RICE UNIVERSITY

Numerically Stable and Statistically Efficient Algorithms for Large Scale Exponential Fitting

by

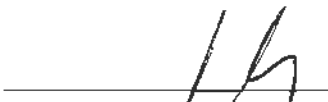
Jeffrey M. Hokanson

A Thesis for the Degree

Doctor of Philosophy

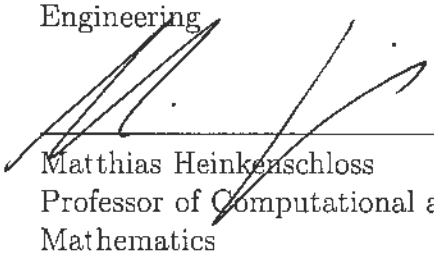
Thesis Committee:

Mark Embree, Co-chair
Professor of Computational and Applied Mathematics



Steven J. Cox, Co-chair
Professor of Computational and Applied Mathematics

Athanasios C. Antoulas
Professor of Electrical and Computer Engineering



Matthias Heinkenschloss
Professor of Computational and Applied Mathematics

Houston, Texas

August, 2013

ABSTRACT

Numerically Stable and Statistically Efficient Algorithms for Large Scale Exponential Fitting

by

Jeffrey M. Hokanson

The exponential fitting problem appears in diverse applications such as magnetic resonance spectroscopy, mechanical resonance, chemical reactions, system identification, and radioactive decay. In each application, the exponential fitting problem decomposes measurements into a sum of exponentials with complex coefficients plus noise. Although exponential fitting algorithms have existed since the invention of Prony's Method in 1795, the modern challenge is to build algorithms that stably recover statistically optimal estimates of these complex coefficients while using millions of measurements in the presence of noise. Existing variants of Prony's Method prove either too expensive, most scaling cubically in the number of measurements, or too unstable. Nonlinear least squares methods scale linearly in the number of measurements, but require well-chosen initial estimates lest these methods converge slowly or find a spurious local minimum.

We provide an analysis connecting the many variants of Prony's Method that have been developed in different fields over the past 200 years. This provides a unified framework that extends our understanding of the numerical and statistical properties of these algorithms.

We also provide two new algorithms for exponential fitting that overcome several

practical obstacles. The first algorithm is a modification of Prony's Method that can recover a few exponential coefficients from measurements containing thousands of exponentials, scaling linearly in the number of measurements. The second algorithm compresses measurements onto a subspace that minimizes the covariance of the resulting estimates and then recovers the exponential coefficients using an existing nonlinear least squares algorithm restricted to this subspace. Numerical experiments suggest that small compression spaces can be effective; typically we need fewer than 20 compressed measurements per exponential to recover the parameters with 90% efficiency. We demonstrate the efficacy of this approach by applying these algorithms to examples from magnetic resonance spectroscopy and mechanical vibration.

Finally, we use these new algorithms to help answer outstanding questions about damping in mechanical systems. We place a steel string inside vacuum chamber and record the free response at multiple pressures. Analyzing these measurements with our new algorithms, we recover eigenvalue estimates as a function of pressure that illuminate the mechanism behind damping.

Acknowledgments

First and foremost, I would like to thank Professor Mark Embree for imparting in me an appreciation for the aesthetics of mathematics and for giving me the freedom to explore the results that became this thesis. I could not have completed this project without his tireless help, countless revisions, and endless patience. Mark took me under his wing when I was a junior and gave me my first real taste of research. Together, we have eaten good cheese, admired fine typography and graphic design, and built lab equipment. In the process we have built a relationship that has shaped me into both a better mathematician and a wiser person.

I would also like to thank the members of my committee for their help in shaping this thesis. Along with Mark Embree, Professor Steve Cox provided the inspiration to try to solve inverse eigenvalue problems using real experimental data. Professor Matthias Heinkenschloss provided invaluable insight into the wide field of optimization methods. A series of conversations with Professor Thanos Antoulas showed me how my research fits into the context of system identification. I am immensely grateful for all their contributions.

I would also like to thank Ion-Victor Gosea for pointing out errors in an early draft of Section 3.5.3.

I would also like to thank the NSF for providing the funding for both the Physics of Strings Seminar and the Eigenvalue Clinic through which the ideas in this thesis were refined. NSF funding also supported the experimental work described in this thesis.

Many people provided assistance in building and operating these experiments. Professor Stan Dodds provided considerable guidance through the perils of experi-

mental science and help in improving the vacuum chamber. His sense of humor and his open door have been an important part of my life at Rice. Jeffrey Bridge designed the improved photodetectors and worked with me to build the apparatus for the vacuum experiment. Jeff was not only a great experimental partner but also a loyal and inspiring friend.

Nor could this experiment have been built without the patience and good humor of many of the shop staff. In particular, I would like to thank Joe Gesenhus for many hours of assistance in the machine shop and Michael Dye for his jokes and his help with electrical components of the experiment. I would also like to thank Dwight Dear for teaching me how to use a machine shop, Carlos Amaro for letting me play with his toys, and Dick Chronister for being awesome.

I would also like to thank the staff of the CAAM department, particularly Daria Lawrence, Ivy Gonzales, Brenda Aune, Eric Aune, and Fran Moshiri. They have provided not only advice and assistance but also warmth and compassion. Second to Mark Embree, Daria has play the greatest role in seeing me and this thesis through.

Finally I would like to thank my friends and family. I am forever grateful for my many wonderful friends who have gotten me into and out of trouble and reminded me that there was a life outside of my thesis. From rocket supplies to telescopes to my education at Rice, my parents have given flight to my dreams and enabled me to become who I am today. Their love and support has been a constant force in my life. Lastly, I would like to thank my girlfriend Kimberly DeBruler who, in addition to providing assistance monitoring the vacuum chamber experiment, has given me an appreciation for many new things in life. Her love and affection have made finishing this thesis possible.

Contents

Illustrations

Tables

Algorithms

Chapter 1

Introduction

Exponential signals abound: from the decay of radioactive isotopes, sine waves in radio, and decaying sinusoids in Magnetic Resonance Spectroscopy (MRS) experiments. In all cases, the signal takes the form

$$\widetilde{y}(t) = g(t) + \sum_{k=0}^{p-1} a_k e^{\omega_k t}, \tag{1.1}$$

where $a_k \in \mathbb{C}$ is the amplitude, $\omega_k \in \mathbb{C}$ the complex frequency, and $g \in \mathbb{C}$ some random noise. The complex frequency ω_k takes several forms: real and negative for radioactive decay, purely imaginary for radio waves, and in the left half of the complex plane for MRS experiments. The goal of exponential fitting is to recover the parameters $\mathbf{a} = [a_0, \cdots, a_{p-1}]^\top$ and $\boldsymbol{\omega} = [\omega_0, \cdots, \omega_{p-1}]^\top$ that best approximate a finite number of samples of $y(t)$. These parameters $\mathbf{a}$ and $\boldsymbol{\omega}$ reveal important information: the presence of isotopes, velocities of radar targets through the Doppler effect, resonant frequencies of mechanical structures, and the structure of organic compounds. Algorithms for exponential fitting descend in two distinct lineages: one beginning with Prony's Method in 1795 [127] and the other beginning with nonlinear least squares circa 1805 (Legendre). Prony's Method has profound numerical and statistical flaws: it yields biased estimates of $\boldsymbol{\omega}$, the covariance of $\boldsymbol{\omega}$ typically exceeds the Cramér-Rao lower bound by several orders of magnitude, and the estimates of $\boldsymbol{\omega}$ are extremely sensitive to rounding errors even for moderate numbers of exponentials ($p \approx 20$). Dozens of variants of Prony's Method have been developed in the

past two hundred years, but none simultaneously solves all these problems without requiring prohibitive amounts of computational work. In contrast, the nonlinear least squares formulation has excellent numerical and statistical properties, but requires potentially many expensive iterations. Exponential fitting is but one example among many nonlinear least squares problems solved by specialized optimization algorithms, such as the Levenberg–Marquardt algorithm. There are no specialized algorithms exclusively for exponential fitting, but *Variable Projection* [55] was developed to reduce the computational work by constructing the gradient and Jacobian for $\boldsymbol{\omega}$ while implicitly solving for $\mathbf{a}$. The modern challenge is to build algorithms that stably recover statistically optimal estimates of the exponential coefficients $\boldsymbol{\omega}$ and $\mathbf{a}$ while using the millions of measurements of y that are easily available with modern hardware. In this thesis we build new algorithms for the exponential fitting problem that require fewer operations and offer improved numerical and statistical properties, and descend from both Prony's Method and nonlinear least squares.

The first contribution of this thesis is a comprehensive linear algebra-based understanding of Prony's Method and its many derivatives in Chapter 2. The variants of Prony's Method are spread across many disciplines (see Table 1.1); we provide a cohesive review across these fields that goes beyond the limited reviews that already exist: [79] in the physics literature, [102] in the structural analysis literature, [9] for the radar detection problem, [66] considering primarily statistical questions in the applied math literature, and [150] for subspace methods in the electrical engineering literature. We also provide a new statistical analysis that yields estimates of the bias and covariance of Prony's Method; these estimates also apply to other variants of Prony's Method. This comprehensive review and analysis of Prony's Method reveals two new Prony-type methods. The first is a *Restarted Prony's Method* that determines

Table 1.1 : A sampling of papers from fields in which exponential fitting occurs. Within each field, papers are listed chronologically.

Field	Cite.	Description
Astronomy	[162]	alternating optimization of $\mathbf{a}$ and $\boldsymbol{\lambda}$, with heuristics for initial conditions
Biology	[122]	Peeling Method
	[33]	compressing measurements for Prony's Method
	[121]	Prony's Method using determinants, rebuked by [140]
	[139]	Prony's Method with extraneous exponentials
Chemistry	[53, 54]	Laplace transform methods
	[78]	Prony's Method with weighted sums of entries
Magnetic Resonance	[11]	copy of [92]
	[12]	copy of [94]
	[75]	total least squares variant of Kung's Method
	[153]	copy of Variable Projection [55]
Mathematics	[69]	minimizing residual with Prony's Method constraint
	[28]	enforcing $\lambda_j, a_j \in \mathbb{R}^+$ using quadratic programming
	[55]	Variable Projection
	[116]	Nullspace method in right norm
	[123]	Prony's Method, expanded to include translated exponentials
	[108]	genetic algorithm for exponential fitting
Radar	[9]	review
Physics	[74]	integral representation
Systems	[92]	Prony's Method with extraneous exponentials and truncated SVD
	[94]	Prony's Method with extraneous exponentials using the left singular subspace
Vibration	[102]	literature review
	[44]	literature review and analysis of algorithms
	[109]	modification of Ibrahim time domain method (a variant of Prony's Method)

a few exponentials among many exponentials, inspired by the Implicitly Restarted Arnoldi Method [142]. The second is the *Maximum Likelihood Prony Method* that is a fast, stable variant Householder's Method that yields maximum likelihood estimates of exponential parameters.

Another contribution of this thesis is the framework of *compression* to solve nonlinear least squares problems (Chapter 4). Rather than solving the full nonlinear least squares problem

$$\widetilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \|\widetilde{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta})\|_2^2, \tag{1.2}$$

this approach compresses (1.2) onto the subspace $\mathcal{W} \subset \mathbb{C}^n$ spanned by $\mathbf{W} \in \mathbb{C}^{n\times m}$, (with $\mathbf{W}^*\mathbf{W} = \mathbf{I}$) and instead solves

$$\widetilde{\boldsymbol{\theta}}_{\mathcal{W}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \|\mathbf{W}^*\widetilde{\mathbf{y}} - \mathbf{W}^*\mathbf{f}(\boldsymbol{\theta})\|_2^2. \tag{1.3}$$

Whereas existing techniques such as Incremental Gradient [50] choose $\mathbf{W}$ from random columns of the identity matrix $\mathbf{I}$, we choose $\mathbf{W}$ deterministically to minimize the covariance of $\widetilde{\boldsymbol{\theta}}$, and measure the increase of the covariance using a generalization of Fisher's *efficiency*. By choosing matrices $\mathbf{W}$ where $\mathbf{W}^*\mathbf{f}(\boldsymbol{\theta})$ has a closed form expression, we are able to reduce the computational burden while maintaining statistical accuracy. For the exponential fitting problem, we build $\mathbf{W}$ from columns of a block diagonal matrix composed of discrete Fourier transform blocks. Typically, twenty columns per exponential ($m = 20p$) are sufficient to recover $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}$ with 90% efficiency.

When solving the exponential fitting problem using optimization, such as in the compressed case discussed above, both the number of exponentials and initial estimates of $\boldsymbol{\omega}$ (and sometimes $\mathbf{a}$) are required. In Chapter 3, we review existing techniques for choosing the number of exponentials and providing initial estimates. One contribution we make is to update the *Peeling Method* [122] to use complex data and

initial parameter estimates provided by the Restarted Prony Method. We also review existing statistical criteria for selecting the number of exponentials using the Akaike Information Criterion and the singular values of a Hankel matrix. In Corollary 3.1 we improve the Hankel matrix singular value criteria by providing a new, accurate estimate of singular values corresponding to noise.

In this chapter, begin by formulating the exponential fitting problem, discussing existence and uniqueness. Then, we use the simple harmonic oscillator to illustrate how many systems described by differential equations have solutions that consist of sums of complex exponentials. By determining the exponential coefficients $\boldsymbol{\omega}$, we can infer parameters of the underlying differential equation. We then show that exponential fitting is a special case of system identification that arises from free response measurements. As history of exponential fitting is intertwined with many fields, including system identification, we provide a brief historical sketch in Section 1.4 including work from the present day.

1.1 Formulation of the Exponential Fitting Problem

Recalling (1.1), the exponential fitting problem seeks to recover estimates $\widetilde{\boldsymbol{\omega}}$ and $\widetilde{\mathbf{a}}$ of the true parameters $\widehat{\boldsymbol{\omega}}$ and $\widehat{\mathbf{a}}$ from the noisy measurements of (1.1) at times t_j:

$$\widetilde{y}(t_j) = g_j + \sum_{k=0}^{p-1} \widehat{a}_k e^{\widehat{\omega}_k t_j} \in \mathbb{C}, \qquad 0 \leq j < n, \tag{1.4}$$

where g_j is some random noise.[1] In this thesis, we make three further common assumptions.

[1] Tildes (e.g., $\widetilde{\omega}_j$) denote perturbed quantities; hats (e.g., $\widehat{\omega}_j$) true values.

First, we assume that $\widetilde{y}$ is sampled at a regular rate, i.e., $t_j = \delta j$, resulting in

$$\widetilde{y}(\delta j) =: \widetilde{y}_j = g_j + \sum_{k=0}^{p-1} \widehat{a}_k e^{\widehat{\omega}_k j} \in \mathbb{C} \quad 0 \leq j < n, \tag{1.5}$$

where δ has been absorbed into $\widehat{\boldsymbol{\omega}}$. Regularly sampled measurements are a by-product of the hardware used to acquire measurements in many fields (e.g., radar, magnetic resonance, mechanical resonance) and consequently, in most large scale problems ($n > 10^4$), y is sampled regularly. This assumption is also necessary to use Prony-type methods. When measurements are sampled irregularly, as in astronomy, we can either use nonlinear least squares methods, which do not require regular sampling, or construct regularly sampled measurements by interpolating the irregular measurements. Interpolation can result in systematic error, especially when $y(\delta(j+1)) - y(\delta j)$ is large.

The second assumption is that noise components $g_j \in \mathbb{C}$ are normally distributed random variables, The third assumption is that we seek maximum likelihood estimates of $\widehat{\boldsymbol{\omega}}$ and $\widehat{\mathbf{a}}$. Hence if $\mathbf{g}$ samples a standard complex normal distribution with zero mean and covariance $\boldsymbol{\Sigma}$, then the maximum likelihood estimates are given by

$$[\widetilde{\boldsymbol{\omega}}, \widetilde{\mathbf{a}}] = \underset{\boldsymbol{\omega}, \mathbf{a} \in \mathbb{C}^p}{\operatorname{argmin}} \|\widetilde{\mathbf{y}} - \mathbf{f}([\boldsymbol{\omega}, \mathbf{a}])\|_{\boldsymbol{\Sigma}}, \tag{1.6}$$

where $[\mathbf{f}([\boldsymbol{\omega}, \mathbf{a}])]_j = \sum_{k=0}^{p-1} a_k e^{\omega_k j}$ and $\|\mathbf{x}\|_{\boldsymbol{\Sigma}}^2 = \mathbf{x}^* \boldsymbol{\Sigma}^{-1} \mathbf{x}$. The assumption of normally distributed noise is both reasonable and practical: reasonable, because if there are many sources of noise, the central limit theorem guarantees the resulting noise will be approximately normally distributed; practical, because the resulting maximum likelihood estimates minimize a weighted ℓ_2 norm and the inner product associated with this norm gives additional structure that aids solving (1.6). If $\mathbf{g}$ samples another distribution, an alternative objective function is required. For example, the

noise in radioactive decay samples a Poisson distribution. However, for large radioactive samples, the Poisson distribution approximates a normal distribution [17, §2.4]. Maximum likelihood estimates are not the only approach for computing estimates $\widetilde{\boldsymbol{\omega}}$ and $\widetilde{\mathbf{a}}$; some authors take a Bayesian approach (see, e.g., [3]) that assumes some prior knowledge about the true parameter values $\widehat{\boldsymbol{\omega}}$ and $\widehat{\mathbf{a}}$ [24].

We can measure the efficiency of any exponential fitting algorithm by comparing the covariance of $\widetilde{\boldsymbol{\theta}} = [\widetilde{\boldsymbol{\omega}}^\top, \widetilde{\mathbf{a}}^\top]^\top$ to the Cramér-Rao lower bound. The Cramér-Rao bound provides a lower bound on the covariance of any technique that estimates parameters $\widetilde{\boldsymbol{\theta}}$ from noisy measurements $\widetilde{\mathbf{y}} = \mathbf{g} + \mathbf{f}(\widehat{\boldsymbol{\theta}})$. If $\widetilde{\boldsymbol{\theta}}$ is an unbiased estimate of $\widehat{\boldsymbol{\theta}}$ and $\mathbf{g}$ samples a complex normal distribution with zero mean and covariance $\boldsymbol{\Sigma}$, then the Cramér-Rao bound gives

$$\mathsf{Cov}[\widetilde{\boldsymbol{\theta}}] := \mathsf{E}[(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}})(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}})^*] \succeq (\mathbf{F}(\widehat{\boldsymbol{\theta}})^* \boldsymbol{\Sigma}^{-1} \mathbf{F}(\widehat{\boldsymbol{\theta}}))^{-1} =: \mathcal{I}(\widehat{\boldsymbol{\theta}})^{-1} \tag{1.7}$$

where $[\mathbf{F}(\boldsymbol{\theta})]_{\cdot,j} = \partial / \partial \theta_j \mathbf{f}(\boldsymbol{\theta})$ [137, eq. (6.51)]. Sections 3.2 and 3.3 give expressions for the *Fisher information matrix* $\mathcal{I}$ particular to the exponential fitting problem. A similar bound for biased estimates of $\widehat{\boldsymbol{\theta}}$ also exists [137, eq. (6.50)], and sometimes biased estimates can have a smaller covariance than unbiased estimates – although this does not appear to be the case for the exponential fitting problem. To measure the effectiveness of an exponential fitting algorithm, we generalize Fisher's *efficiency* [49, §4] to multivariate problems, defining the efficiency of an estimator that yields $\widetilde{\boldsymbol{\theta}}$ as

$$\eta_{\widehat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}} := \frac{\mathsf{Tr}\, \mathcal{I}(\widehat{\boldsymbol{\theta}})^{-1}}{\mathsf{Tr}\, \mathsf{Cov}[\widetilde{\boldsymbol{\theta}}]}. \tag{1.8}$$

This efficiency is typically reported as a percentage. Typically we consider the limit where $\boldsymbol{\Sigma} \to \mathbf{0}$ uniformly. Section 4.1 motivates this definition, as $\mathsf{Tr}\, \mathsf{Cov}[\widetilde{\boldsymbol{\theta}}]$ corresponds to the expected ℓ_2 error of $\widetilde{\boldsymbol{\theta}}$.

Finally, for notation, Prony-type methods recover $\lambda_j := e^{\omega_j}$ instead of ω_j directly. Correspondingly, we can rewrite (1.6) as

$$\widetilde{y}_j = g_j + \sum_{k=0}^{p-1} \widehat{a}_k \widehat{\lambda}_k^j \quad 0 \le j < n. \tag{1.9}$$

We do not use this parameterization when using nonlinear least squares methods because, when n is large, errors accumulate when computing λ_k^j, either by $\lambda_k^j = e^{j \log \lambda_k}$ or by repeated multiplication [64, §2.10].

1.1.1 Special Cases for Real Measurements

Although we primarily deal with the generality of complex measurements for simplicity, most applications have real measurements $\widetilde{\mathbf{y}}$ (see Table 1.2). In this case, the maximum likelihood estimator for $\boldsymbol{\omega}$ and $\mathbf{a}$ remains the same, (1.6), except $\mathbf{g}$ samples a real distribution with covariance $\boldsymbol{\Sigma}$ (see Appendix A). Real measurements enforce one of two special cases on the parameters $\boldsymbol{\omega}$ and $\mathbf{a}$.

The least restrictive case allows $\boldsymbol{\omega}$ and $\mathbf{a}$ to be complex, but if ω_j is non-real, then for some k, $\omega_k = \overline{\omega_j}$ and $a_k = \overline{a_j}$. This case corresponds to damped sines and cosines. These sines and cosines will be undamped if $\boldsymbol{\omega}$ is purely imaginary.

The more restrictive case, occurring in diffusion and radioactive, decay requires $\boldsymbol{\omega} \in \mathbb{R}^-$ and $\mathbf{a} \in \mathbb{R}^+$. Although this can be easily imposed using inequality constraints on the nonlinear least squares problem, there exist specialized algorithms: e.g., [28].

1.1.2 Existence and Uniqueness

Regardless of whether the measurements $\mathbf{y}$ are real or complex, there always exists a solution of (1.6) that such that $\mathbf{f}([\boldsymbol{\omega}, \mathbf{a}]) = \mathbf{y}$. Pisarenko [125] establishes the existence of a solution by invoking a theorem of Carathéodory [60].

Table 1.2 : Applications of exponential fitting. Problems with complex data ($\mathbf{y} \in \mathbb{C}^n$) are marked by $\circ$. Problems with real data ($\mathbf{y} \in \mathbb{R}^n$) and minimal restraints on $\boldsymbol{\omega}$ and $\mathbf{a}$ are marked by $\times$. Problems with real data that enforce $\boldsymbol{\omega} \in \mathbb{R}^-$ and $\mathbf{a} \in \mathbb{R}^+$ are marked by $+$.

Field	Cite.	Type	Applications
Biology	[41]	$\times+$	multi-compartment models
	[76]	$+$	pulse fluorometry
	[122]	$+$	gas absorption
Chemistry	[53]	$+$	multiple species chemical reaction rates
	[152]	$\circ$	magnetic resonance spectroscopy
Electronics	[80]	$+$	power systems
Physics	[88]	$+$	radioactive decay
	[15]	$\times$	lack hole mergers
	[74]	$+$	underwater explosions
	[70]	$\times$	musical instruments
	[100]	$\times$	crack detection
	[132]	$\times$	normal modes of the earth
Vibration	[46]	$\times$	modal analysis

Theorem 1.1 (C. Carathéodory)**.** *Let* $\mathbf{y} \in \mathbb{C}^n$, $\mathbf{y} \neq 0$, $n > 0$. *There exists an integer* $1 \leq p \leq n$ *and constants* $\mathbf{a} \in \mathbb{R}_+^p$, $\boldsymbol{\omega} \in [0, 2\pi]^p$, *with* $\omega_j \neq \omega_k$ *if* $j \neq k$ *and*

$$y_k = \sum_{j=0}^{p-1} a_j e^{ik\omega_j} \qquad k = 0, 1, \ldots, n-1. \tag{1.10}$$

The integer p *and the constants* $\mathbf{a}, \boldsymbol{\omega}$ *are uniquely determined.*

A crude way to construct this solution is to use the discrete Fourier transform. If we choose $\boldsymbol{\omega}$ to be the nth roots of unity ($\omega_j = 2\pi j/n$) then $\mathbf{a}$ is given by the discrete Fourier transform of $\mathbf{y}$, i.e.,

$$\mathbf{a} = \mathbf{F}_n^* \mathbf{y} \tag{1.11}$$

where $[\mathbf{F}_n]_{j,k} = n^{-1/2} e^{2\pi ijk/n}$.

Solutions to the exponential fitting problem are not unique, as we are free to permute the entries of $\boldsymbol{\omega}$ and $\mathbf{a}$ simultaneously. To compare two solutions $\boldsymbol{\omega}$ and $\widetilde{\boldsymbol{\omega}}$,

we always use the *marriage norm* that permutes $\widetilde{\boldsymbol{\omega}}$ to minimize the mismatch, i.e.,

$$\min_{\pi} \|\widetilde{\boldsymbol{\lambda}}_{\pi} - \boldsymbol{\lambda}\|_2, \tag{1.12}$$

where π is a permutation. This permutation can be constructed in at most p^2 operations using the *Marriage Algorithm* developed by Gale and Shapley [51].

In practice, we are not interested in solving (1.6) as stated, but instead solve (1.6) while assuming a statistically justified number of exponentials, as discussed in Section 3.5. Determining the number of exponentials is related to the minimum realization problem in system identification, which seeks the smallest model order p that accurately describes the system; see, e.g., [58].

1.1.3 Nonparametric Methods

Parametric signal processing problems, such as exponential fitting, assume the signal $\widetilde{\mathbf{y}}$ is determined by a known model with some unknown parameters (i.e., $\boldsymbol{\omega}$). Nonparametric methods are an alternative for exploratory data analysis that attempt to extract instantaneous frequency information as a function of time. The most basic of these methods is the spectrogram: taking the discrete Fourier transform of small segments of $\widetilde{\mathbf{y}}$. The time and frequency resolution of the spectrogram is limited by the size of the Fourier transform, with smaller segments yielding finer time resolution but coarser frequency information. Wavelet analysis fixes the resolution problem in the spectrogram, providing a uniform resolution over frequency and time; see, e.g., [48]. The results of wavelet transforms are sometimes hard to interpret, prompting the use of the Hilbert-Huang transform [73]. The Hilbert-Huang transform breaks $\widetilde{\mathbf{y}}$ into a combination of so called 'intrinsic mode functions' and reveals the time-frequency dependence using the Hilbert transform of each of these mode functions, providing a more concentrated representation of energy in frequency and time [73].

Although these nonparametric methods can be used to estimate the exponential parameters $\boldsymbol{\omega}$, direct parametric methods will provide better estimates provided $\widetilde{\mathbf{y}}$ fits the model, in our case, exponentials plus noise (1.1).

1.2 A Prototypical Exponential Fitting Application

The simple harmonic oscillator is a prototypical example that illustrates how exponential fitting originates in many fields, and also shows the infrequent case of degenerate exponentials and the complications that can arise in parameter recovery. Two common forms of the simple harmonic oscillator are the mass-spring-damper system or the resistor-capacitor-inductor circuit, both illustrated in Figure 1.1. We consider the general case where u solves the second order differential equation

$$u''(t) + 2\gamma u'(t) + c^2 u(t) = 0, \qquad u(0) = u_0,\ u'(0) = v_0. \tag{1.13}$$

We can describe u as the sum of two exponentials. Linearizing (1.13) yields the coupled, first order differential equation

$$\begin{bmatrix} u'(t) \\ u''(t) \end{bmatrix} = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}' = \begin{bmatrix} 0 & 1 \\ -c^2 & -2\gamma \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \Leftrightarrow \mathbf{x}' = \mathbf{A}\mathbf{x}, \quad \mathbf{x}(0) = \mathbf{x}_0 = \begin{bmatrix} u_0 \\ v_0 \end{bmatrix}. \tag{1.14}$$

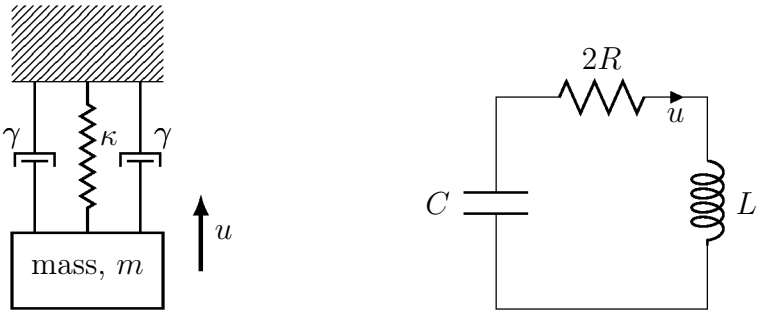
When $\mathbf{A}$ is diagonalizable, we can decompose $\mathbf{A}$ as $\mathbf{AV} = \mathbf{V}\boldsymbol{\Omega}$, where $\mathbf{V}$ is the matrix of eigenvectors and $\boldsymbol{\Omega} = \operatorname{diag}(\omega_+, \omega_-)$ is the matrix of eigenvalues:

$$\omega_\pm = -\gamma \pm \sqrt{\gamma^2 - c^2}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{v}_+ & \mathbf{v}_- \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \omega_+ & \omega_- \end{bmatrix}. \tag{1.15}$$

Then the matrix exponential gives the solution to (1.14):

$$\begin{aligned} \mathbf{x}(t) &= e^{\mathbf{A}t}\mathbf{x}_0 = \mathbf{V}e^{t\boldsymbol{\Omega}}\mathbf{V}^{-1}\mathbf{x}_0 \\ &= \mathbf{v}_+ e^{t\omega_+}\mathbf{w}_+^*\mathbf{x}_0 + \mathbf{v}_- e^{t\omega_-}\mathbf{w}_-^*\mathbf{x}_0 \end{aligned} \qquad \text{where } \mathbf{V}^{-1} = \mathbf{W} = \begin{bmatrix} \mathbf{w}_+^* \\ \mathbf{w}_-^* \end{bmatrix}.$$

Mass-Spring-Damper System Resistor-Inductor-Capacitor Circuit



$mu'' + 2\gamma u' + \kappa u = 0$ $Lu'' + 2Ru' + 1/Cu = 0$

Figure 1.1 : Two equivalent harmonic oscillators.

Since $u(t) = [\mathbf{x}(t)]_0 = \mathbf{e}_0^\top \mathbf{x}(t)$, where $\mathbf{e}_k$ is the k column of the identity matrix, then

$$\begin{aligned} u(t) = \mathbf{e}_0^\top \mathbf{x}(t) &= \left[\mathbf{e}_0^\top \mathbf{v}_+ \mathbf{w}_+^* \mathbf{x}(0)\right] e^{t\omega_+} + \left[\mathbf{e}_0^\top \mathbf{v}_- \mathbf{w}_-^* \mathbf{x}(0)\right] e^{t\omega_-} \\ &= a_+ e^{t\omega_+} + a_- e^{t\omega_-}. \end{aligned} \tag{1.16}$$

Linearization of other differential equations will always result in a sum of exponentials, provided $\mathbf{A}$ is diagonalizable. Then we can recover $\boldsymbol{\omega}$ and $\mathbf{a}$ from noisy measurements $\widetilde{y}_j = u(\delta j) + g_j$ by solving the exponential fitting problem (1.6).

Not every differential equation will result in the sum of exponentials. In the simple harmonic oscillator example when $\gamma = c$, then $\omega_+ = \omega_- = -\gamma$ and $\mathbf{A}$ has a Jordan block. Then the solution for $u(t)$ has polynomial terms as well as exponentials. If $\mathbf{AV} = \mathbf{VJ}$ is the Jordan normal form, where

$$\mathbf{V} = \begin{bmatrix} 1 & 0 \\ -\gamma & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{J} = \begin{bmatrix} -\gamma & 1 \\ 0 & -\gamma \end{bmatrix}, \tag{1.17}$$

then writing $\mathbf{J} = \boldsymbol{\Omega} + \mathbf{N}$, where $\mathbf{N}$ is the nilpotent part of $\mathbf{J}$,

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}_0 = \mathbf{V}e^{\boldsymbol{\Omega}t + \mathbf{N}t}\mathbf{V}^{-1}\mathbf{x}_0 = e^{-\gamma t}\mathbf{V}(\mathbf{I} + t\mathbf{N})\mathbf{V}^{-1}\mathbf{x}_0 = e^{-\gamma t}\mathbf{x}_0 + te^{-\gamma t}(\mathbf{V}\mathbf{N}\mathbf{V}^{-1}\mathbf{x}_0).$$

Consequently, u exhibits exponential decay with a polynomial component,

$$u(t) = (\mathbf{e}_0^\top \mathbf{x}_0)e^{-\gamma t} + (\mathbf{e}_0^\top \mathbf{V}\mathbf{N}\mathbf{V}^{-1}\mathbf{x}_0)te^{-\gamma t} = a_0 e^{-\gamma t} + a_1 te^{-\gamma t}. \tag{1.18}$$

In practice, we are unable to distinguish between non-diagonalizable systems resulting in non-exponential behavior from those that are diagonalizable and do have exponential behavior. The matrix exponential is a continuous function, and as diagonalizable matrices are dense in the set of all matrices, an arbitrary perturbation of measurements $u(t)$ from a non-diagonalizable system will correspond to a diagonalizable system (possibly with exotic transient behavior). However, if this structure is known a priori, then specialized techniques can be used such as modifying $\mathbf{f}$ in (1.6) or exploiting generalizations of the Vandermonde-Hankel decomposition [151, Thm. 4.12].

We can also use exponential fitting to infer γ and c from the original differential equation. Rather than developing a specialized algorithm to find these parameters, we first recover the eigenvalues of $\mathbf{A}$ (the exponential parameters $\boldsymbol{\omega}$) and from these, estimate the desired parameters. For example, to recover γ from noisy measurements $\widetilde{\mathbf{y}}$, we first estimate $\widetilde{\boldsymbol{\omega}}$ and then solve

$$\widetilde{\gamma} = \underset{\gamma}{\operatorname{argmax}} \left\| \begin{bmatrix} \widetilde{\omega}_+ \\ \widetilde{\omega}_- \end{bmatrix} - \begin{bmatrix} -\gamma + \sqrt{\gamma^2 - c^2} \\ -\gamma - \sqrt{\gamma^2 - c^2} \end{bmatrix} \right\|_{\boldsymbol{\Sigma}}, \tag{1.19}$$

where $\boldsymbol{\Sigma}$ is the covariance of $\widetilde{\boldsymbol{\omega}}$. Then the *Principle of Invariance* (§ A.6) guarantees that $\widetilde{\gamma}$ is a maximum likelihood estimate of γ. We must be careful when measurements are real: if $\widetilde{\omega}_+$ is complex, then $\overline{\widetilde{\omega}_-} = \widetilde{\omega}_+$, and these two are correlated and we must consider the covariance of the real and imaginary parts separately. Figure 1.2 provides an example of this recovery.

