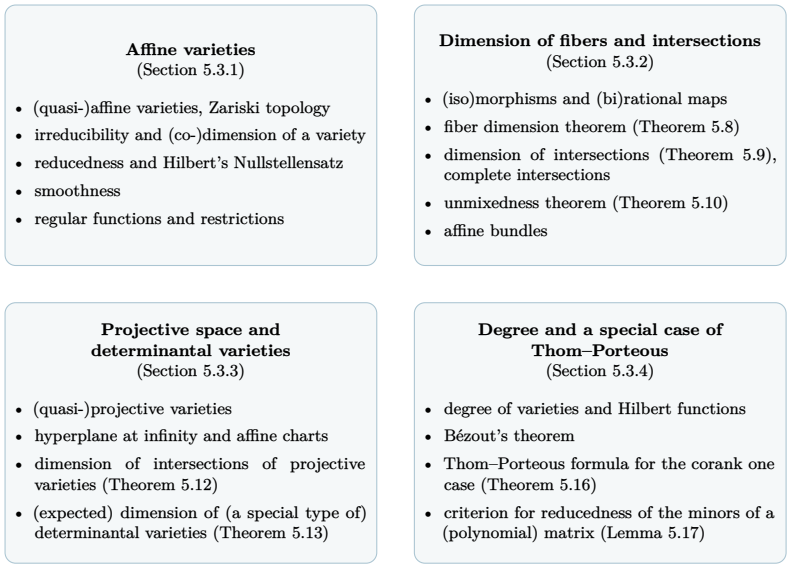


Figure 5.1: Overview of the algebro-geometric notions used throughout the chapter. Section 5.3 introduces these concepts: affine and projective varieties, dimension and intersection theory, determinantal varieties, degrees, and a special case of the Thom–Porteous formula. The exposition is intentionally limited in scope and presents only the essential ideas in an accessible form, with examples included at several points to illustrate the underlying theory. Readers already familiar with these notions may skip the section without loss of continuity.

These closed sets, as defined above, endow $\mathbb{C}^n$ with the so-called *Zariski topology* (see e.g., [152] for a formal treatment). This topology[4] is much coarser than the usual (Euclidean) topology on $\mathbb{C}^n$; for instance, any non-empty Zariski-open subset of an irreducible variety X, e.g., $X = \mathbb{C}^n$, is dense. Indeed, if Y is a Zariski-open subset of X, one may picture Y as X from which only a "thin" algebraic subvariety (e.g., points, a curve or a surface) has been removed. In contrast, in the Euclidean topology there are many non-empty open sets whose closure is still a proper subset (e.g., small open balls in $\mathbb{C}^n$).

An open subset of an affine variety is a *quasi-affine variety*. There are two equivalent ways to view/construct this type of varieties: one can start from an affine variety and remove some lower-dimensional subsets (e.g., exclude certain points or subvarieties), or, equivalently, one can consider the common zero locus of a system of polynomials, but now taken inside a chosen open subset of $\mathbb{C}^n$.

[4] A topology on $\mathbb{C}^N$ is a choice of which subsets are declared "open". The usual Euclidean topology is the one where open sets are unions of open balls $B(\boldsymbol{x}, r) = \{\boldsymbol{y} \in \mathbb{C}^N : |\boldsymbol{y} - \boldsymbol{x}| < r\}$.

Every variety X can be uniquely written as a finite union of its *irreducible components*, in such a way that no component is contained in the union of the others (an *irredundant* decomposition):

$$X = X_1 \cup \cdots \cup X_r, \qquad X_i \text{ irreducible, closed subset } X_i \not\subseteq X_j \text{ for all } i, j.$$

The *dimension* of an irreducible variety X is the length of the longest strictly increasing chain of irreducible subvarieties

$$\dim X = \max \{\, d \in \mathbb{N} \mid \exists Y_0 \subsetneq Y_1 \subsetneq \cdots \subsetneq Y_d = X,\ Y_i \text{ irreducible subvarieties} \,\}.$$

By convention $\dim \emptyset := -1$. Let X be an arbitrary variety, and let $X = X_1 \cup \cdots \cup X_r$ be its decomposition into irreducible components. Then,

$$\dim X := \max_{1 \le i \le r} \dim X_i.$$

It turns out that any chain of irreducible varieties can be extended (by suitably inserting more irreducible subvarieties) to a maximal one, and if X is irreducible, then all such maximal chains have the same length. In particular, any closed proper subvariety of an irreducible variety has strictly smaller dimension than the variety itself.

A variety X is *equidimensional* if every irreducible component $X_i \subseteq X$ has the same dimension. The *codimension* of an irreducible subvariety Y in a variety X is $\operatorname{codim}_X(Y) = \dim X' - \dim Y$, where $X' \subseteq X$ is an irreducible component of largest dimension containing Y. The dimension at a point $\boldsymbol{p} \in X$ is $\operatorname{codim}_X\{\boldsymbol{p}\}$; if X is irreducible then this is just the dimension of X.

The *defining ideal* of a variety X is the set of all polynomials vanishing on X,

$$I(X) = \{\, f \in R_n \mid f(\boldsymbol{x}) = 0 \text{ for all } \boldsymbol{x} \in X \,\}.$$

For a set of polynomials $F \subseteq R_n$, denote by $\langle F \rangle$ the ideal generated by F. Conversely, for an ideal $I \subseteq R_n$, a subset $F = \{f_1, \ldots, f_s\} \subset I$ is called a *generating set* of I if $I = \langle F \rangle$.

For any ideal $J \subseteq R_n$, its *radical* $\sqrt{J}$ is defined by

$$\sqrt{J} := \left\{\, f \in R_n \mid \exists\, k \in \mathbb{N} : f^k \in J \,\right\}.$$

A set of polynomials $F \subseteq R_n$ is said to define a variety X *in a reduced way* if $V(F) = X$ and the ideal $\langle F \rangle$ is radical, i.e., $\sqrt{\langle F \rangle} = \langle F \rangle$. By *Hilbert's Nullstellensatz* we have

$$I(V(F)) = \sqrt{\langle F \rangle},$$

so if $V(F) = X$, then F defines X in a reduced way if and only if $\langle F \rangle = I(X)$, that is, if F is a *generating set* of the defining ideal $I(X)$.

If $F = \{f_1, \dots, f_m\}$ is a generating set of the defining ideal $I(X)$, then the *smooth locus* of X is

$$X^{\mathrm{sm}} = \{\, \boldsymbol{p} \in X \mid \operatorname{rank} \boldsymbol{J_F}(\boldsymbol{p}) \geq n - \operatorname{codim}_X \{\boldsymbol{p}\} \,\},$$

where $\boldsymbol{J}_F(\boldsymbol{p})$ is the Jacobian matrix of F at $\boldsymbol{p}$:

$$\boldsymbol{J}_F(\boldsymbol{p}) = \left[\frac{\partial f_i}{\partial x_j}(\boldsymbol{p}) \right]_{i,j}.$$

The smooth locus is an open dense set $X^{\mathrm{sm}} \subseteq X$, and is isomorphic to a complex manifold of the same dimension as X. A point $\boldsymbol{p} \in X$ is called *singular* if it does not lie in X^{sm}. We denote $X^{\mathrm{sm}}(\mathbb{R}) \coloneqq X^{\mathrm{sm}} \cap \mathbb{R}^n$, that is, the set of smooth points of X with real coordinates. If $\boldsymbol{p} \in X^{\mathrm{sm}}(\mathbb{R})$, then $X^{\mathrm{sm}}(\mathbb{R}) \subseteq \mathbb{R}^n$ is a smooth (real) submanifold in a neighborhood of $\boldsymbol{p}$.

Example 5.5. *A complex matrix $\boldsymbol{A} \in \mathbb{C}^{k \times k}$ is* nilpotent *if $\boldsymbol{A}^n = \boldsymbol{0}$ for some $n > 0$. This is equivalent to its characteristic polynomial being equal to λ^k. For $k = 2$, this gives the algebraic constraint:*

$$\det(\lambda \boldsymbol{I} - \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix}) = \lambda^2 - (x_1 + x_4)\lambda + (x_1 x_4 - x_2 x_3) = \lambda^2.$$

As such, the two polynomials $F = \{x_1 + x_4, x_1 x_4 - x_2 x_3\}$ define the affine variety $\mathcal{N} = V(F) \subset \mathbb{C}^4$ characterizing all nilpotent matrices in $\mathbb{C}^{2\times 2}$. Eliminating x_4 gives an isomorphism $\mathcal{N} \cong Z = \{(x_1, x_2, x_3) \in \mathbb{C}^3 \mid x_2 x_3 = -x_1^2\}$. Because $x_2 x_3 + x_1^2$ is irreducible, this illustrates that $\mathcal{N}$ is irreducible of dimension 2 (one polynomial constraint in a three-dimensional ambient space). Since $\mathcal{N}$ has only one irreducible component, namely itself, it is trivially equidimensional. The closed subset $Y = \{x_2 = 0\} = \left[\begin{smallmatrix} 0 & 0 \\ \mathbb{C} & 0 \end{smallmatrix}\right] \subset \mathcal{N}$ is irreducible of dimension 1, so $\operatorname{codim}_{\mathcal{N}} Y = 1$. Its complement $U = \{x_2 \neq 0\} \subset \mathcal{N}$ is an open subset, for which its closure $\overline{U} = \mathcal{N}$.

The Jacobian matrix of $F = \{x_1 + x_4, x_1 x_4 - x_2 x_3\}$ corresponds to

$$\boldsymbol{J_F} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ x_4 & -x_3 & -x_2 & x_1 \end{bmatrix}.$$

It has rank 2 at every point $\boldsymbol{p} \in \mathcal{N}$ except the origin, where it has rank 1. Since $\dim \mathcal{N} = 2$, the smooth locus is $\mathcal{N}_{\mathrm{sm}} = \mathcal{N} \setminus \{0\}$, making the origin the unique singular point. The defining ideal of the variety is $I(\mathcal{N}) = \langle F \rangle_R$. Since I is radical, the set of generators F defines $\mathcal{N}$ in a reduced way. In contrast, the set $G = \{x_1 + x_4, (x_1 x_4 - x_2 x_3)^2\}$ defines the same variety but does not define it in a reduced way.

For polynomial equations we understand what it means to be defined in a reduced way. We now turn from this global viewpoint to a local one. Globally, we consider functions defined on the entire variety, and reducedness is expressed by the defining ideal being radical. Locally, however, we focus on the behaviour near a single point. Functions in this local setting are defined only on a small neighbourhood and are described algebraically by localizations of rings, which allow denominators that do not vanish at the point. To formulate the local analogue of reducedness, we therefore begin by clarifying what we mean by local polynomial functions.

Let $U \subset \mathbb{C}^n$ be an open subset and consider $F = \{f_1, \dots, f_s\}$. We denote by $\mathcal{O}(U)$ the ring of *regular functions* on U. These are functions $f : U \to \mathbb{C}$ which can locally be written as a quotient g/h of complex polynomials with h nonvanishing on U. In particular, when $U = \mathbb{C}^n$ itself, every regular function is a polynomial, because a denominator that never vanishes on all of $\mathbb{C}^n$ must be a nonzero constant.

For each $f \in F$ we write $f|_U \in \mathcal{O}(U)$ for its *restriction* to U, obtained by evaluating f only on points of U. We say that F defines X *in a reduced way on* U if the ideal in $\mathcal{O}(U)$ generated by the restrictions $f|_U$ coincides with the defining ideal of $X \cap U$ in $\mathcal{O}(U)$, i.e., every function in $\mathcal{O}(U)$ that vanishes on $X \cap U$ can be written (locally) as an $\mathcal{O}(U)$-linear combination of the $f|_U$.

Example 5.6. *Let $X \subset \mathbb{C}^2$ be the x-axis, i.e.,*

$$X = V(y) = \{(x, y) \in \mathbb{C}^2 \mid y = 0\},$$

so that its defining ideal is $I(X) = (y) \subset \mathbb{C}[x, y]$. Consider the set of equations $F = \{xy, y^2\}$. Then the ideal generated by F in $\mathbb{C}[x, y]$, is non-radical. Hence F does not define X in a reduced way.

Now let $U = \{(x, y) \in \mathbb{C}^2 \mid x \neq 0\}$ be the open subset where x does not vanish. On U we can divide by x, so in the ring $\mathcal{O}(U)$ of functions $U \to \mathbb{C}$ which locally can be written as a quotient g/h of polynomials with h nonvanishing, the function x is invertible. Denoting by $xy|_U$ and $y|_U$ the restrictions of xy and y to U, we have $y|_U = \frac{1}{x} \cdot (xy)|_U$ in $\mathcal{O}(U)$, and thus

$$\langle xy, y^2 \rangle_{\mathcal{O}(U)} = \langle y, y^2 \rangle_{\mathcal{O}(U)} = \langle y \rangle_{\mathcal{O}(U)}.$$

Therefore, on U the ideal generated by the restriction of F coincides with the defining ideal of $X \cap U$, i.e., F defines X in a reduced way on U, although it does not do so globally.

5.3.2 Dimension of fibers and intersections

A *morphism* of affine varieties $X \subseteq \mathbb{C}^n$ and $Y \subseteq \mathbb{C}^m$ is a coordinate-wise polynomial map

$$\phi = (\phi_1, \ldots, \phi_m) \colon \mathbb{C}^n \to \mathbb{C}^m, \qquad \phi(X) \subseteq Y.$$

Having polynomial coordinate functions is what sets a morphism apart from general maps; it ensures that ϕ_i is defined on all of $\mathbb{C}^n$ and not merely on X, giving morphisms a global character. The morphism is *dominant* if $\overline{\phi(X)} = Y$. The *fiber* $\phi^{-1}(\boldsymbol{q})$ of $\boldsymbol{q} \in Y$ is the subset $\{\, \boldsymbol{x} \in X \mid \phi(\boldsymbol{x}) = \boldsymbol{q} \,\}$. Equations of $\phi^{-1}(\boldsymbol{q})$ are given by

$$I(X) \cup \{\phi_1 - q_1, \ldots, \phi_m - q_m\} \subseteq \mathbb{C}[x_1, \ldots, x_n],$$

and ϕ has *reduced fibers at* $\boldsymbol{q}$ if this set of equations defines $\phi^{-1}(\boldsymbol{q})$ in a reduced way. An *isomorphism* is a morphism $\phi \colon X \to Y$ that is bijective and whose inverse $\phi^{-1} \colon Y \to X$ is also a morphism; in this case X and Y are said to be algebraically identical.

A *rational map* $f \colon X \dashrightarrow Y$ is a partially defined analogue of a morphism: it is given by regular functions on some dense open subset $U \subseteq X$, whereas a morphism is defined by regular functions everywhere on X. A *birational map* $f \colon X \dashrightarrow Y$ is a rational map for which there exists a rational inverse: there are dense open subsets $U \subseteq X$ and $V \subseteq Y$ and morphisms $U \to V$ and $V \to U$ that are inverse to each other, making $f|_U \colon U \to V$ an isomorphism. If X and Y admit such a map, they are called *birational varieties.* A *birational morphism* is a morphism that is birational as a rational map, that is, a morphism which restricts to an isomorphism between dense open subsets of X and Y. The following example illustrates these concepts.

Example 5.7. *Consider the affine variety $Y = V(xy - 1) \subseteq \mathbb{C}^2$, which is the hyperbola $\{(x, y) \mid xy = 1\}$, and the line $X = \mathbb{C}$. Define*

$$\phi : X \setminus \{0\} \to Y : x \mapsto (x, 1/x).$$

Because the coordinate functions of ϕ are x and $1/x$, this map is regular precisely on those points where the denominator does not vanish. Thus, ϕ is a morphism on the dense open subset $X \setminus \{0\}$, but it fails to be defined at $x = 0$. For that reason, ϕ is not a morphism $X \to Y$; rather, it determines a rational map $X \dashrightarrow Y$, whose domain of definition is $X \setminus \{0\}$.

Now consider the projection

$$\pi \colon Y \to X \colon (x, y) \mapsto x.$$

Its coordinate function is simply x, a polynomial, so π is a morphism defined everywhere on Y. Moreover, because every point $(x, y) \in Y$ satisfies $y = 1/x$, the image satisfies $\pi(Y) = X \setminus \{0\}$. On these dense open subsets, the maps ϕ and π are inverse isomorphisms:

$$\pi(\phi(x)) = x \quad (x \neq 0),$$

$$\phi(\pi(x, y)) = (x, y) \quad ((x, y) \in Y).$$

Thus X and Y are birational varieties and $\pi : Y \to X$ is a birational morphism: it is a morphism everywhere but becomes an isomorphism after restricting to the dense open subsets Y and $X \setminus \{0\}$.

Theorem 5.8 (Fiber dimension theorem [152, Theorem 1.25])**.** *Let $\phi\colon X \to Y$ be a morphism of varieties and X irreducible.*

1. *For every $\boldsymbol{q} \in \phi(X)$, one has $\dim \phi^{-1}(\boldsymbol{q}) \geq \dim X - \dim \overline{\phi(X)}$.*
2. *There is a dense open subset $U \subseteq \overline{\phi(X)}$ such that ϕ has reduced fibers at $\boldsymbol{q} \in U$ and equality holds in the first statement.*

In the context of the previous result, since U is a dense open subset of $\overline{\phi(X)}$, it follows that for a generic point (i.e., "almost all points") in the image of the morphism the fiber is reduced and its dimension is given exactly by the fiber dimension theorem. This result is a key tool for algebraic geometers in analyzing the dimension and reducedness of varieties, and we will use it extensively throughout the remainder of this chapter.

Theorem 5.9 (Dimension of intersection [152, Theorem 1.24])**.**

1. *If $X, Y \subseteq \mathbb{C}^n$ are varieties, then every irreducible component Z of the intersection $X \cap Y$ satisfies*

$$\dim Z \geq \dim X + \dim Y - n.$$

2. *If $f_1, \ldots, f_s \in R_n$, then $V(f_1, \ldots, f_s)$ is empty or each of its irreducible components has dimension $\geq n - s$.*

With minor additional technicalities, the theorem remains valid after replacing $\mathbb{C}^n$ by any non-empty open subset $U \subseteq \mathbb{C}^n$.

Consider a non-empty open subset $U \subseteq \mathbb{C}^n$ and a variety $X \subseteq U$ with $\operatorname{codim}_U X = s \leq n$. We say that X is a *set-theoretic complete intersection in U* if there exist a set $\{f_1, \ldots, f_s\} \subseteq R_n$ such that $X = V(f_1, \ldots, f_s) \cap U$.

In this case, X (up to taking radicals) can be defined by s equations, and no strictly smaller number of equations suffices to cut out X set-theoretically. If, in addition, $f_1, \dots, f_s$ generate the defining ideal $I(X)$ of X in $\mathcal{O}(U)$ (i.e., they define X by a radical ideal in U), then X is an *ideal-theoretic complete intersection in U*.

In practical applications, being able to demonstrate that the defining equations of a variety form an ideal-theoretic complete intersection has several benefits. First, it guarantees that the equations generate the full defining ideal and use the smallest possible number of generators (equal to the codimension), so no equations are missing and none is algebraically superfluous. Second, it ensures a generically well-conditioned Jacobian for numerical continuation. When the variety is zero-dimensional, complete intersections correspond to square systems, where the number of equations equals the number of variables.

Theorem 5.10 (Unmixedness theorem [55, Corollary 18.14])**.** *Let $X = V(f_1, \dots, f_s)$ be a set-theoretic complete intersection. If X is defined in a reduced way somewhere on every irreducible component of X, then it is reduced everywhere, so actually an ideal-theoretic complete intersection.*

Example 5.11 (Example 5.5 continued)**.** *Let $\phi : \mathcal{N} \to \mathbb{C}$, $\phi(\boldsymbol{x}) = x_2$ project a nilpotent matrix, represented by $\boldsymbol{x} = (x_1, x_2, x_3, x_4) \in \mathcal{N}$, onto its $(1,2)$-entry x_2. The image of the morphism is $\mathbb{C}$ since $(0, q, 0, 0) \in \mathcal{N}$ for all $q \in \mathbb{C}$. The fiber $\phi^{-1}(q) = \{\boldsymbol{x} \in \mathcal{N} \mid x_2 = q\}$ has defining equations $F \cup \{x_2 - q\}$. Because of the two linear equations, the fiber is isomorphic to $Z_q = \{(x_1, x_3) \in \mathbb{C}^2 \,|\, q x_3 = -x_1^2\}$. If $q \neq 0$, this is the affine plane curve $x_3 = -x_1^2/q$, which is irreducible, reduced, and of dimension 1. If $q = 0$, the equation becomes $x_1^2 = 0$ so Z_q is defined by the non-reduced ideal $\langle x_1^2 \rangle$, and the underlying set is the line $\{(0, x_3) \in \mathbb{C}^2\}$. In both cases $\dim \phi^{-1}(q) = 1$ such that the fiber dimension theorem holds: $\dim \phi^{-1}(q) = 1 \geq \dim \mathcal{N} - \dim \overline{\phi(\mathcal{N})} = 2 - 1 = 1$. Finally, since the 2-dimensional variety $\mathcal{N} \subset \mathbb{C}^4$ is cut out by the two polynomials $x_1 + x_4$ and $x_1 x_4 - x_2 x_3$, and they generate a radical ideal, $\mathcal{N}$ is a set- and ideal-theoretic complete intersection.*

We now introduce affine bundles as a special and useful class of subvarieties. Take a (quasi) affine variety $B \subseteq \mathbb{C}^m$. Fix an $(\ell \times n)$-matrix $\boldsymbol{M}(\boldsymbol{y})$ whose entries are polynomials in the coordinates $\boldsymbol{y} \in B$. Assume that $\operatorname{rank} \boldsymbol{M}(\boldsymbol{y}) = k$ at every point $\boldsymbol{y} \in B$. Choose a polynomial column vector $\boldsymbol{b}(\boldsymbol{y}) \in \mathbb{C}[\boldsymbol{y}]^\ell$ so that for all $\boldsymbol{y} \in B$, the vector $\boldsymbol{b}(\boldsymbol{y})$ lies in the image of $\boldsymbol{M}(\boldsymbol{y})$, ensuring that the linear system $\boldsymbol{M}(\boldsymbol{y})\boldsymbol{x} = \boldsymbol{b}(\boldsymbol{y})$ admits solutions for every $\boldsymbol{y} \in B$. Under these assumptions, let

$$X = \{\, (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{C}^n \times B \;\mid\; \boldsymbol{M}(\boldsymbol{y})\boldsymbol{x} = \boldsymbol{b}(\boldsymbol{y}) \,\}$$

and consider the projection $\pi : X \to B, (\boldsymbol{x}, \boldsymbol{y}) \to \boldsymbol{y}$. For a chosen point $\boldsymbol{y} \in B$, the fiber

$$\pi^{-1}(\boldsymbol{y}) = \{ (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{C}^n \times \mathbb{C}^m \mid \boldsymbol{M}(\boldsymbol{y})\boldsymbol{x} = \boldsymbol{b}(\boldsymbol{y}) \},$$

is an affine linear subspace of $\mathbb{C}^n$ of dimension $r = n - k$, hence isomorphic to $\mathbb{C}^r$. We say that $\pi : X \to B$ is an *affine bundle of rank* r. In that context, X is called the *total space* of the bundle, B is the *base space*, and π is the *bundle projection*.

Establishing that a given morphism $\pi : X \to B$ is a rank r affine bundle is advantageous. Because all fibers have dimension r, we have $\dim X = \dim B + r$. Moreover, the local product structure of the bundle—namely, that X is locally isomorphic to $B \times \mathbb{C}^r$—ensures that properties such as irreducibility, equidimensionality, reducedness, and smoothness pass from the base B to the total space X. For instance, if B is irreducible, then so is X.

5.3.3 Projective space and determinantal varieties

Projective space $\mathbb{P}^n$ is defined as $(\mathbb{C}^{n+1} \setminus \{0\})$ modulo the equivalence relation that identifies two nonzero vectors $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{C}^{n+1}$ whenever they are scalar multiples of each other. A projective point $(x_0 : x_1 : \cdots : x_n) \in \mathbb{P}^n$ represents the equivalence class $[\boldsymbol{x}]$ of all nonzero vectors $(\lambda x_0, \lambda x_1, \ldots, \lambda x_n) \in \mathbb{C}^{n+1}$ with $\lambda \in \mathbb{C} \setminus \{0\}$. The notation $\mathbb{P}(V)$ denotes the *projectivization* of a vector space V, i.e., the set of one-dimensional linear subspaces of V. In particular, $\mathbb{P}(\mathbb{C}^{n+1}) = \mathbb{P}^n$.

A *projective variety* X is the zero-set $V(F) \subseteq \mathbb{P}^n$ of a set of homogeneous polynomials $F = \{f_1, \ldots, f_s\} \subseteq R_{n+1}$. Its defining ideal $I(X)$ is defined as the defining ideal of the *affine cone* $\widehat{X} = \left\{ \boldsymbol{x} \in \mathbb{C}^{n+1} \,\middle|\, [\boldsymbol{x}] \in X \right\} \cup \{0\}$, and it is generated by all homogeneous polynomials in R_{n+1} vanishing on X. If $X \subseteq \mathbb{P}^n$ is a projective variety and $\widehat{X} \subseteq \mathbb{C}^{n+1}$ denotes its affine cone, then by construction $\dim \widehat{X} = \dim X + 1$ and hence the codimension is preserved:

$$\operatorname{codim}_{\mathbb{P}^n}(X) = n - \dim X = (n+1) - (\dim X + 1) = \operatorname{codim}_{\mathbb{C}^{n+1}}(\widehat{X}).$$

Projective space $\mathbb{P}^n$ can be decomposed as the disjoint union of the hyperplane at infinity $H_\infty \coloneqq V(x_0)$ and its complement

$$U_0 = \{x_0 \neq 0\} = \{ (1 : x_1 : \cdots : x_n) \mid \boldsymbol{x} \in \mathbb{C}^n \},$$

which is called the standard projective chart on $\mathbb{P}^n$. An *affine chart* of a projective variety X is an open subset of X that is identified with an affine variety. In this case, normalizing $x_0 = 1$ gives the isomorphism $U_0 \cong \mathbb{C}^n$.

In this way, we can associate to a projective variety $X \subseteq \mathbb{P}^n$ not contained in H_∞ the affine variety

$$X' = X \cap U_0 = \{\, \boldsymbol{x} \in \mathbb{C}^n \mid (1 : x_1 : \cdots : x_n) \in X \,\}, \qquad I(X') = I(X)|_{x_0=1}.$$

Conversely, if $Y \subseteq \mathbb{C}^n$, then one can consider the closure $\overline{Y} \subseteq \mathbb{P}^n$. Its defining ideal is given by

$$I(\overline{Y}) = \langle f^{\mathrm{h}} = \sum_{i=0}^{d} x_0^{d-i} f^{(i)} \mid f = f^{(0)} + \cdots + f^{(d)} \in I(Y),\ \deg f = d \rangle,$$

where in this case, $f^{(i)}$ denotes the component of f of total degree i. Note, however, that if $Y = V(f_1, \ldots, f_s)$, then it is usually *not* the case that $\overline{Y} = V(f_1^{\mathrm{h}}, \ldots, f_s^{\mathrm{h}})$, even when the f_i define Y in a reduced way.

A *quasi-projective variety* is an open subset of a projective variety. Quasi-projective varieties include all previously introduced types of varieties. Indeed, if $W \subseteq \mathbb{C}^n$ is an affine variety, and $\overline{W} \subseteq \mathbb{P}^n$ denotes its projective closure, then $W = \overline{W} \cap \mathbb{C}^n$, so an affine variety W is an open subset of the projective variety $\overline{W}$ and hence is quasi-projective. It is the complement of the hyperplane at infinity.

An important advantage of working in projective space is that many equations admit solutions that do not exist in affine space. For example, parallel lines in the affine plane intersect at H_∞. Formally, this principle is captured by the following theorem.

Theorem 5.12 ([152, Theorem 1.24])**.** *If $X, Y \subseteq \mathbb{P}^n$ are projective varieties with $\dim X + \dim Y \geq n$, then $X \cap Y \neq \emptyset$ (and $\dim X \cap Y \geq \dim X + \dim Y - n$).*

A particularly nice class of projective varieties is given by *determinantal varieties*, which are varieties defined by the vanishing of the minors of a matrix. In the remainder of this chapter we focus on determinantal varieties associated with matrices whose entries in each column all have the same degree (although this degree may differ from column to column). Fix integers $f \geq e$ and $d_1, \ldots, d_e \geq 0$, and consider such a matrix

$$\boldsymbol{M}(\boldsymbol{x}) = (h_{i,j})_{1 \leq i \leq f,\ 1 \leq j \leq e}, \quad h_{i,j} \in R_{n+1} \text{ is homogeneous of degree } d_j. \tag{5.5}$$

The determinantal variety of rank k associated to $\boldsymbol{M}$ is the projective variety

$$\Sigma_k(\boldsymbol{M}) = \{\, \boldsymbol{x} \in \mathbb{P}^n \mid \operatorname{rank} \boldsymbol{M}(\boldsymbol{x}) \leq k \,\}.$$

It is a closed set for which the equations are given by the $k+1$-minors of $\boldsymbol{M}$ (which are homogeneous by assumption on the degrees of the entries of $\boldsymbol{M}$).

Theorem 5.13 (Dimension of determinantal varieties)**.** *Let $\boldsymbol{M}$ be the matrix polynomial from* (5.5)*. For every irreducible component $Z \subseteq \Sigma_k(\boldsymbol{M})$ one has*

$$\dim Z \geq n - (e-k)(f-k).$$

Note that for $e = 1$ and $k = 0$, one obtains a projective analogue of Theorem 5.9-(ii). In order to prove the theorem, we need the following lemma.

Lemma 5.14 ([56, Theorem 0.2])**.** *Let $f : \mathbb{C}^m \to \mathbb{C}^n$ be a morphism of irreducible varieties. Let $W \subseteq \mathbb{C}^n$ be a subvariety. Then*

$$\operatorname{codim}_{\mathbb{C}^m} f^{-1}(W) \leq \operatorname{codim}_{\mathbb{C}^n} W.$$

Proof. Consider the graph

$$\Gamma_f \coloneqq \{\, (\boldsymbol{x}, f(\boldsymbol{x})) \mid \boldsymbol{x} \in \,\} \subseteq \mathbb{C}^m \times \mathbb{C}^n.$$

The morphism $\mathbb{C}^m \to \Gamma_f$, $\boldsymbol{x} \mapsto (\boldsymbol{x}, f(\boldsymbol{x}))$ is an isomorphism, hence Γ_f is irreducible and $\dim \Gamma_f = \dim \mathbb{C}^m = m$. Let $M \coloneqq \mathbb{C}^m \times W \subseteq \mathbb{C}^m \times \mathbb{C}^n$. We have $\dim M = \dim \mathbb{C}^m + \dim W$. Moreover,

$$\Gamma_f \cap M = \{\, (\boldsymbol{x}, f(\boldsymbol{x})) \mid \boldsymbol{x} \in \mathbb{C}^m,\ f(\boldsymbol{x}) \in W \,\}$$

is isomorphic to $f^{-1}(W)$, via the projection $(\boldsymbol{x}, f(\boldsymbol{x})) \mapsto \boldsymbol{x}$.

The product $\mathbb{C}^m \times \mathbb{C}^n$ is of dimension $m+n$, so that we can apply Theorem 5.9 about the dimension of the intersection to Γ_f and M. Let Z be an irreducible component of $\Gamma_f \cap M$, then

$$\begin{aligned} \dim Z &\geq \dim \Gamma_f + \dim M - \dim(\mathbb{C}^m \times \mathbb{C}^n) \\ &= m + (m + \dim W) - (m+n) = m + \dim W - n. \end{aligned}$$

Since Z is isomorphic to an irreducible component of $f^{-1}(W)$, this shows that every irreducible component Z_0 of $f^{-1}(W)$ satisfies $\dim Z_0 \geq m + \dim W - n$. Rewriting this in terms of codimension:

$$\operatorname{codim}_{\mathbb{C}^m} Z_0 = m - \dim Z_0 \leq m - (m + \dim W - n) = n - \dim W = \operatorname{codim}_{\mathbb{C}^n} W.$$

By definition,

$$\operatorname{codim}_{\mathbb{C}^m} f^{-1}(W) = \min\left\{ \operatorname{codim}_{\mathbb{C}^m} Z_0 \mid Z_0 \text{ irred. component of } f^{-1}(W) \right\},$$

so we conclude $\operatorname{codim}_{\mathbb{C}^m} f^{-1}(W) \leq \operatorname{codim}_{\mathbb{C}^n} W$. □

Proof of Theorem 5.13. Let $\phi : \mathbb{C}^{n+1} \to \mathbb{C}^{ef} : \boldsymbol{x} \mapsto \boldsymbol{M}(\boldsymbol{x})$ be the morphism defined by evaluating all entries of the polynomial matrix $\boldsymbol{M}(\boldsymbol{x})$ defined in (5.5). Let $D_k \subseteq \mathbb{C}^{ef}$ be the standard determinantal variety of (unstructured) $(e \times f)$-matrices of rank at most k,

$$D_k \coloneqq \left\{ \boldsymbol{Q} \in \mathbb{C}^{ef} \mid \operatorname{rank} \boldsymbol{Q} \leq k \right\}.$$

The variety D_k is irreducible and

$$\operatorname{codim}_{\mathbb{C}^{ef}} D_k = (e-k)(f-k),$$

see, e.g., [83, Proposition 12.2]. By construction,

$$\phi^{-1}(D_k) = \left\{ \boldsymbol{x} \in \mathbb{C}^{n+1} \mid \operatorname{rank} \boldsymbol{M}(\boldsymbol{x}) \leq k \right\}.$$

Now use the homogeneity of the entries $h_{i,j}$ of $\boldsymbol{M}$: $h_{i,j}$ is homogeneous of degree d_j, so for any $\lambda \in \mathbb{C} \setminus \{0\}$, we have $h_{i,j}(\lambda \boldsymbol{x}) = \lambda^{d_j} h_{i,j}(\boldsymbol{x})$. This scales the j-th column of $\boldsymbol{M}(\boldsymbol{x})$ by a nonzero scalar λ^{d_j}, hence $\operatorname{rank} \boldsymbol{M}(\lambda \boldsymbol{x}) = \operatorname{rank} \boldsymbol{M}(\boldsymbol{x})$. Therefore, $\phi^{-1}(D_k)$ is a cone: it is invariant under scalar multiplication by λ. This shows that

$$\mathbb{P}(\phi^{-1}(D_k)) = \Sigma_k(\boldsymbol{M}).$$

Thus, $\phi^{-1}(D_k)$ is the affine cone over the projective determinantal variety $\Sigma_k(\boldsymbol{M})$.

Let Z be an irreducible component of $\Sigma_k(\boldsymbol{M})$ and $\widehat{Z} \subseteq \mathbb{C}^{n+1}$ its affine cone, such that $\widehat{Z}$ is an irreducible component of $\phi^{-1}(D_k)$, and $\dim \widehat{Z} = \dim Z + 1$ (taking the cone adds one dimension). Thus,

$$\operatorname{codim}_{\mathbb{C}^{n+1}} \widehat{Z} = (n+1) - \dim \widehat{Z} = (n+1) - (\dim Z + 1) = n - \dim Z = \operatorname{codim}_{\mathbb{P}^n} Z.$$

Applying Lemma 5.14 to the morphism ϕ and the subvariety $D_k \subset \mathbb{C}^{ef}$ shows that for every irreducible component T of $\phi^{-1}(D_k)$,

$$\operatorname{codim}_{\mathbb{C}^{n+1}} T \leq \operatorname{codim}_{\mathbb{C}^{ef}} D_k = (e-k)(e-f).$$

In particular, this holds for $T = \widehat{Z}$:

$$\begin{aligned} &\operatorname{codim}_{\mathbb{P}^n} Z = \operatorname{codim}_{\mathbb{C}^{n+1}} \widehat{Z} \leq (e-k)(f-k) \\ \iff\ &\dim Z \geq n - (e-k)(f-k). \end{aligned}$$

□

A helpful way to interpret Theorem 5.13 is to compare its conclusion with the *expected* geometry of determinantal varieties. In the space of all $(f \times e)$-matrices, the locus of matrices of rank at most k is defined by the $(k+1)$-minors and

has codimension $(f-k)(e-k)$, hence dimension $k(f+e-k)$. One can see this dimension directly by parametrizing a rank-k matrix via a k-dimensional row-space, a k-dimensional column space, and an invertible $(k \times k)$-core, which together give $k(f+e-k)$ degrees of freedom. When the entries of the matrix vary independently, this "expected" dimension is achieved exactly. In contrast, when the entries arise from prescribed homogeneous polynomials (e.g., the matrix $\boldsymbol{M}(\boldsymbol{x})$ from (5.5)) the rank conditions (potentially) impose fewer independent constraints, and the codimension can only become smaller. Theorem 5.13 formalizes this: every component of $\Sigma_k(\boldsymbol{M})$ has codimension no greater than in the unrestricted case.

Example 5.15. *Consider the* 3×2 *Hankel matrix* $\boldsymbol{H} = [x_{i+j}]_{0\le i\le 2,\,0\le j\le 1}$, *constructed from the homogeneous coordinates* $[x_0 : x_1 : x_2 : x_3] \in \mathbb{P}^3$. *The determinantal variety defined by the maximal minors of* $\boldsymbol{H}$ *is then* $X = \Sigma_1(\boldsymbol{H}) = V(x_1^2 - x_0x_2,\ x_0x_3 - x_1x_2,\ x_2^2 - x_1x_3) \subset \mathbb{P}^3$.

On the affine chart $x_0 \neq 0$, *set* $x_0 = 1$; *the first and second equations become* $x_2 = x_1^2$ *and* $x_3 = x_1^3$, *while the third is automatically satisfied. Hence the affine part* $X' = X \cap \mathbb{C}^3$ *is parameterized by* $t \mapsto (t, t^2, t^3)$. *Setting* $x_0 = 0$ *forces* $x_1 = x_2 = 0$ *while leaving* $x_3 \neq 0$; *thus* $X \cap H_\infty = \{[0:0:0:1]\}$ *is a single point at infinity.*

It is clear that $\dim X = 1$. *For any projective plane* $Y \subset \mathbb{P}^3$ *one has* $\dim X + \dim Y = 3$, *so by Theorem 5.12 the intersection* $X \cap Y$ *is non-empty. Finally, the determinantal-variety dimension estimate* $\dim X \ge n - (e-k)(f-k)$ *with* $n = 3,\ e = 2,\ f = 3,\ k = 1$ *yields* $\dim X \ge 1$, *and equality holds in this example.*

5.3.4 Degree and a special case of Thom–Porteous

If $X = V(F) \subseteq \mathbb{C}^n$ is a finite set, then the quotient space $R_n/\langle F\rangle$ has finite dimension. The degree of the ideal, $\deg\langle F\rangle$, is the dimension of this quotient space and we define the *degree* of X as

$$\deg X \coloneqq \deg\langle F\rangle.$$

In this case, $\deg X$ corresponds to the number of points in X, counted with multiplicity, and therefore

$$\deg X = \#X \quad \Longleftrightarrow \quad F \text{ defines } X \text{ in a reduced way.}$$

When X is infinite (its dimension is positive) the definition above is no longer meaningful. In this case, we pick $\dim X$ generic affine-linear polynomials

$\ell_1, \ldots, \ell_{\dim X}$ and define

$$\deg X := \deg\langle I(X), \ell_1, \ldots, \ell_{\dim X}\rangle.$$

The degree is independent of the particular choice of the ℓ_i, and this construction agrees with the familiar notion of the degree of a hypersurface: if $X = V(f)$, then $\deg X = \deg f$.

If no irreducible component of $X \subseteq \mathbb{P}^n$ is contained in the hyperplane at infinity H_∞, then its degree is the degree of $X \cap \mathbb{C}^n$. The degree (and dimension) of a projective variety can alternatively be characterized via the leading coefficient of the *Hilbert function*:

$$h_X(t) := \dim_{\mathbb{C}} \frac{S_t}{I(X)_t} = \frac{\deg X}{(\dim X)!} t^{\dim X} + O(t^{\dim X - 1}),$$

where S_t denotes the homogeneous polynomials of degree t and $I(X)_t$ the subspace of S_t containing all elements vanishing on X. The term $O(t^{\dim X - 1})$ is in the standard big-O notation, meaning that the remainder is bounded by a constant multiple of $t^{\dim X - 1}$ for all sufficiently large t.

A version of Bézout's theorem is that if $X = V(f_1, \ldots, f_n)$ is finite, then

$$\#X \leq \deg X = \deg(f_1) \cdots \deg(f_n).$$

This has a vast generalization (cf. [56, Theorem 12.4]), of which we consider a special case.

Theorem 5.16 (Thom–Porteous formula (corank one))**.** *Let $\boldsymbol{M}(\boldsymbol{x}) \in R_{n+1}^{f \times e}$ $(e \leq f)$ be a matrix of homogeneous polynomials in $\boldsymbol{x} \in \mathbb{P}^n$ whose columns have constant degree $d_1, \ldots, d_e$. Assume that the determinantal variety of rank $e-1$, $\Sigma_{e-1}(\boldsymbol{M})$, has the expected dimension*

$$\dim \Sigma_{e-1}(\boldsymbol{M}) = n - (f - (e-1))(e - (e-1)) = n - (f - e + 1).$$

Then

$$\deg \Sigma_{e-1}(\boldsymbol{M}) = T^{f-e+1}\text{-coefficient of } \frac{1}{(1 - d_1 T) \cdots (1 - d_e T)}.$$

More involved formulae exist for any rank k, and for more general matrices (e.g., $e > f$ and/or mixed degrees, assuming their minors are still homogeneous) [56, Theorem 12.4].

If the dimension of $\Sigma_{e-1}(\boldsymbol{M})$ is zero, that is, if the determinantal variety consists of finitely many elements, then this gives a way to count the number of points

$\boldsymbol{x} \in \mathbb{P}^n$ such that $\boldsymbol{M}(\boldsymbol{x})$ is of sub-maximal rank. This count, however, is only correct if we take the multiplicities into account, and is an overcount if the maximal minors of $\boldsymbol{M}$ do not define $\Sigma_{e-1}(\boldsymbol{M})$ in a reduced way.

A sufficient criterion for the maximal minors to define a determinantal variety in a reduced way is the following situation.

Lemma 5.17 ([100, Theorem 0.4.15])**.** *Let $X \subseteq \mathbb{C}^n$ be an affine variety and consider a matrix polynomial $\boldsymbol{M}(\boldsymbol{x}) \in R_n^{f \times e}$, with $e \leq f$. Define a (quasi-projective) variety $\mathcal{K}$ by the f equations $\boldsymbol{M}(\boldsymbol{x})\boldsymbol{v} = 0$, that is,*

$$\mathcal{K} = \left\{ (\boldsymbol{x}, [\boldsymbol{v}]) \in X \times \mathbb{P}^{e-1} \;\middle|\; \boldsymbol{M}(\boldsymbol{x}) \cdot \boldsymbol{v} = 0 \right\}.$$

If these f equations define $\mathcal{K}$ in a reduced way and if $\forall\, \boldsymbol{x} \in X : \operatorname{rank} \boldsymbol{M}(\boldsymbol{x}) \in \{e, e-1\}$, then the e-minors of $\boldsymbol{M}$ define $\Sigma_{e-1}(\boldsymbol{M})$ in a reduced way as well.

Example 5.18 (Example 5.15 continued)**.** *We compute the degree of $X = \Sigma_1(\boldsymbol{H}) \subset \mathbb{P}^3$ introduced in Example 5.15. Because $\dim X = 1$ and X is not contained in the hyperplane at infinity H_∞, we consider the affine chart $x_0 = 1$ where $X' = X \cap \mathbb{C}^3$ is parametrized by $t \mapsto (t, t^2, t^3)$. Choose a generic linear polynomial $\ell = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \beta$ with independent coefficients. The ideal $\langle I(X'), \ell \rangle$ can be characterized by substitution of the parametrization, leading to the univariate cubic: $\alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \beta$, from which it follows that $\deg X = 3$.*

Because every entry in $\boldsymbol{H}$ is linear, both columns have constant degree one, so the hypotheses of Theorem 5.16 apply with $f = 3$, $e = 2$ and $n = 3$. To find the degree, we expand $(1-T)^{-2} = \sum_{k \geq 0} (k+1) T^k$. The coefficient of $T^{f-e+1} = T^2$ is 3; Hence, $\deg X = 3$, precisely the number obtained in the previous paragraph.

5.4 Finiteness of the Solution Set

In this section, we perform a comprehensive algebro-geometric analysis of the variety defined by the critical equations of the continuous-time, single-input single-output $\mathcal{H}_2$-norm model reduction problem. We stratify its components, derive its dimension using the fiber dimension theorem and establish reducedness of the points of interest (Theorem 5.19). This gives a comprehensive proof that for most given models H, the set of critical points is finite (Corollary 5.21). We will build on these results in the remainder of this chapter (Sections 5.5 and 5.6). The main arguments developed in Sections 5.4-5.6, along with their interdependencies leading to the final result, namely the count of the number of critical points, are summarized schematically in Figure 5.2.

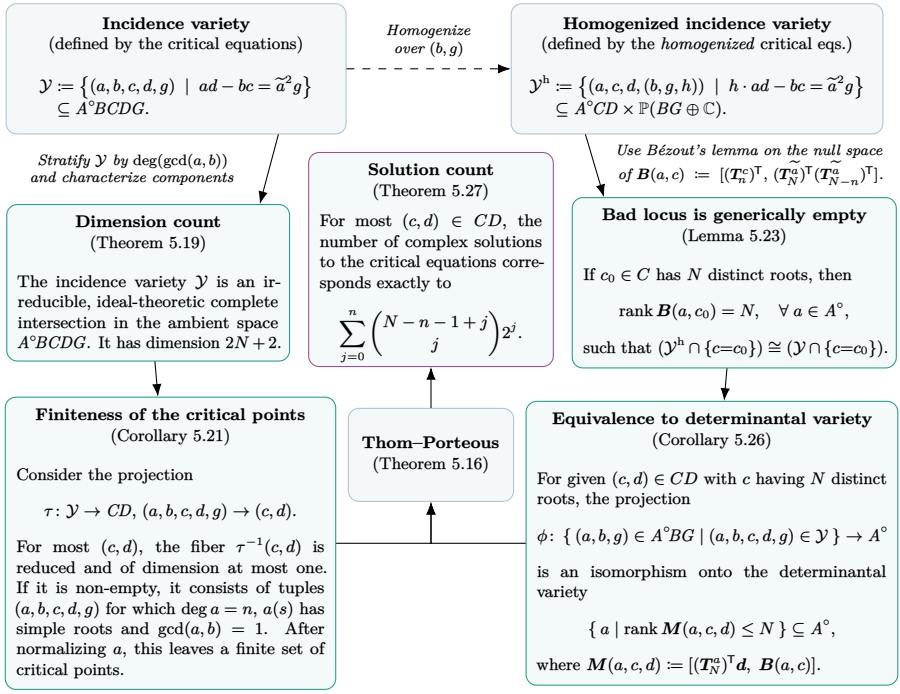


Figure 5.2: Schematic overview of the logical structure and dependencies between the arguments developed in Sections 5.4–5.6. The diagram illustrates how the analysis of the incidence variety and the corresponding interpretation as a determinantal variety lead to the main result of this chapter, which is the exact count of the number of (complex) critical points. All notation appearing here will be formally introduced in the corresponding sections; the figure is intended mainly as a navigational aid which allows the reader to return to it at any stage to see how each step fits into the overall argument.

We will deal with several products of vector spaces. To ease notation, we use the following shorthands:

$$A \coloneqq P_{\leq n}, \quad B \coloneqq P_{\leq n-1}, \quad C \coloneqq P_{\leq N}, \quad D \coloneqq P_{\leq N-1}, \quad G \coloneqq P_{\leq N-n-1}.$$

Additionally, $A^\circ \coloneqq A \setminus 0$, and we use, for example, $A^\circ BG$ to denote the product $(A \setminus 0) \times B \times G$.

Consider the *incidence variety* defined by the critical equations:

$$\mathcal{Y} \coloneqq \left\{ (a,b,c,d,g) \;\middle|\; ad - bc = \widetilde{a}^2 g \right\} \subseteq A^\circ BCDG,$$

that is, the variety consisting of all tuples (a,b,c,d,g) satisfying the given algebraic relation.

Theorem 5.19 (Dimension count)**.** *The subvariety $\mathcal{Y}$ is an irreducible ideal-theoretic complete intersection in $A^\circ BCDG$ of dimension $2N+2$.*

Before proving Theorem 5.19, we point out a useful structure in the equations defining $\mathcal{Y}$. Consider the projection

$$\pi \colon \mathcal{Y} \to ABG, \;\; (a,b,c,d,g) \to (a,b,g).$$

For fixed (a,b,g), the right-hand side of the defining equation of $\mathcal{Y}$ is constant, so the unknowns (c,d) are constrained by an affine system of linear equations:

$$\begin{bmatrix} (\boldsymbol{T}_N^a)^\mathsf{T} & -(\boldsymbol{T}_{N+1}^b)^\mathsf{T} \end{bmatrix} \begin{bmatrix} \boldsymbol{d} \\ \boldsymbol{c} \end{bmatrix} = (\boldsymbol{T}_N^{\widetilde{a}})^\mathsf{T} (\boldsymbol{T}_{N-n}^{\widetilde{a}})^\mathsf{T} \boldsymbol{g}. \tag{5.6}$$

Consequently, each fiber of π is an affine linear subspace:

$$\pi^{-1}(a,b,g) = (a,b) \times \{\, (c,d) \mid (5.6) \text{ is satisfied} \,\} \times (g). \tag{5.7}$$

Lemma 5.20. *For given $(a,b,g) \in A^\circ BG$, the fiber $\pi^{-1}(a,b,g)$ is non-empty if and only if the greatest common divisor of (a,b) divides $\widetilde{a}^2 g$, i.e., $\gcd(a,b) \mid \widetilde{a}^2 g$. In that case, we have that*

$$\begin{aligned} \dim \pi^{-1}(a,b,g) &= N + \deg \gcd(a,b) - \max\{\deg b,\ \deg a - 1\} \\ &= N + \deg \gcd(a^{\mathrm{h}}, b^{\mathrm{h}}) - (n-1). \end{aligned}$$

Here $a^{\mathrm{h}} \in \mathbb{C}[s,t]_{\leq n}$ and $b^{\mathrm{h}} \in \mathbb{C}[s,t]_{\leq n-1}$ denote the homogenization of the polynomials a and b, respectively.

Proof. Let $a' = a/\gcd(a,b)$ and $b' = b/\gcd(a,b)$ such that $\gcd(a',b') = 1$. It is immediate from

$$\gcd(a,b) \cdot (a'd - b'c) = \widetilde{a}^2 g$$

that $\gcd(a,b) \mid \widetilde{a}^2 g$ is necessary for the fiber (5.7) to be non-empty. We now show that it is also sufficient.

Let $\boldsymbol{S}(a,b) \coloneqq \begin{bmatrix} (\boldsymbol{T}_N^a)^\mathsf{T} & -(\boldsymbol{T}_{N+1}^b)^\mathsf{T} \end{bmatrix} \in \mathbb{C}^{(N+n)\times(2N+1)}$. Any element $(\boldsymbol{u}^\mathsf{T}, \boldsymbol{v}^\mathsf{T})^\mathsf{T} \in \mathbb{C}^{2N+1}$ in the null space of $\boldsymbol{S}(a,b)$ corresponds to polynomials $u \in P_{\leq N-1}$ and $v \in P_{\leq N}$ satisfying $au + bv = 0$. Removal of the common factor of a and b gives $a'u + b'v = 0$, which implies $a'u = -b'v$.

Since $\gcd(a',b') = 1$, this can only be satisfied if there exists a polynomial $r \neq 0$ such that $u = b'r$ and $v = -a'r$. Let $k \coloneqq \deg \gcd(a,b)$, then taking the degrees of u and v into account gives

$$\deg r \leq \min\Big\{ \underbrace{(N-1)}_{\deg u} - \underbrace{(\deg b - k)}_{\deg b'},\ \underbrace{(N)}_{\deg v} - \underbrace{(\deg a - k)}_{\deg a'} \Big\}.$$

Because r is determined by $\deg r + 1$ coefficients and each choice of r leads to a unique element in the null space of the matrix $\boldsymbol{S}(a,b)$, we have

$$\begin{aligned} \operatorname{nullity}(\boldsymbol{S}(a,b)) &= \min\big\{N - \deg b + k,\ N - (\deg a - 1) + k\big\} \\ &= N + k - \max\{\deg b,\ \deg a - 1\}. \end{aligned}$$

Taking into account the size of $\boldsymbol{S}(a,b)$ gives[5]

$$\operatorname{rank} \boldsymbol{S}(a,b) = N + \max\{\deg b,\ \deg a - 1\} + 1 - k. \tag{5.8}$$

The image of $\boldsymbol{S}(a,b)$ is the subspace of coefficient vectors corresponding to the following polynomials $\{\, pa + qb : \deg p \leq N-1, \deg q \leq N \,\}$. Each element of this set is divisible by $\gcd(a,b)$, so that

$$\operatorname{image} \boldsymbol{S}(a,b) \subseteq \big\{\, \gcd(a,b) \cdot l \ \mid\ l \in P_{\leq N + \max\{\deg b,\ \deg a - 1\} - k} \,\big\}.$$

We know from (5.8) that the dimension of $\operatorname{image} \boldsymbol{S}(a,b)$ equals $N + \max\{\deg b,\ \deg a - 1\} - k + 1$, so that this inclusion is in fact equality. Therefore, for the right-hand side of (5.6) to lie in $\operatorname{image} \boldsymbol{S}(a,b)$, it is necessary and sufficient that $\gcd(a,b)$ divides $\widetilde{a}^2 g$. When this divisibility holds, the affine space of solutions $\pi^{-1}(a,b,g)$ has dimension equal to the dimension of the null space of $\boldsymbol{S}(a,b)$. □

[5]Note that this rank formula generalizes the familiar fact that the rank of the classical Sylvester matrix of a and b of size $(2n-1) \times (2n-1)$ equals $2n - 1 - \deg \gcd(a,b)$.

Now that we have established Lemma 5.20, we can present a proof of Theorem 5.19.

Proof of Theorem 5.19. The proof will proceed in the following steps: First, we stratify $\mathcal{Y}$ by the degree of the greatest common divisor of a and b, this will allow us to determine the dimension of $\mathcal{Y}$. Then, we show that $\mathcal{Y}$ is a set-theoretic complete intersection and irreducible. Lastly, reducedness will follow from a bundle argument, so $\mathcal{Y}$ is an ideal-theoretic complete intersection.

For $0 \leq k \leq n-1$, consider the following two subsets of $A^\circ BG$:

$$\mathcal{S}_k \coloneqq \mathcal{S}_{\geq k} \setminus \mathcal{S}_{\geq k+1},$$

where $\mathcal{S}_{\geq k} \coloneqq \left\{ (a,b,g) \;\middle|\; \deg \gcd(a^{\mathrm{h}}, b^{\mathrm{h}}) \geq k \right\}$, and

$$\mathcal{T}_k \coloneqq \left\{ (a,b,g) \in \mathcal{S}_k \;\middle|\; \gcd(a,b) \text{ divides } \widetilde{a}^2 g \right\}.$$

The defining equations of $\mathcal{S}_{\geq k}$ are given by the vanishing of $(k+1)$-minors of the (square) Sylvester matrix constructed from the polynomials a and b. Within $\mathcal{S}_k$, the divisibility constraint of $\mathcal{T}_k$ is given by the $(k+1)$-minors of the matrix $\boldsymbol{S}(a,b)$ concatenated with the coefficient vector of $\widetilde{a}^2 g$.

The variety $\mathcal{S}_{\geq k}$ is the image of the polynomial map

$$\underbrace{(P_k \setminus 0)}_{f} \times \underbrace{(P_{n-k} \setminus 0)}_{u} \times \underbrace{P_{n-k-1}}_{v} \times G \to A^\circ BG, \qquad (f,u,v,g) \mapsto (fu, fv, g),$$

hence is irreducible. Given $(a,b,g) \in \mathcal{S}_{\geq k}$, there are only finitely many factorizations $a = fu$, $b = fv$ up to rescaling f by $\lambda \in \mathbb{C} \setminus \{0\}$; Hence, the fibers of this map are 1-dimensional. From this we conclude that $\mathcal{S}_{\geq k}$ is an irreducible variety of dimension $N + n + 1 - k$.

If $k = 0$, then the divisibility constraint is trivial and $\mathcal{T}_0 = \mathcal{S}_0$. For $k \geq 1$, $\mathcal{T}_k \subseteq \mathcal{S}_k$ is a strict subvariety, since there are tuples $(a,b,g) \in \mathcal{S}_k$ that do *not* satisfy the divisibility constraint (e.g., $a = (s-1)^n$, $b = (s-1)^k (s-2)^{n-1-k}$, $g = (s-3)^{N-n-1}$). Therefore $\dim \mathcal{T}_0 = N + n + 1$ and $\dim \mathcal{T}_k \leq N + n - k$ for $k \geq 1$.

Stratifying $\mathcal{Y}$ by the degree of $\gcd(a,b)$ then gives us this (locally closed) decomposition

$$\mathcal{Y} = \bigcup_{k=0}^{n-1} \mathcal{Y}_k, \qquad \mathcal{Y}_k = \left\{ (a,b,c,d,g) \in \mathcal{Y} \mid \deg \gcd(a,b) = k \right\}.$$

By (5.7) and Lemma 5.20, $\mathcal{Y}_k$ is an affine bundle over $\mathcal{T}_k$ of rank at most $N + k - (n-1)$.

In particular, each (non-empty) $\mathcal{Y}_k$ is irreducible of dimension at most

$$\begin{aligned}\dim \mathcal{Y}_k &= \dim \mathcal{T}_k + \operatorname{rank}_{\mathcal{T}_k} \mathcal{Y}_k \\ &\leq (N+n-k+1) + (N+k-(n-1)) = 2N+2,\end{aligned}$$

where equality holds only for $k=0$, so that $\dim \mathcal{Y} = \max_k \dim \mathcal{Y}_k = 2N+2$.

From this we can conclude that $\mathcal{Y}$ is a set-theoretic complete intersection:

$$\operatorname{codim}_{A^\circ BCDG} \mathcal{Y} = \underbrace{(n+1)+n+(N+1)+N+(N-n)}_{\text{ambient dimension}} - \underbrace{2N+2}_{\dim \mathcal{Y}} = \underbrace{N+n}_{\#\text{equations}}.$$

Theorem 5.9 then shows that every irreducible component of $\mathcal{Y}$ has dimension exactly $2N+2$. On the other hand, $\mathcal{Y} = \bigcup_{k=0}^{n-1} \overline{\mathcal{Y}_k}$ is a union of irreducible subvarieties, so none of the $\mathcal{Y}_k$, $k \geq 1$, are components, as they have smaller dimension. In this way we conclude that $\mathcal{Y} = \overline{\mathcal{Y}_0}$ is an irreducible variety.

Finally, we have established that $\mathcal{Y}_0$ is a dense open subset of $\mathcal{Y}$, and it is a bundle over the smooth variety $\mathcal{T}_0$, an open set of $\mathbb{C}^{N+n+1}$. In particular, $\mathcal{Y}$ is reduced (even smooth) in $\mathcal{Y}_0$ and by the unmixedness Theorem 5.10, the variety $\mathcal{Y}$ is reduced everywhere; meaning that the critical equations define $\mathcal{Y}$ in a reduced way. □

Corollary 5.21 (Finiteness of the critical points)**.** *Consider the projection $\tau\colon \mathcal{Y} \to CD$, $(a,b,c,d,g) \to (c,d)$. For almost all $(c,d) \in C \times D$, the fiber $\tau^{-1}(c,d)$ is reduced, of dimension at most 1. If it is non-empty, it consists of tuples (a,b,c,d,g) satisfying*

$$\deg a = n, \quad a(s) \text{ has simple roots}, \quad \gcd(a,b) = 1. \tag{5.9}$$

In particular, for generic $H(s) \in \mathcal{M}_N$, the set of complex solutions (a,b,g) to the critical equations (5.3) is finite up to normalization of a (e.g., fixing $a_n = 1$) and is non-degenerate (i.e., consists of simple solutions without multiplicities).

Proof. We distinguish two cases. When τ is not dominant, there is a dense open subset of points $(c,d) \in CD$ lying outside the image of τ, so that for generic choice of (c,d) we have $\tau^{-1}(c,d) = \emptyset$.

Conversely, if τ is dominant, i.e., $\overline{\tau(\mathcal{Y})} = CD$, the fibers need not be empty. In this case, the fiber dimension theorem (cf. Theorem 5.8) implies that generic fibers are reduced and of dimension

$$\dim \tau^{-1}(c,d) = \dim \mathcal{Y} - \dim CD = (2N+2) - (2N+1) = 1.$$

To prove the three additional conditions in (5.9), we must show that for a generic choice of (c, d), the fiber $\tau^{-1}(c, d)$ does not intersect the subsets of $\mathcal{Y}$ where these conditions fail. The failure of each condition is described by a polynomial equation:

$$a_n = 0, \quad \Delta(a) = 0, \quad \operatorname{res}_s(a, b) = 0, \tag{5.10}$$

respectively, where $\Delta(a)$ denotes the discriminant of a and $\operatorname{res}_s(a, b)$ is the determinant of the Sylvester matrix of a and b.

To show that these conditions from (5.10) are non-trivial on the irreducible variety $\mathcal{Y}$, we construct a point in $\mathcal{Y}$ that satisfies none of them. Let a be a generic polynomial of degree n, so that $a_n \neq 0$ and $\Delta(a) \neq 0$. Let $b(s) = 1$, which ensures that $\gcd(a, b) = 1$ and thus $\operatorname{res}_s(a, b) \neq 0$. Finally, let $g(s) = 0$. The defining equation of $\mathcal{Y}$ then simplifies to $ad - c = 0$. We can satisfy this by choosing any polynomial $d \in P_{\leq N-n}$ and setting $c = ad$. This constructs a family of points $(a, b, c, d, g) \in \mathcal{Y}$ for which none of the conditions in (5.10) hold.

Since $\mathcal{Y}$ is irreducible and none of the conditions from (5.10) holds identically on $\mathcal{Y}$, the subsets where the conditions in (5.9) fail are subvarieties of $\mathcal{Y}$; of dimension at most $2N + 1$. Let

$$\mathcal{W} \coloneqq \mathcal{Y} \cap (\{a_n = 0\} \cup \{\Delta(a) = 0\} \cup \{\operatorname{res}_s(a, b) = 0\})$$

and consider $\pi \colon \mathcal{W} \to CD \colon (a, b, c, d, g) \mapsto (c, d)$. Since the solutions to the defining equations of $\mathcal{W}$ are invariant wrt. the transformation $(a, b, c, d, g) \to (\lambda a, \lambda b, c, d, \frac{1}{\lambda^2} g)$ for any $\lambda \in \mathbb{C} \setminus \{0\}$, all non-empty fibers $\pi^{-1}(c, d)$ have dimension at least one. This then shows by the fiber dimension theorem that the image $\pi(\mathcal{W})$ has dimension strictly smaller than $\dim \mathcal{W}$, hence strictly smaller than the dimension of CD. We can conclude that for a generic choice of (c, d), the fiber $\tau^{-1}(c, d)$ will not intersect $\mathcal{W}$ and thus satisfy the conditions in (5.9).

Finally, since a generic fiber $\tau^{-1}(c, d)$ is one-dimensional and the condition $a_n \neq 0$ holds, we can normalize by setting $a_n = 1$. This reduces the dimension of the generic fiber to 0, which completes the proof. □

Remark 5.22. *We will illustrate in Sections 5.5 and 5.6 that for generic $(c, d) \in CD$, the fibers $\tau^{-1}(c, d)$ are in bijection with the determinantal variety of a matrix on projective space (cf. Corollary 5.26 and subsequent discussion), which ensures that the fibers are non-empty for most (c, d).*

5.5 Multiparameter Eigenvalue Problem

In this section, we characterize the critical points via a determinantal formulation, exploiting the fact that the variables $\boldsymbol{b}$ and $\boldsymbol{g}$ appear only linearly in the critical equations. To obtain the determinantal description, we homogenize the critical equations with respect to b and g. We show that for generic c this homogenization does not introduce spurious components in the resulting incidence variety (Lemma 5.23), so that the determinantal formulation can legitimately be used to analyze the critical points (Corollary 5.26).

Introduce a scalar variable $h \in \mathbb{C}$ and define the *homogenized critical equations* by:

$$h \cdot \left((\boldsymbol{T}_N^a)^\mathsf{T}\boldsymbol{d}\right) - (\boldsymbol{T}_n^c)^\mathsf{T}\boldsymbol{b} - (\widetilde{\boldsymbol{T}_N^a})^\mathsf{T}(\widetilde{\boldsymbol{T}_{N-n}^a})^\mathsf{T}\boldsymbol{g} = \boldsymbol{0},$$

where, as before, $\widetilde{a} \coloneqq a(-s)$. The partially linear structure of these equations allows us to write them more compactly:

$$\boldsymbol{M}(a,c,d)\begin{bmatrix} -h \\ \boldsymbol{b} \\ \boldsymbol{g} \end{bmatrix} = \boldsymbol{0}, \text{ where } \boldsymbol{M}(a,c,d) \coloneqq \begin{bmatrix} (\boldsymbol{T}_N^a)^\mathsf{T}\boldsymbol{d} & (\boldsymbol{T}_n^c)^\mathsf{T} & (\widetilde{\boldsymbol{T}_N^a})^\mathsf{T}(\widetilde{\boldsymbol{T}_{N-n}^a})^\mathsf{T} \end{bmatrix}. \tag{5.11}$$

This system is, by construction, homogeneous of degree one in the variables $(h, \boldsymbol{b}, \boldsymbol{g})$. That is, for any $\lambda \in \mathbb{C} \setminus \{0\}$, the substitution $(h, \boldsymbol{b}, \boldsymbol{g}) \mapsto (\lambda h, \lambda\boldsymbol{b}, \lambda\boldsymbol{g})$ gives the same set of equations, but rescaled by λ.

We now exploit the homogeneity in (5.11) by projectivizing the variables $(\boldsymbol{b}, \boldsymbol{g}, h)$. (It will become clear later why this projectivization is useful for the analysis of the number of critical points.) This leads to the quasi-projective incidence variety

$$\mathcal{Y}^{\mathrm{h}} \coloneqq \{\, (a, c, d, (b, g, h)) \mid (5.11) \text{ is satisfied} \,\} \subseteq A^{\circ}CD \times \mathbb{P}(BG \oplus \mathbb{C}).$$

This construction may be viewed as a partial compactification[6] of the affine incidence variety $\mathcal{Y}$. Indeed, the affine chart $h = 1$ coincides with $\mathcal{Y}$, while the locus $\mathcal{Y}^{\infty} \coloneqq \mathcal{Y}^{\mathrm{h}} \cap \{h = 0\}$ is a new, additional part "at infinity" along the direction of (b, g). Thus, $\mathcal{Y}^{\mathrm{h}}$ decomposes as a disjoint union

$$\mathcal{Y}^{\mathrm{h}} = \mathcal{Y} \cup \mathcal{Y}^{\infty}.$$

The locus $\mathcal{Y}^{\infty}$ does not contain critical points of the model reduction problem; it arises solely from the introduction of the homogenizing variable h. For this

[6] A *partial compactification* of a variety X is a larger space in which X sits as an open subset, obtained by adding some, but not necessarily all, points at infinity.

reason, we refer to $\mathcal{Y}^\infty$, and its restriction to a specific choice of $(c,d) \in CD$, as the *bad locus*[7].

Let

$$\boldsymbol{B}(a,c) \coloneqq \begin{bmatrix} (\boldsymbol{T}_n^c)^\mathsf{T} & (\boldsymbol{T}_N^{\widetilde{a}})^\mathsf{T} (\boldsymbol{T}_{N-n}^{\widetilde{a}})^\mathsf{T} \end{bmatrix},$$

then it follows from (5.11) that

$$\mathcal{Y}^\infty = \left\{ a,c,d,(b,g,h)) \mid \boldsymbol{B}(a,c) \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{g} \end{bmatrix} = \boldsymbol{0} \text{ and } h = 0 \right\} \subseteq A^\circ CD \times \mathbb{P}(BG \oplus \mathbb{C}).$$

This description indicates that inside the bad locus, the maximal minors of $\boldsymbol{B}(a,c)$ must vanish. We will use this insight to show that the bad locus $\mathcal{Y}^\infty$, when restricted to a specific $c \in C$, is empty for most choices c.

Lemma 5.23 (Bad locus is generically empty)**.** *For given $(a,c) \in A^\circ C$,*

$$\text{nullity}\,(\boldsymbol{B}(a,c)) = \max\Big\{0,\ \deg \gcd(c,\widetilde{a}^2) + \min\big\{N - \deg c - n,\ n - 2\cdot\deg a\big\}\Big\}.$$

It follows naturally that if $c_0 \in C$ has N distinct roots, then

$$\operatorname{rank} \boldsymbol{B}(a,c_0) = N, \quad \forall\, a \in A^\circ,$$

such that in this case $(\mathcal{Y}^\mathrm{h} \cap \{c = c_0\}) \cong (\mathcal{Y} \cap \{c = c_0\})$.

Proof. We use Bézout's relation to inspect the nature of $\operatorname{rank} \boldsymbol{B}(a,c)$. For arbitrary $\boldsymbol{u} \in \mathbb{C}^n$, $\boldsymbol{v} \in \mathbb{C}^{N-n}$,

$$\boldsymbol{B}(a,c) \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} = \boldsymbol{0} \iff cu + \widetilde{a}^2 v = 0.$$

Let $f \coloneqq \gcd(c, \widetilde{a}^2)$. Pruning f gives $uc' + va' = 0$, where $c = fc'$ and $\widetilde{a}^2 = fa'$. For this to be satisfied for nonzero v and u, we need $v = -c'w$ and $u = a'w$ for some polynomial $w \neq 0$. Taking into account the maximum degrees of u and v gives

$$\deg w \leq \min\big\{ \underbrace{(N-n-1)}_{\deg v} - \underbrace{(\deg c - \deg f)}_{\deg c'},\ \underbrace{(n-1)}_{\deg u} - \underbrace{(2\cdot\deg a - \deg f)}_{\deg a'} \big\}.$$

Since $w \neq 0$ is parametrized using $\deg w + 1$ coefficients, and each choice of w with the appropriate degree constitutes an element in the null space of $\boldsymbol{B}(a,c)$, we conclude that

$$\text{nullity}\,(\boldsymbol{B}(a,c)) = \max\Big\{0,\ \deg f + \min\big\{N - \deg c - n,\ n - 2\cdot\deg a\big\}\Big\}. \tag{5.12}$$

[7]The appearance of such a "bad locus" and its implications have been visualized schematically in Figure 1.6.

When $c_0 \in C$ is such that $\deg c_0 = N$, the minimum in (5.12) is upper bounded by $-n$. If in addition, c_0 has distinct roots only, $\deg f \leq n$ by construction, so that $\boldsymbol{B}(a, c)$ has full column rank for all $a \in A^\circ$. In this case $\mathcal{Y}^\infty \cap \{c = c_0\} = \emptyset$, which completes the proof. □

Example 5.24 (Multiple poles)**.** *If $\deg c = N$ but c has multiple roots, the set of polynomials a for which $\boldsymbol{B}(a, c)$ fails to have full rank can be finite (modulo normalization, e.g., by fixing $a_n = 1$) or infinite. For illustration, consider $(N, n) = (2, 1)$ with $c = (s - p)^2$ for some $p \in \mathbb{C}$. In this case, $\boldsymbol{B}(a, c)$ drops rank in the case $a = (s + p)$. Alternatively, consider $(N, n) = (4, 3)$ and $c = (s - p_1)^2 (s - p_2)^2$ for some $p_1, p_2 \in \mathbb{C}$. Here, $\boldsymbol{B}(a, c)$ drops rank when $a = (s + p_1)(s + p_2)$ or $a = (s + p_1)(s + p_2)(s - p_3)$ for arbitrary $p_3 \in \mathbb{C}$.*

Remark 5.25. *In the context of the model reduction problem* (5.1)*, the polynomials (c, d) are determined by the given higher-order model H. Since the parameter N is explicitly defined as the degree of c in the setup of the problem, the condition $c_N \neq 0$ is trivial. The condition of c having simple roots only is reasonable in practice: the higher-order model H often originates from system identification experiments using noisy measurement data.*

Lemma 5.23 shows that, as long as c has N distinct roots (an assumption which, as discussed above, is almost always satisfied in practice), the fibers of $\mathcal{Y}$ and of its homogenization $\mathcal{Y}^{\mathrm{h}}$ from their restriction to this choice of c coincide: homogenizing the critical equations over (b, g) does not create an extra component in their solution set (the bad locus is empty for such c). It follows that for c having N distinct roots, we may equally well continue our analysis of the critical points using the null space incidence from (5.11), with the kernel vector (b, g, h) living in projective space. This gives a characterization of the critical points in terms of the rank behavior of $\boldsymbol{M}(a, c, d)$. In particular, points of $\mathcal{Y}$ (restricted to such c) are precisely those for which $\boldsymbol{M}(a, c, d)$ drops rank, and we can exploit this characterization to eliminate the variables (b, g, h) from the subsequent analysis.

Corollary 5.26 (Equivalence to determinantal variety)**.** *For given $(c, d) \in CD$ with c having N distinct roots, the projection*

$$\phi\colon \{\, (a, b, g) \in A^\circ BG \mid (a, b, c, d, g) \in \mathcal{Y} \,\} \to A^\circ$$

is an isomorphism onto the determinantal variety

$$\Sigma_N(\boldsymbol{M}(a, c, d)) = \{\, a \mid \operatorname{rank} \boldsymbol{M}(a, c, d) \leq N \,\} \subseteq A^\circ. \tag{5.13}$$

If (c, d) are generic, then $\Sigma_N(\boldsymbol{M}(a, c, d))$ is a finite set (finite up to scaling of a) of reduced points; that is, the maximal minors of $\boldsymbol{M}(a, c, d)$ define the elements in a reduced way.

Proof. Fix $(c,d) \in CD$ with c having N distinct roots. For any $(a,b,c,d,g) \in \mathcal{Y}$, the defining equations of $\mathcal{Y}$ imply that $(-1,b,g)$ is a nonzero null vector of $\boldsymbol{M}(a,c,d)$, so $\boldsymbol{M}(a,c,d)$ is rank-deficient. Conversely, by Lemma 5.23, for any $a \in \Sigma_N(\boldsymbol{M}(a,c,d))$ we have $\operatorname{rank} \boldsymbol{B}(a,c) = N$, and hence the null space of $\boldsymbol{M}(a,c,d)$ is one-dimensional. Thus there is a unique $(b,g,h) \in \mathbb{P}(BG \oplus \mathbb{C})$ with $h \neq 0$ in this null space. Normalizing to $h = 1$ yields a unique pair (b,g) and therefore a unique point (a,b,g) with $(a,b,c,d,g) \in \mathcal{Y}$. This shows that ϕ is bijective onto $\Sigma_N(\boldsymbol{M}(a,c,d))$. Since ϕ is a morphism and its inverse is obtained algebraically from the one-dimensional kernel, ϕ is an isomorphism onto its image.

Corollary 5.21 further shows that, for generic (c,d), the fiber $\tau^{-1}(c,d) \in \mathcal{Y}$ is finite up to scaling of a. Moreover, the same corollary asserts that for generic (c,d) the critical equations define the fibers $\tau^{-1}(c,d)$ in a reduced way. By Lemma 5.17, this reducedness carries over to the determinantal description. □

Instead of solving the critical equations directly with general-purpose polynomial system solvers, such as homotopy continuation methods (e.g., [20]), one can use the determinantal variety in (5.13) to determine the complex solutions to the critical equations. Assuming the data c, d are such that this variety consists of finitely many points (finite up to normalization of a), its elements can be computed with dedicated multiparameter eigenvalue problem (MEP) solvers (e.g., [90, 176, 178]). Indeed, the polynomial matrix $\boldsymbol{M}(a,c,d)$ corresponds to the *multiparameter eigenvalue problem (MEP)* formulation proposed in our earlier work [109] (cf. Chapter 2).

5.6 Degree of the Solution Set

To obtain a finite set, we restrict our analysis to a specific choice of $(c,d) \in CD$ and consider the projectivization of the determinantal variety $\Sigma_N(\boldsymbol{M}(a,c,d))$. Since all entries of the matrix $\boldsymbol{M}(a,c,d)$ are homogeneous in a, this step is natural and leads to the projective variety

$$\Sigma_{c,d} \coloneqq \mathbb{P}(\Sigma_N(\boldsymbol{M}(a,c,d))) \subseteq \mathbb{P}(A) = \{\, a \in \mathbb{P}(A) \mid \operatorname{rank} \boldsymbol{M}(a,c,d) \leq N \,\}.$$

We are now in a position to count the (complex) critical points. Applying the Thom–Porteous formula to the projective determinantal variety $\Sigma_{c,d}$ yields its degree.

Theorem 5.27 (Solution count)**.** *Let m denote the number critical points, that is, the number of complex solutions to the critical equations* (5.1)*. Assume that*

(c, d) are such that the set $\Sigma_{c,d}$ is finite, then

$$m \leq \deg \Sigma_{c,d} = \sum_{j=0}^{n} \binom{N-n-1+j}{j} 2^j. \tag{5.14}$$

Moreover, if (c, d) are generic, then equality holds.

Proof. The expected dimension of $\Sigma_{c,d}$ corresponds to $n - (N+1-N)(N+n-N) = 0$. Thus, under the assumption that $\Sigma_{c,d}$ is finite, we can apply the Thom–Porteous formula for the corank one case, cf. Theorem 5.16. The coefficient of T^n in the series expansion of $(1-T)^{-1}(1-2T)^{-(N-n)}$ can be determined by expressing each factor as a power series. Specifically,

$$\frac{1}{1-T} = \sum_{j=0}^{\infty} T^j, \qquad \frac{1}{(1-2T)^{N-n}} = \sum_{j=0}^{\infty} \binom{N-n+j-1}{j} (2T)^j,$$

such that by the Cauchy product[8],

$$\frac{1}{(1-T)(1-2T)^{N-n}} = \sum_{k=0}^{\infty} \left(\sum_{j=0}^{k} \binom{N-n+j-1}{j} 2^j \right) T^k.$$

In this series, the coefficient of T^n corresponds to

$$\sum_{j=0}^{n} \binom{N-n-1+j}{j} 2^j.$$

If (c, d) are generic, then by Corollary 5.26 the maximal minors define $\Sigma_{c,d}$ in a reduced way, so that $\deg \Sigma_{c,d} = \#\Sigma_{c,d}$, and $\Sigma_{c,d}$ is in bijection with the solutions of (5.1). □

For generic instances of the model reduction problem (5.1)—in particular, when the given higher-order model H has N distinct poles—the number of critical points coincides with the formula in (5.14). Since the genuine critical points (e.g., local minima, saddle points) of the optimization problem correspond to the subset of real-valued solutions for which $\widehat{H}$ is stable, this count gives an upper bound on the number of genuine critical points of the optimization problem. Theorem 5.27, therefore, achieves the main objective of this chapter. In contrast to the number of complex critical points, the number of real-valued, stable solutions depends on the specific choice of the data c, d and cannot, in general, be determined solely from the problem size (N, n).

[8]Each coefficient of the product is given by the discrete convolution of the coefficient vectors of its factors

Table 5.1: Evaluation of the formula in (5.14) from Theorem 5.27 for various problem sizes (N, n). Under the assumption that c, d are generic, so that c has N distinct roots, each entry gives the number of complex solutions to the critical equations of the corresponding instance of the model reduction problem. Empty cells indicate infeasible problem sizes, i.e., combinations that do not satisfy $N > n$.

(N,n)	1	2	3	4	5	6	7	8	9	10	11
2	3										
3	5	7									
4	7	17	15								
5	9	31	49	31							
6	11	49	111	129	63						
7	13	71	209	351	321	127					
8	15	97	351	769	1023	769	255				
9	17	127	545	1471	2561	2815	1793	511			
10	19	161	799	2561	5503	7937	7423	4097	1023		
11	21	199	1121	4159	10625	18943	23297	18943	9217	2047	
12	23	241	1519	6401	18943	40193	61183	65537	47103	20481	4095

Several values of the expression (5.1) are depicted in Table 5.1. If we reduce the order of the given model by one, i.e., $n = N - 1$, the formula (5.14) reduces to $\sum_{j=0}^{n} 2^j = 2^{n+1} - 1$.

Example 5.28. *In [109, Example 2] the authors reported that the critical equations for $(N, n) = (6, 2)$ have 49 complex solutions, exactly matching the value predicted by the expression in* (5.14)*. For that specific instance, 11 of the solutions were real-valued, 5 of which corresponded to stable models.*

Even when c has multiple roots, so that the bad locus is nonempty, the Thom–Porteous formula still provides an upper bound on the number of complex critical points, as long as the bad locus is finite. Let $\operatorname{mult}_a I_{N+1}(\boldsymbol{M}(a, c, d))$ denote the multiplicity of a with respect to the ideal generated by the $(N+1)$-minors of $\boldsymbol{M}(a, c, d)$ and consider the bad locus

$$\mathcal{B} \coloneqq \left\{ a \in \mathbb{P}(A) \mid \operatorname{rank} \boldsymbol{B}(a, c) < N \right\} \subseteq \Sigma_{c,d}.$$

When $\mathcal{B}$ is non-empty but finite, the number of critical points m satisfies

$$m = \#\Sigma_{c,d} - \#\mathcal{B} \leq \deg \Sigma_{c,d} - \sum_{a \in \mathcal{B}} \operatorname{mult}_a I_{N+1}(\boldsymbol{M}(a, c, d)) \leq \deg \Sigma_{c,d} - \deg \mathcal{B}.$$

Intuitively, part of the generic count $\deg \Sigma_{c,d}$ (cf. Theorem 5.27) is accounted for by points in the bad locus; subtracting their contribution decreases the resulting count accordingly. The impact of each such point in the bad locus

on $\deg \Sigma_{c,d}$ is governed by its multiplicity with respect to the maximal minors of $\boldsymbol{M}(a,c,d)$, which is in turn bounded below by the multiplicity of the point with respect to the minors of $\boldsymbol{B}(a,c)$.

Special cases in which c leads to a finite but nonempty bad locus can be analyzed individually. To conclude this section, we present one such example below.

Theorem 5.29. *Let m denote the number of complex critical points. If (c,d) is such that $\Sigma_{c,d}$ is finite and $\deg c = N$ but c has exactly one double root, then*

$$\#\mathcal{B} = \binom{N-2}{n-1},$$

so that

$$m \leq \sum_{j=0}^{n} \binom{N-n-1+j}{j} 2^j - \binom{N-2}{n-1}.$$

Proof. Write $c = (s-p_1)^2 \prod_{i=2}^{N-1}(s-p_i)$ with $p_1, \ldots, p_{N-1}$ distinct. The set $\Sigma_{c,d}$ can be split into those points that correspond to critical points of (5.1) and bad locus points. The solutions corresponding to $\mathcal{B}$ are those with $\operatorname{rank} \boldsymbol{B}(a,c) < N$. From Lemma 5.23 we deduce that this happens if and only if $\widetilde{a} = (s-p_1)\prod_{i\in I}(s-p_i)$, where $I \subseteq \{2,\ldots,N-1\}$ is a selection of $n-1$ roots of c. Thus $\mathcal{B}$ corresponds of exactly $\binom{N-2}{n-1}$ points, potentially with multiplicity. We conclude using the first part of Theorem 5.27

$$m = \#\Sigma_{c,d} - \#\mathcal{B} \overset{\text{Theorem 5.27}}{\leq} \sum_{j=0}^{n} \binom{N-n-1+j}{j} 2^j - \binom{N-2}{n-1}. \qquad \square$$

Example 5.30 (Given model with multiple roots)**.** *Consider the third-order model with one multiple pole:*

$$H = \frac{z^2 - 5z + 6}{z^3 - 1.8z^2 + 1.05z - 0.2} = \frac{(z-2)(z-3)}{(z-1/2)^2(z-0.8)}.$$

We compute its optimal first-order approximant ($n = 1$). For this setup, the matrix $\boldsymbol{M}(a,c,d)$ corresponds to

$$\boldsymbol{M}(a,c,d) = \begin{bmatrix} 6a_0 & -0.2 & a_1^2 & 0 \\ 6a_1 - 5a_0 & 1.05 & 2a_0a_1 & a_1^2 \\ a_0 - 5a_1 & -1.8 & a_0^2 & 2a_1 \\ a_1 & 1 & 0 & a_0^2 \end{bmatrix},$$

where we substituted $\boldsymbol{c} = [c_0,\, c_1,\, c_2,\, c_3]^\mathsf{T} = [-0.2,\, 1.05,\, -1.8,\, 1]^\mathsf{T}$ and $d = [d_0,\, d_1,\, d_2]^\mathsf{T} = [6,\, -5,\, 1]^\mathsf{T}$.

Working in the affine chart $a_1 = 1$, the intersection of the determinantal variety $\Sigma_{c,d}$ with this chart leads to the following (approximate) values for a_0: $\{0.5396, -2, -0.4234, -1.3962, -0.9464\}$. *These values (rounded for displaying purposes) were obtained using MacaulayLab [178]. We also compute the bad locus*

$$\mathcal{B} := \Big\{ a \in \mathbb{P} \mid \operatorname{rank} \underbrace{\begin{bmatrix} -0.2 & a_1^2 & 0 \\ 1.05 & 2a_0a_1 & a_1^2 \\ -1.8 & a_0^2 & 2a_1 \\ 1 & 0 & a_0^2 \end{bmatrix}}_{\boldsymbol{B}(a,c)} \leq 2 \Big\},$$

which, in the affine chart $a_1 = 1$, consists of the single point $a_0 = -2$.

The critical points of the model reduction problem correspond to the elements in $\Sigma_{c,d} \setminus \mathcal{B}$. Removing the bad locus leaves four critical points:

$$(a_0, a_1) = \{\, (0.5396, 1), (-0.4234, 1), (-1.3962, 1), (-0.9464, 1) \,\}.$$

Table 5.1, which lists the number of critical points for generic $(c,d) \in CD$, shows that there are 5 solutions for $(N,n) = (3,1)$. In our example, however, c fails the genericity condition because it contains a double root. Since $\Sigma_{c,d}$ is finite, we may apply Theorem 5.29 to adjust the generic count: we must subtract $\binom{N-2}{n-1} = \binom{1}{0} = 1$ to account for the multiplicity. This yields 4 critical points, exactly matching our numerical findings.

The fact that the bad locus consists of the single point $a_0 = -2$ is not surprising. Lemma 5.23 states that the nullity of $\boldsymbol{B}(a,c)$ is given by

$$\max \Big\{ 0,\ \deg \gcd(c, \widetilde{a}^2) + \min \big\{ N - \deg c - n,\ n - 2 \cdot \deg a \big\} \Big\}.$$

Here, $\deg c = 3$, so that when $\deg a = 0$ this expression evaluates to zero. Alternatively, if $\deg a = 1$, it becomes

$$\max \Big\{ 0,\ \deg \gcd(c, \widetilde{a}^2) - 1 \Big\},$$

showing that a bad locus occurs precisely when $\deg \gcd(c, \widetilde{a}^2) = 2$. This happens only when $\widetilde{a}(z) = (z - 1/2)$, which implies $a(z) = (z - 2)$.

Example 5.31. *Consider again the model from the previous example, but now compute the optimal second-order approximant ($n = 2$).*

The matrix $\boldsymbol{M}(a,c,d)$ is

$$\boldsymbol{M}(a,c,d) = \begin{bmatrix} 6a_0 & -0.2 & 0 & a_2^2 \\ 6a_1 - 5a_0 & 1.05 & -0.2 & 2a_1a_2 \\ a_0 - 5a_1 + 6a_2 & -1.8 & 1.05 & a_1^2 + 2a_0a_2 \\ a_1 - 0.5a_2 & 1 & -1.8 & 2a_0a_1 \\ a_2 & 0 & 1 & a_0^2 \end{bmatrix}.$$

Working in the affine chart $a_2 = 1$, the intersection of $\Sigma_{c,d}$ with this chart yields the approximate solutions (a_0, a_1, a_2):

$$\big\{(7.2098, -6.6056, 1), (-0.7122, -0.1950, 1), (0.6226, -0.9336, 1),$$
$$(0.6971, -1.1199, 1), (0.7646, -1.7265, 1), (2.5, -3.25, 1)\big\}.$$

All values (rounded for displaying purposes) were obtained using Macaulay-Lab [178]. The last point has multiplicity two with respect to the ideal of maximal minors of $\boldsymbol{M}(a, c, d)$, so that $\#\Sigma_{c,d} = 6$ and $\deg \Sigma_{c,d} = 7$. This last value coincides with the entry in Table 5.1 for $(N, n) = (3, 2)$.

It now remains to investigate the bad locus

$$\mathcal{B} \coloneqq \big\{a \in \mathbb{P}^2 \mid \operatorname{rank} \underbrace{\begin{bmatrix} -0.2 & 0 & a_2^2 \\ 1.05 & -0.2 & 2a_1a_2 \\ -1.8 & 1.05 & a_1^2+2a_0a_2 \\ 1 & -1.8 & 2a_0a_1 \\ 0 & 1 & a_0^2 \end{bmatrix}}_{\boldsymbol{B}(a,c)} \leq 2\big\}.$$

From Lemma 5.23, we know that the nullity of $\boldsymbol{B}(a, c)$ is given by

$$\max\Big\{0,\ \deg \gcd(c, \widetilde{a}^2) + \min\big\{-2,\ 2 - 2 \cdot \deg a\big\}\Big\}.$$

In both cases, $\deg a = 0$ and $\deg a = 1$, this expression evaluates to zero. Alternatively, if $\deg a = 2$, it becomes

$$\max\Big\{0,\ \deg \gcd(c, \widetilde{a}^2) - 2\Big\},$$

showing that a bad locus occurs precisely when $\deg \gcd(c, \widetilde{a}^2) > 2$. This is the case only when $\widetilde{a}(z) = (z - 1/2)(0.8)$, so that $a(z) = (z-2)(z-1/0.8) = 2.5 - 3.25z + z^2$. Numerical computation confirms that, in the affine chart $a_2 = 1$, the bad locus consists of the single point $a = [a_0,\, a_1,\, a_2]^\mathsf{T} = [2.5, -3.25, 1]^\mathsf{T}$. The point has multiplicity two with respect to the maximal minors of $\boldsymbol{B}(a, c)$, so that $\#\mathcal{B} = 1$ and $\deg \mathcal{B} = 2$. We can conclude that there are $\#\Sigma_{c,d} - \#\mathcal{B} = 5$ critical points.

Theorem 5.29 tells us to adjust the generic count, which predicts 7 critical points, by subtracting $\binom{N-2}{n-1} = \binom{1}{1} = 1$, giving an upper bound of 6 critical points. However, the point in the bad locus occurs with multiplicity two in the ideal of maximal minors of $\boldsymbol{M}(a, c, d)$, so that the inequality in Theorem 5.29 is not tight.

In the exceptional case where c is such that the bad locus is infinite, the determinantal variety $\Sigma_{c,d}$ is infinite as well (it is the union of the critical

points and the bad locus). By consequence, the determinantal variety cannot be computed directly using MEP solvers such as MacaulayLab [178], which rely on the assumption that the variety is zero-dimensional. To obtain the critical points in this situation, one must instead apply polynomial root-finding solvers directly to the critical equations (5.4).

5.7 Conclusions and Future Work

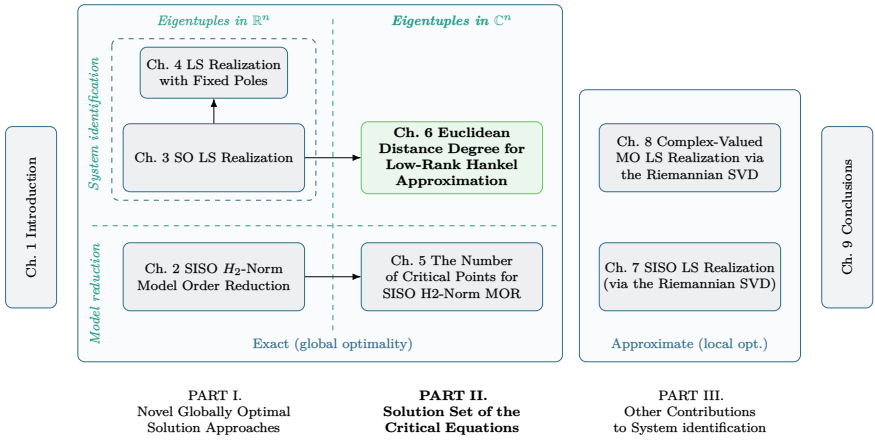
We have shown that the critical equations arising from the SISO $\mathcal{H}_2$-norm model reduction problem admit only finitely many complex solutions and, more importantly, we provided a closed-form formula that, for most higher-order models H, gives the exact number of critical points directly from the order of the original and reduced model. Because each genuine critical point (e.g., local minima) to the model reduction problem must correspond to a real-valued and stable solution to the critical equations, the count establishes an upper bound on the number of genuine critical points the optimization problem can have.

The proposed count has clear practical value. It indicates the intrinsic algebraic complexity that global solvers based on Gröbner bases, homotopy continuation, or specialized MEP algorithms must face, and it offers a benchmark against which the completeness of such solvers can be measured. Moreover, the formula shows that the number of complex solutions grows rapidly with problem size, theoretically confirming that the computational cost of globally optimal approaches grows quickly. These findings are in line with the experimental observations made in Part I of this dissertation. For the iterative algorithms commonly used in engineering practice, knowing the maximal number of critical points possibly helps to asses whether multiple random initializations are likely to discover different local optima or repeatedly converge to the same set of models.

Future work includes the extension of this algebro-geometric analysis to multiple-input multiple-output systems. A complementary goal is to refine the total count by estimating how many solutions are simultaneously real and stable, thereby tightening the bound on the number of true critical points. We also want to extend our analysis to critical points with higher-order poles, which possibly requires an alternative definition of the critical equations.

Chapter 6

The Euclidean Distance Degree for Low-Rank Hankel Matrix Approximation



Chapter abstract

We consider the structured low-rank approximation (SLRA) problem for Hankel matrices, or equivalently, the weighted least squares realization problem for autonomous linear time-invariant models. This problem can be viewed as a special case of a distance minimization problem between a generic point and a given algebraic variety. The number of complex solutions to the first-order optimality conditions (i.e., critical equations) is an invariant. The solutions naturally decompose into two disjoint components: the smooth locus, consisting of points where the associated Hankel matrix exactly satisfies the low-rank constraint, and the singular locus, where the rank drops further and the variety is singular.

We show that the subset of solutions on the smooth locus, the number of which is referred to as the Euclidean Distance (ED) degree of the variety, can be effectively characterized as the solution set to a multiparameter eigenvalue problem (MEP). By applying the Thom–Porteous formula, we recover the expression for the generic ED degree. We further demonstrate that the MEP solutions are generically reduced, a property that is advantageous from a computational perspective. For certain choices of weights, however, the MEP formulation introduces spurious solutions, which we refer to as the bad locus. We prove that this phenomenon occurs only for specific sets of weights. This analysis sheds new light on the so-called ED defect, which is the discrepancy between the ED degree for special weights and its generic value. In many cases, our results provide an algorithmic means to compute the ED degree of the Hankel variety for given weights.

In addition, we provide a count of the solutions to the critical equations lying in the singular locus, thereby complementing the ED degree. This result is particularly relevant for approaches that solve the critical equations directly and thus also encounter solutions in the singular locus. Together with the previously known count for the smooth locus (the ED degree), this leads to, for the first time, a complete enumeration of all complex solutions to the critical equations associated with the SLRA problem for Hankel matrices.

Candidate's contributions and relation to published work

The material presented in this chapter has not been published elsewhere. It is the result of joint work by Leonie Kayser, associated with the Numerical Algebraic Geometry group at the Max Planck Institute for Mathematics in the Sciences (MPI-MiS), Leipzig, and the candidate, with both authors contributing equally. This collaboration originated during a two-week research visit of the candidate to MPI-MiS in October 2024, at the invitation of Dr. Telen, leader of the Numerical Algebraic Geometry group.

The candidate's contributions to this work include formulating the research questions, contributing to the derivation of the theoretical results, and developing illustrative numerical experiments. While Leonie Kayser provided the algebro-geometric expertise, the author contributed his perspective on the underlying structure of the systems-theoretic problem from which the critical equations arise. In addition to validating the results through numerical computations, the candidate played an important role in interpreting the theoretical findings from a systems-theoretic perspective and situating them within the broader engineering literature. By carefully adapting the presentation of the results given in this chapter, he ensured that the results became accessible to a wider engineering audience and established clear connections to the earlier developments in Part I of this dissertation. An alternative exposition of these results can be found in Leonie Kayser's PhD dissertation [100].

6.1 Introduction

Similarly as in Section 3.7, we consider the structured low-rank approximation problem (SLRA) for Hankel matrices, or equivalently, the weighted least squares realization problem for autonomous linear time-invariant models.

6.1.1 Problem formulation

Let $\boldsymbol{y} \in \mathbb{R}^N$, and fix an integer $r \in \mathbb{N}$ such that $1 \leq r \leq (N-1)/2$. The Hankel matrix $\boldsymbol{H}_{N-r}^{\boldsymbol{y}}$ constructed from the sequence $\boldsymbol{y}$ is defined as:

$$\boldsymbol{H}_{N-r}^{\boldsymbol{y}} = \begin{bmatrix} y_0 & y_1 & y_2 & \cdots & y_r \\ y_1 & y_2 & y_3 & \cdots & y_{r+1} \\ \vdots & \vdots & \vdots & & \vdots \\ y_{N-r-1} & y_{N-r} & y_{N-r+1} & \cdots & y_{N-1} \end{bmatrix} \in \mathbb{R}^{(N-r)\times(r+1)}.$$

The set of *Hankel matrices of rank* $\leq r$ constitutes an algebraic variety:

$$\mathcal{X}_{N,r} = \left\{ \widehat{\boldsymbol{y}} \;\middle|\; \operatorname{rank} \boldsymbol{H}_{N-r}^{\widehat{\boldsymbol{y}}} \leq r \right\} \subseteq \mathbb{C}^N, \tag{6.1}$$

It is defined by the vanishing of all maximal minors of $\boldsymbol{H}_{N-r}^{\widehat{\boldsymbol{y}}}$.

Consequently, for some positive integer r and data $\boldsymbol{y} \in \mathbb{R}^N \setminus \mathcal{X}_{N,r}(\mathbb{R})$, the SLRA problem for Hankel matrices can be formulated as follows:

$$\underset{\widehat{\boldsymbol{y}} \in \mathbb{R}^N}{\text{minimize}} \; Q(\widehat{\boldsymbol{y}} - \boldsymbol{y}) \qquad \text{subject to } \widehat{\boldsymbol{y}} \in \mathcal{X}_{N,r}(\mathbb{R}), \tag{6.2}$$

where $Q \colon \mathbb{R}^N \to \mathbb{R} : \boldsymbol{x} \mapsto Q(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{x}$ is a quadratic form defined by a symmetric positive definite matrix

$$\boldsymbol{\Lambda} \in \operatorname{Sym}_{\mathbb{R}}(N) \coloneqq \left\{ \boldsymbol{M} \in \mathbb{R}^{N\times N} \mid \boldsymbol{M}^{\mathsf{T}} = \boldsymbol{M} \right\} \text{ and } \boldsymbol{\Lambda} \succ 0.$$

This form induces a weighted Euclidean norm on $\mathbb{R}^N$, with $Q(\boldsymbol{x}) = \frac{1}{2}\left\|\boldsymbol{x}\right\|_{\boldsymbol{\Lambda}}^2$.

We will refer to the first-order optimality conditions of (6.2) as the *critical equations*, the real solutions of which are the genuine critical points (e.g., local minima, saddle points) of the optimization problem (6.2). By a slight abuse of terminology, we will refer to *all complex solutions* of the critical equations as *critical points*.

These critical points can be partitioned into two disjoint components, based on the rank of the associated Hankel matrix. A subset of the solutions lies in the *smooth locus*

$$\mathcal{X}_{N,r}^{\text{sm}} \coloneqq \mathcal{X}_{N,r} \setminus \mathcal{X}_{N,r-1} \subseteq \mathbb{C}^N,$$

describing data $\widehat{\boldsymbol{y}}$ for which the corresponding Hankel matrix $\boldsymbol{H}_{N-r}^{\widehat{\boldsymbol{y}}}$ has rank exactly r. Its complement, the *singular locus*

$$\mathcal{X}_{N,r}^{\text{sing}} \coloneqq \mathcal{X}_{N,r-1} \subseteq \mathbb{C}^N,$$

contains data $\widehat{\boldsymbol{y}}$ for which the Hankel matrix has rank strictly less than r. It is the set of points at which the variety fails to be smooth.

The SLRA problem is an instance of a *Euclidean distance minimization problem*, which searches for the points on a given algebraic variety (cf. $\mathcal{X}_{N,r}$) that are closest, in the Euclidean metric, to a given external point in the ambient space (cf. $\boldsymbol{y}$). In this context, the number of complex critical points that lie inside the smooth locus of the given variety has been shown to depend only on the problem size and the chosen set of weights, defining the metric. This number is called the *Euclidean Distance (ED) degree* of the variety [51] and it has been studied for low-rank Hankel matrices in [102, 132, 133].

A formula for the ED degree of the variety $\mathcal{X}_{N,r}$ with respect to generic weights has been given in [132, Theorem 3.7]. Although their approach is effective, it draws on advanced machinery from intersection theory, placing the arguments somewhat out of reach for readers not already familiar with these tools[1]. By contrast, the remainder of this chapter relies only on basic notions of algebraic geometry, in particular those introduced in the previous chapter (cf. Section 5.3).

6.1.2 Contributions and outline

We show that for generic $\boldsymbol{y}$, the eigentuples of the multiparameter eigenvalue problem (MEP) formulation introduced in Chapter 3 characterize all complex critical points of the SLRA problem (6.2) lying inside the smooth locus. Following a strategy similar to Chapter 5, we apply the fiber dimension theorem and the Thom–Porteous formula to derive an exact count of the number of

[1] In their approach, Ottaviani and collaborators use the characterization of the generic ED degree as the sum of the *polar degrees* of the so-called *conormal variety* (essentially [51, Theorem 5.4]). The conormal variety is the geometric object that records, for each smooth point of the variety, all hyperplanes tangent to it. To obtain the polar degrees, they first desingularize the conormal variety and then compute the degrees as products of *Segre classes*, which may be viewed as intersection-theoretic invariants that measure how a subvariety sits inside a larger ambient space.

these solutions in the context of generic weights $\mathbf{\Lambda}$, thereby reproducing the formula for the generic ED degree described in [132, Theorem 3.7].

This count is particularly relevant in a computational setting, as solvers for polynomial systems and/or MEPs typically operate over the complex field $\mathbb{C}$. Additionally, our algebro-geometric analysis of the solution set of the MEP shows that, generically, the solutions are reduced, i.e., multiplicity-free. This property is computationally favorable, as multiplicities typically cause ill-conditioning and increased cost in numerical computations.

Compared to the analysis presented in [132], our framework provides novel geometric insights into the structure of the problem. Our approach is direct, providing an explicit genericity condition on the weights that must be satisfied to obtain the generic ED degree. Moreover, using the notion of a bad locus, we illustrate how the number of complex critical points may change under special choices of weights $\mathbf{\Lambda}$ and, in many cases, our framework gives an algorithmic way to compute the ED degree wrt. the Hankel variety for specific weights. Because our analysis uses only elementary tools from algebraic geometry, it provides a transparent and accessible perspective.

In addition to these results on the solutions to the critical equations in the smooth locus, a preliminary investigation of the singular locus $\mathcal{X}_{N,r}^{\text{sing}}$ further leads to an expression for the number of solutions inside the singular locus. No such result was previously available in the literature. While this analysis is exploratory, the count is relevant for approaches that solve the critical equations directly and therefore also compute the solutions inside the singular locus, e.g., via the Riemannian SVD formulation proposed in [38, 40]. When combined with the previously available count for the smooth locus (the ED degree), this provides, for the first time, a complete count of all solutions to the critical equations of the SLRA problem for Hankel matrices.

Outline We briefly recall the concept of the ED degree of an algebraic variety (Section 6.2). We then formulate the critical equations for the structured low-rank approximation problem in the Hankel setting (Section 6.3), followed by a reformulation into the *substituted* critical equations. This alternative formulation excludes solutions in the singular locus and thereby provides a direct means to study the ED degree (Section 6.4). The solution set of these equations is analyzed in Section 6.5, where we establish its finiteness, and in Section 6.6 we show that this set admits a determinantal representation: It is defined by the same MEP as encountered in Section 3.7. In Section 6.7, we apply the Thom–Porteous formula to obtain an explicit count of the number of complex critical points inside the smooth locus. Section 6.8 extends the analysis to the singular locus. Finally, Section 6.9 summarizes the results of

the chapter and outlines directions for future research. A schematic overview of the Euclidean distance minimization problem for the variety $\mathcal{X}_{N,r}$, together with the related concepts that will be introduced in Section 6.3, is given in Figure 6.1.

Notation We use similar conventions to Chapter 5. Let $P_{\leq k} \coloneqq \{f(z) = \sum_{i=0}^{k} f_i z^i \mid f_i \in \mathbb{C}\}$ denote the $(k+1)$-dimensional space of complex polynomials of degree at most k. For each polynomial $f \in P_{\leq k}$, we write its coefficient vector in boldface $\boldsymbol{f} = (f_0, \ldots, f_k)^{\mathsf{T}} \in \mathbb{C}^{k+1}$, and conversely, any $\boldsymbol{f} \in \mathbb{C}^{k+1}$ defines a unique polynomial $f \in P_{\leq k}$. For a given (polynomial) matrix $\boldsymbol{M} \in R_n^{p \times q}$, let $I_{r+1}(\boldsymbol{M})$ denote the ideal generated by the $(r+1)$-minors of $\boldsymbol{M}$.

6.2 Euclidean Distance Degree

This section reviews the Euclidean Distance (ED) degree. It is defined as the number of complex solutions to the first-order optimality conditions of the squared Euclidean distance function from a generic point to (the smooth locus of) a given algebraic variety. In computational practice, the ED degree provides an upper bound on the number of candidate solutions that must be determined by numerical solvers when approximating observed data with points on the variety. It therefore serves as a valuable indicator for the intrinsic computational complexity associated with such approximation problems, making it an important concept in numerical algebraic geometry. In what follows, we restrict attention to the most fundamental results and omit proofs. For a complete treatment of the general theory, the reader is referred to [51] or [19].

Let $\emptyset \neq X(\mathbb{R}) \subseteq \mathbb{R}^n$ be a real algebraic variety, $\boldsymbol{x} \in \mathbb{R}^n$ a given data point and fix a positive definite matrix $\boldsymbol{\Lambda} \in \mathrm{Sym}_{\mathbb{R}}(N)$. Let $d^2_{\boldsymbol{x}}(\widehat{\boldsymbol{x}}) = \boldsymbol{Q}(\boldsymbol{x} - \widehat{\boldsymbol{x}})$. We consider the following fundamental optimization problem:

$$\underset{\widehat{\boldsymbol{x}} \in X(\mathbb{R})}{\text{minimize}}\, d^2_{\boldsymbol{x}}(\widehat{\boldsymbol{x}}) = \|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|^2_{\boldsymbol{\Lambda}}\,.$$

If $X(\mathbb{R})$ is smooth, then every minimizer $\widehat{\boldsymbol{x}}^*$ satisfies the first-order optimality conditions for the restriction of $d^2_{\boldsymbol{x}}(\widehat{\boldsymbol{x}})$ to $X(\mathbb{R})$. Equivalently, the gradient

$$\nabla d^2_{\boldsymbol{x}}(\widehat{\boldsymbol{x}}^*) = 2\boldsymbol{\Lambda}(\boldsymbol{x} - \widehat{\boldsymbol{x}}^*)$$

is orthogonal to the tangent space $T(\widehat{\boldsymbol{x}}^*)X$. If $X(\mathbb{R})$ is not smooth, the tangent space computed from the Jacobian does not correctly represent the local geometry of X at singular points, such that a minimizer lying in the singular locus doesn't necessarily satisfy the standard condition. Therefore, in

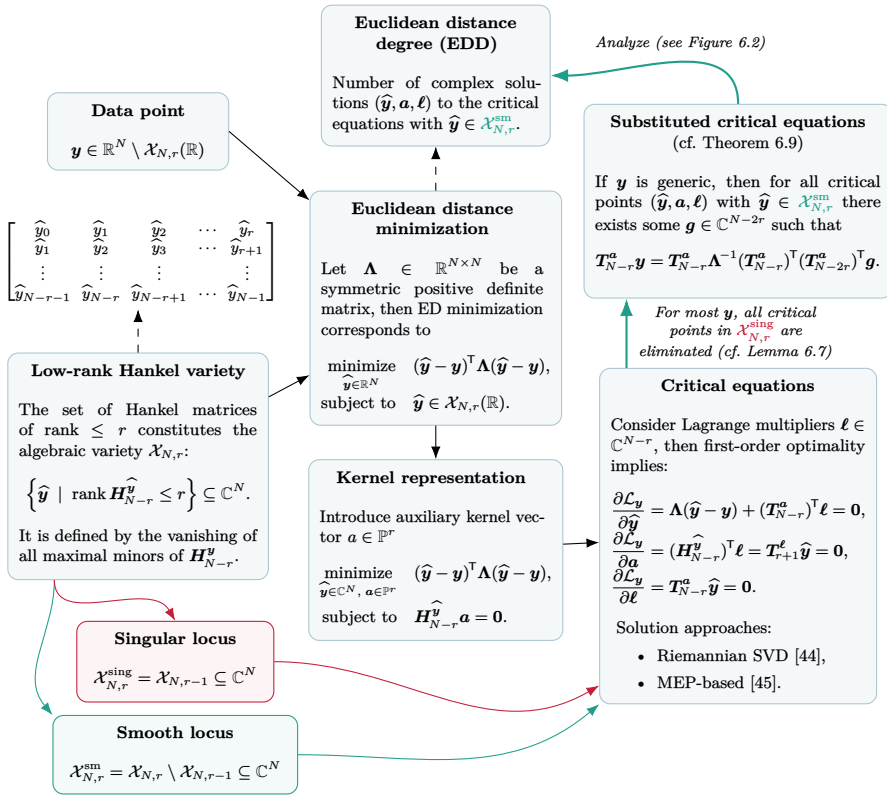


Figure 6.1: Schematic overview of the Euclidean distance minimization problem for the variety $\mathcal{X}_{N,r}$ of low-rank Hankel matrices, together with the related concepts introduced in Sections 6.1 and 6.3. The formulation of the optimization problem is "desingularized" by incorporating the kernel vector $\boldsymbol{a} \in \mathbb{P}^n$ as an auxiliary variable in the analysis. From this augmented setup we construct the corresponding first-order optimality conditions—the critical equations—whose solutions, the critical points, may lie in either the smooth or the singular locus of $\mathcal{X}_{N,r}$. We show that for most $\boldsymbol{y} \in \mathbb{C}^N$, rewriting the critical equations using the substitution $\boldsymbol{\ell} = (\boldsymbol{T}^{\boldsymbol{a}}_{N-2r})^{\mathsf{T}}\boldsymbol{g}$ removes all solutions in the singular locus. The resulting incidence over $(\boldsymbol{y}, a, g)$ provides a direct way to study the ED degree, defined as the number of critical points lying in the smooth locus of $\mathcal{X}_{N,r}$ (see Figure 6.2). All notation appearing here will be formally introduced in the corresponding sections; the figure is intended mainly as a navigational aid.

the general case, one considers a stratification of X into smooth components and requires criticality with respect to the stratum containing the point $\widehat{\boldsymbol{x}}^*$.