Recovering parameters from the underlying differential equation for other problems follows a similar procedure: compute maximum likelihood estimates $\widetilde{\boldsymbol{\omega}}$ (eigen-

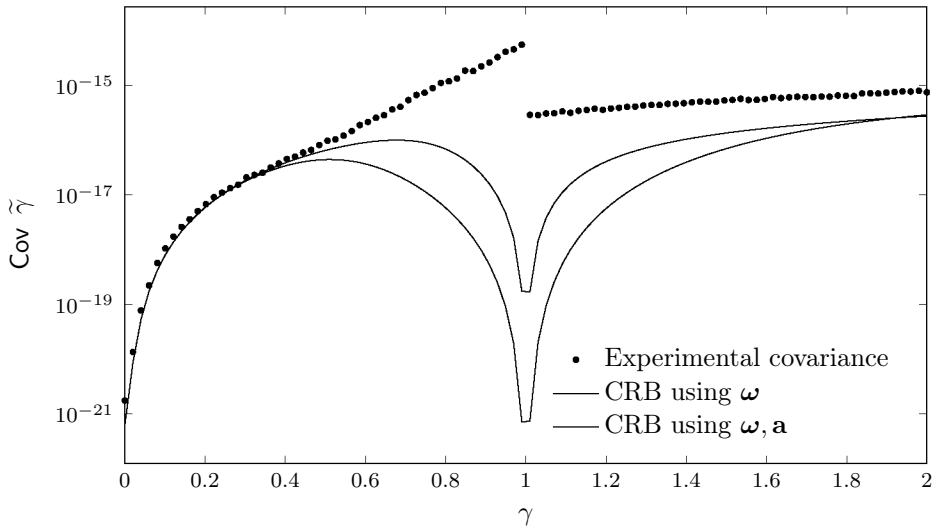


Figure 1.2 : Covariance of $\widetilde{\gamma}$ recovered from $\widetilde{\boldsymbol{\omega}}$ via Monte-Carlo experiments using (1.19) as compared to the Cramér-Rao lower bound (denoted CRB). In this example, we used initial conditions $u_0 = 1$ and $v_0 = 0$ with $c = 1$ and record $n = 100$ measurements sampled at $t_j = j$ polluted by $\mathbf{g}$ with covariance $\boldsymbol{\Sigma} = 10^{-16}\mathbf{I}$. Ignoring the additional information encoded by the initial conditions and revealed by $\mathbf{a}$, causes the Cramér-Rao bound to increase. The numerical experiments do not match the Cramér-Rao bound, as the computation of the bound in this example ignores the fact $\widetilde{\boldsymbol{\omega}}$ comes in conjugate pairs.

values of $\mathbf{A}$) from measurements $\widetilde{\mathbf{y}}$ and, from knowledge of the structure of $\mathbf{A}$, compute the desired underlying parameters. This second step is called an *inverse eigenvalue problem*, for which there are many specialized algorithms, see, e.g., Chu and Golub [30]. In Chapter 5, we are concerned with recovering a damping model for a vibrating string. Assuming the damping originates from a variable viscous damping symmetric about the center of the string, Cox and Embree provide an algorithm to recover this unknown damping function [37]. However, most authors are concerned only with the inverse eigenvalue problem and do not attempt to estimate eigenvalues from actual measurements; a notable exception is Cox, Embree, and Hokanson who reconstructed the mass distribution of a beaded string from experimental data [38]. However, none of these inverse eigenvalue algorithms use the covariance of $\widetilde{\boldsymbol{\omega}}$ to inform the recovery of the underlying system as (1.19) does. Hence, although an inverse eigenvalue algorithm may recover the underlying parameters exactly in the absence of noise, in the presence of noise, they will not yield maximum likelihood estimates of the sought-after parameters, and hence the covariance of their estimates will exceed the Cramér-Rao bound.

1.3 Exponential Fitting and System Identification

Exponential fitting is a special case of the deterministic realization or system identification problem in systems theory (see, e.g., [83, §2.2] or [154]). This problem seeks to recover matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ from a known input $\mathbf{u}$ and measured output $\mathbf{y}$ related by either the continuous time model

$$\begin{aligned} \mathbf{x}'(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \qquad \mathbf{x}(0) = \mathbf{x}_0 \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \end{aligned} \tag{1.20}$$

or the discrete time model

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k \qquad \mathbf{x}_0 \text{ given}, \quad k \geq 0. \\ \mathbf{y}_k &= \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k \end{aligned} \tag{1.21}$$

The realization problem is equivalent to exponential fitting if $\mathbf{y} \in \mathbb{C}$ and if there is no input, as in the case of 'free response' measurements where $\mathbf{u} = \mathbf{0}$ and $\mathbf{x}_0 \neq \mathbf{0}$. Then assuming $\mathbf{A}$ is diagonalizable, the output $\mathbf{y}$ is a sum of exponentials. In the continuous time case, using the eigendecomposition of $\mathbf{A}$, $\mathbf{AV} = \mathbf{V\Omega}$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{V}e^{\mathbf{\Omega}t}\mathbf{V}^{-1}\mathbf{x}_0 = \sum_{k=0}^{p-1} \left([\mathbf{CV}]_{\cdot,k} [\mathbf{V}^{-1}\mathbf{x}_0]_k \right) e^{\omega_j t}, \tag{1.22}$$

or in the discrete case, with eigendecomposition $\mathbf{AV} = \mathbf{V\Lambda}$ (to correspond with (1.9))

$$\mathbf{y}_k = \mathbf{C}\mathbf{V}\mathbf{\Lambda}^k\mathbf{V}^{-1}\mathbf{x}_0 = \sum_{k=0}^{p-1} \left([\mathbf{CV}]_{\cdot,k} [\mathbf{V}^{-1}\mathbf{x}_0]_k \right) \lambda_j^k. \tag{1.23}$$

The simple harmonic oscillator from the previous section fits into the continuous time system theory framework using the same $\mathbf{A}$ and setting $\mathbf{C} = [1\ 0]^\top$; both $\mathbf{B}$ and $\mathbf{D}$ are empty matrices as there is no input. In addition to free response measurements, 'impulse response' measurements can also be used. Impulse response measurements assume the initial conditions are zero ($\mathbf{x}_0 = \mathbf{0}$) and the input vanishes at all times past $t > 0$; i.e., input $\mathbf{u}(t)$ is a Dirac delta in the continuous time case, $\mathbf{u}(t) = \delta_{0,t}\mathbf{u}_0$, or a Kronecker delta input in the discrete time case, $\mathbf{u}_k = \delta_{0,k}\mathbf{u}_0$.

Due to the relation between exponential fitting and system identification, many exponential fitting algorithms are parallelled by results in the systems literature. For example, the Ho-Kalman Method for system identification [85, §6.2] is related to Kung's Method [94], as both estimate $\mathbf{A}$ using the left singular vectors of a Hankel matrix of observations, but Kung's Method goes one step further, estimating the exponential coefficients $\boldsymbol{\lambda}$ as eigenvalues of $\mathbf{A}$.

1.4 A Brief History of Algorithms for Exponential Fitting

The history of the exponential fitting problem begins in 1795 with the publication of the first exponential fitting algorithm (for $\boldsymbol{\lambda} \in \mathbb{R}^p$) by Gaspard Riche de Prony in *Journal de l'Ecole Polytechnique* [127] and continues to the modern day. Exponential fitting algorithms descend from the 18th century in two distinct lineages: one beginning with Prony's Method and the other that solves the nonlinear least squares problem (1.6) using an optimization algorithm. The first use of least squares to first fit a function to measurements is the subject of a priority dispute between Legendre, who published first 1805, and Gauss, who claimed to have developed the method in 1795 but did not publish his result until 1809; see [47] for a history. However, Gauss did develop the justification of least squares: minimizing the sum of squares yields a maximum likelihood estimate when noise is normally distributed. Over the next two hundred years, other authors would modify each of these methods. Prony's Method is both numerically unstable and yields very inefficient estimates (in the sense of (1.8)); subsequent modifications focus on improving these properties. In contrast, nonlinear least squares methods are both numerically stable and statistically efficient, but require many, potentially expensive, iterations to converge; consequently, most improvements attempt to reduce the computational cost.

The earliest modification of Prony's Method was developed by Yule (1927) [166] and Walker (1931) [156] to determine the period of sunspot numbers; see [104, §1.2]. Their method is similar to Prony's Method, except it computes an autoregressive model using autocorrelation measurements rather than the direct measurements $\widetilde{\mathbf{y}}$ (see Section 2.6). This spawned a series of algorithms that parallel developments in exponential fitting, such as Pisarenko's Method [125] (paralleling Prony's Method), MUSIC [10] (paralleling the Kumaresan and Tufts Method [92]), and ESPRIT [134]

(paralleling Kung's Method [94]). These papers by Yule and Walker are also the origin of the Yule-Walker equations used to construct autoregressive models given autocorrelation measurements.

Until digital computers became common for research during the 1950s, Prony's Method appeared in many numerical methods textbooks of the period with similar prominence to polynomial interpolation; for example: Whittaker and Robinson (1924) [159, §180], Lanczos (1956) [95, §IV.23], and Hildebrand (1956) [65, §9.4]. However, by 1959, Prony's Method and its variants had fallen out favor, replaced by nonlinear least squares methods using gradient descent [54, §II].

In 1949, Householder not only identified that Prony's Method yields inefficient estimates of $\widehat{\boldsymbol{\omega}}$ (a fact that still eludes some authors), but also provided an iteration that yields the efficient, maximum likelihood estimate of $\widehat{\boldsymbol{\omega}}$. Unfortunately these results went entirely unnoticed and similar maximum likelihood Prony Methods were reinvented twice: once in 1975 by Osborne [116] and again in 1986 by Bresler and Macovski [23] (see Section 2.4).

In 1957, there was the first publication of a nonlinear least squares approach for exponential fitting [88]. However, early methods like this one only used gradient descent to solve (1.6). Later developments during the 1960s and 1970s provided more sophisticated algorithms to solve the nonlinear least squares problem such as the Levenberg-Marquardt Method; see, e.g., [40]. Variable Projection began as a specialized optimization algorithm for exponential fitting that implicitly solved for the linear parameters $\mathbf{a}$ [138] (1970). Then in 1973 This approach was generalized by Golub and Pereyra for arbitrary separable nonlinear least squares problems [55].

During the 1960s and 1970s, exponential fitting frequently appeared in the application literature; a 1978 bibliography counts 116 papers concerning applications of

exponential fitting [84]. Most algorithms that appear are either variations of Prony's Method or original graphical techniques. For example, Perl reported the *Peeling Method* that fits multiple real exponentials by inspection in 1960 [122]; additional information is provided in [79, VI.B.1]. However, a similar, more robust approach by Cornell in 1956 [34, §II.3], predates Perl's example. Prony's Method reappears, both directly, reinvented by Parsons [121], and indirectly, where the role of $\widetilde{\mathbf{y}}$ as in Prony's Method has been replaced by linear combinations of $\widetilde{\mathbf{y}}$ in Cornell's Method [33] and the Method of Moments (1973) [78] (see Section 2.8). During this period, some authors attempted to solve the exponential fitting problem by finding peaks of the inverse Laplace transform of $y(t)$ [54, 53]. However, these techniques proved ultimately unsuccessful as computing the inverse Laplace transform is *exponentially ill-conditioned*; see [45, Fig. 6.4]

Starting in 1980 and continuing until the mid 1995, there was a burst of interest in the exponential fitting problem in the electrical engineering literature. Two forces conspired to push exponential fitting to the fore: increasing digital signal processing capabilities and funding motivated by military the applications of radar target identification [9] and the direction of arrival problem [136]. As speed is critical in digital signal processing, these methods were variants of Prony's Method, but improved the efficiency. The main breakthrough was to include spurious exponentials and using a low rank SVD solution by Kumaresan and Tufts in 1982 [92]. A system theoretic approach by Kung, Arun, and Rao in 1983 [94] used the same ingredients, but removed the problem of separating spurious from non-spurious exponentials. The theoretical basis for these algorithms was later developed by Rao in 1988 [129]. It was during this time that the autocorrelation based algorithms MUSIC and ESPRIT were developed for the direction of arrival problem. The final major algorithm in this vein is

Hankel Total Least Squares (HTLS), a total least squares variant of Kung's Method developed in 1994 [75].

Since then, exponential fitting continues to appear in many application areas: magnetic resonance spectroscopy [152] and harmonic suppression in power systems [80] are particularly active. However, many application fields remain unaware of the advancements in exponential fitting. For example [132] (2008) discusses an optimization approach to exponential fitting that includes polynomial terms that could be improved using Variable Projection; [155] (2012) uses Prony's Method without any modifications that almost certainly results in a poor fit, and [15] (2007) provides a literature review for fitting exponentials to numerical simulations without noise, but omits significant methods such as Householder's Method, Osborne's Method, Kung's Method, and Variable Projection.

In addition to these applications of exponential fitting using physical measurements, there is continued interest exponential fitting from a theoretical perspective. Exponential fitting is deeply connected to Padé approximation [157] and appears in many other contexts in applied math [4]. Related to these underyling problems, several recent papers investigate exponential fitting in the ℓ_∞ norm; e.g., [18] (2005), [123] (2011), and [14] (2013).

Chapter 2

Variations on a Theme of Prony

Prony's Method has spawned many variants and improvements since its initial development in 1795 [127]. These algorithms share a common property: they identify the exponential coefficients using algebraic relationships — eigenvalues, roots of polynomials, etc. Some of these methods go unrecognized as variants of Prony's Method, and the lack of a common derivation hinders the study of each of these variant's numerical and statistical properties. In this chapter we present a new, common derivation of each of these methods, beginning with Prony's Method in Section 2.1. This improves and extends an earlier unification of Prony's Method, Pisarenko's Method and the Matrix Pencil Method by Ouibrahim [118].

The numerous variations of Prony's Method point to the profound statistical and numerical problems with Prony's Method. In the presence of noise, Prony's Method provides biased estimates of $\boldsymbol{\omega}$ whose covariance is several orders of magnitude larger than the Cramér-Rao lower bound on the covariance of $\boldsymbol{\omega}$. This was first noticed and corrected by Householder in 1949 [69], and was subsequently rediscovered by Osborne in 1975 [116] and Bresler and Macovski in 1986 [23]. These three methods are discussed in Section 2.4. Prony's Method is also unstable when there are many exponentials (e.g., $p > 20$) due to the sensitivity of roots of a polynomial to perturbations of its coefficients. This instability affects many variations of Prony's Method, unless these methods implicitly include additional, extraneous roots to reduce the sensitivity of the true roots to perturbation, as discussed in Section 2.5.

In this chapter, we also provide an analysis of the bias and covariance of Prony's Method, including a new second order bias estimate in Section 2.3, extending the work of Hua [71]. Due to the common derivation, we can then extend these estimates to numerically stable variants of Prony's Method, such as Kung's Method.

In organizing these many variants of these Prony's Method, we also built several new variations. One variation provides numerically stable maximum likelihood estimates by including extraneous exponentials. Another new variant, the Orthogonalized Matrix Pencil Method in Section 2.8, can isolate a few exponentials among many by orthogonalizing against unwanted exponentials.

2.1 Derivation of Prony's Method

Prony's Method consists of two steps: building an autoregressive model that explains the measurements and then recovering the exponential parameters from this model.

An autoregressive model of order p assumes that the value of y_k depends linearly on the preceding p values in $\mathbf{y}$ via the formula

$$y_k = -\sum_{j=0}^{p-1} \alpha_{p-j} y_{k-j} \quad k \geq p. \tag{2.1}$$

The first step of Prony's Method recovers the *lag coefficients* $\boldsymbol{\alpha}$ by combining p copies of (2.1) with different k values into a $p \times p$ system of linear equations[1]

$$\begin{bmatrix} y_0 & y_1 & y_2 & \cdots & y_{p-1} \\ y_1 & y_2 & y_3 & \cdots & y_p \\ \vdots & \vdots & & & \vdots \\ y_{p-1} & y_p & y_{p+1} & \cdots & y_{2p-2} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{p-1} \end{bmatrix} = - \begin{bmatrix} y_p \\ y_{p+1} \\ \vdots \\ y_{2p-1} \end{bmatrix},$$

[1]Some authors (e.g., [120]) prefer flipping the order of $\boldsymbol{\alpha}$, thus flipping the order of columns in $\mathbf{H}$ and resulting in a matrix with Toeplitz structure rather than the Hankel structure.

which we write as

$$\mathbf{H}\boldsymbol{\alpha} = -\mathbf{h}. \tag{2.2}$$

Given this structure, $\mathbf{H}$ is called a Hankel matrix.

The second step of Prony's Method recovers the exponential parameters from $\boldsymbol{\alpha}$. We note the autoregressive model (2.1) can be rewritten as a first order difference equation in state-space form

$$\begin{bmatrix} y_{k-p+1} \\ y_{k-p+2} \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -\alpha_0 & \cdots & -\alpha_{p-2} & -\alpha_{p-1} \end{bmatrix} \begin{bmatrix} y_{k-p} \\ y_{k-p+1} \\ \vdots \\ y_{k-1} \end{bmatrix},$$

which we abbreviate as

$$k \geq p \tag{2.3}$$

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} \tag{2.4}$$

with the initial condition $\mathbf{x}_p$. Then measurements y_k correspond to the output of the system viewed through $\mathbf{C} = \mathbf{e}_p^\top$ (see (1.21)) ; i.e., $y_k = \mathbf{C}\mathbf{x}_k$. If, for the moment, we assume that $\mathbf{A}$ is diagonalizable, then $\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}$, and

$$\mathbf{x}_k = \mathbf{A}^{k-p}\mathbf{x}_p = \mathbf{V}\boldsymbol{\Lambda}^{k-p}\mathbf{V}^{-1}\mathbf{x}_p.$$

Hence y_k is

$$y_k = \mathbf{C}\mathbf{x}_k = \mathbf{e}_p^\top\mathbf{x}_k = \sum_{j=0}^{p-1} \left(\lambda^{-p}[\mathbf{e}_p^*\mathbf{V}]_j[\mathbf{V}^{-1}\mathbf{x}_p]_j\right)\lambda_j^k = \sum_{j=0}^{p-1} a_j\lambda_j^k. \tag{2.5}$$

Thus, the desired exponential parameters $\lambda_j = e^{\omega_j}$ are the eigenvalues of $\mathbf{A}$.

As $\mathbf{A}$ is a *companion matrix*, its eigenvalues are the roots of the polynomial

$$q(t) = t^p + \alpha_{p-1}t^{p-1} + \cdots + \alpha_1 t + \alpha_0 = \sum_k (t-\lambda_k)^{r_k}, \quad \sum_k r_k = p. \tag{2.6}$$

Algorithm 2.1: Prony's Method

Input : Measurements $y_0, y_1, \ldots, y_{2p-1} \in \mathbb{C}$ and model order p.
Output: Exponential parameters $\lambda_0, \lambda_1, \ldots, \lambda_{p-1}$.
1 Form $\mathbf{H} \in \mathbb{C}^{p \times p}$ with $[\mathbf{H}]_{j,k} = y_{j+k}$ and $\mathbf{h} \in \mathbb{C}^p$ with $[\mathbf{h}]_j = y_{j+p}$;
2 Compute $\boldsymbol{\alpha} = -\mathbf{H}^{-1}\mathbf{h}$;
3 Find roots λ_k of $q(t) = t^p + \alpha_{p-1}t^{p-1} + \alpha_{p-2}t^{p-2} + \cdots + \alpha_1 t + \alpha_0$.

Prony's Method, as classically posed [127], found the roots of $q(t)$ rather than the eigenvalues of $\mathbf{A}$, as is now much more common. The polynomial (2.6) is both the characteristic and minimal polynomial of $\mathbf{A}$. If $q(t)$ has any repeated roots, these correspond to Jordan blocks of dimension r_k in $\mathbf{A}$ [68, Thm. 3.3.6]. Hence $\mathbf{A}$ is diagonalizable only when the roots λ_k are distinct. In the exponential fitting setting (1.1) , the assumption that $\mathbf{A}$ is diagonalizable always holds. Measurements y_k are of the form $y_k = \sum_j a_j \lambda_j^k$; if one exponential is repeated, say, $\lambda_j = \lambda_\ell$, we can combine these two exponentials, dropping the ℓth term and setting $a_j \leftarrow a_j + a_\ell$. Thus the eigenvalues of $\mathbf{A}$ will always be distinct in the absence of noise.

Prony's Method is summarized in Algorithm 2.1. As the number of exponentials, p, grows, Prony's Method becomes increasingly susceptible to the severe ill-conditioning that can famously arise in root finding computations [160]. Polynomial roots are extremely sensitive perturbations of the monomial coefficients $\boldsymbol{\alpha}$. If $\boldsymbol{\alpha}$ is perturbed to $\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + \epsilon\boldsymbol{\alpha}^{(1)}$, then the roots $\widetilde{\lambda}_j$ of the perturbed polynomial

$$\widetilde{q}(t) = t^p + \widetilde{\alpha}_{p-1}t^{p-1} + \cdots + \widetilde{\alpha}_1 t + \widetilde{\alpha}_0$$

obey, asymptotically, in the limit of $\epsilon \to 0$ [161, §7.4],

$$\widetilde{\lambda}_j = \lambda_j - \epsilon \frac{\sum_{k=0}^{p-1} \alpha_k^{(1)} \lambda_j^k}{\prod_{j \neq k}(\lambda_j - \lambda_k)} + \mathcal{O}(\epsilon^2). \tag{2.7}$$

In exponential fitting, unperturbed roots will typically be in the neighborhood of the unit circle. Hence, the numerator is typically small, on the order of $\|\boldsymbol{\alpha}^{(1)}\|_1$.

However, the denominator can be arbitrarily small if roots λ_j cluster, causing $\widetilde{\lambda}_k$ to be extremely sensitive to perturbations.[2]

2.2 Equivalent Reformulations of Prony's Method

Prony's Method has deep connections with many existing algorithms. These connections often go unnoticed (with the exception of [118]) because there are four equivalent ways to reformulate Prony's Method. This section reviews these four methods, noting that despite their differences, each becomes increasingly ill-conditioned as the number exponentials grows, just like Prony's Method. This ill-conditioning is very strong: even in the absence of noise, round off error in double precision arithmetic prevents the recovery of $p = 20$ exponentials. The ill-conditioning of these methods can sometimes be mitigated adding extraneous exponentials, described in Section 2.5. The presence of noise in the measurements adds further complications as discussed in Section 2.3. None of these methods obtains the Cramér-Rao lower bound on the covariance of the recovered, noisy $\widetilde{\boldsymbol{\lambda}}$. Moreover, these methods have a second order bias that can make the mean of $\widetilde{\boldsymbol{\lambda}}$ far away from its true value.

2.2.1 Prony Least Squares

As Prony's Method (1795) predated the development of least squares (circa 1805), Prony found the lag coefficients $\boldsymbol{\alpha}$ with a square $\mathbf{H} \in \mathbb{C}^{p \times p}$, neglecting any information in $\mathbf{y}$ past the $2p$ entry. A natural extension of Prony's Method includes this additional information by appending additional rows to the matrix $\mathbf{H}$ in (2.2), finding the least

[2] In Wilkinson's polynomial, the ill-conditioning emerges as a result of the high polynomial order rather than clustering roots. There, he takes $\lambda_j = j + 1$ for $j = 0, 1, \ldots, 19$; then the perturbation of λ_{19} contains terms like $20^{19}/(19!)\alpha_{19}^{(1)}$.

Algorithm 2.2: Prony Least Squares

Input : Measurements $y_0, y_1, \ldots, y_{n-1} \in \mathbb{C}$ and model order p.

Output: Exponential parameters $\lambda_0, \lambda_1, \ldots, \lambda_{p-1}$.

1 Form $\mathbf{H} \in \mathbb{C}^{(n-p)\times p}$ with $[\mathbf{H}]_{j,k} = y_{j+k}$ and $\mathbf{h} \in \mathbb{C}^{n-p}$ with $[\mathbf{h}]_j = y_{j+p}$;

2 Compute $\boldsymbol{\alpha} = -\mathbf{H}^{+}\mathbf{h}$;

3 Find roots λ_k of $q(t) = t^p + \alpha_{p-1}t^{p-1} + \alpha_{p-2}t^{p-2} + \cdots + \alpha_1 t + \alpha_0$.

squares estimate of the overdetermined system:

$$\min_{\boldsymbol{\alpha}\in\mathbb{C}^p} \left\| \underbrace{\begin{bmatrix} y_0 & y_1 & \cdots & y_{p-1} \\ y_1 & y_2 & \cdots & y_p \\ \vdots & \vdots & & \vdots \\ y_{n-p-1} & y_{n-p} & \cdots & y_{n-2} \end{bmatrix}}_{\mathbf{H}\in\mathbb{C}^{(n-p)\times p}} \underbrace{\begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{p-1} \end{bmatrix}}_{\boldsymbol{\alpha}\in\mathbb{C}^p} + \underbrace{\begin{bmatrix} y_p \\ y_{p+1} \\ \vdots \\ y_{n-1} \end{bmatrix}}_{\mathbf{h}\in\mathbb{C}^{n-p}} \right\|_2^2 . \tag{2.8}$$

It is unknown who first combined Prony's Method with least squares, but this *Prony Least Squares* approach (Algorithm 2.2) appears in several early textbooks; e.g., Whittaker and Robinson (1924) [159, §180] and Hildebrand (1956) [65, §9.4].

In the absence of noise, Prony Least Squares is equivalent to Prony's Method, as $\mathbf{h}$ is in the range of $\mathbf{H}$. However, the additional rows tend to improve the condition number of $\mathbf{H}$, fortifying Prony Least Squares against round off error as compared to Prony's Method as Figure 2.1 illustrates. This formulation also exposes why Prony's Method provides poor estimates of $\boldsymbol{\lambda}$ in the presence of noise: the norm (2.8) does not minimize the mismatch between model and measurements, but instead correlates errors in $\mathbf{y}$ with $\boldsymbol{\alpha}$. Section 2.3 analyzes how this affects the estimates of $\boldsymbol{\lambda}$ in the presence noise.

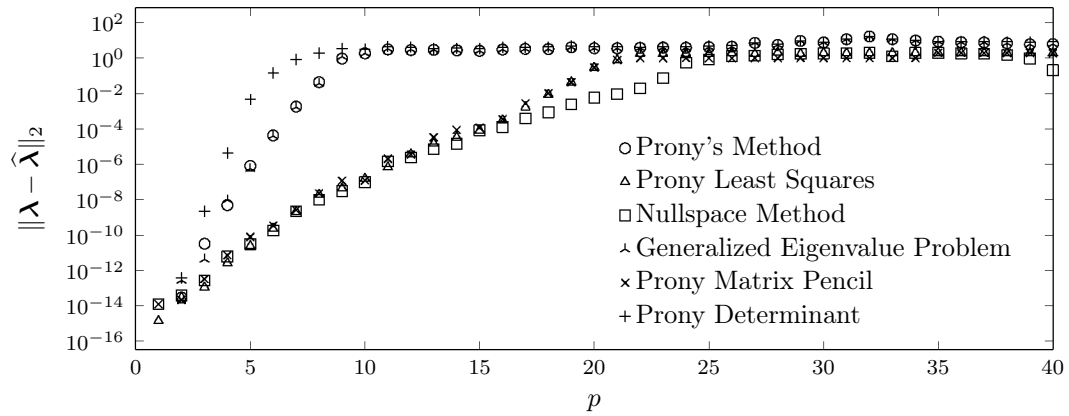


Figure 2.1 : An illustration of the sensity of several Prony-type methods to round off errors. Each method attempted to recover p exponentials from $n = 10^3$ measurements without any noise. The true exponential parameters were first p entries in $\widehat{\boldsymbol{\lambda}} = e^{\widehat{\boldsymbol{\omega}}}$, where $\widehat{\boldsymbol{\omega}} = [1i\ 1.1i\ 1.2i\ 1.3i\ \ldots\ 5.9i\ 6i]^\top$. In comparison, a nonlinear least squares method (Section 3.1.4) converged exactly to $\widehat{\boldsymbol{\lambda}}$ in double precision arithmetic.

2.2.2 Nullspace Method

The Nullspace Method rearranges the equations in Prony Least Squares so the lag coefficients emerge as the nullspace of an expanded Hankel matrix $\underline{\mathbf{H}}$. The structure of (2.8) enforces the condition that q is monic, but we can relax this condition finding so as to find a scalar multiple of q, i.e.,

$$\underline{q}(t) = \underline{\alpha}_p t^p + \underline{\alpha}_{p-1} t^{p-1} + \cdots + \underline{\alpha}_1 t + \underline{\alpha}_0$$

corresponding to the autoregressive model (cf. (2.1))

$$\underline{\alpha}_p y_k = -\sum_{j=1}^{p} \underline{\alpha}_{p-j} y_{k-j}, \quad k \geq p.$$

We then find the coefficients $\underline{\boldsymbol{\alpha}} \in \mathbb{C}^{p+1}$ by finding the nullspace of $\underline{\mathbf{H}} = [\mathbf{H}, \mathbf{h}]$. Starting from Prony Least Squares

$$\min_{\boldsymbol{\alpha} \in \mathbb{C}^p} \left\| \begin{bmatrix} \mathbf{H} & \mathbf{h} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ -1 \end{bmatrix} \right\|_2 \quad \text{is equivalent to} \quad \min_{\underline{\boldsymbol{\alpha}} \in \mathbb{C}^{p+1}} \|\underline{\mathbf{H}}\, \underline{\boldsymbol{\alpha}}\|_2 . \tag{2.9}$$

Algorithm 2.3: Nullspace Method

Input : Measurements $y_0, y_1, \ldots, y_{n-1} \in \mathbb{C}$ and model order p.

Output: Exponential parameters $\lambda_0, \lambda_1, \ldots, \lambda_{p-1}$.

1 Form $\underline{\mathbf{H}} \in \mathbb{C}^{(n-p)\times(p+1)}$ with $[\mathbf{H}]_{j,k} = y_{j+k}$;

2 Compute SVD of $\underline{\mathbf{H}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ where $[\mathbf{\Sigma}]_{j,j} = \sigma_j$ and $\sigma_j \geq \sigma_k$ for all $j > k$;

3 Set $\underline{\boldsymbol{\alpha}} = [\mathbf{V}]_{\cdot,p+1} \in \mathbb{C}^{p+1}$;

4 Find roots λ_k of $\underline{q}(t) = \underline{\alpha}_p t^p + \underline{\alpha}_{p-1} t^{p-1} + \underline{\alpha}_{p-2} t^{p-2} + \cdots + \underline{\alpha}_1 t + \underline{\alpha}_0$.

There are an infinite number of solutions to the problem on the right; Prony's Method chooses the one corresponding to $\underline{\alpha}_{p+1}$, while the Nullspace Method chooses the one where $\|\underline{\boldsymbol{\alpha}}\|_2 = 1$. In the presence of noise, $\underline{\mathbf{H}}$ is unlikely to have a nullspace, so we estimate the true nullspace as the smallest right singular vector of $\underline{\mathbf{H}}$ (see Algorithm 2.3 for details).

The Nullspace Method only shows up once in the literature for exponential fitting methods that directly use measurements $\mathbf{y}$. When Golub was asked by Dudly in 1977 how to improve the stability of Prony's Method, he suggested normalizing $\|\underline{\boldsymbol{\alpha}}\|_2 = 1$ [43, p.17]. As Figure 2.1 illustrates, this does improve the conditioning: the error using the Nullspace Method is less than the error using Prony Least Squares. However, this idea is prominent in methods that use autocovariance information about $\mathbf{y}$ such as Pisarenko's Method (described in Section 2.6.2) and the *Minimum Norm Method* [93].

2.2.3 Prony Matrix Pencil Method

A third reformulation of Prony Least Squares (2.8) finds the exponential parameters $\boldsymbol{\lambda}$ directly as eigenvalues of a rectangular matrix pencil, rather than as roots of a polynomial. Although we might hope this avoids the numerical instabilities of Prony's Method, the matrix pencil approach is equally flawed.

The matrix pencil approach starts by appending columns from $\mathbf{H}$ to the left of $\mathbf{h}$, resulting in the linear system

$$\begin{bmatrix} y_0 & y_1 & \cdots & y_{p-1} \\ y_1 & y_2 & \cdots & y_p \\ \vdots & \vdots & & \vdots \\ y_{n-p-1} & y_{n-p} & \cdots & y_{n-2} \end{bmatrix} \begin{bmatrix} 0 & & & -\alpha_0 \\ 1 & \ddots & & \vdots \\ & \ddots & 0 & -\alpha_{p-2} \\ & & 1 & -\alpha_{p-1} \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & \cdots & y_p \\ y_2 & y_3 & \cdots & y_{p+1} \\ \vdots & \vdots & & \vdots \\ y_{n-p} & y_{n-p+1} & \cdots & y_{n-1} \end{bmatrix}. \tag{2.10}$$

We call the first Hankel matrix $\mathbf{H}_0 \in \mathbb{C}^{(n-p)\times p}$ ($[\mathbf{H}_0]_{j,k} = y_{j+k}$), the second Hankel matrix $\mathbf{H}_1 \in \mathbb{C}^{(n-p)\times p}$ ($[\mathbf{H}_1]_{j,k} = y_{j+k+1}$), and the companion matrix $\mathbf{C}$. Notice that $\mathbf{C} = \mathbf{A}^\top$, where $\mathbf{A}$ is the same matrix from (2.4); as such, the eigenvalues of $\mathbf{C}$ are the exponential parameters $\boldsymbol{\lambda}$. One approach to compute $\boldsymbol{\lambda}$ would be to form $\mathbf{C}$ directly as $\mathbf{C} = \mathbf{H}_0^+\mathbf{H}_1$ and then compute the spectrum of $\mathbf{C}$. Another approach would be to note the eigenvalues of $\mathbf{C}$ are also the eigenvalues of the matrix pencil

$$\lambda\mathbf{H}_0\mathbf{x} = \mathbf{H}_1\mathbf{x}, \tag{2.11}$$

since if $(\lambda_j, \mathbf{x}_j)$ is an eigenpair of $\mathbf{C}$, then

$$\mathbf{H}_0\mathbf{C}\mathbf{x}_j = \lambda_j\mathbf{H}_0\mathbf{x}_j = \mathbf{H}_1\mathbf{x}_j.$$

Most modern Prony-type methods form a matrix similar to $\mathbf{C}$ from $\mathbf{H}_0$ and $\mathbf{H}_1$, and then compute the eigenvalues of $\mathbf{C}$ to recover $\boldsymbol{\lambda}$ as described in Algorithm 2.4. For example, the *Matrix Pencil Method* [72] forms $\mathbf{C}$ using a rank-truncated pseudoinverse of $\mathbf{H}_0$. *Kung's Method* [94, 12] is similar, computing $\mathbf{C}$ from left singular vectors of $\mathbf{H}_0$. More details on these methods are provided in Section 2.5.

The direct matrix pencil formulation (2.11) is not often used in practice, because rectangular matrix pencil problems are fraught with numerical challenges. Although in the absence of noise there exist p distinct solutions λ_j, infinitesimal perturbations

Algorithm 2.4: Prony Matrix Pencil Method

Input : Measurements $y_0, y_1, \ldots, y_{n-1} \in \mathbb{C}$ and model order p.

Output: Exponential parameters $\lambda_0, \lambda_1, \ldots, \lambda_{p-1}$.

1 Form $\mathbf{H}_0 \in \mathbb{C}^{(n-p)\times p}$ with $[\mathbf{H}_0]_{j,k} = y_{j+k}$;

2 Form $\mathbf{H}_1 \in \mathbb{C}^{(n-p)\times p}$ with $[\mathbf{H}_1]_{j,k} = y_{j+k+1}$;

3 Solve $\mathbf{C} = \mathbf{H}_0^{+}\mathbf{H}_1$;

4 Find eigenvalues λ_k of $\mathbf{C}$.

Algorithm 2.5: Prony Generalized Eigenvalue Problem

Input : Measurements $y_0, y_1, \ldots, y_{2p-1} \in \mathbb{C}$ and model order p.

Output: Exponential parameters $\lambda_0, \lambda_1, \ldots, \lambda_{p-1}$.

1 Form $\mathbf{H}_0 \in \mathbb{C}^{p\times p}$ with $[\mathbf{H}_0]_{j,k} = y_{j+k}$;

2 Form $\mathbf{H}_1 \in \mathbb{C}^{p\times p}$ with $[\mathbf{H}_1]_{j,k} = y_{j+k+1}$;

3 Find eigenvalues λ_j of the generalized eigenvalue problem $\lambda\mathbf{H}_0\mathbf{x} = \mathbf{H}_1\mathbf{x}$.

will, in general, cause there to be no solution (cf., [164]). One approach correcting this is to find λ_j as an eigenvalue of the smallest perturbation of $\mathbf{H}_0$ and $\mathbf{H}_1$ [22]; e.g.,

$$\begin{aligned} &\min_{\mathbf{E}_0,\mathbf{E}_1,\lambda,\mathbf{x}} \|\mathbf{E}_0\|_{\mathrm{F}}^2 + \|\mathbf{E}_1\|_{\mathrm{F}}^2 \\ &\text{such that } \lambda(\mathbf{H}_0 + \mathbf{E}_0)\mathbf{x} = (\mathbf{H}_1 + \mathbf{E}_1)\mathbf{x} \quad \text{and} \quad \|\mathbf{x}\|_2 = 1. \end{aligned} \tag{2.12}$$

Not only does this approach not enforce the Hankel structure that $\mathbf{E}_0$ and $\mathbf{E}_1$ must have, but it uses different perturbations $\mathbf{E}_0$ and $\mathbf{E}_1$ for each λ_j. We can avoid the difficulties of solving the matrix pencil problem (2.11) by truncating the bottom rows of $\mathbf{H}_0$ and $\mathbf{H}_1$, resulting in a square *generalized eigenvalue problem*, yielding λ_j as described in Algorithm 2.5.

Although this approach does not explicitly find the roots of a polynomial, it still suffers from the same ill-conditioning whether we estimate $\mathbf{C}$ or directly solve the generalized eigenvalue problem. We can analyze either approach by using the rectangular matrix pencil and finding $\mathbf{C} = \mathbf{H}_0^{+}\mathbf{H}_1$ ($\mathbf{C} = \mathbf{H}_0^{-1}\mathbf{H}_1$ in the generalized

eigenvalue case).[3] The eigenvalue condition number for eigenvalue λ_j in $\mathbf{C}$ is

$$\kappa(\lambda_j) = \frac{\|\mathbf{z}_j\|\|\mathbf{x}_j\|}{|\mathbf{z}_j^*\mathbf{x}_j|}, \tag{2.13}$$

where $\mathbf{z}_j$ and $\mathbf{x}_j$ are the left and right eigenvectors of $\mathbf{C}$. Since $\mathbf{C}$ is a companion matrix, the left and right eigenvectors are [149, eq. (10), (11)]

$$\mathbf{x}_j = \begin{bmatrix} 1 & \lambda & \lambda^2 & \cdots & \lambda^{p-1} \end{bmatrix}^\top \quad \text{and}$$
$$\mathbf{z}_j = \begin{bmatrix} b_0 & b_1 & b_2 & \cdots & b_{p-1} \end{bmatrix}^\top \quad \text{where} \quad \sum_{j=0}^{p-1} b_j z^j = \frac{q(z) - q(\lambda)}{z - \lambda}.$$

This special structure gives $\mathbf{z}_j^*\mathbf{x}_j = q'(\lambda_j)$ and hence

$$\kappa(\lambda_j) = \frac{\|\mathbf{z}_j\|}{|q'(\lambda_j)|} \sqrt{\frac{1 - |\lambda_j|^{2p}}{1 - |\lambda_j|^2}}. \tag{2.14}$$

This condition number does not include the ill-conditioning that derives from noisy data, as perturbations to $\mathbf{H}_0$ and $\mathbf{H}_1$ ($\mathbf{E}_0$ and $\mathbf{E}_1$) yield amplified perturbations to $\mathbf{C}$:

$$\widetilde{\mathbf{C}} = (\mathbf{H}_0 + \mathbf{E}_0)^+(\mathbf{H}_1 + \mathbf{E}_1) = \mathbf{C} - \mathbf{H}_0^+\mathbf{E}_0\mathbf{C} + \mathbf{H}_0^+\mathbf{E}_1 + \mathcal{O}(\|\mathbf{E}_0\|^2). \tag{2.15}$$

The resulting condition number comparable to that for Prony Least Squares (2.7), as illustrated in the computations in Figure 2.1.

2.2.4 Prony Determinant Method

A fourth variation of Prony's Method computes the coefficients of the polynomial $\underline{q}$ using the determinant. First note that the Hankel matrix $\mathbf{H} \in \mathbb{C}^{(p+1)\times p}$ can be decomposed into the product of two Vandermonde matrices $\mathbf{V}_m(\boldsymbol{\lambda}) \in \mathbb{C}^{m\times p}$ where $[\mathbf{V}(\boldsymbol{\lambda})]_{j,k} = \lambda_k^j$ and a diagonal matrix $\mathbf{A} = \mathrm{diag}(\mathbf{a})$:

$$\mathbf{H}_{n,m} = \mathbf{V}_n(\boldsymbol{\lambda})\mathbf{A}\mathbf{V}_m(\boldsymbol{\lambda})^\top \in \mathbb{C}^{n\times m}. \tag{2.16}$$

[3] Generalized eigenvalue perturbation analysis is more intricate and does not apply to the rectangular case; see, e.g., [143, Ch. 6].

This is an immediate result of matrix multiplication and can be generalized to arbitrary Hankel matrices using confluent Vandermonde matrices [151, Ch. 4]. Then, consider the determinate of the matrix

$$\underline{q}(t) = \det \begin{bmatrix} \mathbf{H}_{p+1,p} & \mathbf{V}_{p+1}(t) \end{bmatrix} = \det \begin{bmatrix} \mathbf{V}_{p+1}(\boldsymbol{\lambda}) \mathbf{A} \mathbf{V}_p(\boldsymbol{\lambda})^\top & \mathbf{V}_{p+1}(t) \end{bmatrix}. \tag{2.17}$$

As there are p distinct exponentials, $\mathbf{V}_p(\boldsymbol{\lambda})$ has full rank; hence the rank of this matrix is only decreased when t is one of the exponential coefficients λ_j. Thus q has roots λ_j. The polynomial q is given by

$$\underline{q}(t) = \det \begin{bmatrix} y_0 & y_1 & \cdots & y_{p-1} & t^0 \\ y_1 & y_2 & \cdots & y_p & t^1 \\ \vdots & \vdots & & \vdots & \vdots \\ y_p & y_{p+1} & \cdots & y_{2p-1} & t^p \end{bmatrix} = \sum_{j=0}^{p} t^j \det \mathbf{H}_j \tag{2.18}$$

where $\mathbf{H}_j$ is $\mathbf{H}_{p+1,p}$ with the jth row deleted.

This approach shows up once in the *Method of Moments* [78, App. I], although in that algorithm linear combinations of measurements y_j are used for each entry in the Hankel matrix; see Section 2.7.2 for more details. Although similar, this is not Prony's Method with $\boldsymbol{\alpha}$ found using Cramer's Rule [68, § 0.8.3]. As might be expected, this method is very numerically unstable. The additional round off errors accumulated in computing the determinant result in an even more numerically sensitive algorithm (given in Algorithm 2.6), as Figure 2.1 illustrates. Every other method in this section can be improved by adding extraneous exponentials, as discussed in Section 2.5, but this determinant method fails, as $\mathbf{H}_{p+2,p+1}$ is rank deficient.

Algorithm 2.6: Prony Determinant Method

Input : Measurements $y_0, y_1, \ldots, y_{2p-1} \in \mathbb{C}$ and model order p.

Output: Exponential parameters $\lambda_0, \lambda_1, \ldots, \lambda_{p-1}$.

1 Form $\mathbf{H} \in \mathbb{C}^{(p+1)\times p}$ with $[\mathbf{H}]_{j,k} = y_{j+k}$ and $\mathbf{h} \in \mathbb{C}^p$ with $[\mathbf{h}]_j = y_{j+p}$;

2 **for** $j = 0, \ldots, p-1$ **do**

3 Form $\mathbf{H}_j$ by deleting the jth row of $\mathbf{H}$;

4 $\underline{\alpha}_j = \det \mathbf{H}_j$;

5 Find roots λ_k of $\underline{q}(t) = \underline{\alpha}_p t^p + \underline{\alpha}_{p-1} t^{p-1} + \underline{\alpha}_{p-2} t^{p-2} + \cdots + \underline{\alpha}_1 t + \underline{\alpha}_0$.

2.3 Statistics of Prony's Method

The existing analysis of Prony's Method treats the large n limit, invoking the *Law of Large Numbers* to place very weak restrictions on the classes of perturbations considered (e.g., [82]). In contrast, we are interested in the performance of Prony's Method for a finite n under the assumption that measurements $\mathbf{y}$ are perturbed by additive noise sampling a normal distribution. We record noisy measurements $\widetilde{\mathbf{y}}$:

$$\widetilde{\mathbf{y}} = \mathbf{y} + \epsilon \mathbf{g} \quad \text{where} \quad \mathbf{g} \sim \mathcal{N}(\, 0, \boldsymbol{\Sigma}). \tag{2.19}$$

Then, the perturbation $\widetilde{\mathbf{y}}$ results in perturbations of $\mathbf{H}$ and $\mathbf{h}$, $\widetilde{\mathbf{H}}$ and $\widetilde{\mathbf{h}}$. These yield the perturbed Prony Least Squares estimate

$$\widetilde{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \, \|\widetilde{\mathbf{H}}\boldsymbol{\alpha} + \widetilde{\mathbf{h}}\|_2. \tag{2.20}$$

In this section, we derive estimates of $\widetilde{\boldsymbol{\alpha}}$ in the limit $\epsilon \to 0$, expanding $\widetilde{\boldsymbol{\alpha}}$ in a power series

$$\widetilde{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}} + \epsilon \boldsymbol{\alpha}^{(1)} + \epsilon^2 \boldsymbol{\alpha}^{(2)} + \mathcal{O}(\epsilon^3),$$

where, for clarity, $\widehat{\boldsymbol{\alpha}}$ is the true, unperturbed value of the lag coefficients. These estimates show that in the limit $\epsilon \to 0$, the covariance of $\widetilde{\boldsymbol{\alpha}}$ fails to obtain the

Cramér-Rao bound and $\mathsf{E}[\boldsymbol{\alpha}^{(2)}] \neq \mathbf{0}$. Both effects are amplified by the ill-conditioning of roots $\widetilde{\boldsymbol{\lambda}}$ to perturbations of the lag coefficients $\widetilde{\boldsymbol{\alpha}}$; see (2.7). This leads to estimates $\widetilde{\boldsymbol{\lambda}}$ where $\|\widetilde{\boldsymbol{\lambda}} - \widehat{\boldsymbol{\lambda}}\|$ is large and the covariance of $\widetilde{\boldsymbol{\lambda}}$ is orders of magnitude larger than the Cramér-Rao bound.

Our derivation follows Osborne (1975) [116] who first corrected Prony's Method to work in the proper norm (see Section 2.4). Although he did not derive covariance estimates for $\widetilde{\boldsymbol{\alpha}}$ and $\widetilde{\boldsymbol{\lambda}}$, these are latent in his work. We also extend these results, providing a new second order bias estimate.

2.3.1 First Order

To obtain the first order perturbation $\boldsymbol{\alpha}^{(1)}$, we first rearrange (2.8)

$$\widetilde{\mathbf{H}}\boldsymbol{\alpha} + \widetilde{\mathbf{h}} = (\mathbf{H} + \epsilon\mathbf{E})\boldsymbol{\alpha} + (\mathbf{h} + \epsilon\mathbf{e}) = \mathbf{H}\boldsymbol{\alpha} + \mathbf{h} + \epsilon(\mathbf{E}\boldsymbol{\alpha} + \mathbf{e}),$$

where $[\mathbf{E}]_{j,k} = g_{j,k}$ and $[\mathbf{e}]_j = g_{j+p}$. Then, $\mathbf{E}\boldsymbol{\alpha} + \mathbf{e}$ is a linear combination of $\mathbf{g}$:

$$\begin{bmatrix} g_0 & g_1 & \cdots & g_{p-1} \\ g_1 & g_2 & \cdots & g_p \\ \vdots & \vdots & & \vdots \\ g_{n-p-1} & g_{n-p} & \cdots & g_{n-2} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{p-1} \end{bmatrix} + \begin{bmatrix} g_p \\ g_{p+1} \\ \vdots \\ g_{n-1} \end{bmatrix} = \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_{p-1} & 1 & & \\ & \alpha_0 & \alpha_1 & \cdots & \alpha_{p-1} & 1 & \\ & & \ddots & \ddots & & \ddots & \ddots \\ & & & \alpha_0 & \alpha_1 & \cdots & \alpha_{p-1} & 1 \end{bmatrix} \begin{bmatrix} g_0 \\ g_1 \\ \vdots \\ g_{n-1} \end{bmatrix} \tag{2.21}$$

We call the right matrix $\mathbf{T}(\boldsymbol{\alpha})$. Hence, in the presence of noise, Prony Least Squares solves

$$\widetilde{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{H}\boldsymbol{\alpha} + \mathbf{h} + \epsilon\mathbf{T}(\boldsymbol{\alpha})\mathbf{g}\|_2. \tag{2.22}$$

To compute $\boldsymbol{\alpha}^{(1)}$, we apply the perturbation expansion of $\widetilde{\boldsymbol{\alpha}}$

$$\min_{\boldsymbol{\alpha}^{(1)}} \|\mathbf{H}(\widehat{\boldsymbol{\alpha}} + \epsilon\boldsymbol{\alpha}^{(1)}) + \mathbf{h} + \epsilon\mathbf{T}(\widehat{\boldsymbol{\alpha}})\mathbf{g} + \mathcal{O}(\epsilon^2)\|_2,$$

and match orders, yielding

$$\boldsymbol{\alpha}^{(1)} = -\mathbf{H}^{+}\mathbf{T}(\widehat{\boldsymbol{\alpha}})\mathbf{g}. \tag{2.23}$$

As $\boldsymbol{\alpha}^{(1)}$ is a linear combination of normally distributed random variables, it is also a normally distributed random variable. As such $\boldsymbol{\alpha}^{(1)}$ is completely described by its mean

$$\mathsf{E}\left[\boldsymbol{\alpha}^{(1)}\right] = -\mathbf{H}^{+}\mathbf{T}(\widehat{\boldsymbol{\alpha}})\,\mathsf{E}[\mathbf{g}] = \mathbf{0}. \tag{2.24}$$

and its covariance

$$\begin{aligned}\mathsf{Cov}\left[\boldsymbol{\alpha}^{(1)}\right] &= \mathsf{E}\left[\boldsymbol{\alpha}^{(1)}(\boldsymbol{\alpha}^{(1)})^{*}\right] = \mathbf{H}^{+}\mathbf{T}(\widehat{\boldsymbol{\alpha}})\,\mathsf{E}\left[\mathbf{g}\mathbf{g}^{*}\right]\mathbf{T}^{*}(\widehat{\boldsymbol{\alpha}})\mathbf{H}^{+*} \\ &= \mathbf{H}^{+}\mathbf{T}(\widehat{\boldsymbol{\alpha}})\boldsymbol{\Sigma}\mathbf{T}^{*}(\widehat{\boldsymbol{\alpha}})\mathbf{H}^{+*}.\end{aligned} \tag{2.25}$$

From these perturbation estimates of $\widetilde{\boldsymbol{\alpha}}$, we can derive perturbation estimates of $\widetilde{\boldsymbol{\lambda}} = \widehat{\boldsymbol{\lambda}} + \epsilon\boldsymbol{\lambda}^{(1)} + \mathcal{O}(\epsilon^2)$ using (2.7),

$$\boldsymbol{\lambda}^{(1)} = \mathbf{L}\boldsymbol{\alpha}^{(1)} = \begin{bmatrix} \prod_{\substack{j=0 \\ j\neq 0}}^{p-1}(\widehat{\lambda}_0 - \widehat{\lambda}_j)^{-1} & & \\ & \ddots & \\ & & \prod_{\substack{j=0 \\ j\neq p-1}}^{p-1}(\widehat{\lambda}_{p-1} - \widehat{\lambda}_j)^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\lambda}_0^0 & \cdots & \widehat{\lambda}_0^{p-1} \\ \vdots & & \vdots \\ \widehat{\lambda}_{p-1}^0 & \cdots & \widehat{\lambda}_{p-1}^{p-1} \end{bmatrix} \boldsymbol{\alpha}^{(1)}. \tag{2.26}$$

Then, we can apply this perturbation estimate to estimate the asymptotic covariance of $\widetilde{\boldsymbol{\lambda}}$:

$$\mathsf{Cov}[\widetilde{\boldsymbol{\lambda}}] = \epsilon^2\mathbf{L}\,\mathsf{Cov}[\boldsymbol{\alpha}^{(1)}]\mathbf{L}^{*} + \mathcal{O}(\epsilon^3) = \epsilon^2\mathbf{L}\mathbf{H}^{+}\mathbf{T}(\widehat{\boldsymbol{\alpha}})\boldsymbol{\Sigma}\mathbf{T}(\widehat{\boldsymbol{\alpha}})^{*}\mathbf{H}^{+*}\mathbf{L}^{*} + \mathcal{O}(\epsilon^3). \tag{2.27}$$

When two true roots $\widehat{\boldsymbol{\lambda}}$ are close, $\mathbf{L}$ will be large, amplifying perturbations to $\widehat{\boldsymbol{\alpha}}$. Figure 2.2 illustrates that this effect can be extreme; the Prony Least Squares estimate of $\widetilde{\boldsymbol{\lambda}}$ has a covariance two orders of magnitude larger than the Cramér-Rao bound for a well conditioned $\widehat{\boldsymbol{\lambda}}$ and six orders of magnitude larger for an ill-conditioned $\widehat{\boldsymbol{\lambda}}$.

2.3.2 Second Order Bias

In addition to the large covariance of $\widetilde{\boldsymbol{\lambda}}$, Prony Least Squares has a second order bias; i.e., $\mathsf{E}[\boldsymbol{\alpha}^{(2)}] \neq \mathbf{0}$. We note this by solving Prony Least Squares using the normal equations

$$\widetilde{\mathbf{H}}^{*}\widetilde{\mathbf{H}}\widetilde{\boldsymbol{\alpha}} = -\widetilde{\mathbf{H}}^{*}\widetilde{\mathbf{h}}. \tag{2.28}$$

Expanding $\widetilde{\boldsymbol{\alpha}}$ in a power series, we obtain:

$$\begin{aligned}
\mathbf{H}^{*}\mathbf{H}\widehat{\boldsymbol{\alpha}} &= -\mathbf{H}^{*}\mathbf{h}, \\
\mathbf{H}^{*}\mathbf{H}\boldsymbol{\alpha}^{(1)} &= -\mathbf{H}^{*}\mathbf{e} - \mathbf{E}^{*}\mathbf{h} - (\mathbf{E}^{*}\mathbf{H} + \mathbf{H}^{*}\mathbf{E})\widehat{\boldsymbol{\alpha}}, \\
\mathbf{H}^{*}\mathbf{H}\boldsymbol{\alpha}^{(2)} &= -\mathbf{E}^{*}\mathbf{e} - (\mathbf{E}^{*}\mathbf{H} + \mathbf{H}^{*}\mathbf{E})\boldsymbol{\alpha}^{(1)} - \mathbf{E}^{*}\mathbf{E}\widehat{\boldsymbol{\alpha}}.
\end{aligned}$$

As expectation is a linear operator, $(\mathbf{H}^{*}\mathbf{H})^{-1}$ can be applied at the last step as to recover $\boldsymbol{\alpha}^{(2)}$. Thus, to determine $\mathsf{E}[\boldsymbol{\alpha}^{(2)}]$, we compute the expectation of the terms on the right side of the ϵ^2 equation. For the first term $\mathbf{E}^{*}\mathbf{e}$,

$$\mathsf{E}[\mathbf{E}^{*}\mathbf{e}]_j = \mathsf{E}\left[\begin{bmatrix} g_0 & g_1 & \cdots & g_{p-1} \\ g_1 & g_2 & \cdots & g_p \\ \vdots & \vdots & & \vdots \\ g_{m-1} & g_m & \cdots & g_{m+p-2} \end{bmatrix}^{*} \begin{bmatrix} g_p \\ g_{p+1} \\ \vdots \\ g_{m+p-1} \end{bmatrix}\right]_j = \mathsf{E}\left[\sum_{i=0}^{m-1} g_{i+j}^{*} g_{i+p}\right] = m\delta_{j,p},$$

for $0 \leq j < p$; hence $\mathsf{E}[\mathbf{E}^{*}\mathbf{e}] = \mathbf{0}$. The expectation of the $\mathbf{E}^{*}\mathbf{E}$ term follows a similar pattern

$$\mathsf{E}[\mathbf{E}^{*}\mathbf{E}]_{j,k} = \mathsf{E}\left[\sum_{i=0}^{m-1} g_{i+j}^{*} g_{i+k}\right] = m\delta_{j,k}.$$

Hence $\mathsf{E}[\mathbf{E}^{*}\mathbf{E}] = m\mathbf{I}$. To compute the expectation of the middle term, we replace $\boldsymbol{\alpha}^{(1)}$ using the $\mathcal{O}(\epsilon)$ equation:

$$(\mathbf{E}^{*}\mathbf{H} + \mathbf{H}^{*}\mathbf{E})\boldsymbol{\alpha}^{(1)} = (\mathbf{E}^{*}\mathbf{H} + \mathbf{H}^{*}\mathbf{E})(\mathbf{H}^{*}\mathbf{H})^{-1}(-\mathbf{H}^{*}\mathbf{e} - \mathbf{E}^{*}\mathbf{h} - (\mathbf{E}^{*}\mathbf{H} + \mathbf{H}^{*}\mathbf{E})\widehat{\boldsymbol{\alpha}}).$$

Upon taking the expectation, only terms with g^* and g remain,

$$\mathsf{E}[(\mathbf{E}^*\mathbf{H}+\mathbf{H}^*\mathbf{E})\boldsymbol{\alpha}^{(1)}] = -\,\mathsf{E}[\mathbf{E}^*\mathbf{H}(\mathbf{H}^*\mathbf{H})^{-1}\mathbf{H}^*(\mathbf{e}+\mathbf{E}\widehat{\boldsymbol{\alpha}})] - \mathsf{E}[\mathbf{H}^*\mathbf{E}(\mathbf{H}^*\mathbf{H})^{-1}\mathbf{E}^*(\mathbf{h}+\mathbf{H}\widehat{\boldsymbol{\alpha}})].$$

Since $\mathbf{H}\widehat{\boldsymbol{\alpha}} = -\mathbf{h}$, the second term vanishes. To evaluate the first term, define $\mathbf{P} = \mathbf{H}(\mathbf{H}^*\mathbf{H})^{-1}\mathbf{H}^*$ (the orthogonal projector onto the range of $\mathbf{H}$). Then,

$$\begin{aligned}
[\mathbf{w}]_i &= [\mathbf{E}^*\mathbf{P}(\mathbf{e}+\mathbf{E}\widehat{\boldsymbol{\alpha}})]_i = \sum_{j=0}^{m-1} g^*_{i+j}[\mathbf{P}(-\mathbf{e}+\mathbf{E}\widehat{\boldsymbol{\alpha}})]_j \\
&= \sum_{j=0}^{m-1}\sum_{k=0}^{m-1} g^*_{i+j}[\mathbf{P}]_{j,k}([\mathbf{e}]_k + [\mathbf{E}\widehat{\boldsymbol{\alpha}}]_k) = \sum_{j,k,\ell=0}^{m-1,m-1,p-1} g^*_{i+j}[\mathbf{P}]_{j,k}(g_{k+p} + [\mathbf{E}]_{k,\ell}\widehat{\alpha}_\ell) \\
&= \sum_{j,k,\ell=0}^{m-1,m-1,p-1} g^*_{i+j}[\mathbf{P}]_{j,k}(g_{k+p} + g_{k+\ell}\widehat{\alpha}_\ell) = \sum_{j=0}^{m-1}[\mathbf{P}]_{j,i+j-p} + \sum_{j,\ell=0}^{m-1,p-1}[\mathbf{P}]_{j,i+j-\ell}\widehat{\alpha}_\ell.
\end{aligned}$$

Combining these results gives the second order bias:

$$\begin{aligned}
\mathsf{E}[\boldsymbol{\alpha}^{(2)}] &= -m(\mathbf{H}^*\mathbf{H})^{-1}\widehat{\boldsymbol{\alpha}} + (\mathbf{H}^*\mathbf{H})^{-1}\mathbf{w} \\
\text{where } [\mathbf{w}]_i &= \sum_{j,\ell=0}^{m-1,p-1}[\mathbf{P}]_{j,i+j-\ell}\widehat{\alpha}_\ell + \sum_{j=0}^{m-1}[\mathbf{P}]_{j,i+j-p}.
\end{aligned} \tag{2.29}$$

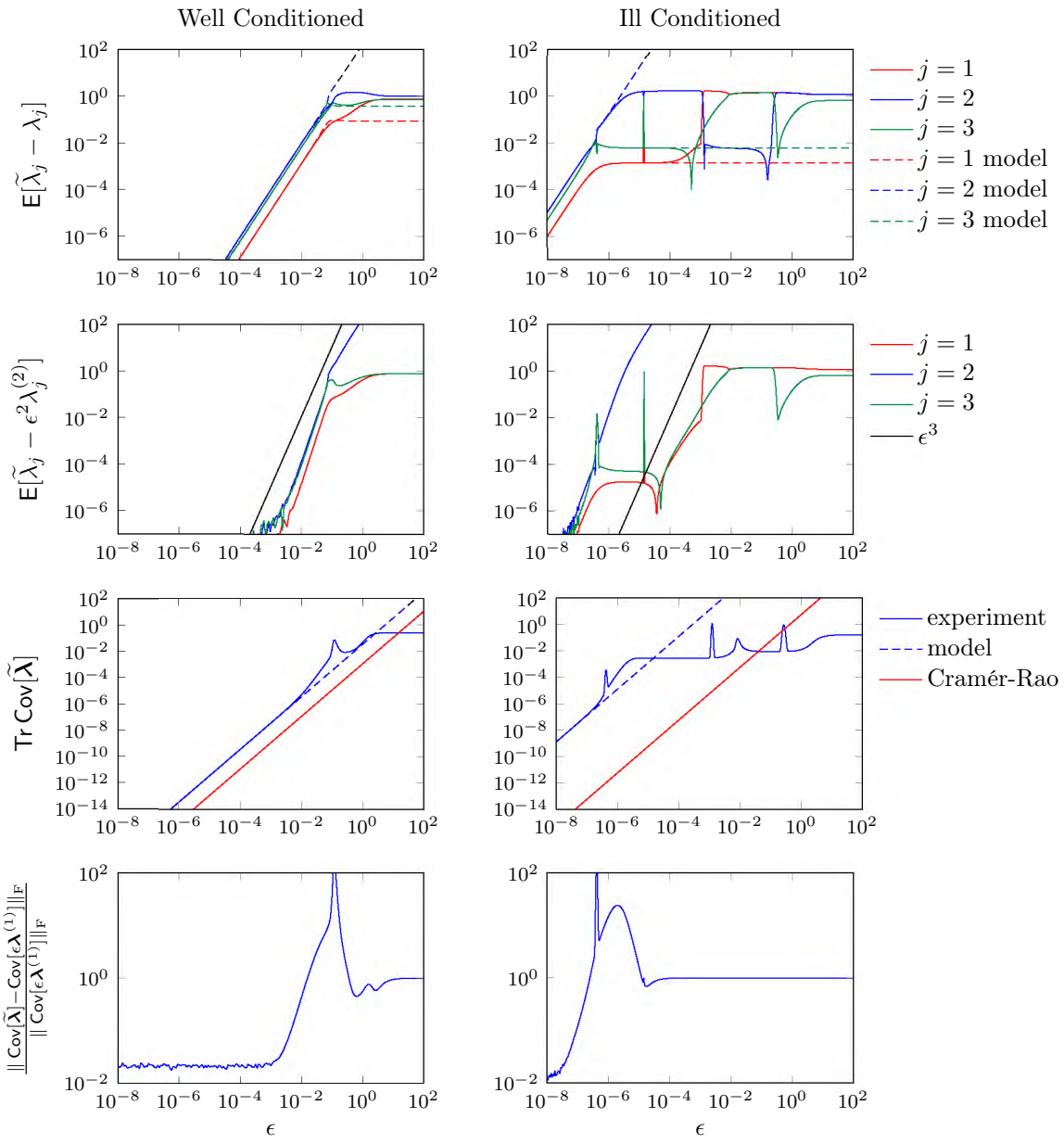
For Prony's Method, which is Prony Least Squares with $m = p$, then $\mathbf{P} = \mathbf{I}$ and $\mathbf{w} = p\widehat{\boldsymbol{\alpha}}$, hence

$$\mathsf{E}[\boldsymbol{\alpha}^{(2)}] = \mathbf{0} \quad \text{if} \quad m = p. \tag{2.30}$$

As Figure 2.2 illustrates, this bias has a dramatic effect on the roots $\widetilde{\boldsymbol{\lambda}}$. In the well-conditioned example, perturbations of size $\epsilon = 10^{-1}$ lead to perturbations of roots of order 1. The ill-conditioned example is even worse; perturbations of size $\epsilon = 10^{-5}$ move one the recovered roots $\widetilde{\boldsymbol{\lambda}}$ to the other size of the complex plane from its true location – an error of approximately 2.

2.4 Maximum Likelihood Prony Methods

Prony Least Squares fails to achieve the Cramér-Rao bound since (2.8) estimates $\widetilde{\boldsymbol{\alpha}}$ in the wrong norm. By changing the norm in (2.8), we can modify Prony Least Squares



$\log(\lambda_1) = \omega_1 = 1i - 0.01$ $\quad \log(\lambda_1) = \omega_1 = 1i - 0.01$

$\log(\lambda_2) = \omega_2 = 2i - 0.03$ $\quad \log(\lambda_2) = \omega_2 = 1i - 0.03$

$\log(\lambda_3) = \omega_3 = 3i - 0.05$ $\quad \log(\lambda_3) = \omega_3 = 1i - 0.05$

Figure 2.2 : Perfomance of asymptotic estimates for Prony's Method. Apperent discontinuities occur where the marriage of $\widetilde{\boldsymbol{\lambda}}$ to $\boldsymbol{\lambda}$ pushes one index far away. The first row of plots confirms our asymptotic bias estimates and the second row shows the next term is $\mathcal{O}(\epsilon^3)$. Likewise the third and four rows confirm the same for the covariance estimates. Here $n = 256$.

to provide a maximum likelihood estimate of $\widetilde{\boldsymbol{\alpha}}$. Suppose instead of solving (2.8) in the ℓ_2-norm, we do so in the yet unspecified $\boldsymbol{\Gamma}$-norm

$$\min_{\boldsymbol{\alpha}} \|\widetilde{\mathbf{H}}\boldsymbol{\alpha} + \widetilde{\mathbf{h}}\|_{\boldsymbol{\Gamma}}^2 = \min_{\boldsymbol{\alpha}} \|\boldsymbol{\Gamma}^{-1/2}(\widetilde{\mathbf{H}}\boldsymbol{\alpha} + \mathbf{h})\|_2^2.$$

Expanding the error following (2.22),

$$\boldsymbol{\Gamma}^{-1/2}(\widetilde{\mathbf{H}}\boldsymbol{\alpha} + \widetilde{\mathbf{h}}) = \boldsymbol{\Gamma}^{-1/2}(\mathbf{H}\boldsymbol{\alpha} + \mathbf{h}) + \epsilon\boldsymbol{\Gamma}^{-1/2}\mathbf{T}(\boldsymbol{\alpha})\mathbf{g}.$$

If $\boldsymbol{\alpha}$ is a maximum likelihood estimate, then the residual $\boldsymbol{\Gamma}^{-1/2}\mathbf{T}(\boldsymbol{\alpha})\mathbf{g}$ should be normally distributed with zero mean and unit covariance $\mathbf{I}$; i.e.,

$$\mathbf{I} = \mathsf{Cov}[\boldsymbol{\Gamma}^{-1/2}\mathbf{T}(\boldsymbol{\alpha})\mathbf{g}] = \boldsymbol{\Gamma}^{-1/2}\mathbf{T}(\boldsymbol{\alpha})\boldsymbol{\Sigma}\mathbf{T}(\boldsymbol{\alpha})^*\boldsymbol{\Gamma}^{-1/2}. \tag{2.31}$$

This constraint is satisfied if we choose

$$\boldsymbol{\Gamma}(\boldsymbol{\alpha}) := \mathbf{T}(\boldsymbol{\alpha})\boldsymbol{\Sigma}\mathbf{T}(\boldsymbol{\alpha})^*. \tag{2.32}$$

Thus, the maximum likelihood estimate $\widetilde{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ is

$$\widetilde{\boldsymbol{\alpha}} = \operatorname*{argmin}_{\boldsymbol{\alpha}} \|\widetilde{\mathbf{H}}\boldsymbol{\alpha} + \widetilde{\mathbf{h}}\|_{\boldsymbol{\Gamma}(\boldsymbol{\alpha})}^2 = \operatorname*{argmin}_{\boldsymbol{\alpha}} \|\boldsymbol{\Gamma}(\boldsymbol{\alpha})^{-1/2}(\widetilde{\mathbf{H}}\boldsymbol{\alpha} + \widetilde{\mathbf{h}})\|_2^2. \tag{2.33}$$

Since $\widetilde{\boldsymbol{\alpha}}$ is a maximum likelihood estimate, then $\widetilde{\boldsymbol{\lambda}}$ is also a maximum likelihood estimate by the Principle of Invariance [167]. There are two approaches to solve (2.33): either using a gradient decent method or a fixed point iteration where, given $\boldsymbol{\alpha}_k$, the next iterate is

$$\boldsymbol{\alpha}_{k+1} = -(\boldsymbol{\Gamma}(\boldsymbol{\alpha}_k)^{-1/2}\widetilde{\mathbf{H}})^+(\boldsymbol{\Gamma}(\boldsymbol{\alpha}_k)^{-1/2}\widetilde{\mathbf{h}}). \tag{2.34}$$

We call this second approach the *Maximum Likelihood Prony Method*, given in Algorithm 2.7.

The Maximum Likelihood Prony Method is a protype for reformulations of Prony's Method yielding maximum likelihood estimates of $\boldsymbol{\alpha}$. Similar approaches were developed independently three times: by Householder in 1949 [69], Osborne in 1975 [116],

Algorithm 2.7: Maximum Likelihood Prony Method

Input : Measurements $y_0, y_1, \ldots, y_{n-1} \in \mathbb{C}$, model order p, covariance $\mathbf{\Sigma}$, and convergence tolerance τ

Output: Exponential parameters $\lambda_0, \lambda_1, \ldots, \lambda_{p-1}$.