Suppose that X is defined in a neighborhood of $\widehat{\boldsymbol{x}}$ by the polynomials $f_1, \ldots, f_s \subset \mathbb{R}[x_1, \ldots, x_n]$. Then the tangent space $T(\widehat{\boldsymbol{x}})X$ at a nonsingular point $\widehat{\boldsymbol{x}} \in X^{\mathrm{sm}} = X \setminus X^{\mathrm{sing}}$ is given by

$$T(\widehat{\boldsymbol{x}})X = \left\{ \boldsymbol{v} \in \mathbb{C}^n \mid \nabla f_i(\widehat{\boldsymbol{x}})^{\mathsf{T}} \boldsymbol{v} = 0 \text{ for all } i = 0, \ldots, s \right\},$$

that is, the right null space of the Jacobian

$$\boldsymbol{J_f}(\widehat{\boldsymbol{x}}) = [\nabla f_1(\widehat{\boldsymbol{x}}), \ldots, \nabla f_s(\widehat{\boldsymbol{x}})]^{\mathsf{T}}.$$

It has dimension $\dim X$. Hence, the criticality condition can equivalently be written as

$$\boldsymbol{\Lambda}(\boldsymbol{x} - \widehat{\boldsymbol{x}}^*) \perp T(\widehat{\boldsymbol{x}}^*)X \iff \boldsymbol{\Lambda}(\boldsymbol{x} - \widehat{\boldsymbol{x}}) \in \operatorname{row}\left(\boldsymbol{J_f}(\widehat{\boldsymbol{x}})\right),$$

that is, the residual must lie the $\operatorname{codim}(X)$-dimensional row space of $\boldsymbol{J_f}(\widehat{\boldsymbol{x}})$.

This motivates the definition of the *ED correspondence* of X. It is the incidence variety of pairs $(\boldsymbol{x}, \widehat{\boldsymbol{x}})$ for which $\widehat{\boldsymbol{x}} \in X^{\mathrm{sm}}$ is a critical point wrt. the distance $d_{\boldsymbol{x}}^2$:

$$\mathcal{E}_X \coloneqq \overline{\left\{ (\boldsymbol{x}, \widehat{\boldsymbol{x}}) \in \mathbb{C}^n \times X^{\mathrm{sm}} \;\middle|\; \operatorname{rank} \begin{bmatrix} (\boldsymbol{\Lambda}(\widehat{\boldsymbol{x}} - \boldsymbol{x}))^{\mathsf{T}} \\ \boldsymbol{J_f}(\widehat{\boldsymbol{x}}) \end{bmatrix} \leq \operatorname{codim} X \right\}} \subseteq \mathbb{C}^n \times \mathbb{C}^n. \tag{6.3}$$

Lemma 6.1 ([51, Theorem 4.1])**.** *The restriction $\mathcal{E}_X^\circ = \mathcal{E}_X \cap (\mathbb{C}^n \times X^{\mathrm{sm}})$ is a smooth determinantal variety of dimension n, defined by the $(\operatorname{codim}(X) + 1)$-minors of the augmented Jacobian matrix appearing in (6.3). The projection*

$$\pi_{\widehat{\boldsymbol{x}}} \colon \mathcal{E}_X \to X \colon (\boldsymbol{x}, \widehat{\boldsymbol{x}}) \mapsto (\widehat{\boldsymbol{x}})$$

is an affine bundle over X^{sm}.

To find the critical points to $d_{\boldsymbol{x}}^2$ on X^{sm} means to compute the fiber of

$$\pi_{\boldsymbol{x}} \colon \mathcal{E}_X \to \mathbb{C}^n \colon (\boldsymbol{x}, \widehat{\boldsymbol{x}}) \mapsto (\boldsymbol{x})$$

and discarding those points that lie in the singular locus X^{sing}.

Lemma 6.2 ([51, Theorem 4.1])**.** *For a generic point $\boldsymbol{x} \in \mathbb{C}^n$ (or $\mathbb{R}^n$), the fiber $\pi_{\boldsymbol{x}}^{-1}$ is finite. If $X^{\mathrm{sm}}(\mathbb{R}) \neq 0$, then $\pi_{\boldsymbol{x}}$ is a dominant map, such that the fibers are non-empty.*

The *Euclidean distance degree* of a (complex) variety $X \subseteq \mathbb{C}^n$ with respect to a positive definite weight matrix $\mathbf{\Lambda}$, denoted $\mathrm{EDD}_{\mathbf{\Lambda}}(X)$, is defined as the number of points in the fiber $\pi_{\boldsymbol{x}}^{-1} \subseteq \mathcal{E}_X$ for generic $\boldsymbol{x} \in \mathbb{C}^n$. When the weighting matrix is the identity, we write simply $\mathrm{EDD}(X)$ instead of $\mathrm{EDD}_{\boldsymbol{I}_n}(X)$.

The *generic Euclidean distance degree* of a variety $\mathrm{EDD}_{\mathrm{gen}}(X)$ is the ED degree of X with respect to a generic choice of weight matrix $\mathbf{\Lambda}$. For a fixed $\mathbf{\Lambda}$, the *ED defect* measures the deviation of the weighted ED degree from the generic one, and is defined as:

$$\mathrm{EDDdef}_{\mathbf{\Lambda}}(X) = \mathrm{EDD}_{\mathrm{gen}}(X) - \mathrm{EDD}_{\mathbf{\Lambda}}(X) \geq 0.$$

6.3 The Critical Equations

The optimization problem (6.2) is precisely a nearest-point problem onto the Hankel variety $\mathcal{X}_{N,r}$ with respect to a weighted Euclidean norm. Consequently, it is associated with a well-defined ED degree. Specifically, the number of complex critical points $\widehat{\boldsymbol{y}} \in \mathcal{X}_{N,r}^{\mathrm{sm}}$ of problem (6.2) coincides with $\mathrm{EDD}_{\mathbf{\Lambda}}(\mathcal{X}_{N,r})$.

Reconsider $\mathcal{X}_{N,r}$, the variety of low-rank Hankel matrices introduced in (6.1). We summarize its most relevant properties.

Theorem 6.3. *Let $\mathcal{X}_{N,r}$ be the variety as defined in* (6.1)*, then:*

1. *$\mathcal{X}_{N,r}$ is irreducible, has dimension $2r$ and $\mathcal{X}_{N,r}^{\mathrm{sing}} = \mathcal{X}_{N,r-1}$.*
2. *The prime[2] ideal $I(\mathcal{X}_{N,r}) \subseteq \mathbb{C}[y_0, \ldots, y_d]$ is generated by the $(r+1)$-minors $I_{r+1}(\boldsymbol{H}_{N-r}^{\boldsymbol{y}})$.*
3. *The ideal of minors $I_{r+1}(\boldsymbol{H}_j^{\boldsymbol{y}})$ is independent of the choice $r+1 \leq j \leq N-r$ (so that $\boldsymbol{H}_j^{\boldsymbol{y}}$ has at least $r+1$ rows and columns).*

Statements (ii) and (iii) correspond to Theorem 1 and Proposition 5 in [185], respectively. Note that (iii) closely resembles Lemma 3.16. We will prove the first statement of Theorem 6.3, as the argument is instructive for understanding the structure of $\mathcal{X}_{N,r}$. The proof relies on two geometric concepts: affine bundles and birational morphisms, introduced in Section 5.3.2.

Proof of Theorem 6.3 (i). Our argument follows the presentation in [95, Theorem 1.45]. Note that for all $\widehat{\boldsymbol{y}} = [\widehat{y}_0, \ldots, \widehat{y}_{N-1}] \in \mathbb{C}^N$ and $\boldsymbol{a} = [a_0, \ldots, a_r] \in$

[2] A prime ideal cuts out a variety that is irreducible.

$\mathbb{P}^r$,

$$\boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r}\boldsymbol{a} = \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\widehat{\boldsymbol{y}}, \tag{6.4}$$

allowing us to use the Toeplitz matrix $\boldsymbol{T}^{\boldsymbol{a}}_{N-r}$ to enforce the low-rank structure on $\widehat{\boldsymbol{y}}$. Consider the incidence variety

$$\mathcal{Z} \coloneqq \left\{ (\boldsymbol{a}, \widehat{\boldsymbol{y}}) \;\middle|\; \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\widehat{\boldsymbol{y}} = 0 \right\} \subseteq \mathbb{P}^r \times \mathbb{C}^N.$$

The projection

$$\pi_{\boldsymbol{a}} \colon \mathcal{Z} \to \mathbb{P}^r \colon (\boldsymbol{a}, \widehat{\boldsymbol{y}}) \mapsto \boldsymbol{a}$$

gives $\mathcal{Z}$ the structure of a projective bundle over $\mathbb{P}^r$ with fibers isomorphic to $\operatorname{null}\left(\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\right) \subseteq \mathbb{C}^N$. Indeed, since $\boldsymbol{T}^{\boldsymbol{a}}_{N-r}$ has full row rank for all $\boldsymbol{a} \in \mathbb{P}^r$, its kernel is a linear subspace of constant dimension r. Hence, the fibers are isomorphic to $\mathbb{C}^r$. It follows that $\mathcal{Z}$ is smooth and irreducible of dimension $2r$.

The projection $\pi_{\widehat{\boldsymbol{y}}}(\mathcal{Z})$ coincides with the Hankel variety $\mathcal{X}_{N,r} \subseteq \mathbb{C}^N$. Moreover, the restricted map

$$\pi_{\widehat{\boldsymbol{y}}}|_{\mathcal{Z}} : \mathcal{Z} \to \mathcal{X}_{d,r} : (\boldsymbol{a}, \widehat{\boldsymbol{y}}) \mapsto \widehat{\boldsymbol{y}}$$

is a birational morphism: its inverse is defined on the open subset $\mathcal{X}^{\mathrm{sm}}_{N,r} = \mathcal{X}_{N,r} \setminus \mathcal{X}_{N,r-1}$ by

$$\widehat{\boldsymbol{y}} \mapsto (\operatorname{null}\left(\boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r}\right), \widehat{\boldsymbol{y}}).$$

This establishes that $\mathcal{X}_{N,r}$ has the same dimension as $\mathcal{Z}$ and is irreducible.

Since $\mathcal{X}_{N,r}$ is a determinantal variety, it is singular at those points where $\operatorname{rank} \boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r} < r$. Hence, its singular locus is exactly $\mathcal{X}_{N,r-1}$. This completes the proof of (i). □

The variety $\mathcal{X}_{N,r}$ can be described algebraically as the zero set of all $(r+1)$-minors of the Hankel matrix $\boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r}$. While this characterization is theoretically elegant, it is not particularly convenient: the number of such minors, each of which corresponds to a polynomial of degree exactly $r+1$, equals $\binom{N-r}{r+1}$ and thus grows quickly with N. In addition, $\mathcal{X}_{N,r}$ is not smooth, with singularities occurring at $\mathcal{X}_{N,r-1}$.

From both a geometric and computational perspective, $\mathcal{Z}$ is a useful representation of $\mathcal{X}_{N,r}$. By introducing the auxiliary variables $\boldsymbol{a} \in \mathbb{P}^r$, which reside in the null space of $\boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r}$, the singularities of $\mathcal{X}_{N,r}$ are "spread out" in such a way that the incidence $\mathcal{Z}$ becomes smooth. Additionally, $\mathcal{Z}$ comes with a simpler description, being defined by only $N-r$ quadratic equations (cf. (6.4)).

This equivalence between $\mathcal{X}_{N,r}$ and $\mathcal{Z}$ allows us to rephrase the constrained optimization problem (6.2) over the real locus of the smooth variety $\mathcal{Z}(\mathbb{R})$:

$$\underset{\substack{\widehat{\boldsymbol{y}}\in\mathbb{C}^N \\ \boldsymbol{a}\in\mathbb{P}^r}}{\text{minimize}}\ Q(\widehat{\boldsymbol{y}}-\boldsymbol{y}) \qquad \text{subject to } \boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r}\boldsymbol{a} = \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\widehat{\boldsymbol{y}} = \boldsymbol{0}. \tag{6.5}$$

The corresponding Lagrangian is

$$\mathcal{L}_{\boldsymbol{y}}(\widehat{\boldsymbol{y}},\boldsymbol{a},\boldsymbol{\ell}) = Q(\widehat{\boldsymbol{y}}-\boldsymbol{y}) + \boldsymbol{\ell}^{\mathsf{T}}\boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r}\boldsymbol{a} = Q(\widehat{\boldsymbol{y}}-\boldsymbol{y}) + \boldsymbol{\ell}^{\mathsf{T}}\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\widehat{\boldsymbol{y}},$$

where the unknown vector of Lagrange multipliers $\boldsymbol{\ell}$ resides in $\mathbb{C}^{N-r}$. By differentiating $\mathcal{L}_{\boldsymbol{y}}$ with respect to the variables, we obtain a system of polynomial equations characterizing the critical points of (6.5).

Definition 6.4. *The* critical equations *of the reparametrized SLRA problem* (6.5) *correspond to:*

$$\frac{\partial\mathcal{L}_{\boldsymbol{y}}}{\partial\widehat{\boldsymbol{y}}} = \boldsymbol{\Lambda}(\widehat{\boldsymbol{y}}-\boldsymbol{y}) + (\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}\boldsymbol{\ell} = \boldsymbol{0}, \tag{6.6a}$$

$$\frac{\partial\mathcal{L}_{\boldsymbol{y}}}{\partial\boldsymbol{a}} = (\boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r})^{\mathsf{T}}\boldsymbol{\ell} = \boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{r+1}\boldsymbol{\ell} = \boldsymbol{T}^{\boldsymbol{\ell}}_{r+1}\widehat{\boldsymbol{y}} = \boldsymbol{0}, \tag{6.6b}$$

$$\frac{\partial\mathcal{L}_{\boldsymbol{y}}}{\partial\boldsymbol{\ell}} = \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\widehat{\boldsymbol{y}} = \boldsymbol{0}. \tag{6.6c}$$

Projecting the complex solution set $(\widehat{\boldsymbol{y}},\boldsymbol{a},\boldsymbol{\ell})$ of these critical equations onto the first factor, i.e., $(\widehat{\boldsymbol{y}},\boldsymbol{a},\boldsymbol{\ell}) \mapsto \widehat{\boldsymbol{y}}$, associates each solution with a candidate critical point $\widehat{\boldsymbol{y}}$ of the original problem (6.2).

When restricted to $\widehat{\boldsymbol{y}} \in \mathcal{X}^{\mathrm{sm}}_{N,r}$, this projection precisely identifies the critical points counted by the ED degree for the Hankel variety $\mathcal{X}_{N,r}$. However, the critical equations also admit solutions with $\widehat{\boldsymbol{y}} \in \mathcal{X}^{\mathrm{sing}}_{N,r}$. These additional solutions do not contribute to the ED degree, and therefore, the system as formulated above has strictly more solutions than $\mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r})$.

We restate Lemma 3.22 in the present notation.

Lemma 6.5. *Let* $\boldsymbol{y} \in \mathbb{C}^N \setminus \mathcal{X}_{N,r}(\mathbb{R})$ *and* $\widehat{\boldsymbol{y}} \in \mathcal{X}_{N,r-1}(\mathbb{R})$. *Then* $\widehat{\boldsymbol{y}}$ *is* not *a (local) minimizer of the function* $\boldsymbol{x} \mapsto Q(\boldsymbol{y}-\boldsymbol{x})$ *on* $\mathcal{X}_{N,r}(\mathbb{R})$.

This result indicates that, also from the real optimization point of view, solutions in $\mathcal{X}^{\mathrm{sing}}_{N,r}$ are not relevant for identifying the (global) minimizers of the SLRA problem (6.2).

6.4 Removal of the Singular Locus

In this section, we present a reformulation of the critical equations (6.6a)-(6.6c). We illustrate that for a generic data point $\boldsymbol{y} \in \mathbb{C}^N$, the proposed reformulation eliminates solutions lying in the singular locus of the Hankel variety such that the solution set consists solely of critical points in the smooth locus $\mathcal{X}_{N,r}^{\text{sm}}$. The alternative formulation therefore provides an elegant framework for studying the ED degree of the Hankel variety, $\text{EDdeg}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r})$.

Rearranging (6.6a) gives $\widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{\Lambda}^{-1}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}}\boldsymbol{\ell}$. In Section 3.7.3, where we restricted attention to real-valued solutions, we concluded that $\widehat{\boldsymbol{y}}$ is therefore determined entirely by $\boldsymbol{a}$ and $\boldsymbol{\ell}$. The situation here differs, however, since the matrix $\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}}$ is not guaranteed to be invertible for all complex $\boldsymbol{a} \in \mathbb{P}^r$.

Lemma 6.6. *Assuming the matrix $\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}}$ is invertible, then*

$$\boldsymbol{\ell} = (\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}})^{-1}\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{y}$$

$$\widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{\Lambda}^{-1}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}})^{-1}\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{y}.$$

This assumption is satisfied for real, nonzero $\boldsymbol{a}$ and $\boldsymbol{\Lambda}$ positive definite. In that case, $\widehat{\boldsymbol{y}}$ is the $\boldsymbol{\Lambda}$-orthogonal projection of $\boldsymbol{y}$ onto $\text{null}\left(\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\right)$.

Proof. We can substitute (6.6a) into (6.6c) to obtain

$$\boldsymbol{0} = \boldsymbol{T}_{N-r}^{\boldsymbol{a}}\widehat{\boldsymbol{y}} = \boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{y} + \boldsymbol{T}_{N-r}^{\boldsymbol{a}}(\widehat{\boldsymbol{y}} - \boldsymbol{y}) = \boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{y} - \boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}}\boldsymbol{\ell}.$$

Multiplying with $(\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}})^{-1}$ leads to the expression for $\boldsymbol{\ell}$. The second is obtained by substituting the first into (6.6a).

If $\boldsymbol{a}$ is real, then $\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}}$ is positive definite (hence invertible), since $(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}}$ has full column-rank for any $\boldsymbol{a} \neq \boldsymbol{0}$. Lastly, the formula for $\widehat{\boldsymbol{y}}$ is exactly the formula of $\boldsymbol{\Lambda}$-orthogonal projection onto the null space of $\boldsymbol{T}_{N-r}^{\boldsymbol{a}}$. □

A straightforward but unsatisfactory strategy would be to apply Lemma 6.6 in an attempt to eliminate $\widehat{\boldsymbol{y}}$ and $\boldsymbol{\ell}$ from the critical equations. This approach would introduce the inverse of the matrix $\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}})^{\mathsf{T}}$, thereby leaving the purely polynomial setting we wish to maintain. Matrix-inversions are also unfavorable from a computational point of view. The preferred way forward builds on the insights developed in Chapter 3.

For generic data $\boldsymbol{y}$, solutions $(\widehat{\boldsymbol{y}}, \boldsymbol{a}, \boldsymbol{\ell})$ of the critical equations (6.6a)-(6.6c) for which $\widehat{\boldsymbol{y}} \in \mathcal{X}_{N,r}^{\text{sing}}$ can be detected by the Lagrange vector $\boldsymbol{\ell}$, as follows.

Lemma 6.7. *Assume* $(\widehat{\boldsymbol{y}}, \boldsymbol{a}, \boldsymbol{\ell})$ *satisfy* (6.6a)-(6.6c) *and consider the polynomials* $a \in P_{\leq r} \setminus \{0\}$ *and* $\ell \in P_{\leq N-r-1}$ *corresponding to the coefficient vectors* $\boldsymbol{a}$ *and* $\boldsymbol{\ell}$*, respectively. Then,*

1. *If* $\widehat{\boldsymbol{y}} \in \mathcal{X}^{\mathrm{sm}}_{N,r}$*, i.e.,* $\operatorname{rank} \boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r} = r$*, then* ℓ *is a multiple of* a*;*
2. *Conversely, if* ℓ *is a multiple of* a *and the data vector* $\boldsymbol{y} \in \mathbb{C}^N$ *is sufficiently generic, then* $\widehat{\boldsymbol{y}} \in \mathcal{X}^{\mathrm{sm}}_{N,r}$*.*

Proof. When $\operatorname{rank} \boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r} = r$, the rows of $\boldsymbol{T}^{\boldsymbol{a}}_{N-2r} \in \mathbb{C}^{(N-2r)\times(N-r)}$ span a basis for the $(N-2r)$-dimensional left null space of $\boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r}$. Combined with (6.6b), this illustrates that $\boldsymbol{\ell} = (\boldsymbol{T}^{\boldsymbol{a}}_{N-2r})^{\mathsf{T}} \boldsymbol{g}$ for some $\boldsymbol{g} \in \mathbb{C}^{N-2r}$. This proves the first statement.

For the converse we need some genericity of $\boldsymbol{y}$. Consider the incidence variety

$$\begin{aligned}\mathcal{I} = \{\, (\boldsymbol{y}, \widehat{\boldsymbol{y}}, a, g) \mid \ell = a \cdot g \text{ and (6.6a)-(6.6c) hold} \,\} \\ \subseteq \mathbb{C}^N \times \mathbb{C}^N \times (P_{\leq r} \setminus 0) \times P_{\leq N-2r-1}.\end{aligned}$$

The projection of $\mathcal{I}$ onto $(\widehat{\boldsymbol{y}}, a, g)$ gives the variety (note that $\ell = a \cdot g$ implies that (6.6b) and (6.6a) can be satisfied for any $(\widehat{\boldsymbol{y}}, a, g)$)

$$\begin{aligned}\mathcal{P} &= \left\{ (\widehat{\boldsymbol{y}}, a, g) \;\middle|\; 0 \neq \boldsymbol{a} \in \operatorname{null}\left(\boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r}\right),\ g \in P_{\leq N-2r-1} \right\} \\ &= \bigcup_{i=1}^{r} \underbrace{\left\{ (\widehat{\boldsymbol{y}}, a, g) \;\middle|\; \widehat{\boldsymbol{y}} \in \widehat{\mathcal{X}}_{N,i} \setminus \widehat{\mathcal{X}}_{N,i-1},\ \boldsymbol{a} \in \operatorname{null}\left(\boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r}\right),\ g \in P_{\leq N-2r-1} \right\}}_{\dim = 2i+(r-i+1)+(N-2r) = N+1+i-r}.\end{aligned}$$

This shows that $\dim \mathcal{P} = N+1$. Since inside $\mathcal{I}$, each $(\widehat{\boldsymbol{y}}, a, g) \in \mathcal{P}$ determines a unique vector $\boldsymbol{y}$ via (6.6a), we have $\dim \mathcal{I} = dim \mathcal{P}$. Moreover, the stratification of $\mathcal{P}$ further shows that the set of points inside $\mathcal{X}^{\mathrm{sing}}_{N,r}$, i.e., for which $\operatorname{rank} \boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r} \leq r-1$, has dimension $\leq N$.

The projection $\mathcal{I} \to \mathbb{C}^N$ onto the $\boldsymbol{y}$-component is dominant, since $\operatorname{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r}) > 0$ by Lemma 6.2. Its generic fibers are one-dimensional in $\mathcal{X}^{\mathrm{sm}}_{N,r}$ and finite in the locus $\mathcal{X}^{\mathrm{sing}}_{N,r}$. But since $a \neq 0$, any solution to (6.6a)-(6.6c) gives a 1-dimensional family, replacing a, ℓ with $\lambda a, \frac{1}{\lambda}\ell$. In particular, for generic $\boldsymbol{y}$ there is no solution with $\operatorname{rank} \boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r} < r$ if $\ell = a \cdot g$. □

Thus, for generic $\boldsymbol{y} \in \mathbb{C}^N$, substitution of $\ell = a \cdot g$ for $g \in P_{\leq d-2r}$ removes the lower-rank solutions from the solution set of the critical equations. This transforms the critical equations as follows:

$$\mathbf{0} = \boldsymbol{\Lambda}(\widehat{\boldsymbol{y}} - \boldsymbol{y}) + (\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2r})^{\mathsf{T}}\boldsymbol{g} \tag{6.7a}$$

$$\mathbf{0} = \boldsymbol{T}^{\boldsymbol{g}}_{r+1}\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\widehat{\boldsymbol{y}} \tag{6.7b}$$

$$\mathbf{0} = \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\widehat{\boldsymbol{y}}. \tag{6.7c}$$

In this formulation, (6.7b) is a direct consequence of (6.7c), so we may omit it. We can rearrange (6.7a) to $\widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2r})^{\mathsf{T}}\boldsymbol{g}$ and substitute it into (6.7c) to arrive at the reformulation also obtained in Section 3.7.

Definition 6.8. *We refer to the following system as the* substituted critical equations *of the reparametrized SLRA problem* (6.5) *correspond to*

$$\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{y} = \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2r})^{\mathsf{T}}\boldsymbol{g}. \tag{6.8}$$

The solution set to the substituted critical equations excludes solutions arising from the singular locus $\mathcal{X}^{\text{sing}}_{N,r}$. This is formalized as follows.

Theorem 6.9. *For given generic $\boldsymbol{y} \in \mathbb{C}^N$, the complex solutions $(\widehat{\boldsymbol{y}}, \boldsymbol{a}, \boldsymbol{\ell})$ of the critical equations* (6.6a)-(6.6c) *for which* $\operatorname{rank} \boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r} = r$ *are in bijection to the solutions $(\boldsymbol{a}, \boldsymbol{g})$ of the system in* (6.8)*. The bijection is given by $\ell = a \cdot g$ and $\widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2r})^{\mathsf{T}}\boldsymbol{g}$. If $\boldsymbol{y}$ is real, then this bijection preserves real solutions.*

In particular, the number of complex solutions to (6.8) *equals* $\operatorname{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r})$.

Proof. The statements are immediate from Lemma 6.7, the derivations above, and the definition of the ED degree (cf. Section 6.2). Indeed, for each (a, g) satisfying (6.8), the bijection $\ell = a \cdot g$ and $\widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2r})^{\mathsf{T}}\boldsymbol{g}$ defines a unique tuple $(\widehat{\boldsymbol{y}}, \boldsymbol{a}, \boldsymbol{\ell})$ satisfying the critical equations (6.6a)-(6.6c). For generic $\boldsymbol{y}$, it follows from Lemma 6.7 that these $\widehat{\boldsymbol{y}}$ reside in the smooth locus $\mathcal{X}^{\text{sm}}_{N,r}$, i.e., $\operatorname{rank} \boldsymbol{H}^{\widehat{\boldsymbol{y}}}_{N-r} = r$. □

By Theorem 6.9, the system of equations (6.8) provides a particularly convenient formulation for studying $\operatorname{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r})$.

6.5 The Substituted ED Correspondence

In the standard ED correspondence (cf. Section 6.2), one considers pairs of data $\boldsymbol{y}$ and the corresponding critical points $\widehat{\boldsymbol{y}}$. In this section, we make the connection between the substituted critical equations (6.8) and the ED correspondence precise.

The logical flow of our approach is summarized in Figure 6.2, which maps the dependencies between the arguments developed in Sections 6.5-6.7. All notation appearing in the figure will be introduced in the appropriate sections. The purpose of the figure is to serve as a navigational guide to the structure of the main argument.

We will deal with several products of vector spaces. To ease notation, we use the shorthand

$$A \coloneqq P_{\leq n} \quad \text{and } G \coloneqq P_{\leq N-2r-1}.$$

Additionally, $A^\circ \coloneqq A \setminus \{0\}$, and we use, for example, $A^\circ G$ to denote the product $(A \setminus \{0\}) \times G$.

Define the matrix

$$\boldsymbol{B}_{\boldsymbol{\Lambda}}(a) \coloneqq \boldsymbol{T}^{\boldsymbol{a}}_{N-r} \boldsymbol{\Lambda}^{-1} (\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}} (\boldsymbol{T}^{\boldsymbol{a}}_{N-2r})^{\mathsf{T}}$$

and consider the incidence variety corresponding to the substituted critical equations:

$$\mathcal{Y}_{\boldsymbol{\Lambda}} \coloneqq \left\{ \, (\boldsymbol{y}, a, g) \;\middle|\; \boldsymbol{T}^{\boldsymbol{a}}_{N-r} \boldsymbol{y} = \boldsymbol{B}_{\boldsymbol{\Lambda}}(a)\, \boldsymbol{g} \, \right\} \subseteq \mathbb{C}^N \times A^\circ G.$$

Recall the notion of an "ideal-theoretic complete intersection" from Section 5.3.2. This allows to state the following result concerning $\mathcal{Y}_{\boldsymbol{\Lambda}}$.

Lemma 6.10. *$\mathcal{Y}_{\boldsymbol{\Lambda}}$ is an irreducible ideal-theoretic complete intersection in the ambient space $A^\circ G$ of dimension $N+1$ and codimension $N-r$.*

Proof. The projection of $\mathcal{Y}_{\boldsymbol{\Lambda}}$ onto (a, g) yields the open set

$$\mathcal{V} = \{ \, (a, g) \mid a \in A^\circ, \; g \in G \, \}.$$

This turns $\mathcal{Y}_{\boldsymbol{\Lambda}}$ into a rank r affine bundle over $\mathcal{V}$, as any choice $(a, g) \in \mathcal{V}$ determines $\boldsymbol{y}$ as a solution to the affine linear system $\boldsymbol{T}^{\boldsymbol{a}}_{N-r} \boldsymbol{y} = \boldsymbol{B}_{\boldsymbol{\Lambda}}(a)\, \boldsymbol{g}$. Consequently, $\mathcal{Y}_{\boldsymbol{\Lambda}}$ is a smooth variety of dimension

$$\dim A^\circ + \dim G + r = (r+1) + (N-2r) + r = N+1.$$

Moreover, the system has $N-r$ equations in an ambient space of dimension $(N) + (r+1) + (N-2r) = 2N-r+1$, such that $\mathcal{Y}_{\boldsymbol{\Lambda}}$ is globally defined by codimension many equations. □

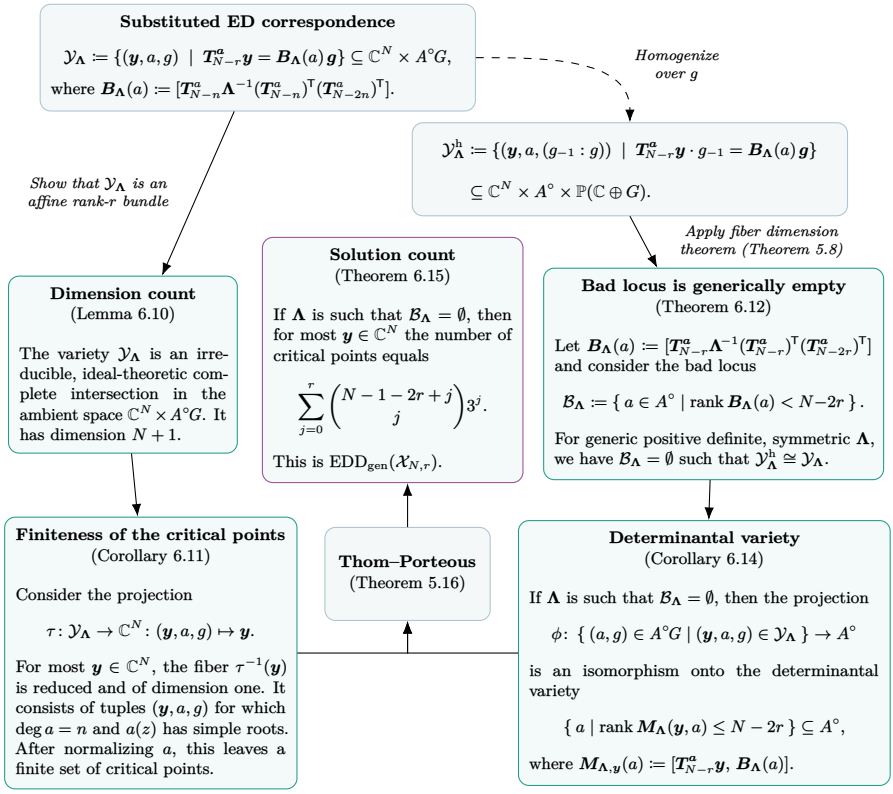


Figure 6.2: Schematic overview of the logical structure and dependencies between the arguments developed in Sections 6.5–6.7. The diagram illustrates how the analysis of the incidence variety $\mathcal{Y}_{\boldsymbol{\Lambda}}$ and the corresponding interpretation as a determinantal variety lead to a formula for the generic ED degree. All notation appearing here will be formally introduced in the corresponding sections; the figure is intended mainly as a navigational aid which allows the reader to return to it at any stage to see how each step fits into the overall argument.

Corollary 6.11 (Finiteness of the critical points)**.** *Consider the projection*

$$\tau\colon \mathcal{Y}_{\mathbf{\Lambda}} \to \mathbb{C}^N\colon (\boldsymbol{y}, a, g) \mapsto \boldsymbol{y}.$$

For almost all $\boldsymbol{y} \in \mathbb{C}^N$*, the fiber* $\tau^{-1}(\boldsymbol{y})$ *is reduced, of dimension* 1*. In particular, it consists of tuples* $(\boldsymbol{y}, a, g)$ *satisfying*

$$\deg a = r \quad \textit{and} \quad a(z) \textit{ has simple roots.} \tag{6.9}$$

Equivalently, for generic $\boldsymbol{y} \in \mathbb{C}^N$*, the set of complex solutions* (a, g) *to* (6.8) *is finite up to normalization of* a *(e.g., fixing* $a_r = 1$*) and reduced (i.e., it consists of simple solutions without multiplicities).*

Proof. Dominance of the projection $\tau\colon \mathcal{Y}_{\mathbf{\Lambda}} \to \mathbb{C}^N$ follows directly from the general ED framework [51] (cf. Lemma 6.2), so that $\overline{\tau(\mathcal{Y}_{\mathbf{\Lambda}})} = \mathbb{C}^N$. Hence, by the fiber dimension theorem (cf. Theorem 5.8) we have for most $\boldsymbol{y} \in \mathbb{C}^N$:

$$\dim \tau^{-1}(\boldsymbol{y}) = \dim \mathcal{Y}_{\mathbf{\Lambda}} - \dim \mathbb{C}^N = (N+1) - N = 1.$$

To prove the additional conditions in (6.9), note that the negation of each one is given by a polynomial equation in a:

$$a_r = 0 \quad \text{and} \quad \Delta(a) = 0, \tag{6.10}$$

respectively, where $\Delta(a)$ denotes the discriminant of a, that is, the polynomial in the coefficients of a that vanishes if a has multiple roots.

We show that the conditions in (6.10) are non-trivial on the irreducible variety $\mathcal{Y}_{\mathbf{\Lambda}}$ by constructing a point in $\mathcal{Y}_{\mathbf{\Lambda}}$ that satisfies none of them. Let a be a generic polynomial of degree n, so that $a_n \neq 0$ and $\Delta(a) \neq 0$. Then, via (6.8), every choice of $g \in G$ determines an n-dimensional linear subspace of $\boldsymbol{y}$'s for which $(\boldsymbol{y}, a, g) \in \mathcal{Y}_{\mathbf{\Lambda}}$.

Since $\mathcal{Y}_{\mathbf{\Lambda}}$ is irreducible and none of the conditions in (6.10) holds identically on $\mathcal{Y}_{\mathbf{\Lambda}}$, it follows that

$$\mathcal{W} \coloneqq \mathcal{Y}_{\mathbf{\Lambda}} \cap (\{a_n = 0\} \cap \{\Delta(a) = 0\})$$

is a subvariety of $\mathcal{Y}_{\mathbf{\Lambda}}$; hence of dimension at most N. Consider the morphism $\pi\colon \mathcal{W} \to \mathbb{C}^N\colon (\boldsymbol{y}, a, g) \mapsto \boldsymbol{y}$. Since the solutions to the defining equations of $\mathcal{W}$ are invariant wrt. the transformation $(\boldsymbol{y}, a, g) \to (\boldsymbol{y}, \lambda a, \frac{1}{\lambda^2} g)$ for any $\lambda \in \mathbb{C} \setminus \{0\}$, all non-empty fibers $\pi^{-1}(\boldsymbol{y})$ have dimension at least one. Via Theorem 5.8 this implies that the image $\pi(\mathcal{W})$ has dimension strictly smaller than $\dim \mathcal{W}$, hence strictly smaller than the dimension of $\mathbb{C}^N$, the ambient space dimension

of $\boldsymbol{y}$. We can conclude that for a generic choice of $\boldsymbol{y}$, the fiber $\tau^{-1}(\boldsymbol{y})$ will not intersect $\mathcal{W}$ and thus satisfy the conditions in (6.9).

Finally, setting $a_r = 1$ reduces the dimension of $\tau^{-1}(\boldsymbol{y})$ to 0. This completes the proof. □

Notice the close parallel between, on the one hand, Theorem 5.19 and Corollary 5.21, which establish finiteness of the complex solution set of the critical equations for the SISO $\mathcal{H}_2$-norm model reduction problem, and, on the other hand, Lemma 6.10 together with Corollary 6.11, which state an analogous finiteness result for the solutions of the substituted critical equations (6.8).

By the bijection established in Theorem 6.9, each pair $(a, g) \in \tau^{-1}(\boldsymbol{y})$ corresponds to a unique solution $(\widehat{\boldsymbol{y}}, \boldsymbol{a}, \boldsymbol{\ell})$ of the critical equations (6.6a)–(6.6c), with $\widehat{\boldsymbol{y}} \in \mathcal{X}^{\mathrm{sm}}_{N,r}$. Corollary 6.11 ensures that, for generic $\boldsymbol{y} \in \mathbb{C}^N$, these fibers consist of finitely many reduced points (up to normalization of a). This conclusion is in full agreement with the generic finiteness statement for the ED degree given in Lemma 6.2.

6.6 Multiparameter Eigenvalue Problem

Our approach in the remaining sections follows a similar strategy to the one developed in Sections 5.5-5.6, adapted to the setting of the SLRA problem (6.2).

To make the substituted critical equations (6.8) linear in $g \in G$, we homogenize with the new variable $g_{-1} \in \mathbb{C}$. This leads to the quasi-projective incidence variety:

$$\mathcal{Y}^{\mathrm{h}}_{\boldsymbol{\Lambda}} \coloneqq \left\{ (\boldsymbol{y}, a, (g_{-1} : g)) \;\middle|\; \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{y} \cdot g_{-1} = \boldsymbol{B}_{\boldsymbol{\Lambda}}(a)\, \boldsymbol{g} \right\} \subseteq \mathbb{C}^N \times A^\circ \times \mathbb{P}(\mathbb{C} \oplus G). \tag{6.11}$$

which can be interpreted as a partial compactification of the affine incidence variety $\mathcal{Y}_{\boldsymbol{\Lambda}}$. Indeed, the affine chart $g_{-1} = 1$ coincides with $\mathcal{Y}_{\boldsymbol{\Lambda}}$ and the locus $\mathcal{Y}^{\infty}_{\boldsymbol{\Lambda}} \coloneqq \mathcal{Y}_{\boldsymbol{\Lambda}} \cap \{g_{-1} = 0\}$ is added at infinity along the direction of g, so that $\mathcal{Y}^{\mathrm{h}}_{\boldsymbol{\Lambda}}$ is the disjoint union $\mathcal{Y}^{\mathrm{h}}_{\boldsymbol{\Lambda}} = \mathcal{Y}_{\boldsymbol{\Lambda}} \cup \mathcal{Y}^{\infty}_{\boldsymbol{\Lambda}}$.

Define the matrix

$$\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{y}, a) \coloneqq \begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{y} & \boldsymbol{B}_{\boldsymbol{\Lambda}}(a) \end{bmatrix}.$$

With this notation, the defining equations of $\mathcal{Y}^{\mathrm{h}}_{\boldsymbol{\Lambda}}$ can be compactly written as

$$\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{y}, a) \cdot \begin{bmatrix} g_{-1} \\ \boldsymbol{g} \end{bmatrix} = \boldsymbol{0}. \tag{6.12}$$

In Chapter 5, it was possible to prove that the spurious component $\mathcal{Y}_{\boldsymbol{\Lambda}}^{\infty}$ vanishes for generic problem instances (cf. Lemma 5.23). Here, the existence of $\mathcal{Y}_{\boldsymbol{\Lambda}}^{\infty}$ does not depend on the particular data vector $\boldsymbol{y}$. Rather, its appearance is determined by the choice of weight matrix $\boldsymbol{\Lambda}$.

To formalize this, we define the *bad locus*

$$\mathcal{B}_{\boldsymbol{\Lambda}} \coloneqq \{\, a \mid \operatorname{rank} \boldsymbol{B}_{\boldsymbol{\Lambda}}(a) < N - 2r \,\} \subseteq A^{\circ},$$

and its complement $\mathcal{G}_{\boldsymbol{\Lambda}} = A^{\circ} \backslash \mathcal{B}_{\boldsymbol{\Lambda}}$, which we refer to as the *good locus*. Remember that the appearance of such a "bad locus" and its implications have been visualized schematically in Figure 1.6.

From (6.12), it follows that the spurious component $\mathcal{Y}_{\boldsymbol{\Lambda}}^{\infty}$ added by projectivizing $\mathcal{Y}_{\boldsymbol{\Lambda}}$ is empty over $\mathcal{G}_{\boldsymbol{\Lambda}}$, i.e., $\mathcal{Y}_{\boldsymbol{\Lambda}}^{\infty} \cup \{a \in \mathcal{G}_{\boldsymbol{\Lambda}}\} = \emptyset$. Indeed, if a is such that the matrix $\boldsymbol{B}_{\boldsymbol{\Lambda}}(a)$ has full column rank, the homogenized system of equations has no solutions with $g_{-1} = 0$.

One might wonder whether the good locus could be empty—equivalently, whether the bad locus could coincide with all of A°, meaning that the $\boldsymbol{B}_{\boldsymbol{\Lambda}}(a)$ fails to have full column rank for every $a \in A^{\circ}$. However, the proof of Lemma 3.26 shows that when the weight matrix $\boldsymbol{\Lambda}$ is positive definite, the matrix $\boldsymbol{B}_{\boldsymbol{\Lambda}}(a)$ has full column rank for all real $a \in A^{\circ}$. Under this positive definiteness assumption, which we imposed as the outset of the chapter, the good locus necessarily contains all such real points and is therefore nonempty: $\mathcal{G}_{\boldsymbol{\Lambda}} \neq \emptyset$.

The following result goes further: it shows that, for almost all weights, the bad locus is in fact empty.

Theorem 6.12. *The subset of positive definite $\boldsymbol{\Lambda} \in \operatorname{Sym}(N)$ for which the bad locus is empty, i.e., $\mathcal{B}_{\boldsymbol{\Lambda}} = \emptyset$, is a dense open set.*

To be able to prove Theorem 6.12, we need a small linear algebra lemma.

Lemma 6.13. *Pick $\boldsymbol{v} \in \mathbb{P}(\mathbb{C}^N)$ and let $W \subseteq \mathbb{C}^N$ be an r-dimensional linear subspace of $\mathbb{C}^N$. Then the set $\{\, \boldsymbol{\Lambda} \in \operatorname{Sym}(N) \mid \boldsymbol{\Lambda}\boldsymbol{v} \in W \,\}$ is a linear subspace of dimension $r + \binom{N}{2}$.*

Proof. If $\boldsymbol{v} = \boldsymbol{e_1} \coloneqq [1, 0, \ldots, 0]^{\mathsf{T}}$, then the condition $\boldsymbol{\Lambda}\boldsymbol{v} \in W$ means that the first column of $\boldsymbol{\Lambda}$ is a vector from W. The other columns can be chosen arbitrarily as long as the resulting matrix is symmetric:

$$\boldsymbol{\Lambda} = \left[\begin{array}{c|ccc} w_0 & w_1 & \ldots & w_d \\ \hline w_1 & & & \\ \vdots & & \boldsymbol{\Lambda}' & \\ w_d & & & \end{array}\right]$$

The dimension of the solution set is thus $\dim W + \dim \operatorname{Sym}(N-1) = r + \binom{N}{2}$. In the general case, apply a basis transformation $\boldsymbol{\Lambda} \mapsto \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^{\mathsf{T}}$ that transforms the given vector $\boldsymbol{v}$ into $\boldsymbol{e_1}$, and the same conclusion holds. □

Proof of Theorem 6.12. Let $\mathcal{S} \subset \operatorname{Sym}(N)$ denote the open subset of invertible symmetric matrices. Consider the null space incidence variety

$$\mathcal{N} = \{\, (a, \boldsymbol{\Lambda}, g) \mid \boldsymbol{B}_{\boldsymbol{\Lambda}}(a)\boldsymbol{g} = \boldsymbol{0} \,\} \subseteq \mathbb{P}(A) \times \mathcal{S} \times \mathbb{P}(G).$$

and recall that $\boldsymbol{B}_{\boldsymbol{\Lambda}}(a) = \boldsymbol{T}^{\boldsymbol{a}}_{N-r}\boldsymbol{\Lambda}^{-1}(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2r})^{\mathsf{T}}$.

The fibers over the projection $\mathcal{N} \to \mathbb{P}(A) \times \mathbb{P}(G) \colon (a, \boldsymbol{\Lambda}, g) \mapsto (a, g)$ are matrices $\boldsymbol{\Lambda}^{-1}$ which map the vector $(\boldsymbol{T}^{\boldsymbol{a}}_{N-r})^{\mathsf{T}}(\boldsymbol{T}^{\boldsymbol{a}}_{N-2r})^{\mathsf{T}}\boldsymbol{g}$ into null $\left(\boldsymbol{T}^{\boldsymbol{a}}_{N-r}\right)$. By the previous lemma, these fibers have dimension $r + \binom{N}{2}$, so that the incidence $\mathcal{N}$ has dimension

$$\underbrace{r}_{\dim \mathbb{P}(A)} + \underbrace{(N-2r-1)}_{\dim \mathbb{P}(G)} + r + \binom{N}{2} = \binom{N+1}{2} - 1 = \dim \mathcal{S} - 1.$$

Hence, the image of the projection $\mathcal{N} \to \mathcal{S} \colon (a, \boldsymbol{\Lambda}, g) \mapsto \boldsymbol{\Lambda}$ has dimension strictly smaller than $\dim \mathcal{S}$. In particular, its image has codimension at least one, and thus does not contain a generic matrix $\boldsymbol{\Lambda} \in \operatorname{Sym}(N)$. □

When the weights are such that the bad locus $\mathcal{B}_{\boldsymbol{\Lambda}}$ is empty, which by Theorem 6.12 is almost always the case, the varieties $\mathcal{Y}_{\boldsymbol{\Lambda}}$ and $\mathcal{Y}^{\mathrm{h}}_{\boldsymbol{\Lambda}}$ are equivalent: homogenizing the substituted critical equations over g does not create an extra component with $g_{-1} = 0$ (the bad locus). Thus, for these weights, we may continue our analysis of the ED degree using the null space incidence from (6.12), with the kernel vector (g_{-1}, g) living in projective space.

This insight allows us to characterize the critical points as a determinantal variety: the incidence (6.12) illustrates that the critical points can be described purely in terms of the rank of $\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{y}, a)$, so that we can eliminate the variables g from our analysis.

Corollary 6.14 (Equivalence to determinantal variety)**.** *When $\boldsymbol{\Lambda}$ is such that the bad locus is empty ($\mathcal{B}_{\boldsymbol{\Lambda}} = \emptyset$), the projection*

$$\phi \colon \{\, (a, g) \in A^{\circ}G \mid (\boldsymbol{y}, a, g) \in \mathcal{Y}_{\boldsymbol{\Lambda}} \,\} \to A^{\circ}$$

is an isomorphism onto the determinantal variety

$$\Sigma_{N-2r}(\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{y}, a)) = \{\, a \mid \operatorname{rank} \boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{y}, a) \leq N - 2r \,\} \subseteq A^{\circ}.$$

If $\boldsymbol{y} \in \mathbb{C}^N$ is generic, then $\Sigma_{N-2r}(\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{y}, a))$ is a finite set (finite up to scaling of a) of reduced points; that is, the maximal minors of $\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{y}, a)$ define the elements in a reduced way.

Proof. Let $\mathbf{\Lambda}$ be such that $\mathcal{B}_{\mathbf{\Lambda}} = \emptyset$. For any $(\boldsymbol{y}, a, g) \in \mathcal{Y}_{\mathbf{\Lambda}}$, the defining equations of $\mathcal{Y}_{\mathbf{\Lambda}}$ imply that $(-1, g)$ is a nonzero null vector of $\boldsymbol{M}_{\mathbf{\Lambda}}(\boldsymbol{y}, a)$, so that this matrix is rank-deficient. Conversely, because $\mathcal{B}_{\mathbf{\Lambda}} = \emptyset$ we have for any $a \in \Sigma_{N-2r}(\boldsymbol{M}_{\mathbf{\Lambda}}(\boldsymbol{y}, a))$ that $\operatorname{rank} \boldsymbol{B}_{\mathbf{\Lambda}}(a) = N - 2r$, and hence the null space of $\boldsymbol{M}_{\mathbf{\Lambda}}(\boldsymbol{y}, a)$ is one-dimensional. Thus there is a unique $(g, g_{-1}) \in \mathbb{P}(G \oplus \mathbb{C})$ with $g_{-1} \neq 0$ in this null space. Normalizing this vector to $g_{-1} = 1$ yields a unique g and therefore a unique point (a, g) for which $(\boldsymbol{y}, a, g) \in \mathcal{Y}_{\mathbf{\Lambda}}$. This shows that ϕ is bijective onto $\Sigma_{N-2r}(\boldsymbol{M}_{\mathbf{\Lambda}}(\boldsymbol{y}, a))$. Since ϕ is a morphism and its inverse is obtained algebraically from the one-dimensional kernel, ϕ is an isomorphism onto its image.

Corollary 6.11 further shows that for generic $\boldsymbol{y} \in \mathbb{C}^N$, the fiber $\tau^{-1}(\boldsymbol{y}) \in \mathcal{Y}_{\mathbf{\Lambda}}$ is finite up to scaling of a, and that the substituted critical equations define it in a reduced way. By Lemma 5.17, this reducedness carries over to the determinantal description. □

6.7 Degree of the Solution Set

If the weights $\mathbf{\Lambda}$ are such that $\mathcal{B}_{\mathbf{\Lambda}} = \emptyset$, application of the Thom–Porteous formula to the determinantal variety

$$\Sigma_{\boldsymbol{y},\mathbf{\Lambda}} \coloneqq \mathbb{P}(\Sigma_{N-2r}(\boldsymbol{M}_{\mathbf{\Lambda}}(\boldsymbol{y}, a))) \subseteq \mathbb{P}(A)$$

provides a direct computational approach to determine $\operatorname{EDD}_{\mathbf{\Lambda}}(\mathcal{X}_{N,r})$. Note that the projectivization of $\Sigma_{N-2r}(\boldsymbol{M}_{\mathbf{\Lambda}}(\boldsymbol{y}, a))$ is natural as all entries of the matrix $\boldsymbol{M}_{\mathbf{\Lambda}}(\boldsymbol{y}, a)$ are homogeneous in a.

Theorem 6.15. *Assume that the weight matrix $\mathbf{\Lambda}$ is such that $\mathcal{B}_{\mathbf{\Lambda}} = \emptyset$ and thus $\mathcal{G}_{\mathbf{\Lambda}} = \mathbb{P}^r$. Then for generic $\boldsymbol{y} \in \mathbb{C}^N$, the degree of $\Sigma_{\boldsymbol{y},\mathbf{\Lambda}}$ is given by*

$$\deg \Sigma_{\boldsymbol{y},\mathbf{\Lambda}} = \sum_{j=0}^{r} \binom{N-1-2r+j}{j} 3^j.$$

Proof. The expected dimension of $\Sigma_{\boldsymbol{y},\mathbf{\Lambda}}$ equals

$$r - (N - r - (N - 2r + 1) + 1) = 0,$$

so that, under the assumption that $\Sigma_{\boldsymbol{y},\mathbf{\Lambda}}$ is finite, we can apply Thom–Porteous formula for the corank one case (cf. Theorem 5.16).

Since $\boldsymbol{M}_{\mathbf{\Lambda}}(\boldsymbol{y}, a)$ has size $(N - r) \times (N - 2r + 1)$ and, except for the first column (which has degree one), all columns are of degree 3 in a, the computation reduces

to determining the coefficient of T^r (indeed, $N - r - (N - 2r + 1) + 1 = r$) in the series expansion of $(1-T)^{-1}(1-3T)^{-(N-2r)}$. Expanding each factor into a power series gives[3]

$$\frac{1}{1-T} = \sum_{j=0}^{\infty} T^j, \quad \frac{1}{(1-3T)^{N-2r}} = \sum_{j=0}^{\infty} (-3T)^j \binom{2r-N}{j}.$$

Taking the Cauchy product leads to

$$\frac{1}{(1-T)(1-3T)^{N-2r}} = \sum_{k=0}^{\infty} \left(\sum_{j=0}^{k} (-3)^j \binom{2r-N}{j} \right) T^k,$$

of which the rth coefficient corresponds to:

$$\sum_{j=0}^{r} (-3)^j \binom{2r-N}{j} = \sum_{j=0}^{r} 3^j \binom{N-1-2r+j}{j}. \qquad \square$$

We can use Theorem 6.15 to obtain the generic ED degree.

Corollary 6.16 (Generic ED degree)**.** *If the weight matrix $\boldsymbol{\Lambda}$ is such that $\mathcal{B}_{\boldsymbol{\Lambda}} = \emptyset$ and thus $\mathcal{G}_{\boldsymbol{\Lambda}} = \mathbb{P}^r$ (i.e., the matrix $\boldsymbol{B}_{\boldsymbol{\Lambda}}(a)$ never drops rank on $a \in \mathbb{P}^r$), then*

$$\mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r}) = \mathrm{EDD}_{\mathrm{gen}}(\mathcal{X}_{N,r}) = \sum_{j=0}^{r} \binom{N-1-2r+j}{j} 3^j. \tag{6.13}$$

Proof. The condition of $\boldsymbol{B}_{\boldsymbol{\Lambda}}(a)$ not dropping rank is satisfied for generic $\boldsymbol{\Lambda}$ (cf. Theorem 6.12), and so this condition is sufficient for having generic ED degree. $\square$

Let us investigate the set of weights $\boldsymbol{\Lambda}$ for which $\mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r})$ is strictly smaller than the generic ED degree. If the bad locus $\mathcal{B}_{\boldsymbol{\Lambda}}$ is non-empty but finite, then its degree gives a bound on $\mathrm{EDDdef}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r})$.

Theorem 6.17 (Bounds on the ED defect)**.** *Assume that $\mathcal{B}_{\boldsymbol{\Lambda}}$ is non-empty but finite. Then $\mathrm{EDDdef}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r}) > 0$. More precisely, for generic $\boldsymbol{y}$ we have*

$$\mathrm{EDDdef}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r}) = \sum_{a \in \mathcal{B}_{\boldsymbol{\Lambda}}} \mathrm{mult}_a \, I_{N-2r+1}(\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{y}, a)) \geq \deg \mathcal{B}_{\boldsymbol{\Lambda}}, \tag{6.14}$$

[3]Binomial coefficients extend to negative upper indices via the falling-factorial formula $\binom{n}{k} = n(n-1)\cdots(n-k+1)/k!$ for $k \geq 0$. In particular, for $m \in \mathbb{N}$ one has $\binom{-m}{k} = (-1)^k \binom{m+k-1}{k}$, which follows from the generalized binomial theorem.

where $I_{N-2r+1}(\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{y}, a))$ *denotes the ideal generated by the* $(N-2r+1)$*-minors of* $\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{y}, a)$ *and* $\operatorname{mult}_a I$ *denotes the multiplicity of the point* a *with respect to the ideal* I*.*

Proof. If $\mathcal{B}_{\boldsymbol{\Lambda}}$ is finite, then we can still apply Theorem 5.16 to the ideal of minors $I \coloneqq I_{N-2r+1}(\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{y}, a))$, whose zero set in $\mathbb{P}(A)$ is finite for generic $\boldsymbol{y}$. We can partition the degeneracy locus $\Sigma_{\boldsymbol{y},\boldsymbol{\Lambda}}$ into two sets, the solutions in $\mathcal{G}_{\boldsymbol{\Lambda}}$ and the solutions supported in the bad locus $\mathcal{B}_{\boldsymbol{\Lambda}}$:

$$\operatorname{EDD}_{\text{gen}}(\mathcal{X}_{N,r}) = \deg I = \sum_{a \in \mathcal{G}_{\boldsymbol{\Lambda}}} \operatorname{mult}_a I + \sum_{a \in \mathcal{B}_{\boldsymbol{\Lambda}}} \operatorname{mult}_a I.$$

To formally establish

$$\sum_{a \in \mathcal{G}_{\boldsymbol{\Lambda}}} \operatorname{mult}_a I = \operatorname{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r}),$$

we must show that for generic $\boldsymbol{y}_0 \in \mathbb{C}^N$

$$\phi(\mathcal{Y}_{\boldsymbol{\Lambda}} \cap \{\boldsymbol{y} = \boldsymbol{y}_0\}) \cap \mathcal{B}_{\boldsymbol{\Lambda}} = \emptyset.$$

The exceptional set of $\boldsymbol{y}_0$ for which the above intersection is non-empty has codimension equal to that of $\mathcal{B}_{\boldsymbol{\Lambda}}$ in $\mathbb{P}(A)$. Since $\mathcal{B}_{\boldsymbol{\Lambda}}$ is assumed to be finite, we have $\operatorname{codim} \mathcal{B}_{\boldsymbol{\Lambda}} = r$, and thus this exceptional set is avoided for generic choice of $\boldsymbol{y}$. This shows the first formula.

Furthermore, as $\boldsymbol{M}(\boldsymbol{y}, a) = [\boldsymbol{T}^{\boldsymbol{a}}_{N-r} \mid \boldsymbol{B}_{\boldsymbol{\Lambda}}(a)]$, it follows that $I \subseteq I_{N-r}(\boldsymbol{B}_{\boldsymbol{\Lambda}}(a))$, such that $\deg I \geq \deg I_{N-2r}(B_{\boldsymbol{\Lambda}}(a)) = \deg \mathcal{B}_{\boldsymbol{\Lambda}}$. This gives the claimed inequality. □

We now present explicit values of the ED degree of the Hankel variety $\mathcal{X}_{d,r}$, for three classes of diagonal weight matrices:

$$\text{Unit: } \boldsymbol{\Lambda} = \operatorname{diag}(1, \ldots, 1),$$

$$\text{Frobenius: } \boldsymbol{\Lambda} = \operatorname{diag}\left((\min\{i+1, N-i\})_{i=0,\ldots,N-1}\right),$$

$$\text{Bombieri: } \boldsymbol{\Lambda} = \operatorname{diag}\left(\tbinom{N-1}{0}, \tbinom{N-1}{1}, \ldots, \tbinom{N-1}{N-1}\right).$$

The results of the numerical and symbolic computations are depicted in Tables 6.1-6.3, which are consistent with [132, Table 4].

The observed values for unit weights and Frobenius weights lead to the following conjectures.

Table 6.1: Values of $\mathrm{EDD}_\Lambda(\mathcal{X}_{N,r})$ for unit weights. Entries shown in bold coincide with the generic ED degree $\mathrm{EDD}_{\mathrm{gen}}(\mathcal{X}_{N,r})$.

$N \setminus r$	1	2	3	4
2	**1**			
3	**4**			
4	**7**	**1**		
5	**10**	**13**		
6	**13**	**34**	**1**	
7	**16**	**64**	**40**	
8	**19**	**103**	**142**	**1**
9	**22**	**151**	**334**	**121**
10	**25**	**208**	**643**	**547**

Table 6.2: Values of $\mathrm{EDD}_\Lambda(\mathcal{X}_{N,r})$ for Frobenius weights. Entries shown in bold coincide with the generic ED degree $\mathrm{EDD}_{\mathrm{gen}}(\mathcal{X}_{N,r})$. The ED defect describes how a non-empty bad locus 'absorbs' solutions from $\mathrm{EDD}_{\mathrm{gen}}(\mathcal{X}_{N,r})$.

$N \setminus r$	1	2	3	4
2	**1**			
3	2			
4	**7**	**1**		
5	6	9		
6	**13**	**34**	**1**	
7	10	38	34	
8	**19**	**103**	**142**	**1**
9	14	103	246	113
10	**25**	**208**	**643**	543

Conjecture 6.18. *For unit weights $\mathbf{\Lambda} = \mathrm{diag}\,(1, \ldots, 1)$, the ED degree of $\mathcal{X}_{N,r}$ coincides with the generic value:*

$$\mathrm{EDD}(\mathcal{X}_{N,r}) = \mathrm{EDD}_{\mathrm{gen}}(\mathcal{X}_{N,r}).$$

Conjecture 6.19. *Frobenius weights are defective for odd N (but also for $\mathrm{EDD}_\Lambda(\mathcal{X}_{10,4})$) and their defect can be explained by the bad locus.*

Example 6.20. *Conjecture 6.18 is consistent with the numerical observations reported in Examples 3.13 and 4.10. Example 3.13 reported 739 complex solutions, which agrees with the evaluation of the generic expression* (6.13)

Table 6.3: Values of $\mathrm{EDD}_\Lambda(\mathcal{X}_{N,r})$ for Bombieri weights. Entries shown in bold coincide with the generic ED degree $\mathrm{EDD}_{\mathrm{gen}}(\mathcal{X}_{N,r})$. The ED defect describes how a non-empty bad locus 'absorbs' solutions from $\mathrm{EDD}_{\mathrm{gen}}(\mathcal{X}_{N,r})$. Entries for which the inequality from Theorem 6.17 is observed to be strict, and hence computation can not "definitively" prove the stated EDD, are depicted by a range (where the underlined numbers are estimates), whereas '+' indicates a positive dimensional bad locus.

$N \setminus r$	1	2	3	4
2	**1**			
3	2			
4	3	**1**		
5	4	7		
6	5	16	**1**	
7	6	28	20	
8	7	43..45	$\underline{62}$..64	**1**
9	8	61..65	$\underline{134}$..142	+
10	9	82..88	$\underline{243}$..263	+

for $N = 16$ and $r = 2$:

$$\sum_{j=0}^{2} \binom{15-4+j}{j} 3^j = 739.$$

Similarly, Example 4.10 reported 6466 complex solutions, matching the evaluation of (6.13) *for $N = 16$ and $r = 3$:*

$$\sum_{j=0}^{3} \binom{15-6+j}{j} 3^j = 6466.$$

For Frobenius weights, the defect seems to be completely captured by the points and multiplicities in the finite bad locus (i.e., equality holds in (6.14)).

Example 6.21. *Consider the case $N = 3$ and $r = 1$. For Frobenius weights, we obtain $\mathbf{\Lambda} = \mathrm{diag}\,(1, 2, 1)$, which in this case coincides with the Bombieri weights. After dehomogenizing $a \in \mathbb{P}(\mathbb{C}^2)$ by setting $a_0 = 1$, the matrix*

$$\boldsymbol{B}_{\mathbf{\Lambda}}(a) = \begin{bmatrix} a_0(a_0^2 + a_1^2/2) + (a_0 a_1^2)/2 \\ a_1(a_0^2/2 + a_1^2) + (a_0^2 a_1)/2 \end{bmatrix}$$

vanishes precisely for $(a_0, a_1) = (1, \pm j)$. Hence, the bad locus has $\#\mathcal{B}_{\mathbf{\Lambda}} = 2$. Numerical computations further show that the determinantal ideal $I_1(\boldsymbol{B}_{\mathbf{\Lambda}}(a))$ is reduced, so that both points occur with multiplicity one.

If we increase the problem size to $N = 5$, the bad locus still consists of the same two points, but they occur with multiplicity two wrt. $I_3(\boldsymbol{B}_{\boldsymbol{\Lambda}}(a))$.

The Bombieri weights are exceptional in that they exhibit a very large ED degree defect (conjecturally the largest possible [102]). However, the degree of the bad locus accounts for only part of this defect: for $N \geq 8$, equality in (6.14) seems not to be attained. In other words, the points of the bad locus appear with higher multiplicity with respect to the determinantal ideal of $\boldsymbol{M}_{\boldsymbol{\Lambda}}(\boldsymbol{y}, a)$ than with respect to that of $\boldsymbol{B}_{\boldsymbol{\Lambda}}(a)$.

In such cases, Table 6.3 gives a range of possible values for the ED degree. The upper bound of this range is given by the degree of the bad locus (cf. Theorem 6.17), while the lower bound is given by the number of affine solutions obtained from numerically solving[4] the substituted critical equations (Definition 6.8) for generic data $\boldsymbol{y} \in \mathbb{C}^N$. Entries in the tables marked with a "+" correspond to instances where the bad locus has positive dimension; in this situation Theorem 6.17 does not apply.

Example 6.22. *Consider the generic data*

$$\boldsymbol{y} = \begin{pmatrix} -0.07058313 \\ 0.53147675 \\ -0.80685232 \\ 2.45699133 \end{pmatrix} \in \mathbb{R}^4.$$

The dehomogenized common roots of the defining equations of $\mathcal{Y}^{\mathrm{h}}_{\boldsymbol{\Lambda}}$ from (6.11), for the case of $r = 1$ and Bombieri weights $\boldsymbol{\Lambda} = \operatorname{diag}(1, 3, 3, 1)$, are given in Table 6.4.

By Corollary 6.14, each of the common roots occurs as a point in the determinantal variety $\Sigma_{\boldsymbol{y},\boldsymbol{\Lambda}}$. The common roots for which $g_{-1} = 0$ come from the spurious component $\mathcal{Y}^{\infty}_{\boldsymbol{\Lambda}}$, which is introduced by the homogenization of the critical equations. For these solutions $a \in \mathcal{B}_{\boldsymbol{\Lambda}}$.

The total number of solutions corresponds to $\deg \Sigma_{\boldsymbol{y},\boldsymbol{\Lambda}} = \mathrm{EDD}_{\mathrm{gen}}(\mathcal{X}_{4,1}) = 7$. The bad locus $\mathcal{B}_{\boldsymbol{\Lambda}}$ consists of 2 points, each occurring with multiplicity 2, such that $\mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{4,1}) = 3$.

In this case, the multiplicity structure of $\mathcal{B}_{\boldsymbol{\Lambda}}$ is equivalent to the multiplicity structure of the bad locus points in $\Sigma_{\boldsymbol{y},\boldsymbol{\Lambda}}$. Denote the multiplicity of the bad locus points in the ideal of maximal minors of $\boldsymbol{B}_{\boldsymbol{\Lambda}}(a)$ as

$$m_{\mathcal{B}} = \sum_{a \in \mathcal{B}_{\boldsymbol{\Lambda}}} \operatorname{mult}_a I_{N-2r}(\boldsymbol{B}_{\boldsymbol{\Lambda}}(a)),$$

[4]We used the homotopy continuation implementation provided by [20].

Table 6.4: The dehomogenized common roots of the defining equations of $\mathcal{Y}^{\mathrm{h}}_{\boldsymbol{\Lambda}}$ from (6.11) (restricted to the problem setup considered in Example 6.22).

i	$\boldsymbol{a}$	$(g_{-1}, \boldsymbol{g})$
1	$1,\ 2.571$	$1,\ 0.010\,57,\ 0.069\,04$
2 - 3	$1,\ -0.4217 \pm 0.2799j$	$1,\ -0.4153 \mp 1.179j,\ 2.461 \pm 0.032\,21j$
4 - 5	$1,\ -j$	$0,\ -0.199\,20 - 0.678\,46j,\ -0.678\,46 + 0.199\,202j$
6 - 7	$1,\ j$	$0,\ -0.199\,20 + 0.678\,46j,\ -0.678\,46 - 0.199\,202j$

then this leads to the following characterization:

$$\underbrace{\mathrm{EDD}_{\mathrm{gen}}(\mathcal{X}_{4,1})}_{7} - \underbrace{\#\mathcal{B}_{\boldsymbol{\Lambda}}}_{2} \geq \underbrace{\mathrm{EDD}_{\mathrm{gen}}(\mathcal{X}_{4,1})}_{7} - \underbrace{m_{\mathcal{B}}}_{4} \geq \underbrace{\mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{4,1})}_{3}.$$

Example 6.23. *Tables 6.1 and 6.3 are consistent with the numerical observations reported in Example 3.31, in which Bombieri weights are considered. It was reported that the MEP had 208 solutions (which corresponds to the* $\mathrm{EDD}_{\mathrm{gen}}(\mathcal{X}_{10,2})$*), whereas the system of equations* (6.8) *had only 82 affine solutions (which falls within the range for* $\mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{10,2})$ *reported in Table 6.3). The other 126 solutions to the MEP correspond to points inside the bad locus, explaining why these solution paths tracked by homotopy continuation diverged to infinity along the direction of* $\boldsymbol{g}$*.*

6.8 The Singular Locus

In this section, we revisit the critical equations (cf. Definition 6.4) and focus on the subset of solutions $(\widehat{\boldsymbol{y}}, \boldsymbol{a}, \boldsymbol{\ell})$ for which $\widehat{\boldsymbol{y}}$ lies in the singular locus $\mathcal{X}^{\mathrm{sing}}_{N,r}$. We consider generic $\boldsymbol{y} \in \mathbb{C}^N$.

Recall the natural stratification of the Hankel variety:

$$\mathcal{X}_{N,r} \supseteq \mathcal{X}_{N,r-1} \supseteq \cdots \supseteq \mathcal{X}_{N,0}.$$

We begin with the most degenerate case, namely $\mathcal{X}_{N,0}$, which consists of the single point $\widehat{\boldsymbol{y}} = \boldsymbol{0}$. For this point, the equations (6.6b) and (6.6c) are trivially satisfied. By Lemma 6.6, the remaining equations (6.6a) are satisfied when $\widehat{\boldsymbol{y}}$ is the $\boldsymbol{\Lambda}$-orthogonal projection of $\boldsymbol{y}$ onto the null space of $\boldsymbol{T}^{\boldsymbol{a}}_{N-r}$. This null space is an r-dimensional linear subspace spanned by the (confluent) Vandermonde vectors

$$\boldsymbol{p_i} = [1, p_1, \ldots, p_i^{N-1}]^{\mathsf{T}}$$

corresponding to the roots of the polynomial a.

We denote the $\boldsymbol{\Lambda}$-orthogonal projection onto null $\left(\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\right)$ by:

$$\widehat{\boldsymbol{y}} = \boldsymbol{P}_{\boldsymbol{\Lambda},\boldsymbol{a}} = \left(\boldsymbol{I} - (\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1})^{\mathsf{T}}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}(\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1})^{\mathsf{T}})^{-1}\boldsymbol{T}_{N-r}^{\boldsymbol{a}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\right)\boldsymbol{y}.$$

For the resulting $\widehat{\boldsymbol{y}}$ to be identically zero, the vector $\boldsymbol{y}$ must be $\boldsymbol{\Lambda}$ orthogonal to each of the Vandermonde vectors $\boldsymbol{p_i}$:

$$\langle \boldsymbol{y}, \boldsymbol{p_i} \rangle_{\boldsymbol{\Lambda}} = \boldsymbol{y}^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{p_i} = \boldsymbol{0}, \quad i = 1, \ldots, r.$$

Because of the Vandermonde structure, these orthogonality conditions are satisfied if and only if the polynomial a has roots $p_1, \ldots, p_r$ selected (without repetition) from the set of roots of the polynomial

$$y_{\boldsymbol{\Lambda}}(z) = \sum_{j=0}^{N-1} c_j z^j, \qquad c_j = \sum_{i=1}^{N} \boldsymbol{\Lambda}_{(i,j+1)} y_i,$$

where $\boldsymbol{\Lambda}_{(i,j)}$ denotes the (i,j)th entry of $\boldsymbol{\Lambda}$. In the diagonal case, this expression reduces to

$$y_{\boldsymbol{\Lambda}}(z) = \boldsymbol{\Lambda}_{(0,0)} y_0 + \boldsymbol{\Lambda}_{(1,1)} y_1 z + \cdots + \boldsymbol{\Lambda}_{(N-1,N-1)} y_{N-1} z^{N-1}.$$

Assuming that $y_{\boldsymbol{\Lambda}}(z)$ has $N-1$ distinct roots, there are precisely $\binom{N-1}{r}$ possible choices of r roots to construct a. Consequently, exactly $\binom{N-1}{r}$ solutions to the critical equations lie in the stratum $\mathcal{X}_{N,0}$.

Assume $r > r' \geq 1$ and consider the stratum $\mathcal{X}_{N,r'}$. Let $(\widehat{\boldsymbol{y}}', \boldsymbol{a}', \boldsymbol{\ell}')$ be one of the $\mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r'})$ solutions of the substituted critical equations for rank r'. Since these are smooth solutions, Lemma 6.7 guarantees that the corresponding Lagrange multiplier satisfies

$$\ell' = a' \cdot g', \quad g' \in P_{\leq N-2r'}.$$

From each such solution for the smaller problem (rank r'), we can construct singular solutions to the critical equations for the larger problem (rank $r > r'$). Define

$$a = a' \cdot \prod_{i=r'+1}^{r} (z - p_i),$$

where the additional roots p_i are are chosen (without repetition) from the set of roots of the misfit polynomial

$$\widetilde{\boldsymbol{y}}_{\boldsymbol{\Lambda}}(z) = \boldsymbol{y}_{\boldsymbol{\Lambda}}(z) - \widehat{\boldsymbol{y}}_{\boldsymbol{\Lambda}}(z) = (a')^2 \cdot g'.$$

Here we used relation (6.7a), namely

$$\boldsymbol{\Lambda}(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = (\boldsymbol{T}_{N-r'}^{\boldsymbol{a}'})^{\mathsf{T}} (\boldsymbol{T}_{N-2r'}^{\boldsymbol{a}'})^{\mathsf{T}} \boldsymbol{g}'.$$

This construction of the point $(\widehat{\boldsymbol{y}}, \boldsymbol{a}, \boldsymbol{\ell})$ ensures that

$$\widehat{\boldsymbol{y}} = \boldsymbol{P}_{\boldsymbol{\Lambda},\boldsymbol{a}} \boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{\Lambda},\boldsymbol{a}'} \boldsymbol{y} = \widehat{\boldsymbol{y}}',$$

so the projected solution remains unchanged. Conceptually, we transfer some roots of the roots of $\ell' = a' \cdot g'$ into the new polynomial a (in particular, some of the roots of g'), so that the novel Lagrange multipliers take the form $\ell = a' \cdot f$ for some $f \in P_{\leq N-2r'-(r-r')}$.

It follows that the new triple $(\widehat{\boldsymbol{y}}, \boldsymbol{a}, \boldsymbol{\ell})$ is a solution to the critical equations for rank r.

- Equations (6.6a) and (6.6c) remain satisfied, since $\widehat{\boldsymbol{y}}$ is still obtained from the orthogonal projection of Lemma 6.6.
- Equation (6.6b) is also satisfied, because $\boldsymbol{H}_{r+1}^{\widehat{\boldsymbol{y}}} \boldsymbol{\ell} = 0$ holds: the factor a' in ℓ ensures that the recurrence relation defined by ℓ vanishes for $\widehat{\boldsymbol{y}}$.

Thus, every solution of the rank-r' problem generates a family of singular solutions for the rank-r problem. The number of such solutions contained in $\mathcal{X}_{N,r'}$ corresponds to

$$\mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,1}) \cdot \binom{N-1-2r'}{r-r'}.$$

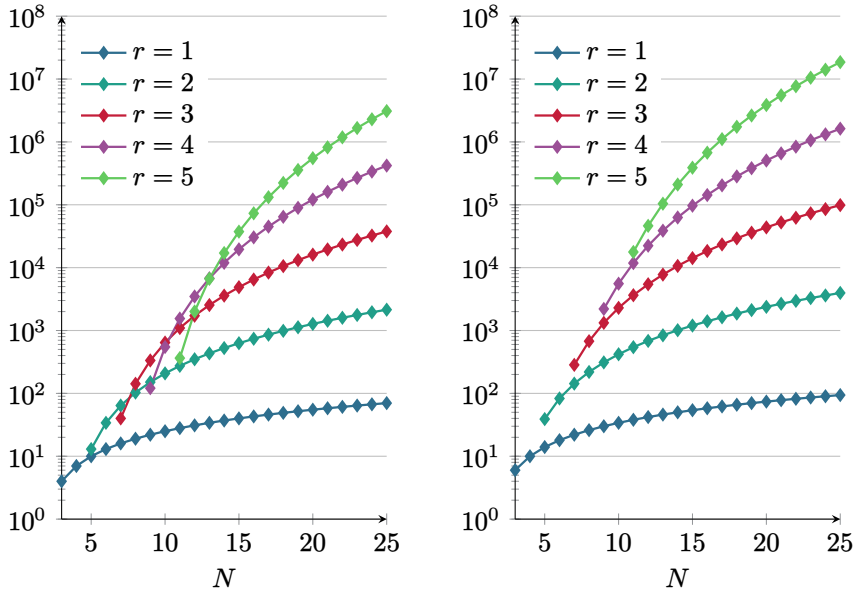
We can now state the general result.

Proposition 6.24. *We define the* complete Euclidean distance degree *of the Hankel variety, denoted* $\mathrm{CEDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r})$*, as the total number of complex solutions to the critical equations* (6.6a)-(6.6c)*, including those lying in the singular locus. By combining the contributions from all strata, we obtain for* $r \geq 1$*:*

$$\mathrm{CEDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r}) = \sum_{i=0}^{r} \mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,i}) \binom{N-1-2i}{r-i},$$

where $\mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,0}) = 1$.

The growth of $\mathrm{EDD}_{\mathrm{gen}}(\mathcal{X}_{N,r})$ and $\mathrm{CEDD}_{\mathrm{gen}}(\mathcal{X}_{N,r})$ for different problem sizes (N, r) is illustrated in Figure 6.3.



(a) $\mathrm{EDD}_{\mathrm{gen}}(\mathcal{X}_{N,r})$ (b) $\mathrm{CEDD}_{\mathrm{gen}}(\mathcal{X}_{N,r})$

Figure 6.3: The generic Euclidean distance degree $\mathrm{EDD}_{\mathrm{gen}}(\mathcal{X}_{N,r})$ and the complete Euclidean distance degree $\mathrm{CEDD}_{\mathrm{gen}}(\mathcal{X}_{N,r})$ for several problem sizes (N, r). Both quantities grow rapidly for larger problem sizes.

Example 6.25. *The closed-form expression for* $\mathrm{CEDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{N,r})$ *agrees with the numerical findings of Example 3.31, where the Bombieri weights are used. Specifically, for* $N = 10$ *and* $r = 2$*, we obtain*

$$\mathrm{CEDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{10,2}) = \mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{10,0})\binom{9}{2} + \mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{10,1})\binom{7}{1} + \mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{10,2})\binom{5}{0}$$

$$= 1 \cdot 36 \;+\; 9 \cdot 7 \;+\; 82 \cdot 1 = 181.$$

This matches the reported number of solutions of the Riemannian SVD (3.44) *formulation, which is equivalent to solving the critical equations* (6.6a)–(6.6c)*. Here we used* $\mathrm{EDD}_{\boldsymbol{\Lambda}}(\mathcal{X}_{10,2}) = 82$*, consistent with the number of affine common roots of the system* (6.8) *reported in Example 3.31.*

Example 6.26. *In the numerical example from [45, Section 8.2], 10 complex solutions were reported for the problem instance:* $\boldsymbol{y} = [4, 3, 2, 1]^{\mathsf{T}}$*, unit weights*

and $r = 1$. *This is in agreement with*

$$\mathrm{CEDD}(\mathcal{X}_{4,1}) = \mathrm{EDD}(\mathcal{X}_{4,0})\binom{3}{1} + \mathrm{EDD}(\mathcal{X}_{4,1})\binom{1}{0} = 1 \cdot 3 + 7 \cdot 1 = 10.$$

Indeed, the MEP-based approach in [45] solves for all complex solutions of the critical equations, including those that lie inside the singular locus. This is one of the main advantages of the novel MEP proposed in Chapter 3: For generic data $\boldsymbol{y}$*, its solution set consists of all the complex solutions lying inside the smooth locus, which, by Lemma 6.5, contains all (local) minimizers of the SLRA problem* (6.2).

6.9 Conclusions and Future Work

In this chapter, we analyzed the complex solutions to the critical equations of the structured low-rank approximation (SLRA) problem for Hankel matrices. We showed that the solutions decompose into a smooth and singular locus. We rederived the expression for the generic Euclidean distance degree of the low-rank Hankel variety and developed a framework that provides new insights into the ED defect, i.e., the difference between the Euclidean distance degree associated with a specific weight matrix and the generic ED degree. Finally, we complemented the ED degree with a count of the solutions in the singular locus, establishing, for the first time, a complete enumeration of all complex solutions to the critical equations of the SLRA problem. These novel insights are valuable both from a theoretical perspective and for practical applications, where the complexity of solving the system (or MEP) is determined by the number of such solutions.

The generic ED degree grows rapidly with problem size, such that computing all critical points quickly becomes computationally infeasible in practice. Moreover, we observed that the total number of solutions, counting both the smooth and singular loci (cf. the complete ED degree), grows even faster than the generic ED degree alone. This confirms the computational advantage of the globally optimal MEP-based formulation proposed in Chapter 3, since, for generic data, it excludes all singular-locus solutions from its solution set and thus reduces the effective solution count compared to directly solving the critical equations.

In future research, we would like to focus on further understanding and refining these results. First, we aim to analyze in detail specific classes of weights for which the ED defect is nonzero, with the goal of obtaining sharper bounds on the defect by studying the geometry of the corresponding bad loci. Second, although our analysis has been carried out over the complex numbers, only

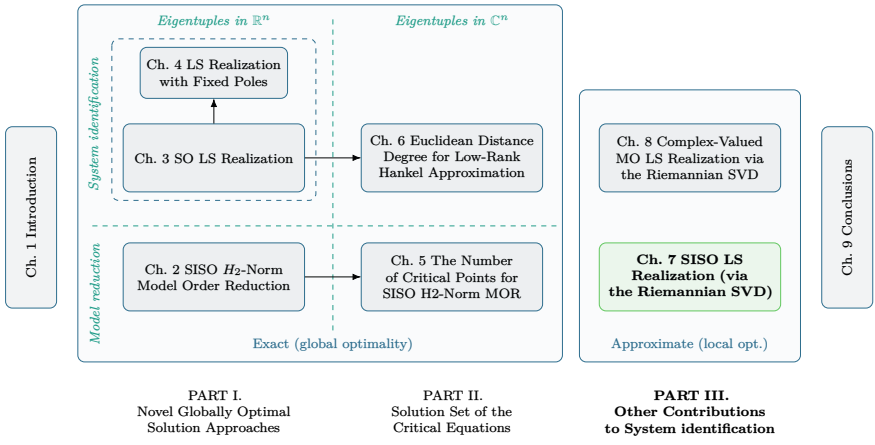
the real solutions correspond to genuine critical points (e.g., local minima and saddle points) of the SLRA problem. Since the number of real solutions depends on the specific data instance, no general closed-form expression is available. Nevertheless, extensive numerical experiments could provide insight into the typical proportion of real solutions within the full solution set. Alternatively, one could investigate ways to predict the number of real solutions for a given data instance, for example via the study of data discriminants.

Part III

Other Contributions to System Identification

Chapter 7

Least Squares Projection Onto the Behavior for SISO LTI Models



Chapter abstract

We first consider the least squares projection onto the behavior for linear time-invariant (LTI) single-input single-output (SISO) models, in which the observed input-output data are modified in a least squares (LS) sense by subtracting so-called misfits, so that the modified data satisfy a given linear dynamic relation. We show that the LS-criterion of the projection problem induces an orthogonal decomposition of the ambient data space and we characterize this decomposition by means of banded block-Toeplitz matrices, the elements of which are the coefficients of the difference equation describing the SISO LTI dynamics. We thereby generalize earlier results in the literature on autonomous LTI models to the more complicated SISO case. Additionally, we illustrate that the novel characterization is equivalent (up to a change of model representation) to results derived using (isometric) state-space representations in the literature on behavioral systems theory.

We then explore several characterizations of the locally optimal solutions of the full SISO LS realization problem, in which also the model parameters have to be estimated from the given data. Directly solving the system of equations arising from the kernel-based Riemannian SVD formulation emerged as the most promising candidate for a globally optimal methodology, for which no prior results are available in the literature. However, as all variables appear nonlinearly in these equations, a direct MEP reformulation is not possible. Several alternative strategies for obtaining global optimality were considered, but each encountered significant (computational) obstacles.

Candidate's contributions and relation to published work

This work has been previously published as:

- **Lagauw, S.** and De Moor, B. (2024). "Least Squares Projection Onto the Behavior for SISO LTI Models". In: *IFAC-PapersOnLine, 58.15. Part of special issue: 20th IFAC Symposium on System Identification (SYSID)*. Northeastern University, Boston, United States, pp. 336–341.

The candidate led the theoretical development, implemented the code, conducted the experiments, and authored the text.

This chapter differs slightly from the original article:

- The notation and terminology was adapted to ensure consistency with the conventions used throughout the remainder of this dissertation.
- Section 7.4 is novel compared to the earlier published work. It considers several characterizations of the locally optimal solutions of the full SISO LS realization problem, in which also the model parameters have to be estimated from the given data.

7.1 Introduction

System identification translates observed time-series data into a mathematical model, which typically belongs to a user-specified model class. In practice, however, the observed data as such are almost never *model-compliant*, i.e., the data do not belong to the *behavior* of the model, which is the set of all trajectories that exactly satisfy the model dynamics [187, 86]. There could be many reasons for this, e.g., measurement inaccuracies, missing data, observational errors, and model mismatch. Stochastics are often introduced to explain this discrepancy between the observed data and the model [11, 120]. However, unless one has a priori information about the nature of the stochastic disturbances in the observed data, one might want to avoid making any statistical assumption(s). Instead, a more natural way to proceed is to modify the observed data as little as possible, by subtracting so-called *misfits*, so that the modified data become model-compliant. The smaller these misfits, the better the 'fit' between the model and the observed data.

7.1.1 Problem formulation

We consider discrete-time causal single-input single-output (SISO) linear time-invariant (LTI) models of finite order n. For this model class, input data $\widehat{\boldsymbol{u}} = [\widehat{u}_0, \dots, \widehat{u}_{N-1}]^\mathsf{T} \in \mathbb{R}^N$ and output data $\widehat{\boldsymbol{y}} = [\widehat{y}_0, \dots, \widehat{y}_{N-1}]^\mathsf{T} \in \mathbb{R}^N$ are model-compliant if and only if they are exactly related by an nth order LTI input-output relation,

$$\sum_{i=0}^{n} a_i \widehat{y}_{k+i} - \sum_{i=0}^{n} b_i \widehat{u}_{k+i} = 0, \ \forall k \in \{0, \dots, N-n-1\}, \tag{7.1}$$

where we assume that $N > 3n + 1$. The $2(n+1)$ coefficients that appear in (7.1) are the model parameters of the SISO LTI model, which we will group in the vectors $\boldsymbol{a} = [a_0, \dots, a_n]^\mathsf{T} \in \mathbb{R}^{n+1}$ and $\boldsymbol{b} = [b_0, \dots, b_n]^\mathsf{T} \in \mathbb{R}^{n+1}$. As such, given an observed input signal $\boldsymbol{u} = [u_0, \dots, u_{N-1}]^\mathsf{T} \in \mathbb{R}^N$ and output signal $\boldsymbol{y} = [y_0, \dots, y_{N-1}]^\mathsf{T} \in \mathbb{R}^N$, the identification problem consists of the minimization of the 2-norm of the difference between the observed and model-compliant data sequences $\widetilde{\boldsymbol{u}} = \boldsymbol{u} - \widehat{\boldsymbol{u}}$ and $\widetilde{\boldsymbol{y}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$, the so-called *misfits*, over the variables $\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{y}}$, and the model parameters $\boldsymbol{a}, \boldsymbol{b}$,

$$\begin{aligned} \underset{\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{y}}, \boldsymbol{a}, \boldsymbol{b}}{\text{minimize}} \quad & J = \frac{1}{2} \left\| \begin{bmatrix} \boldsymbol{y} - \widehat{\boldsymbol{y}} \\ \boldsymbol{u} - \widehat{\boldsymbol{u}} \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} \widetilde{\boldsymbol{y}} \\ \widetilde{\boldsymbol{u}} \end{bmatrix} \right\|_2^2, \\ \text{subject to} \quad & \text{Eq. (7.1) and } g(\boldsymbol{a}, \boldsymbol{b}) = 1, \end{aligned} \tag{7.2}$$

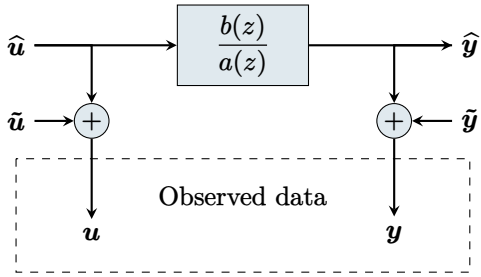


Figure 7.1: The observed input-output signals $\boldsymbol{u}, \boldsymbol{y} \in \mathbb{R}^N$, respectively, are decomposed into a model-compliant part $\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{y}}$, and so-called misfits $\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{y}}$, the 2-norm of which is to be minimized. The coefficients $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{n+1}$ of the nth degree polynomials $a(z)$ and $b(z)$ in the forward-shift z, are the model parameters.

where $g(\boldsymbol{a}, \boldsymbol{b})$ formulates a normalization constraint to ensure uniqueness of the difference relation in (7.1), e.g., $g(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}^\mathsf{T}\boldsymbol{a} + \boldsymbol{b}^\mathsf{T}\boldsymbol{b}$ or $g(\boldsymbol{a}, \boldsymbol{b}) = a_n$.

To ease the notation, we introduce the observed *trajectory* $\boldsymbol{w} \in \mathbb{R}^{2N}$, which comprises the input-output data,

$$\boldsymbol{w} = \begin{bmatrix} \boldsymbol{w}_0^\mathsf{T} & \dots & \boldsymbol{w}_{N-1}^\mathsf{T} \end{bmatrix}^\mathsf{T} = \begin{bmatrix} y_0 & u_0 & \dots & y_{N-1} & u_{N-1} \end{bmatrix}^\mathsf{T},$$

such that $\boldsymbol{w}_k = \begin{bmatrix} y_k & u_k \end{bmatrix}^\mathsf{T}$, and we define $\widehat{\boldsymbol{w}}$ and $\widetilde{\boldsymbol{w}}$ similarly. The partial derivative of a scalar function $f(\dots)$ with respect to $\boldsymbol{w}_k$ will be denoted as,

$$\frac{\partial f(\dots)}{\partial \boldsymbol{w}_k} = \begin{bmatrix} \frac{\partial f(\dots)}{\partial y_k} & \frac{\partial f(\dots)}{\partial u_k} \end{bmatrix}^\mathsf{T}.$$

7.1.2 Literature overview

This *misfit modeling* setup, which is depicted in Figure 7.1, corresponds to the *pure-misfit*-case of the more general misfit-versus-latency framework proposed in [114]. The problem is also known in the literature on behavioral systems theory [186], where it is described as a particular instance of the *global total least squares* (GTLS) problem [144]: the GTLS formulation is more general than (7.2) in the sense that it considers the broader class of multiple-input multiple-output (MIMO) models.

Plain nonlinear solvers could be used to tackle the constrained nonlinear least squares problem (7.2), but also dedicated solution approaches exist. The problem can be reformulated to a structured total least squares problem, where the (locally) optimal model parameters are to be retrieved via the non-trivial

null space of a structured matrix approximation of the double-Hankel matrix containing the observed input-output data [38, 39, 126, 125].

Alternatively, the methods proposed in [143] exploit the observation that (7.2) is a double minimization problem:

1. for given observed data and model parameters, the minimal norm misfits that make the observed data model-compliant are to be retrieved (the inner minimization);
2. the model parameters itself are to be optimized (the outer minimization).

In particular, the fact that the optimal solution to the inner-problem is given by a linear transformation (projection), allows one to substitute the closed-form solution into the outer problem, thereby eliminating the $2N$ decision variables $\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{y}}$ from the optimization problem. A similar approach has been taken in Chapter 3, where the misfit modeling problem for autonomous LTI models is considered: the optimal misfits are expressed as the orthogonal projection of the observed data onto a linear subspace, which is completely determined[1] by the parameters of the autonomous model.

7.1.3 Contributions and outline

In this chapter, we study the inner minimization problem: the *least squares projection onto the behavior for SISO LTI models.* Our contribution is threefold:

1. In Section 7.2, we show that the least squares criterion induces an orthogonal decomposition of the ambient data space. We completely characterize this decomposition using banded (block-) Toeplitz matrices, the elements of which are the coefficients of the difference equation describing the SISO LTI dynamics (7.1). We thereby generalize the results from [45, Section 4] to the more complicated SISO case. Several optimality properties are obtained (e.g., the optimal misfits are structured, in the sense that they are itself generated by an nth order LTI model), which shed new light on results obtained in previous work [39, 46].
2. In Section 7.3, we reconsider the projection onto the behavior using an (isometric) state-space representation to model the SISO LTI dynamics,

[1] The behavior of the model corresponds to the right null space of a banded Toeplitz matrix, which is constructed from the coefficients of the difference equation describing the autonomous LTI dynamics. The least squares optimal misfits are shown to belong to the orthogonal complementary subspace of this behavior, i.e., the row space of the banded Toeplitz matrix [45].

instead of the difference relation in (7.1), similar to the framework proposed in [144]. We demonstrate that the orthogonal decomposition of the ambient space as established in Section 7.2 can also be characterized in this setting, thereby providing a state-space based alternative to the results obtained in Section 7.2. The combination of both perspectives allows for a more intuitive understanding of the obtained results.

3. In Section 7.4, we address the full SISO LS realization problem, where the model parameters must also be estimated from the observed data. We derive several alternative characterizations of the locally optimal solutions, including a novel image representation based Riemannian SVD formulation, and use these to explore potential strategies for obtaining a globally optimal methodology. As no prior results on globally optimal LS realization for SISO models exist in the literature, the results of this section provide a foundation for future progress.

7.2 Projection onto the Behavior

In the projection problem, a given sequence of observed data $\boldsymbol{w}$ has to be projected onto the linear dynamics of a given model (i.e., the model parameters $\boldsymbol{a}, \boldsymbol{b}$ of (7.1) are known), so that the 2-norm of the obtained misfits $\widetilde{\boldsymbol{w}}$ is minimal. With $\boldsymbol{\ell} = [\ell_0, \ldots, \ell_{N-n-1}]^\mathsf{T} \in \mathbb{R}^{N-n}$ Lagrange multipliers, the Lagrangian of this projection problem becomes,

$$\mathcal{L}(\widehat{\boldsymbol{w}}, \boldsymbol{\ell}) = \frac{1}{2} \sum_{k=0}^{N-1} \|\boldsymbol{w}_k - \widehat{\boldsymbol{w}}_k\|_2^2 + \sum_{k=0}^{N-n-1} \ell_k \left(\sum_{i=0}^{n} a_i \widehat{y}_{k+i} - \sum_{i=0}^{n} b_i \widehat{u}_{k+i} \right). \quad (7.3)$$

The first-order necessary conditions for optimality, which can be obtained by equating all partial derivatives of (7.3) to zero, allow to derive the following results.

Writing out the linear relation from (7.1) for all $k = 0, \ldots, N-1$, gives,

$$\widetilde{\boldsymbol{T}}^\mathsf{T} \widehat{\boldsymbol{w}} = \boldsymbol{0}, \quad (7.4)$$

where $\widetilde{\boldsymbol{T}} \in \mathbb{R}^{(2N)\times(N-n)}$ is a banded block-Toeplitz matrix, the transpose of which is defined[2] as,

$$\widetilde{\boldsymbol{T}}^{\mathsf{T}} = \begin{bmatrix} a_0 & -b_0 & a_1 & -b_1 & \dots & \dots & a_n & -b_n & & & & \\ & & a_0 & -b_0 & \dots & \dots & a_{n-1} & -b_{n-1} & a_n & -b_n & & \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & & a_0 & -b_0 & \dots & \dots & a_{n-1} & -b_{n-1} & a_n & -b_n \end{bmatrix}.$$

7.2.1 The behavior

Equation (7.4) shows that the model-compliant data $\widehat{\boldsymbol{w}}$ must lie in the (right) null space of $\widetilde{\boldsymbol{T}}^{\mathsf{T}}$, i.e., the behavior of the model from (7.1) corresponds to $\operatorname{null}\left(\widetilde{\boldsymbol{T}}^{\mathsf{T}}\right)$. The latter is an $(N+n)$-dimensional subspace of $\mathbb{R}^{2N}$, since the normalization constraint $g(\boldsymbol{a},\boldsymbol{b})$ together with the structure of the matrix ensures that $\widetilde{\boldsymbol{T}}$ is of full-column rank. The banded block-Toeplitz structure of $\widetilde{\boldsymbol{T}}^{\mathsf{T}}$ allows us to construct a basis for its null space by hand, leading to the matrix $\widehat{\boldsymbol{T}} \in \mathbb{R}^{(2N)\times(N+n)}$:

$$\widehat{\boldsymbol{T}} = \left[\begin{array}{ccc|cccc|ccc} b_0 & \dots & b_{n-1} & b_n & & & & & & \\ a_0 & \dots & a_{n-1} & a_n & & & & & & \\ & \ddots & \vdots & b_{n-1} & b_n & & & & & \\ & \ddots & \vdots & a_{n-1} & a_n & \ddots & & & & \\ & & b_0 & \vdots & \vdots & \ddots & b_n & & & \\ & & a_0 & \vdots & \vdots & & a_n & & & \\ & & & b_0 & b_1 & & \vdots & b_n & & \\ & & & a_0 & a_1 & \ddots & \vdots & a_n & & \\ & & & & b_0 & \ddots & b_1 & \vdots & \ddots & \\ & & & & a_0 & \ddots & a_1 & \vdots & \ddots & \\ & & & & & \ddots & b_0 & b_1 & \dots & b_n \\ & & & & & & a_0 & a_1 & \dots & a_n \end{array}\right].$$

If we assume that the model in (7.1) is minimal, such that the polynomials $a(z)$ and $b(z)$ are coprime, the columns of $\widehat{\boldsymbol{T}}$ are guaranteed to be linearly independent (this is a consequence of [113, Lemma 3.3]). Combined with (7.4),

[2] Notice that this banded block-Toeplitz matrix is the SISO generalization of the banded Toeplitz matrix derived in [45].

we find that,

$$\widehat{\boldsymbol{w}} \in \text{null}\left(\widetilde{\boldsymbol{T}}^{\mathsf{T}}\right) = \text{range}\left(\widehat{\boldsymbol{T}}\right).$$

We can conclude that the $(N{+}n)$-dimensional subspace range $\left(\widehat{\boldsymbol{T}}\right)$ corresponds to the *behavior* associated to the model described by the input-output dynamics from (7.1). Unsurprisingly, the dimensionality of the behavior corresponds to the degrees of freedom (dofs) that one has to construct the trajectory $\widehat{\boldsymbol{w}}$: when simulating an nth order SISO model for $k = 0, \ldots, N{-}1$, there are n initial conditions and N inputs that can be chosen freely.

7.2.2 The misfit space

It is straightforward, though notationally tedious, to show that equating $\partial\mathcal{L}(\ldots)/\widehat{\boldsymbol{w}}_k$ to zero for all $k = 0, \ldots, N{-}1$, leads to,

$$\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{T}}\boldsymbol{\ell}, \tag{7.5}$$

indicating that the optimal output misfits $\widetilde{\boldsymbol{y}}$ can be expressed as the convolution of the vectors $\boldsymbol{a}$ and $\boldsymbol{\ell}$, and similarly that the optimal input misfits $\widetilde{\boldsymbol{u}}$ are obtained from the convolution of $-\boldsymbol{b}$ and $\boldsymbol{\ell}$. Furthermore, it can be seen that the misfit trajectory $\widetilde{\boldsymbol{w}}$ 'resides' in the $(N{-}n)$-dimensional subspace spanned by the columns of the matrix $\widetilde{\boldsymbol{T}}$. We will call this subspace the *misfit space*.

The orthogonal complementarity between the row and (right) null space of a matrix imply that the model- and misfit spaces are orthogonal complementary subspaces in $\mathbb{R}^{2N}$. Indeed, the matrix product $\widetilde{\boldsymbol{T}}^{\mathsf{T}}\widehat{\boldsymbol{T}}$ is zero by construction. By consequence, the first-order necessary conditions for optimality of the data projection problem, and thus also of (7.2), induce an orthogonal decomposition of the ambient data space,

$$\mathbb{R}^{2N} = \underbrace{\text{range}\left(\widehat{\boldsymbol{T}}\right)}_{\text{behavior}} \oplus \underbrace{\text{range}\left(\widetilde{\boldsymbol{T}}\right)}_{\text{misfits}}. \tag{7.6}$$

This decomposition implies that the optimal trajectories $\widehat{\boldsymbol{w}}$ and $\widetilde{\boldsymbol{w}}$ are mutually orthogonal,

$$\widehat{\boldsymbol{w}} \perp \widetilde{\boldsymbol{w}}. \tag{7.7}$$

As shown in [39], one can alternatively derive (7.7) by substituting (7.5) in (7.4), after multiplying (7.4) from the left with $\boldsymbol{\ell}^{\mathsf{T}}$.

Counterintuitively, a natural consequence of the data orthogonality (7.7) is that besides the model-compliant data, also the optimal input-output misfits $\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{y}}$,

are structured. In particular, the $(n{+}1)$th equation up until the Nth equation[3] in,

$$\widehat{\boldsymbol{T}}^{\mathsf{T}}\widetilde{\boldsymbol{w}} = \boldsymbol{0}, \tag{7.8}$$

show that the input-output misfits satisfy a linear difference relation of the following form,

$$\sum_{i=0}^{n} b_{n-i}\widetilde{y}_{k+i} + \sum_{i=0}^{n} a_{n-i}\widetilde{u}_{k+i} = 0, \ \forall k \in \{0, \dots, N{-}n{-}1\}.$$

Characterizing these misfit dynamics in the z-domain (where z denotes the forward-shift: $zy_k = y_{k+1}$) leads to,

$$\widetilde{H}(z) = -\frac{a_r(z)}{b_r(z)} = \frac{a_0 z^n + \cdots + a_{n-1}z + a_n}{b_0 z^n + \cdots + b_{n-1}z + b_n}, \tag{7.9}$$

with the 'reversed-coefficients'-polynomials defined as $a_r(z) = z^n a(z^{-1})$ and $b_r(z) = z^n b(z^{-1})$. The transfer function in (7.9) mapping the inputs $\widetilde{\boldsymbol{u}}$ to the outputs $\widetilde{\boldsymbol{y}}$ can be interpreted as the 'inverse' of the transfer function of the model governed by (7.1),

$$\widehat{H}(z) = \frac{b(z)}{a(z)} = \frac{b_n z^n + \cdots + b_1 z + b_0}{a_n z^n + \cdots + a_1 z + a_0}. \tag{7.10}$$

Alternatively, assuming that the given model (7.1) is stable, one can derive (7.9) immediately from (7.7) by expressing the orthogonality[4] in the z-domain, giving,

$$\begin{aligned} &\langle \frac{b(z)}{a(z)}\widehat{U}(z), \widetilde{Y}(z)\rangle + \langle \widehat{U}(z), \widetilde{U}(z)\rangle = 0 \\ \iff &\frac{1}{2\pi i}\oint_{|z|=1}\left[\widetilde{Y}(z)\frac{b(z^{-1})}{a(z^{-1})} + \widetilde{U}(z)\right]\widehat{U}(z^{-1})\mathrm{d}z = 0 \\ \iff &\widetilde{Y}(z) = \frac{-a(z^{-1})}{b(z^{-1})}\widetilde{U}(z) = \frac{-a_r(z)}{b_r(z)}\widetilde{U}(z). \end{aligned}$$

[3] Notice the difference compared to the model-compliant data: the first and last n equations in (7.8) additionally constrain $\widetilde{\boldsymbol{w}}$. We will see in Section 7.3 that these constraints ensure that the initial ($k = 0$) and final ($k = N{-}1$) state of the misfit model is zero.

[4] The transfer (matrix) function of causal, stable, discrete-time LTI models can be considered as an element of the Hardy space $\mathcal{H}_2$ of functions analytic on the exterior of the unit-disc. The inner-product $\langle \boldsymbol{F}(z), \boldsymbol{G}(z)\rangle$ of two (matrix) functions $\boldsymbol{F}(z), \boldsymbol{G}(z) \in \mathcal{H}_2$ is defined as:

$$\langle \boldsymbol{F}(z), \boldsymbol{G}(z)\rangle = \frac{1}{2\pi i}\oint_{\mathbb{T}} \boldsymbol{F}^{\mathsf{H}}(1/z^{\mathsf{H}})\boldsymbol{G}(z)\frac{dz}{z}, \tag{7.11}$$

with $\mathbb{T}$ the unit circle: $|z| = 1$. Note that in the context of LTI systems, the function $\boldsymbol{F}(z)$ corresponds to a transfer (matrix) function with real-valued impulse response, such that $\boldsymbol{F}^{\mathsf{H}}(1/z^{\mathsf{H}}) = \boldsymbol{F}^{\mathsf{T}}(1/z)$ [182].

The relation in (7.9) was also obtained in [46, 39], albeit without proof.

7.2.3 Optimal projection

It is a consequence of the orthogonal decomposition of the ambient space in (7.6) that the least squares optimal decomposition of the observed input-output data $\boldsymbol{w} = \widehat{\boldsymbol{w}} + \widetilde{\boldsymbol{w}}$ for a given model $(\boldsymbol{a}, \boldsymbol{b})$ boils down to an orthogonal projection problem: the minimal norm misfits $\widetilde{\boldsymbol{w}}$ are obtained via the orthogonal projection of $\boldsymbol{w}$ onto the misfit space, range $\left(\widetilde{\boldsymbol{T}}\right)$,

$$\widetilde{\boldsymbol{w}} = (\widetilde{\boldsymbol{T}}^{\mathsf{T}})^{\dagger}\widetilde{\boldsymbol{T}}^{\mathsf{T}}\boldsymbol{w} = \widetilde{\boldsymbol{T}}(\widetilde{\boldsymbol{T}}^{\mathsf{T}}\widetilde{\boldsymbol{T}})^{-1}\widetilde{\boldsymbol{T}}^{\mathsf{T}}\boldsymbol{w}, \tag{7.12}$$

or equivalently, the optimal regular data $\widehat{\boldsymbol{w}}$ can be computed via the orthogonal projection of $\boldsymbol{w}$ onto range $\left(\widehat{\boldsymbol{T}}\right)$,

$$\begin{aligned}\widehat{\boldsymbol{w}} &= \widehat{\boldsymbol{T}}(\widehat{\boldsymbol{T}}^{\mathsf{T}}\widehat{\boldsymbol{T}})^{-1}\widehat{\boldsymbol{T}}^{\mathsf{T}}\boldsymbol{w}, \\ &= (\boldsymbol{I} - (\widetilde{\boldsymbol{T}}^{\mathsf{T}})^{\dagger}\widetilde{\boldsymbol{T}}^{\mathsf{T}})\boldsymbol{w}.\end{aligned}$$

7.3 The State-Space Dual

Inspired by the framework developed in [144], in which the least squares misfit modeling problem has been studied from a 'behavioral' point of view, we reconsider the projection onto the behavior in this section using a state-space model to parametrize the SISO LTI dynamics, instead of the difference equation from (7.1). Unsurprisingly, we find the same results as in Section 7.2, albeit that they appear in a different form due to the state-space representation.

7.3.1 Behavioral system theoretic models

Alternative to the relation in (7.1), we can enforce the nth order SISO LTI dynamics on the model-compliant data $\widehat{\boldsymbol{w}}$ by requiring the data to satisfy the following behavioral state-space recurrence relations,

$$\begin{aligned}\widehat{\boldsymbol{x}}_{k+1} &= \boldsymbol{A}\widehat{\boldsymbol{x}}_k + \boldsymbol{B}\widehat{v}_k \\ \widehat{\boldsymbol{w}}_k &= \boldsymbol{C}\widehat{\boldsymbol{x}}_k + \boldsymbol{D}\widehat{v}_k,\end{aligned} \qquad \text{for } k = 0, \ldots, N-1, \tag{7.13}$$

with $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, $\boldsymbol{B} \in \mathbb{R}^{n}$, $\boldsymbol{C} \in \mathbb{R}^{2 \times n}$, $\boldsymbol{D} \in \mathbb{R}^{2}$. Contrary to a standard input-output state-space representation, which maps inputs $\widehat{\boldsymbol{u}}$ to outputs $\widehat{\boldsymbol{y}}$, the model

in (7.13) maps a sequence of auxiliary variables $\widehat{\boldsymbol{v}} \in \mathbb{R}^N$ to the trajectory $\widehat{\boldsymbol{w}}$. The auxiliary variables $\widehat{\boldsymbol{v}}$ have no immediate physical interpretation, and one should not consider them as 'inputs' [186]. Rather, they represent the degree of freedom in the trajectory $\widehat{\boldsymbol{w}}$, at each time step k, that originates from the SISO structure of (7.1).

Besides the standard basis change of the state vector $\widehat{\boldsymbol{x}}_k$, which makes up the similarity transform of standard input-output state-space models, also the addition of static feedback and a change of basis[5] of the auxiliary variables $\widehat{v}_k$ can be performed without altering the behavior of (7.13) [144]. In particular, it has been shown that, assuming minimality and stability of the original input-output relation modeled by (7.13), one can always perform a similarity transformation on the behavioral model (7.13) so that the resulting model becomes *isometric* [144],

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = \boldsymbol{I}_{n+1}.$$

Therefore, to simplify our derivations, we can assume without loss of generality that the behavioral state-space model $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$ in (7.13) is isometric.

Given an observed trajectory $\boldsymbol{w}$ and model $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$, the projection onto the behavior now consists of finding the minimal norm $\widetilde{\boldsymbol{w}}$ such that the obtained $\widehat{\boldsymbol{w}} = \boldsymbol{w} - \widetilde{\boldsymbol{w}}$ is model-compliant with respect to (7.13). Using the Lagrange multipliers $\boldsymbol{\lambda}_k \in \mathbb{R}^n$ and $\boldsymbol{\mu}_k \in \mathbb{R}^2$ for $k = 0, ..., N-1$, we can write the Lagrangian of the projection problem as,

$$\begin{aligned} \mathcal{L}(\widehat{\boldsymbol{w}}, \widehat{\boldsymbol{x}}, \widehat{\boldsymbol{v}}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{2} \sum_{k=0}^{N-1} \|\widehat{\boldsymbol{w}}_k - \boldsymbol{w}_k\|_2^2 \\ + \sum_{k=0}^{N-1} \boldsymbol{\lambda}_k^{\mathsf{T}} (\widehat{\boldsymbol{x}}_{k+1} - \boldsymbol{A}\widehat{\boldsymbol{x}}_k - \boldsymbol{B}\widehat{v}_k) + \sum_{k=0}^{N-1} \boldsymbol{\mu}_k^{\mathsf{T}} (\widehat{\boldsymbol{w}}_k - \boldsymbol{C}\widehat{\boldsymbol{x}}_k - \boldsymbol{D}\widehat{v}_k). \end{aligned}$$

The structure of the behavioral state-space model allows us to easily construct a basis for the behavior in which $\widehat{\boldsymbol{w}}$ resides. Writing out the recurrence relation in (7.13) for $k = 0, \ldots, N-1$, gives,

$$\widehat{\boldsymbol{w}} = \widehat{\boldsymbol{H}}\widehat{\boldsymbol{v}} + \boldsymbol{\Gamma}\widehat{\boldsymbol{x}}_0,$$

[5] Note that the auxiliary variables are scalar in the SISO case, thereby excluding the possibility of this change of basis.

where $\widehat{\boldsymbol{H}} \in \mathbb{R}^{2N \times N}$ is the matrix build from the Markov parameters of the model,

$$\widehat{\boldsymbol{H}} = \begin{bmatrix} \boldsymbol{D} & 0 & \dots & \dots & \dots & 0 \\ \boldsymbol{CB} & \boldsymbol{D} & 0 & \dots & \dots & 0 \\ \boldsymbol{CAB} & \boldsymbol{CB} & \boldsymbol{D} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & \ddots & 0 \\ \boldsymbol{CA}^{N-2}\boldsymbol{B} & \boldsymbol{CA}^{N-3}\boldsymbol{B} & \dots & \dots & \boldsymbol{CB} & \boldsymbol{D} \end{bmatrix} \tag{7.14}$$

and $\boldsymbol{\Gamma} \in \mathbb{R}^{2N \times n}$ denotes its (extended) observability matrix,

$$\boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{C}^{\mathsf{T}} & (\boldsymbol{CA})^{\mathsf{T}} & \dots & (\boldsymbol{CA}^{N-1})^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}.$$

Because of the isometry and minimality of (7.13) we know that the rank of $\left[\boldsymbol{\Gamma}, \widehat{\boldsymbol{H}}\right]$ is equal to $N+n$, indicating that the union of range $(\boldsymbol{\Gamma})$ and range $\left(\widehat{\boldsymbol{H}}\right)$ corresponds to the $(N+n)$-dimensional behavior of the model in (7.13),

$$\widehat{\boldsymbol{w}} \in \left[\text{range}\left(\widehat{\boldsymbol{H}}\right) \oplus \text{range}\left(\boldsymbol{\Gamma}\right)\right]. \tag{7.15}$$

7.3.2 Structured misfits

Equating the partial derivatives of the Lagrangian with respect to the state vectors $\widehat{\boldsymbol{x}}_k$ and the auxiliary inputs $\widehat{v}_k$ to zero for all $k = 0, \dots, N-1$, immediately shows that the sequence of Lagrange multipliers $\boldsymbol{\mu}_k$ equals the misfits $\widetilde{\boldsymbol{w}}_k = \boldsymbol{w}_k - \widehat{\boldsymbol{w}}_k$. In turn, we can use the partial derivatives with respect to the state $\widehat{\boldsymbol{x}}_{\boldsymbol{k}}$ and the auxiliary variables $\widehat{v}_k$ to show that the Lagrange multipliers $\boldsymbol{\lambda}_{\boldsymbol{k}}$ satisfy a particular backward recursion,

$$\begin{bmatrix} \partial\mathcal{L}/\partial\widehat{\boldsymbol{x}}_{\boldsymbol{k}} \\ \partial\mathcal{L}/\partial\widehat{v}_k \end{bmatrix} = \boldsymbol{0} \iff \begin{bmatrix} \boldsymbol{\lambda}_{\boldsymbol{k-1}} \\ 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}^{\mathsf{T}} & \boldsymbol{C}^{\mathsf{T}} \\ \boldsymbol{B}^{\mathsf{T}} & \boldsymbol{D}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda}_{\boldsymbol{k}} \\ \widetilde{\boldsymbol{w}}_{\boldsymbol{k}} \end{bmatrix}, \tag{7.16}$$

where we substituted $\boldsymbol{\mu}_{\boldsymbol{k}}$ by $\widetilde{\boldsymbol{w}}_{\boldsymbol{k}}$ and defined $\boldsymbol{\lambda}_{\boldsymbol{(-1)}} = \boldsymbol{0}$, so that the relation in (7.16) holds for $k = 0, ..., N-1$. Additionally, $\partial\mathcal{L}/\partial\widehat{\boldsymbol{x}}_{\boldsymbol{N}} = 0$ implies that $\boldsymbol{\lambda}_{\boldsymbol{N-1}} = 0$.