Form $\mathbf{H} \in \mathbb{C}^{(n-p)\times p}$ with $[\mathbf{H}]_{j,k} = y_{j+k}$ and $\mathbf{h} \in \mathbb{C}^{n-p}$ with $[\mathbf{h}]_j = y_{j+p}$;
Compute $\boldsymbol{\alpha} = -\mathbf{H}^{+}\mathbf{h}$;
Compute the Cholesky decomposition $\mathbf{L}\mathbf{L}^{*} \leftarrow \mathbf{\Sigma}$;
while $\|\boldsymbol{\alpha}' - \boldsymbol{\alpha}\| > \tau$ **do**
    $\boldsymbol{\alpha}' \leftarrow \boldsymbol{\alpha}$;
    Compute the $\mathbf{R}$ in the QR-decomposition $\mathbf{Q}\mathbf{R} \leftarrow \mathbf{L}^{*}\mathbf{T}(\boldsymbol{\alpha})^{*}$;
    $\boldsymbol{\alpha} \leftarrow -(\mathbf{R}^{-*}\mathbf{H})^{-1}(\mathbf{R}^{-*}\mathbf{h})$;
Find roots λ_k of $q(t) = t^p + \alpha_{p-1}t^{p-1} + \alpha_{p-2}t^{p-2} + \cdots + \alpha_1 t + \alpha_0$.

and Bresler and Macovski in 1986 [23]. Householder's derivation yields an iteration similar to (2.34) using the normal equations; Osborne's solves the equivalent of (2.33) in the Nullspace Method setting using gradient decent yielding a nonlinear eigenvalue problem iteration; Bresler and Marcovski's *Iterative Quadratic Maximum Likelihood* also solves the equivalent of (2.33) in the Nullspace Method setting, but uses a fixed point iteration similar to (2.34). Despite these three independent discoveries, these corrections to Prony's Method never supplanted methods that work in the wrong norm.

In this section, we derive both Householder's Method and Osborne's Method (Bresler and Macovski's Method is a special case). These methods are subject to the same instability problems as other variants of Prony's Method as they still compute the roots of q; Figure 2.3 provides an example of this behavior. When noise is small, each method obtains Cramér-Rao lower bound the covariance of $\widetilde{\boldsymbol{\lambda}}$, but when noise grows, these methods suddenly loose accuracy at a critical point when the perturbations are sufficient to prevent convergence.

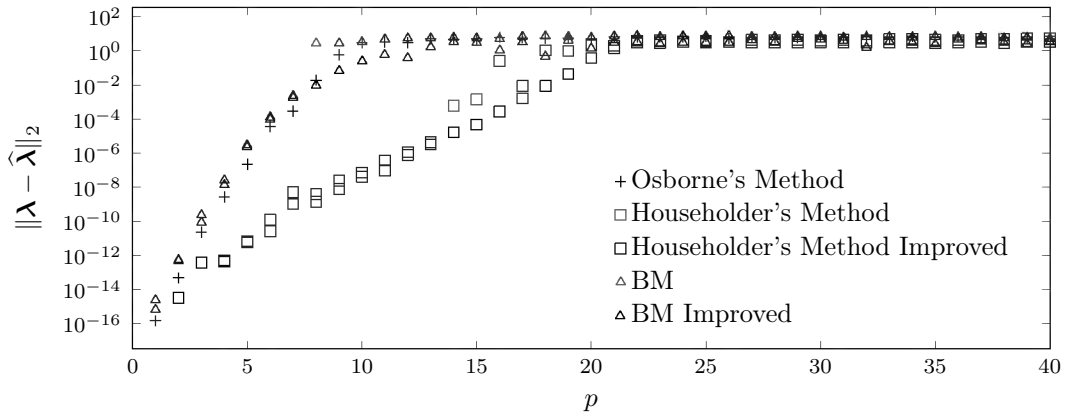


Figure 2.3 : An illustration of the sensitivity of Maximum Likelihood Prony Methods to round off errors. This uses the same parameters as in Figure 2.1.

2.4.1 Householder's Method

The basic approach of Householder's Method [69] is to minimize the mismatch between measurements $\widetilde{\mathbf{y}}$ and the model $\mathbf{y}(\boldsymbol{\alpha})$ subject to an equality constraint forcing the residual $\mathbf{r}(\boldsymbol{\alpha}) = \widetilde{\mathbf{y}} - \mathbf{y}(\boldsymbol{\alpha})$ to satisfy the autoregressive model (2.1) exactly. The equality constraint is enforced by requiring $\mathbf{T}(\boldsymbol{\alpha})\mathbf{r}(\boldsymbol{\alpha}) = \mathbf{0}$; hence we wish to solve

$$\min_{\boldsymbol{\alpha}} \|\mathbf{r}(\boldsymbol{\alpha})\|_{\boldsymbol{\Sigma}}^2 \quad \text{such that} \quad \mathbf{T}(\boldsymbol{\alpha})\mathbf{r}(\boldsymbol{\alpha}) = \mathbf{0}. \tag{2.35}$$

To solve (2.35), we introduce Lagrange multipliers $\boldsymbol{\ell} \in \mathbb{C}^{n-p}$ and solve the augmented problem

$$\min_{\boldsymbol{\alpha},\boldsymbol{\ell}} \|\mathbf{r}(\boldsymbol{\alpha})\|_{\boldsymbol{\Sigma}}^2 + \boldsymbol{\ell}^* \mathbf{T}(\boldsymbol{\alpha})\mathbf{r}(\boldsymbol{\alpha}). \tag{2.36}$$

Definite $\phi(\boldsymbol{\alpha}, \boldsymbol{\ell})$ as the quantity minimized above. Although we can directly solve (2.36), Householder uses several clever properties to yield a fixed point iteration for $\boldsymbol{\alpha}$ directly.

At the minima, the derivative of ϕ with the respect to the residual should be zero,

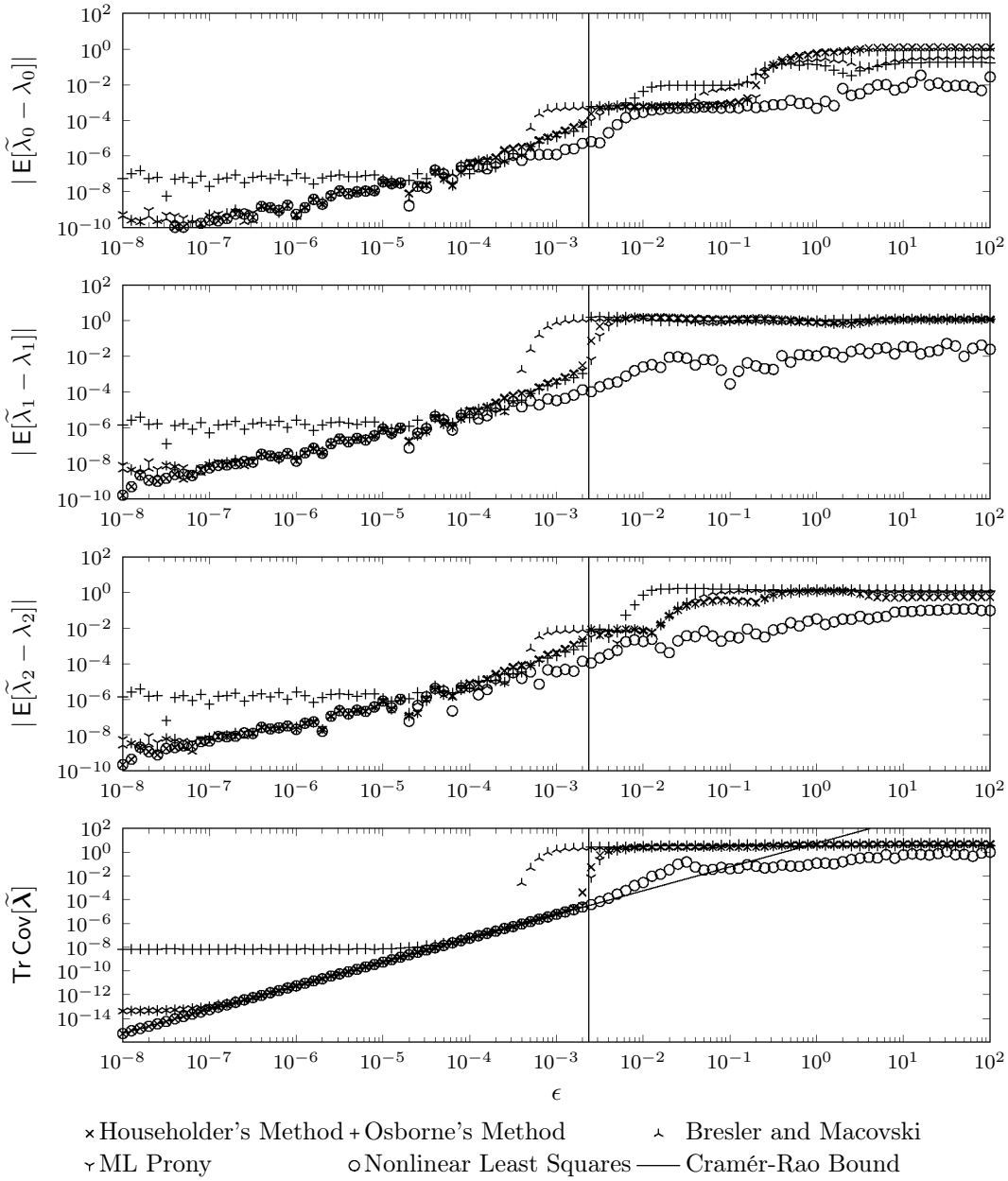


Figure 2.4 : Comparison of Maximum Likelihood Prony Methods applied to the ill-conditioned example from Figure 2.2. Prior to the break down of ML-Prony Methods around $\epsilon = 2 \times 10^{-3}$, the bias in $\widetilde{\boldsymbol{\lambda}}$ grows sublinearly. The gray vertical line shows a heuristic for estimating the breakdown of Prony-Type methods introduced in Section 2.5: the point where $\mathsf{E}[\|\mathbf{E}\|_2]$ exceeds the smallest singular value of $\mathbf{H}$ 50% of the time.

hence[4],

$$\mathbf{0} = \frac{\partial \phi}{\partial \mathbf{r}}(\boldsymbol{\alpha}, \boldsymbol{\ell}) = \mathbf{r}^* \boldsymbol{\Sigma}^{-1} + \boldsymbol{\ell}^* \mathbf{T}(\boldsymbol{\alpha}). \tag{2.37}$$

This corresponds to [69, eq. (16a)]. Next we consider a perturbation expansion of $\mathbf{T}(\widetilde{\boldsymbol{\alpha}})\mathbf{y}(\widetilde{\boldsymbol{\alpha}})$ about $\boldsymbol{\alpha}$

$$\begin{aligned} \mathbf{T}(\widetilde{\boldsymbol{\alpha}})\mathbf{y}(\widetilde{\boldsymbol{\alpha}}) =& \mathbf{T}(\boldsymbol{\alpha})\mathbf{y}(\boldsymbol{\alpha}) + \mathbf{T}(\boldsymbol{\alpha})\left(\mathbf{y}(\widetilde{\boldsymbol{\alpha}}) - \mathbf{y}(\boldsymbol{\alpha})\right) \\ &+ \sum_{j=0}^{p-1} \widehat{\mathbf{T}}(\mathbf{e}_j)\mathbf{y}(\boldsymbol{\alpha})(\widetilde{\alpha}_j - \alpha_j) + \mathcal{O}(\|\widetilde{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_j\|_2^2) \end{aligned} \tag{2.38}$$

where $\widehat{\mathbf{T}}(\mathbf{e}_j) \in \mathbb{C}^{(n-p)\times n}$ is zero except for ones along the jth super diagonal. We simplify this expression noting the sum gives $\mathbf{H}(\boldsymbol{\alpha})(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$, where $\mathbf{H}(\boldsymbol{\alpha})$ is the Hankel matrix constructed from $\mathbf{y}(\boldsymbol{\alpha})$ instead of $\widetilde{\mathbf{y}}$. Further, $\mathbf{T}(\boldsymbol{\alpha})\mathbf{y}(\boldsymbol{\alpha}) = \mathbf{0}$ since $\mathbf{y}(\boldsymbol{\alpha})$ satisfies the equality constraints exactly. To enforce the equality constraint, we seek $\widetilde{\boldsymbol{\alpha}}$ such that $\mathbf{T}(\boldsymbol{\alpha})\widetilde{\mathbf{y}} = \mathbf{T}(\boldsymbol{\alpha})\mathbf{y}(\boldsymbol{\alpha})$. Combining this with the above equation and making the assumption that the next step provides the exact solution (i.e., $\widetilde{\mathbf{y}} = \mathbf{y}(\boldsymbol{\alpha}^{(1)})$) so $\mathbf{r}(\boldsymbol{\alpha}) = \mathbf{y}(\boldsymbol{\alpha}^{(1)}) - \mathbf{y}(\boldsymbol{\alpha}^{(0)})$, then

$$\mathbf{T}(\widetilde{\boldsymbol{\alpha}})\widetilde{\mathbf{y}} = \mathbf{T}(\boldsymbol{\alpha})\mathbf{r}(\boldsymbol{\alpha}) + \mathbf{H}(\boldsymbol{\alpha})(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}). \tag{2.39}$$

This equation corresponds to [69, eq. (16b)]. Using (2.37), we note $\mathbf{r} = \boldsymbol{\Sigma}\mathbf{T}(\boldsymbol{\alpha})^*\boldsymbol{\ell}$ and hence

$$\mathbf{T}(\widetilde{\boldsymbol{\alpha}})\widetilde{\mathbf{y}} = \mathbf{T}(\boldsymbol{\alpha})\boldsymbol{\Sigma}\mathbf{T}(\boldsymbol{\alpha})^*\boldsymbol{\ell} + \mathbf{H}(\boldsymbol{\alpha})(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}).$$

Rearranging and inverting gives the Lagrange multipliers

$$\boldsymbol{\ell} = \boldsymbol{\Gamma}(\boldsymbol{\alpha})^{-1}\left[\mathbf{T}(\widetilde{\boldsymbol{\alpha}})\widetilde{\mathbf{y}} - \mathbf{H}(\boldsymbol{\alpha})(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha})\right] \tag{2.40}$$

where $\boldsymbol{\Gamma}(\boldsymbol{\alpha}) = \mathbf{T}(\boldsymbol{\alpha})\boldsymbol{\Sigma}\mathbf{T}(\boldsymbol{\alpha})^*$. The Lagrange multipliers must satisfy the first order necessary conditions,

$$\mathbf{0} = \frac{\partial \phi}{\partial \boldsymbol{\alpha}} = \frac{\partial \phi}{\partial \mathbf{r}}\frac{\partial \mathbf{r}}{\partial \boldsymbol{\alpha}} = (\mathbf{r}^*\boldsymbol{\Sigma}^{-1} + \boldsymbol{\ell}^*\mathbf{T}(\boldsymbol{\alpha}))\frac{\partial \mathbf{r}}{\partial \boldsymbol{\alpha}}.$$

[4]See the [137, App. A2.2] for information on complex derivatives.

In the limit of a small residual, the first term vanishes and we can compute the second term implicitly from (2.38),

$$\mathbf{T}(\boldsymbol{\alpha})\frac{\partial \phi}{\partial \boldsymbol{\alpha}} = \boldsymbol{\ell}^* \sum_{j=0}^{p-1} \widehat{\mathbf{T}}(\mathbf{e}_j)\mathbf{y}(\boldsymbol{\alpha}) = \boldsymbol{\ell}^* \mathbf{H}(\boldsymbol{\alpha}). \tag{2.41}$$

Thus we set $\boldsymbol{\ell}^*\mathbf{H}(\boldsymbol{\alpha}) = \mathbf{0}$ and combining with (2.40) yields

$$\mathbf{0} = \mathbf{H}(\boldsymbol{\alpha})^*\boldsymbol{\ell} = \mathbf{H}(\boldsymbol{\alpha})^*\boldsymbol{\Gamma}(\boldsymbol{\alpha})^{-1}\left[\mathbf{T}(\widetilde{\boldsymbol{\alpha}})\widetilde{\mathbf{y}} - \mathbf{H}(\boldsymbol{\alpha})(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha})\right]. \tag{2.42}$$

Then, making the approximation $\mathbf{H}(\boldsymbol{\alpha}) \to \widetilde{\mathbf{H}}$ and $\mathbf{T}(\widetilde{\boldsymbol{\alpha}}) \to \mathbf{T}(\boldsymbol{\alpha})$, yields a fixed point iteration updating $\boldsymbol{\alpha} \to \widetilde{\boldsymbol{\alpha}}$:

$$\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + (\widetilde{\mathbf{H}}^*\boldsymbol{\Gamma}(\boldsymbol{\alpha})^{-1}\widetilde{\mathbf{H}})^{-1}\widetilde{\mathbf{H}}^*\boldsymbol{\Gamma}(\boldsymbol{\alpha})^{-1}\mathbf{T}(\boldsymbol{\alpha})\widetilde{\mathbf{y}}. \tag{2.43}$$

Algorithm 2.8: Householder's Method

Input : Measurements $y_0, y_1, \ldots, y_{2p-1} \in \mathbb{C}$, model order p, noise covariance $\boldsymbol{\Sigma}$, and covergence tolerance τ.

Output: Exponential parameters $\lambda_0, \lambda_1, \ldots, \lambda_{p-1}$.

1 Form $\mathbf{H} \in \mathbb{C}^{(n-p)\times p}$ with $[\mathbf{H}]_{j,k} = y_{j+k}$ and $\mathbf{h} \in \mathbb{C}^{n-p}$ with $[\mathbf{h}]_j = y_{j+p}$;
2 Compute $\boldsymbol{\alpha} = -\mathbf{H}^{-1}\mathbf{h}$;
3 **while** $\|\boldsymbol{\alpha}^{(1)}\| > \tau$ **do**
4 Form $\boldsymbol{\Gamma}(\boldsymbol{\alpha}) = \mathbf{T}(\boldsymbol{\alpha})\boldsymbol{\Sigma}\mathbf{T}(\boldsymbol{\alpha})^*$;
5 Solve $\widetilde{\mathbf{H}}^*\boldsymbol{\Gamma}(\boldsymbol{\alpha})^{-1}\widetilde{\mathbf{H}}\boldsymbol{\alpha}^{(1)} = \widetilde{\mathbf{H}}^*\boldsymbol{\Gamma}(\boldsymbol{\alpha})^{-1}\widetilde{\mathbf{y}}$ for $\boldsymbol{\alpha}^{(1)}$;
6 $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + \boldsymbol{\alpha}^{(1)}$;
7 Find roots λ_k of $q(t) = t^p + \alpha_{p-1}t^{p-1} + \alpha_{p-2}t^{p-2} + \cdots + \alpha_1 t + \alpha_0$.

Householder used Prony Least Squares to provide the initial estimate of $\boldsymbol{\alpha}$ for his algorithm, given in Algorithm 2.8. Although this fixed point iteration converges rapidly and yields maximum likelihood estimates of $\boldsymbol{\alpha}$ for moderate amounts of noise, this method went unnoticed except for a few authors, including Osborne [116, 117] who failed to notice the connections between his method and Householder's Method.

2.4.2 Osborne's Method

Osborne's Method similarly fixes the norm in (2.8), but unlike Householder's Method, begins from the Nullspace Method rather than Prony Least Squares. Recall the Nullspace Method solves

$$\min_{\substack{\underline{\boldsymbol{\alpha}}\in\mathbb{C}^{p+1}\\ \|\underline{\boldsymbol{\alpha}}\|_2=1}} \|\widetilde{\underline{\mathbf{H}}}\,\underline{\boldsymbol{\alpha}}\|_2 \tag{2.44}$$

where $\widetilde{\underline{\mathbf{H}}}$ are noisy measurements $\widetilde{\underline{\mathbf{H}}} = [\widetilde{\mathbf{H}}, \widetilde{\mathbf{h}}]$. Separating noise $\mathbf{g}$, following (2.22),

$$\widetilde{\underline{\mathbf{H}}}\,\underline{\boldsymbol{\alpha}} = \underline{\mathbf{H}}\,\underline{\boldsymbol{\alpha}} + \epsilon\widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}})\mathbf{g} \tag{2.45}$$

where $\widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}})$ is a modification of $\mathbf{T}(\boldsymbol{\alpha})$ from (2.21),

$$\widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}}) := \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_{p-1} & \alpha_p & & \\ & \alpha_0 & \alpha_1 & \cdots & \alpha_{p-1} & \alpha_p & \\ & & \ddots & \ddots & & \ddots & \ddots \\ & & & \alpha_0 & \alpha_1 & \cdots & \alpha_{p-1} & \alpha_p \end{bmatrix}. \tag{2.46}$$

This implies that the error $\epsilon\widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}})\mathbf{g}$ is normally distributed with zero mean and covariance $\widehat{\boldsymbol{\Gamma}}(\underline{\boldsymbol{\alpha}}) = \widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}})\boldsymbol{\Sigma}\widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}})^*$ (c.f., Section 2.3 for Prony Least Squares). To minize errors with covariance $\widehat{\boldsymbol{\Gamma}}(\underline{\boldsymbol{\alpha}})$, we work in the $\widehat{\boldsymbol{\Gamma}}(\underline{\boldsymbol{\alpha}})$ norm minimizing,

$$\|\widetilde{\underline{\mathbf{H}}}\boldsymbol{\alpha}\|_{\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\alpha})} = \|\widehat{\boldsymbol{\Gamma}}(\underline{\boldsymbol{\alpha}})^{-1/2}\widetilde{\underline{\mathbf{H}}}\boldsymbol{\alpha}\|_2 = \|\widehat{\boldsymbol{\Gamma}}(\underline{\boldsymbol{\alpha}})^{-1/2}\underline{\mathbf{H}}\boldsymbol{\alpha} + \widehat{\boldsymbol{\Gamma}}(\underline{\boldsymbol{\alpha}})^{-1/2}\widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}})\mathbf{g}\|_2. \tag{2.47}$$

Hence, we now wish to solve $\min_{\underline{\boldsymbol{\alpha}}} \phi(\underline{\boldsymbol{\alpha}})$ where

$$\phi(\underline{\boldsymbol{\alpha}}) := \underline{\boldsymbol{\alpha}}^*\,\widetilde{\underline{\mathbf{H}}}^*\widehat{\boldsymbol{\Gamma}}(\underline{\boldsymbol{\alpha}})^{-1}\widetilde{\underline{\mathbf{H}}}\,\underline{\boldsymbol{\alpha}} = \|\widetilde{\underline{\mathbf{H}}}\,\underline{\boldsymbol{\alpha}}\|^2_{\widehat{\boldsymbol{\Gamma}}(\underline{\boldsymbol{\alpha}})}. \tag{2.48}$$

This is a nonlinear eigenvalue problem for the smallest eigenvector of $\widetilde{\underline{\mathbf{H}}}^*\boldsymbol{\Gamma}(\underline{\boldsymbol{\alpha}})^{-1}\widetilde{\underline{\mathbf{H}}}$. A simple iteration to solve (2.48) was suggested by Bresler and Macovski [23]: given $\underline{\boldsymbol{\alpha}}_k$ find the next iterate $\underline{\boldsymbol{\alpha}}_{k+1}$ as the eigenvector of smallest eigenvalue of

$$\mu\underline{\boldsymbol{\alpha}}_{k+1} = \widetilde{\underline{\mathbf{H}}}^*\boldsymbol{\Gamma}(\underline{\boldsymbol{\alpha}}_k)^{-1}\widetilde{\underline{\mathbf{H}}}\boldsymbol{\alpha}_{k+1}. \tag{2.49}$$

Bresler and Macovski call this the *Iterative Quadratic Maximum Likelihood* method; see Algorithm 2.9. Experience shows this iteration converges rapidly.

Osborne proposed an alternative to this iteration that instead chooses $\underline{\boldsymbol{\alpha}}_{k+1}$ using gradient descent. This implies finding a stationary point of ϕ where $\partial\phi/\partial\underline{\boldsymbol{\alpha}} = \mathbf{0}$. Invoking Wirtinger Calculus,

$$\frac{\partial\phi}{\partial\underline{\boldsymbol{\alpha}}}(\underline{\boldsymbol{\alpha}}) = \underline{\boldsymbol{\alpha}}^*\widetilde{\underline{\mathbf{H}}}^*\boldsymbol{\Gamma}(\underline{\boldsymbol{\alpha}})^{-1}\widetilde{\underline{\mathbf{H}}} + \sum_{k=0}^{p}\left(\underline{\boldsymbol{\alpha}}^*\widetilde{\underline{\mathbf{H}}}^*\boldsymbol{\Gamma}(\underline{\boldsymbol{\alpha}})^{-1}\frac{\partial\boldsymbol{\Gamma}(\underline{\boldsymbol{\alpha}})}{\partial\alpha_k}\boldsymbol{\Gamma}(\underline{\boldsymbol{\alpha}})^{-1}\widetilde{\underline{\mathbf{H}}}\underline{\boldsymbol{\alpha}}\right)\mathbf{e}_k^* \quad \text{and,}$$

$$\frac{\partial\boldsymbol{\Gamma}(\underline{\boldsymbol{\alpha}})}{\partial\alpha_k} = \frac{\partial}{\partial\alpha_k}\widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}})\boldsymbol{\Sigma}\widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}})^* = \widehat{\mathbf{T}}(\mathbf{e}_k)\boldsymbol{\Sigma}\widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}})^*.$$

To simplify these expressions, we define

$$\mathbf{z}(\underline{\boldsymbol{\alpha}}) := \widehat{\boldsymbol{\Gamma}}(\underline{\boldsymbol{\alpha}})^{-1}\widetilde{\underline{\mathbf{H}}}\underline{\boldsymbol{\alpha}} \quad \text{and} \quad \mathbf{Z}(\underline{\boldsymbol{\alpha}}) := \sum_{k=0}^{p}\widehat{\mathbf{T}}(\mathbf{e}_k)^*\overline{\mathbf{v}(\underline{\boldsymbol{\alpha}})}\mathbf{e}_k^*. \tag{2.50}$$

Then we write the second term as

$$\begin{aligned}\sum_{k=0}^{p}\mathbf{z}(\underline{\boldsymbol{\alpha}})^*\widehat{\mathbf{T}}(\mathbf{e}_k)\boldsymbol{\Sigma}\widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}})^*\mathbf{z}(\underline{\boldsymbol{\alpha}})\mathbf{e}_k^* &= \sum_{k=0}^{p}\mathbf{z}(\underline{\boldsymbol{\alpha}})^\top\overline{\widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}})}\boldsymbol{\Sigma}^\top\widehat{\mathbf{T}}(\mathbf{e}_k)^\top\overline{\mathbf{z}(\underline{\boldsymbol{\alpha}})}\mathbf{e}_k^* \\ &= \mathbf{z}(\underline{\boldsymbol{\alpha}})^\top\overline{\widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}})}\boldsymbol{\Sigma}^\top\mathbf{Z}(\underline{\boldsymbol{\alpha}}).\end{aligned}$$

We extract an $\underline{\boldsymbol{\alpha}}$ on the left, noting:

$$\begin{aligned}\left(\mathbf{z}(\underline{\boldsymbol{\alpha}})^\top\overline{\widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}})}\right)^* &= \widehat{\mathbf{T}}(\boldsymbol{\alpha})^\top\overline{\mathbf{z}(\underline{\boldsymbol{\alpha}})} = \sum_{k=0}^{p}\widehat{\mathbf{T}}(\mathbf{e}_k)^\top\overline{\mathbf{z}(\underline{\boldsymbol{\alpha}})}\alpha_k \\ &= \sum_{k=0}^{p}\widehat{\mathbf{T}}(\mathbf{e}_k)^\top\overline{\mathbf{z}(\underline{\boldsymbol{\alpha}})}\mathbf{e}_k\underline{\boldsymbol{\alpha}} = \mathbf{Z}(\underline{\boldsymbol{\alpha}})\underline{\boldsymbol{\alpha}}.\end{aligned}$$

Hence,

$$\frac{\partial\phi}{\partial\underline{\boldsymbol{\alpha}}}(\underline{\boldsymbol{\alpha}}) = \underline{\boldsymbol{\alpha}}^*\widetilde{\underline{\mathbf{H}}}^*\boldsymbol{\Gamma}(\underline{\boldsymbol{\alpha}})^{-1}\widetilde{\underline{\mathbf{H}}} + \underline{\boldsymbol{\alpha}}^*\mathbf{Z}(\underline{\boldsymbol{\alpha}})^*\boldsymbol{\Sigma}^\top\mathbf{Z}(\underline{\boldsymbol{\alpha}}).$$

Imposing the constraint that $\|\underline{\boldsymbol{\alpha}}\| = 1$, adds the Lagrangian multiplier $-\underline{\boldsymbol{\alpha}}^*$ leading to the nonlinear eigenvalue problem (after taking the complex conjugate)

$$\left[\widetilde{\underline{\mathbf{H}}}^*\boldsymbol{\Gamma}(\underline{\boldsymbol{\alpha}}_k)^{-1}\widetilde{\underline{\mathbf{H}}} + \mathbf{Z}(\underline{\boldsymbol{\alpha}}_k)^*\overline{\boldsymbol{\Sigma}}\mathbf{Z}(\underline{\boldsymbol{\alpha}}_k) - \mu\mathbf{I}\right]\underline{\boldsymbol{\alpha}}_{k+1} = \mathbf{0}. \tag{2.51}$$

Algorithm 2.9: Bresler and Macovski

Input : Measurements $y_0, y_1, \ldots, y_{2p-1} \in \mathbb{C}$, model order p, noise covariance $\mathbf{\Sigma}$.

Output: Exponential parameters $\lambda_0, \lambda_1, \ldots, \lambda_{p-1}$.

Form $\underline{\mathbf{H}} \in \mathbb{C}^{(n-p)\times(p+1)}$ with $[\mathbf{H}]_{j,k} = y_{j+k}$;
Compute SVD of $\underline{\mathbf{H}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ where $[\mathbf{\Sigma}]_{j,j} = \sigma_j$ and $\sigma_j \geq \sigma_k$ for all $j > k$;
Set $\underline{\boldsymbol{\alpha}} = [\mathbf{V}]_{\cdot,p+1} \in \mathbb{C}^{p+1}$;
repeat
    $\mathbf{A} \leftarrow \underline{\mathbf{H}}^* \mathbf{\Gamma}(\underline{\boldsymbol{\alpha}})^{-1} \underline{\mathbf{H}}$;
    Compute the smallest eigenvector of $\mathbf{A}$, $\underline{\mathbf{v}}$;
    $\underline{\boldsymbol{\alpha}} \leftarrow \underline{\mathbf{v}}$;
until *convergence*;

Algorithm 2.10: Osborne's Method

Input : Measurements $y_0, y_1, \ldots, y_{n-1} \in \mathbb{C}$, model order p, and convergence tolerance τ; optionally $\boldsymbol{\lambda}$

Output: Exponential parameters $\lambda_0, \lambda_1, \ldots, \lambda_{p-1}$.

Form $\underline{\mathbf{H}} \in \mathbb{C}^{n-p\times p+1}$ with $[\mathbf{H}]_{j,k} = y_{j+k}$;
if $\boldsymbol{\lambda}$ *provided* **then**
    $\sum_{k=0}^{p} \alpha_k t^k = \prod_{j=0}^{p-1}(t - \lambda_j)$;
else
    Set $\underline{\boldsymbol{\alpha}} = [0\ 0\ 0\ \ldots\ 0\ 1]^\top$;
$i \leftarrow 0$;
while $\mu > \tau\|\mathbf{A}\|$ *and* i *less than some maximum number of iterations* **do**
    $\mathbf{L}\mathbf{L}^* \leftarrow \widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}})\mathbf{\Sigma}\widehat{\mathbf{T}}(\underline{\boldsymbol{\alpha}})^*$ Cholesky decomposition;
    $\mathbf{K} \leftarrow \mathbf{L}^{-1}\mathbf{H}$;
    $\mathbf{A} \leftarrow \mathbf{K}^*\mathbf{K}$;
    if $i > 0$ *or inital estimates of* $\boldsymbol{\lambda}$ *provided* **then**
        $\mathbf{z} \leftarrow \mathbf{L}^{-*}(\mathbf{K}\underline{\boldsymbol{\alpha}})$;
        Form $\mathbf{Z}$ according to (2.50);
        $\mathbf{A} \leftarrow \mathbf{A} + \mathbf{V}^*\overline{\mathbf{\Sigma}}\mathbf{V}$;
    Find smallest eigenvalue of $\mathbf{A}$, μ: $\mathbf{A}\underline{\boldsymbol{\alpha}} = \mu\underline{\boldsymbol{\alpha}}$;
    $i \leftarrow i + 1$;
Find roots λ_k of $q(t) = \alpha_p t^p + \alpha_{p-1}t^{p-1} + \alpha_{p-2}t^{p-2} + \ldots + \alpha_1 t + \alpha_0$.

This is a similar eigenvalue problem to (2.49), with a correction to include the variable norm that vanishes near the minima. Successive iterations solving (2.51) as described in Algorithm 2.10 tend to converge rapidly (see [116, §5]). However, the addition of the variable norm term does not improve the resulting estimates of $\underline{\boldsymbol{\alpha}}$ when ϵ is small as Figure 2.4 suggests. Rather, this addition increases sensitivity to rounding errors, preventing estimates of $\underline{\boldsymbol{\alpha}}$ from converging to the same accuracy as methods without this correction (e.g., Householder's Method, Maximum Likelihood Prony).

2.5 Extraneous Exponentials

One approach to fix numerical ill-conditioning of Prony's Method is to include an additional p' extraneous exponentials to improve the recovery of the p exponentials present in the data. Simon was the first to note this in 1970 [139, §3], recommending including an extra $p' = p$ exponentials. However, Simon's observation went unnoticed until reintroduced independently by Kumaresan and Tufts in the electrical engineering literature in 1982 [92], generalized to include an arbitrary number of extraneous exponentials. Although Kumaresan and Tufts hint at why adding extraneous exponentials improves the recovery of the non-extraneous exponentials, Rao (1988) [129, §III] was the first to provide a rigorous explanation: the recovered roots $\widetilde{\boldsymbol{\lambda}}$ are sensitive to perturbations in $\widetilde{\boldsymbol{\alpha}}$ (2.7),

$$\widetilde{\lambda}_j = \lambda_j - \epsilon \frac{\sum_{k=0}^{p-1} \alpha_k^{(1)} \lambda_j^k}{\prod_{j \neq k} (\lambda_j - \lambda_k)} + \mathcal{O}(\epsilon^2). \tag{2.52}$$

The extraneous exponentials tend to be uniformly distributed inside the unit circle (see, e.g., [91, 19]) and hence increase the denominator, making $\widetilde{\boldsymbol{\lambda}}$ more robust to perturbations. In this section we review these results, adding new insight that suggests the extraneous roots are placed near the roots of unity. We also demonstrate methods

like Kung's Method [94], the Matrix Pencil Method [72], and the Hankel Total Least Squares Method [75] benefit by implicitly including these extraneous exponentials but compute only the non-spurious roots.

A prototype for these algorithms that include extraneous exponentials begins with Prony Least Squares for $p + p'$ exponentials:

$$\mathbf{H}\boldsymbol{\alpha} = -\mathbf{h}, \qquad \mathbf{H} \in \mathbb{C}^{m\times(p+p')},\ \boldsymbol{\alpha} \in \mathbb{C}^{p+p'},\ \mathbf{h} \in \mathbb{C}^{m}. \tag{2.53}$$

In the absence of noise $\mathbf{H}$ has rank p, and as $\mathbf{h}$ is in the range of $\mathbf{H}$ (see (2.16)), there is a p' dimensional subspace defining solutions for $\boldsymbol{\alpha}$. One solution can be found by partitioning $\mathbf{H}$ into two blocks: $\mathbf{H}_0 \in \mathbb{C}^{m\times p'}$ and $\mathbf{H}_{p'} \in \mathbb{C}^{m\times p}$,

$$\begin{bmatrix}\mathbf{H}_0 & \mathbf{H}_{p'}\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ \boldsymbol{\alpha}_{p'}\end{bmatrix} = -\mathbf{h}, \tag{2.54}$$

where $\boldsymbol{\alpha}_{p'}$ are coefficients of q from Prony Least Squares using only p exponentials. The remaining solutions add a vector from the nullspace of $\mathbf{H}$; if $\mathbf{H}$ has the SVD

$$\mathbf{H} = \begin{bmatrix}\mathbf{U}_p & \mathbf{U}_{p'}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_p & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_p^*\\ \mathbf{V}_{p'}^*\end{bmatrix}, \tag{2.55}$$

then solution set for $\boldsymbol{\alpha}$ is

$$\boldsymbol{\alpha}(\mathbf{x}) = \begin{bmatrix}\mathbf{0}\\ \boldsymbol{\alpha}_{p'}\end{bmatrix} + \mathbf{V}_{p'}^*\mathbf{x}, \qquad \mathbf{x} \in \mathbb{C}^{p'}. \tag{2.56}$$

Figure 2.5 shows one dimension of this solution space. Finally, this method chooses $\boldsymbol{\alpha}(\mathbf{x})$ with the smallest norm and computes the roots of the degree $p + p'$ polynomial q separating spurious roots from non-spurious roots using a heuristic.