By exploiting the fact that (7.13) is isometric, all solutions of (7.16) can be described as,

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda}_{\boldsymbol{k-1}} \\ 0 \end{bmatrix} + \begin{bmatrix} \widetilde{\boldsymbol{B}} \\ \widetilde{\boldsymbol{D}} \end{bmatrix} \widetilde{v}_k = \begin{bmatrix} \boldsymbol{\lambda}_{\boldsymbol{k}} \\ \widetilde{\boldsymbol{w}}_{\boldsymbol{k}} \end{bmatrix} \tag{7.17}$$

where $\widetilde{\boldsymbol{v}} \in \mathbb{R}^N$ and the matrices $\widetilde{\boldsymbol{B}} \in \mathbb{R}^n$ and $\widetilde{\boldsymbol{D}} \in \mathbb{R}^2$ come from the unitary completion of the behavioral model in (7.13), such that,

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} & \widetilde{\boldsymbol{B}} \\ \boldsymbol{C} & \boldsymbol{D} & \widetilde{\boldsymbol{D}} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} & \widetilde{\boldsymbol{B}} \\ \boldsymbol{C} & \boldsymbol{D} & \widetilde{\boldsymbol{D}} \end{bmatrix}^{\mathsf{T}} = \boldsymbol{I}_{n+2}. \tag{7.18}$$

Substitution of $\widetilde{\boldsymbol{x}}_{\boldsymbol{k}} = \boldsymbol{\lambda}_{\boldsymbol{k-1}}$ for $k = 0, \ldots, N-1$, and rewriting (7.17) leads to the state-space representation of the *misfit model*,

$$\begin{bmatrix} \widetilde{\boldsymbol{x}}_{\boldsymbol{k+1}} \\ \widetilde{\boldsymbol{w}}_{\boldsymbol{k}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \widetilde{\boldsymbol{B}} \\ \boldsymbol{C} & \widetilde{\boldsymbol{D}} \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{x}}_{\boldsymbol{k}} \\ \widetilde{v}_k \end{bmatrix}, \quad k = 0, \ldots, N-1. \tag{7.19}$$

Similarly as in Section 7.2, we can conclude that the optimal misfit is highly structured: the misfit itself is generated by an nth order LTI model.

Equation (7.19) allows us to derive a basis for the linear subspace in which the misfit $\widetilde{\boldsymbol{w}}_{\boldsymbol{k}}$ resides. Contrary to (7.15), we do not need to consider the observability matrix of the misfit model because we know that $\widetilde{\boldsymbol{x}}_{\boldsymbol{0}} = \boldsymbol{0}$. That leaves us with the matrix $\widetilde{\boldsymbol{H}} \in \mathbb{R}^{2N \times N}$, which is defined similarly as in (7.14) but using the matrices $\boldsymbol{A}, \widetilde{\boldsymbol{B}}, \boldsymbol{C}$ and $\widetilde{\boldsymbol{D}}$. However, we cannot take any linear combination of its columns to generate $\widetilde{\boldsymbol{w}}$ as there is an additional constraint that needs to be satisfied: $\widetilde{\boldsymbol{x}}_{\boldsymbol{N}} = 0$. For this constraint to be true, $\widetilde{\boldsymbol{v}}$ needs to be an element of the null space of $\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}$, where $\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}} \in \mathbb{R}^{n \times N}$ is the reversed extended controllability matrix of the misfit model,

$$\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}} = \begin{bmatrix} \boldsymbol{A}^{N-1}\widetilde{\boldsymbol{B}} & \ldots & \boldsymbol{A}\widetilde{\boldsymbol{B}} & \widetilde{\boldsymbol{B}} \end{bmatrix}.$$

Minimality of the input-output relation in (7.10) implies minimality of (7.9). Therefore, knowing that the misfit model (7.19) is an isometric behavioral realization of (7.9), the matrix $\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}$ will be of rank n, such that its null space $\boldsymbol{Z} \in \mathbb{R}^{N \times (N-n)}$,

$$\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}} \boldsymbol{Z} = \boldsymbol{0},$$

is $(N-n)$-dimensional. Finally, we find that the $(N-n)$-dimensional misfit space corresponds to range $\left(\widetilde{\boldsymbol{H}}\boldsymbol{Z}\right)$.

We conclude that the first-order necessary conditions for optimality of the state-space projection onto the behavior problem once more induce an orthogonal decomposition of the ambient space,

$$\mathbb{R}^{2N} = \underbrace{\left(\text{range}\left(\boldsymbol{H}\right) \ \oplus \ \text{range}\left(\boldsymbol{\Gamma}\right)\right)}_{\text{behavior}} \oplus \underbrace{\text{range}\left(\widetilde{\boldsymbol{H}}\boldsymbol{Z}\right)}_{\text{misfits}}. \tag{7.20}$$

The orthogonal complementarity of both subspaces follows from the isometric construction of the model (7.13) and the misfit model (7.19).

7.3.3 Optimal projection

Because of (7.20), we can express the minimal norm $\widetilde{\boldsymbol{w}}$ as the orthogonal projection of $\boldsymbol{w}$ onto the column space of $\widetilde{\boldsymbol{H}}\boldsymbol{Z}$,

$$\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{H}}\boldsymbol{Z}[(\widetilde{\boldsymbol{H}}\boldsymbol{Z})^{\mathsf{T}}\widetilde{\boldsymbol{H}}\boldsymbol{Z}]^{-1}(\widetilde{\boldsymbol{H}}\boldsymbol{Z})^{\mathsf{T}}\boldsymbol{w} = \widetilde{\boldsymbol{H}}\boldsymbol{Z}(\widetilde{\boldsymbol{H}}\boldsymbol{Z})^{\mathsf{T}}\boldsymbol{w}, \tag{7.21}$$

where we used the fact that $\widetilde{\boldsymbol{H}}\boldsymbol{Z}$ is a semi-orthogonal matrix, such that: $(\widetilde{\boldsymbol{H}}\boldsymbol{Z})^{\mathsf{T}}\widetilde{\boldsymbol{H}}\boldsymbol{Z} = \boldsymbol{I}_{N-n}$.

From (7.19) and the fact that $\widetilde{\boldsymbol{x}}_{\boldsymbol{0}} = \boldsymbol{0}$, we know that $\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{v}}$. Combine this with (7.21) to see that we can compute the sequence $\widetilde{\boldsymbol{v}}$ which minimizes the 2-norm of $\widetilde{\boldsymbol{w}}$ via,

$$\widetilde{\boldsymbol{v}} = \boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}\widetilde{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{w}. \tag{7.22}$$

There is a alternative way to interpret (7.22). By combining (7.13) and (7.19) and taking $\boldsymbol{x}_{\boldsymbol{k+1}} = \widehat{\boldsymbol{x}}_{\boldsymbol{k+1}} + \widetilde{\boldsymbol{x}}_{\boldsymbol{k+1}}$, we get that,

$$\begin{bmatrix} \boldsymbol{x}_{\boldsymbol{k+1}} \\ \boldsymbol{w}_{\boldsymbol{k}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} & \widetilde{\boldsymbol{B}} \\ \boldsymbol{C} & \boldsymbol{D} & \widetilde{\boldsymbol{D}} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{\boldsymbol{k}} \\ \widehat{v}_k \\ \widetilde{v}_k \end{bmatrix}, \quad k = 0, \ldots, N-1.$$

We can now use the isometry (7.18) to derive that the sequences of auxiliary variables $\widehat{\boldsymbol{v}}$ and $\widetilde{\boldsymbol{v}}$ can be obtained from a backward recursion[6] using the observed data $\boldsymbol{w}$,

$$\begin{bmatrix} \boldsymbol{x}_k \\ \widehat{v}_k \\ \widetilde{v}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}^{\mathsf{T}} & \boldsymbol{C}^{\mathsf{T}} \\ \boldsymbol{B}^{\mathsf{T}} & \boldsymbol{D}^{\mathsf{T}} \\ \widetilde{\boldsymbol{B}}^{\mathsf{T}} & \widetilde{\boldsymbol{D}}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{\boldsymbol{k+1}} \\ \boldsymbol{w}_{\boldsymbol{k}} \end{bmatrix}, \quad k = 0, \ldots, N-1. \tag{7.23}$$

Writing out the backward recursion from (7.23) gives,

$$\widetilde{\boldsymbol{v}} = \widetilde{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{w} + (\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^{\mathsf{T}}\widehat{\boldsymbol{x}}_{\boldsymbol{N}}, \tag{7.24}$$

where we substituted $\boldsymbol{x}_{\boldsymbol{N}}$ with $\widehat{\boldsymbol{x}}_{\boldsymbol{N}}$ since $\widetilde{\boldsymbol{x}}_{\boldsymbol{N}} = 0$. Since $\|\widetilde{\boldsymbol{w}}\|_2^2 = \|\widetilde{\boldsymbol{v}}\|_2^2$ due to the isometry, the optimal $\widetilde{\boldsymbol{v}}$, which generates the minimal norm misfit $\widetilde{\boldsymbol{w}}$, can

[6] We call the state-space models

$$(\boldsymbol{A}^{\mathsf{T}}, \boldsymbol{C}^{\mathsf{T}}, \boldsymbol{B}^{\mathsf{T}}, \boldsymbol{D}^{\mathsf{T}}) \quad \text{and} \quad (\boldsymbol{A}^{\mathsf{T}}, \boldsymbol{C}^{\mathsf{T}}, \widetilde{\boldsymbol{B}}^{\mathsf{T}}, \widetilde{\boldsymbol{D}}^{\mathsf{T}}),$$

the *adjoint model* and *adjoint misfit model*, respectively.

be found by choosing $\widehat{\boldsymbol{x}}_{\boldsymbol{N}}$ such that $\|\widetilde{\boldsymbol{v}}\|_2^2$ is minimal,

$$\begin{aligned} \mathbf{0} &= \frac{\partial \|\widetilde{\boldsymbol{v}}\|_2^2}{\partial \widehat{\boldsymbol{x}}_{\boldsymbol{N}}} \\ \iff \mathbf{0} &= \widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^{\mathsf{T}}\widehat{\boldsymbol{x}}_{\boldsymbol{N}} + \widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}\widetilde{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{w} \\ \iff \widehat{\boldsymbol{x}}_{\boldsymbol{N}} &= -(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^{\mathsf{T}})^{-1}\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}\widetilde{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{w}, \end{aligned} \tag{7.25}$$

where we used that,

$$\|\widetilde{\boldsymbol{v}}\|_2^2 = \boldsymbol{w}^{\mathsf{T}}\widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{w} + \widehat{\boldsymbol{x}}_{\boldsymbol{N}}^{\mathsf{T}}\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^{\mathsf{T}}\widehat{\boldsymbol{x}}_{\boldsymbol{N}} + 2\boldsymbol{w}^{\mathsf{T}}\widetilde{\boldsymbol{H}}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^{\mathsf{T}}\widehat{\boldsymbol{x}}_{\boldsymbol{N}}.$$

Substituting (7.25) in (7.24) shows that the optimal $\widetilde{\boldsymbol{v}}$ is equal to,

$$\begin{aligned} \widetilde{\boldsymbol{v}} &= \widetilde{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{w} - (\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^{\mathsf{T}}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^{\mathsf{T}})^{-1}\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}\widetilde{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{w}, \\ &= \left(\boldsymbol{I} - (\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^{\mathsf{T}}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}})^{\mathsf{T}})^{-1}\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}\right)\widetilde{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{w}. \end{aligned}$$

Thus, the optimal $\widetilde{\boldsymbol{v}}$ can be obtained as the orthogonal projection of the backwards-filtered observed data $\boldsymbol{w}$ onto range$(\boldsymbol{Z})$, which corresponds to the orthogonal complement of row $\left(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{r}}\right)$, leading to the same conclusion as (7.22).

7.3.4 The z-domain

It now comes as no surprise that, similarly as in Section 7.2, the 'reversed polynomial' relation between the z-transforms of $\widetilde{\boldsymbol{y}}$ and $\widetilde{\boldsymbol{u}}$ is also found when one constructs the transfer matrices of the state-space models discussed above. It is straightforward to show that when the isometric model in (7.13) describes the same dynamics as the input-output relation in (7.1), the transfer matrix $\widehat{\boldsymbol{H}}(z)$ of the isometric model (7.13), mapping the z-transform of the auxiliary inputs $\boldsymbol{v}$ to the z-transform of the model-compliant trajectory $\widehat{\boldsymbol{w}}$, is given as,

$$\widehat{\boldsymbol{H}}(z) = \frac{c}{d(z)}\begin{bmatrix} b(z) \\ a(z) \end{bmatrix},$$

where $c \in \mathbb{R}$ and $d(z)$ a monic polynomial of degree n, satisfying $d(z)d_r(z) = c^2(a_r(z)a(z) + b_r(z)b(z))$. The latter ensures that $\widehat{\boldsymbol{H}}(z)$ is *lossless*, in the sense that,

$$\langle \widehat{\boldsymbol{H}}(z), \widehat{\boldsymbol{H}}(z) \rangle = 1,$$

with the inner-product defined as in (7.11). The transfer function of the misfit model (7.19) follows from the *lossless embedding* $\overline{\boldsymbol{H}}(z)$ [66] of the rational matrix

function $\widehat{\boldsymbol{H}}(z)$,

$$\overline{\boldsymbol{H}}(z) = \begin{bmatrix} \widehat{\boldsymbol{H}}(z) & \widetilde{\boldsymbol{H}}(z) \end{bmatrix} = \frac{c}{d(z)} \begin{bmatrix} b(z) & a_r(z) \\ a(z) & -b_r(z) \end{bmatrix},$$

such that,

$$\langle \overline{\boldsymbol{H}}(z), \overline{\boldsymbol{H}}(z) \rangle = \boldsymbol{I}_2.$$

Since $\widetilde{\boldsymbol{H}}(z)$ describes the mapping from the auxiliary variables to the misfit $\widetilde{\boldsymbol{w}}$, the mapping from $\widetilde{\boldsymbol{u}}$ to $\widetilde{\boldsymbol{y}}$ can be derived as:

$$\begin{bmatrix} \widetilde{Y}(z) \\ \widetilde{U}(z) \end{bmatrix} = \widetilde{\boldsymbol{H}}(z)\widetilde{V}(z) = \frac{c}{d(z)} \begin{bmatrix} a_r(z) \\ -b_r(z) \end{bmatrix} \widetilde{V}(z)$$

$$\iff \widetilde{Y}(z) = -\frac{a_r(z)}{b_r(z)} \widetilde{U}(z),$$

which agrees with our findings from Section 7.2.

7.4 Least Squares Realization for SISO LTI Models

This section, which was not part of the original publication, consider the full SISO LS realization problem, in which the model parameters themselves must also be estimated from observed sequences of input-output data. We begin by briefly revisiting the main results from the previous sections, but instead of using the newly introduced matrices $\widehat{\boldsymbol{T}}$ and $\widetilde{\boldsymbol{T}}$, we reformulate the results using the banded Toeplitz matrices $\boldsymbol{T}^{\boldsymbol{a}}_{N-n} \in \mathbb{R}^{(N-n)\times N}$ and $\boldsymbol{T}^{\boldsymbol{b}}_{N-n} \in \mathbb{R}^{(N-n)\times N}$, which have been used consistently throughout the previous chapters. We then show that all critical points of the problem can be characterized as solutions of a Riemannian SVD formulation. Finally, we propose and explore several possible approaches toward achieving a globally optimal methodology.

7.4.1 Notational reformulation

The model equations (7.1) can be expressed conveniently as

$$\begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-n} & -\boldsymbol{T}^{\boldsymbol{b}}_{N-n} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{y}} \\ \widehat{\boldsymbol{u}} \end{bmatrix} = \boldsymbol{0}. \tag{7.26}$$

This is equivalent to (7.4), up to a permutation of the columns. Consider $\widehat{\boldsymbol{Y}}_{\boldsymbol{N-n}}, \widehat{\boldsymbol{U}}_{\boldsymbol{N-n}} \in \mathbb{R}^{(N-n)\times(n+1)}$,the Hankel matrices constructed from the data

$\widehat{\boldsymbol{y}}$, $\widehat{\boldsymbol{u}}$, respectively, then (7.26) is equivalent to

$$\begin{bmatrix}\widehat{\boldsymbol{Y}}_{\boldsymbol{N-n}} & -\widehat{\boldsymbol{U}}_{\boldsymbol{N-n}}\end{bmatrix}\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} = \boldsymbol{0}. \tag{7.27}$$

This alternative notation allows for an elegant characterization of the complementarity between the subspaces spanned by the columns of $\widehat{\boldsymbol{T}}$ and $\widetilde{\boldsymbol{T}}$.

Lemma 7.1 (Commuting Banded Toeplitz (CBT))**.** *Consider the matrices* $\boldsymbol{T}_N^{\boldsymbol{a}}, \boldsymbol{T}_N^{\boldsymbol{b}} \in \mathbb{R}^{N\times(N+n)}$*, then,*

$$\begin{bmatrix}\boldsymbol{T}_{N-n}^{\boldsymbol{a}} & -\boldsymbol{T}_{N-n}^{\boldsymbol{b}}\end{bmatrix}\begin{bmatrix}\boldsymbol{T}_N^{\boldsymbol{b}}\\ \boldsymbol{T}_N^{\boldsymbol{a}}\end{bmatrix} = \boldsymbol{0}. \tag{7.28}$$

If the model in (7.1) *is minimal, such that the polynomials* $a(z)$ *and* $b(z)$ *are coprime, then,*

$$\operatorname{rank}\begin{bmatrix}(\boldsymbol{T}_N^{\boldsymbol{b}})^\mathsf{T} & (\boldsymbol{T}_N^{\boldsymbol{a}})^\mathsf{T}\end{bmatrix} = N+n,$$

such that the columns of the matrix $\begin{bmatrix}(\boldsymbol{T}_N^{\boldsymbol{b}})^\mathsf{T} & (\boldsymbol{T}_N^{\boldsymbol{a}})^\mathsf{T}\end{bmatrix}^\mathsf{T}$ *are linearly independent.*

Proof. The identity in (7.28) follows directly from writing out the matrix product on the left-hand side. Under the assumption of $a(z)$ and $b(z)$ being coprime, linear independence of the columns of $\begin{bmatrix}(\boldsymbol{T}_N^{\boldsymbol{b}})^\mathsf{T} & (\boldsymbol{T}_N^{\boldsymbol{a}})^\mathsf{T}\end{bmatrix}^\mathsf{T}$ follows from [113, Lemma 3.3]. □

Combining (7.26) and Lemma 7.1 implies the existence of a unique vector $\boldsymbol{z} = [z_0, \dots, z_{N+n-1}]^\mathsf{T} \in \mathbb{R}^{N+n}$ such that

$$\begin{bmatrix}\widehat{\boldsymbol{y}}\\ \widehat{\boldsymbol{u}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{T}_N^{\boldsymbol{a}}\\ \boldsymbol{T}_N^{\boldsymbol{b}}\end{bmatrix}\boldsymbol{z}, \tag{7.29}$$

indicating that the model-compliant data arises as the equation error of an unknown, *common driving variable* $\boldsymbol{z}$ with respect to the linear difference relations defined by $a(z)$ and $b(z)$.

Notice that, in contrast to the autonomous LTI case considered in Chapter 3, the SISO setting leads to an image representation (7.29) parametrized by the model parameters directly.

7.4.2 First-order optimality

We assume that the given data $\boldsymbol{y}, \boldsymbol{u}$ is not model-compliant, such that

$$\operatorname{rank}\begin{bmatrix}\boldsymbol{Y}_{\boldsymbol{N-n}} & \boldsymbol{U}_{\boldsymbol{N-n}}\end{bmatrix} = 2n+2. \tag{7.30}$$

For this to be satisfied, we need $N > 3n + 1$, since otherwise the double-Hankel matrix is not of full column rank by construction.

In the SISO LS realization problem, we want to identify the model parameters $\boldsymbol{a}, \boldsymbol{b}$ for which the optimal misfits $\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{u}}$, that modify these given data such that the modified data $\widehat{\boldsymbol{y}}, \widehat{\boldsymbol{u}}$ become model-compliant, have minimal 2-norm, leading to the optimization problem (7.2).

Consider the Lagrangian of (7.2),

$$\mathcal{L}(\boldsymbol{a}, \boldsymbol{b}, \widehat{\boldsymbol{y}}, \widehat{\boldsymbol{u}}, \boldsymbol{\ell}, \lambda) = J + \boldsymbol{\ell}^\mathsf{T} \left(\boldsymbol{T}^{\boldsymbol{a}}_{N-n} \widehat{\boldsymbol{y}} - \boldsymbol{T}^{\boldsymbol{b}}_{N-n} \widehat{\boldsymbol{u}} \right) + \lambda \left(g(\boldsymbol{a}, \boldsymbol{b}) - 1 \right), \tag{7.31}$$

with $\boldsymbol{\ell} = [\ell_0, \ldots, \ell_{N-n-1}]^\mathsf{T} \in \mathbb{R}^{N-n}$ and $\lambda \in \mathbb{R}$ Lagrange multipliers. The first-order optimality conditions (FONC) then correspond to

$$\frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{y}}} = -(\boldsymbol{y} - \widehat{\boldsymbol{y}}) + (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^\mathsf{T} \boldsymbol{\ell} = 0 \tag{7.32a}$$

$$\frac{\partial \mathcal{L}}{\partial \widehat{\boldsymbol{u}}} = -(\boldsymbol{u} - \widehat{\boldsymbol{u}}) + (\boldsymbol{T}^{\boldsymbol{b}}_{N-n})^\mathsf{T} \boldsymbol{\ell} = 0 \tag{7.32b}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{a}} = \widehat{\boldsymbol{Y}}^\mathsf{T}_{\boldsymbol{N-n}} \boldsymbol{\ell} + \lambda \frac{\partial g(\boldsymbol{a}, \boldsymbol{b})}{\partial \boldsymbol{a}} = \boldsymbol{T}^{\boldsymbol{\ell}}_{n+1} \widehat{\boldsymbol{y}} + \lambda \frac{\partial g(\boldsymbol{a}, \boldsymbol{b})}{\partial \boldsymbol{a}} = 0, \tag{7.32c}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}} = \widehat{\boldsymbol{U}}^\mathsf{T}_{\boldsymbol{N-n}} \boldsymbol{\ell} + \lambda \frac{\partial g(\boldsymbol{a}, \boldsymbol{b})}{\partial \boldsymbol{b}} = \boldsymbol{T}^{\boldsymbol{\ell}}_{n+1} \widehat{\boldsymbol{u}} + \lambda \frac{\partial g(\boldsymbol{a}, \boldsymbol{b})}{\partial \boldsymbol{b}} = 0, \tag{7.32d}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\ell}} = \boldsymbol{T}^{\boldsymbol{a}}_{N-n} \widehat{\boldsymbol{y}} - \boldsymbol{T}^{\boldsymbol{b}}_{N-n} \widehat{\boldsymbol{u}} = \boldsymbol{0}, \tag{7.32e}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = g(\boldsymbol{a}, \boldsymbol{b}) - 1 = 0. \tag{7.32f}$$

By combining (7.32c), (7.32d) and (7.32e), it is easily shown that at the critical points, $\lambda = 0$. This is true for both suggested choices of $g(\boldsymbol{a}, \boldsymbol{b})$.

Data optimality Let us briefly reconsider the results obtained in Section 7.2 and consider the implications of the vanishing of the partial derivatives of the Lagrangian with respect to $\widehat{\boldsymbol{y}}, \widehat{\boldsymbol{u}}$ and $\boldsymbol{\ell}$:

$$\partial \mathcal{L} / \partial \widehat{\boldsymbol{y}} = 0 \iff \widetilde{\boldsymbol{y}} = (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^\mathsf{T} \boldsymbol{\ell}, \quad \text{and} \quad \partial \mathcal{L} / \partial \widehat{\boldsymbol{u}} = 0 \iff \widetilde{\boldsymbol{u}} = -(\boldsymbol{T}^{\boldsymbol{b}}_{N-n})^\mathsf{T} \boldsymbol{\ell}.$$

Substitution of these expressions in the model equations gives

$$\begin{aligned} &\boldsymbol{T}^{\boldsymbol{a}}_{N-n}(\boldsymbol{y}-\widetilde{\boldsymbol{y}}) - \boldsymbol{T}^{\boldsymbol{b}}_{N-n}(\boldsymbol{u}-\widetilde{\boldsymbol{u}}) = \boldsymbol{0} \\ \iff &\begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-n} & -\boldsymbol{T}^{\boldsymbol{b}}_{N-n} \end{bmatrix} \left(\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{u} \end{bmatrix} - \begin{bmatrix} (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}} \\ -(\boldsymbol{T}^{\boldsymbol{b}}_{N-n})^{\mathsf{T}} \end{bmatrix} \boldsymbol{\ell} \right) = \boldsymbol{0}, \end{aligned}$$

such that, in the critical points of (7.2), the vector $\boldsymbol{\ell}$ corresponds to

$$\boldsymbol{\ell} = \left[\boldsymbol{T}^{\boldsymbol{a}}_{N-n} (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}} + \boldsymbol{T}^{\boldsymbol{b}}_{N-n} (\boldsymbol{T}^{\boldsymbol{b}}_{N-n})^{\mathsf{T}} \right]^{-1} \begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-n} & -\boldsymbol{T}^{\boldsymbol{b}}_{N-n} \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{u} \end{bmatrix}. \tag{7.33}$$

In turn, the optimal misfits correspond to

$$\begin{bmatrix} \widetilde{\boldsymbol{y}} \\ \widetilde{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-n} & -\boldsymbol{T}^{\boldsymbol{b}}_{N-n} \end{bmatrix}^{\dagger} \begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-n} & -\boldsymbol{T}^{\boldsymbol{b}}_{N-n} \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{u} \end{bmatrix}, \tag{7.34}$$

where

$$\begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-n} & -\boldsymbol{T}^{\boldsymbol{b}}_{N-n} \end{bmatrix}^{\dagger} = \begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-n} & -\boldsymbol{T}^{\boldsymbol{b}}_{N-n} \end{bmatrix}^{\mathsf{T}} \left(\left[\boldsymbol{T}^{\boldsymbol{a}}_{N-n} (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}} + \boldsymbol{T}^{\boldsymbol{b}}_{N-n} (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}} \right] \right)^{-1}.$$

This result is equivalent to the projection observed in (7.12). Combining (7.34) with Lemma 7.1 illustrates that

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{u} \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{T}^{\boldsymbol{b}}_{N} \\ \boldsymbol{T}^{\boldsymbol{a}}_{N} \end{bmatrix} \boldsymbol{z}}_{\text{model}} + \underbrace{\begin{bmatrix} (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}} \\ -(\boldsymbol{T}^{\boldsymbol{b}}_{N-n})^{\mathsf{T}} \end{bmatrix} \boldsymbol{\ell}}_{\text{misfit}},$$

where the optimal $\boldsymbol{z} \in \mathbb{R}^{N+n}$ is given as

$$\boldsymbol{z} = \left(\left[(\boldsymbol{T}^{\boldsymbol{b}}_{N})^{\mathsf{T}} \boldsymbol{T}^{\boldsymbol{b}}_{N} + (\boldsymbol{T}^{\boldsymbol{a}}_{N})^{\mathsf{T}} \boldsymbol{T}^{\boldsymbol{a}}_{N} \right] \right)^{-1} \begin{bmatrix} (\boldsymbol{T}^{\boldsymbol{b}}_{N})^{\mathsf{T}} & (\boldsymbol{T}^{\boldsymbol{a}}_{N})^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{u} \end{bmatrix}. \tag{7.35}$$

Because for any real-valued matrix $\boldsymbol{A}$:

$$\operatorname{rank} \boldsymbol{A} = \operatorname{rank} \boldsymbol{A}\boldsymbol{A}^{\mathsf{T}} = \operatorname{rank} \boldsymbol{A}^{\mathsf{T}}\boldsymbol{A},$$

we have

$$\operatorname{rank} \left(\begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-n} & -\boldsymbol{T}^{\boldsymbol{b}}_{N-n} \end{bmatrix} \begin{bmatrix} (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}} \\ -(\boldsymbol{T}^{\boldsymbol{b}}_{N-n})^{\mathsf{T}} \end{bmatrix} \right) = N - n$$

and, under the assumption that $a(z)$ and $b(z)$ are coprime,

$$\operatorname{rank} \left(\begin{bmatrix} (\boldsymbol{T}^{\boldsymbol{b}}_{N})^{\mathsf{T}} & (\boldsymbol{T}^{\boldsymbol{a}}_{N})^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{T}^{\boldsymbol{b}}_{N} \\ \boldsymbol{T}^{\boldsymbol{a}}_{N} \end{bmatrix} \right) = N + n,$$

such that the matrix inverses in (7.34) and (7.35) are well-defined.

Model optimality Requiring the partial derivatives of the Lagrangian (7.31) with respect to the model parameters to vanish, leads to

$$\boldsymbol{\ell}^{\mathsf{T}} \begin{bmatrix} \widehat{\boldsymbol{Y}}_{\boldsymbol{N-n}} & \widehat{\boldsymbol{U}}_{\boldsymbol{N-n}} \end{bmatrix} = \mathbf{0}, \tag{7.36}$$

indicating that, in the critical points, $\boldsymbol{\ell}$ resides in the left null space of the model-compliant data double-Hankel matrix. The rank of the $(N-n)\times(2n+2)$ double-Hankel matrix is upper-bounded by $2n+1$, since the model compliancy of the data ensures that there exists a nontrivial vector in its null space, see e.g., (7.27). As such, its left null space has dimension at least $N-3n-1$. These are novel insights, not encountered previously in Sections 7.2-7.3

Consider the banded-Toeplitz matrix $\boldsymbol{T}_{n+1}^{\boldsymbol{\ell}} \in \mathbb{R}^{(n+1)\times N}$ generated by $\boldsymbol{\ell}$, then (7.36) is equivalent to

$$\boldsymbol{T}_{n+1}^{\boldsymbol{\ell}} \widehat{\boldsymbol{y}} = \boldsymbol{T}_{n+1}^{\boldsymbol{\ell}} \widehat{\boldsymbol{u}} = \mathbf{0}.$$

It appears that in the critical points, both $\widehat{\boldsymbol{y}}$ and $\widehat{\boldsymbol{u}}$ correspond to the output of an $(N-n-1)$'th order autonomous LTI model determined by the characteristic polynomial $\ell(z)$, constructed from the coefficients $\boldsymbol{\ell}$.

Remark 7.2. *In Chapter 3, the optimality of the model parameters allowed us to derive the double-FIR-filter property (cf. Theorem 3.3). This property exploits the observation that, in the case of autonomous LTI models, the left null space of the model-compliant data Hankel matrix $\widehat{\boldsymbol{Y}}$ is spanned by $N-2n$ forward shifts of the model parameter vector $\boldsymbol{a}$, so that the Lagrange multipliers $\boldsymbol{\ell}$ can be expressed as a linear combination of these forward shifts.*

The characterization proved to be advantageous for two reasons: 1. For generic observed data $\boldsymbol{y}$, it removed the so-called 'lower-order' critical points from the solution set; 2. It allowed to eliminate the $N-n$ variables $\boldsymbol{\ell}$, replacing them with $\boldsymbol{g}$, containing only $N-2n$ variables.

In the SISO setting, however, it is unclear whether the analogous relation (7.36) *can be exploited to achieve similar benefits. At first sight, it does not appear possible to construct a basis for the left null space of the double-Hankel matrix, in which $\boldsymbol{\ell}$ resides, using a structured matrix composed from the model parameters $\boldsymbol{a}$ and $\boldsymbol{b}$.*

The novel image representation ($\widehat{\boldsymbol{y}} = \boldsymbol{T}_N^{\boldsymbol{b}}\boldsymbol{z}$ and $\widehat{\boldsymbol{u}} = \boldsymbol{T}_N^{\boldsymbol{a}}\boldsymbol{z}$) leads to an alternative formulation of (7.36). It is straightforward, though notationally tedious, to show that the following relations hold:

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{N-n}} = \boldsymbol{T}_{N-n}^{\boldsymbol{b}} \boldsymbol{Z}_{\boldsymbol{N}} \quad \text{and} \quad \widehat{\boldsymbol{U}}_{\boldsymbol{N-n}} = \boldsymbol{T}_{N-n}^{\boldsymbol{a}} \boldsymbol{Z}_{\boldsymbol{N}}, \tag{7.37}$$

where $\boldsymbol{Z_N} \in \mathbb{R}^{N\times(n+1)}$ corresponds to the Hankel matrix constructed from $\boldsymbol{z}$:

$$\boldsymbol{Z_N} = \begin{bmatrix} z_0 & z_1 & \cdots & z_n \\ z_1 & z_2 & \cdots & z_n \\ \vdots & \vdots & & \vdots \\ z_{N-1} & z_N & \cdots & z_{N+n-1} \end{bmatrix}.$$

It follows from (7.37), that (7.36) is equivalent to:

$$\boldsymbol{Z_N^\mathsf{T}}(\boldsymbol{T_{N-n}^b})^\mathsf{T}\boldsymbol{\ell} = -\boldsymbol{Z_N^\mathsf{T}}\widetilde{\boldsymbol{u}} = \boldsymbol{0}, \tag{7.38}$$

$$\boldsymbol{Z_N^\mathsf{T}}(\boldsymbol{T_{N-n}^a})^\mathsf{T}\boldsymbol{\ell} = \boldsymbol{Z_N^\mathsf{T}}\widetilde{\boldsymbol{y}} = \boldsymbol{0}. \tag{7.39}$$

We will make use of these expressions in the next subsection.

7.4.3 Riemannian SVD

The SISO LS realization can be naturally interpreted as a structured low-rank approximation (SLRA) problem (cf. Section 3.7). Given a full-rank double-Hankel matrix constructed from the observed data (7.30), the objective is to compute the closest, in the least squares sense, structured low-rank double-Hankel matrix:

$$\operatorname{rank}\begin{bmatrix}\widehat{\boldsymbol{Y}} & \widehat{\boldsymbol{U}}\end{bmatrix} \leq 2n+1. \tag{7.40}$$

This SLRA interpretation lies at the basis of the Riemannian Singular Value Decomposition (SVD), which, combined with the standard variable metric inverse iteration algorithm, provides an heuristic approach for solving (7.2), see e.g., [38, 46, 39] and related work.

After briefly revisiting the Riemannian SVD proposed in [46, 39], we derive an alternative Riemannian SVD-like formulation that builds on the novel image representation developed throughout this chapter.

Kernel representation We exploit the properties of the critical points, which are enforced by the first-order optimality conditions (7.32), to eliminate the variables $\widehat{\boldsymbol{y}}$ and $\widehat{\boldsymbol{u}}$ from the obtained system of equations. By rewriting the Toeplitz-vector product in (7.33) as a Hankel-vector product, we get

$$\begin{bmatrix}\boldsymbol{Y_{N-n}} & \boldsymbol{U_{N-n}}\end{bmatrix}\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} = \begin{bmatrix}\boldsymbol{T_{N-n}^a} & -\boldsymbol{T_{N-n}^b}\end{bmatrix}\begin{bmatrix}(\boldsymbol{T_{N-n}^a})^\mathsf{T}\\ -(\boldsymbol{T_{N-n}^b})^\mathsf{T}\end{bmatrix}\boldsymbol{\ell}. \tag{7.41}$$

On the other hand, (7.36) implies:

$$\begin{aligned} \boldsymbol{T}_{n+1}^{\boldsymbol{\ell}} \widehat{\boldsymbol{y}} = \boldsymbol{T}_{n+1}^{\boldsymbol{\ell}} \left[\boldsymbol{y} - (\boldsymbol{T}_{n+1}^{\boldsymbol{\ell}})^{\mathsf{T}} \boldsymbol{a} \right] &= \boldsymbol{0}, \\ \iff \boldsymbol{T}_{n+1}^{\boldsymbol{\ell}} \boldsymbol{y} = \boldsymbol{T}_{n+1}^{\boldsymbol{\ell}} (\boldsymbol{T}_{n+1}^{\boldsymbol{\ell}})^{\mathsf{T}} \boldsymbol{a}, \\ \iff \boldsymbol{Y}_{\boldsymbol{N-n}}^{\mathsf{T}} \boldsymbol{\ell} = \boldsymbol{T}_{n+1}^{\boldsymbol{\ell}} (\boldsymbol{T}_{n+1}^{\boldsymbol{\ell}})^{\mathsf{T}} \boldsymbol{a}, \end{aligned} \tag{7.42}$$

where we used $\widetilde{\boldsymbol{y}} = (\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{T}} \boldsymbol{\ell} = (\boldsymbol{T}_{n+1}^{\boldsymbol{\ell}})^{\mathsf{T}} \boldsymbol{a}$. A similar derivation leads to

$$\boldsymbol{U}_{\boldsymbol{N-n}}^{\mathsf{T}} \boldsymbol{\ell} = -\boldsymbol{T}_{n+1}^{\boldsymbol{\ell}} (\boldsymbol{T}_{n+1}^{\boldsymbol{\ell}})^{\mathsf{T}} \boldsymbol{b}, \tag{7.43}$$

such that the combination of (7.42) and (7.43) gives

$$\begin{bmatrix} \boldsymbol{Y}_{\boldsymbol{N-n}}^{\mathsf{T}} \\ \boldsymbol{U}_{\boldsymbol{N-n}}^{\mathsf{T}} \end{bmatrix} \boldsymbol{\ell} = \begin{bmatrix} \boldsymbol{T}_{n+1}^{\boldsymbol{\ell}} (\boldsymbol{T}_{n+1}^{\boldsymbol{\ell}})^{\mathsf{T}} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{T}_{n+1}^{\boldsymbol{\ell}} (\boldsymbol{T}_{n+1}^{\boldsymbol{\ell}})^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix}. \tag{7.44}$$

Expressions (7.41) and (7.44), which describe a system of $N + n + 2$ cubic polynomial equations in the $N + n + 2$ variables $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\ell})$, form the basis of the Riemannian SVD proposed in [46, Theorem 1].

Image representation Consider the extended data vectors $\boldsymbol{y_e}, \boldsymbol{u_e} \in \mathbb{R}^{N+2n}$, obtained by appending n zeros at the start and end of the given data $\boldsymbol{y}, \boldsymbol{u}$, respectively, e.g.,

$$\boldsymbol{y_e} = \begin{bmatrix} 0 & \dots & 0 & \boldsymbol{y}^{\mathsf{T}} & 0 & \dots & 0 \end{bmatrix}^{\mathsf{T}}.$$

From (7.35), we get

$$\begin{bmatrix} \boldsymbol{Y_e} & \boldsymbol{U_e} \end{bmatrix} \begin{bmatrix} \boldsymbol{b_r} \\ \boldsymbol{a_r} \end{bmatrix} = \left[(\boldsymbol{T}_N^{\boldsymbol{b}})^{\mathsf{T}} \boldsymbol{T}_N^{\boldsymbol{b}} + (\boldsymbol{T}_N^{\boldsymbol{a}})^{\mathsf{T}} \boldsymbol{T}_N^{\boldsymbol{a}} \right] \boldsymbol{z}, \tag{7.45}$$

where $\boldsymbol{Y_e}, \boldsymbol{U_e} \in \mathbb{R}^{(N+n)\times(n+1)}$ are Hankel matrices constructed from the extended data $\boldsymbol{y_e}$, $\boldsymbol{u_e}$ and $\boldsymbol{a_r}, \boldsymbol{b_r}$ correspond to the coefficient-reversed parameter vectors $\mathrm{rev}(\boldsymbol{a}) = [a_n, a_{n-1}, \dots, a_0]$ and $\mathrm{rev}(\boldsymbol{b})$, respectively. This established the first set of equations compromising the novel Riemannian SVD.

It follows from the optimality of the model parameters, expressed in (7.38), that

$$\boldsymbol{Z}_{\boldsymbol{N}}^{\mathsf{T}} \widetilde{\boldsymbol{u}} = \boldsymbol{Z}_{\boldsymbol{N}}^{\mathsf{T}} [\boldsymbol{u} - \boldsymbol{T}_N^{\boldsymbol{a}} \boldsymbol{z}] = \boldsymbol{0} \iff \boldsymbol{Z}_{\boldsymbol{N}}^{\mathsf{T}} \boldsymbol{u} = \boldsymbol{Z}_{\boldsymbol{N}}^{\mathsf{T}} \boldsymbol{Z}_{\boldsymbol{N}} \boldsymbol{a}. \tag{7.46}$$

where we used (7.29). Similarly, it follows from (7.39) that

$$\boldsymbol{Z}_{\boldsymbol{N}}^{\mathsf{T}} \widetilde{\boldsymbol{y}} = \boldsymbol{Z}_{\boldsymbol{N}}^{\mathsf{T}} [\boldsymbol{y} - \boldsymbol{T}_N^{\boldsymbol{b}} \boldsymbol{z}] = \boldsymbol{0} \iff \boldsymbol{Z}_{\boldsymbol{N}}^{\mathsf{T}} \boldsymbol{y} = \boldsymbol{Z}_{\boldsymbol{N}}^{\mathsf{T}} \boldsymbol{Z}_{\boldsymbol{N}} \boldsymbol{b}, \tag{7.47}$$

The combination of (7.46) and (7.47), in which we rewrite the Hankel-vector product as a Toeplitz-vector product, gives

$$\begin{bmatrix} \boldsymbol{T}_{n+1}^{\boldsymbol{y}} \\ \boldsymbol{T}_{n+1}^{\boldsymbol{u}} \end{bmatrix} \boldsymbol{z} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{Z}_{\boldsymbol{N}}^{\mathsf{T}} \boldsymbol{Z}_{\boldsymbol{N}} \\ \boldsymbol{Z}_{\boldsymbol{N}}^{\mathsf{T}} \boldsymbol{Z}_{\boldsymbol{N}} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix}, \tag{7.48}$$

where $\boldsymbol{T}_{n+1}^{\boldsymbol{u}}, \boldsymbol{T}_{n+1}^{\boldsymbol{y}} \in \mathbb{R}^{(n+1)\times N+n}$ are banded-Toeplitz.

We can now exploit the fact that

$$\boldsymbol{P}\boldsymbol{T}_{n+1}^{\boldsymbol{u}} = \boldsymbol{U}_{\boldsymbol{e}}^{\mathsf{T}},$$

where the matrix $\boldsymbol{P} \in \mathbb{R}^{(n+1)\times(n+1)}$ is the following row-perturbation matrix:

$$\boldsymbol{P} = \begin{bmatrix} & & & 1 \\ & & 1 & \\ & \iddots & & \\ 1 & & & \end{bmatrix}.$$

Pre-multiplication of (7.48) with $\boldsymbol{P}$ gives us the second set of equations

$$\begin{bmatrix} \boldsymbol{Y}_{\boldsymbol{e}}^{\mathsf{T}} \\ \boldsymbol{U}_{\boldsymbol{e}}^{\mathsf{T}} \end{bmatrix} \boldsymbol{z} = \begin{bmatrix} \boldsymbol{P}\boldsymbol{Z}_{\boldsymbol{N}}^{\mathsf{T}} \boldsymbol{Z}_{\boldsymbol{N}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}\boldsymbol{Z}_{\boldsymbol{N}}^{\mathsf{T}} \boldsymbol{Z}_{\boldsymbol{N}} \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{a} \end{bmatrix}$$

$$\iff \begin{bmatrix} \boldsymbol{Y}_{\boldsymbol{e}}^{\mathsf{T}} \\ \boldsymbol{U}_{\boldsymbol{e}}^{\mathsf{T}} \end{bmatrix} \boldsymbol{z} = \begin{bmatrix} \boldsymbol{P}\boldsymbol{Z}_{\boldsymbol{N}}^{\mathsf{T}} \boldsymbol{Z}_{\boldsymbol{N}} \boldsymbol{P} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}\boldsymbol{Z}_{\boldsymbol{N}}^{\mathsf{T}} \boldsymbol{Z}_{\boldsymbol{N}} \boldsymbol{P} \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_{\boldsymbol{r}} \\ \boldsymbol{a}_{\boldsymbol{r}} \end{bmatrix}, \tag{7.49}$$

where we used that $\boldsymbol{P}\boldsymbol{P} = \boldsymbol{I}$.

The combination of (7.45) and (7.49), which, up to the author's knowledge, has not yet appeared in the literature, clearly has the structure of a Riemannian SVD. It corresponds to a system of $N + 3n + 2$ cubic polynomial equations in the $N+3n+2$ variables $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{z})$, the real-valued solutions of which characterize the critical points of the SISO LS realization problem.

A (computational) investigation into the applicability of the inverse-iteration algorithm [38, 46, 40] to this novel Riemannian SVD formulation is left for future work. Instead, the next subsection attempts to formulate a globally optimal methodology for the SISO LS realization problem.

7.4.4 Global optimality

Due to the non-convex nature of the misfit modeling problem (7.2), the solution approaches described in the literature [38, 39, 143, 126, 125] are heuristic in

Table 7.1: Model parameter values $\boldsymbol{a}$ and $\boldsymbol{b}$ of the real-valued solutions of the system of polynomial equations described in Example 7.3, computed using the homotopy continuation implementation from [20] and normalized by enforcing $a_1 = 1$. The system has 36 solutions, 6 of which real-valued. For each solution, σ_i denotes the singular values of the corresponding block Hankel matrix constructed from the model-compliant data.

i	$\|\widetilde{\boldsymbol{w}}\|_2^2$	a_0	b_0	b_1	σ_i
1	0.031444	0.71678	-1.6872	-2.096	$\{0.2207, 0.101, 0.00175, 1.386 \times 10^{-19}\}$
2	2.2471	0.076922	8.1474	4.8605	$\{1.789, 0.7543, 0.03141, 8.4 \times 10^{-18}\}$
3	6.3402	-1.5025	0.10402	0.030477	$\{2.902, 1.636, 0.1001, 1.684 \times 10^{-17}\}$
4	10.42	1.7576	1.1326	0.15318	$\{3.836, 2.16, 0.2192, 7.306 \times 10^{-17}\}$
5	11.37	6.647	1.1542	1.8941	$\{3.754, 2.732, 0.1368, 1.44 \times 10^{-16}\}$
6	12.505	0.42644	0.55571	0.2344	$\{3.775, 2.816, 1.487, 4.546 \times 10^{-16}\}$

the sense that convergence to the globally optimal model(s) can generally not be guaranteed. In line with the overarching objective of this dissertation, we investigate whether the framework developed in this chapter can be used to devise a globally optimal approach for solving the SISO LS realization problem. We discuss several alternative strategies aimed at computing the entire set of critical points.

Riemannian SVD The most straightforward route to global optimality is to compute all common roots to the system of equations resulting from one of the Riemannian SVD formulations introduced earlier. Among these, the kernel-based formulation, parametrized in terms of the variables $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{\ell}$, is the more attractive option from a computational point of view, as it involves comparatively the smallest number of variables.

Example 7.3. *Consider the observed data ($N = 5$):*

$$\boldsymbol{y} = \begin{pmatrix} -0.458 \\ 0.2801 \\ 0.2059 \\ 3.1133 \\ 0.4679 \end{pmatrix}, \quad \textit{and} \quad \boldsymbol{u} = \begin{pmatrix} -1.1439 \\ 0.7251 \\ -0.9888 \\ -0.8094 \\ -0.6277 \end{pmatrix}.$$

For $n = 1$, the system of equations obtained from (7.44) and (7.45), in which we substituted $a_1 = 1$, consists of 8 polynomial equations in the variables a_0, b_0, b_1 and $\ell_0, \ell_1, \ell_2, \ell_3$. The real-valued solutions of this system of cubic polynomial equations, corresponding to the critical points of this instance of the SISO LS realization problem, are depicted in Table 7.1.

Example 7.4. *Consider the observed data ($N = 8$):*

$$\boldsymbol{y} = \begin{pmatrix} 0.039268 \\ 0.91689 \\ -1.3062 \\ -0.16728 \\ -0.082792 \\ 0.60245 \\ -0.49965 \\ 0.60873 \end{pmatrix}, \quad \textit{and} \quad \boldsymbol{u} = \begin{pmatrix} -0.11724 \\ 0.1858 \\ 0.025232 \\ 0.1186 \\ 0.7998 \\ 0.056306 \\ 1.2748 \\ -0.49047 \end{pmatrix}.$$

For $n = 2$, the system of equations obtained from (7.44) and (7.45), in which we substituted $a_2 = 1$, consists of 12 polynomial equations in the variables $a_0, a_1, b_0, b_1 b_2$ and $\boldsymbol{\ell} \in \mathbb{R}^6$. The real-valued solutions, corresponding to the critical points of this instance of the SISO LS realization problem, are depicted in Table 7.2.

For $n = 1$ and $n = 2$, our numerical experiments indicate that the corresponding Riemannian SVD equations have finitely many solutions over $\mathbb{C}$. While this is encouraging, the question of whether finiteness holds for arbitrary model orders n remains open. Establishing such a result would require an algebraic characterization of the solution set of the system of equations. Similarly to the analysis we did in Chapters 5 and 6, one could attempt to derive a structural description of the variety defined by these equations, identify possible degenerate components, and, if feasible, determine the exact number of complex solutions in the case of generic data $\boldsymbol{y}, \boldsymbol{u}$. This definitely constitutes an interesting direction for future research.

An interesting observation that can be drawn from Tables 7.1 and 7.2 concerns the model-compliant data associated to each of the obtained critical points. The right-most column of the tables indicates that, at least for these two examples, the phenomenon of so-called 'lower-order' dynamics that we encountered in Chapter 3, does not occur in the SISO case. Indeed, for all critical points, equality holds in (7.40).

The drawback of both Riemannian SVD formulations, however, is that all variables appear nonlinearly in (some of) the equations. This makes it impossible to immediately eliminate a subset of the variables (those appearing linearly only) from the problem setup by reformulating the system of equations to an MEP. As a result, the computational complexity of solving the system grows rapidly with increasing problem sizes (N, n), due to the (relatively) large number of variables involved.

Table 7.2: Model parameter values $\boldsymbol{a}$ and $\boldsymbol{b}$ of the real-valued solutions of the system of polynomial equations described in Example 7.4, computed using the homotopy continuation implementation from [20] and normalized by enforcing $a_2 = 1$. The system has 902 solutions, 42 of which real-valued. For each solution, σ_i denotes the two smallest singular values of the corresponding block Hankel matrix constructed from the model-compliant data.

i	$\|\widetilde{\boldsymbol{w}}\|_2^2$	a_0	a_1	b_0	b_1	b_2	σ_i
1	0.00067662	1.6495	0.66952	4.5434	−0.2784	−3.0143	$\{0.881, 1.287 \times 10^{-16}\}$
2	0.62845	0.32408	0.78785	−0.39069	0.49202	0.041491	$\{0.4415, 8.731 \times 10^{-17}\}$
3	0.85238	1.0557	−0.11082	−0.97199	1.6764	−1.0849	$\{0.6005, 7.05 \times 10^{-17}\}$
4	0.85372	0.92606	0.043018	−0.94089	1.4696	−0.79991	$\{0.602, 3.325 \times 10^{-17}\}$
5	1.1226	−0.8991	1.2701	−2.226	2.7733	3.8358	$\{0.005507, 3.515 \times 10^{-17}\}$
6	1.6803	−1.5034	−0.61302	0.37706	0.31791	0.030817	$\{0.1757, 1.924 \times 10^{-16}\}$
7	1.7167	−2.4183	−1.7123	1.111	1.4553	0.45568	$\{0.396, 1.695 \times 10^{-16}\}$
8	1.7249	−0.88837	0.012508	0.069949	0.14249	−0.056713	$\{0.05707, 1.756 \times 10^{-16}\}$
9	1.8053	−0.89802	0.012485	0.19613	0.46612	0.20419	$\{0.2692, 1.867 \times 10^{-17}\}$
10	1.8876	−2.8486	−0.23745	1.1354	−0.99572	0.83938	$\{0.2774, 8.966 \times 10^{-17}\}$
11	1.9068	−6.7446	−2.7386	2.0634	−2.5886	1.2434	$\{0.1449, 7.195 \times 10^{-19}\}$
12	2.0362	−0.45855	1.7266	0.31161	−1.0497	0.65831	$\{0.1464, 2.687 \times 10^{-16}\}$
13	2.0672	−0.18557	−1.4138	−0.27961	3.2267	−0.60852	$\{0.1039, 1.507 \times 10^{-16}\}$
14	2.0789	−1.2275	1.2771	0.50472	−1.049	0.69614	$\{0.03017, 4.56 \times 10^{-17}\}$
15	2.0963	0.14406	1.9794	−0.29294	−1.7452	0.071394	$\{0.04562, 1.259 \times 10^{-16}\}$
16	2.1003	−0.50322	−0.48624	−0.21407	1.3626	−0.27913	$\{0.2156, 5.132 \times 10^{-17}\}$
17	2.1582	−1.533	3.4758	1.7811	−2.2666	4.5067	$\{0.2016, 1.448 \times 10^{-16}\}$
18	2.1635	−2.1235	3.3077	1.8359	−2.9878	4.3233	$\{0.1251, 5.13 \times 10^{-17}\}$
19	2.3788	−1.028	−0.29316	−0.13337	−0.53716	−0.64875	$\{0.3063, 8.486 \times 10^{-17}\}$
20	2.461	−1.7075	3.8254	4.6336	2.9233	4.9461	$\{0.03367, 3.906 \times 10^{-17}\}$
21	2.4882	3.8284	13.27	8.0239	3.1775	11.044	$\{0.06408, 7.677 \times 10^{-17}\}$
22	2.489	1.5006	7.4606	4.753	2.3476	6.0506	$\{0.05441, 1.549 \times 10^{-16}\}$
23	2.5472	8.322	0.0047581	−7.5018	−3.2366	−6.7006	$\{0.06271, 1.377 \times 10^{-16}\}$
24	2.6369	−1.7386	0.51622	0.39299	−1.8464	−0.15532	$\{0.148, 3.653 \times 10^{-17}\}$
25	2.9342	−0.41334	−0.82779	−0.78589	0.62321	−1.3263	$\{0.04317, 6.703 \times 10^{-17}\}$
26	3.5095	0.87029	−2.2397	−0.095545	−0.87257	1.1739	$\{0.3303, 8.851 \times 10^{-17}\}$
27	3.5491	1.4926	−4.7454	0.23634	−1.3025	1.6221	$\{0.2324, 2.75 \times 10^{-16}\}$
28	3.7158	1.0641	−0.46054	−1.066	−1.0328	1.6663	$\{0.3664, 1.653 \times 10^{-16}\}$
29	3.7282	0.83998	−0.15314	−1.4548	−0.8983	1.8489	$\{0.2516, 4.008 \times 10^{-17}\}$
30	3.7821	0.7703	−0.74322	−0.1691	−0.801	0.85154	$\{0.4173, 1.077 \times 10^{-16}\}$
31	3.7979	8.2378	0.022901	−4.357	−4.7949	5.0453	$\{0.03622, 5.441 \times 10^{-17}\}$
32	3.8003	7.0082	1.2537	−4.5227	−3.8481	4.7582	$\{0.04058, 3.846 \times 10^{-17}\}$
33	3.8886	2.2765	−1.3152	−0.4678	−1.7714	0.83803	$\{0.1283, 8.431 \times 10^{-17}\}$
34	3.9609	4.0927	2.3263	0.18782	3.2928	5.0782	$\{0.002441, 1.433 \times 10^{-16}\}$
35	4.0564	−0.073732	−0.49372	−0.028298	−0.30806	0.52445	$\{0.003297, 1.777 \times 10^{-16}\}$
36	4.0792	0.01641	−0.70619	0.077784	−0.067316	0.81679	$\{0.01177, 1.15 \times 10^{-16}\}$
37	4.0812	−0.043818	−0.61536	0.0063584	−0.13204	0.75066	$\{0.003691, 1.821 \times 10^{-16}\}$
38	4.6029	0.30762	−1.2969	0.28189	−0.79229	0.097862	$\{0.03214, 3.966 \times 10^{-17}\}$
39	4.6099	0.556	−1.5288	0.3949	−0.80898	0.11257	$\{0.002091, 1.586 \times 10^{-16}\}$
40	4.6116	0.057568	−1.038	0.1007	−0.76835	0.10249	$\{0.02366, 1.857 \times 10^{-16}\}$
41	4.8594	1.1298	1.6416	0.14251	0.42952	0.56407	$\{0.0001614, 3.894 \times 10^{-17}\}$
42	4.8982	2.22	−1.8022	1.1569	−0.62205	1.8325	$\{0.01561, 2.698 \times 10^{-17}\}$

In the remainder of this chapter, we explore alternative approaches, aiming either to reduce the number of variables in the system to be solved or to obtain an MEP-based characterization of the critical points.