Using minimum norm solution of $\boldsymbol{\alpha} \in \mathbb{C}^{p+p'}$ has two advantages. One benefit is the minimum norm solution corresponds to the pseudoinverse of $\mathbf{H}$; i.e., in the

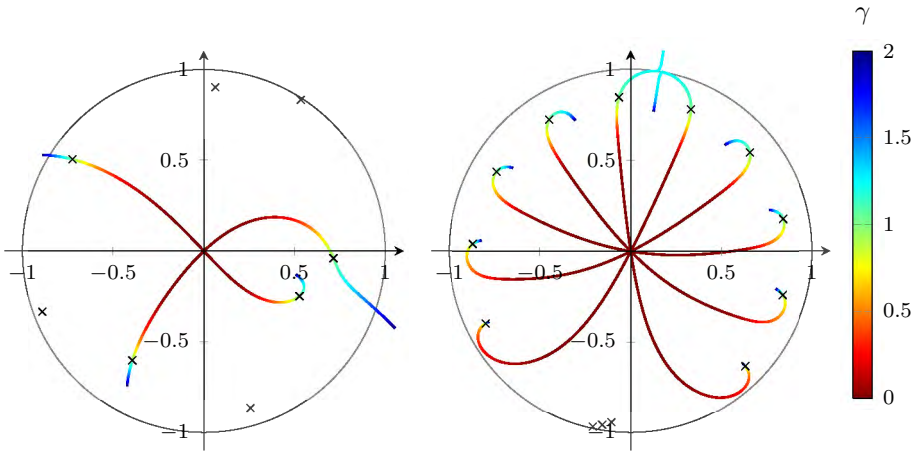


Figure 2.5 : Two examples of extraneous roots along a one dimensional subspace. The least squares solution for $\widetilde{\boldsymbol{\alpha}}$ gives roots marked by '×'. Thei colored curves show the roots resulting from $\boldsymbol{\alpha}(\gamma\mathbf{x})$ (2.54) for $\gamma \in [0, 2]$ and with $\mathbf{x}$ chosen such that $\boldsymbol{\alpha}(\mathbf{x})$ is the least squares solution.

presence of noise

$$\widetilde{\boldsymbol{\alpha}} = \widetilde{\mathbf{V}}_p \widetilde{\boldsymbol{\Sigma}}_p^{-1} \widetilde{\mathbf{U}}_p^* \widetilde{\mathbf{h}}. \tag{2.57}$$

The left and right singular vectors of $\widetilde{\mathbf{H}}$ are relatively stable to small perturbations $\epsilon\mathbf{E} = \widetilde{\mathbf{H}} - \mathbf{H}$. We can bound the diagonal matrix canonical angles $\boldsymbol{\Phi} \in \mathbb{C}^{p\times p}$ between $\mathrm{Ran}\mathbf{U}_p$ and $\mathrm{Ran}\widetilde{\mathbf{U}}_p$ and the canonical angles $\boldsymbol{\Theta} \in \mathbb{C}^{p\times p}$ between $\mathrm{Ran}\mathbf{V}_p$ and $\mathrm{Ran}\widetilde{\mathbf{V}}_p$ using a combination of Wedin's Theorem [143, Thm. V.4.1] and Weyl's Theorem,

$$\|\sin\boldsymbol{\Phi}\|_{\mathrm{F}}^2 + \|\sin\boldsymbol{\Theta}\|_{\mathrm{F}}^2 \leq \frac{2\epsilon^2\|\mathbf{E}\|_{\mathrm{F}}^2}{\sigma_p - \epsilon\|\mathbf{E}\|_2}. \tag{2.58}$$

Using Corollary 3.1 we can estimate the expected value of this bound

$$\mathsf{E}\left[\frac{2\|\mathbf{E}\|_{\mathrm{F}}^2}{\sigma_p - \|\mathbf{E}\|_2}\right] \lesssim \frac{2\epsilon^2(n-p-p')(p+p')}{\sigma_p - \epsilon\sqrt{-2n\log(1-\sqrt[n]{r})}} \quad \text{with probability } r. \tag{2.59}$$

Provided ϵ is small, the left and right singular subspaces remain relatively unperturbed, and hence in the presence of noise, the rank-p pseudoinverse acts to remove perturbations of $\widetilde{\mathbf{h}}$ that are not in the range of $\mathbf{H}$.

The other benefit of the minimum norm choice of $\boldsymbol{\alpha}$ is that it tends to place spurious roots near a complex rotation of $p+p'$ roots of unity and this acts to decrease the sensitivity of $\widetilde{\boldsymbol{\lambda}}$ to perturbations of $\widetilde{\boldsymbol{\alpha}}$. Previous authors have demonstrated that these roots must fall inside the unit circle, e.g., [91] from a filter perspective and [19, Prop. 2.6] from a linear algebra perspective, and in the limit $p' \to \infty$ the spurious roots tend to be uniformly distributed inside the unit circle [119]. We suggest a new intuitive explanation why the spurious roots are placed near the roots of unity (evident in Figure 2.5 and [92, Fig. 1(h)]). If q be the degree $p+p'$ polynomial defined by $\boldsymbol{\alpha}$, then we can express $\boldsymbol{\alpha}$ as the solution to the linear system

$$\begin{bmatrix} 1 & t_0 & t_0^2 & \cdots & t_0^{p+p'} \\ 1 & t_1 & t_1^2 & \cdots & t_1^{p+p'} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_{p+p'} & t_{p+p'}^2 & \cdots & t_{p+p'}^{p+p'} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{p+p'} \end{bmatrix} = \begin{bmatrix} q(t_0) \\ q(t_1) \\ \vdots \\ q(t_{p+p'}) \end{bmatrix} \tag{2.60}$$

where $\alpha_{p+p'} = 1$. By choosing $t_j = e^{i\theta} e^{2\pi i j/(p+p'+1)}$ (rotated $p+p'+1$ roots of unity), the Vandermonde matrix on the left is a multiple of the discrete Fourier transform matrix $\mathbf{F}_{p+p'+1}$. As $\mathbf{F}_{p+p'+1}$ is unitary, then inverting $\mathbf{F}_{p+p'+1}$ gives

$$1 + \|\boldsymbol{\alpha}\|_2 = \sqrt{p+p'+1} \left\| e^{-i\theta} \mathbf{F}_{p+p'+1}^* \begin{bmatrix} q(t_0) \\ q(t_1) \\ \vdots \\ q(t_{p+p'}) \end{bmatrix} \right\|_2 = \sqrt{p+p'+1} \left\| \begin{bmatrix} q(t_0) \\ q(t_1) \\ \vdots \\ q(t_{p+p'}) \end{bmatrix} \right\|_2 . \tag{2.61}$$

Hence minimizing $\boldsymbol{\alpha}$ minimizes the values of q at the rotated $p+p'+1$ roots of unity. In practice, spurious roots tend to align at the p' points furthest away from the non-spurious roots as seen on the right side of Figure 2.5. The rotation $e^{i\theta}$ provides an additional degree of freedom that further explains the placement seen.

One remaining question is how many extraneous exponentials to include? Most authors suggest setting $p + p' = \lfloor n/2 \rfloor$ in an attempt to maximize the smallest singular value of $\widetilde{\mathbf{H}}$, however Hua and Sarkar, recommend $p + p' = \lfloor n/3 \rfloor$ based on a perturbation analysis of the $p = 1$ case [72]. We can extend Hua and Sarkar's analysis to the multiple exponential case by using the perturbation analysis from Section 2.3 where $\mathbf{H}$ has been replaced with its rank-p approximation. Figure 2.6 shows the bias and covariance of the desired exponentials as a function of the number of extraneous exponentials p'. This confirms Hua and Sarkar's observation that the choice $p + p' = \lfloor n/3 \rfloor$ yields minimum covariance estimates for this class of algorithms.

Finally, before reviewing methods that use extraneous exponentials, we note that Rao suggested that $\|\mathbf{\Delta}\|_2 < \sigma_p$ is a sufficient condition for SVD-methods to succeed [129, p. 1031]. This follows from the singular vector perturbation estimate (2.59) and seems to accurately predict where these methods fail in Figure 2.8 (using $p + p' = \lfloor n/2 \rfloor$).

2.5.1 Kumaresan and Tufts Method

The Kumaresan and Tufts Method [92] follows the outline of the method above except that they reverse the order of $\widetilde{\mathbf{y}}$. Then, instead of seeing exponential decay, we see exponential growth. Since the extraneous exponentials are inside the unit circle and correspond to exponential decay, this provides a criteria for separating spurious and non-spurious exponentials. This method is numerically stable (Figure 2.7), and yields reasonable estimates of $\boldsymbol{\lambda}$ (Figure 2.8) however the covariance of these estimates is larger than the Cramér-Rao bound, corresponding to an efficiency of about 70% for this ill-conditioned example (Figure 2.9).

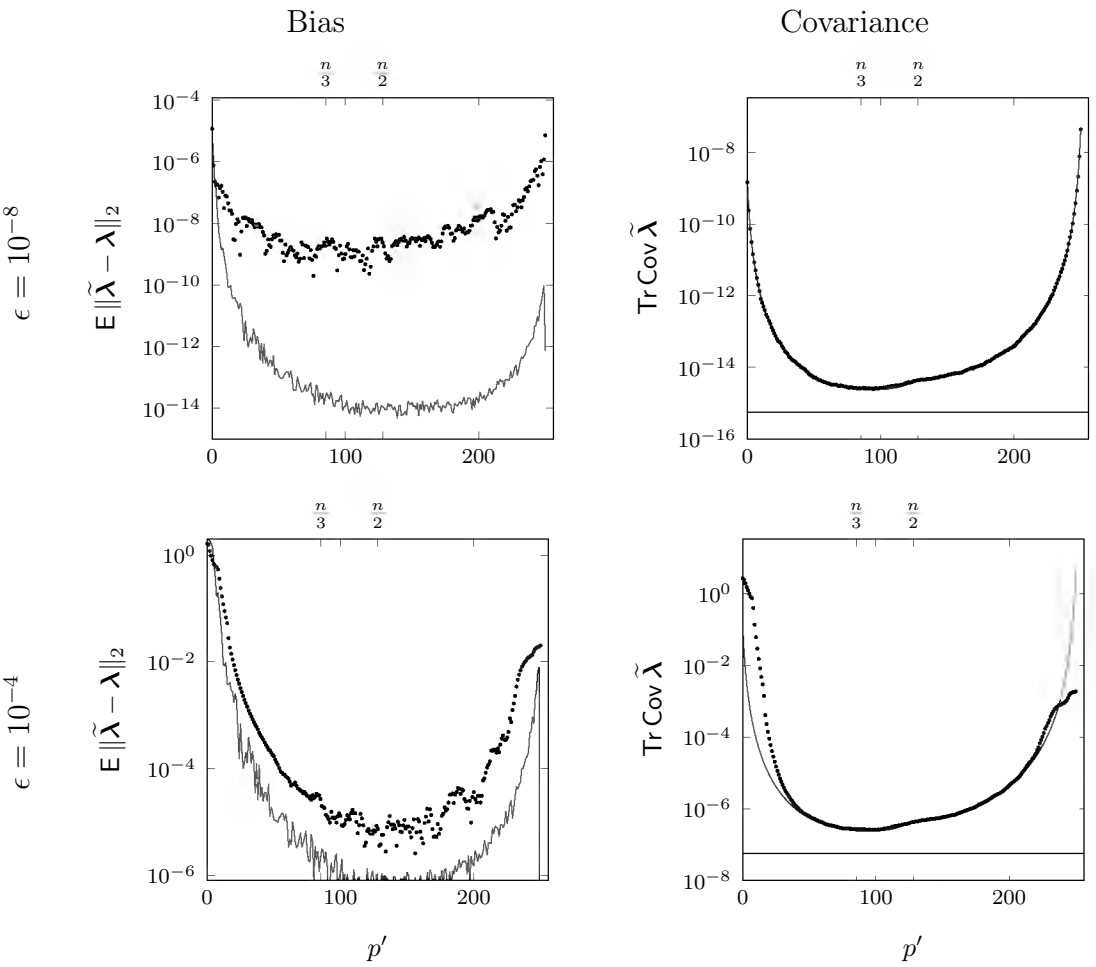


Figure 2.6 : Using the ill-conditioned example from Figure 2.2 with different numbers of extraneous exponentials, we see that the bias and covariance obey the estimates from Section 2.3. The asymptotic estimates of the bias from (2.29) (gray line) matches the bias estimated using a Monte-Carlo approach (black dots). Similarly, the asymptotic covariance (2.27) (gray line) matches the Monte-Carlo covariance (black dots); however, the covariance exceeds the Cramér-Rao bound (black line).

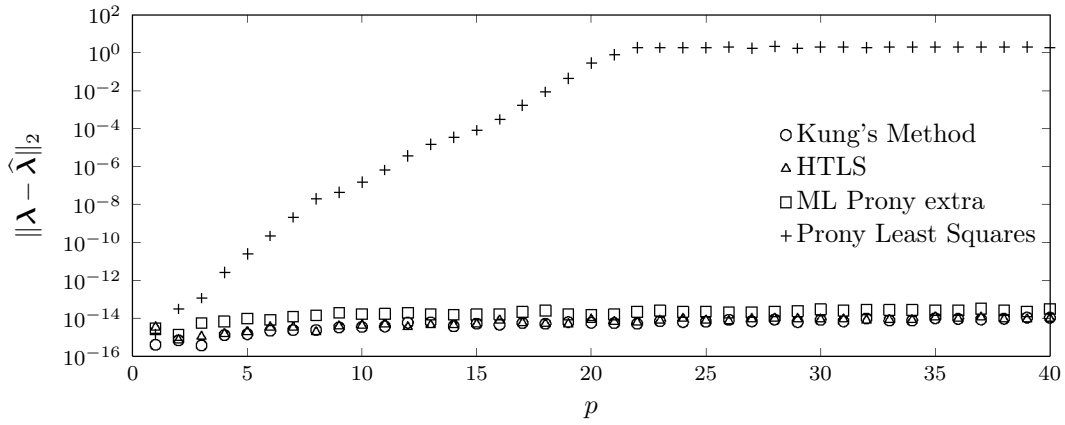


Figure 2.7 : An illustration of the sensity of several Prony-type methods that include extraneous exponentials to round off errors. This uses the same parameters as in Figure 2.1. "ML Prony extra" refers to the Maximum Likelihood Prony Method using extraneous exponentials to improve stability.

2.5.2 Matrix Pencil Method

The Matrix Pencil Method of Hua and Sarkar [72] follows the matrix pencil variant of Prony's Method in Section 2.2.3, however it enlarges the matrices $\mathbf{H}_0$ and $\mathbf{H}_1$ to be in $\mathbb{C}^{(n-\ell)\times\ell}$ rather than $\mathbb{C}^{(n-p)\times p}$. Rather than directly computing the generalized eigenvalues, this method uses a rank-truncated pseudoinverse to construct a matrix with a related spectrum, $\widetilde{\mathbf{C}}$. Starting with the rank-p SVD $\widetilde{\mathbf{H}}_0 = \widetilde{\mathbf{U}}_p\widetilde{\mathbf{\Sigma}}_p\widetilde{\mathbf{V}}_p^*$, this method forms $\widetilde{\mathbf{C}}$ as

$$\widetilde{\mathbf{H}}_0\widetilde{\mathbf{C}} = \widetilde{\mathbf{H}}_1$$
$$\widetilde{\mathbf{C}} := \widetilde{\mathbf{\Sigma}}_p^{+}\widetilde{\mathbf{U}}_p^*\widetilde{\mathbf{H}}_1\widetilde{\mathbf{V}}_p \in \mathbb{C}^{p\times p}.$$

This formulation fixes one of the hardest parts of the Kumaresan and Tufts Method: determining which roots (or in this case eigenvalues) are significant. By using a rank-p pseudoinverse, $\widetilde{\mathbf{C}}$ has only p eigenvalues that we assume correspond to the non-spurious exponentials.

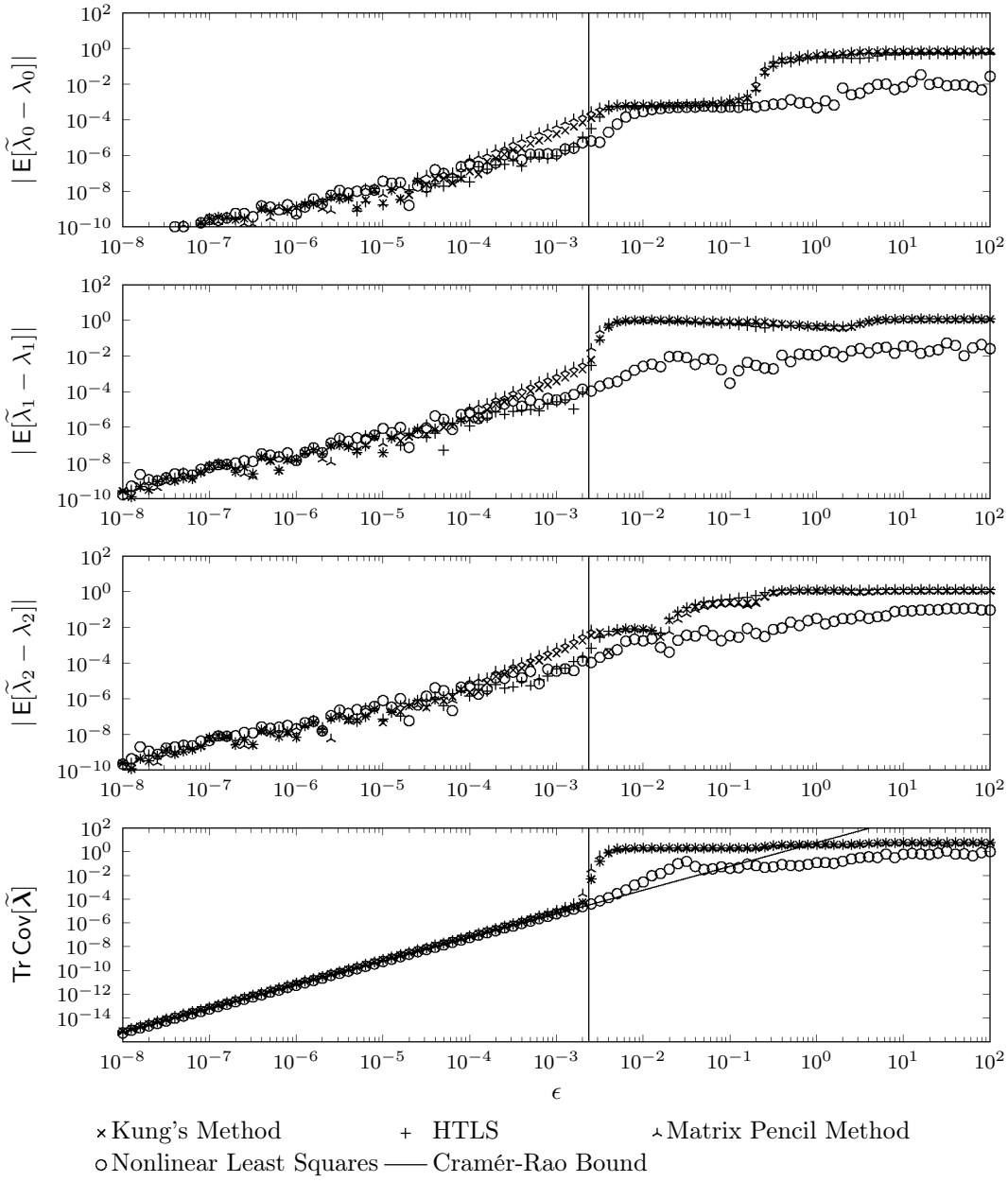


Figure 2.8 : Comparison of Prony Methods using extraneous exponentials applied to an ill-conditioned example from Figure 2.2 (see also Figure 2.4). Prior to the break down of ML-Prony Methods around $\epsilon = 2\times10^{-3}$, the bias in $\widetilde{\boldsymbol{\lambda}}$ grows sublinearly. The gray vertical line shows Rao's sufficient condition for recovery where $\mathsf{E}[\|\mathbf{E}\|_2]$ exceeds the smallest singular value of $\mathbf{H}$ 50% of the time.

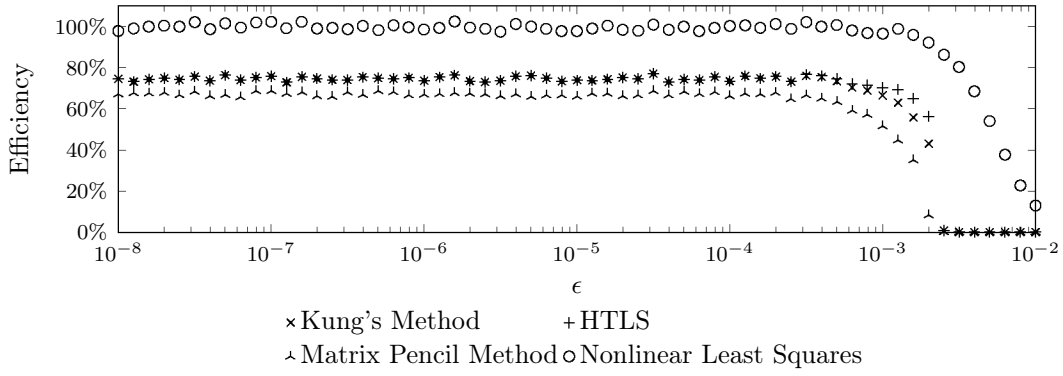


Figure 2.9 : Statistical efficiency of variants of Prony's Method that include extraneous exponentials. This figure uses the same data as Figure 2.8.

In comparison to other methods in this class, the Matrix Pencil Method yields worse estimates of $\widetilde{\boldsymbol{\lambda}}$; see Figures 2.8 and 2.9.

2.5.3 Kung's Method

In the Matrix Pencil Method, the SVD is only applied to one of the matrices, not both. Kung's Method [94] (also known as HSVD [12]), developed prior to the Matrix Pencil Method, corrects this and further removes noise, by estimating $\widetilde{\mathbf{C}}$ using only the left singular values of $\widetilde{\mathbf{H}} \in \mathbb{C}^{(n+1-\ell)\times\ell}$. Combining the Vandermonde decomposition of a Hankel Matrix (2.16) with the matrix pencil formulation of Prony's Method, we have

$$\mathbf{V}_{n-\ell}(\boldsymbol{\lambda})\mathbf{A}\mathbf{V}_{\ell}(\boldsymbol{\lambda})^{\top}\mathbf{C} = \mathbf{V}_{n-\ell}(\boldsymbol{\lambda})\mathbf{A}\boldsymbol{\Lambda}\mathbf{V}_{\ell}(\boldsymbol{\lambda})^{\top}. \tag{2.62}$$

By assumption both $\mathbf{A}$ and $\boldsymbol{\Lambda}$ are full rank, and since $\boldsymbol{\lambda}$ has distinct entries, $\mathbf{V}_{\ell}(\boldsymbol{\lambda})^{\top}$ also has full rank. As such, both are invertible, and instead of finding the eigenvalues of $\mathbf{C}$, we can find the eigenvalues of the similar matrix $\mathbf{D}$, where

$$\mathbf{V}_{n-\ell}(\boldsymbol{\lambda})\mathbf{D} = \mathbf{V}_{n-\ell}(\boldsymbol{\lambda})\boldsymbol{\Lambda}. \tag{2.63}$$

Rather than finding $\mathbf{V}_{n-\ell}(\boldsymbol{\lambda})$ and $\mathbf{V}_{n-\ell}(\boldsymbol{\lambda})\boldsymbol{\Lambda}$, Kung's Method estimates similar matrices using the left singular vectors of $\widetilde{\mathbf{H}} \in \mathbb{C}^{(n+1-\ell)\times\ell}$. If $\widetilde{\mathbf{U}}_p$ are the first p left singular vectors of $\widetilde{\mathbf{H}}$, then

$$\mathbf{V}_{n-\ell}(\boldsymbol{\lambda}) \approx \widetilde{\mathbf{U}}_p^{\uparrow}\mathbf{R}$$
$$\mathbf{V}_{n-\ell}(\boldsymbol{\lambda})\boldsymbol{\Lambda} \approx \widetilde{\mathbf{U}}_p^{\downarrow}\mathbf{R},$$

where $^{\uparrow}$ denotes removing the bottom row of a matrix and $^{\downarrow}$ the top row. This saves us from having to compute two SVDs: one of $\widetilde{\mathbf{H}}_0$ and $\widetilde{\mathbf{H}}_1$. As a result of the additional filtering and since both $\widetilde{\mathbf{U}}_p^{\uparrow}$ and $\widetilde{\mathbf{U}}_p^{\downarrow}$ are well conditioned (nearly unitary), the estimate

$$\widetilde{\mathbf{D}} = (\widetilde{\mathbf{U}}_p^{\uparrow})^{+}\widetilde{\mathbf{U}}_p^{\downarrow} \tag{2.64}$$

approximates $\mathbf{D}$ in a numerically stable fashion. This procedure can be computationally simplified by using the Sherman-Morrison formula to compute the pseudoinverse of $\widetilde{\mathbf{U}}_p^{\uparrow}$. Noting $\widetilde{\mathbf{U}}_p^{\uparrow}$ only subtracts the last row of $\widetilde{\mathbf{U}}_p$, $\widetilde{\mathbf{u}}^*$, we have

$$(\widetilde{\mathbf{U}}_p^{\uparrow})^*\widetilde{\mathbf{U}}_p^{\uparrow} = (\widetilde{\mathbf{U}}_p - \mathbf{e}_{n+1-\ell}\widetilde{\mathbf{u}}^*)^*(\widetilde{\mathbf{U}}_p - \mathbf{e}_{n+1-\ell}\widetilde{\mathbf{u}}^*) = \widetilde{\mathbf{U}}_p^*\widetilde{\mathbf{U}}_p - \widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^* = \mathbf{I} - \widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^*. \tag{2.65}$$

Hence,

$$\widetilde{\mathbf{D}} = \left(\mathbf{I} + \frac{\widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^*}{1 - \widetilde{\mathbf{u}}^*\widetilde{\mathbf{u}}}\right)(\widetilde{\mathbf{U}}_p^{\uparrow})^*\widetilde{\mathbf{U}}_p^{\downarrow}. \tag{2.66}$$

The steps in this method are summarized in Algorithm 2.11.

2.5.4 Hankel Total Least Squares

A further modification of Kung's Method is to construct $\widetilde{\mathbf{D}}$ using the total least squares solution [56, §12.3]of (2.64) instead of the least squares solution. This accounts for both errors in the left and right hand side of (2.64), but does so in the wrong norm. The performance of this method is similar to Kung's Method, but yields less biased

and more efficient estimates of $\widetilde{\boldsymbol{\lambda}}$ in a narrow range from $\epsilon = 10^{-4}$ to $\epsilon = 2 \cdot 10^{-3}$ in Figures 2.8 and 2.9.

2.5.5 Prony Maximum Likelihood with Extraneous Exponentials

In Section 2.4, we saw that every Prony-type method that yields maximum likelihood estimates of $\widetilde{\boldsymbol{\lambda}}$ is numerically unstable (Figure 2.3). Including spurious exponentials and using a rank-p pseudoinverse instead of $\widetilde{\mathbf{H}}^{+}$ can make these methods numerically stable as we illustrate with one variant in Figure 2.7.

2.6 Autocovariance-Based Prony Variants

In this section, we discuss two variants of Prony's Method that use autocorrelation measurements rather than $\widetilde{\mathbf{y}}$. These two variants parallel Prony's Method and the Nullspace Method; generalizations of these approaches that include extraneous exponentials exist, such as MUSIC [10] and ESPRIT [134].

2.6.1 Yule-Walker Method

The *Yule-Walker Method* is Prony's Method where the determination of the autoregressive model by (2.2) has been replaced with the Yule-Walker equations [144, §4.4]. Unlike Prony's Method, the Yule-Walker Method can be biased even in the absence of noise.

The Yule-Walker equations determine the lag coefficients $\boldsymbol{\alpha}$ using measurements of the autocorrelation of $\mathbf{y}$

$$c_k = \mathsf{E}\left[y_{j+k}\overline{y_j}\right]. \tag{2.67}$$

This assumes that $\{y_j\}_{j=0}^{n-1}$ is a stationary sequence (i.e., independent of k). Under

these assumptions, the Yule-Walker equations for the lag coefficents $\boldsymbol{\alpha}$ solve

$$\begin{bmatrix} c_{q-p+1} & c_{q-p+2} & \cdots & c_q \\ c_{q-p+2} & c_{q-p+3} & \cdots & c_{q-p+2} \\ \vdots & \vdots & \ddots & \vdots \\ c_q & c_{q+1} & \cdots & c_{q+p-1} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{p-1} \end{bmatrix} = - \begin{bmatrix} c_{q+1} \\ c_{q+2} \\ \vdots \\ c_{q+p} \end{bmatrix}; \tag{2.68}$$

see, e.g., [104, § 6.5].[5] Calling the matrix on the left $\mathbf{C}_q \in \mathbb{C}^{p\times p}$ and the vector on the right $\mathbf{c}_q$, we have a system that resembles (2.2):

$$\mathbf{C}_q\boldsymbol{\alpha} = \mathbf{c}_q. \tag{2.69}$$

We could immediately solve for $\boldsymbol{\alpha}$ using (2.69); however, in the presence of noise, we must make a special choice of q. Let $\widetilde{y}_j$ be noisy measurements of y_j obeying the autoregressive model

$$\widetilde{y}_k = \sum_{j=1}^{p} -\alpha_{p-j}\widetilde{y}_{k-j} + \sigma g_k, \quad k \geq p, \tag{2.70}$$

where g_k is an independent standard complex Gaussian random variable.[6] Then the corresponding autocorrelation obeys $\widetilde{c}_j = c_j + \delta_{0,j}\sigma^2$. So, if we solved for $\boldsymbol{\alpha}$ using $q = 0$, then the answer would be spurious:

$$\widetilde{\mathbf{C}}_0\boldsymbol{\alpha} = (\mathbf{C}_0 + \sigma^2\mathbf{J})\boldsymbol{\alpha} = \mathbf{c}_0 = \widetilde{\mathbf{c}}_0, \tag{2.71}$$

where $\mathbf{J}$ is the flipped identity matrix. Instead we choose $q = p$ so that c_0 does not appear in $\widetilde{\mathbf{C}}_q$ or $\widetilde{\mathbf{c}}_q$; hence $\widetilde{\mathbf{C}}_p = \mathbf{C}_p$.

---

[5] The method was originally presented in [166, 156].

[6] Everywhere else in this thesis, $\widetilde{y}_j$ refers to *additive* Gaussian noise, $\widetilde{y}_j = y_j + g_j$.

If we have a single sequence of measurements $\widetilde{\mathbf{y}}$, we can estimate c_q by one of two approaches [144, § 2.2.2]:

$$\text{a biased estimate:} \quad c_j \approx \frac{1}{n} \sum_{k=0}^{n-j-1} y_{k+j}\overline{y_j}, \quad \text{or} \tag{2.72}$$

$$\text{an unbiased estimate:} \quad c_j \approx \frac{1}{n-k} \sum_{k=0}^{n-j-1} y_{k+j}\overline{y_j}. \tag{2.73}$$

To recover $\boldsymbol{\alpha}$, we fill $\mathbf{C}_p$ with these estimates of autocorrelations c_q and solve for $\boldsymbol{\alpha}$.

We can arrive at an approximation of $\mathbf{C}_p$ starting from a modified Hildebrand's Method, and this reveals why the Yule-Walker Method is biased. Hildebrand's Method applies no matter where in the sequence of $\{y_j\}_{j=0}^{n-1}$ we start. Starting with the pth entry, we solve

$$\mathbf{H}_p\boldsymbol{\alpha} = \begin{bmatrix} \mathbf{h}_p & \mathbf{h}_{p+1} & \cdots & \mathbf{h}_{2p-1} \end{bmatrix} \boldsymbol{\alpha} = -\mathbf{h}_{2p}, \tag{2.74}$$

where $[\mathbf{H}_p]_{j,k} = y_{j+k+p}$ and $[\mathbf{h}_{2p}]_j = y_{j+2p}$. Next, we multiply on the left by a matrix by $\mathbf{H}_0$ with its columns flipped,

$$\mathbf{T}_0^*\mathbf{H}_p\boldsymbol{\alpha} = \begin{bmatrix} \mathbf{h}_{p-1}^* \\ \mathbf{h}_{p-2}^* \\ \vdots \\ \mathbf{h}_0^* \end{bmatrix} \begin{bmatrix} \mathbf{h}_p & \mathbf{h}_{p+1} & \cdots & \mathbf{h}_{2p-1} \end{bmatrix} \boldsymbol{\alpha} = - \begin{bmatrix} \mathbf{h}_{p-1}^* \\ \mathbf{h}_{p-2}^* \\ \vdots \\ \mathbf{h}_0^* \end{bmatrix} \mathbf{h}_{2p}$$

$$\begin{bmatrix} \mathbf{h}_{p-1}^*\mathbf{h}_p & \mathbf{h}_{p-1}^*\mathbf{h}_{p+1} & \cdots & \mathbf{h}_{p-1}^*\mathbf{h}_{2p-1} \\ \mathbf{h}_{p-2}^*\mathbf{h}_p & \mathbf{h}_{p-2}^*\mathbf{h}_{p+1} & \cdots & \mathbf{h}_{p-1}^*\mathbf{h}_{2p-1} \\ \vdots & & & \vdots \\ \mathbf{h}_0^*\mathbf{h}_p & \mathbf{h}_0^*\mathbf{h}_{p+1} & \cdots & \mathbf{h}_0^*\mathbf{h}_{2p-1} \end{bmatrix} \boldsymbol{\alpha} = - \begin{bmatrix} \mathbf{h}_{p-1}^*\mathbf{h}_{2p} \\ \mathbf{h}_{p-2}^*\mathbf{h}_{2p} \\ \vdots \\ \mathbf{h}_0^*\mathbf{h}_{2p} \end{bmatrix}.$$

Examining the entries of the matrix on the left, we notice that these approximate a scalar multiple of the autocovariance:

$$\mathbf{h}_j^*\mathbf{h}_k = \sum_{\ell=0}^{n-2p-1} \overline{y_{j+\ell}} y_{k+\ell} \approx n c_{k-j}. \tag{2.75}$$

Thus, $\mathbf{T}_0^*\mathbf{H}_p$ is a scaled approximation of $\mathbf{C}_p$. Since the solution of $\mathbf{T}_0^*\mathbf{H}_p\boldsymbol{\alpha} = \mathbf{T}_0^*\mathbf{h}_{2p}$ is identical to (2.74), we conclude that the Yule-Walker method is biased because $n\mathbf{C}_p$ only approximates $\mathbf{T}_0^*\mathbf{H}_p$. In the limit $n \to \infty$, both estimates of the autocovariance coverge to $\frac{1}{n}\mathbf{h}_k^*\mathbf{h}_{k+j}$,

As solving Hildebrand's Method directly does not have the bias present in the Yule-Walker equations, we cannot advise using the Yule-Walker Method unless the autocovariance measurements are directly given.