Double-FIR-filter-like property In the autonomous LTI setting, the *double-FIR-filter property* allowed the formulation of an MEP-based characterization of all critical points (cf. Chapter 3). Ideally, we would seek a similar property in the SISO setting, allowing to construct a (rectangular) multiparameter eigenvalue (MEP) characterization describing (a subset of) the critical points of the SISO LS realization problem. However, despite various attempts, we have thus far been unable to identify such a property in the SISO setting. The absence of a closed-form relation expressing the left null space of the double-Hankel matrix, appearing in (7.36), in terms of the model parameters $\boldsymbol{a}$ and $\boldsymbol{b}$ is the key obstacle in extending the MEP-based approach for autonomous LTI models presented in this dissertation to the more general SISO setting.

Elimination of the Lagrange multipliers In a master's thesis project [97], we explored whether the Lagrange multipliers $\boldsymbol{\ell}$ in the system of equations derived from the FONC could be eliminated analytically. While such an elimination is indeed possible, the resulting reformulation did not yield a satisfactory new method for computing the complete set of critical points. The limitation is that, over the complex field $\mathbb{C}$, the resulting system of polynomial equations has a positive-dimensional solution set[7].

In what follows, we briefly describe this approach and point to the origin of this positive dimensionality. For more details on these findings, we refer the reader to [97, Ch. 4].

To ease the derivations, we introduce the short-hand notation

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-n} & -\boldsymbol{T}^{\boldsymbol{b}}_{N-n} \end{bmatrix} \begin{bmatrix} (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}} & -(\boldsymbol{T}^{\boldsymbol{b}}_{N-n})^{\mathsf{T}} \end{bmatrix}$$

and

$$\boldsymbol{e} = \begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-n} & -\boldsymbol{T}^{\boldsymbol{b}}_{N-n} \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{u} \end{bmatrix}, \tag{7.50}$$

so that (7.33) becomes $\boldsymbol{\ell} = \boldsymbol{D}^{-1}\boldsymbol{e}$. The variables $\boldsymbol{\ell}$ can be eliminated from the FONC (7.32) by substituting this expression into (7.32c) and (7.32d). This

[7] Algebraic polynomial and MEP solvers are generally applicable only to systems with a finite number of isolated solutions. Homotopy continuation methods can, in principle, still be applied in this context, as they can track solution paths converging to generic points on each component of a positive-dimensional variety. Nevertheless, the presence of such components complicates the interpretation of the computed solutions.

leads to

$$\begin{bmatrix} \frac{\partial \mathcal{L}}{\partial \boldsymbol{a}} = \left\{ \boldsymbol{e}^{\mathsf{T}} \boldsymbol{D}^{-1} \dfrac{\partial (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})}{\partial a_i} \left[\boldsymbol{y} - (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}} \boldsymbol{D}^{-1} \boldsymbol{e} \right] \right\}_{i=0,\ldots,n} \\ \frac{\partial \mathcal{L}}{\partial \boldsymbol{b}} = \left\{ \boldsymbol{e}^{\mathsf{T}} \boldsymbol{D}^{-1} \dfrac{\partial (\boldsymbol{T}^{\boldsymbol{b}}_{N-n})}{\partial b_i} \left[\boldsymbol{u} + (\boldsymbol{T}^{\boldsymbol{b}}_{N-n})^{\mathsf{T}} \boldsymbol{D}^{-1} \boldsymbol{e} \right] \right\}_{i=0,\ldots,n} \end{bmatrix} = \boldsymbol{0},$$

which, using $\boldsymbol{D}^{-1} = \operatorname{adj}(\boldsymbol{D}) / \det(\boldsymbol{D})$, corresponds to a system of $2(n+1)$ rational equations in the variables $\boldsymbol{a}$ and $\boldsymbol{b}$. After multiplying out the denominators, we obtain the following system of polynomial equations:

$$\begin{bmatrix} \left\{ \boldsymbol{e}^{\mathsf{T}} \operatorname{adj}(\boldsymbol{D}) \dfrac{\partial (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})}{\partial a_i} \left[\det(\boldsymbol{D}) \boldsymbol{y} - (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{T}} \operatorname{adj}(\boldsymbol{D}) \boldsymbol{e} \right] \right\}_{i=0,\ldots,n} \\ \left\{ \boldsymbol{e}^{\mathsf{T}} \operatorname{adj}(\boldsymbol{D}) \dfrac{\partial (\boldsymbol{T}^{\boldsymbol{b}}_{N-n})}{\partial b_i} \left[\det(\boldsymbol{D}) \boldsymbol{u} + (\boldsymbol{T}^{\boldsymbol{b}}_{N-n})^{\mathsf{T}} \operatorname{adj}(\boldsymbol{D}) \boldsymbol{e} \right] \right\}_{i=0,\ldots,n} \end{bmatrix} = \boldsymbol{0}, \quad (7.51)$$

the real-valued solutions of which characterize the critical points of the SISO LS realization problem.

For real-valued, nonzero parameters $\boldsymbol{a}, \boldsymbol{b}$, the matrix $\boldsymbol{D}$ is always positive definite. However, if $\boldsymbol{a}, \boldsymbol{b}$ are allowed to take values in $\mathbb{C}$, this property no longer holds, and $\boldsymbol{D}$ can become singular or even vanish entirely.

Example 7.5. *Consider the smallest nontrivial case $(N, n) = (5, 1)$, for which*

$$\boldsymbol{D} = \begin{bmatrix} \sum_{i=0}^{1} (a_i^2 + b_i^2) & a_0 a_1 + b_0 b_1 & & \\ a_0 a_1 + b_0 b_1 & \sum_{i=0}^{1} (a_i^2 + b_i^2) & a_0 a_1 + b_0 b_1 & \\ & a_0 a_1 + b_0 b_1 & \sum_{i=0}^{1} (a_i^2 + b_i^2) & a_0 a_1 + b_0 b_1 \\ & & a_0 a_1 + b_0 b_1 & \sum_{i=0}^{1} (a_i^2 + b_i^2) \end{bmatrix}.$$

Assume that we normalize the model parameters by setting $a_1 = 1$, then, there is a one-dimensional set in $\mathbb{C}^3$, corresponding to the common roots of

$$(b_0^2 + 1)(b_1^2 + 1) = 0 \quad \textit{and} \quad a_0 = -b_0 b_1,$$

which describes all triples (a_0, b_0, b_1) for which $\boldsymbol{D}$ becomes the zero matrix. One simple example is $(a_0, b_0, b_1) = (0, j, 0)$.

Consequently, over $\mathbb{C}$, there exists a positive-dimensional set for which $\boldsymbol{D} = \boldsymbol{0}$. Since the adjugate of the zero matrix is itself the zero matrix, it follows that this entire component is contained in the solution set of (7.51).

A possible workaround is to explicitly remove these degenerate cases by appending the following polynomial constraint to the system (7.51):

$$\det(\boldsymbol{D_{N-n}}) \cdot \phi - 1 = 0, \quad (7.52)$$

for which a new auxiliary variable $\phi \in \mathbb{C}$ has to be introduced. This constraint enforces $\det(\boldsymbol{D_{N-n}}) \neq 0$, thereby guaranteeing that $\boldsymbol{D_{N-n}}$ is nonsingular. It is an application of Rabinowitsch's trick[8], which expresses the inequality $\det(\boldsymbol{D_{N-n}}) \neq 0$ in purely algebraic form. From a geometric viewpoint, (7.52) maps the positive-dimensional components corresponding to $\det(\boldsymbol{D_{N-n}}) = 0$ to infinity in the affine space along the direction of ϕ. These "solutions at infinity" can then be safely discarded, for example, by removing divergent paths in homotopy continuation or by a column compression in the block-Macaulay framework.

While conceptually clean, appending (7.52) to the system of equations has computational drawbacks: it increases the number of decision variables by one, and the determinant is typically a very high-degree polynomial, making the system significantly more sensitive to perturbations and numerical inaccuracies.

Linearization In the autonomous LTI setting, [45] suggests an MEP formulation that provides an alternative to the approach developed in Chapter 3. The key insight in that work was to linearize the FONC by introducing (many) auxiliary variables, thereby converting it into a set of polynomial equations in which only the model parameters $\boldsymbol{a}$ appeared nonlinearly. The remaining variables, appearing solely in linear terms, could then be absorbed into the eigenvector, giving an MEP formulation. While theoretically elegant, Chapter 3 demonstrated a major limitation of this approach: the introduction of the auxiliary variables results in large coefficient matrices in the MEP, limiting its computational practicality.

The same linearization strategy can be applied in the SISO setting. We omit the full derivation of the MEP here and focus instead on demonstrating that it also admits a positive-dimensional solution set. More details can be found the master's thesis project [97].

[8]The introduction of an auxiliary variable to encode an inequality constraint as a polynomial identity, known as Rabinowitsch's trick, was first employed by Rabinowitsch in his proof of Hilbert's Nullstellensatz [140]. See e.g., [33, Ch. 2, Theorem 2] for more information.

Applying the linearization strategy proposed in [45] to the FONC (7.32) gives the following MEP:

$$\underbrace{\begin{bmatrix} \boldsymbol{D} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{e} \\ \left[\frac{\partial(\boldsymbol{D})}{\partial a_i}\right]_i & \boldsymbol{I}\otimes\boldsymbol{D} & \boldsymbol{0} & \left[\frac{\partial(\boldsymbol{T}^{\boldsymbol{a}}_{N-n})}{\partial a_i}\boldsymbol{y}\right]_i \\ \left[\frac{\partial(\boldsymbol{D})}{\partial b_i}\right]_i & \boldsymbol{0} & \boldsymbol{I}\otimes\boldsymbol{D} & -\left[\frac{\partial(\boldsymbol{T}^{\boldsymbol{b}}_{N-n})}{\partial b_i}\boldsymbol{y}\right]_i \\ \left[(\frac{\partial(\boldsymbol{T}^{\boldsymbol{a}}_{N-n})}{\partial a_i}\boldsymbol{y})^{\mathsf{T}}\right]_i & \boldsymbol{I}\otimes\boldsymbol{e}^{\mathsf{T}} & \boldsymbol{0} & \boldsymbol{0} \\ \left[(\frac{\partial(\boldsymbol{T}^{\boldsymbol{b}}_{N-n})}{\partial b_i}\boldsymbol{y})^{\mathsf{T}}\right]_i & \boldsymbol{0} & \boldsymbol{I}\otimes\boldsymbol{e}^{\mathsf{T}} & \boldsymbol{0} \end{bmatrix}}_{\boldsymbol{M}(\boldsymbol{a},\boldsymbol{b})} \underbrace{\begin{bmatrix} \boldsymbol{\ell} \\ \boldsymbol{g_1} \\ \boldsymbol{g_2} \\ -1 \end{bmatrix}}_{\boldsymbol{v}} = \boldsymbol{0}, \tag{7.53}$$

where $[\dots]_i$ denotes vertical stacking over $i = 0, \dots, n$, $\boldsymbol{I}$ is the identity matrix, and $\otimes$ denotes the Kronecker product.

In deriving the MEP, we implicitly assume that the last entry of the eigenvector $\boldsymbol{v}$ is nonzero, ensuring that the eigenvector can always be appropriately normalized to retrieve the Lagrange multipliers $\boldsymbol{\ell}$. However, for a subset of the eigentuples, which are the tuples $\boldsymbol{a}, \boldsymbol{b}$ for which $\boldsymbol{M}(\boldsymbol{a}, \boldsymbol{b})$ drops rank, the corresponding eigenvectors violate this assumption of the last element of $\boldsymbol{v}$ begin nonzero. We can refer to this subset of eigentuples as the 'bad locus', in analogy with the bad loci encountered in Chapters 5 and 6.

Unfortunately, this particular bad locus can be shown to be positive-dimensional. The positive dimensionality is caused by the same set of (complex-valued) parameter tuples $(\boldsymbol{a}, \boldsymbol{b})$ as encountered before: those for which the matrix $\boldsymbol{D}$ vanishes identically[9]. For every such tuple $(\boldsymbol{a}, \boldsymbol{b})$, there exists a family of eigenvectors of the form

$$\boldsymbol{v} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{g_1}^{\mathsf{T}} & \boldsymbol{g_2}^{\mathsf{T}} & \boldsymbol{0} \end{bmatrix}^{\mathsf{T}} \tag{7.54}$$

that satisfy (7.53). Indeed, if $\boldsymbol{D} = \boldsymbol{0}$ and $\boldsymbol{v}$ is chosen as in (7.54), the MEP (7.53) imposes the following constraints on the remaining elements of $\boldsymbol{v}$:

$$\begin{cases} \boldsymbol{I}\otimes\boldsymbol{e}^{\mathsf{T}}\boldsymbol{g_1} = 0, \\ \boldsymbol{I}\otimes\boldsymbol{e}^{\mathsf{T}}\boldsymbol{g_2} = 0, \end{cases}$$

which are satisfied for all $\boldsymbol{g_1}$ and $\boldsymbol{g_2}$ corresponding to vertical stacks of vectors orthogonal to the equation error $\boldsymbol{e}$, from (7.50).

[9]We have not carried out a comprehensive analysis of the set of eigentuples, so it is possible that it contains additional positive-dimensional components beyond the one described here.

Consequently, despite the fact that the linearization procedure from [45] extends directly to the SISO case, the reformulation introduces a positive dimensional "spurious' component into the solution set of the obtained MEP.

7.5 Conclusions and Future Work

In the first part of the chapter, we considered the projection onto the behavior for linear time-invariant (LTI) single-input single-output (SISO) models, which arises in the context of misfit modeling. We characterized the orthogonal decomposition of the ambient data space, which is induced by the least squares optimality criterion, using banded block-Toeplitz matrices (Section 7.2). Then, in Section 7.3, we showed that similar results are obtained when an (isometric) behavioral state-space representation is used to model the SISO LTI dynamics instead of the difference equation that was used in Section 7.2, thereby establishing two equivalent (up to a change in the model parametrization) descriptions of the optimal solution. Depending on the particular use-case, one of these two alternative formulations can be preferred: e.g., the difference equation considered in Section 7.2 has, if one fixes $a_0 = 1$, a minimal number of parameters $(2n+1)$ which can be favorable in the context of optimization. On the other hand, whereas a state-space model is inherently overparametrized, the characterization developed in Section 7.3 allows for a more intuitive interpretation of the results compared to the formulation given in Section 7.2.

The second part of this section examined the SISO LS realization problem, mostly from the perspective of global optimization. The kernel-based Riemannian SVD formulation emerged as the most promising approach. For $n = 1$ and $n = 2$, numerical experiments suggest a finite number of solutions over $\mathbb{C}$; whether this holds for general n remains open. Unlike in the autonomous LTI case, no lower-order dynamics were observed: in all computed examples, the double-Hankel matrix at each critical point attained the target rank $2n + 1$. The main limitation of the Riemannian SVD approach is that all variables appear nonlinearly, preventing direct MEP reformulation. Furthermore, the relatively large number of variables causes the computational complexity to increase rapidly with the problem size (N, n).

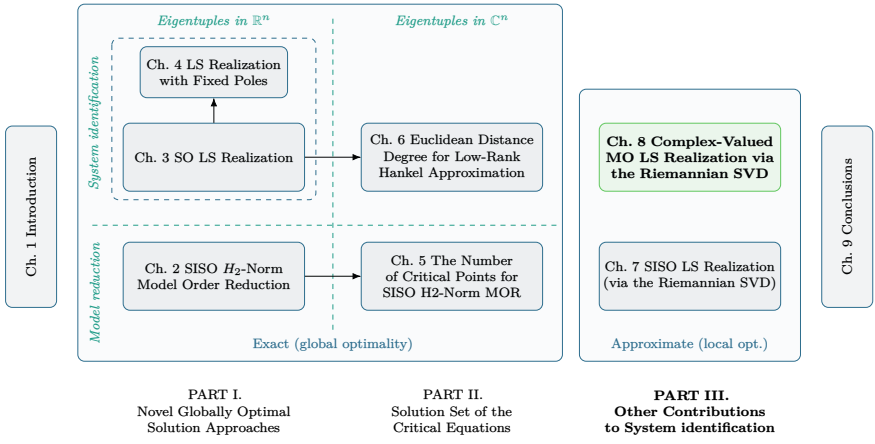
Several alternative strategies were considered, but each encountered significant obstacles. The double-FIR-filter property from the autonomous case could not be extended to SISO due to the absence of a closed-form basis for the relevant double-Hankel null space. Eliminating the Lagrange multipliers produces polynomial systems with positive-dimensional solution sets over $\mathbb{C}$, and while Rabinowitsch-type constraints can remove these components, they add

substantial computational overhead. Likewise, auxiliary-variable linearization leads to an MEP but introduces spurious positive-dimensional components that limit its usefulness.

In future work, we want to continue the exploratory investigation presented in Section 7.4.4 in an attempt to develop a methodology that allows to identify the globally optimal solutions of (7.2). Another potential direction for future work is to determine whether the set of critical points is finite for arbitrary problem sizes (N, n), potentially through an algebro-geometric analysis of the complex solution set of the kernel Riemannian SVD formulation, along similar lines as the work in Part II of this dissertation. Lastly, since the characterization obtained in [144] holds for the broader class of MIMO models, we believe that we must be able to generalize the 'banded block-Toeplitz' framework presented in Section 7.2 from SISO to MIMO models.

Chapter 8

Least Squares Realization of Complex-Valued Multiple-Output LTI Systems via the Riemannian Singular Value Decomposition



Chapter abstract

We present a generalization of least squares realization of multiple-output linear time-invariant models to the complex domain, thereby enabling broader applicability in areas such as signal processing, control, and system identification. We discuss the challenges inherent to optimization over complex variables and formulate the realization problem as a structured total least squares problem. A key contribution is the derivation of a complex-valued nonlinear generalized singular value decomposition. We adapt an existing heuristic algorithm that works over the real domain to solve the complex nonlinear generalized singular value decomposition. We validate our framework through illustrative examples and demonstrate its use in a nuclear magnetic resonance modeling problem and a direction of arrival estimation problem.

Candidate's contributions and relation to published work

This work is accepted for publication in *IEEE Control Systems Letters*:

- Viroux, S., **Lagauw, S.**, and De Moor, B. (2025). "Least Squares Realization of Complex-Valued Multiple-Output Autonomous LTI Systems via the Riemannian Singular Value Decomposition". *IEEE Control Systems Letters*, 9, pp. 2181–2186.

The candidate was a co-author of the original article, contributing equally with Sem Viroux to the research and writing.

This chapter differs slightly from the original article:

- The notation and terminology was adapted to ensure consistency with the conventions used throughout the remainder of this dissertation.
- Remarks 8.1 and 8.2, about the convergence behavior of the proposed algorithm and related topics, were added.

8.1 Introduction

We consider complex-valued multiple-output autonomous linear time-invariant (LTI) causal lumped parameter dynamical models in discrete-time, with a pre-specified model order n. The output data $\widehat{\boldsymbol{y}}_i = [\widehat{y}_{i,0} \ \cdots \ \widehat{y}_{i,N-1}]^{\mathsf{T}} \in \mathbb{C}^N$ (with $N > n$) satisfy a difference equation of the form:

$$a_0\widehat{y}_{i,k} + a_1\widehat{y}_{i,k+1} + \cdots + a_n\widehat{y}_{i,k+n} = \boldsymbol{a}(z)\widehat{y}_{i,k} = 0, \tag{8.1}$$

where $\widehat{\boldsymbol{y}}_i$ is the ith output signal and $\boldsymbol{a}(z)$ is a degree n polynomial in the forward-shift operator z (i.e., $zy_{i,k} = y_{i,k+1}$), the coefficients of which correspond to the complex model parameters $\boldsymbol{a} = [a_0 \ a_1 \ \cdots \ a_{n-1} \ a_n]^{\mathsf{T}} \in \mathbb{C}^{n+1}$. The data $\widehat{\boldsymbol{y}}_i$ are said to be *model-compliant* since they satisfy the difference equation (8.1) exactly for all $k = 0, \ldots, N{-}n{-}1$.

To ensure uniqueness of the parametrization and without loss of generality, we fix $a_n = 1$ for the remainder of this chapter, so that the model $a(z)$ is completely determined by n parameters. We will therefore denote the vector of unknown parameters as $\boldsymbol{a}_{\mathrm{tr}} = [a_0 \ a_1 \ \cdots \ a_{n-1}]^{\mathsf{T}} \in \mathbb{C}^n$.

In the least squares (LS) realization problem, one is given p sequences of data $\boldsymbol{y}_i \in \mathbb{C}^N$ that do not comply with (8.1). The objective is to find the parameters $\boldsymbol{a}$ that minimize the 2-norm of the *misfits* $\widetilde{\boldsymbol{y}}_i = \boldsymbol{y}_i - \widehat{\boldsymbol{y}}_i$, which is the modification needed to make the data $\boldsymbol{y}$ model-compliant (see Figure 8.1):

$$\underset{\substack{\widehat{\boldsymbol{y}}_1,\ldots,\widehat{\boldsymbol{y}}_p \,\in\, \mathbb{C}^N \\ \boldsymbol{a}_{\mathrm{tr}} \,\in\, \mathbb{C}^n}}{\text{minimize}} \quad \frac{1}{2}\sum_{i=1}^{p} \|\boldsymbol{y}_i - \widehat{\boldsymbol{y}}_i\|_2^2 = \frac{1}{2}\sum_{i=1}^{p} \|\widetilde{\boldsymbol{y}}_i\|_2^2\,, \tag{8.2a}$$

$$\text{subject to} \quad \boldsymbol{a}(z)\widehat{y}_{i,k} = 0, \ k = 0, \ldots, N{-}n{-}1. \tag{8.2b}$$

From a systems-theoretic perspective, the difference relation (8.1) is equivalent to the characteristic (or minimal) polynomial of a linear system, and thus relates subsequent samples of model-compliant output data from an autonomous LTI system or an LTI impulse response. The literature about estimating the coefficients of (8.1) from data–in a real-valued setting–is vast, see e.g., [45, 8] and references therein. The problem traces back to the work of Prony [48] and has since appeared in numerous formulations across signal processing, control, and system identification. The estimation problem (8.2) can be interpreted as a structured [38, 44] and/or constrained [1] total least squares (TLS) problem. This connection to low-rank matrix approximation was already implicit in the early system identification literature [118]. When a-priori statistical assumptions are imposed on the misfits, (8.2) naturally leads to a maximum likelihood setting,

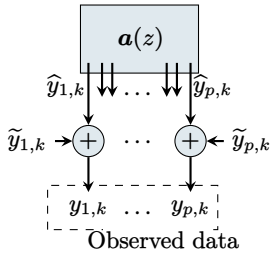


Figure 8.1: Schematic overview of the autonomous LTI multiple-output model. The given output data $\boldsymbol{y}_i \in \mathbb{C}^N$ are modified using the so-called misfits $\widetilde{\boldsymbol{y}}_i \in \mathbb{C}^N$, so that the modified data $\widehat{\boldsymbol{y}}_i \in \mathbb{C}^N$ are model-compliant. The coefficients $\boldsymbol{a}_{\mathrm{tr}} \in \mathbb{C}^n$ of the nth degree polynomial $\boldsymbol{a}(z)$ in the forward-shift operator z, are the unknown model parameters.

see e.g., dynamic errors in variables modeling [154] and the iterative quadratic maximum likelihood methods [21, 116].

Systems theory and identification is, however, not inherently limited to a real-valued setting, and working with *complex-valued signals and model parameters* in (8.2) allows for a broader class of applications. Indeed, complex-valued signals arise naturally in numerous domains, including digital communication applications [160], nuclear magnetic resonance (NMR) modeling [172], rotor bearing systems identification [137], and direction of arrival (DOA) estimation in array processing [148]. Recently, complex-valued representations have also been explored in the neural network community [112]. For a rigorous treatment of signal processing of complex-valued data, we refer the reader to [151].

In this chapter, we extend the structured TLS framework previously developed for real-valued problems [38, 44] to the complex domain, resulting in a novel methodology for computing locally optimal solutions to the complex-valued LS realization problem (8.2). This constitutes our *main contribution.* In Section 8.2, we address the key intricacies associated with optimization over complex decision variables. Characterizing first-order optimality leads to a nonlinear generalized singular value decomposition (coined Riemannian SVD) in the complex field, which is presented in Section 8.3. Section 8.4 introduces an extension of the variable metric inverse iteration algorithm to the complex domain. In Section 8.5, we work out two numerical examples, drawn from NMR modeling and DOA estimation, to validate the proposed methodology and illustrate its performance. We conclude with remarks and future work in Section 8.6.

Notation We denote scalars, vectors and matrices by lowercase, boldface lowercase and boldface uppercase letters, respectively. We use j to denote the imaginary unit $\sqrt{-1}$. The complex conjugate, transpose, and Hermitian transpose of a complex vector $\boldsymbol{x} = \boldsymbol{x}^{\mathrm{Re}} + \mathrm{j}\boldsymbol{x}^{\mathrm{Im}} \in \mathbb{C}^d$ are indicated by $\overline{\boldsymbol{x}}$, $\boldsymbol{x}^{\mathsf{T}}$, and $\boldsymbol{x}^{\mathsf{H}}$, respectively. Subscript i and k of a time series denote the output and time indices, respectively, e.g., $y_{i,k}$.

8.2 Optimization over Complex Decision Variables

In this section we give a brief overview of the technicalities involved with equality-constrained optimization over complex decision variables. We first note that in the context of minimization over complex decision variables, the objective function must be real ($f : \mathbb{C}^d \to \mathbb{R} : \boldsymbol{x} \mapsto f(\boldsymbol{x}), d \in \mathbb{Z}_+$). Indeed, for the optimization problem to be well-defined, $f(\boldsymbol{x})$ must be such that there exists an ordering on its function values[1]: one has to be able to conclude whether a solution $\boldsymbol{x}_1$ performs *better* than another solution $\boldsymbol{x}_2$ by means of the objective function value, e.g., $f(\boldsymbol{x}_1) < f(\boldsymbol{x}_2)$.

The first-order necessary conditions for optimality (FONC) of an optimization problem are obtained via the first-order derivatives of the corresponding Lagrangian. The *Cauchy–Riemann equations* serve as a necessary and sufficient condition for a function to be *analytic*, which means that its derivative is well defined for all points in the complex domain. Since real-valued objective functions, except for $f(\boldsymbol{x}) = c \in \mathbb{R}$, do not satisfy the Cauchy–Riemann equations, they are not complex differentiable.

Thus, we need a manner of redefining differentiation of a real-valued function $f(\boldsymbol{x})$ over complex variables $\boldsymbol{x} \in \mathbb{C}^d$. The most straightforward way of doing this, is by decoupling the derivatives w.r.t. the real and the imaginary parts, such that the real-valued objective function $f(\boldsymbol{x})$ can be interpreted as a function of two independent real variables $f(\boldsymbol{x}^{\mathrm{Re}}, \boldsymbol{x}^{\mathrm{Im}})$. This approach makes the problem completely real-valued, allowing for the use of real differential calculus to formulate the FONC for optimality.

The aforementioned reformulation of the optimization problem over complex variables into a real-valued problem is, however, often impractical, since it tends to obscure or destroy the structure of the original complex-variable problem.

[1] Even though it would be possible to impose an ordering on complex numbers which allows for a convenient comparison between two complex numbers, concepts such as the gradient and higher order derivatives would lose their meaning. The problem is that, although the complex numbers can be ordered in various ways, they cannot be made into an ordered field with the usual addition and multiplication operators [104].

An alternative approach involves the use of $\mathbb{CR}$-calculus, also referred to as *Wirtinger derivatives* [104], which provides an equivalent but more elegant formulation of the FONC. This approach is based on the observation that a function $f(\boldsymbol{x}^{\mathrm{Re}}, \boldsymbol{x}^{\mathrm{Im}})$ that is analytic over the vector space $\mathbb{R}^{d\times 2}$, is also analytic over the vector space $\{(\boldsymbol{x}, \overline{\boldsymbol{x}})\} \subset \mathbb{C}^{d\times 2}$.

For such a *polyanalytic* function $f(\boldsymbol{x}, \overline{\boldsymbol{x}})$ the Wirtinger derivatives towards $\boldsymbol{x}$ and $\overline{\boldsymbol{x}}$ are defined via the chain rule:

$$\frac{\partial f(\boldsymbol{x}, \overline{\boldsymbol{x}})}{\partial \boldsymbol{x}} = \frac{\partial f(\boldsymbol{x}^{\mathrm{Re}}, \boldsymbol{x}^{\mathrm{Im}})}{\partial \boldsymbol{x}^{\mathrm{Re}}} \frac{\partial \boldsymbol{x}^{\mathrm{Re}}}{\partial \boldsymbol{x}} + \frac{\partial f(\boldsymbol{x}^{\mathrm{Re}}, \boldsymbol{x}^{\mathrm{Im}})}{\partial \boldsymbol{x}^{\mathrm{Im}}} \frac{\partial \boldsymbol{x}^{\mathrm{Im}}}{\partial \boldsymbol{x}},$$

$$\frac{\partial f(\boldsymbol{x}, \overline{\boldsymbol{x}})}{\partial \overline{\boldsymbol{x}}} = \frac{\partial f(\boldsymbol{x}^{\mathrm{Re}}, \boldsymbol{x}^{\mathrm{Im}})}{\partial \boldsymbol{x}^{\mathrm{Re}}} \frac{\partial \boldsymbol{x}^{\mathrm{Re}}}{\partial \overline{\boldsymbol{x}}} + \frac{\partial f(\boldsymbol{x}^{\mathrm{Re}}, \boldsymbol{x}^{\mathrm{Im}})}{\partial \boldsymbol{x}^{\mathrm{Im}}} \frac{\partial \boldsymbol{x}^{\mathrm{Im}}}{\partial \overline{\boldsymbol{x}}},$$

which in practice boils down to differentiation of $f(\boldsymbol{x}, \overline{\boldsymbol{x}})$ with respect to $\boldsymbol{x}$ (resp. $\overline{\boldsymbol{x}}$), while considering the variable $\overline{\boldsymbol{x}}$ (resp. $\boldsymbol{x}$) as a constant, i.e., $\boldsymbol{x}$ and $\overline{\boldsymbol{x}}$ are treated as independent variables for the purpose of differentiation.

Nevertheless, one must be careful when working with constrained optimization problems over complex variables [104, 151]. To construct the Lagrangian when using Wirtinger derivatives, the original constraint as well as its complex conjugate must be incorporated into the Lagrangian using a single set of complex-conjugated Lagrange multipliers.

8.3 Complex Riemannian SVD (RiSVD)

8.3.1 Model-compliance

Writing out (8.1) for all $k = 0, \ldots, N-n-1$, illustrates that constraint (8.2b) of the realization problem can be expressed as

$$\begin{bmatrix} \widehat{\boldsymbol{Y}}_1 \\ \widehat{\boldsymbol{Y}}_2 \\ \vdots \\ \widehat{\boldsymbol{Y}}_p \end{bmatrix} \boldsymbol{a} = \begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-n} & & & \\ & \boldsymbol{T}^{\boldsymbol{a}}_{N-n} & & \\ & & \ddots & \\ & & & \boldsymbol{T}^{\boldsymbol{a}}_{N-n} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{y}}_1 \\ \widehat{\boldsymbol{y}}_2 \\ \vdots \\ \widehat{\boldsymbol{y}}_p \end{bmatrix} = \boldsymbol{0}, \tag{8.3}$$

where $\widehat{\boldsymbol{Y}}_i \in \mathbb{C}^{(N-n)\times(n+1)}$ is the Hankel matrix constructed from the elements of $\widehat{\boldsymbol{y}}_i$, and $\boldsymbol{T}^{\boldsymbol{a}}_{N-n} \in \mathbb{C}^{(N-n)\times N}$ is a banded-Toeplitz matrix, where the first row of $\boldsymbol{T}^{\boldsymbol{a}}_{N-n}$ is given by the model parameters $\boldsymbol{a}$ appended with $N-n-1$ zeros, and all elements below the diagonal are zero. Notice the equality between

the Hankel-vector products and the (banded) Toeplitz-vector products in (8.3). This reformulation follows from the Hankel/Toeplitz structure of the involved matrices and will be used extensively throughout this chapter.

Relation (8.3) illustrates that the model parameters $\boldsymbol{a}$ associated to a known sequence of model-compliant output data can be retrieved via the null space of the model-compliant data matrix: $\widehat{\boldsymbol{Y}} = [\widehat{\boldsymbol{Y}}_1^{\mathsf{T}} \ \cdots \ \widehat{\boldsymbol{Y}}_p^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{p(N-n)\times(n+1)}$. Conversely, when the model parameters $\boldsymbol{a}$ are known, the *behavior* of the model[2] corresponds to the (pn)-dimensional null space of the block-diagonal matrix constructed from $\boldsymbol{T}_{N-n}^{\boldsymbol{a}}$.

The given data $\boldsymbol{y}$, however, are not compliant with the model class under consideration, such that the given data block-Hankel matrix $\boldsymbol{Y} = [\boldsymbol{Y}_1^{\mathsf{T}} \ \cdots \ \boldsymbol{Y}_p^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{p(N-n)\times(n+1)}$, where each $\boldsymbol{Y}_i$ is a Hankel matrix, is of full column rank[3].

The identification problem (8.2) corresponds to a structured low-rank approximation problem: for a given data block-Hankel matrix $\boldsymbol{Y}$ that is of full column-rank, i.e., containing data that is not compliant with the model class under consideration, one wants to identify the best approximating, in a (weighted) least squares sense, low-rank block-Hankel matrix $\widehat{\boldsymbol{Y}}$. This rank-deficiency $\operatorname{rank}\widehat{\boldsymbol{Y}} \leq n$ is enforced by the constraint (8.2b). If $\operatorname{rank}\widehat{\boldsymbol{Y}} = n$, the model parameters corresponding to the model-compliant data $\widehat{\boldsymbol{y}}$ are unique.

8.3.2 Data optimality

We approach the realization problem (8.2) by first rewriting the constraints (8.2b) in terms of the Hankel matrices, as in (8.3), and constructing the Lagrangian, for which we rely on the concepts presented in Section 8.2.

$$\mathcal{L}(\boldsymbol{a}_{\mathrm{tr}}, \widehat{\boldsymbol{y}}, \boldsymbol{\ell}) = \sum_{i=1}^{p} \left[(\boldsymbol{y}_i - \widehat{\boldsymbol{y}}_i)^{\mathsf{H}} (\boldsymbol{y}_i - \widehat{\boldsymbol{y}}_i) + \boldsymbol{\ell}_i^{\mathsf{H}} \boldsymbol{T}_{N-n}^{\boldsymbol{a}} \widehat{\boldsymbol{y}}_i + \boldsymbol{\ell}_i^{\mathsf{T}} \overline{\boldsymbol{T}_{N-n}^{\boldsymbol{a}} \widehat{\boldsymbol{y}}_i} \right],$$

where the vector $\boldsymbol{\ell} = [\boldsymbol{\ell}_1^{\mathsf{T}} \ \ldots \ \boldsymbol{\ell}_p^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{p(N-n)}$ contains the Lagrange multipliers. The system of equations that characterizes the first-order optimal solutions

[2] This is the set of all model-compliant data sequences.

[3] Remark that the matrix of given data $\boldsymbol{Y}$ has a non-trivial null space by construction when $p(N-n) < n+1$, such that $N \geq \lceil n + (n+1)/p \rceil$ is a necessary condition for the identification problem to be well-defined.

follows from equating the partial derivatives of the Lagrangian to zero:

$$\frac{\partial \mathcal{L}}{\partial \overline{\boldsymbol{\ell}_i}} = \boldsymbol{T}^{\boldsymbol{a}}_{N-n}\widehat{\boldsymbol{y}}_i = \boldsymbol{0}, \quad \forall i \in \{1, \ldots, p\} \tag{8.4}$$

$$\frac{\partial \mathcal{L}}{\partial \overline{\widehat{\boldsymbol{y}}_i}} = \widehat{\boldsymbol{y}}_i - \boldsymbol{y}_i + (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{H}}\boldsymbol{\ell}_i = \boldsymbol{0}, \quad \forall i \in \{1, \ldots, p\} \tag{8.5}$$

$$\frac{\partial \mathcal{L}}{\partial \overline{\boldsymbol{a}_{\mathrm{tr}}}} = \sum_{i=1}^{p} \widehat{\boldsymbol{Y}}_i^{\mathsf{H}}\boldsymbol{\ell}_i = \boldsymbol{0}. \tag{8.6}$$

Since $\mathcal{L}$ is purely real, the partial derivatives w.r.t. $\boldsymbol{\ell}_i$, $\widehat{\boldsymbol{y}}_i$, and $\boldsymbol{a}_{\mathrm{tr}}$ correspond to the complex conjugates of (8.4), (8.5), and (8.6), respectively.

Equating the partial derivatives of the Lagrangian w.r.t. the model-compliant data to zero (cfr. (8.5)) illustrates that for each output i, the misfits $\widetilde{\boldsymbol{y}}_i$ lie in the row space of the complex conjugate of the Toeplitz matrix $\boldsymbol{T}^{\boldsymbol{a}}_{N-n}$

$$\begin{aligned} -(\boldsymbol{y}_i - \widehat{\boldsymbol{y}}_i) + (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{H}}\boldsymbol{\ell}_i &= \boldsymbol{0} \\ \iff \widetilde{\boldsymbol{y}}_i &= (\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{H}}\boldsymbol{\ell}_i. \end{aligned} \tag{8.7}$$

Combining this with the constraints (8.3) and $\boldsymbol{y}_i = \widehat{\boldsymbol{y}}_i + \widetilde{\boldsymbol{y}}_i$ gives

$$\begin{aligned} \boldsymbol{T}^{\boldsymbol{a}}_{N-n}\widehat{\boldsymbol{y}}_i = \boldsymbol{T}^{\boldsymbol{a}}_{N-n}(\boldsymbol{y}_i - \widetilde{\boldsymbol{y}}_i) &= \boldsymbol{0} \\ \iff \boldsymbol{T}^{\boldsymbol{a}}_{N-n}\boldsymbol{y}_i &= \boldsymbol{T}^{\boldsymbol{a}}_{N-n}\widetilde{\boldsymbol{y}}_i \\ \iff \boldsymbol{T}^{\boldsymbol{a}}_{N-n}\boldsymbol{y}_i &= \boldsymbol{T}^{\boldsymbol{a}}_{N-n}(\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{H}}\boldsymbol{\ell}_i. \end{aligned}$$

Writing this out for all outputs $i = 1, \ldots, p$ gives

$$\begin{bmatrix} \boldsymbol{T}^{\boldsymbol{a}}_{N-n} & & & \\ & \boldsymbol{T}^{\boldsymbol{a}}_{N-n} & & \\ & & \ddots & \\ & & & \boldsymbol{T}^{\boldsymbol{a}}_{N-n} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{y}}_1 \\ \widehat{\boldsymbol{y}}_2 \\ \vdots \\ \widehat{\boldsymbol{y}}_p \end{bmatrix} = \boldsymbol{D}_{\boldsymbol{a}} \begin{bmatrix} \boldsymbol{\ell}_1 \\ \vdots \\ \boldsymbol{\ell}_p \end{bmatrix}, \tag{8.8}$$

where the square matrix $\boldsymbol{D}_{\boldsymbol{a}} \in \mathbb{C}^{p(N-n)\times p(N-n)}$ corresponds to a block-diagonal matrix of p blocks:

$$\boldsymbol{D}_{\boldsymbol{a}} = \mathrm{blockdiag}\left(\boldsymbol{T}^{\boldsymbol{a}}_{N-n}(\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{H}}, \boldsymbol{T}^{\boldsymbol{a}}_{N-n}(\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{H}}, \ldots, \boldsymbol{T}^{\boldsymbol{a}}_{N-n}(\boldsymbol{T}^{\boldsymbol{a}}_{N-n})^{\mathsf{H}}\right).$$

Using (8.3) we can rewrite (8.8) as,

$$\begin{bmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_p \end{bmatrix} \begin{bmatrix} \boldsymbol{a}_{\mathrm{tr}} \\ 1 \end{bmatrix} = \boldsymbol{D}_{\boldsymbol{a}}\boldsymbol{\ell}, \tag{8.9}$$

where $\boldsymbol{\ell} = [\boldsymbol{\ell}_1^{\mathsf{T}}, \ldots, \boldsymbol{\ell}_p^{\mathsf{T}}]^{\mathsf{T}}$. It follows from (8.8) that the optimal Lagrange multipliers $\boldsymbol{\ell}_i$ correspond to

$$\boldsymbol{\ell}_i = (\boldsymbol{T}_{N-n}^{\boldsymbol{a}}(\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{H}})^{-1}\boldsymbol{T}_{N-n}^{\boldsymbol{a}}\boldsymbol{y}_i$$

such that the optimal misfits $\widetilde{\boldsymbol{y}}_i$ are obtained via the orthogonal projection of the given data $\boldsymbol{y}_i$ onto row $\left(\boldsymbol{T}_{N-n}^{\boldsymbol{a}}\right)$:

$$\widetilde{\boldsymbol{y}}_i = (\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{H}}(\boldsymbol{T}_{N-n}^{\boldsymbol{a}}(\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{H}})^{-1}\boldsymbol{T}_{N-n}^{\boldsymbol{a}}\boldsymbol{y}_i.$$

8.3.3 Model optimality

Optimality of the model parameters $\boldsymbol{a}$ is obtained when we equate the partial derivatives of the Lagrangian w.r.t. the model-parameters (8.6) to zero

$$\sum_{i=1}^{p} \widehat{\boldsymbol{Y}}_i^{\mathsf{H}}\boldsymbol{\ell}_i = \sum_{i=1}^{p}\left(\boldsymbol{Y}_i^{\mathsf{H}} - \widetilde{\boldsymbol{Y}}_i^{\mathsf{H}}\right)\boldsymbol{\ell}_i = \boldsymbol{0}$$

$$\iff \begin{bmatrix}\boldsymbol{Y}_1^{\mathsf{H}} & \cdots & \boldsymbol{Y}_2^{\mathsf{H}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\ell}_1 \\ \vdots \\ \boldsymbol{\ell}_p\end{bmatrix} = \sum_{i=1}^{p}\boldsymbol{T}_{\boldsymbol{\ell}_i}\overline{\widetilde{\boldsymbol{y}}_i}.$$

Use now that $\overline{\widetilde{\boldsymbol{y}}_i} = (\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{T}}\overline{\boldsymbol{\ell}_i}$ which follows from the complex conjugate of (8.7) and $(\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{T}}\overline{\boldsymbol{\ell}_i} = \boldsymbol{T}_{\boldsymbol{\ell}_i}^{\mathsf{H}}\boldsymbol{a}$, such that

$$\begin{bmatrix}\boldsymbol{Y}_1^{\mathsf{H}} & \cdots & \boldsymbol{Y}_2^{\mathsf{H}}\end{bmatrix}\boldsymbol{\ell} = \boldsymbol{D}_{\boldsymbol{\ell}}\boldsymbol{a}, \tag{8.10}$$

where the square matrix $\boldsymbol{D}_{\boldsymbol{\ell}} \in \mathbb{C}^{(n+1)\times(n+1)}$ corresponds to

$$\boldsymbol{D}_{\boldsymbol{\ell}} = \sum_{i=1}^{p}\boldsymbol{T}_{\boldsymbol{\ell}_i}\boldsymbol{T}_{\boldsymbol{\ell}_i}^{\mathsf{H}}.$$

8.3.4 Complex Riemannian SVD (RiSVD)

Combining (8.9) with (8.10) leads to the following system:

$$\boldsymbol{Y}\boldsymbol{a} = \boldsymbol{D}_{\boldsymbol{a}}\boldsymbol{\ell},$$

$$\boldsymbol{Y}^{\mathsf{H}}\boldsymbol{\ell} = \boldsymbol{D}_{\boldsymbol{\ell}}\boldsymbol{a}.$$

Substituting $\boldsymbol{\ell}$ via $\boldsymbol{\ell} = \tau \boldsymbol{u}$, with $\tau \in \mathbb{R}$ such that $\boldsymbol{u}^{\mathsf{H}} \boldsymbol{D}_{\boldsymbol{a}} \boldsymbol{u} = 1$ and $\boldsymbol{D}_{\boldsymbol{\ell}} = \tau^2 \boldsymbol{D}_{\boldsymbol{u}}$, we obtain the set of equations:

$$\begin{aligned} \boldsymbol{Y}\boldsymbol{a} &= \boldsymbol{D}_{\boldsymbol{a}} \boldsymbol{u} \tau, \qquad & \boldsymbol{a} &= \begin{bmatrix} \boldsymbol{a}_{\mathrm{tr}} \\ 1 \end{bmatrix}, \\ \boldsymbol{Y}^{\mathsf{H}} \boldsymbol{\ell} &= \boldsymbol{D}_{\boldsymbol{u}} \boldsymbol{a} \tau, & \boldsymbol{u}^{\mathsf{H}} \boldsymbol{D}_{\boldsymbol{a}} \boldsymbol{u} &= 1. \end{aligned} \tag{8.11}$$

This is a nonlinear generalized SVD with $\boldsymbol{D}_{\boldsymbol{a}}$ and $\boldsymbol{D}_{\boldsymbol{u}}$ positive definite weighting matrices, quadratic in $\boldsymbol{a}$ and $\boldsymbol{u}$, respectively. On the basis of our analysis, we conclude that the original real Riemannian SVD (RiSVD) generalizes to the complex plane by replacing the transposes with Hermitian conjugates. The value of τ^2 is equal to the objective function, i.e., the squared 2-norm of the misfits, just as in [38].

In other words, the identification problem (8.2) can be solved in a locally optimal way by computing the triplet $(\boldsymbol{a}, \tau, \boldsymbol{u}) \in \mathbb{C}^{n+1} \times \mathbb{R} \times \mathbb{C}^{p(N-n)}$ that solves the RiSVD (8.11) with smallest possible τ^2. The corresponding model-compliant data $\widehat{\boldsymbol{y}}$ can be obtained from the triplet $(\boldsymbol{a}, \tau, \boldsymbol{u})$ via

$$\begin{bmatrix} \widehat{\boldsymbol{y}}_1 \\ \vdots \\ \widehat{\boldsymbol{y}}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_p \end{bmatrix} - \tau \begin{bmatrix} (\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{H}} & & \\ & \ddots & \\ & & (\boldsymbol{T}_{N-n}^{\boldsymbol{a}})^{\mathsf{H}} \end{bmatrix} \boldsymbol{u}. \tag{8.12}$$

8.4 An Inverse-Iteration Method

To solve for a minimizing triplet $(\boldsymbol{a}, \tau, \boldsymbol{u})$ of the real-valued RiSVD, the standard variable metric inverse iteration algorithm can be used [38]. With a few minor adjustments, this algorithm generalizes to the complex-valued RiSVD (8.11): All transposes in the original algorithm should be changed to Hermitian transposes; and to ensure uniqueness of equivalent obtained triplets $\boldsymbol{a}^{(k)}$ and $\boldsymbol{u}^{(k)}$ are projected such that $a_0^{(k)} = 1$ (line 13 and 14). Algorithm 1 shows the modified algorithm. Note that all variables in Algorithm 1 are complex.

Like alternating least squares (ALS) methods, this algorithm iteratively solves the two equations in the first column of (8.11) by linearizing them freezing $\boldsymbol{D}_{\boldsymbol{u}}$ and $\boldsymbol{D}_{\boldsymbol{a}}$ for the current values of $\boldsymbol{u}$ and $\boldsymbol{a}$. In each subsequent iteration, the metrics $\boldsymbol{D}_{\boldsymbol{u}}$ and $\boldsymbol{D}_{\boldsymbol{a}}$ are updated with the new values of the intermediate solutions $\boldsymbol{u}^{(k)}$ and $\boldsymbol{a}^{(k)}$ and the same (linearized) equations are solved again. The algorithm splits $\boldsymbol{u}$ into two orthogonal components $\boldsymbol{Q}_1 \boldsymbol{z}$ and $\boldsymbol{Q}_2 \boldsymbol{w}$. The output of the algorithm depends on the initial guesses $\boldsymbol{u}^{(k)}$ and $\boldsymbol{a}^{(k)}$, which makes this a heuristic approach. Convergence is tested by measuring the difference between subsequent values $\tau^{(k)}$.

Algorithm 1: Variable metric inverse iteration algorithm on $\mathbb{C}$, based on [38].

Input: Initial guess $\boldsymbol{a}^{(0)}$, $\boldsymbol{u}^{(0)}$;
$\boldsymbol{Q} \in \mathbb{C}^{p(N-n)\times p(N-n)}, \boldsymbol{R} \in \mathbb{C}^{p(N-n)\times(n+1)} \leftarrow \mathrm{QR}(\boldsymbol{Y})$
1 $\boldsymbol{Q}_1, \boldsymbol{Q}_2 \leftarrow \boldsymbol{Q}_{[1:n+1]}, \boldsymbol{Q}_{[n+2:p(N-n)]}$
`/* subscript denotes a range of columns */`
2 $k \leftarrow 0$
3 **repeat**
4 $\quad k \leftarrow k+1$
5 $\quad \boldsymbol{z}^{(k)} \leftarrow \boldsymbol{R}^{-\mathsf{H}} \boldsymbol{D}_{\boldsymbol{u}^{(k-1)}} \boldsymbol{a}^{(k-1)}$
6 $\quad \boldsymbol{w}^{(k)} \leftarrow -\left(\boldsymbol{Q}_2^{\mathsf{H}} \boldsymbol{D}_{\boldsymbol{a}^{(k-1)}} \boldsymbol{Q}_2\right)^{-1} \left(\boldsymbol{Q}_2^{\mathsf{H}} \boldsymbol{D}_{\boldsymbol{a}^{(k-1)}} \boldsymbol{Q}_1\right) \boldsymbol{z}^{(k)}$
7 $\quad \boldsymbol{u}^{(k)} \leftarrow \boldsymbol{Q}_1 \boldsymbol{z}^{(k)} + \boldsymbol{Q}_2 \boldsymbol{w}^{(k)}$
8 $\quad \boldsymbol{u}^{(k)} \leftarrow \frac{\boldsymbol{u}^{(k)}}{\left\|\boldsymbol{u}^{(k)}\right\|_2}$
9 $\quad \boldsymbol{a}^{(k)} \leftarrow \boldsymbol{R}^{-1} \boldsymbol{Q}_1^{\mathsf{H}} \boldsymbol{D}_{\boldsymbol{a}^{(k-1)}} \boldsymbol{u}^{(k)}$
10 $\quad \tau^{(k)} \leftarrow \left\|\boldsymbol{a}^{(k)}\right\|_2$
11 $\quad \boldsymbol{a}^{(k)} \leftarrow \frac{\boldsymbol{a}^{(k)}}{\tau^{(k)}}$
12 **until** *convergence*
13 $\boldsymbol{u}^{(k)} \leftarrow \boldsymbol{u}^{(k)} a_0^{(k)}$
14 $\boldsymbol{a}^{(k)} \leftarrow \boldsymbol{a}^{(k)} / a_0^{(k)}$

As for computational complexity, the main cost of the algorithm is the matrix inversion[4] of the symmetric positive definite $[p(N-n)-(n+1)] \times [p(N-n)-(n+1)]$ matrix $\boldsymbol{Q}_2^{\mathsf{H}} \boldsymbol{D}_{\boldsymbol{a}} \boldsymbol{Q}_2$, which is computed in each iteration (line 6). A single QR decomposition of $\boldsymbol{Y}$ is required at initialization, with negligible cost compared to the iterative steps.

Remark 8.1 (Convergence behavior)**.** *At present, no proof of convergence for the proposed algorithm is known. It can be viewed as a* self-consistent field (SCF) iteration *(see e.g., [145]): the quantities defining the linearized equations depend on the solution one is trying to compute, and are repeatedly updated until a fixed point is reached. SCF-type iterations are well known for exhibiting notoriously difficult-to-predict convergence behavior, and the present method is no exception. Numerical experiments sometimes display periods of stagnation followed by sudden rapid convergence, while in other cases the iterates settle more smoothly. Nevertheless, in most practical situations, and especially when good initial conditions are provided, the algorithm converges rapidly to a fixed point.*

Remark 8.2 (Optimal solution approaches)**.** *The focus of this chapter is theoretical: we extend the structured TLS framework to the complex domain and,*

[4]While symmetry and positive definiteness can be exploited, such implementation considerations fall outside the scope of this letter.

Table 8.1: Exact parameter values of the simulated NMR signal used in (8.13) in example 8.3.

m	f_m [Hz]	d_m [rad/s]	A_m [1]	ϕ_m [rad]
1	-1379	208	6.1	$15\pi/180$
2	-685	256	9.9	$15\pi/180$
3	-271	197	6.0	$15\pi/180$
4	353	117	2.8	$15\pi/180$
5	478	808	17.0	$15\pi/180$

as a proof of concept, present an algorithm to solve the equations. Future work must address the question of how to compute solutions to these novel Riemannian SVD equations in an optimal and numerically robust manner. Alternatives to the presented algorithm exist. For instance, the Riemannian SVD can also be viewed as an eigenvalue problem with an eigenvector nonlinearity, a problem class for which new solution techniques are actively being developed, see e.g., [29, 30].

8.5 Numerical Experiments

8.5.1 Nuclear magnetic resonance (NMR) modeling

Complex signals are prevalent in nuclear magnetic resonance (NMR) modeling, where they allow storing phase information obtained via quadrature detection. We closely follow an experiment described in [172], where a signal produced by an nth-order autonomous single-output system is modeled as:

$$y_{\text{exact},k} = \sum_{m=1}^{n} A_m e^{\mathrm{j}(\phi_0+\phi_m)} e^{(-d_m+\mathrm{j}2\pi f_m)t_k} \tag{8.13}$$

where $t_k = kT_\text{s}$ with T_s the sampling period. The frequencies f_m and damping factors d_m can be retrieved via the roots z_m of the characteristic polynomial $\boldsymbol{a}(z)$ of the LTI model.