.

2.6.2 Pisarenko's Method

Pisarenko's Method [125] is the analog of the Nullspace Method using the autocovariance c_j in place of y_j. We can see this by noting the autocorrelation also obeys the same autoregressive model:

$$y_{k+j} = \sum_{\ell=1}^{p} -\alpha_{p-\ell} y_{k+j-\ell}$$

$$c_k = \mathsf{E}[y_{k+j}\overline{y_j}] = \sum_{\ell=1}^{p} -\alpha_{p-\ell}\,\mathsf{E}[y_{k+j-\ell}\overline{y_j}] = \sum_{\ell=1}^{p} -\alpha_{p-\ell} c_{k-\ell}.$$

Hence,

$$\begin{bmatrix} c_{-p} & c_{-p+1} & \cdots & c_{-1} & c_0 \\ c_{-p+1} & c_{-p+2} & \cdots & c_0 & c_1 \\ \vdots & \vdots & & \vdots & \vdots \\ c_{-1} & c_0 & \cdots & c_{p-2} & c_{p-1} \\ c_0 & c_1 & \cdots & c_{p-1} & c_p \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{p-1} \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}. \tag{2.76}$$

If we allow the scaling of the left vector to be free, we have

$$\mathbf{C}_{p+1}\underline{\boldsymbol{\alpha}} = \mathbf{0}. \tag{2.77}$$

Hence in the absence of noise, we compute the nullspace of $\mathbf{C}_{p+1}$ and find the roots of $\underline{q}$, just as with the Nullspace Method.

When there is noise this method requires slight modification. Rather than computing the nullspace of $\mathbf{C}_{p+1}$ we equivalently seek the eigenvector corresponding to the zero eigenvalue. However, the smallest eigenvector of may have changed, since

$$\widetilde{\mathbf{C}}_{p+1}\underline{\boldsymbol{\alpha}} = (\widetilde{\mathbf{C}}_{p+1} + \sigma^2\mathbf{J})\underline{\boldsymbol{\alpha}}.$$

Instead, we flip the order of columns in $\mathbf{C}_{p+1}$ and $\widetilde{\mathbf{C}}_{p+1}$, yielding $\mathbf{D}_{p+1}$ and $\widetilde{\mathbf{D}}_{p+1}$, and flip the rows in $\underline{\boldsymbol{\alpha}}$, forming $\underline{\boldsymbol{\beta}}$. Then,

$$\widetilde{\mathbf{D}}_{p+1}\underline{\boldsymbol{\beta}} = \mathbf{D}_{p+1}\underline{\boldsymbol{\beta}} + \sigma^2\mathbf{I}\underline{\boldsymbol{\beta}}.$$

The eigenvector corresponding to the smallest eigenvalue of $\widetilde{\mathbf{D}}_{p+1}$ is the zero eigenvector of $\mathbf{D}_{p+1}$. Hence, Pisarenko's Method computes the smallest eigenvector of $\widetilde{\mathbf{D}}_{p+1}$, $\underline{\boldsymbol{\beta}}$ and flips the order to yield the coefficients of $\underline{\boldsymbol{\alpha}}$, as summarized in Algorithm 2.12. As $\widetilde{\mathbf{D}}_{p+1}$ is Hermitian matrix, we can use an eigendecomposition rather than a SVD.

As with the Yule-Walker Method, Pisarenko's Method approximates $\underline{\boldsymbol{\alpha}}$ when the autocovariance is estimated using either (2.72) or (2.73) for finite n. We notice $\mathbf{C}_{p+1}$ is approximately porportional to $\underline{\mathbf{H}}^*\underline{\mathbf{H}}$,

$$\mathbf{C}_{p+1} = \begin{bmatrix} c_0 & c_{-1} & \cdots & c_{-p} \\ c_1 & c_0 & \cdots & c_{-p+1} \\ \vdots & \vdots & \ddots & \vdots \\ c_p & c_{p-1} & \cdots & c_0 \end{bmatrix} \propto \begin{bmatrix} \mathbf{h}_0^*\mathbf{h}_0 & \mathbf{h}_1^*\mathbf{h}_0 & \cdots & \mathbf{h}_p^*\mathbf{h}_0 \\ \mathbf{h}_0^*\mathbf{h}_1 & \mathbf{h}_1^*\mathbf{h}_1 & \cdots & \mathbf{h}_p^*\mathbf{h}_1 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_0^*\mathbf{h}_p & \mathbf{h}_1^*\mathbf{h}_p & \cdots & \mathbf{h}_p^*\mathbf{h}_p \end{bmatrix} = \underline{\mathbf{H}}^*\underline{\mathbf{H}}. \tag{2.78}$$

Because most authors seem unware of what the covariance matrix should approximate in the limit of small n, they often propose exotic schemes for 'improving' the estimates of $\underline{\boldsymbol{\alpha}}$. For example, Stoica and Moses form $\widetilde{\mathbf{C}}_{p+1}$ as $\underline{\mathbf{H}}^*\underline{\mathbf{H}}$ ([144,

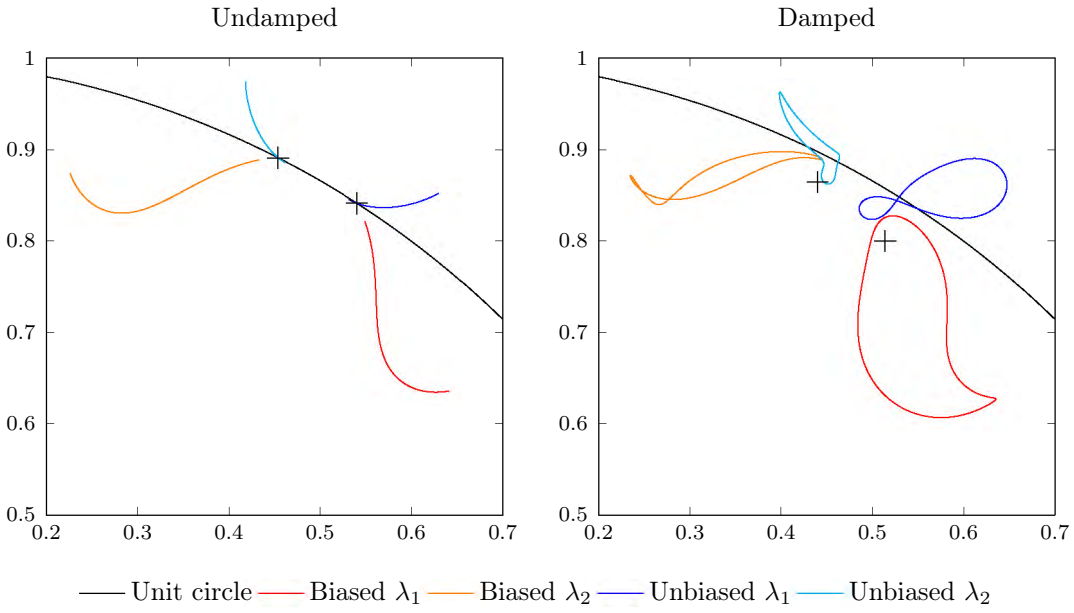


Figure 2.10 : Bias in Pisarenko's method for small n. In this example there are two exponentials $\lambda_1 = e^{1i}$ and $\lambda_2 = e^{1.1i}$ on the left and $\lambda_1 = e^{1i-0.05}$ and $\lambda_2 = e^{1i-0.03}$ on the right (marked by +). In both cases $a_1 = 1$ and $a_2 = e^{i\theta}$. As θ swept through 0 to 2π the recovered values of $\boldsymbol{\lambda}$ using Pisarenko's method using both biased and unbiased covariance estimates generated the curves seen above. Here $n = 10$.

eqs. (4.5.14), (4.8.1)]), but then noting it does not have the Toeplitz structure expected, recommend averaging the diagonal entries to enforce this structure [144, p. 159]. This can only make performance worse, because $\underline{\mathbf{H}}^*\underline{\mathbf{H}}$ already provides exact estimates of $\underline{\boldsymbol{\alpha}}$ in the absence of noise.

2.7 Compressed Prony's Method

Compressed variants of Prony's Method, such as the Method of Moments, combine measurements $\widetilde{\mathbf{y}}$ resulting in new pseudo-measurements $\widetilde{\mathbf{z}}$ and use a Prony-type method to then recover the exponential parameters $\widetilde{\boldsymbol{\lambda}}$. This has two advantages. During the 1960s and 1970s, computing resources were limited; combining measure-

Algorithm 2.11: Kung's Method

Input : Measurements $y_0, y_1, \ldots, y_{n-1} \in \mathbb{C}$, model order p

Output: Exponential parameters $\lambda_0, \lambda_1, \ldots, \lambda_{p-1}$.

1 Form $\underline{\mathbf{H}} \in \mathbb{C}^{(n+1-\ell)\times \ell}$ with $[\mathbf{H}]_{j,k} = y_{j+k}$;

2 Compute the SVD: $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^* \leftarrow \underline{\mathbf{H}}$;

3 Form $\widetilde{\mathbf{D}} \leftarrow (\widetilde{\mathbf{U}}_p^{\uparrow})^{+}\widetilde{\mathbf{U}}_p^{\downarrow}$;

4 Compute the eigenvalues $\boldsymbol{\lambda}$ of $\widetilde{\mathbf{D}}$;

Algorithm 2.12: Pisarenko's Method

Input : Measurements $y_0, y_1, \ldots, y_{n-1} \in \mathbb{C}$ and model order p.

Output: Exponential parameters $\lambda_0, \lambda_1, \ldots, \lambda_{p-1}$.

1 Estimate covariances using either $c_j = \frac{1}{n}\sum_{k=0}^{n-j-1} y_{k+j}\overline{y_j}$ (biased) or $c_j = \frac{1}{n-k}\sum_{k=0}^{n-j-1} y_{k+j}\overline{y_j}$ (unbiased) for $j = -p, \ldots, p$;

2 Form $\mathbf{C} \in \mathbb{C}^{p+1\times p+1}$ where $[\mathbf{C}]_{j,k} = c_{j-k}$ ($c_{-k} = \overline{c_k}$);

3 Compute eigendecomposition of $\mathbf{C}$, $\mathbf{C}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}$ where $[\mathbf{\Lambda}]_{j,j} = \lambda_j$ and $\lambda_j \geq \lambda_k \forall j > k$;

4 Set $\underline{\boldsymbol{\alpha}} = [\mathbf{V}]_{\cdot,p+1} \in \mathbb{C}^{p+1}$;

5 Find roots λ_k of $q(t) = \alpha_p t^p + \alpha_{p-1}t^{p-1} + \alpha_{p-2}t^{p-2} + \ldots + \alpha_1 t + \alpha_0$.

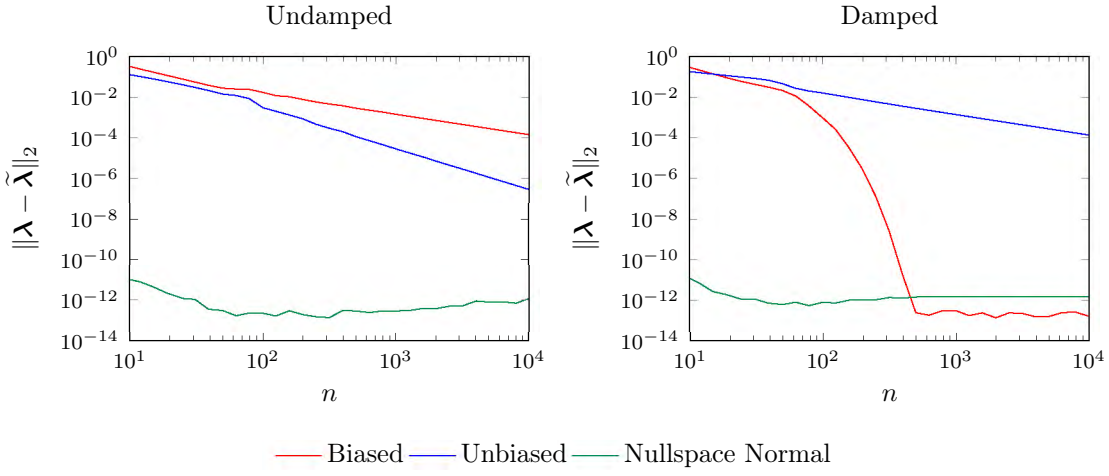


Figure 2.11 : Error in Pisarenko's method as a function of n using the undamped and damped parameters in Figure 2.10. Nullspace normal refers to taking $\mathbf{C} = \underline{\mathbf{H}}^*\underline{\mathbf{H}}$.

ments was a promising alternative to solving a potentially large least squares problem. Also, by combining measurements, the covariance of the pseudo-measurements $\widetilde{z}_j$ is reduced relative to $\widetilde{y}_j$ and if errors in $\widetilde{\mathbf{y}}$ are not normally distributed, the summation of measurements tends to make errors normally distributed as a result of the Central Limit Theorem.

A prototype of these methods begins with Prony Least Squares

$$\widetilde{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\widetilde{\mathbf{H}}\boldsymbol{\alpha} + \widetilde{\mathbf{h}}\|_2 \qquad \widetilde{\mathbf{H}} \in \mathbb{C}^{m\times p},\ \widetilde{\mathbf{h}} \in \mathbb{C}^m \tag{2.79}$$

and compresses these measurements by multiplying on the left by a full rank matrix $\mathbf{W} \in \mathbb{C}^{m\times p}$, yielding the compressed estimate

$$\widetilde{\boldsymbol{\alpha}}_{\mathbf{W}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{W}^*\widetilde{\mathbf{H}}\boldsymbol{\alpha} + \mathbf{W}^*\widetilde{\mathbf{h}}\|_2. \tag{2.80}$$

Following (2.23), the first term in the perturbation expansion of $\widetilde{\boldsymbol{\alpha}}$ is

$$\boldsymbol{\alpha}^{(1)}_{\mathbf{W}} = -(\mathbf{W}^*\mathbf{H})^+\mathbf{W}^*\mathbf{T}(\widehat{\boldsymbol{\alpha}})\mathbf{g}. \tag{2.81}$$

One trivial choice picks $\mathbf{W}$ that spans the range of $\widetilde{\mathbf{H}}$ with orthogonal columns; this simply yields the least squares estimate for $\widetilde{\boldsymbol{\alpha}}$. Another choice would take $\mathbf{W}$ to be columns of the discrete Fourier matrix $\mathbf{F}_m$ corresponding to frequencies of interest; Mandelshtam takes a similar approach in Filter Diagonalization [103]. Other, more exotic choices of $\mathbf{W}$ are possible, such as Finite Impulse Response filters [29].

In the remainder of this section, we review existing compressed techniques and discuss a new variation of the Maximum Likelihood Prony Method that uses compression to reduce the cost of the norm correction step.

2.7.1 Cornell's Method

Cornell's Method [33] is a modification of Prony's Method that partitions $\widetilde{\mathbf{y}}$ into sums of $m = \lfloor n/(2p) \rfloor$ adjacent measurements, yielding p pseudo-measurements

$$\widetilde{z}_k = \sum_{j=0}^{m-1} \widetilde{y}_{j+km} \qquad 0 \le k < 2p. \tag{2.82}$$

The method then continues with Prony's Method, replacing $\widetilde{\mathbf{y}}$ with $\widetilde{\mathbf{z}}$

$$\begin{bmatrix} \widetilde{z}_0 & \widetilde{z}_1 & \widetilde{z}_2 & \cdots & \widetilde{z}_{p-1} \\ \widetilde{z}_1 & \widetilde{z}_2 & \widetilde{z}_3 & \cdots & \widetilde{z}_p \\ \vdots & \vdots & & & \vdots \\ \widetilde{z}_{p-1} & \widetilde{z}_p & \widetilde{z}_{p+1} & \cdots & \widetilde{z}_{2p-2} \end{bmatrix} \begin{bmatrix} \widetilde{\alpha}_0 \\ \widetilde{\alpha}_1 \\ \vdots \\ \widetilde{\alpha}_{p-1} \end{bmatrix} = - \begin{bmatrix} \widetilde{z}_p \\ \widetilde{z}_{p+1} \\ \vdots \\ \widetilde{z}_{2p-1} \end{bmatrix}, \tag{2.83}$$

and recovers roots $m\widetilde{\boldsymbol{\lambda}}$ of the polymial q.

2.7.2 Method of Moments

The *Method of Moments* was developed by Isenberg and Dyson in 1969 to study fluoresce of biological compounds [77], and was expanded upon in 1973 [78] and 1983 [76]. This method uses the determinant formulation of Prony's Method discussed in Section 2.2.4 applied to scaled moments μ_k. The moments are defined with respect to the continuous signal $y(t)$,

$$\mu_k := \int_0^\infty t^k y(t)\,\mathrm{d}t = \int_0^\infty t^k \sum_{j=0}^{p-1} a_j \lambda_j^t \,\mathrm{d}t = \sum_{j=0}^{p-1} \frac{a_j k!}{(\log \lambda_j)^{k+1}}. \tag{2.84}$$

Then the Method of Moments computes the scaled moments

$$g_k := \frac{\mu_k}{k!} = \sum_{j=0}^{p-1} a_j \tau^k \tag{2.85}$$

by numerically integrating $y(t)$ using samples $y(t_j)$. These scaled moments then reveal inverse exponential coefficient $\tau_j = \lambda_j^{-1}$ as the roots of the polynomial

$$q(\tau) = \det \begin{bmatrix} \tau^0 & \tau^1 & \tau^2 & \cdots & \tau^p \\ g_0 & g_1 & g_2 & \cdots & g_p \\ g_1 & g_2 & g_3 & \cdots & g_{p+1} \\ \vdots & & & & \vdots \\ g_{p-1} & g_p & g_{p+1} & \cdots & g_{2p-1} \end{bmatrix}. \tag{2.86}$$

As Isenberg points out, this method performs better than nonlinear least squares methods for non-normally distributed noise. This is a result of the central limit theorem – the numerical integration that estimates μ_k adds elements of g_j, and, under mild assumptions as $n \to \infty$, these errors approximate a normal distribution.

2.7.3 Compressed Maximum Likelihood Prony Method

Computing the Q-less QR factorization is the dominant cost of the Maximum Likelihood Prony Method given in Algorithm 2.7, requiring $\mathcal{O}(n^3)$ operations if dense algorithms are used. We can reduce this cost by compressing onto $\mathbf{U} \in \mathbb{C}^{m \times p}$ from the short-form SVD $\widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}^* = \widetilde{\mathbf{H}}$. Modifying the derivation in Section 2.4, the correct norm for this compressed problem is

$$\mathbf{\Gamma}(\boldsymbol{\alpha}) = \mathbf{U}^*\mathbf{T}(\boldsymbol{\alpha})\mathbf{\Sigma}\mathbf{T}(\boldsymbol{\alpha})^*\mathbf{U}, \tag{2.87}$$

cf, (2.32). Consequently, instead of computing the QR factorization of $\mathbf{T}(\boldsymbol{\alpha}) \in \mathbb{C}^{m \times n}$, we compute the QR factorization of the smaller matrix $\mathbf{U}^*\mathbf{T}(\boldsymbol{\alpha}) \in \mathbb{C}^{p \times n}$. The modified algorithm is given in Algorithm 2.13.

Algorithm 2.13: Compressed Maximum Likelihood Prony Method

Input : Measurements $y_0, y_1, \ldots, y_{n-1} \in \mathbb{C}$, model order p, covariance $\mathbf{\Sigma}$, and convergence tolerance τ

Output: Exponential parameters $\lambda_0, \lambda_1, \ldots, \lambda_{p-1}$.

Form $\mathbf{H} \in \mathbb{C}^{(n-p)\times p}$ with $[\mathbf{H}]_{j,k} = y_{j+k}$ and $\mathbf{h} \in \mathbb{C}^{n-p}$ with $[\mathbf{h}]_j = y_{j+p}$;
Compute $\boldsymbol{\alpha} = -\mathbf{H}^{+}\mathbf{h}$;
Compute the short form SVD: $\mathbf{U\Sigma V}^{*} = \mathbf{H}$;
Form $\widehat{\mathbf{H}} \leftarrow \mathbf{U}^{*}\mathbf{H}$ and $\widehat{\mathbf{h}} \leftarrow \mathbf{U}^{*}\mathbf{h}$;
Compute the Cholesky decomposition $\mathbf{LL}^{*} \leftarrow \mathbf{\Sigma}$;
while $\|\boldsymbol{\alpha}' - \boldsymbol{\alpha}\| > \tau$ **do**
    $\boldsymbol{\alpha}' \leftarrow \boldsymbol{\alpha}$;
    Compute the $\mathbf{R}$ in the QR-decomposition $\mathbf{QR} \leftarrow \mathbf{L}^{*}\mathbf{T}(\boldsymbol{\alpha})^{*}\mathbf{U}$;
    $\boldsymbol{\alpha} \leftarrow -(\mathbf{R}^{-*}\widehat{\mathbf{H}})^{-1}(\mathbf{R}^{-*}\widehat{\mathbf{h}})$;
Find roots λ_k of $q(t) = t^p + \alpha_{p-1}t^{p-1} + \alpha_{p-2}t^{p-2} + \cdots + \alpha_1 t + \alpha_0$.

2.8 Compressed Matrix Pencil Methods for Localized Exponentials

Oftentimes we are interested in determining a few critical exponential coefficients among thousands in our dataset. Determining all these exponential coefficients would be impractical since the computational cost Prony-type methods grows cubically in the number of exponentials (nonlinear least squares methods grow quadratically). In this section, we develop two new algorithms based on the matrix pencil formulation of Prony's Method that use a pair of compression matrices $\mathbf{W}$ and $\mathbf{Y}$ to recover only a desired set of exponentials located near a target frequency.

We start with the matrix pencil formulation, building overdetermined matrices $\mathbf{H}_0$ and $\mathbf{H}_1$ (there is no requirement for them to be square), and then reducing the dimension of the matrix pencil with using left and right compression matrices $\mathbf{W}$ and

$\mathbf{Y}$ that yield a square generalized eigenvalue problem

$$\lambda \mathbf{W}^* \mathbf{H}_0 \mathbf{Y} \mathbf{x} = \mathbf{W}^* \mathbf{H}_1 \mathbf{Y} \mathbf{x}. \tag{2.88}$$

We choose the two matrices $\mathbf{W}$ and $\mathbf{Y}$ by noting that the Hankel matrices $\mathbf{H}_0$ and $\mathbf{H}_1$ decompose into a product of a Vandermonde matrix $[\mathbf{V}(\boldsymbol{\lambda})]_{j,k} = \lambda_k^j$ and two diagonal matrices, $\mathbf{A} = \text{diag}(\mathbf{a})$ and $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$:

$$\mathbf{H}_0 = \mathbf{V}(\boldsymbol{\lambda}) \mathbf{A} \mathbf{V}(\boldsymbol{\lambda})^\top, \qquad \mathbf{H}_1 = \mathbf{V}(\boldsymbol{\lambda}) \mathbf{A} \boldsymbol{\Lambda} \mathbf{V}(\boldsymbol{\lambda})^\top. \tag{2.89}$$

If we seek some subset of wanted eigenvalues $\boldsymbol{\lambda}_w$, we would ideally choose $\mathbf{W}$ to have range $\mathbf{V}(\boldsymbol{\lambda}_w)$ and nullspace $\mathbf{V}(\boldsymbol{\lambda}_u)$, where $\boldsymbol{\lambda}_u$ are the unwanted eigenvalues. Similarly, we would find $\mathbf{Y}$ with range $\overline{\mathbf{V}(\boldsymbol{\lambda})}$ and nullspace $\overline{\mathbf{V}(\boldsymbol{\lambda}_u)}$.

2.8.1 Filtered Matrix Pencil Method

If we know the desired exponential parameters $\boldsymbol{\lambda}_w$ fall in some frequency range, we can choose $\mathbf{W}$ and $\mathbf{Y}$ to contain columns of the Fourier matrix with a similar frequency. In particular, if $\mathcal{I}$ is a set of indices near $\arg \boldsymbol{\lambda}_w$ (i.e., $\mathbf{F}_{\cdot,\mathcal{I}}^* \mathbf{V}(\boldsymbol{\lambda}_w)$ is large), we choose $\mathbf{W} = \mathbf{F}_{\cdot,\mathcal{I}}$ and $\mathbf{Y} = \overline{\mathbf{F}}_{\cdot,\mathcal{I}}$. This choice approximately filters out exponentials from outside the frequency range selected, as $\mathbf{F}_{\cdot,\mathcal{I}}^* \mathbf{V}(\lambda)$ is small, but preserves those frequencies inside it. As our previous discussions have noted, including additional, spurious exponentials improves numerical stability and removes error due to exponentials outside the desired frequency region. If additional exponentials are included, an additional SVD step can be used to remove the spurious exponentials as described in Algorithm 2.14. However, this simple approach can never remove the corrupting influence of other exponentials outside the desired range.

We can see this corrupting influence in Figure 2.13. Here, we consider $\mathbf{y}$ with 100 exponentials that grow increasingly close in imaginary part as shown in Figure 2.12.

Algorithm 2.14: Filtered Matrix Pencil Method

Input : Measurements $\widetilde{\mathbf{y}} \in \mathbb{C}^n$, desired number of exponentials p_w, an undesired number of exponentials $p_u \geq 1$, a target frequency $\mathring{\omega}$.

Output: $\widetilde{\boldsymbol{\omega}} \in \mathbb{C}^{p_w}$

$p \leftarrow p_u + p_w$;

$m \leftarrow \lfloor n/2 \rfloor$;

Form $\mathbf{h}_0 \in \mathbb{C}^{m \times m}$ and $\mathbf{h}_1 \in \mathbb{C}^{m \times m}$ from $\widetilde{\mathbf{y}}$;

Find indices $\mathcal{I}$ corresponding to the p smallest values of $|e^{\phi_k} - e^{\mathring{\omega}}|$ where $\phi_k = 2\pi ik/m$ and $0 \leq k < m$;

Form $\mathbf{W} \leftarrow [\mathbf{F}_m]_{\cdot,\mathcal{I}}$;

Form $\mathbf{Y} \leftarrow \overline{[\mathbf{F}_m]_{\cdot,\mathcal{I}}}$;

Compute the rank p_w skinny SVD: $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^* \leftarrow \mathbf{W}^*\mathbf{H}_0\mathbf{Y}$;

Compute eigenvalues $\boldsymbol{\lambda}$ of $\boldsymbol{\Sigma}^{1/2}\mathbf{U}^*\mathbf{W}^*\mathbf{H}_1\mathbf{Y}\mathbf{V}\boldsymbol{\Sigma}^{1/2}$;

$\widetilde{\boldsymbol{\omega}} \leftarrow \log(\boldsymbol{\lambda})$;

Using the Filtered Matrix Pencil Method, we recover the single exponential ω_j with an error about 10^{-3} that gets worse as the exponentials get closer together. One approach to remove this error leds to the Orthogonalized Matrix Pencil Method that orthogonalizes against the unwanted exponentials.

2.8.2 Orthogonalized Matrix Pencil Method

Rather than use a fixed pair of compression matrices, the Orthogonalized Matrix Pencil Method updates $\mathbf{W}$ and $\mathbf{Y}$ using the estimates of the wanted exponentials $\boldsymbol{\lambda}_w$ and the unwanted exponentials $\boldsymbol{\lambda}_u$. Starting from columns of the Fourier matrix as in the Filter Matrix Pencil Method, we remove the recovered, unwanted exponentials $\boldsymbol{\lambda}_u$ by orthogonalizing $\mathbf{W}$ against them:

$$\widehat{\mathbf{W}} \leftarrow (\mathbf{I} - \mathbf{Q}\mathbf{Q}^*)\mathbf{W} \quad \text{where } \mathbf{Q}\mathbf{R} = \mathbf{V}(\boldsymbol{\lambda}_u)$$

$$\widehat{\mathbf{Y}} \leftarrow (\mathbf{I} - \widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^*)\mathbf{Y} \quad \text{where } \widehat{\mathbf{Q}}\widehat{\mathbf{R}} = \overline{\mathbf{V}(\boldsymbol{\lambda}_u)}.$$

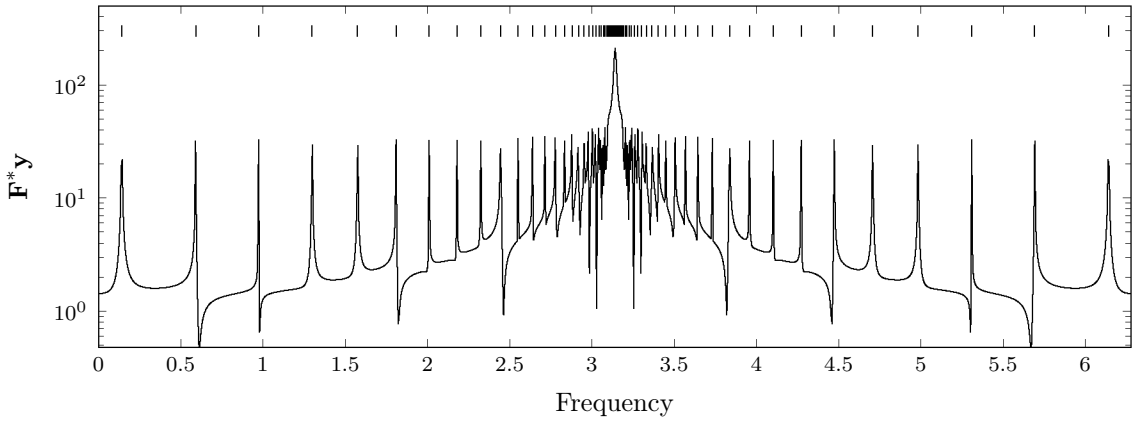


Figure 2.12 : Fourier transform of $\mathbf{y}$ with 100 exponentials where $\omega_j = i(\pi - 3 \cdot 0.85^{j-1})$ for $0 \leq j < 50$ and $\omega_j = i(\pi + 3 \cdot 0.85^{99-j})$ for $50 \leq j \leq 99$. For all j, $a_j = 1$. The black ticks above denote the location of $\boldsymbol{\omega}$.

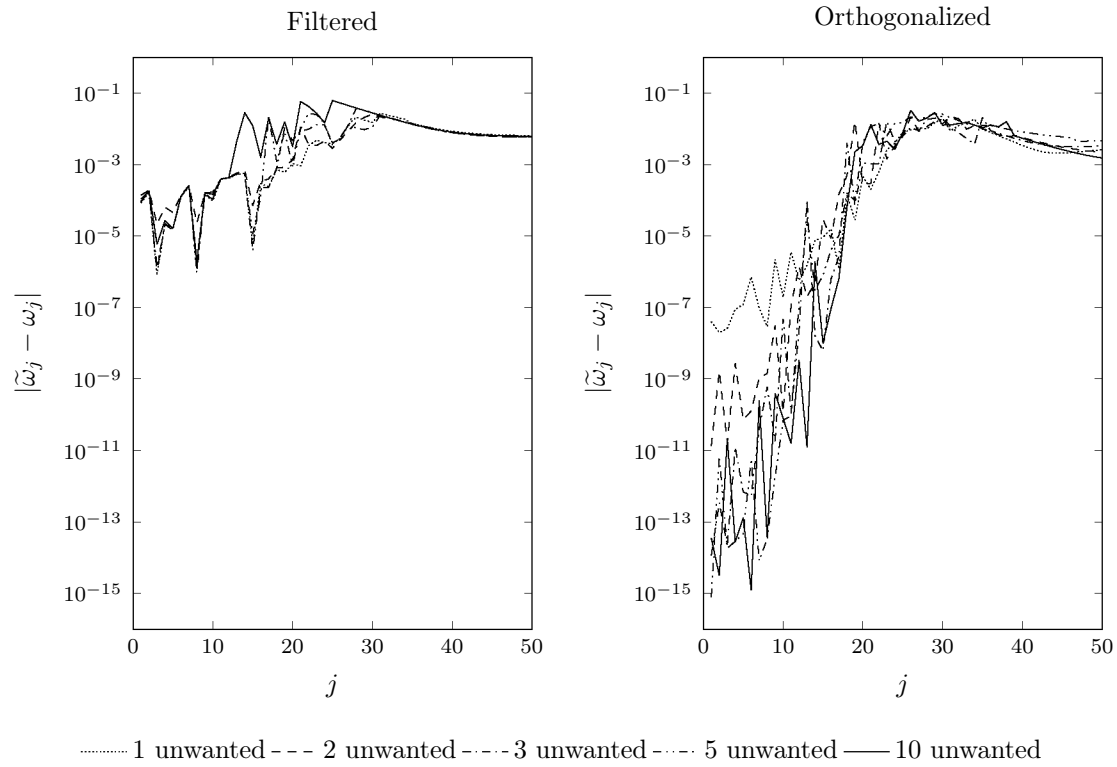


Figure 2.13 : Performance of Filtered and Orthogonalized Matrix Pencil Methods. Each algorithm was tasked with finding the single exponential with frequency ω_j from $\mathbf{y}$ given in Figure 2.12 using between one and ten unwanted exponentials. As j increases, the exponentials become more closely packed, and for $j > 30$, error decreases due to the packing.

Then we reorthogonalize $\widehat{\mathbf{W}}$ and $\widehat{\mathbf{Y}}$ for numerical stability by applying another QR factorization. It is important that $\mathbf{Y}$ is orthogonalized against $\overline{\mathbf{V}(\boldsymbol{\lambda}_u)}$, as in the Vandermonde decomposition of $\mathbf{H}_0$ and $\mathbf{H}_1$, the product $\mathbf{V}(\boldsymbol{\lambda})^\top\mathbf{Y}$ appears. As we repeat this procedure, $\mathbf{W}$ and $\mathbf{Y}$ become increasingly orthogonal to the unwanted elements of $\boldsymbol{\lambda}$, even though each iteration may not compute all the λ_j present. These steps are summarized in Algorithm 2.15. This algorithm assumes a target frequency $\mathring{\omega}$, but more sophisticated rules for separating wanted and unwanted exponentials can be applied.

As we see in Figure 2.13, the Orthogonalized Matrix Pencil Method can yield estimates of ω_j accurate to rounding error when the exponentials are well separated and a number of unwanted exponentials are included. However, as the exponentials get closer together as j increases, this proceedure yields increasingly inaccurate estimates. Part of the complication is $\mathbf{V}(\boldsymbol{\lambda}_u)$ is close to parallel to $\mathbf{V}(\boldsymbol{\lambda}_w)$, leading $\mathbf{W}$ and $\mathbf{Y}$ to capture little of the desired exponentials. One correction would be to restart $\mathbf{W}$ and $\mathbf{Y}$ using the current estimates of $\boldsymbol{\lambda}_w$; i.e., build $\mathbf{W}$ and $\mathbf{Y}$ from $\mathbf{V}(\boldsymbol{\lambda}_w)$ and near by columns of the Fourier matrix. As with all restarting, this would sacrifice information gained about the unwanted $\boldsymbol{\lambda}$.

Algorithm 2.15: Orthogonalized Matrix Pencil Method

Input : Measurements $\widetilde{\mathbf{y}} \in \mathbb{C}^n$, desired number of exponentials p_w, an undesired number of exponentials $p_u \geq 1$, a target frequency $\mathring{\omega}$.

Output: $\widetilde{\boldsymbol{\omega}} \in \mathbb{C}^{p_w}$

$p \leftarrow p_u + p_w$;

$m \leftarrow \lfloor n/2 \rfloor$;

Form $\mathbf{H}_0 \in \mathbb{C}^{m \times m}$ and $\mathbf{H}_1 \in \mathbb{C}^{m \times m}$ from $\widetilde{\mathbf{y}}$;

Find indices $\mathcal{I}$ corresponding to the p smallest values of $|e^{\phi_k} - e^{\mathring{\omega}}|$ where $\phi_k = 2\pi ik/m$ and $0 \leq k < m$;

Form $\mathbf{W} \leftarrow [\mathbf{F}_m]_{\cdot,\mathcal{I}}$;

Form $\mathbf{Y} \leftarrow \overline{[\mathbf{F}_m]_{\cdot,\mathcal{I}}}$;

repeat

  Solve $\lambda \mathbf{W}^* \mathbf{H}_0 \mathbf{W} \mathbf{v} = \mathbf{W}^* \mathbf{H}_1 \mathbf{W} \mathbf{v}$, yielding eigenvalues $\boldsymbol{\lambda}$;

  Separate $\boldsymbol{\lambda}$ into: wanted $\boldsymbol{\lambda}_w$ nearest $e^{\mathring{\omega}}$; place the remainder in $\boldsymbol{\lambda}_u$;

  QR-Factorization $\mathbf{QR} \leftarrow \mathbf{V}(\boldsymbol{\lambda}_u)$;

  $\widehat{\mathbf{W}} \leftarrow (\mathbf{I} - \mathbf{Q}\mathbf{Q}^*)\mathbf{W}$;

  Orthogonalize $\widehat{\mathbf{W}}$, replacing $\mathbf{W}$;

  $\widehat{\mathbf{Y}} \leftarrow (\mathbf{I} - \overline{\mathbf{Q}}\mathbf{Q}^\top)\mathbf{Y}$;

  Orthogonalize $\widehat{\mathbf{Y}}$, replacing $\mathbf{Y}$;

until *convergence*;

$\widetilde{\boldsymbol{\omega}} \leftarrow \log(\boldsymbol{\lambda})$.

Chapter 3

Nonlinear Least Squares Methods

Maximum likelihood methods are the alternative to Prony's Method and its modifications. The basic premise, as the name implies, is to maximize the likelihood of the parameters $\boldsymbol{\theta}$ given noisy measurements $\widetilde{\mathbf{y}}$. The likelihood function is defined as

$$\mathcal{L}(\boldsymbol{\theta}) := p(\widetilde{\mathbf{y}}|\boldsymbol{\theta}), \tag{3.1}$$

where p is the probability distribution associated with the noise (see Appendix A). This leads to the maximum likelihood estimate of $\boldsymbol{\theta}$:

$$\widetilde{\boldsymbol{\theta}}_{\mathrm{ML}} := \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \mathcal{L}(\boldsymbol{\theta}). \tag{3.2}$$

In practice, we often assume the noise is additive, e.g., $\widetilde{\mathbf{y}} = \mathbf{y} + \mathbf{g} = \mathbf{f}(\widehat{\boldsymbol{\theta}}) + \mathbf{g}$, and $\mathbf{g}$ samples a proper complex normal distribution, $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})$. This transforms the maximum likelihood problem (3.2) into an equivalent nonlinear least squares problem:

$$\widetilde{\boldsymbol{\theta}}_{\mathrm{ML}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \|\widetilde{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta})\|_{\boldsymbol{\Gamma}}^2, \quad \text{where} \quad \|\mathbf{x}\|_{\boldsymbol{\Gamma}}^2 = \mathbf{x}^* \boldsymbol{\Gamma}^{-1} \mathbf{x}. \tag{3.3}$$

Although normally distributed noise is common (a result of the Central Limit Theorem), the practical benefit is that the least squares problem (3.3) is easier to solve than (3.2). Many algorithms exploit the additional structure to reduce computational costs; e.g., the Gauss-Newton and Levenberg-Marquard algorithms. Additionally, in this setting maximum likelihood methods obtain the Cramér-Rao lower bound on the covariance of $\widetilde{\boldsymbol{\theta}}$. Unfortunately, the optimization algorithms used to solve (3.2) and

(3.3) require initial estimates of $\boldsymbol{\theta}$, something Prony's Method does not require. Poor initial estimates can cause optimization algorithms to converge to the wrong value, which we observe in our practical computations.

In this chapter, we briefly review the important parts of the optimization algorithms for nonlinear least squares problems: the Gauss-Newton approximation of the second derivative and the Variable Projection functional for separable problems. This overview provides the framework for asymptotic estimates for both the general and separable least squares problems. These show that we do indeed obtain the Cramér-Rao bound. Importantly, these asymptotic estimates lay the foundation for computing the efficiency of the compressed problem in the next chapter. Finally, we address a common complaint about nonlinear least squares methods for exponential fitting: that nonlinear least squares methods require initial estimates of $\boldsymbol{\omega}$ and are prone to converge to spurious local minima (e.g., [39]). The final three sections discuss how these concerns can be addressed using various heuristics. Section 3.4 compares various techniques to provide initial parameter estimates for optimization tailored to exponential fitting. Section 3.5 discusses three different techniques for determining the number of exponentials in a signal $\mathbf{y}$ and develops a new bound based on the singular values of a Hankel matrix. Finally, Section 3.6, updates the *Peeling Method*, originally developed by Perl [122], that sequentially removes exponentials. This new version treats complex data, using the Akaike Information Criterion to determine the number of exponentials and a filtered Matrix Pencil Method to provide initial estimates. Combined, these techniques significantly address all the complaints about nonlinear least squares methods, excepting speed: for small n, Prony-type methods are often faster. Compression, a technique developed in the next chapter, speeds the solution exponential fitting by sacrificing a small amount of statistical accuracy.

3.1 Nonlinear Least Squares Algorithms

There exists a vast array of algorithms for solving nonlinear least squares problems like (3.3): e.g., Gradient Descent, Conjugate Gradient, BGFS, Gauss-Newton, and Levenberg-Marquard; see Björck [20, Ch. 9] or Nocedal and Wright [114, Ch. 10] for more detail. For our purposes, we briefly provide sufficient background for the perturbation analysis in the next section.

In solving (3.3), we minimize the function

$$\phi(\boldsymbol{\theta}) = \mathbf{r}(\boldsymbol{\theta})^* \boldsymbol{\Gamma}^{-1} \mathbf{r}(\boldsymbol{\theta}), \quad \text{where} \quad \mathbf{r}(\boldsymbol{\theta}) = \widetilde{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta}). \tag{3.4}$$

For simplicity, we can remove $\boldsymbol{\Gamma}$ by equivalently placing it onto the residual vector $\mathbf{r}$, $\mathbf{r} \leftarrow \boldsymbol{\Gamma}^{-1/2}\mathbf{r}$, allowing us to take $\phi = \mathbf{r}^*\mathbf{r}$. The derivative of $\mathbf{r}$ is the Jacobian, denoted $\mathbf{J}(\boldsymbol{\theta})$, where

$$[\mathbf{J}(\boldsymbol{\theta})]_{\cdot,k} = \frac{\partial}{\partial \theta_k} \mathbf{r}(\boldsymbol{\theta}). \tag{3.5}$$

3.1.1 Gradient Descent

Using the first-order Taylor expansion of ϕ (see Appendix B),

$$\phi(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta}_0) + 2\text{Re}\,[\mathbf{r}(\boldsymbol{\theta}_0)^* \mathbf{J}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)] + \mathcal{O}(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2), \tag{3.6}$$

we notice that ϕ decreases most in the opposite direction of the gradient $\mathbf{g}(\boldsymbol{\theta}_0) = \mathbf{J}(\boldsymbol{\theta}_0)^* \mathbf{r}(\boldsymbol{\theta}_0)$. Using this fact, starting from our initial estimate $\boldsymbol{\theta}_0$, we advance the estimate of $\boldsymbol{\theta}$ by following $\mathbf{p}_k = -\mathbf{g}(\boldsymbol{\theta}_k)$,

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_k \mathbf{p}_k \tag{3.7}$$

using α_k to control the step-size and ensure convergence via, e.g., the Wolfe conditions [114, eq. 1.7]. This method is called *Gradient Descent* and $\boldsymbol{\theta}_k$ converges at best linearly.

3.1.2 Newton's Method

Including the second derivative of ϕ yields an improved quadratic model. The first three terms in the Taylor expansion of ϕ about $\boldsymbol{\theta}_0$ are

$$\phi(\boldsymbol{\theta}+\mathbf{p}) = \phi(\boldsymbol{\theta}) + \operatorname{Re}\left[2\mathbf{r}^*\mathbf{J}\mathbf{p} + \mathbf{p}^*\mathbf{J}^*\mathbf{J}\mathbf{p} + \sum_k \overline{p_k}\mathbf{r}^*\frac{\partial \mathbf{J}}{\partial \theta_k}\mathbf{p}\right] + \mathcal{O}(\|\mathbf{p}\|^3). \tag{3.8}$$

The second derivative matrix implicit above is called the Hessian, $\mathbf{H}(\boldsymbol{\theta})$,

$$[\mathbf{H}(\boldsymbol{\theta})]_{j,k} = \frac{\partial^2 \phi}{\partial \theta_j \partial \overline{\theta_k}}. \tag{3.9}$$

This provides a quadratic model of ϕ,

$$\psi_k(\mathbf{p}) = \phi(\boldsymbol{\theta}) + \mathbf{g}(\boldsymbol{\theta})^*\mathbf{p} + \frac{1}{2}\mathbf{p}^*\mathbf{H}(\boldsymbol{\theta})\mathbf{p} = \phi(\boldsymbol{\theta}+\mathbf{p}) + \mathcal{O}(\|\mathbf{p}\|^3). \tag{3.10}$$

If $\mathbf{H}$ is positive definite, $\mathbf{p}_k = -\mathbf{H}^{-1}\mathbf{g}$ is the unique minimizer of ψ_k. We then iterate with $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \mathbf{p}_k$. This algorithm, *Newton's Method*, converges quadratically.

However, $\mathbf{H}$ may not always be positive definite, nor the quadratic model ψ accurate if $\mathbf{p}$ is large. One approach is to limit the motion of $\mathbf{p}$ to be within a *trust region* $\|\mathbf{p}\|_2 \leq \Delta$. Finding a $\mathbf{p}$ in this trust region is equivalent to finding a nonnegative scalar λ that regularizes the problem, equivalently solving

$$\begin{aligned} (\mathbf{H}+\lambda\mathbf{I})\mathbf{p} &= -\mathbf{g} \\ \lambda(\Delta - \|\mathbf{p}\|_2) &= 0 \\ \mathbf{H}+\lambda\mathbf{I} &\succeq \mathbf{0}; \end{aligned} \tag{3.11}$$

see Sorensen [141, Lem. 4.2] for further details.

For exponential fitting, we also need to impose constraints on $\boldsymbol{\omega}$. The function $\mathbf{f}$ rapidly diverges as $\boldsymbol{\omega}$ enters the right half plane, causing the estimate of ψ_k to be inaccurate. To avoid this, we apply box constraints to $\boldsymbol{\omega}$, which allow only slight movement into the right half plane. Coleman and Li provide an easy way to apply these box constraints by modifying the Hessian in the trust region subproblem [31].

3.1.3 Gauss-Newton Approximation

The Gauss-Newton approximation of the Hessian neglects the pure second derivatives $\partial \mathbf{J}/\partial \theta_k$, leaving

$$\mathbf{H}_{\mathrm{GN}} = \mathbf{J}^*\mathbf{J}. \tag{3.12}$$

Often this approximation is used to avoid explicitly computing the second derivative. Although for exponential fitting computing additional is possible, the Gauss-Newton approximation reduces the condition number of the trust region subproblem. Rather than solving $(\mathbf{H} + \lambda \mathbf{I})\mathbf{p} = -\mathbf{g}$ to compute $\mathbf{p}$, we may instead use the least squares estimate

$$\mathbf{p} = -\begin{bmatrix} \mathbf{J} \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix}^+ \begin{bmatrix} \mathbf{r} \\ \mathbf{0} \end{bmatrix}; \tag{3.13}$$

see, e.g., [114, eq. (10.41)].

3.1.4 Variable Projection

In the exponential fitting problem, there are two kinds of variables: $\boldsymbol{\omega}$, a truly nonlinear parameter and $\mathbf{a}$, that enters linearly. To see this, we write $\mathbf{f}(\boldsymbol{\theta})$ in terms of $\boldsymbol{\omega}$ and $\mathbf{a}$ separately; $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{V}(\boldsymbol{\omega})\mathbf{a}$ where $[\mathbf{V}(\boldsymbol{\omega})]_{j,k} = e^{j\omega_k}$. Then, in place of (3.3), we solve (taking $\boldsymbol{\Gamma} = \mathbf{I}$ for simplicity)

$$\min_{\boldsymbol{\omega},\mathbf{a}} \|\widetilde{\mathbf{y}} - \mathbf{V}(\boldsymbol{\omega})\mathbf{a}\|_2^2. \tag{3.14}$$

Such a problem is called a *separable* nonlinear least squares problem.

Variable Projection (VARPRO) solves separable problems by implicitly solving for the linear variable $\mathbf{a}$. If $\boldsymbol{\omega}$ is fixed, the pseudoinverse gives the solution for $\mathbf{a}$; $\mathbf{a} = \mathbf{V}(\boldsymbol{\omega})^+\widetilde{\mathbf{y}}$. Applying this result we extract the projector $\mathbf{P}(\boldsymbol{\omega})$:

$$\widetilde{\mathbf{y}} - \mathbf{V}(\boldsymbol{\omega})\mathbf{V}(\boldsymbol{\omega})^+\widetilde{\mathbf{y}} = (\mathbf{I} - \mathbf{V}(\boldsymbol{\omega})\mathbf{V}(\boldsymbol{\omega})^+)\widetilde{\mathbf{y}} = \mathbf{P}(\boldsymbol{\omega})\widetilde{\mathbf{y}}.$$

	Code	Description	Operation Count
1.	$\mathbf{V} \leftarrow \mathbf{V}(\boldsymbol{\omega})$	assemble vectors	$\mathcal{O}(np)$
2.	$\mathbf{Q}, \mathbf{R} \leftarrow \mathbf{V}$	skinny QR of $\mathbf{V}$	$\mathcal{O}(np^2)$
3.	$\mathbf{b} \leftarrow \mathbf{Q}^*\mathbf{y}$		$\mathcal{O}(np)$
4.	$\mathbf{r} \leftarrow \mathbf{y} - \mathbf{Q}\mathbf{b}$	compute residual	$\mathcal{O}(np)$
5.	$\mathbf{a} \leftarrow \mathbf{R}^{-1}\mathbf{b}$	linear parameter computation	$\mathcal{O}(np + p^2)$
6.	$\dot{\mathbf{V}} \leftarrow \dot{\mathbf{V}}(\boldsymbol{\omega})$	assemble vector derivatives	$\mathcal{O}(np)$
7.	$\mathbf{L} \leftarrow \left[\dot{\mathbf{V}} - \mathbf{Q}\mathbf{Q}^*\dot{\mathbf{V}}\right] \operatorname{diag}(\mathbf{a})$		$\mathcal{O}(np^2)$
8.	$\mathbf{K} \leftarrow \mathbf{Q}\mathbf{R}^{-*}\operatorname{diag}(\dot{\mathbf{V}}^*\mathbf{r})$		$\mathcal{O}(np^2 + p^3)$
9.	$\mathbf{J} \leftarrow -\mathbf{L} - \mathbf{K}$	form Jacobian	$\mathcal{O}(np)$
		total cost:	$\mathcal{O}(np^2 + p^3)$
		with Kaufman simplification:	$\mathcal{O}(np^2 + p^2)$

Algorithm 3.1: Pseudocode computing the Variable Projection residual and Jacobian of a separable nonlinear least squares problem when each column of $\mathbf{V}(\boldsymbol{\omega})$ depends on the corresponding entry of $\boldsymbol{\omega}$. In place of the derivative tensor we form $\dot{\mathbf{V}}(\boldsymbol{\omega}) = [\mathbf{V}'(\omega_1)\,,\, \mathbf{V}'(\omega_2)\,,\, \ldots\,,\, \mathbf{V}'(\omega_p)]$. The Kaufman approximation (setting $\mathbf{K} = \mathbf{0}$) removes $\mathcal{O}(p^3)$ cost but decreases the accuracy of the Jacobian away from the minimum. If only the residual is required, the algorithm terminates after step 4.

Thus we can minimize over $\boldsymbol{\omega}$ alone:

$$\min_{\boldsymbol{\omega} \in \mathbb{C}^p} \|\mathbf{P}(\boldsymbol{\omega})\widetilde{\mathbf{y}}\|_2. \tag{3.15}$$

The residual is $\mathbf{r}(\boldsymbol{\omega}) = \mathbf{P}(\boldsymbol{\omega})\widetilde{\mathbf{y}}$ and the Jacobian is

$$[\mathbf{J}(\boldsymbol{\omega})]_{\cdot,j} = -\left[\left(\mathbf{P}(\boldsymbol{\omega})\frac{\partial \mathbf{V}(\boldsymbol{\omega})}{\partial \omega_j}\mathbf{V}(\boldsymbol{\omega})^-\right) + \left(\mathbf{P}(\boldsymbol{\omega})\frac{\partial \mathbf{V}(\boldsymbol{\omega})}{\partial \omega_j}\mathbf{V}(\boldsymbol{\omega})^-\right)^*\right]\widetilde{\mathbf{y}}. \tag{3.16}$$

Here $\mathbf{V}^-$ is any 'least squares' inverse of $\mathbf{V}$ satisfying $\mathbf{V}\mathbf{V}^-\mathbf{V} = \mathbf{V}$ and $\mathbf{V}\mathbf{V}^- = (\mathbf{V}\mathbf{V}^-)^*$ [27, Ch. 6]. This approach for removing $\mathbf{a}$ originated in a thesis by Scolnik for the exponential fitting problem [138] and was generalized by Golub and Pereyra under the name Variable Projection [55]. Near the solution, the second term in (3.16) becomes small, so Kaufman recommends neglecting it [86].

Asymptotically, iterations of Variable Projection converge faster than using the Gauss-Newton method [135, Cor. 3.2]. Additionally, as Variable Projection optimizes

over half the number of variables as the general nonlinear least squares problem (3.3), each iteration typically requires fewer operations. Variable Projection is also faster than *alternating methods* that switch between fixing $\boldsymbol{\omega}$ and solving for $\mathbf{a}$ using linear least squares and then fixing $\mathbf{a}$ and solving for $\boldsymbol{\omega}$ using nonlinear least squares. For details, see the numerical experiments of Ruhe and Wedin [135][1].

3.1.5 Box Constraints

The exponential fitting problem contains one additional challenge: $\mathbf{V}(\omega)$ diverges as $n \to \infty$ if $\operatorname{Re}\omega > 0$ so even small excursions of ω into the right half plane can cause breakdown. To avoid this, we impose box constraints on the value of ω such that $\operatorname{Re}\omega < \omega_{\max}$ following [31]. Experience shows that ω must be allowed to slightly enter the right half plane to avoid slow convergence, and $\omega_{\max}$ is proscribed to keep $\|\mathbf{V}(\omega)\|_\infty$ below a fixed value (≈ 100).

3.2 Perturbation Analysis for Nonlinear Least Squares

We seek a perturbation expansion of the recovered solution $\widetilde{\boldsymbol{\theta}}$ as a function of $\mathbf{g}$ in the limit that $\mathbf{g}$ is small. That is, we seek a power series expansion of

$$\widetilde{\boldsymbol{\theta}}(\epsilon\mathbf{g}) := \operatorname*{argmin}_{\boldsymbol{\theta}} \|\mathbf{f}(\widehat{\boldsymbol{\theta}}) + \mathbf{g} - \mathbf{f}(\boldsymbol{\theta})\|_{\mathbf{\Gamma}}. \tag{3.17}$$

Using this expansion, we then estimate the asymptotic covariance.

[1] Ruhe and Wedin call Variable Projection 'Algorithm I,' Variable Projection with Kaufman's simplification 'Algorithm II,' and alternating methods 'Algorithm III'.

3.2.1 Asymptotic Perturbation

First, we write a power series expansion of $\widetilde{\boldsymbol{\theta}}$: $\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}} + \epsilon\boldsymbol{\theta}^{(1)} + \mathcal{O}(\epsilon^2)$ and expand $\mathbf{f}(\boldsymbol{\theta})$ in a Taylor series, $\mathbf{f}(\widetilde{\boldsymbol{\theta}}) = \mathbf{f}(\widehat{\boldsymbol{\theta}}) + \epsilon\mathbf{F}(\widehat{\boldsymbol{\theta}})\boldsymbol{\theta}^{(1)} + \mathcal{O}(\epsilon^2)$, where $[\mathbf{F}(\boldsymbol{\theta})]_{j,k} = \partial f_j(\boldsymbol{\theta})/\partial\theta_k$. Substituting in (3.17) we obtain

$$\|\mathbf{f}(\widehat{\boldsymbol{\theta}}) + \epsilon\mathbf{g} - \mathbf{f}(\widehat{\boldsymbol{\theta}})\|_{\boldsymbol{\Gamma}} = \epsilon\|\boldsymbol{\Gamma}^{-1/2}\mathbf{g} - \boldsymbol{\Gamma}^{-1/2}\mathbf{F}(\widehat{\boldsymbol{\theta}})\boldsymbol{\theta}^{(1)}\|_2 + \mathcal{O}(\epsilon^2).$$

The first order correction minimizes the norm, yielding

$$\boldsymbol{\theta}^{(1)} = \left(\boldsymbol{\Gamma}^{-1/2}\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{+}\boldsymbol{\Gamma}^{-1/2}\mathbf{g} = \left(\mathbf{F}(\widehat{\boldsymbol{\theta}})^*\boldsymbol{\Gamma}^{-1}\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{-1}\mathbf{F}(\widehat{\boldsymbol{\theta}})^*\boldsymbol{\Gamma}^{-1/2}\mathbf{g}. \tag{3.18}$$

(The last step above requires $\mathbf{F}$ to have full column rank — a condition that holds for exponential fitting provided $\boldsymbol{\theta}$ is distinct.) When $\boldsymbol{\Gamma} = \mathbf{I}$, this reduces to the least squares estimate,

$$\boldsymbol{\theta}^{(1)} = \mathbf{F}^{+}\mathbf{g}. \tag{3.19}$$

3.2.2 Asymptotic Covariance

The asymptotic covariance follows directly from the asymptotic perturbation;

$$\begin{aligned}
\mathsf{Cov}[\widetilde{\boldsymbol{\theta}}] &= \mathsf{E}[(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}})(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}})^*] \\
&= \epsilon^2\left(\boldsymbol{\Gamma}^{-1/2}\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{+}\boldsymbol{\Gamma}^{-1/2}\,\mathsf{E}[\mathbf{g}\mathbf{g}^*]\boldsymbol{\Gamma}^{-1/2}\left(\boldsymbol{\Gamma}^{-1/2}\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{+*} + \mathcal{O}(\epsilon^3) \\
&= \epsilon^2\left(\boldsymbol{\Gamma}^{-1/2}\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{+}\left(\boldsymbol{\Gamma}^{-1/2}\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{+*} + \mathcal{O}(\epsilon^3) \\
&= \epsilon^2\left(\mathbf{F}(\widehat{\boldsymbol{\theta}})^*\boldsymbol{\Gamma}^{-1}\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{-1} + \mathcal{O}(\epsilon^3).
\end{aligned} \tag{3.20}$$

For this problem, the Fisher information matrix is (see Appendix A.5)

$$\mathcal{I}(\widehat{\boldsymbol{\theta}}) := (\nabla\mathbf{f}(\widehat{\boldsymbol{\theta}}))^*(\epsilon^2\boldsymbol{\Gamma})^{-1}(\nabla\mathbf{f}(\widehat{\boldsymbol{\theta}})) = \epsilon^{-2}\mathbf{F}(\widehat{\boldsymbol{\theta}})^*\boldsymbol{\Gamma}^{-1}\mathbf{F}(\widehat{\boldsymbol{\theta}}), \tag{3.21}$$

and the Cramér-Rao bound states

$$\mathsf{Cov}[\widetilde{\boldsymbol{\theta}}] \succeq \mathcal{I}(\widehat{\boldsymbol{\theta}})^{-1} = \epsilon^2\left(\mathbf{F}(\widehat{\boldsymbol{\theta}})^*\boldsymbol{\Gamma}^{-1}\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{-1}. \tag{3.22}$$

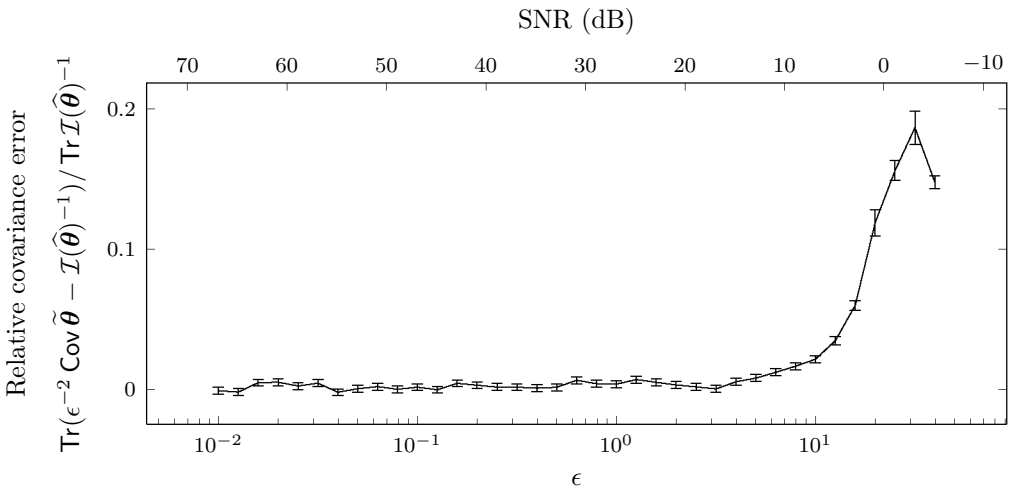


Figure 3.1 : The relative accuracy of the asymptotic covariance estimate (3.20) for large noise in an MRS example [153, Table 1] with $p = 11$ exponentials. Error bars denote error the sample covariance. The covariance estimate remains accurate up until the signal to noise ratio (SNR) reaches about 3 (10 dB).

Hence, in the limit $\epsilon \to 0$, $\mathsf{Cov}[\widetilde{\boldsymbol{\omega}}] = \mathcal{I}(\widehat{\boldsymbol{\theta}})^{-1}$ – the maximum likelihood estimator asymptotically obtains the Cramér-Rao bound. In our numerical experiments, the asymptotic estimate (3.20) holds even for fairly large values of ϵ, as seen in Figure 3.1.

3.3 Perturbation Analysis for Separable Nonlinear Least Squares

Computing the perturbation expansion for the separable problem follows the same approach as the general problem in the previous section. We define

$$\widetilde{\boldsymbol{\omega}}(\mathbf{g}) := \operatorname*{argmin}_{\boldsymbol{\omega}} \|\mathbf{P}(\boldsymbol{\omega})(\mathbf{V}(\widehat{\boldsymbol{\omega}})\widehat{\mathbf{a}} + \mathbf{g})\|_2^2 = \operatorname*{argmin}_{\boldsymbol{\omega}} \|\mathbf{P}(\boldsymbol{\omega})\widetilde{\mathbf{y}}\|_2^2 \tag{3.23}$$

and seek a power series expansion for $\widetilde{\boldsymbol{\omega}}$ in the small $\mathbf{g}$ limit. The covariance of $\widetilde{\boldsymbol{\omega}}$ corresponds to the previous result, but this expression is useful in selecting compression spaces in the next chapter.

3.3.1 Asymptotic Perturbation

As before, we expand $\widetilde{\boldsymbol{\omega}}$ in a power series: $\widetilde{\boldsymbol{\omega}} = \widehat{\boldsymbol{\omega}} + \epsilon\boldsymbol{\omega}^{(1)} + \mathcal{O}(\epsilon^2)$. We begin with the Taylor expansion of $\mathbf{P}(\boldsymbol{\omega})$ given in [55, Lem. 4.1] using the tensor product notation of Kolda and Bader [90, §2.5],[2]

$$\mathbf{P}(\widehat{\boldsymbol{\omega}} + \epsilon\boldsymbol{\omega}^{(1)}) = \mathbf{P}(\widehat{\boldsymbol{\omega}}) + \epsilon\,\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\omega}})\,\bar{\times}_3\,\boldsymbol{\omega}^{(1)} + \mathcal{O}(\epsilon^2) \quad \text{as} \quad \epsilon \to \mathbf{0}, \tag{3.24}$$

$$\text{where } [\boldsymbol{\mathcal{J}}(\boldsymbol{\omega})]_{\cdot,\cdot,j} = -\left(\mathbf{P}(\boldsymbol{\omega})\frac{\partial\mathbf{V}(\boldsymbol{\omega})}{\partial\omega_j}\mathbf{V}(\boldsymbol{\omega})^{-}\right) - \left(\mathbf{P}(\boldsymbol{\omega})\frac{\partial\mathbf{V}(\boldsymbol{\omega})}{\partial\omega_j}\mathbf{V}(\boldsymbol{\omega})^{-}\right)^{*}. \tag{3.25}$$

From this 3-tensor $\boldsymbol{\mathcal{J}} \in \mathbb{C}^{n\times n\times m}$ we can compute the Jacobian $\mathbf{J}(\boldsymbol{\omega})$ as

$$\mathbf{J}(\boldsymbol{\omega}) = \boldsymbol{\mathcal{J}}(\boldsymbol{\omega})\,\bar{\times}_2\,\widetilde{\mathbf{y}}. \tag{3.26}$$

Next, we apply this expansion to the perturbed measurements $\widetilde{\mathbf{y}} = \mathbf{V}(\widehat{\boldsymbol{\omega}})\widehat{\mathbf{a}} + \epsilon\mathbf{g}$,

$$\mathbf{P}(\widetilde{\boldsymbol{\omega}})\widetilde{\mathbf{y}} = \mathbf{P}(\boldsymbol{\omega})\left[\mathbf{V}(\boldsymbol{\omega})\mathbf{a} + \epsilon\mathbf{g}\right] + \epsilon\left(\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\omega}})\,\bar{\times}_3\,\boldsymbol{\omega}^{(1)}\right)\mathbf{V}(\boldsymbol{\omega})\mathbf{a} + \mathcal{O}(\epsilon^2). \tag{3.27}$$

Immediately we notice $\mathbf{P}(\widehat{\boldsymbol{\omega}})\mathbf{V}(\widehat{\boldsymbol{\omega}})$ vanishes as $\mathbf{P}(\widehat{\boldsymbol{\omega}})$ is a projector onto the space perpendicular to $\mathbf{V}(\widehat{\boldsymbol{\omega}})$. The second term can be rearranged using the rules of tensor multiplication,

$$\left(\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\omega}})\,\bar{\times}_3\,\boldsymbol{\omega}^{(1)}\right)\mathbf{V}(\widehat{\boldsymbol{\omega}})\mathbf{a} = \left(\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\omega}})\,\bar{\times}_2\,(\mathbf{V}(\widehat{\boldsymbol{\omega}})\mathbf{a})\right)\boldsymbol{\omega}^{(1)}.$$

Expanding the product $\boldsymbol{\mathcal{J}}(\boldsymbol{\omega})\,\bar{\times}_2\,\mathbf{V}(\boldsymbol{\omega})\mathbf{a}$,

$$\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\omega}})\,\bar{\times}_2\,\mathbf{V}(\widehat{\boldsymbol{\omega}})\widehat{\mathbf{a}} = \sum_{j=1}^{m} -\mathbf{P}(\widehat{\boldsymbol{\omega}})\frac{\partial\mathbf{V}(\widehat{\boldsymbol{\omega}})}{\partial\omega_j}\mathbf{V}(\widehat{\boldsymbol{\omega}})^{-}\mathbf{V}(\widehat{\boldsymbol{\omega}})\widehat{\mathbf{a}} - \mathbf{V}(\widehat{\boldsymbol{\omega}})^{-*}\left(\frac{\partial\mathbf{V}(\widehat{\boldsymbol{\omega}})}{\partial\omega_j}\right)^{*}\mathbf{P}(\widehat{\boldsymbol{\omega}})\mathbf{V}(\widehat{\boldsymbol{\omega}})\widehat{\mathbf{a}}$$

[2]The symbol $\bar{\times}_j$ denotes the inner product along the jth mode; for example

$$[\boldsymbol{\mathcal{J}}(\boldsymbol{\lambda})\,\bar{\times}_2\,\mathbf{y}]_{j,\ell} = \sum_{k=1}^{n}[\boldsymbol{\mathcal{J}}(\boldsymbol{\lambda})]_{j,k,\ell}\;y_k.$$

we note that the second is zero, as $\mathbf{P}(\widehat{\boldsymbol{\omega}})\mathbf{V}(\widehat{\boldsymbol{\omega}}) = \mathbf{0}$ (this fact leads to Kaufman's simplification). Assuming $\mathbf{V}(\widehat{\boldsymbol{\omega}})$ has full column rank, then $\mathbf{V}(\widehat{\boldsymbol{\omega}})^{-}\mathbf{V}(\widehat{\boldsymbol{\omega}}) = \mathbf{I}$ [27, Thm. 6.3.1] and, consequently,

$$\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\omega}}) \,\bar{\times}_2\, (\mathbf{V}(\widehat{\boldsymbol{\omega}})\widehat{\mathbf{a}}) = -\sum_{j=1}^{m} \mathbf{P}(\widehat{\boldsymbol{\omega}}) \frac{\partial \mathbf{V}(\widehat{\boldsymbol{\omega}})}{\partial \omega_j} \widehat{\mathbf{a}} = -\mathbf{P}(\widehat{\boldsymbol{\omega}})\boldsymbol{\mathcal{V}}(\widehat{\boldsymbol{\omega}}) \,\bar{\times}_2\, \widehat{\mathbf{a}},$$

where $\boldsymbol{\mathcal{V}} \in \mathbb{C}^{n\times n\times m}$ is the gradient of $\mathbf{V}(\boldsymbol{\omega})$: $[\boldsymbol{\mathcal{V}}(\boldsymbol{\omega})]_{:,:,j} = \partial \mathbf{V}(\boldsymbol{\omega})/\partial \omega_j$. Combining all these results, we rewrite (3.27) as

$$\mathbf{P}(\widetilde{\boldsymbol{\omega}})\widetilde{\mathbf{y}} = \epsilon \mathbf{P}(\widehat{\boldsymbol{\omega}})\mathbf{g} - \epsilon \left(\mathbf{P}(\widehat{\boldsymbol{\omega}})\boldsymbol{\mathcal{V}}(\widehat{\boldsymbol{\omega}}) \,\bar{\times}_2\, \widehat{\mathbf{a}}\right) \boldsymbol{\omega}^{(1)} + \mathcal{O}(\epsilon^2). \tag{3.28}$$

The least squares solution for $\boldsymbol{\omega}^{(1)}$ gives the minimum value of (3.23):

$$\boldsymbol{\omega}^{(1)} = -\left(\mathbf{P}(\boldsymbol{\omega})\boldsymbol{\mathcal{V}}(\boldsymbol{\omega}) \,\bar{\times}_2\, \mathbf{a}\right)^{+} \mathbf{P}(\boldsymbol{\omega})\mathbf{g}. \tag{3.29}$$