Example 8.3. *The performance of the method is tested using simulated data obtained from a model with parameters A_m, ϕ_m, d_m, and f_m as shown in Table 8.1, and $\phi_0 = 0$. Note that the frequencies f_4 and f_5 lie close to each other, making the estimation problem quite ill conditioned.*

The model is evaluated for $N = 128$ data points, uniformly sampled with sampling time $T_\text{s} = 10^{-4}$ s. The data sequence is perturbed with complex zero

mean Gaussian noise $\epsilon_k \in \mathbb{C}$ with variance $4\nu^2$, i.e., where

$$\epsilon_k = \epsilon_k^{\mathrm{Re}} + \mathrm{j}\epsilon_k^{\mathrm{Im}}, \qquad \epsilon_k^{\mathrm{Re}}, \epsilon_k^{\mathrm{Im}} \sim \mathcal{N}\left(0, \nu^2\right). \tag{8.14}$$

The best total least squares (TLS) estimator of the model parameters $\boldsymbol{a}_{TLS}$ is used to determine an initial guess for the inverse iteration algorithm (Algorithm 1): (1) $\boldsymbol{a}^{(0)}$ is taken as the TLS estimate $\boldsymbol{a}_{\mathrm{TLS}}$; (2) $\boldsymbol{u}^{(0)}$ is then calculated as $\boldsymbol{u}^{(0)} = \left(\boldsymbol{T}^{\mathsf{H}}_{\boldsymbol{a}^{(0)}}\right)^{\dagger}(\boldsymbol{y} - \widehat{\boldsymbol{y}}_{\mathrm{TLS}})$, where $\widehat{\boldsymbol{y}}_{\mathrm{TLS}}$[5] is the model-compliant data as estimated using TLS.

The method is evaluated across 200 independent noisy signal realizations $\boldsymbol{y}_i$ at varying noise levels ν, using: a single sensor (p=1); two sensors (p=2); and four sensors (p=4). For p=2 and p=4, the data is grouped into sets of two and four realizations, respectively, resulting in 100 and 50 executions of the algorithm.

Figs. 8.2-8.4 shows the distribution of the misfits and the relative errors on the hard to estimate frequencies f_4 and f_5 and damping factors d_4 and d_5. Each of the obtained solutions is a local minimum of the realization problem (8.2).

It is noticeable that for larger noise levels, the given data $\boldsymbol{y}$ stray further away from the model-compliant data space. *Since this space is merely a small subspace of the large ambient space $\mathbb{C}^{pN}$ and $\widetilde{\boldsymbol{y}}$ corresponds to the residual of the orthogonal projection of $\boldsymbol{y}$ onto this subspace, it is not unexpected to observe $\|\widetilde{\boldsymbol{y}}\|_2^2$ to increase with ν.*

When the number of observations p increases, both $\|\widetilde{\boldsymbol{y}}\|_2^2$ and the accuracy of the estimated parameters improve on average and the variance of the estimates across multiple independent runs/noise realizations decreases. These results empirically support the intuition that concurrently using multiple independent sequences enhances estimation robustness. We observe that the runtime of Algorithm 1 scales approximately as $\mathcal{O}(p^3)$ for a single execution. However, when the total amount of data is fixed, increasing p reduces the number of independent runs proportionally, resulting in an overall computational cost that grows roughly quadratically with p.

It is further clear from Figs. 8.2-8.4 that for larger noise levels ν the parameter f_4 is not correctly estimated. This is due to the RiSVD method overfitting on the noise and consequently interpreting the contributions of f_4 and f_5 as contributions done by only a single frequency. Overfitting on the noise is a natural consequence of the objective function (8.2).

[5] Note that $\widehat{\boldsymbol{y}}_{\mathrm{TLS}}$ is not a critical point of (8.2), since it disregards the Hankel structure of the given data matrix $\boldsymbol{Y}$ when reducing its rank.

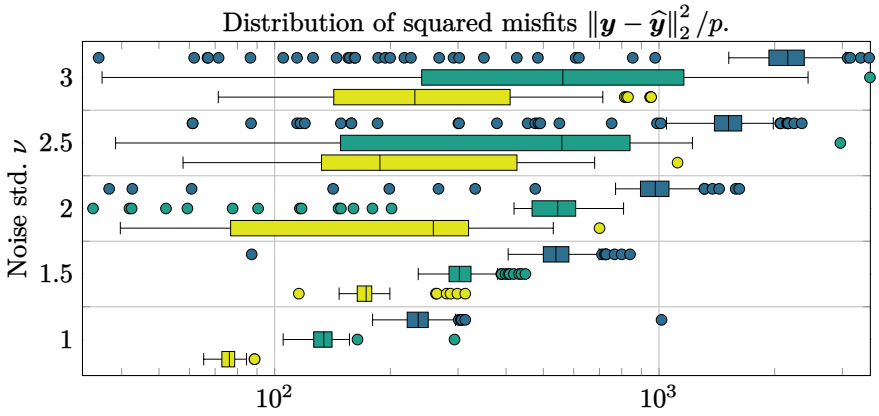


Figure 8.2: Distribution of the squared 2-norm of the misfits for 200 noisy realizations for various noise levels and various number of sensors p, corresponding to Experiment 8.3. Legend: $p = 1$, $p = 2$, $p = 4$.

8.5.2 Direction of arrival (DOA) estimation

In a second experiment we relate the autonomous realization problem to a direction of arrival (DOA) estimation problem. Estimating the DOA is a common problem in radar, sonar, and wireless communication applications [129, 148].

Consider a setup of p sensors, evenly distributed along a line as illustrated in Figure 8.5. A signal $s(t)$ is transmitted and received by the sensor array under an angle θ, where $s(t)$ is modeled as a sum of n constant amplitude complex exponentials:

$$s(t) = \sum_{m=1}^{n} R_m e^{\mathrm{j}\psi_m} e^{\mathrm{j}2\pi f_m t}.$$

With $R_m, \psi_m, f_m \in \mathbb{R}$. The signal source is assumed to be far enough such that the DOA θ is virtually the same for each sensor. The ith antenna receives a phase-shifted discrete-time version of $s(t)$ perturbed by zero mean Gaussian measurement noise:

$$y_{i,k} = \sum_{m=1}^{n} R_m e^{\mathrm{j}\psi_m} e^{\mathrm{j}2\pi f_m (kT_\mathrm{s} + \phi_0 + \phi_i)} + \epsilon_{i,k}. \tag{8.15}$$

We assume ϕ_0 to be zero. The phase-shift of the ith sensor is then given by:

$$\phi_i = i \frac{\Delta \cos(\theta)}{c} \tag{8.16}$$

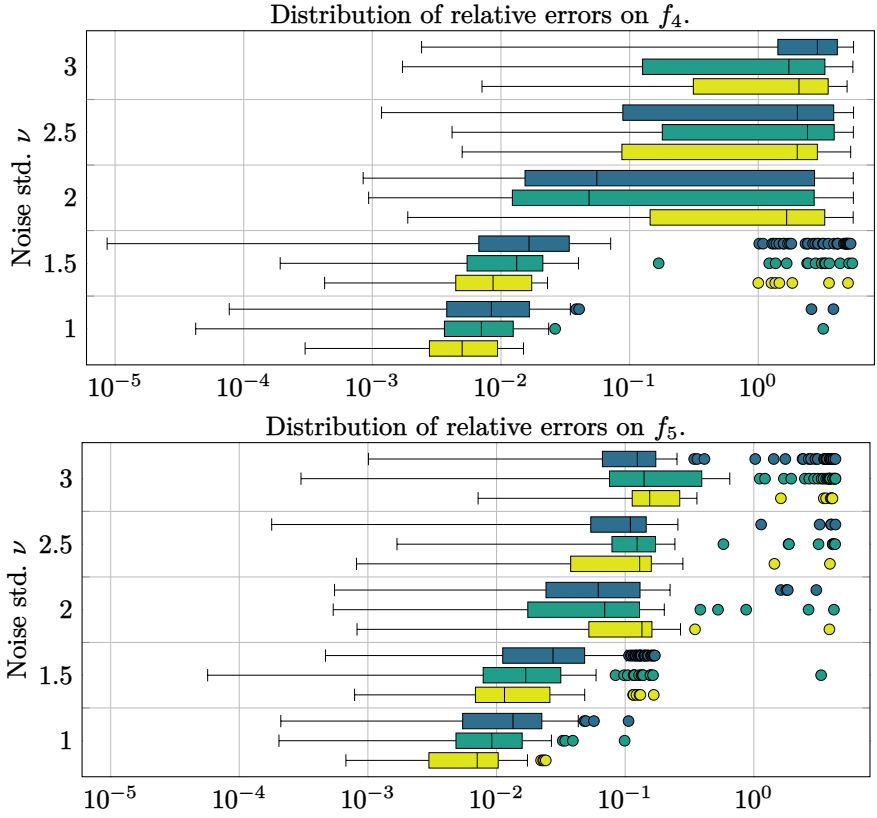


Figure 8.3: Distribution of the relative errors on f_4, f_5 for 200 noisy realizations for various noise levels and various number of sensors p, corresponding to Experiment 8.3. Legend: $p = 1$, $p = 2$, $p = 4$.

where c is the speed of light and Δ is the distance between two successive sensors. After calculating the best fitting model parameters $\boldsymbol{a}$ corresponding to the received data $\boldsymbol{y}_i$ (8.15), the corresponding $\widehat{R}_m$, $\widehat{\psi}_m$ and $\widehat{f}_m$ can be determined and consequently θ can be estimated using (8.16).

Example 8.4. *We consider a setup of $p = 4$ sensors, where the distance between two successive sensors is $\Delta = 250\,\mathrm{m}$ and the (unknown) DOA $\theta = \pi/4$. We generate 200 different signals by randomly generating 200 sets of $\boldsymbol{R}, \boldsymbol{f} \in \mathbb{R}^n$ and $\boldsymbol{\psi} = \boldsymbol{0}$ and evaluate the model (8.15) for $N = 128$ samples, where $T_\mathrm{s} = 1\,\mu\mathrm{s}$. To ensure the frequencies f_m being sufficiently distinct, they are drawn from a normal distribution $\mathcal{N}(\frac{m-1}{2nT_\mathrm{s}}, \frac{1}{(2nT_\mathrm{s})^2})$. The amplitudes R_m are drawn from a uniform distribution on the range $[5, 15]$. The experiment*

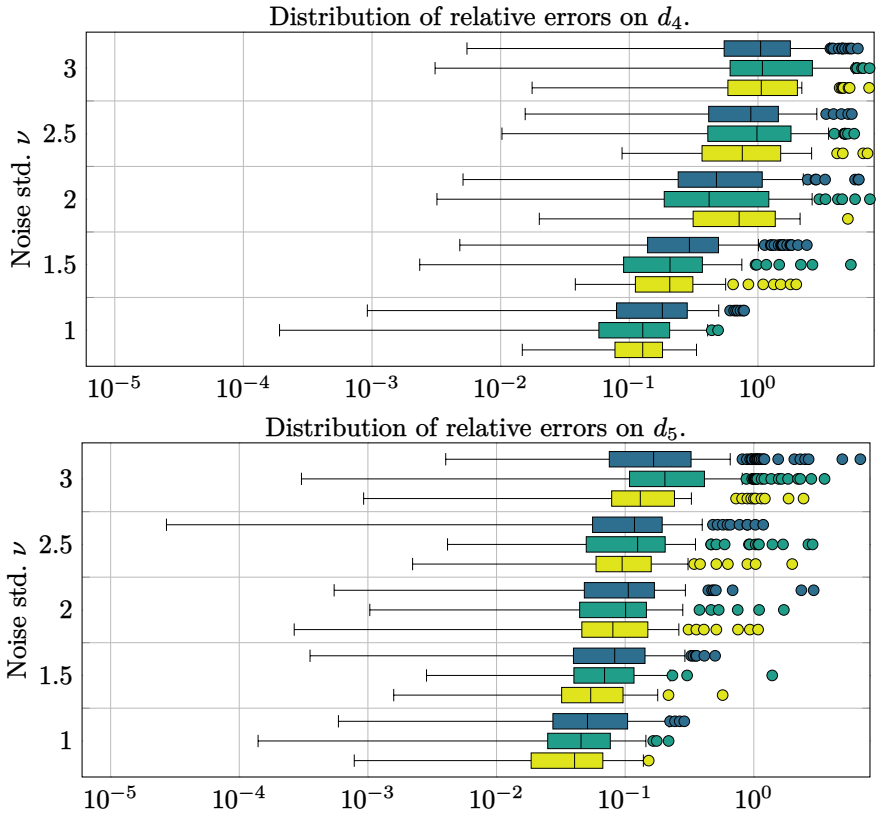


Figure 8.4: Distribution of the relative errors on d_4, and d_5 for 200 noisy realizations for various noise levels and various number of sensors p, corresponding to Experiment 8.3. Legend: $p = 1$, $p = 2$, $p = 4$.

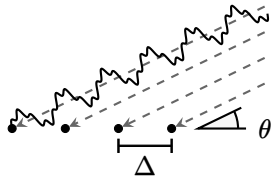


Figure 8.5: A sensor array consisting of $p = 4$ sensors on a line receiving the same signal that arrives in an angle θ w.r.t. the sensor array. Each sensor is represented by a dot. The distance between two successive sensors is Δ.

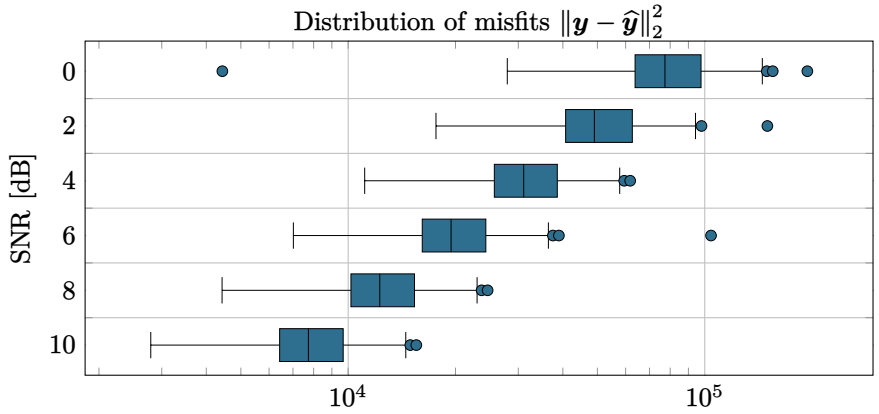


Figure 8.6: Distribution of misfits for 200 different noisy realizations for various noise levels corresponding to Experiment 8.4.

is performed for different signal to noise ratios $\text{SNR} = 0, 2, \ldots, 10$ *dB. For each* SNR*, zero mean Gaussian noise with variance* $4\nu^2$ (8.14) *is added, such that* $\text{SNR} = 10 \log_{10}\left(P_{\text{avg}}/4\nu^2\right)$ *where* $P_{\text{avg}} = \|y_{\text{exact}}\|_2^2/pN$ *is the average power of the noiseless signal. The resulting data is gathered in a vector* $\boldsymbol{y}$ *and structured in a Hankel matrix* $\boldsymbol{Y}$ *. Using Algorithm 1, the optimal RiSVD triplet* $(\boldsymbol{u}, \tau, \boldsymbol{a})$ *is determined with corresponding model-compliant data* $\widehat{\boldsymbol{y}}$ *via* (8.12)*. The parameter estimates* $\widehat{R}_m, \widehat{\psi}_m, \widehat{f}_m$ *corresponding to the model-compliant data* $\widehat{\boldsymbol{y}}$ *can be retrieved via the model parameters* $\boldsymbol{a}$*, and subsequently the DOA* θ *can be estimated.*

Figure 8.6 shows the distribution of the misfits. Similarly as in Experiment 8.3, we observe increasing misfits for increasing noise levels. Figure 8.7 shows histograms of the estimated DOA $\widehat{\theta}$ *of the 200 different noisy signals for each of the various noise levels* $\text{SNR} = 0, 2, \ldots, 10$ *dB. The histograms are constructed using bins of width* $\pi/50$ *rad. The estimated DOA shows a peak around the true DOA* $\theta = \pi/4$ *for all noise levels, where the peak is more pronounced for smaller noise levels. This experiment shows that for smaller noise levels the RiSVD method is able to estimate* θ *within a reasonable accuracy, and even for larger noise levels the estimations still show a peak near the true DOA* θ*.*

8.6 Conclusions and Future Work

We show that the (locally) optimal solutions of the complex LS realization problem follow from a nonlinear generalized singular value decomposition over

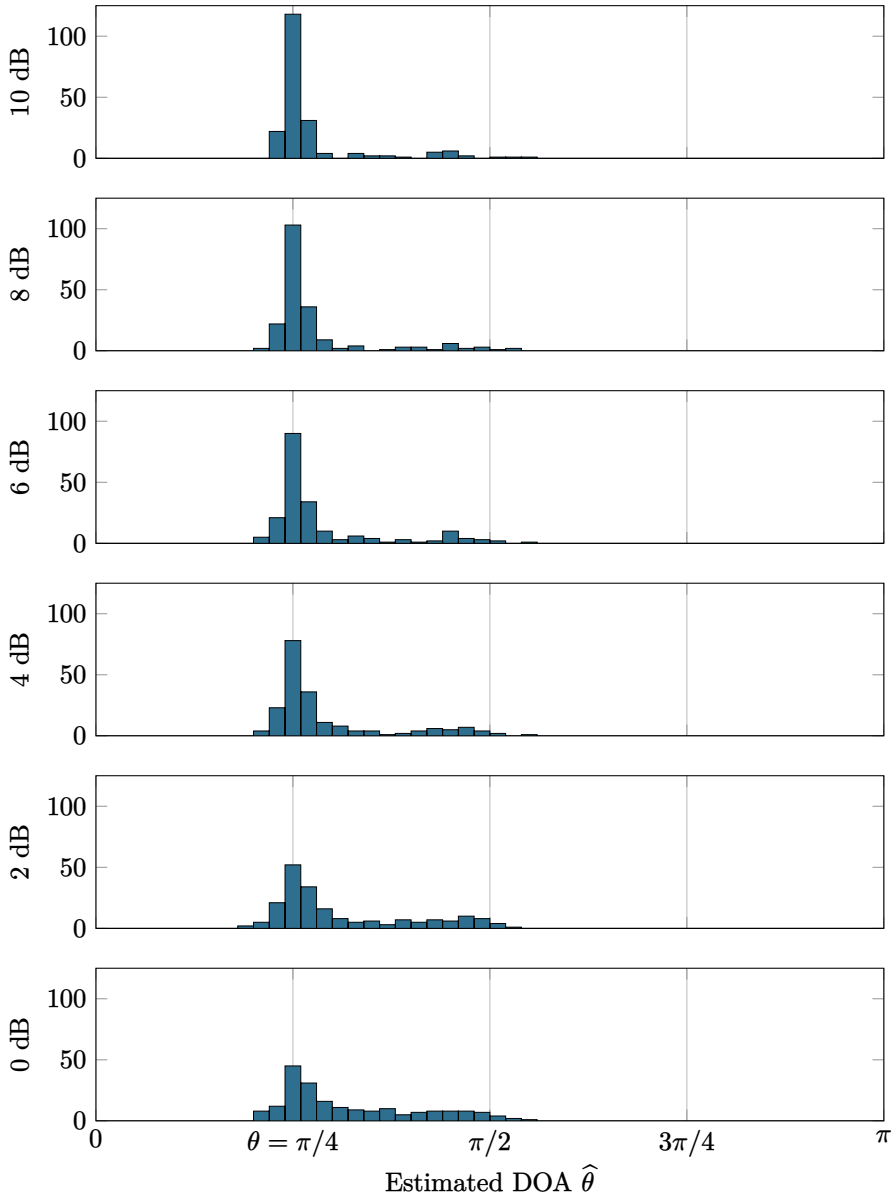


Figure 8.7: Histograms of the estimated DOA $\widehat{\theta}$ for 200 different noisy signals for SNR $= 0, 2, \ldots, 10$ dB corresponding to Experiment 8.4.

the complex field and that with few minor adjustments the standard variable metric inverse iteration algorithm can be used to solve for minimizing singular triplets. The numerical experiments illustrate the use of the realization algorithm for two experimental setups.

While the focus of this chapter is theoretical—extending the structured TLS framework to the complex domain—an important direction for future research is to assess the practical performance of the proposed approach in comparison with the state of the art. We have not aimed to establish competitiveness in this regard, but rather to provide a rigorous foundation for further methodological exploration.

Conclusion

Chapter 9

Contributions, Discussion and Perspectives for Future Work

We begin this chapter with a high-level synthesis of the main contributions of this dissertation (Section 9.1). We then broaden the perspective in Section 9.2, where we reflect on several overarching themes related to the frameworks developed in this dissertation and situate the results within their broader scientific context. The chapter concludes in Section 9.3 with an outlook on promising directions for future research.

9.1 Research Contributions

In this section, we provide a high-level overview of the main research contributions of this dissertation, restricting ourselves to broad themes since the specific contributions have already been described in detail at the beginning of each chapter.

9.1.1 SISO $\mathcal{H}_2$-norm model reduction

We introduced a novel methodology for globally optimal model reduction in the single-input single-output (SISO) setting, where the objective is to approximate a high-order stable linear time-invariant (LTI) system by a reduced-order model that minimizes the $\mathcal{H}_2$-norm of the approximation error.

Achieving global optimality

In **Chapter 2**, the first-order optimality conditions associated with this problem were, by exploiting Walsh's Theorem, reformulated as a quadratic multiparameter eigenvalue problem (MEP). This formulation characterizes all local and global minimizers of the non-convex optimization landscape. Notably, the degree of the resulting MEP is independent of the model orders and its coefficient matrices have a relatively small size, offering a computational advantage over alternative globally optimal methods in the literature, whose complexity typically scales worse with increasing system sizes.

Quantifying algebraic complexity

Building on this framework, **Chapter 5** undertook a rigorous investigation into the algebraic complexity of the solution set of the critical equations of the $\mathcal{H}_2$-norm model reduction problem. By relating the solution set of the critical equations to a corank-one determinantal variety, we were able to apply the Thom–Porteous formula to derive an explicit closed-form expression for the total number of complex solutions. This result provides the first general upper bound on the number of genuine critical points (e.g., local minima, saddle points) for the model reduction problem, which corresponds to a subset of these complex solutions.

The derived count is of both theoretical and practical importance. From a theoretical standpoint, it reveals the intrinsic algebraic complexity of the problem and characterizes the structure of its solution set. From a practical perspective, the upper bound offers a benchmark for assessing the completeness of numerical solvers, such as homotopy continuation, when applied to the critical equations. Furthermore, it informs the design and evaluation of local optimization algorithms specialized for the model reduction problem by quantifying the maximum number of distinct locally optimal points such methods may encounter.

Summary

Taken together, these contributions formally establish a novel globally optimal methodology for SISO $\mathcal{H}_2$-norm model reduction, advancing the state of the art and laying the groundwork for extensions to more complex model structures and/or scalable solution approaches.

9.1.2 Least squares realization of autonomous LTI models

The least squares (LS) realization problem for autonomous single-output linear time-invariant (LTI) dynamical models is a non-convex optimization task, where the goal is to identify model parameters that minimize the misfit between given data and model-compliant output sequences. Although extensively studied in the statistical and system identification literature, the problem has remained partially unresolved with respect to its *global* solution.

Achieving global optimality

In **Chapter 3**, we demonstrated that first-order necessary conditions for optimality of the problem can be reformulated as an MEP. This characterization provides a deterministic procedure to compute all local and global minimizers, thereby establishing a globally optimal solution method for LS realization. Central to this MEP-based formulation is the so-called *double-FIR-filter property* satisfied by (a subset of) the locally optimal models, which serves as a finite-dimensional analogue of Walsh's Theorem encountered in Chapter 2.

We extended the MEP formulation by incorporating non-uniform weights into the misfit norm, thereby generalizing the framework to settings where certain samples in the time series are regarded as more reliable or informative than others. While introducing this weighted formulation, we highlighted the equivalence of the problem to structured low-rank approximation for Hankel matrices. Numerical experiments demonstrated that the incorporation of non-uniform weights can substantially affect the resulting models, indicating the practical relevance of this extension.

In **Chapter 4**, we extended the globally optimal LS realization framework to incorporate a priori fixed pole information in the identification. By embedding partial spectral knowledge directly into the MEP formulation, we established a globally optimal methodology for estimating the remaining unknown poles. This constrained formulation reduces the number of free parameters and thereby lowers the computational complexity compared to the unconstrained case. Numerical experiments confirmed the validity of the approach and further demonstrated that incorporating fixed pole information can improve the statistical accuracy of the estimations: the output sequences more closely approximate the underlying noise-free signal.

Quantifying algebraic complexity

Building on the weighted LS realization framework, **Chapter 6** examined the algebraic complexity of the complex solution set to the critical equations of the structured low-rank approximation problem for Hankel matrices. Focusing on the solutions inside the smooth locus, we rederived the expression for the generic Euclidean distance (ED) degree of the Hankel variety. Because our analysis uses only on elementary tools from intersection theory, it provides a more accessible perspective than the existing treatments of the ED degree.

Furthermore, our framework provides a novel characterization of the ED defect, which is the deviation of the ED degree for specific weight matrices from the generic value. We expressed the defect in terms of a bad locus, defined as the degeneracy locus of a matrix polynomial depending only on the chosen weights, offering a direct tool to investigate the conditions under which this defect arises. In addition, we established, for the first time, a complete count of the solutions to the critical equations inside the singular locus. Combined with the ED degree, this establishes a full characterization of the solution set.

Summary

Taken together, these contributions establish the first deterministic and globally optimal framework for the weighted LS realization problem, or equivalently, the structured low-rank approximation problem for Hankel matrices, thereby advancing the state of the art in (globally optimal) system identification.

9.1.3 Other contributions

In addition to the core developments on globally optimal model reduction and LS realization, this dissertation pursued two further directions that, while more exploratory in scope, provide valuable insights. These are presented in Chapters 7 and 8, the main contributions of which are summarized here.

Least squares realization for SISO LTI models

Chapter 7 addressed the LS realization problem for SISO LTI models. The analysis began with the study of the LS projection onto the behavior of a known model, where observed input-output data are adjusted in a LS sense to satisfy a prescribed difference equation. We showed that this projection induces an orthogonal decomposition of the data space, thereby extending a part of the

results obtained for autonomous systems in Chapter 3 to the more general SISO setting. The orthogonal decomposition identified in the difference-equation setting carries over naturally to the state-space-based behavioral framework, thereby providing an alternative but equivalent characterization of the optimal solutions. Taken together, these dual perspectives give a better understanding of the least squares projection problem.

We subsequently investigated the full LS realization problem, in which the model parameters themselves must be inferred from the observed data. This problem also admits an interpretation as a structured low-rank approximation problem, formulated in terms of a block Hankel matrix obtained by concatenating the input and output Hankel matrices. Among the considered approaches, the kernel-based Riemannian SVD formulation emerged as the most promising. Numerical experiments indicated that, in contrast to the autonomous case, its solution set does not contain solutions with so-called lower-order dynamics: every computed critical point corresponded to a data Hankel matrix of the maximally allowed rank.

Least squares realization of complex-valued MO LTI models

Chapter 8 extended the LS realization problem to autonomous, *complex-valued*, multiple-output LTI systems, thereby broadening its applicability. The main contribution of this chapter is primarily theoretical: it establishes a rigorous extension of the structured total least squares framework to the complex domain. In particular, we developed a complex-valued Riemannian SVD that characterizes all locally optimal solutions, and showed that the variable metric inverse iteration algorithm, with minor adjustments, provides an effective means of computing the minimizing singular triplets. The methodology was validated through examples from nuclear magnetic resonance modeling and direction-of-arrival estimation.

9.2 Discussion

Beyond the chapter-specific contributions, we reflect on several broader topics that are relevant in the context of this dissertation. In what follows, we discuss computational complexity, the interplay between polynomial systems and multiparameter eigenvalue problems, and the implications of the least squares criterion and global optimality for practical system identification.

9.2.1 Global optimality and computational complexity

Relatively speaking, the frameworks proposed in Part I-II represent a substantial improvement over the alternative globally optimal approaches available in the literature. In particular, the obtained MEPs have more compact coefficient matrices than previously known formulations and have a fixed degree. It may even be conjectured that these MEP-based descriptions are optimal, insofar as no more compact characterization relying solely on first-order information is likely to exist.

In absolute terms, however, the computational cost of achieving global optimality remains considerable. This cost constitutes the main bottleneck that prevents the widespread use of the proposed methodologies in practice. Solving large systems of polynomial equations or MEPs is intrinsically hard, and the complexity grows rapidly with the problem size. Progress in the development of numerical solvers and advances in high-performance computing may partially alleviate some of these difficulties, but they cannot overcome the inherent growth in problem complexity.

Indeed, the solution counts obtained in Chapters 5 and 6 indicate a combinatorial increase in the number of solutions with respect to problem sizes. This demonstrates that our "brute-force" strategy for achieving global optimality—computing and analyzing all (regular) critical points—is inherently not scalable. Only under the hypothetical assumption of solvers with negligible runtime would it be feasible to enumerate and examine the full set of candidate solutions.

The main conclusion is therefore that the proposed MEP-based framework is most valuable from a theoretical standpoint. It provides a rigorous understanding of the structure of the problems under consideration and establishes benchmarks against which the performance of heuristic and locally optimal algorithms can be assessed. While its probability of adoption in routine engineering applications is small, its role as a reference framework for theory and benchmarking remains highly significant.

9.2.2 Polynomial system solving vs. MEPs

A recurring theme throughout this dissertation has been the interplay between two related formulations: on the one hand, systems of polynomial equations arising from the critical conditions, and on the other, multiparameter eigenvalue problems (MEPs) obtained when eliminating a subset of the variables from the polynomial systems. Each formulation offers distinct advantages and limitations, both computationally and theoretically.

From a computational perspective, both formulations must ultimately be addressed in a floating-point setting, since the observed data are assumed to be noisy such that one naturally operates in a floating point environment. The homotopy continuation (HC) method, used in this dissertation for solving polynomial systems, has the notable advantage of being naturally parallelizable: the independent tracking of solution paths can be distributed across processors, substantially reducing computational time for larger problem instances. In practice, we observed that as the problem size grows, the serial mixed-volume computations, which are required to construct an appropriate start system, become nearly as expensive as the subsequent path-tracking itself, as long as sufficiently many processors are provided for the latter.

The block Macaulay-based approach used for solving the obtained MEPs does not allow for such parallelization. While for small-scale problems the MEP solver is competitive, most likely because the MEP formulation often involves only a subset of the variables appearing in the polynomial system, the computational cost increases more steeply for larger systems.

For these larger and more complex problem instances, a challenge in HC is the reliable detection of solutions at infinity: solution paths corresponding to genuine affine solutions may be falsely terminated, leading to incomplete solution sets. The block Macaulay framework allows affine and infinite solutions to be separated using well-established numerical linear algebra tools. In particular, the rank structure of the block Macaulay null space, which can be obtained via an SVD, allows to distinguish between affine solutions and those at infinity. However, we observed that for some larger problem instances, correctly determining this rank can also become challenging. Thus, in large-scale instances, both the Macaulay-based approach and homotopy continuation face challenges in reliably distinguishing affine solutions from those at infinity.

From a theoretical perspective, the two formulations are complementary rather than interchangeable. The polynomial systems provide a direct and complete representation of the critical equations, free from the spurious solutions that may arise in the MEP formulation (as illustrated by the bad locus phenomenon). At the same time, the MEP perspective isolates the nonlinear structure of the problem in an elegant form. In fact, it is precisely the MEP formulation, whose solution set can be interpreted as a determinantal variety, that enabled the application of the Thom–Porteous formula in Chapters 5 and 6, leading to exact counts of the number of solutions.

9.2.3 Evaluating model fit

In practice, observed time-series data are almost never perfectly model-compliant; that is, they do not exactly satisfy the dynamics of the candidate model. From a behavioral perspective, this means that the observed trajectory does not belong to the model's behavior, which is defined as the set of all trajectories that satisfy the model dynamics. Such discrepancies may arise from various sources, including measurement inaccuracies, missing data, observational errors, or model mismatch.

Stochastic frameworks account for this gap by introducing probabilistic noise models, but the validity of these statistical assumptions is often unverifiable in practice. In this dissertation, we therefore adopted a deterministic alternative: the observed data are modified as little as possible by subtracting misfits, so that the corrected data satisfy the prescribed model dynamics. The smaller the misfits, the better the "fit" between model and data. The LS criterion, which evaluates the 2-norm of the misfits, thus emerges as a natural and non-parametric choice. It avoids unverifiable statistical assumptions, accommodates any model representation, and provides a purely descriptive measure of fit.

A similar perspective applies in the context of model reduction. A high-order model may possess dynamics of unnecessarily large order, either due to inherent model mismatch or because it was identified from noisy experimental data. Here, the $\mathcal{H}_2$-norm of the error function measures the difference between the impulse responses of the original and reduced-order models, serving as the analogue of the misfit criterion used in the context of system identification.

From a statistical point of view, the LS criterion is most meaningful under the assumption that the disturbances affecting the data are independent, identically distributed Gaussian random variables. In this case, the LS cost coincides with the maximum likelihood formulation. If the noise distribution were known to deviate from Gaussianity, alternative criteria would be preferable. In practice, however, the central limit theorem suggests that with sufficiently many independent experiments, the aggregate effect of disturbances approaches a Gaussian distribution. This consideration naturally raises the question of experimental design: whether one should collect a single long data sequence, or—as explored in Chapters 4 and 8—conduct multiple experiments with distinct, shorter, observed sequences.

From a practical perspective, one may question whether the pursuit of global optimality is always justified. While the methodologies developed in this dissertation provide deterministic procedures for computing globally minimizing solutions, the obtained solutions are necessarily connected to the particular noise realization present in the observed data. As a result, they may overfit this

specific instance of noise rather than capture the dynamics of the true underlying system. In such cases, a locally optimal solution, or even an approximation that does not satisfy first-order optimality conditions, may provide a more accurate representation of the system's intrinsic behavior. This highlights a tradeoff, well known in the context of machine learning, between mathematical optimality and practical relevance: global optimality guarantees theoretical completeness, but it does not necessarily ensure superior performance in realistic scenarios.

9.2.4 Hankel matrix approximation and Walsh's Theorem

A unifying perspective on Chapters 2-6 is that they all address variants of the low-rank approximation problem for Hankel matrices. In the LS realization setting, the data sequence is finite, leading to a finite Hankel matrix and a mean-square error criterion. In discrete-time model reduction, by contrast, the data sequence extends to infinite length (cf. the impulse response of the system), corresponding to a Hankel matrix with an infinite number of rows and an objective function expressed in the ℓ_2 norm (or, in continuous time, the $\mathcal{H}_2$-norm).

Walsh's Theorem, originating in rational approximation theory, was fundamental to the developments in this dissertation. In Chapter 2, it enabled an algebraic characterization of the critical points of the $\mathcal{H}_2$-norm model reduction problem, which is an optimization problem that, at first sight, is not governed by polynomial equations. In Chapter 3, where we considered the LS realization problem, the finite-dimensional analogue of Walsh's Theorem, the double-FIR-filter property, allowed us to eliminate the singular locus from the solution set of the critical equations, leading to a significantly more compact formulation. In this way, a classical result from the early twentieth century underlies much of the theoretical development presented in this dissertation.

9.3 Perspectives for Future Work

In what follows, we outline several directions for future research that build on the results presented here.

A natural first direction for future research is to extend the globally optimal methodology for model order reduction proposed in Chapter 2 to the frequency-weighted setting. In many applications, the primary interest lies in accurately approximating the system dynamics around specific operating frequencies rather than over the full spectrum. This motivates replacing the standard $\mathcal{H}_2$-norm

error measure with its weighted generalization, in which a frequency-dependent weighting function $W(s)$ is incorporated. It remains to be investigated whether the generalized critical equations admit a reformulation as an MEP, and if so, how the structure of such an MEP compares to that of the unweighted case. Moreover, the algebro-geometric analysis developed in Chapter 5 would become substantially more intricate: in particular, the characterization of the bad locus is likely to be far less transparent in the presence of non-uniform weights.

A complementary direction concerns the analysis of existing iterative methods for $\mathcal{H}_2$-norm model reduction, such as Iterative Rational Krylov Approximation (IRKA). Although IRKA performs remarkably well in practice, no general proof of convergence is known, except in the special case of symmetric models. The globally optimal framework developed in this dissertation offers new tools to study IRKA's dynamics. By comparing its iterates with globally optimal solutions, one may gain structural insight into when IRKA converges, why it fails to do so in certain settings, and how it might be modified to improve robustness.

Another promising avenue for future research is the exploration of alternative cost functions in both model order reduction and LS realization. While the present work has been focused on the LS criterion and the $\mathcal{H}_2$-norm, other norms such as the ℓ_1 or ℓ_∞ norms may be more appropriate in certain applications: e.g., when robustness to outliers or uniform error bounds across time or frequency are of concern. A key open question is whether these alternative norms admit an algebraic (polynomial) formulation that would allow the development of globally optimal methodologies analogous to those presented in this dissertation. Understanding how the critical equations associated with these norms can be characterized, and whether they can be cast into MEPs would significantly broaden the scope of the proposed framework.

In the discussion of the previous section, we highlighted the tradeoff between performing multiple globally optimal experiments on smaller independent data sets and conducting a single larger-scale experiment. From a computational standpoint, the former is clearly more tractable, yet the precise implications of this tradeoff remain to be fully understood. Pursuing precise answers to these questions would provide valuable insights. In this dissertation, we deliberately avoided a statistical viewpoint and instead concentrated on the underlying mathematical framework. Developing the associated statistical theory, and embedding it within the context of our globally optimal formulations, would strengthen the bridge from theoretical advances to practical applications.

All of the methodologies developed in this dissertation rely exclusively on first-order information to characterize the critical points of the optimization problems under consideration. An important direction for future research

is the incorporation of second-order information, such as the Hessian, into the framework. This would make it possible to distinguish minimizers from maximizers or saddle points at the characterization stage, thereby avoiding the need to compute and analyze non-minimizing critical points. Such an extension might substantially enhance the efficiency of the globally optimal algorithms.

The explicit characterizations of the critical points obtained in Chapter 7 provide a foundation for developing a globally optimal framework for SISO misfit modeling. Although numerical experiments using the kernel Riemannian SVD formulation suggest, for low-order cases ($n = 1, 2$), that the set of complex solutions is finite, it remains to be investigated whether this holds for all problem sizes. Addressing this question will likely require tools from algebraic geometry, similarly to the work in Part II in this dissertation.

Appendix A

Appendices

A.1 Gentle Introduction to Optimization and Systems & Control

This section introduces the reader to mathematical optimization, systems theory, and control. It is written for an audience no prior exposure to these subjects. Our goal is to convey the basic ideas in an accessible, intuitive way and to sketch, at a conceptual level, why these disciplines play such an important role in today's world.

A.1.1 Mathematical optimization

Mathematical optimization concerns the problem of determining the best solution to a well-defined quantitative task. In essence, an optimization problem is specified by a criterion quantifying the quality of a solution, and a set of feasible solution candidates. Formally, given an *objective function* $f(\boldsymbol{x})$, the goal is to find $\boldsymbol{x}^\star \in \mathcal{X}$ that minimizes or maximizes f, where $\mathcal{X}$ denotes the *feasible set* of the decision variables $\boldsymbol{x}$. This set may be the entire space $\mathbb{R}^n$ or a subset defined by equality and/or inequality constraints.

Optimization arises naturally in quantitative disciplines, most prominently in engineering, computer science, operations research, and economics, but also in fields such as biology, finance, and quantum mechanics [18]. The availability of large-scale data, driven by trends such as the internet of things and large online platforms, has further amplified the role of optimization

in data analysis, artificial intelligence, system identification, network design, and control. Training deep neural networks, computing maximum likelihood estimators, implementing model predictive control, or performing least squares (LS) system realization fundamentally correspond to solving an optimization problem, see e.g., [14, 37, 84, 120, 131, 154] and references therein. Given its pervasive role, advances in optimization methods often translate into significant progress across multiple scientific and engineering domains.

Although early mathematical optimization can be traced to the work of Fourier (1823) and de la Vallée Poussin (1911) [36], optimization emerged as a distinct field in the mid-20th century. Dantzig's development of the simplex method in 1947 marked a breakthrough for linear programming [34]. In parallel, von Neumann and collaborators contributed foundational results in game theory and economic modeling [180]. The formalization of the Karush–Kuhn–Tucker (KKT) conditions around 1951 provided a cornerstone for nonlinear programming [106]. From these origins, the field expanded into diverse subdisciplines, including integer and combinatorial optimization, convex programming, stochastic optimization, and calculus of variations.

In the unconstrained setting, a point $\boldsymbol{x}^\star$ is called a *critical point* of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ if the gradient vanishes at that point, i.e., $(\nabla f(\boldsymbol{x}^\star) = \mathbf{0})$. Vanishing of the gradient implies that the objective f is "flat" in a sufficiently small neighborhood around the critical point. As such, in non-convex problems, critical points may correspond to local minima, local maxima, or saddle points. These types can be distinguished by examining second-order information: the eigenvalues of the Hessian matrix at the critical point, which describe the curvature of f. In the context of constrained optimization, the notion of a critical point is generalized using the KKT conditions. These conditions characterize points where the gradient of the objective function can be expressed as a linear combination of the gradients of the active constraints.

One of the most common and intuitive applications of optimization arises in the context of data fitting. In particular, many estimation tasks in science and engineering can be naturally framed as optimization problems, where the goal is to find model parameters that best explain observed data according to some criterion. The following example illustrates a particular use-case: *linear regression.*

Example A.1 (Linear Regression)**.** *In this didactical toy example, we consider the linear regression model. Suppose that a scientist hypothesizes that a cell culture grows linearly with time. To test the hypothesis, the scientist measures the culture size over a period of* $N = 40$ *days, yielding data* $\{(t_k, y_k)\}_{k=1,\dots,N}$ *with* $t_k = k$ *(days* 1 *through* N*) and* $y_k \in \mathbb{R}$ *the measured size.*

The scientist assumes the following linear model

$$y_k = \beta_0 + \beta_1 t_k + \varepsilon_k, \qquad k = 1, \dots, N,$$

where $\beta_0, \beta_1 \in \mathbb{R}$ are unknown parameters (the intercept and slope, respectively) and ε_k are zero-mean disturbances, assumed i.i.d. with variance σ^2, caused by inaccuracies in the measurement equipment.

To determine the best-fitting line, one would formulate the following optimization problem: we seek the parameters (β_0, β_1) that minimize the squared difference between the model predictions and the observed data. This leads to the minimization of the residual sum of squares (RSS),

$$\underset{\beta_0, \beta_1}{\text{minimize}} \;\; \text{RSS}(\beta_0, \beta_1) = \frac{1}{2} \sum_{k=1}^{N} \big(y_k - \beta_0 - \beta_1 t_k\big)^2 = \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \right\|_2^2,$$

where $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix}^\mathsf{T}$, the data $\boldsymbol{y} = [y_1, \cdots, y_N]^\mathsf{T}$, and the design matrix $\boldsymbol{X}$ corresponds to

$$\boldsymbol{X} = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_N \end{bmatrix} \in \mathbb{R}^{N \times 2}.$$

This is a convex quadratic optimization problem in the parameters (β_0, β_1), which admits a unique global minimizer.

The first-order optimality conditions, which are necessary and sufficient for a global minimizer by convexity, correspond to

$$\nabla f(\boldsymbol{\beta}) = -\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = -\boldsymbol{X}^\mathsf{T}\boldsymbol{y} + \boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0},$$

leading to the normal equations $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\mathsf{T}\boldsymbol{y}$. *Under the assumption that the matrix $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ is invertible, the unique global minimizer of the optimization problem corresponds to:*

$$\boldsymbol{\beta} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}.$$

Figure A.1 shows an example of measured data along with the corresponding optimal LS fit. For the data depicted in the figure, the regression model appears to support the underlying hypothesis given that the global minimum is relatively small, i.e., the linear fit seems a reasonable approximation. However, the data also exhibits a periodic oscillatory component that the linear model fails to capture.

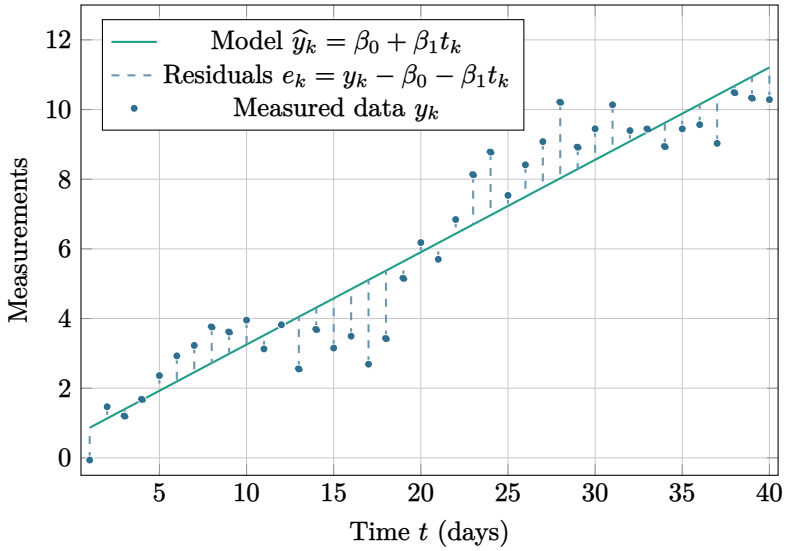


Figure A.1: Visualization of the LS fitting problem considered in Example A.1. The straight line represents the linear model $\widehat{y}_k = \beta_0 + \beta_1 t_k$. The points depict the measured data $\{(t_k, y_k)\}_{k=1,\dots,N}$. The vertical dashed segments are the residuals e_i, whose squared lengths sum to the objective $\mathrm{RSS}(\beta_0, \beta_1) = \sum_{k=1}^{N} (y_k - \beta_0 - \beta_1 t_k)^2$ to be minimized.

A.1.2 Systems theory and control

Systems theory concerns the study of the behavior, interactions, and properties of complex dynamical phenomena. Its focus lies not on the physical world directly, but on the *systems* that generate such phenomena. The notion of a system is intentionally broad: it denotes an interconnected set of components acting together in a dynamical manner. More abstractly, a dynamical system may be defined by the set of all phenomena or *signals* it is capable of generating [186]. Systems may be physical, such as industrial processes, biological networks, or mechanical structures, but can equally be abstract constructs in fields such as economics, sociology, or computer science.

The common framework across these various contexts is the use of mathematical models: equations that describe the system dynamics and/or signals as functions of variables such as time or space. As opposed to a static system, a dynamical system is characterized by the evolution of its internal state over time, typically governed by differential or difference equations. When such a system can be influenced by external excitations, these excitations are referred to as *inputs*.

Conversely, the observable quantities or measurable phenomena generated by the system in response to its internal dynamics and external excitations are the so-called *outputs*. These mathematical models describing the relationship between inputs, states, and outputs form the foundation for simulation, analysis, prediction, monitoring, and control in numerous scientific and engineering disciplines.

The notion of *control* refers to the manipulation of a system's inputs or internal configuration to influence its behavior in a desired manner. Control objectives may include stabilizing unstable dynamics, tracking reference trajectories, rejecting disturbances, optimizing performance, or enforcing safety constraints. In that sense, systems theory is both descriptive, explaining observed behaviors (analysis), and prescriptive, guiding the design and modification of systems (synthesis) [9].

The roots of systems theory trace back to early work in mechanics and physics, where mathematical models were developed to explain motion, energy, and interactions. In the mid-20th century, the rise of electronic computing and the needs of aerospace and industrial automation led to the formalization of control theory as a distinct discipline. Significant contributions included the development of feedback control principles, frequency-domain methods, and state-space formulations. See e.g., [9, 137] and references therein.

From the 1970s onward, modern control theory integrated linear algebra, numerical optimization, and systems analysis. Control became deeply intertwined with optimization, with techniques such as model predictive control exploiting advances in real-time convex optimization. Simultaneously, thanks to the universality of its mathematical abstractions, systems theory expanded beyond engineering into areas such as biology and economics.

Developing a mathematical model of a dynamical system can be done in various ways, depending on the available information and intended application. The three main paradigms are:

1. *White-box modeling*: Models are derived from first principles, grounded in physical laws such as Newton's equations, Maxwell's equations, or conservation principles. All parameters have a direct physical interpretation.

2. *Black-box modeling*: Models are constructed purely from observed input-output data, without explicit usage of underlying physical laws. This approach uses statistical and optimization techniques to fit a parameterized model structure to given data.

3. *Grey-box modeling*: Also known as semi-physical modeling, this approach combines the strengths of white- and black-box methods. A partial physical model is formulated, and unknown parameters, or even certain dynamical components, are estimated from data. This allows incorporation of known physics while retaining flexibility to account for unmodeled dynamics or parameter uncertainty. It is particularly valuable in engineering applications where partial knowledge exists.

The paradigms represent a trade-off between physical interpretability, data requirements, and modeling flexibility. White-box models come with strong interpretability and extrapolation capabilities. However, they require extensive domain knowledge and may become difficult to construct for complex systems. Black-box models are comparatively quick to develop when sufficient high-quality data are available and can capture complex dynamics without detailed physical insight or user intervention, yet they often lack interpretability and may not generalize well, that is, they perform poorly outside the domain of the training data. Grey-box models seek a balance: by embedding partial physical knowledge within a data-driven estimation framework, they preserve interpretability where possible while improving adaptability to effects not modeled by the considered first principles. This hybrid nature, however, may still require substantial modeling effort and careful parameter estimation to avoid inconsistencies between the physical and data-driven components.

The following example revisits the measured data from Example A.1, this time modeling it as the output of a known linear dynamical system driven by the recorded room temperature. We show that, with appropriate initialization, the dynamical single-input single-output (SISO) system accurately reproduces the observed measurements.

Example A.2 (Linear Dynamical SISO System)**.** *We revisit the measured data* $(t_k, y_k)_{k=1,\dots,N}$ *from the previous example, representing the size of a cell culture observed over* $N = 40$ *days. Unlike the earlier static model, we now account for the fact that culture size is dynamically influenced by ambient temperature. In particular, the culture size on day* k *depends not only on the temperature that day but also on previous temperatures and the cell culture sizes measured in the previous days.*

First principles indicate that the measured culture size y_k *can be effectively modeled as the output of a discrete-time linear time-invariant (LTI) system, described by the following second-order difference equation:*

$$y_k = 0.3y_{k-1} - 0.02y_{k-2} + u_k + 1.1u_{k-1} + 0.1u_{k-2}, \quad k = 3, \dots, N,$$

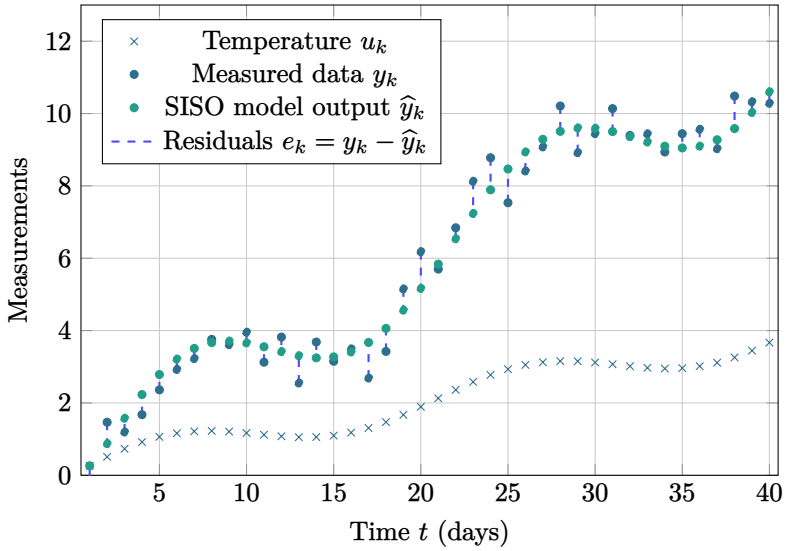


Figure A.2: Comparison between measured data y_k and the output $\widehat{y}_k$ of a SISO linear dynamical system excited by the temperature u_k as input signal, corresponding to the experimental setup described in Example A.2. Vertical dotted lines indicate the residuals between the model predictions and the actual measurements.

where $u_k \in \mathbb{R}$ denotes the recorded room temperature on day k. Equivalently, the system can be represented by the discrete-time SISO transfer function

$$H(z) = \frac{(z+0.1)(z+1)}{(z-0.1)(z-0.2)}.$$

Simulating this model with the recorded temperature data as input and with appropriate initial conditions y_1, y_2 produces an output signal that closely matches the measured culture size (see Figure A.2). The model captures both the overall growth trend and the periodic oscillations in the data, providing a substantially better fit than the static LS model shown in Figure A.1. This illustrates that the observed process is more accurately described by a dynamical system than by a purely static model.

Contrary to the previous example, in which the parameters of the SISO model were derived using first principles, *system identification* builds mathematical models of dynamical systems from measured data [120]. In its broadest sense, the task is to determine, from observed input-output data, a parametrized

model structure and associated parameters so that the resulting model captures the dynamical behavior of the true system. The model is typically chosen from a user-specified model class, reflecting prior knowledge about the system's nature and complexity, e.g., linear vs. nonlinear, or time-variant vs. time-invariant.

A dynamical system is therefore identified using the following four stages:

1. *Experiment design*: Deciding which variables to measure, under what excitation conditions, and over what time interval.
2. *Model selection*: Choosing a class of models (e.g., autoregressive models or input-output state-space representations) that can represent the system's dynamics.
3. *Parameter estimation*: Determining the model parameters that align best with the observed data.
4. *Validation*: Assessing the predictive capability of the identified model on independent data sets to verify that it captures the essential system dynamics without overfitting measurement noise or other disturbances.

If the validation indicates bad performance, the identification loop iterates by revisiting earlier steps, refining the experiment design, reconsidering the model class, or adjusting the parameter estimation procedure, until the model achieves the desired accuracy and robustness criteria.

Observe that in the context of system identification, all elements of a mathematical optimization problem are naturally present. The model equations of the chosen class serve as the constraints, defining the feasible set of all admissible parameter-data combinations. The objective function corresponds to the criterion used to assess model quality, such as the magnitude of the misfits or the prediction error. The decision variables are the parameters of the selected model class, and, when misfits are used, also the modified data points. Solving the identification problem thus amounts to finding the parameter values (and, if applicable, modified data) that minimize the chosen error measure while strictly satisfying the governing model equations. This observation clearly situates system identification within the realm of mathematical optimization.

A.2 Polynomial Systems and MEPs

This section offers a concise introduction to polynomial systems and multiparameter eigenvalue problems (MEPs). Our goal is not to provide a textbook-style

treatment, but rather to give an intuitive account of how to pass from one formulation to the other.

A.2.1 Polynomial system solving

The fundamental observation underlying several results in this dissertation is that many of the optimization problems encountered in system identification and model reduction are polynomial in nature. I.e., they can be expressed in the following form:

$$\begin{aligned} &\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad f(\boldsymbol{x}) \\ &\text{subject to} \quad g_i(\boldsymbol{x}) = 0 \quad \text{for} \quad i \in \{1, ..., k\}, \end{aligned}$$

with $f(\boldsymbol{x})$ and $g_i(\boldsymbol{x})$ for all $i \in \{1, ..., k\}$ polynomials in the variables $\boldsymbol{x} \in \mathbb{R}^n$. In such cases, the first-order necessary conditions for optimality, obtained via the method of Lagrange multipliers, reduce to a system of multivariate polynomial equations:

$$\begin{cases} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\ell}) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) + \sum_{i=1}^{k} \ell_i \nabla_{\boldsymbol{x}} g_i(\boldsymbol{x}) = 0 \\ \partial \mathcal{L}(\boldsymbol{x}, \boldsymbol{\ell}) / \partial \ell_i = g_i(\boldsymbol{x}) = 0 \quad \text{for} \quad i \in \{1, ..., k\}, \end{cases} \tag{A.1}$$

where $\boldsymbol{\ell} = [\ell_0, \ldots, \ell_{k-1}]^{\mathsf{T}} \in \mathbb{R}^k$ are Lagrange multipliers. This constitutes a system of $n + k$ multivariate polynomial equations in $n + k$ unknowns. Solving this system for its set of real-valued common roots leads to the complete set of critical points of the original optimization problem.

Since the global minimizer must be among these critical points, globally optimal polynomial optimization can thus be reformulated as a polynomial root-finding problem: once all the real-valued solutions of (A.1) are computed, evaluating the objective function at each candidate suffices to identify the globally optimal solution(s).

Algebraic geometry is the mathematical discipline that studies polynomials and the geometric objects they define. In this setting, the solution set of a system of polynomial equations is regarded as a geometric object, an algebraic variety, and root-finding amounts to computing the points (or more generally, the components) that make up this variety.

Example A.3. *Consider the following variety:*

$$V = \{(x, y, z) : xy^2 - 10y - 78z = 0\} \subseteq \mathbb{R}^3.$$

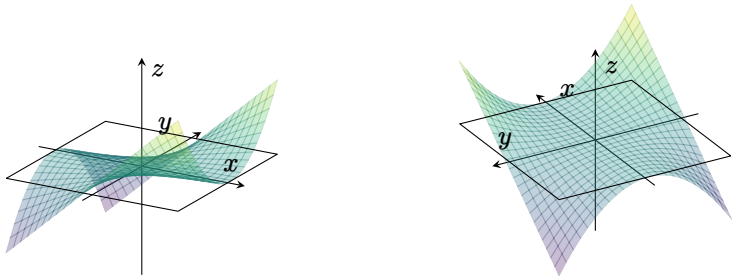


Figure A.3: Surface plots of the variety V over the domain $(x, y) \in [-10, 10]^2$, where $V = \{(x, y, z) \in \mathbb{R}^3 : xy^2 - 10y - 78z = 0\}$. The black square marks the projection of the domain onto the plane $z = 0$.

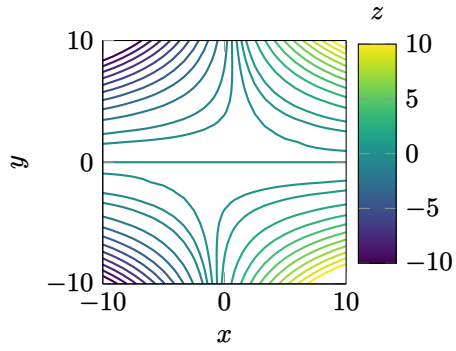


Figure A.4: Contour plot of a function.

It is defined by a single polynomial equation in three variables, hence it is a two-dimensional surface embedded in the three-dimensional space $\mathbb{R}^3$, as visualized in Figures A.3 and A.4.

Most engineering applications are concerned with zero-dimensional varieties, i.e., systems of polynomial equations for which the solution set consists of finitely many isolated points[1]. Over the past decades, a spectrum of approaches for computing these zero-dimensional varieties has emerged, spanning from purely symbolic to numerical techniques. While we provide only a high-level overview, readers seeking a more detailed introduction to (numerical) algebraic geometry may consult [33] and the references therein. For polynomial system solving in engineering applications, an excellent entry point is [155].

[1]Reconsider for example the surface in Figure A.3. Intersecting it with a generic line (which amounts to adding two additional independent equations to the defining equations of the variety V), would lead to a zero-dimensional solution set.

One of the oldest computational tools is Buchberger's Gröbner basis algorithm [22], which symbolically reduces a polynomial system to a canonical form from which the solutions can be derived. Gröbner basis methods are exact and theoretically powerful, but their reliance on symbolic manipulation makes them poorly suited for floating-point arithmetic and inexact data. In applied settings, this issue often results in numerical instability. Recent work has sought to improve robustness (see, e.g., [130]).

A breakthrough at the end of the 20th century came with *homotopy continuation* [155, 179], establishing the field of *numerical algebraic geometry*. These methods solve a polynomial system by embedding it into a continuous deformation (homotopy) from a simpler system with known roots. The solutions of the start system are tracked along continuous paths until they reach the solutions of the target system. Homotopy methods guarantee, in principle, that all isolated roots are found and are naturally parallelizable, since each solution path can be followed independently.

The main challenges lie in constructing effective start systems and in maintaining numerical stability along ill-conditioned paths. Modern implementations combine predictor-corrector steps, adaptive step sizes, and certification routines to address these difficulties. Despite the challenges, homotopy continuation is one of the most powerful general-purpose approaches for polynomial system solving.

Another line of development stems from Stetter's observation [159] that solving a polynomial system can be reduced to an eigenvalue problem. This insight led to *symbolic-numeric methods*, which use symbolic computations to construct so-called multiplication matrices, the eigenvalues of which reveal the coordinates of the solution to the system of equations. These multiplication matrices represent the action of variables on the quotient algebra defined by the system of polynomial equations. While elegant, modern symbolic-numeric methods often depend on Gröbner bases for construction of the multiplication matrices, thereby coming with similar problems as the purely symbolic approaches. To address this, so-called border basis methods were developed as a more numerically robust alternative to Gröbner bases for constructing multiplication matrices, see e.g., [130].

More recently, purely numerical eigenvalue-based approaches have emerged, centered around the *Macaulay matrix* [52, 173]. In this framework, the multiplication matrices are no longer built through symbolic computation, but are recovered directly from the null space of a large, structured coefficient matrix. This null space can be computed robustly using the singular value decomposition (SVD), which is a cornerstone of numerical linear algebra with established, highly optimized implementations. By exploiting shift-invariance

properties of the obtained numerical basis for the null space, one constructs multiplication matrices whose eigenvalues correspond to the solutions of the system. In this way, polynomial system solving reduces to an SVD and eigenvalue computations, both of which are fully compatible with floating-point arithmetic. The approach builds on classical algebraic insights, tracing back to Sylvester's 19th-century work on resultants [164] and Macaulay's introduction of structured matrices constructed from the coefficients of polynomials [121].

In this dissertation, we will not focus on the development of solvers themselves, but rather employ existing implementations. Specifically, we make use of the homotopy continuation techniques from [20] and a "blocked" variant of the Macaulay matrix method [173].

A.2.2 From polynomial system to MEP

A limitation of Macaulay matrix based root-finding methods is their scalability: the size of the coefficient matrix increases combinatorially with the number of variables appearing in the polynomial equations, making the null-space computations prohibitively expensive for larger systems. One way to (at least partially) alleviate this bottleneck is to exploit the linear structure that is often present within the system. Specifically, one can partition the variables $\boldsymbol{x} \in \mathbb{R}^n$ into two groups: those that appear only linearly, denoted by $\boldsymbol{x}_{\text{lin}} \in \mathbb{R}^\ell$, and those that enter nonlinearly, denoted by $\boldsymbol{x}_{\text{non}} \in \mathbb{R}^{n-\ell}$. By "appearing linearly", we mean that the equations involve no powers of the variables in $\boldsymbol{x}_{\text{lin}}$ and no cross-products between them, i.e., each variable in $\boldsymbol{x}_{\text{lin}}$ occurs with degree at most one and without interaction with the other variables in $\boldsymbol{x}_{\text{lin}}$:

$$\begin{cases} f_1(\boldsymbol{x}_{\text{lin}}, \boldsymbol{x}_{\text{non}}) = h_1(\boldsymbol{x}_{\text{non}}) + (\boldsymbol{x}_{\text{lin}})^\mathsf{T} \boldsymbol{g}_1(\boldsymbol{x}_{\text{non}}) = 0 \\ \vdots \\ f_k(\boldsymbol{x}_{\text{lin}}, \boldsymbol{x}_{\text{non}}) = h_k(\boldsymbol{x}_{\text{non}}) + (\boldsymbol{x}_{\text{lin}})^\mathsf{T} \boldsymbol{g}_k(\boldsymbol{x}_{\text{non}}) = 0 \end{cases}$$

where for each equation $\boldsymbol{g}_i(\boldsymbol{x}_{\text{non}}) = [g_{i,1}(\boldsymbol{x}_{\text{non}}), \ldots, g_{i,\ell}(\boldsymbol{x}_{\text{non}})]^\mathsf{T} \in \mathbb{R}[\boldsymbol{x}_{\text{non}}]^\ell$ is a vector of polynomial entries in the variables $\boldsymbol{x}_{\text{non}}$, and $h_i \in \mathbb{R}[\boldsymbol{x}_{\text{non}}]$.

By isolating the variables that appear only linearly and collecting them into an eigenvector, the system can be formulated as a matrix-vector product, thereby constituting an eigenvalue problem:

$$\boldsymbol{M}(\boldsymbol{x}_{\text{non}}) \begin{bmatrix} 1 \\ \boldsymbol{x}_{\text{lin}} \end{bmatrix} = \begin{bmatrix} h_1(\boldsymbol{x}_{\text{non}}) & g_{1,1}(\boldsymbol{x}_{\text{non}}) & \cdots & g_{1,\ell}(\boldsymbol{x}_{\text{non}}) \\ h_2(\boldsymbol{x}_{\text{non}}) & g_{2,1}(\boldsymbol{x}_{\text{non}}) & \cdots & g_{2,\ell}(\boldsymbol{x}_{\text{non}}) \\ \vdots & \vdots & & \vdots \\ h_k(\boldsymbol{x}_{\text{non}}) & g_{k,1}(\boldsymbol{x}_{\text{non}}) & \cdots & g_{k,\ell}(\boldsymbol{x}_{\text{non}}) \end{bmatrix} \begin{bmatrix} 1 \\ \boldsymbol{x}_{\text{lin}} \end{bmatrix} = \boldsymbol{0}.$$

This reformulation constitutes a (rectangular) multiparameter eigenvalue problem (MEP). The matrix $\boldsymbol{M}(\boldsymbol{x}_{\text{non}}) \in \mathbb{R}^{k\times(\ell+1)}$, which is a matrix polynomial over $\boldsymbol{x}_{\text{non}}$, has typically more rows than columns—since the number of equations k exceeds the number of linear variables ℓ—such that it is of full column rank for most $\boldsymbol{x}_{\text{non}}$. Indeed, for the original system of polynomials to admit a zero-dimensional solution set, it is necessary that the number of equations exceeds the number of variables, i.e., $k > n$, and $n \geq \ell$.

Solving the original polynomial system is now equivalent to identifying those values of the nonlinear variables $\boldsymbol{x}_{\text{non}}$ for which the matrix $\boldsymbol{M}(\boldsymbol{x}_{\text{non}})$ drops rank. These values for $\boldsymbol{x}_{\text{non}}$ are referred to as the *eigentuples* of the MEP. Once an eigentuple has been found, the corresponding values of the linear variables $\boldsymbol{x}_{\text{lin}}$ are recovered by substituting the eigentuple into $\boldsymbol{M}(\boldsymbol{x}_{\text{non}})$, computing a vector in its null space, and normalizing it so that the first entry equals one.

Example A.4. *Consider the polynomial system in variables* $\{a_0, b_0, g_0, g_1\}$*:*

$$
\begin{aligned}
0 = f_0 &= -g_0 {a_0}^2 - 10a_0 - 78b_0, \\
0 = f_1 &= 9a_0 - 49b_0 + 2a_0 g_0 - {a_0}^2 g_1 - 10, \\
0 = f_2 &= a_0 - 12b_0 - g_0 + 2a_0 g_1 + 9, \\
0 = f_3 &= 1 - g_1 - b_0.
\end{aligned}
$$

This system of equations will be derived in Example 2.10.

The variables g_0, g_1 *and* b_0 *appear linearly only, whereas the variable* a_0 *enters nonlinearly. Hence we set*

$$
\boldsymbol{x}_{\text{lin}} := \begin{bmatrix} b_0 \\ g_0 \\ g_1 \end{bmatrix} \in \mathbb{R}^3, \qquad \boldsymbol{x}_{\text{non}} := \begin{bmatrix} a_0 \end{bmatrix} \in \mathbb{R}.
$$

Collecting the constant terms and the coefficients of u *and* v *in each equation, we rewrite the system of equations as a matrix–vector product*

$$
\underbrace{\begin{bmatrix} h_1(a_0) & g_{1,1}(a_0) & g_{1,2}(a_0) & g_{1,3}(a_0) \\ h_2(a_0) & g_{2,1}(a_0) & g_{2,2}(a_0) & g_{2,3}(a_0) \\ h_3(a_0) & g_{3,1}(a_0) & g_{3,2}(a_0) & g_{3,3}(a_0) \\ h_4(a_0) & g_{3,1}(a_0) & g_{3,2}(a_0) & g_{4,3}(a_0) \end{bmatrix}}_{=:\ \boldsymbol{M}(a_0)\in\mathbb{R}^{4\times 4}} \begin{bmatrix} 1 \\ b_0 \\ g_0 \\ g_1 \end{bmatrix} = \mathbf{0},
$$

where

$$\boldsymbol{M}(a_0) = \begin{bmatrix} 10a_0 & 78 & a_0^2 & 0 \\ 10 - 9a_0 & 49 & -2a_0 & a_0^2 \\ -9 - a_0 & 12 & 1 & -2a_0 \\ -1 & 1 & 0 & 1 \end{bmatrix}.$$

Finding solutions of the original polynomial system is reduced to finding eigentuples (a_0) *for which* $\boldsymbol{M}(a_0)$ *drops rank, i.e.,* $\operatorname{rank} \boldsymbol{M}(a_0) < 4$. *For any such eigentuple, the corresponding* (b_0, g_0, g_1) *are found by computing a nonzero vector in the right null space of* $\boldsymbol{M}(a_0)$ *and normalizing it so that the first entry equals* 1.

Notice that when the solution set of the original polynomial system is zero-dimensional, that is, it consists of a finite number of isolated points, the rank of $\boldsymbol{M}(\boldsymbol{x}_{\text{non}})$ can drop by at most one at each eigentuple, ensuring that every eigentuple $\boldsymbol{x}_{\text{non}}$ admits a unique eigenvector. Indeed, if instead, $\boldsymbol{M}(\boldsymbol{x}_{\text{non}})$ has an m-dimensional null space with $m > 1$, then each eigentuple gives rise to an infinite family of solutions, as $\boldsymbol{x}_{\text{lin}}$ can be chosen from an arbitrary vector within the null space.

Also note that in the aforementioned reformulation of a system of polynomial equations as an MEP, the equivalence of the two solution sets holds only when, for each eigentuple, the unique eigenvector of the MEP has a nonzero leading entry, so that it can be appropriately normalized to recover the corresponding linear variables $\boldsymbol{x}_{\text{lin}}$. In Part II of this dissertation, particular attention is paid to verify these conditions.

A.2.3 MEP solvers

The eigentuples of an MEP can, in principle, be obtained as the common roots of the polynomial system defined by setting all maximal minors of the matrix $\boldsymbol{M}(\boldsymbol{x}_{\text{non}})$ to zero. In practice, however, determinant-based formulations are notoriously ill-conditioned under finite-precision arithmetic and thus unsuitable for reliable numerical computation. More robust alternatives have emerged in recent years.

The (block) Macaulay framework [173, 175, 176], for instance, constructs a structured matrix from the coefficients of $\boldsymbol{M}(\boldsymbol{x}_{\text{non}})$. Because the number of effective variables is reduced from n to $n - \ell$, the block-Macaulay matrix grows more moderately in size than the standard Macaulay matrix used for the

original polynomial system, significantly decreasing the computational costs while preserving the numerical stability of eigenvalue-based methods.[2]

An alternative line of work [90, 138] employs linearization and projection techniques to solve an MEP. Both approaches can be effectively used to compute the eigentuples. As with polynomial solvers, these methods naturally operate over $\mathbb{C}$, so that some of the retrieved eigentuples must be discarded when only real solutions are of interest.

While the discussion above presented MEPs as a reformulation tool for systems of polynomial equations, this perspective ignores their broader significance. In practice, MEPs arise naturally across applied mathematics and engineering, where they generalize the classical polynomial eigenvalue problem (PEP), which corresponds to the case described above when $\boldsymbol{x}_{\text{non}}$ consists of a single element. The PEP is a cornerstone in the study of mechanical vibrations, circuit dynamics, and other physical phenomena, and the multivariate generalization given by MEPs introduces richer problem classes, with instances encountered in e.g., separation-of-variables techniques for boundary value problems and cryptographic analysis. Unlike in the context of polynomial systems, the associated eigenvectors in such applications need not be normalized to enforce a specific nonzero entry. We refer the reader to [177] for a more detailed treatment of the (rectangular) MEP.

[2] A further reduction of the size of the obtained matrix is theoretically possible by exploiting the Newton polytope of the system [16], though in practice the polytope computation often becomes the dominant bottleneck.

Bibliography

[1] Abatzoglou, T. J., Mendel, J. M., and Harada, G. The constrained total least squares technique and its applications to harmonic superresolution. *IEEE Transactions on Signal Processing 39*, 5 (1991), 1070–1087.

[2] Adcock, R. J. Note on the method of least squares. *The Analyst 4*, 6 (1877), 183–184.

[3] Adcock, R. J. A problem in least squares. *The Analyst 5*, 2 (1878), 53–54.

[4] Agudelo, O. M., Vermeersch, C., and De Moor, B. Globally optimal H_2-norm model reduction: A numerical linear algebra approach. *IFAC PapersOnLine (Special Issue: Proceedings of MTNS 2020) 54*, 9 (2021), 564–571.

[5] Ahmad, M., Frangos, M., and Jaimoukha, I. Second order $\mathcal{H}_2$ optimal approximation of linear dynamical systems. *IFAC Proceedings Volumes 44*, 1 (2011), 12767–12771.

[6] Aigrain, P., and Williams, E. Synthesis of n-reactance networks for desired transient response. *Journal of Applied Physics 20*, 6 (1949), 597–600.

[7] Alsubaie, F. F. *$\mathcal{H}_2$ Optimal Model Reduction for Linear Dynamical Systems and the Solution of Multiparameter Matrix Pencil Problems.* PhD thesis, Imperial College London, London, UK, 2019.

[8] Ando, S. Frequency-domain prony method for autoregressive model identification and sinusoidal parameter estimation. *IEEE Transactions on Signal Processing 68* (2020), 3461–3470.

[9] Antoulas, A. C. *Mathematical System Theory.* Berlin, Germany: Springer, 1991, ch. Introduction, pp. 1–8.

[10] Antoulas, A. C. *Approximation of Large-Scale Dynamical Systems.* SIAM, Philadelphia, PA, 2005.

[11] Aoki, M., and Yue, P. C. On a priori error estimates of some identification methods. *IEEE Transactions on Automatic Control 15*, 5 (1970), 541–548.

[12] Auton, J. R., and Van Blaricum, M. L. Investigation of procedures for automatic resonance extraction from noisy transient electromagnetics data, vol. ii – appendices. Tech. rep., Defense Technical Information Center, Effects Technology Inc., Santa Barbara, CA, US, 1981.

[13] Baratchart, L. Existence and generic properties of l_2 approximants for linear systems. *IMA Journal of Mathematical Control & Information 3*, 2-3 (1986), 89–101.

[14] Barratt, . H., and Boyd, S. P. *Linear controller design : limits of performance.* Prentice Hall information and system sciences series. Prentice Hall, 1991.

[15] Batselier, K., Dreesen, P., and De Moor, B. Prediction error method identification is an eigenvalue problem. In *Proceedings 16th IFAC Symposium on System Identification, SYSID* (Brussels, Belgium, 2012), pp. 221–226.

[16] Bender, M. R., and Telen, S. Yet another eigenvalue algorithm for solving polynomial systems. Tech. rep., 2022. Preprint at `https://arxiv.org/abs/2105.08472`.

[17] Beurling, A. On two problems concerning linear transformations in Hilbert space. *Acta Mathematica 81* (1949), 239–255.

[18] Boyd, S., and Vandenberghe, L. *Convex Optimization.* Cambridge University Press, 2004.

[19] Breiding, P., Kohn, K., and Sturmfels, B. *Metric Algebraic Geometry.* Oberwolfach Seminars. Birkhäuser Cham, 2024.

[20] Breiding, P., and Timme, S. HomotopyContinuation.jl: A package for homotopy continuation in Julia. In *International Congress on Mathematical Software* (2018), pp. 458–465.

[21] Bresler, Y., and Macovski, A. Exact maximum likelihood parameter estimation of superimposed exponential signals in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing 34*, 5 (1986), 1081–1089.

[22] Buchberger, B. *Ein Algorithmus zum Auffinden der Basiselemente des Restklassenringes nach einem nulldimensionalen Polynomideal.* PhD thesis, University of Innsbruck, 1965.

[23] Bultheel, A., and De Moor, B. Rational approximation in linear systems and control. *Journal of Computational and Applied Mathematics 121*, 1–2 (2000), 355–378.

[24] Bunse-Gerstner, A., Kubalińska, D., Vossen, G., and Wilczek, D. h_2-norm optimal model reduction for large scale discrete dynamical MIMO systems. *Journal of Computational and Applied Mathematics 223*, 5 (2010), 1202–1216.

[25] Cadzow, J. Signal enhancement: A composite property mapping algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing 36*, 2 (1988), 49–62.

[26] Chen, H., Van Huffel, S., Van Den Boom, A., and Van Den Bosch, P. Subspace-based parameter estimation of exponentially damped sinusoids using prior knowledge of frequency and phase. *Signal Processing 59*, 1 (1997), 129–136.

[27] Chen, H., Van Huffel, S., and Vandewalle, J. Improved methods for exponential parameter estimation in the presence of known poles and noise. *IEEE Transactions on Signal Processing 45*, 5 (1997), 1390–1393.

[28] Cifuentes, D. A convex relaxation to compute the nearest structured rank deficient matrix. *SIAM Journal on Matrix Analysis and Applications 42*, 2 (2021), 708–729.

[29] Claes, R., Jarlebring, E., Meerbergen, K., and Upadhyaya, P. Linearizable eigenvector nonlinearities. *SIAM Journal on Matrix Analysis and Applications 43*, 2 (2022), 764–786.

[30] Claes, R., Meerbergen, K., and Telen, S. Contour integration for eigenvector nonlinearities. *SIAM Journal on Matrix Analysis and Applications 44*, 4 (2023), 1619–1644.

[31] Collins, E., Phillips, D., and Hyland, D. Robust decentralized control laws for the ACES structure. *IEEE Control Systems Magazine 11*, 3 (1991), 62–70.

[32] Cox, D. A., Little, J., and O'Shea, D. *Using Algebraic Geometry*, 2nd ed. Graduate Texts in Mathematics. Springer, 2005.

[33] Cox, D. A., Little, J., and O'Shea, D. *Ideals, Varieties, and Algorithms.* Undergraduate Texts in Mathematics. Springer Cham, 2015.

[34] Dantzig, G. B. Linear programming. *Operations Research 50*, 1 (2002), 42–47.

[35] De Cock, K., and De Moor, B. Multiparameter eigenvalue problems and shift-invariance. *IFAC-PapersOnLine (Special Issue: Proceedings of MTNS 2020) 54*, 9 (2021), 159–165.

[36] de la Vallée Poussin, C. J. Sur la méthode de l'approximation minimum. *Annales de la Société Scientifique de Bruxelles, deuxième partie. 35* (1911), 1–16.

[37] De Mello, R, F., and Ponti, M. A. *A Practical Approach on the Statistical Learning Theory.* Springer Nature Switzerland AG, 2018.

[38] De Moor, B. Structured total least squares and l_2 approximation problems. *Linear Algebra an Its Applications 188–189* (1993), 163–205.

[39] De Moor, B. Dynamic total linear least squares. In *Proc. of the 10th IFAC Symposium on System Identification* (Copenhagen, Denmark, July 1994), vol. 10-3, pp. 159 – 164.

[40] De Moor, B. Total least squares for affinely structured matrices and the noisy realization problem. *IEEE Transactions on Signal Processing 42*, 11 (1994), 3104–3113.

[41] De Moor, B. Continuous-time algorithms for the riemannian SVD. In *Pro. of the 34th Conference on Decision & Control* (New Orleans, LA, USA, Dec. 1995), pp. 1084–1090.

[42] De Moor, B. The riemannian singular value decomposition. In *SVD and Signal Processing III*, M. Moonen and B. De Moor, Eds., Algorithms, Architectures and Applications. Elsevier, 1995, pp. 61–78.

[43] De Moor, B. Linear system identification, structured total least squares and the riemannian SVD. In *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*, S. Van Huffel, Ed. SIAM, Philadelphia, USA, 1997, pp. 225–238.

[44] De Moor, B. Structured total least squares for hankel matrices. In *Communications, Computation, Control and Signal Processing*, A. Paulraj, V. Roychowdhury, and C. D. Schaper, Eds. Springer New York, NY, 1997, pp. 243–258.

[45] De Moor, B. Least squares optimal realisation of autonomous LTI systems is an eigenvalue problem. *Communications in Information and Systems 20*, 2 (2020), 163–207.

[46] De Moor, B., and Roorda, B. L_2-optimal linear system identification: Structured total least squares for SISO systems. In *Proc. of the 33rd IEEE Conference on Decision and Control (CDC)* (Lake Buena Vista, Florida, USA, 1994), pp. 2874 – 2879.

[47] De Moor, B., Van Overschee, P., and Schelfhout, G. H_2 model reduction for SISO systems. In *Proceedings of the 12th IFAC World Congress* (Sydney, Australia, July 1993), pp. 227–230.

[48] de Prony, G. C. F. M. R. Essai expérimental et analytique sur les lois de la dilatabilité des fluides élastiques et sur celles de la force expansive de la vapeur d'eau et de la vapeur de l'alkool à différentes températures. *Journal de l'Ecole Polytechnique de Paris 1* (1795), 24–76.

[49] Domanov, I., and De Lathauwer, L. Generic uniqueness of a structured matrix factorization and applications in blind source separation. *Journal of Selected Topics in Signal Processing 10*, 4 (2016), 701–711.

[50] Dowling, E. M., DeGroat, R. D., and Linebarger, D. Exponential parameter estimation in the presence of known components and noise. *IEEE Transactions on Antennas and Propagation 42*, 5 (1994), 590–599.

[51] Draisma, J., Horobeţ, E., Ottaviani, G., Sturmfels, B., and Rekha, R. T. The euclidean distance degree of an algebraic variety. *Foundations of Computational Mathematics 16*, 1 (2016), 99–149.

[52] Dreesen, P., Batselier, K., and De Moor, B. Back to the roots: Polynomial system solving, linear algebra, systems theory. *IFAC Proceedings Volumes: Proceedings of the 16th IFAC Symposium on System Identification 45*, 16 (July 2012), 1203–1208.

[53] Dreesen, P., and De Moor, B. *Model-Based Control: Bridging Rigorous Theory and Advanced Technology.* Springer US, 2009, ch. Polynomial Optimization Problems are Eigenvalue Problems, pp. 49–68.

[54] Eckart, C., and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika 1*, 3 (1936), 211–218.

[55] Eisenbud, D. *Commutative Algebra: With a View Toward Algebraic Geometry.* Springer New York, 1995.

[56] Eisenbud, D., and Harris, J. *3264 and All That: A Second Course in Algebraic Geometry.* Cambridge University Press, 2016.

[57] Elad, M., Milanfar, P., and Golub, G. H. Shape from moments—an estimation theory perspective. *IEEE Transactions on Signal Processing 52*, 7 (2004), 1814–1829.

[58] Evans, A. E., and Fischl, R. Optimal least-squares time-domain synthesis of recursive digital filters. *IEEE Transactions on Audio and Electroacoustics 21*, 1 (1973), 61–65.

[59] Everitt, N., Galrinho, M., and Hjalmarsson, H. Open-loop asymptotically efficient model reduction with the Steiglitz-McBride method. *Automatica 89* (2018), 221–234.

[60] Evert, E., and De Lathauwer, L. Guarantees for existence of a best canonical polyadic approximation of a noisy low-rank tensor. *SIAM Journal on Matrix Analysis and Applications 43*, 1 (2022), 328–369.

[61] Evert, E., and De Lathauwer, L. On best low rank approximation of positive definite tensors. *SIAM Journal on Matrix Analysis and Applications 44*, 2 (2023), 867–893.

[62] Evert, E., Vandecappelle, M., and De Lathauwer, L. Canonical polyadic decomposition via the generalized schur decomposition. *IEEE Signal Processing Letters 29* (2022), 937–941.

[63] Evert, E., Vandecappelle, M., and De Lathauwer, L. A recursive eigenspace computation for the canonical polyadic decomposition. *SIAM Journal on Matrix Analysis and Applications 43*, 1 (2022), 274–300.

[64] Fan, H., and Doroslovački, M. On "global convergence" of Steiglitz–McBride adaptive algorithm. *IEEE Transactions on Circuits and Systems–II: Analog and Digital Signal Processing 40*, 2 (1993), 73–87.

[65] Flagg, G., Beattie, C., and Gugercin, S. Convergence of the Iterative Rational Krylov Algorithm. *Systems & Control Letters 61*, 6 (2012), 688–691.

[66] Genin, Y., Van Dooren, P., Kailath, T., Delosme, J. M., and Morf, M. On Σ-lossless transfer functions and related questions. *Linear Algebra and its Applications 50* (1983), 251 – 275.

[67] Gillard, J., and Usevich, K. Structured low-rank matrix completion for forecasting in time series analysis. *International Journal of Forecasting 34*, 4 (2018), 582–597.

[68] Gillard, J., and Usevich, K. Hankel low-rank approximation and completion in time series analysis and forecasting: a brief review. *Statistics and Its Interface 16*, 2 (2023), 287–303.

[69] Gillard, J., and Zhigljavsky, A. Optimization challenges in the structured low rank approximation problem. *Journal of Global Optimization 57*, 3 (2013), 733–751.

[70] Gillard, J., and Zhigljavsky, A. Stochastic algorithms for solving structured low-rank matrix approximation problems. *Communications in Nonlinear Science and Numerical Simulation 21*, 1-3 (2015), 70–88.

[71] Gillis, N., and Glineur, F. Low-rank matrix approximation with weights or missing data is NP-hard. *SIAM Journal on Matrix Analysis and Applications 32*, 4 (2011), 1149–1165.

[72] Glover, K. All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds. *International Journal of Control 39*, 6 (1984), 1115–1193.

[73] Golub, G. H., and Pereyra, V. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis 10*, 2 (1973), 413–432.

[74] Golub, G. H., and Pereyra, V. Separable nonlinear least squares: The variable projection method and its applications. *Inverse Problems 19*, 2 (2003), R1–R26.

[75] Golub, G. H., and Van Loan, C. *Matrix Computations*, 4th ed. Johns Hopkins University Press, Baltimore, US, 2013.

[76] Golyandina, N., and Zhigljavsky, A. *Singular Spectrum Analysis for Time Series*, 2nd ed., vol. 2. Springer Berlin Heidelberg, 2020.

[77] Gugercin, S., Antoulas, A., and Beattie, C. $\mathcal{H}_2$ model reduction for large-scale linear dynamical systems. *SIAM Journal on Matrix Analysis and Applications 30*, 2 (2008), 609–638.

[78] Haardt, M. Structured least squares to improve the performance of ESPRIT-type algorithms. *IEEE Transactions on Signal Processing 45*, 3 (1997), 792–799.

[79] Halmos, P. R. Shifts on Hilbert spaces. *Journal für die reine und angewandte Mathematik 208* (1960), 102–112.

[80] Hanzon, B., Maciejowski, J., and Chou, C. T. Optimal h2 order-one reduction by solving eigenproblems for polynomial equations, 2007.

[81] Hanzon, B., Maciejowski, J. M., and Chou, C. T. Optimal H2 order-one reduction by solving eigenproblems for polynomial equations, 2007.

[82] Hanzon, B., Peeters, R., and Bleylevens, I. On a finiteness result for the number of critical points of the H2 approximation criterion. In *16th IFAC Symposium on System Identification* (July 2012), pp. 728–733.

[83] Harris, J. *Algebraic Geometry: A first course.* Graduate Texts in Mathematics. Springer, New York, 1995.

[84] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer, NY, 2017.

[85] Hauer, J. F., Demeure, C. J., and Scharf, L. L. Initial results in Prony analysis of power system response signals. *IEEE Transactions on Power Systems 5*, 1 (1990), 80–89.

[86] Heij, C. *Deterministic Identification of Dynamical Systems*, vol. 127 of *Lecture Notes in Control and Information Sciences.* Springer, Berlin, Germany, 1989.

[87] Heinig, G., and Rost, K. *Algebraic Methods for Toeplitz-like Matrices and Operators.* Birkhäuser, Boston, 1984.

[88] Hildebrand, F. B. *Introduction to Numerical Analysis*, 2nd ed. Dover Publications, New York, 1974, pp. 457–462.

[89] Ho, B. L., and Kalman, R. E. Effective construction of linear, state-variable models from input/output functions. *Regelungstechnik 14* (1966), 545–548.

[90] Hochstenbach, M., Košir, T., and Plestenjak, B. Numerical methods for rectangular multiparameter eigenvalue problems, with applications to finding optimal ARMA and LTI models. *Numerical Linear Algebra with Applications 31*, 2 (2024).

[91] Holt, J. N., and Antill, R. J. Determining the number of terms in a Prony algorithm exponential fit. *Mathematical Biosciences 36* (1977), 319–332.

[92] Householder, A. S., and Young, G. Matrix approximation and latent roots. *American Mathematical Monthly 45*, 3 (1938), 165–171.

[93] Hua, Y., and Sarkar, T. K. On SVD for estimation generalized eigenvalues of singular matrix pencil in noise. *IEEE Transactions on Signal Processing 39*, 4 (1991), 892–900.

[94] Hyland, D., and Bernstein, D. The optimal projection equations for model reduction and the relationschips among the methods of Wilson, Skelton, and Moore. *IEEE Transactions on Automatic Control 30*, 12 (1985), 1201–1211.

[95] Iarobbino, A., and Kanev, V. *Power sums, Gorenstein Algebras and Determinantal Loci.* Lecture Notes in Mathematics 1721. Springer, 1999.

[96] Ishteva, M., Usevich, K., and Markovsky, I. Factorization approach to structured low-rank approximation with applications. *SIAM Journal on Matrix Analysis and Applications 35*, 3 (2014), 1180–1204.

[97] Janssens, V. The quest for global optimality within linear system identification. Master's thesis, Faculty of Engineering Science, KU Leuven, 2024.

[98] Kailath, T. *Linear Systems.* Information and System Science Series. Prentice-Hall, Englewood Cliffs, 1980.

[99] Kay, S. M. Accurate frequency estimation at low SNR. *IEEE Transactions on Acoustics, Speech and Signal Processing 32*, 3 (1984), 540–547.

[100] Kayser, E. L. *Computation and Complexity in Algebraic Geometry.* Phd thesis, Universität Leipzig, 2025.

[101] Kayser, L. The Waring problem for polynomials: Geometry and applications. Master's thesis, Institute of Algebraic Geomtery, Leibniz Universität Hannover, 2022.

[102] Kozhasov, K., Muniz, A., Qi, Y., and Sodomaco, L. On the minimal algebraic complexity of the rank-one approximation problem for general inner products, 2024.

[103] Krajewski, W. Optimality conditions in multivariable L_2 model reduction. *Journal of the Franklin Institute 330*, 3 (1993), 431–439.

[104] Kreutz-Delgado, K. The complex gradient operator and the $\mathbb{CR}$-calculus, 2009.

[105] Kronecker, L. Zur Theorie der Elimination einer Variablen aus zwei algebraischen Gleichungen. *Monatsberichte der Königlich Preussischen Akademie der Wissenschaften (Berlin)* (1881), 535–600. Reprinted as pp. 113–192 in Mathematische Werke, vol. 2, B.G. Teubner, Leipzig, 1897.

[106] Kuhn, H. W. The Hungarian method for the assignment problem. *Naval Research Logistics 52*, 1 (2005), 7–21.

[107] Kung, S. Y. A new identification and model reduction algorithm via singular value decomposition. In *Proceedings of the 12th Asilomar Conference on Circuits, Systems and Computers, Institute of Electrical and Electronics Engineers, New York* (1978), pp. 705–714.

[108] Kung, S. Y., Arun, K. S., and Rao, D. V. B. State-space and singular-value decomposition-based approximation methods for the harmonic retrieval problem. *Journal of the Optical Society of America 73*, 12 (1983), 1799–1811.

[109] Lagauw, S., Agudelo, O. M., and De Moor, B. Globally optimal SISO H_2-norm model reduction using Walsh's Theorem. *IEEE Control Systems Letters 7* (2023), 1670–1675.

[110] Lagauw, S., Vanpoucke, L., and De Moor, B. Exact solution to the least squares realization problem as a multiparameter eigenvalue problem. *Automatica 185* (2026), 112792.

[111] Lax, P. D. Translation invariant spaces. *Acta Mathematica 101* (1959), 163–178.

[112] Lee, C., Hasegawa, H., and Gao, S. Complex-valued neural networks: A comprehensive survey. *IEEE/CAA Journal of Automatica Sinica 9*, 8 (2022), 1406–1426.

[113] Legat, B., Yuan, C., and Parrilo, P. Low-rank univariate sum of squares has no spurious local minima. *SIAM Journal on Optimization 33*, 3 (2023), 2041–2061.

[114] Lemmerling, P., and De Moor, B. Misfit versus latency. *Automatica 37* (2001), 2057 – 2067.

[115] Lemmerling, P., De Moor, B., and Van Huffel, S. On the equivalence of constrained total least squares and structured total least squares. *IEEE Transactions on Signal Processing 44*, 11 (1996), 2908–2911.

[116] Lemmerling, P., Vanhamme, L., Van Huffel, S., and De Moor, B. IQML-like algorithms for solving structured total least squares problems: a unified view. *Signal Processing 81* (2001), 1935–1945.

[117] Lepschy, A., Mian, G. A., Pinato, G., and Viaro, U. Rational L_2 approximation: A non-gradient algorithm. In *Proceedings of the 30th IEEE Conference on Decision and Control* (Dec. 1991), pp. 2321–2323.

[118] Levin, M. Estimation of a system pulse transfer function in the presence of noise. *IEEE Transactions on Automatic Control 9*, 3 (1964), 229–235.

[119] Li, T., and Wang, X. The BKK root count in C^n. *Mathematics of Computation 65*, 216 (1996), 1477–1484.

[120] Ljung, L. *System Identification: Theory for the User*, 2nd ed. Information and System Sciences Series. Prentice-Hall, Upper Saddle River, NJ, USA, 1999.

[121] Macaulay, F. S. Some formulae in elimination. In *Proc. of the London Mathematical Society* (1902), vol. 1.1, pp. 3–27.

[122] Majda, G., and Wei, M. A simple procedure to eliminate known poles from a time series. *IEEE Transactions on Antennas and Propagation 37*, 10 (1989), 1343–1344.

[123] Markovsky, I. *Low Rank Approximation: Algorithms, Implementation, Applications*, 2nd ed. Communications in Control Engineering. Springer, 2019.

[124] Markovsky, I., and Van Huffel, S. Overview of total least-squares methods. *Signal Processing 87*, 10 (2007), 2283–2302.

[125] Markovsky, I., Willems, J. C., Van Huffel, S., and De Moor, B. *Exact and Approximate Modeling of Linear Systems: A Behavioral Approach.* Monographs on Mathematical Modeling and Computation. SIAM, 2006.

[126] Markovsky, I., Willems, J. C., Van Hufffel, S., De Moor, B., and Pintelon, R. Application of structured total least squares for system identification. *IEEE Transactions on Automatic Control 50*, 10 (2005), 1490–1500.

[127] McClellan, J. H., and Lee, D. Exact equivalence of the Steiglitz–McBride iteration and IQML. *IEEE Transactions on Signal Processing 39*, 2 (1991), 509–512.

[128] Meier, L., and Luenberger, D. Approximation of linear constant systems. *IEEE Transactions on Automatic Control 12*, 5 (Oct. 1967), 585–588.

[129] Molaei, A. M., Zakeri, B., Andargoli, S. M. H., Abbasi, M. A. B., Fusco, V., and Yurduseven, O. A comprehensive review of direction-of-arrival estimation and localization approaches in mixed-field sources scenario. *IEEE Access 12* (2024), 65883–65918.

[130] Mourrain, B., Telen, S., and Van Barel, M. Truncated normal forms for solving polynomial systems: Generalized and efficient algorithms. *Journal of symbolic computation 102*, 1 (2021), 63–85.

[131] Nocedal, J., and Wright, S. J. *Numerical Optimization.* Springer Series in Operations Research and Financial Engineering. Springer, New York, NY, 1999.

[132] Ottaviani, G., Spaenlehauer, P.-J., and Sturmfels, B. Exact solutions in structured low-rank approximation. *SIAM Journal on Matrix Analysis and Applications 35*, 4 (2014), 1521–1542.

[133] Panah, B. S. The computation of the Euclidean distance degree for the middle Catalacticant for the binary forms. *Communications in Algebra* (2025), 1–9.

[134] Paulraj, A. R., Roy, R., and Kailath, T. A subspace rotation approach to signal parameter estimation. *Proceedings of the IEEE 74*, 7 (1986), 1044–1046.

[135] Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science (6th series)) 2*, 11 (1901), 559–572.

[136] Peeters, R. L. M., Bleylevens, I. W. M., and Hanzon, B. On algebraic and linear algebraic aspects of co-order three h2 model order reduction. *IFAC Proceedings Volumes 45* (2012), 704–709.

[137] Pintelon, R., and Schoukens, J. *System Identification: A Frequency Domain Approach.*, 2nd ed. Wiley-IEEE Press, 2012.

[138] Plestenjak, B. MultiParEig. MATLAB Central File Exchange (`https://www.mathworks.com/matlabcentral/fileexchange/47844-multipareig`). Retrieved Jan. 24, 2023.

[139] Porteous, J. R. *Singularities—Symposium I*, vol. 192 of *Lecture Notes in Mathematics.* Springer, Berlin, University of Liverpool, 1971, ch. Simple singularities of maps, pp. 286–307.

[140] Rabinowitsch, J. L. Zum Hilbertschen Nullstellensatz. *Mathematische Annalen 102* (1930), 520–520.

[141] Regalia, P. A. *Adaptive IIR Filtering in Signal Processing and Control.* Marcel Dekker Inc., New York, NY, 1995.

[142] Rojas, J. F., and Mendoza, R. A note on the fiber dimension theorem. *Proyecciones 28*, 1 (2009), 57–73.

[143] Roorda, B. Algorithms for global total least square modelling of finite multivariable time series. *Automatica 31*, 3 (1995), 391–404.

[144] Roorda, B., and Heij, C. Global total least squares modeling of multivariable time series. *IEEE Transactions on Automatic Control 40*, 1 (Jan. 1995), 50–63.

[145] Roothaan, C. C. J. New developments in molecular orbital theory. *Reviews of Modern Physics 23* (1951), 69–89.

[146] Rosen, J., Park, H., and Glick, J. Total least norm formulation and solution for structured problems. *SIAM Journal on matrix analysis and applications 17*, 1 (1996), 110–126.

[147] Roy, R., and Kailath, T. ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing 37*, 7 (1989), 984–995.

[148] Roy, R., and Kailath, T. ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing 37*, 7 (1989), 984–995.

[149] Rump, S. Structured perturbations part i: normwise distances. *SIAM Journal on Matrix Analysis and Applications 25*, 1 (2003), 1–30.

[150] Scharf, L. L., Kumaresan, R., and Shaw, A. K. An algorithm for pole-zero modeling and spectral analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing 34*, 3 (1986), 637–640.

[151] Schreier, P. J., and Scharf, L. L. *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals.* Cambridge University Press, 2010.

[152] Shafarevich, I. R. *Basic Algebraic Geometry 1: Varieties in Projective Space*, 3 ed. Springer Berlin Heidelberg, 2013.

[153] Shen, H., and Huang, J. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis 99*, 6 (2008), 1015–1034.

[154] Söderström, T. *Errors-in-variables Methods in System Identification.* Communications and Control Engineering (CCE). Springer Cham, 2018.

[155] Sommese, A. J., and Wampler, C. W. *The Numerical Solution of Systems of Polynomials Arising in Engineering and Science.* World Scientific, 2005.

[156] Spanos, J., Milman, M., and Mingori, D. A new algorithm for L_2 optimal model reduction. *Automatica 28*, 5 (1992), 897–909.

[157] Srebro, N., and Shraibman, A. Weighted low-rank approximations. In *Proc. of the 20th International Conference on Machine Learning (ICML-03)* (2003), pp. 720–727.

[158] Steiglitz, K., and McBride, L. E. A technique for identification of linear systems. *IEEE Transactions on Automatic Control AC-10* (1965), 461–464.

[159] Stetter, H. *Numerical Polynomial Algebra.* SIAM, Philadelphia, PA, USA, 2004.

[160] Stoica, P., and Moses, R. L. *Spectral Analysis of Signals.* Pearson Prentice Hall, 2005.

[161] Stoica, P., and Sharman, K. C. Maximum likelihood methods for direction-of-arival estimation. *IEEE Transactions on Acoustics, Speech, and Signal Processing 38*, 7 (1990), 1132–1143.

[162] Stoica, P., and Söderström, T. The Steiglitz–McBride identification algorithm revisited–convergence analysis and accuracy aspects. *IEEE Transactions on Automatic Control 26*, 3 (1981), 712–717.

[163] Sylvester, J. Sur une extension dun théorème de Clebsch relatif aux courbes du quatrième degré. *Comptes Rendus, Math. Acad. Sci 102* (1886), 1532–1534.

[164] Sylvester, J. J. On a theory of the syzygetic relations of two rational integral functions, comprising an application to the theory of Sturms functions, and that of the greatest algebraical common measure. *Philosophical Transactions of the Royal Society of London 143* (1853), 407–548.

[165] Telen, S., Van Barel, M., and Verschelde, J. A robust numerical path tracking algorithm for polynomial homotopy continuation. *SIAM Journal on Scientific Computing 42*, 6 (2020), 3610–3637.

[166] Umesh, S., and Trufts, D. W. Estimation of parameters of exponentially damped sinusoids using fast maximum likelihood estimation with applications to NMR spectroscopy data. *IEEE Transactions on Signal Processing 44*, 9 (1996), 2245–2259.

[167] Usevich, K., and Markovsky, I. Variable projection for affinely structured low-rank approximation in weighted 2-norms. *Journal of Computational and Applied Mathematics 272* (2014), 430–448.

[168] Van Dooren, P., Gallivan, K., and Absil, P.-A. $\mathcal{H}_2$-optimal model reduction with higher-order poles. *SIAM Journal on Matrix Analysis and Applications 31*, 5 (2010), 2738–2753.

[169] Van Huffel, S., P., H., and Rosen, J. Formulation and solution of structured total least norm problems for parameter estimation. *IEEE Transactions on Signal Processing 44*, 10 (1996), 2464–2474.

[170] Van Huffel, S., and Vandewalle, J. *The Total Least Squares Problem: Computational Aspects and Analysis.* Fronties in Applied Mathematics; 9. SIAM Philadelphia, USA, 1991.

[171] Van Overschee, P., and De Moor, B. *Subspace Identification for Linear Systems: Theory – Implementation – Applications.* Kluwer Academic Publishers, 1996.

[172] Vanhuffel, S., Chen, H., Decanniere, C., and Vanhecke, P. Algorithm for time-domain nmr data fitting based on total least squares. *Journal of Magnetic Resonance 110*, 2 (Oct. 1994), 228–237.

[173] Vermeersch, C. *The (Block) Macaulay Matrix: Solving Systems of Multivariate Polynomial Equations and Multiparameter Eigenvalue Problems.* PhD thesis, Faculty of Engineering, KU Leuven, Leuven, Belgium, 2023.

[174] Vermeersch, C., and De Moor, B. Globally optimal least-squares ARMA model identification is an eigenvalue problem. *IEEE Control Systems Letters 3*, 4 (2019), 1062–1067.

[175] Vermeersch, C., and De Moor, B. Two complementary block Macaulay matrix algorithms to solve multiparameter eigenvalue problems. *Linear Algebra and its Applications 654* (2022), 177–209.

[176] Vermeersch, C., and De Moor, B. Two double recursive block Macaulay matrix algorithms to solve multiparameter eigenvalue problems. *IEEE Control Systems Letters 7* (2023), 319–324.

[177] Vermeersch, C., and De Moor, B. On the rectangular multiparameter eigenvalue problem. Tech. rep., STADIUS-ESAT, KU Leuven, Leuven, Belgium, 2025.

[178] Vermeersch, C., and De Moor, B. Solving multivariate polynomial systems and rectangular multiparameter eigenvalue problems with MacaulayLab. Tech. rep., KU Leuven, Leuven, Belgium, 2025.

[179] VERSCHELDE, J. Algorithm 795: PHCpack: A general-purpose solver for polynomial systems by homotopy continuation. *ACM Transactions on Mathematical Software 25*, 2 (1999), 251–276.

[180] VON NEUMANN, J., AND MORGENSTERN, O. *Theory of Games and Economic Behavior*, 3rd ed. Princeton University Press, Princeton (NJ), 1980.

[181] ŽIGIĆ, D., WATSON, L. T., COLLINS, E., AND BERNSTEIN, D. S. Homotopy methods for solving the optimal projection equations for the H_2 reduced order model problem. *International Journal of Control 56*, 1 (1992), 173–191.

[182] WAHLBERG, B. Orthogonal rational functions: A transformation analysis. *SIAM Review 45*, 4 (2003), 689 – 705.

[183] WALSH, J. L. *Interpolation and Approximation by Rational Functions in the Complex Domain*, 1 ed. American Mathematical Society, 1935.

[184] WALSH, J. L. *Interpolation and Approximation by Rational Functions in the Complex Domain*, 3rd ed. American Mathematical Society, 1960.

[185] WATANABE, J. Hankel matrices and Hankel ideals. In *Proc. of the School of Science, Tokai University* (1997), no. 32, pp. 11–21.

[186] WILLEMS, J. C. From time series to linear system–Part II. Exact modelling. *Automatica 22*, 6 (1986), 675–694.

[187] WILLEMS, J. C. From time series to linear system–Part III. Approximate modelling. *Automatica 23*, 1 (1987), 87–115.

[188] WILSON, D. Optimum solution of model-reduction problem. *Proceedings of the Institution of Electrical Engineers 117*, 6 (1970), 1161–1165.

[189] YOUNG, G. Matrix approximation and subspace fitting. *Psychometrika 2*, 1 (1937), 21–25.

[190] YOUNG, G. Maximum likelihood estimation and factor analysis. *Psychometrika 6*, 1 (1940), 49–53.

[191] ZEIGER, H., AND MCEWEN, A. Approximate linear realizations of given dimension via ho's algorithm. *IEEE Transactions on Automatic Control 19*, 2 (1974), 153–153.

[192] ZHANG, R., AND PLONKA, G. Optimal approximation with exponential sums by a maximum likelihood modification of Pronys method. *Advances in Computational Mathematics 45*, 3 (2019), 1657–1687.

[193] Zvonarev, N., and Golyandina, N. Low-rank signal subspace: parametrization, projection and signal estimation. *Statistics and Its Interface 16*, 1 (2023), 117–132.

Statement on the use of Generative AI

I did not use generative AI assistance tools (ChatGPT) during the research/writing process of my thesis, except for mere language assistance.
The text, code, and images in this thesis are my own (unless otherwise specified). Generative AI has only been used in accordance with the KU Leuven guidelines and appropriate references have been added. I have reviewed and edited the content as needed and I take full responsibility for the content of the thesis.

List of publications

Peer-Reviewed Articles in International Journals

- **Lagauw, S.**, Vanpoucke, L., and De Moor, B. (2026). "Exact Solution to the Least Squares Realization Problem as a Multiparameter Eigenvalue Problem". *Automatica* 185, pp. 112792.
- Viroux, S., **Lagauw, S.**, and De Moor, B. (2025). "Least Squares Realization of Complex-Valued Multiple-Output Autonomous LTI Systems via the Riemannian Singular Value Decomposition". *IEEE Control Systems Letters*, 9, pp. 2181–2186.
- **Lagauw, S.**, Agudelo, O. M., and De Moor, B. (2023). "Globally Optimal SISO H2-Norm Model Reduction Using Walsh's Theorem". *IEEE Control Systems Letters* 7, pp. 1670–1675.

Peer-Reviewed Publications in International Conference Proceedings

- **Lagauw, S.** and De Moor, B. (2024). "Least Squares Projection Onto the Behavior for SISO LTI Models". In: *IFAC-PapersOnLine, 58.15. Part of special issue: 20th IFAC Symposium on System Identification (SYSID)*. Northeastern University, Boston, United States, pp. 336–341.
- **Lagauw, S.**, Vanpoucke, L., and De Moor, B. (2024). "Exact Characterization of the Global Optima of Least Squares Realization of Autonomous LTI Models as a Multiparameter Eigenvalue Problem". In: *Proceedings of the 22nd European Control Conference (ECC)*. KTH Royal Institute of Technology, Stockholm, Sweden, pp. 3439– 3444.

- Vermeersch, C., **Lagauw, S.**, and De Moor, B. (2023). "Multivariate Polynomial Optimization in Complex Decision Variables Is a (Rectangular) Multiparameter Eigenvalue Problem". In: *Proceedings of the 62nd Conference on Decision and Control (CDC)*. Marina Bay Sands, Singapore, pp. 7299–7305.

Preprints

- Vermeersch, C., **Lagauw, S.**, and De Moor, B. (2025). "Incorporating Fixed Pole Information in the Data-Driven Least Squares Realization Problem". A preprint of the manuscript is available online at: `https://arxiv.org/abs/2509.09394`.

Manuscripts in Preparation

- **Lagauw, S.**, Kayser, L., and De Moor, B. (2026). "Porteous meets Walsh; Counting the number of critical points for H_2-norm model reduction using the algebraic geometry behind multiparameter eigenvalue problems".
- Kayser, L., **Lagauw, S.**, and De Moor, B. (2026). "Generic and specific Euclidean distance degrees of secant varieties to the rational normal curve".

Presentations and Invited Seminars

- **Lagauw, S.**, Kayser, L., and De Moor, B., "The critical points of structured low-rank approximation for Hankel matrices". Presentation at the *STADIUS seminar series*, KU Leuven, Leuven, Belgium, May 2025.
- **Lagauw, S.** and De Moor, B., "Polynomial equations from systems and control theory". Guest lecture at the *Numerical Algebraic Geometry Group*, Max Planck Institute for Mathematics in the Sciences (MPI MiS), Leipzig, Germany, Oct. 2024.
- **Lagauw, S.** and De Moor, B., "Least Squares Projection Onto the Behavior for SISO LTI Models". Presentation at the *20th IFAC Symposium on System Identification (SYSID)*, Northeastern University, Boston, United States, July 2024.
- **Lagauw, S.**, Vanpoucke, L., and De Moor, B., "Exact Characterization of the Global Optima of Least Squares Realization of Autonomous LTI

Models as a Multiparameter Eigenvalue Problem". Presentation at the *22nd European Control Conference (ECC), Invited session: "Perspectives on System Identification Methods for Model Reduction"*, KTH Royal Institute of Technology, Stockholm, Sweden, June 2024.

- **Lagauw, S.**, Vermeersch, C., and De Moor, B., "Solving Applications from Systems Theory Via Efficient Numerical Linear Algebra Root-Finding Algorithms". Presentation at the *SIAM Conference on Applied Linear Algebra (LA24)*, Sorbonne Université, Paris, France, May 2024.
- **Lagauw, S.**, Agudelo, O. M., and De Moor, B., "Globally Optimal H_2-Norm Model Reduction Using Walsh's Theorem". Presentation at the *62nd IEEE Conference on Decision and Control (CDC)*, Marina Bay Sands, Singapore, Dec. 2023.
- **Lagauw, S.** and De Moor, B., "Globally Optimal SISO H_2-Norm Model Reduction". Presentation at the *42nd Benelux Meeting on Systems and Control (BENELUX)*, Hotel Mennorode, Elspeet, The Netherlands, April 2023. (Shortlisted for Best Presentation Award)
- **Lagauw, S.** and De Moor, B., "Globally optimal misfit modelling for SISO systems". Presentation at the *41st Benelux Meeting on Systems and Control (BENELUX)*, Université Libre de Bruxelles (ULB), Brussels, Belgium, June 2022.

Poster Presentations

- **Lagauw, S.** and De Moor, B., "Least squares optimal realisation of SISO LTI systems is an eigenvalue problem". Poster presentation at the *30th European Research Network System Identification (ERNSI) Workshop*, KU Leuven, Leuven, Belgium, Sept. 2022.
- **Lagauw, S.** and De Moor, B., "Globally Optimal SISO H_2-Norm Model Reduction". Poster presentation at the *SeLMA closing event*, ESAT, KU Leuven, Leuven, Belgium, June 2022.

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
Kasteelpark Arenberg 10
3001 Leuven (België)
sibren.lagauw@esat.kuleuven.be