3.3.2 Covariance

Again, the asymptotic covariance $\widetilde{\boldsymbol{\omega}}$ follows from the first-order perturbation estimate (3.29),

$$\begin{aligned} \mathsf{Cov}\,[\widetilde{\boldsymbol{\omega}}] &= \epsilon^2 \left(\mathbf{P}(\widehat{\boldsymbol{\omega}})\boldsymbol{\mathcal{V}}(\widehat{\boldsymbol{\omega}}) \,\bar{\times}_2\, \widehat{\mathbf{a}}\right)^{+} \mathbf{P}(\widehat{\boldsymbol{\omega}})\, \mathsf{E}[\mathbf{g}\mathbf{g}^*] \mathbf{P}(\widehat{\boldsymbol{\omega}}) \left(\mathbf{P}(\widehat{\boldsymbol{\omega}})\boldsymbol{\mathcal{V}}(\widehat{\boldsymbol{\omega}}) \,\bar{\times}_2\, \widehat{\mathbf{a}}\right)^{+*} + \mathcal{O}(\epsilon^3) \\ &= \epsilon^2 \left(\mathbf{P}(\widehat{\boldsymbol{\omega}})\boldsymbol{\mathcal{V}}(\widehat{\boldsymbol{\omega}}) \,\bar{\times}_2\, \widehat{\mathbf{a}}\right)^{+} \mathbf{P}(\widehat{\boldsymbol{\omega}}) \left(\mathbf{P}(\widehat{\boldsymbol{\omega}})\boldsymbol{\mathcal{V}}(\widehat{\boldsymbol{\omega}}) \,\bar{\times}_2\, \widehat{\mathbf{a}}\right)^{+*} + \mathcal{O}(\epsilon^3). \end{aligned} \tag{3.30}$$

As before, we can further simplify this expression. Writing $\mathbf{A} = \boldsymbol{\mathcal{V}}(\widehat{\boldsymbol{\omega}}) \,\bar{\times}_2\, \widehat{\mathbf{a}}$ and taking the skinny SVD $\mathbf{P}(\widehat{\boldsymbol{\omega}})\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{Z}^*$, then $\mathbf{P}(\widehat{\boldsymbol{\omega}})\mathbf{U} = \mathbf{U}$ as $\mathbf{U}$ is in the range of $\mathbf{P}(\widehat{\boldsymbol{\omega}})\mathbf{A}$ which is in the range of $\mathbf{P}(\widehat{\boldsymbol{\omega}})$, and

$$\begin{aligned} (\mathbf{P}(\widehat{\boldsymbol{\omega}})\mathbf{A})^{+}\mathbf{P}(\widehat{\boldsymbol{\omega}})(\mathbf{P}(\widehat{\boldsymbol{\omega}})\mathbf{A})^{+*} &= \mathbf{Z}\boldsymbol{\Sigma}^{+}\mathbf{U}^*\mathbf{P}(\widehat{\boldsymbol{\omega}})\mathbf{U}\boldsymbol{\Sigma}^{+}\mathbf{Z}^* = \mathbf{Z}\boldsymbol{\Sigma}^{+2}\mathbf{Z}^* \\ &= \left[(\mathbf{P}(\widehat{\boldsymbol{\omega}})\mathbf{A})^*(\mathbf{P}(\widehat{\boldsymbol{\omega}})\mathbf{A})\right]^{-1} = \left[\mathbf{A}^*\mathbf{P}(\widehat{\boldsymbol{\omega}})\mathbf{A}\right]^{-1}. \end{aligned} \tag{3.31}$$

This covariance estimate is naively below the Cramér-Rao bound if we only consider $\boldsymbol{\omega}$; then $\mathcal{I}(\widehat{\boldsymbol{\omega}}) = \mathbf{A}^*\mathbf{A}$ and $\mathbf{A}^*\mathbf{A} \succeq \mathbf{A}^*\mathbf{P}(\widehat{\boldsymbol{\omega}})\mathbf{A}$. However, this erroneously neglects

the latent variables $\mathbf{a}$, assuming they are constant. Correctly computing the Cramér-Rao bound shows $\mathsf{Cov}\,\widetilde{\boldsymbol{\omega}}$ obtains this bound. Considering the general nonlinear least squares covariance (3.20), taking $\boldsymbol{\theta} = [\boldsymbol{\omega}, \mathbf{a}]$, the gradient is

$$\mathbf{F}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\mathcal{V}}(\boldsymbol{\omega}) \,\bar{\times}_2\, \mathbf{a} & \mathbf{V}(\boldsymbol{\omega}) \end{bmatrix}.$$

Adopting the shorthand for $\mathbf{A}$ above, the covariance splits into four blocks:

$$\begin{aligned} \mathsf{Cov}\,\widetilde{\boldsymbol{\theta}} &= \epsilon^2 \left[\mathbf{F}(\widehat{\boldsymbol{\theta}})^* \mathbf{F}(\widehat{\boldsymbol{\theta}})\right]^{-1} + \mathcal{O}(\epsilon^3) \\ &= \epsilon^2 \begin{bmatrix} \mathbf{A}^*\mathbf{A} & \mathbf{A}^*\mathbf{V} \\ \mathbf{V}^*\mathbf{A} & \mathbf{V}^*\mathbf{V} \end{bmatrix}^{-1} + \mathcal{O}(\epsilon^3) = \epsilon^2 \begin{bmatrix} (\mathbf{A}^*\mathbf{P}_{\mathbf{V}}^{\perp}\mathbf{A})^{-1} & * \\ * & (\mathbf{V}^*\mathbf{P}_{\mathbf{A}}^{\perp}\mathbf{V})^{-1} \end{bmatrix} + \mathcal{O}(\epsilon^3), \end{aligned} \tag{3.32}$$

where $\mathbf{P}_{\mathbf{A}}^{\perp}$ is the orthogonal projector onto the space perpendicular to the range of $\mathbf{A}$. We immediately identify the $(1,1)$ block with the covariance computed above, as $\mathbf{P}(\widehat{\boldsymbol{\omega}})$ is, by definition, $\mathbf{P}_{\mathbf{V}}^{\perp}$.

3.4 Heuristics for Initial Conditions

The main disadvantage of nonlinear least squares methods, when compared to Prony-type methods, is their requirement for initial estimates of $\boldsymbol{\omega}$. Without good initial estimates of $\boldsymbol{\omega}$, the optimization of $\boldsymbol{\omega}$ can easily be lead astray into spurious local minima. When there is a single exponential without noise, the minimum is very clear (see Figure 3.4) and convergence to the right minimum is typical regardless of the initial estimate. However, as Figure 3.3 illustrates, as few as two exponentials and small amounts of noise can cause poor initial estimates to converge to a spurious minium. We can see the appearance of spurious minima in Figure 3.2; there is even a local minimum in the absence of noise at the average of $\boldsymbol{\omega}$. Spurious local minima become increasingly problematic as the number of exponentials increases. For example in Ta-

Table 3.1 : Failure rate for increasing numbers of exponentials p. The table shows the precentage of failures from 100 random trials with $n = 10^3$, where samples $\omega_j \in [-0.1, 0] \times [0, 2\pi)i$ uniformly, and $|a_j| = 1$ with uniform random phase.

p	Roots of Unity	Largest Fourier	Peeling
1	13	0	0
2	23	12	0
3	43	20	0
4	56	31	0
5	58	36	0
6	68	45	0
7	73	66	0
8	71	61	1
9	76	57	0
10	81	73	0

ble 3.1, using the p roots of unity as initial estimates causes convergence to spurious minima with increasing probability as the number of exponentials, p, increases.

Good initial estimates of $\boldsymbol{\omega}$ combat this, helping the nonlinear least squares solver converge to the true parameter values by placing the initial estimate near the true minimum. Assuming the number of exponentials is already known, we examine one of five different strategies for choosing the initial estimate $\boldsymbol{\omega}_0$ for the exponential fitting problem: choosing $\boldsymbol{\omega}_0$ randomly, setting $\boldsymbol{\omega}_0$ to be the pth roots of unity, picking $\boldsymbol{\omega}_0$ as peaks of the Fourier transform, solving a Prony-type method to provide $\boldsymbol{\omega}_0$, and by using prior knowledge about $\boldsymbol{\omega}$.

3.4.1 Random Imaginary Part

The easiest approach for choosing the initial estimate $\boldsymbol{\omega}_0$ is to randomly guess. For example, choosing each $\boldsymbol{\omega}_0$ independently and uniformly distributed in $[-1, 0] \times [0, 2\pi)i$

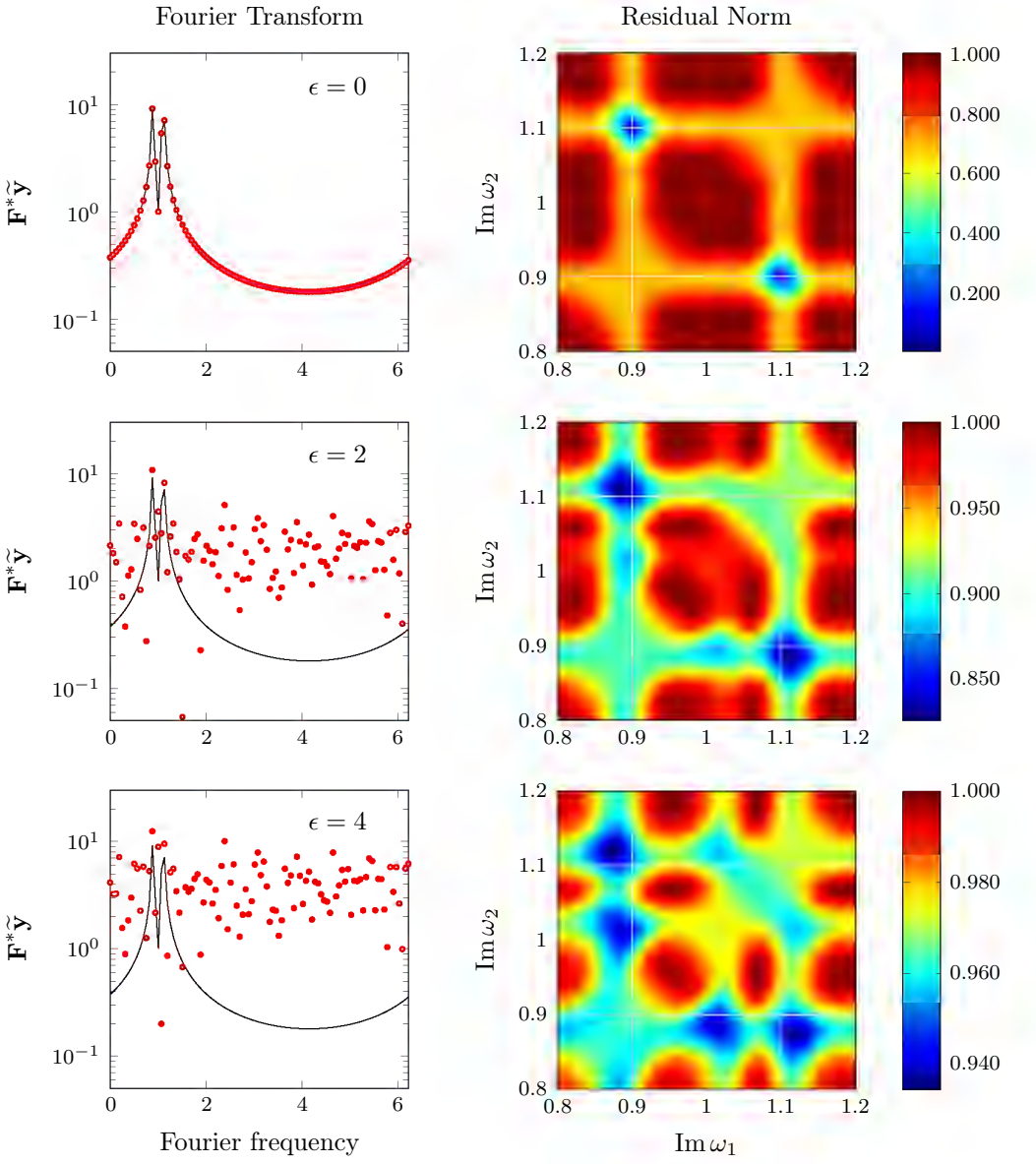


Figure 3.2 : The appearance of local minima as noise increases. For fixed random vector $\mathbf{g}$, we perturb $\widetilde{\mathbf{y}} = \mathbf{y} + \epsilon\mathbf{g}$. The left plots show the Fourier transform of the noisy signal $\mathbf{F}^*\widetilde{\mathbf{y}}$ as red dots and the original unperturbed transform as a black line. The right plots show the norm of the residual $\|\mathbf{P}(\boldsymbol{\omega})\widetilde{\mathbf{y}}\|_2$. As ϵ increases, the true minimum becomes shallower and spurious minima appear. (Note the changing colorbar scale.)

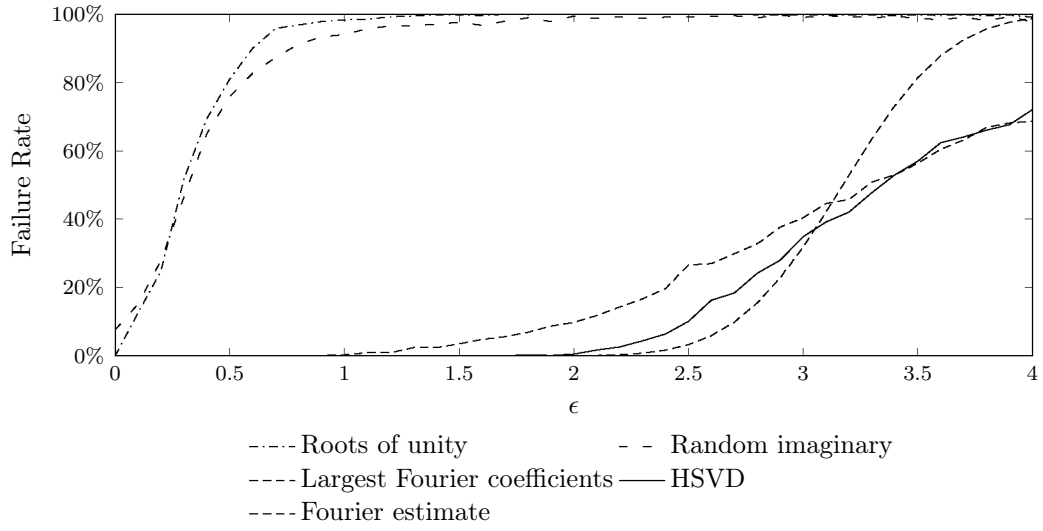


Figure 3.3 : Failure rate for initial estimate heuristics. The unperturbed data consisted of $n = 100$ measurements of $\widehat{\boldsymbol{\omega}} = [0.9i,\ 1.1i]$ with $\widehat{\mathbf{a}} = [1,\ 1]$. For each realization of noise $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, a trust region Variable Projection algorithm attempted to recover $\boldsymbol{\omega}$ from $\widetilde{\mathbf{y}} = \mathbf{y} + \epsilon\mathbf{g}$ starting with the initial estimate given by one of the four methods. Fourier estimate referrs to probability that the largest peaks in the Fourier transform correspond to noise as given by (3.34). The initial conditions are considered to fail if the recovered $\widetilde{\boldsymbol{\omega}}$ falls ten times further away from $\widehat{\boldsymbol{\omega}}$ than the recovered $\widetilde{\boldsymbol{\omega}}$ starting from the true parameters $\widehat{\boldsymbol{\omega}}$ did from $\widehat{\boldsymbol{\omega}}$. There were 10^3 trials for each value of ϵ.

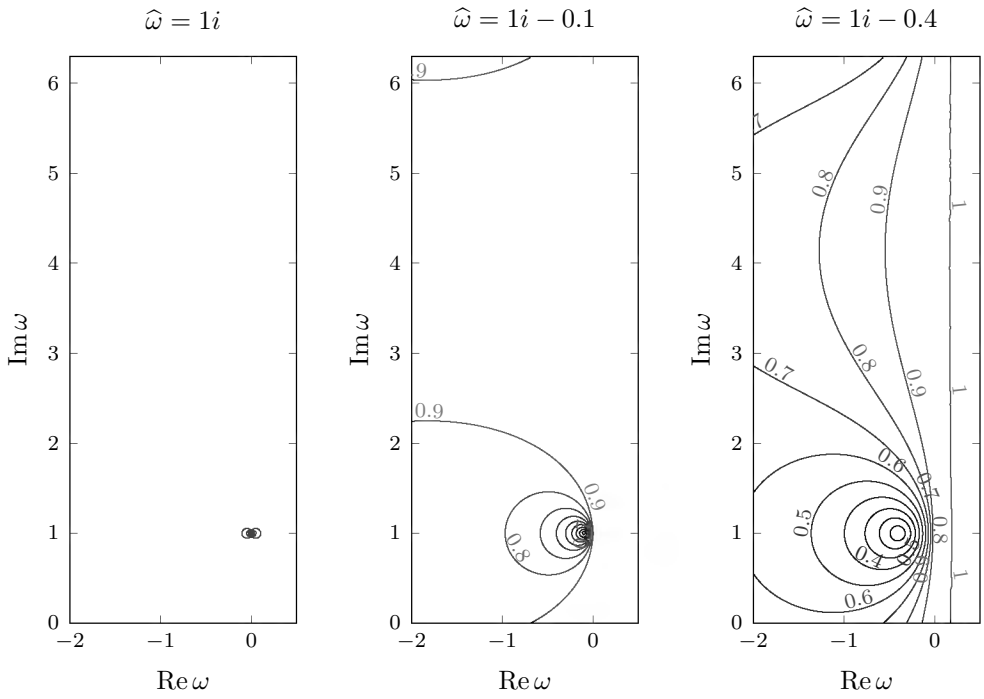


Figure 3.4 : The norm of the residual for a single exponential. Each plot shows contours of $\|\mathbf{P}(\omega)\mathbf{V}(\widehat{\omega})\|_2/\|\mathbf{V}(\widehat{\omega})\|_2$ with spacing 0.1.

of the complex plane. However, if $\boldsymbol{\omega}_0$ ventures too far away from the imaginary axis, then $\mathbf{V}(\boldsymbol{\omega}_0)$ is ill-conditioned and can lead to erroneous Newton steps. Enforce $\operatorname{Re}\boldsymbol{\omega} = \mathbf{0}$ can avoid this form of ill-conditioning, but $\mathbf{V}(\boldsymbol{\omega}_0)$ can still be ill-conditioned if $|[\boldsymbol{\omega}_0]_j - [\boldsymbol{\omega}_0]_k|$ is small. Hence, this approach is prone to failure. As Figure 3.3 shows, even when there is no noise, randomly choosing $\operatorname{Im}\boldsymbol{\omega}_0$ leads to convergence away from $\widehat{\boldsymbol{\omega}}$ in 7% of the simulations.

3.4.2 Roots of Unity

Osborne notes that his method implicitly uses the initial estimate that $\boldsymbol{\omega}$ are the pth roots of unity [116], i.e., $[\boldsymbol{\omega}_0]_j = 2\pi ij/p$ for $j = 0, \ldots, p-1$. Unlike randomly choosing $\boldsymbol{\omega}_0$, choosing the roots of unity makes $\mathbf{V}(\boldsymbol{\omega}_0)$ well conditioned, as its columns are approximately orthogonal (exactly orthogonal if n is an integer multiple of p). However, this approach fails even in the absence of noise; Table 3.1 provides one example. Once noise is introduced, this approach provides even worse initial estimates; for example, in Figure 3.3. The origin of the failure is simple; $\boldsymbol{\omega}_0$ might be far away from true value $\widehat{\boldsymbol{\omega}}$, placing $\boldsymbol{\omega}_0$ such that optimization converges to local minimum near $\boldsymbol{\omega}_0$ rather than finding the minimum associated with $\widehat{\boldsymbol{\omega}}$.

3.4.3 Peaks of the Fourier Transform

A well known fact in signal processing is that peaks of the Fourier transform of $\mathbf{y}$ correspond to the imaginary part of $\boldsymbol{\omega}$. This is a result of the geometric sum formula applied to $\mathbf{F}^*\mathbf{V}(\omega)$ (cf., (4.37)):

$$[\mathbf{F}^*\mathbf{V}(\omega)]_k = \frac{1}{\sqrt{n}} \frac{1 - e^{\omega n - 2\pi ik}}{1 - e^{\omega - 2\pi ik/n}} \quad \omega \neq 2\pi ik/n. \tag{3.33}$$

There are 'peaks' in $\mathbf{F}^*\mathbf{V}(\omega)$ near where the denominator vanishes, $\operatorname{Im}\omega \approx 2\pi ik/n$. As $\mathbf{y} = \mathbf{V}(\widehat{\boldsymbol{\omega}})\widehat{\mathbf{a}}$, then $\mathbf{F}^*\mathbf{y}$ is a sum of these peaks; i.e., $\mathbf{F}^*\mathbf{y} = \mathbf{F}^*\mathbf{V}(\widehat{\boldsymbol{\omega}})\widehat{\mathbf{a}}$. (See

Figure 3.5 for an illustration.) If $\omega = 2\pi ik/n$, then the peak will be a single large entry with all others in $\mathbf{F}^*\mathbf{V}(\omega)$ zero; otherwise the peak will be broad, and widen as as $|\mathrm{Re}\,\omega|$ grows. The problem with this approach is the difficulty of identifying peaks. The largest entry of $\mathbf{F}^*\mathbf{y}$ certainly qualifies, but what about the second? It could either be part of the tail of the first peak, or a peak in its own right. Moreover if two ω_k values are close, they can appear to be one peak instead of two. Figure 3.5 illustrates some of these pathologies. Even with these pathologies, this approach provides good initial estimates in the numerical experiment shown in Figure 3.3. There, as $\widehat{\omega}_1$ and $\widehat{\omega}_2$ are sufficiently separated and of roughly equal magnitude, we simply take the largest two entries of $\mathbf{F}^*\widetilde{\mathbf{y}}$ to form $\boldsymbol{\omega}_0$.

Choosing the initial estimates based on the largest Fourier coefficients works well as the largest entry is robust to noise if damping is light. We can estimate the failure rate of this method by assuming the method fails if the smallest peak is surpassed by another entry in $\mathbf{F}^*\widetilde{\mathbf{y}}$. Assuming noise is uniform, $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \epsilon^2\mathbf{I})$, then $\mathbf{F}^*\mathbf{g}$ also follows the same distribution $\mathbf{F}^*\mathbf{g} = \mathbf{h} \sim \mathcal{N}(\mathbf{0}, \epsilon^2\mathbf{I})$. Thus, we have the probability of failure

$$p_{\text{fail}} = p\left(|[\mathbf{F}^*\mathbf{y}]_{k'} + h_{k'}| \geq |[\mathbf{F}^*\mathbf{y}]_k + h_k| : \ \forall k' \neq k\right).$$

Ideally, we could evaluate this using the *Rice distribution*[3], but for simplicity, we use the crude estimate – that the size of the non-peak entries is trivial and noise does not affect the size of the peak yielding the approximation

$$p_{\text{fail}} \approx p\left(|\widehat{g}_{k'}| \geq |[\mathbf{F}^*\mathbf{y}]_k| : \ \forall k' \neq k\right) = 1 - (1 - e^{-|[\mathbf{F}^*\mathbf{y}]_k|^2/(2\sigma^2)})^{n-1}. \tag{3.34}$$

This estimates the true value where the largest Fourier entries cease to provide good

[3]The Rice distribution describes the distribution of $|z|$ if z a complex normally distributed random variable $z \sim \mathcal{N}(a, \epsilon^2)$; see [130],[158].

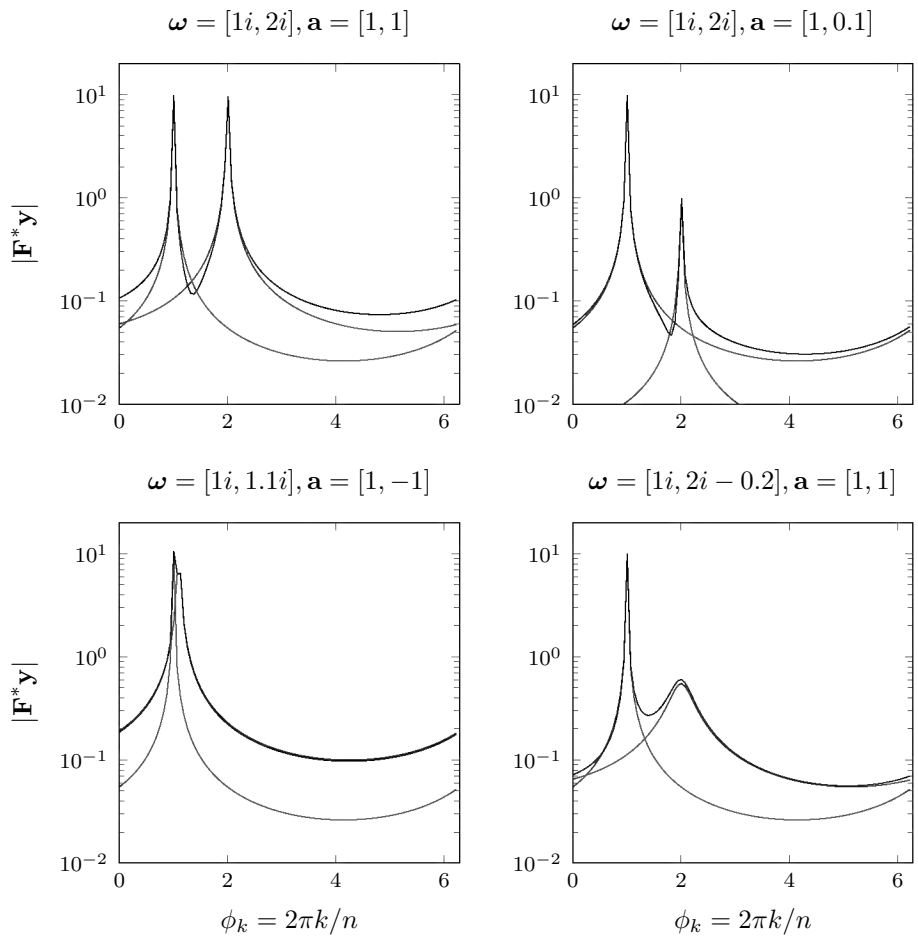


Figure 3.5 : Examples of peaks in the Fourier transform. Clockwise from the upper left: the ideal case with well separated, equal magnitude peaks; well separated peaks of different amplitude; two close peaks; and a broad peak resulting from strong decay. In these examples $n = 100$. The grey curves show $\mathbf{F}^*\mathbf{V}(\omega_1)$ and $\mathbf{F}^*\mathbf{V}(\omega_2)$ while the black curve shows $\mathbf{F}^*\mathbf{V}(\boldsymbol{\omega})\mathbf{a} = \mathbf{F}^*\mathbf{y}$.

initial estimates to within a factor of two, as seen in Figure 3.3, and over estimates the failure rate as noise increases.

Finally note we note this method must produce a well conditioned $\mathbf{V}(\boldsymbol{\omega}_0)$. The initial estimates $\boldsymbol{\omega}_0$ are Fourier frequencies, and hence, $\mathbf{V}(\boldsymbol{\omega}_0)$ consists of the orthogonal columns of $\mathbf{F}$.

3.4.4 Prony-type Methods

Since Prony-type methods do not require initial estimates, they provide an obvious source for the initial estimates needed by optimization routines. This second computation of the exponentials might seem redundant, but it can be beneficial when using one of the compressed Prony variants from Sections 2.8 and 2.8. We can easily obtain fast, crude, estimates of $\boldsymbol{\omega}_0$ using the Filtered Matrix Pencil method in a region indicated by peaks in the Fourier transform.

Although we are the first to propose use of a compressed Prony-type method, the idea had its origin's in Householder's Method, where the full Prony's Method was used to provide initial estimates for the maximum likelihood method for lag coefficients $\boldsymbol{\alpha}$ [69]. Others would later use the High Order Yule-Walker method [145] to provide initial estimates for a Newton method for optimizing $\boldsymbol{\omega}$.

3.4.5 Prior Knowledge

Exponential fitting rarely occurs in the absence of other information about the underlying problem. This frequently yields *prior knowledge* about $\boldsymbol{\omega}$ from the structure of the problem. For example, in the case of damped wave equation in one dimension, we know the imaginary part of $\boldsymbol{\omega}$ are approximately integer multiples of some fundamental frequency, while the real part of $\boldsymbol{\omega}$ is negative. For other lightly damped

mechanical systems like drums, beams, or plates we have similar knowledge about the structure of $\operatorname{Im}\boldsymbol{\omega}$. In general, we also may assume that $\operatorname{Re}\boldsymbol{\omega}$ is in the left half plane as a result of conservation of energy.

3.5 Estimating the Number of Exponentials

Although these five methods are sufficient to provide initial estimates of $\boldsymbol{\omega}$, we still need to know the size of $\boldsymbol{\omega}$, i.e., p, the number of exponentials present. There are three standard approaches for determining p: one motivated by the Hankel construction seen Section 2.2.3, one statistically motivated, and one by the size of a_j, The last two approaches are conducive to an approach called *peeling*, where exponentials are appended sequentially until no longer justified (see Section 3.6).

3.5.1 Hankel Singular Values

In Section 2.2.3, we saw that the Hankel matrix of measurements $[\mathbf{H}]_{j,k} = y_{j+k}$ can be decomposed as $\mathbf{H} = \mathbf{V}(\boldsymbol{\omega})\operatorname{diag}(\mathbf{a})\mathbf{V}(\boldsymbol{\omega})^\top$. Hence, the range of $\mathbf{H}$ is the range of $\mathbf{V}(\boldsymbol{\omega})$ and thus $\mathbf{H}$ has exactly p non-zero singular values (provided $\boldsymbol{\omega}$ has distinct entries). This leads to a widely used heuristic: choose p as the number of large singular values. However noise increases the zero singular values, threatening the ability distinguish the large singular values revealing p from singular values associated with noise. The following new result allow us to bound the growth of the singular values associated with noise. The first result bounds the perturbation of singular values, obtaining a bound on the norm of a finite dimensional Hankel Matrix using a method reminiscent of that used for the infinite dimensional case [21, p. 13]. The second result provides a bound on the expected value of the norm of this random Hankel matrix.

Theorem 3.1. *If* $\mathbf{H}, \widetilde{\mathbf{H}} \in \mathbb{C}^{m\times n}$ *are both Hankel matrices where* $\mathbf{E} := \widetilde{\mathbf{H}} - \mathbf{H}$ *has*

entries $[\mathbf{E}]_{j,k} = g_{j+k}$*, then the singular values* $\widetilde{\sigma}_j$ *and* σ_j *of* $\widetilde{\mathbf{H}}$ *and* $\mathbf{H}$ *obey*

$$|\widetilde{\sigma}_j - \sigma_j| \leq \sqrt{n+m-1}\ \max_k |[\mathbf{F}_{n+m}\mathbf{g}]_k|. \tag{3.35}$$

Proof. From Weyl's Theorem [143, Cor. 4.9], we have

$$|\widetilde{\sigma}_j - \sigma_j| \leq \|\widetilde{\mathbf{H}} - \mathbf{H}\|_2 = \|\mathbf{E}\|_2.$$

We bound the norm of $\mathbf{E}$ by embedding $\mathbf{E}$ inside a circulant matrix $\mathbf{C}$ with flipped columns

$$\mathbf{C} = \begin{bmatrix} g_{m-1} & \cdots & g_0 & g_{n+m-2} & \cdots & g_m \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ g_{n+m-2} & \cdots & g_{n-1} & g_{n-2} & \cdots & g_0 \\ g_0 & \cdots & g_n & g_{n-1} & \cdots & g_1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ g_{m-2} & \cdots & g_{n+m-2} & g_0 & \cdots & g_{n-1} \end{bmatrix} \in \mathbb{C}^{(n+m-1)\times(n+m-1)}. \tag{3.36}$$

Using the definition of the ℓ_2 norm,

$$\|\mathbf{E}\|_2 = \max_{\mathbf{x}\in\mathbb{C}^n} \frac{\|\mathbf{E}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \max_{\substack{\mathbf{x}\in\mathbb{C}^{n+m} \\ \mathbf{x}_j = 0\ \forall j \geq n}} \frac{\|\mathbf{C}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \|\mathbf{C}\|_2.$$

Circulant matrices are diagonalized by the Fourier transform with eigenvalues $\boldsymbol{\lambda} = \sqrt{n+m-1}\mathbf{F}_{n+m-1}\mathbf{c}$, where $\mathbf{c}$ is the first column of $\mathbf{C}$; i.e., $\mathbf{C} = \mathbf{F}^*_{n+m-1}\boldsymbol{\Lambda}\mathbf{F}_{n+m-1}$ for $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$. As $\mathbf{F}_{n+m-1}$ is unitary then,

$$\|\mathbf{C}\|_2 = \|\boldsymbol{\Lambda}\|_2 = \max_k |\lambda_k|.$$

Finally, as we are only interested in the magnitude of λ_k, we can apply the DFT shift theorem to write λ_k in terms of the Fourier transform of $\mathbf{g} \in \mathbb{C}^{n+m-1}$. If

$\mathbf{D}_\ell \in \mathbb{C}^{(n+m-1)\times(n+m-1)}$ is a diagonal matrix with $[\mathbf{D}_\ell]_{k,k} = e^{2\pi i k\ell/(n+m-1)}$ and $\mathbf{S}$ is the circulant shift matrix $[\mathbf{Sg}]_k = g_{k-1 \bmod n+m-1}$, then the DFT shift theorem implies

$$[\mathbf{F}_{n+m-1}\mathbf{S}_\ell\mathbf{x}]_k = [\mathbf{D}_\ell\mathbf{F}_{n+m-1}\mathbf{x}]_k.$$

Hence, taking $\ell = n-1$,

$$\max_k |\lambda_k| = \sqrt{n+m-1}\max_k |[\mathbf{D}_\ell\mathbf{F}_{n+m-1}\mathbf{c}]_k| = \sqrt{n+m-1}\max_k |[\mathbf{F}_{n+m-1}\mathbf{g}]_k|.$$

□

Using this theorem, the following Corollary evaluates the expected value of this bound on $|\widetilde{\sigma}_j - \sigma_j|$ when $\mathbf{g}$ is a complex random vector.

Corollary 3.1. *If* $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \epsilon^2\mathbf{I})$*, then*

$$|\widetilde{\sigma}_j - \sigma_j| \leq \sqrt{n+m-1}\,\gamma \quad \textit{with probability} \quad p(\gamma) = (1 - e^{-\gamma^2/(2\epsilon^2)})^{n+m}. \tag{3.37}$$

Proof. As $\mathbf{g}$ is normally distributed with mean $\mathbf{0}$ and covariance $\sigma^2\mathbf{I}$ and $\mathbf{F}_{n+m-1}$ is a unitary transform, then $\mathbf{h} = \mathbf{F}_{n+m-1}\mathbf{g}$ is also normally distributed with mean $\mathbf{0}$ and covariance $\epsilon^2\mathbf{I}$. Then $|h_k|$ follows a follows a *Rayleigh distribution*[4] with covariance ϵ^2 such that $p(h_k < x) = 1 - e^{-x^2/(2\epsilon^2)}$. As the h_k are independent, the probability that $n+m-1$ of these are bounded above by γ is given by $p(h_k < \gamma)^{n+m-1}$. □

Using this criteria, singular values associated with noise $\mathbf{g}$ using a total of n measurements are bounded above by

$$\widetilde{\sigma}_k < \epsilon\sqrt{-2n\log(1 - \sqrt[n]{r})} \quad k > p \tag{3.38}$$

with probability r. Thus p can be estimated as the number of singular values breaking this bound. Figure 3.6 provides an example of this bound in practice, illustrating its power at distinguishing signal from noise.

[4] A special case of the χ distribution with two degrees of freedom.

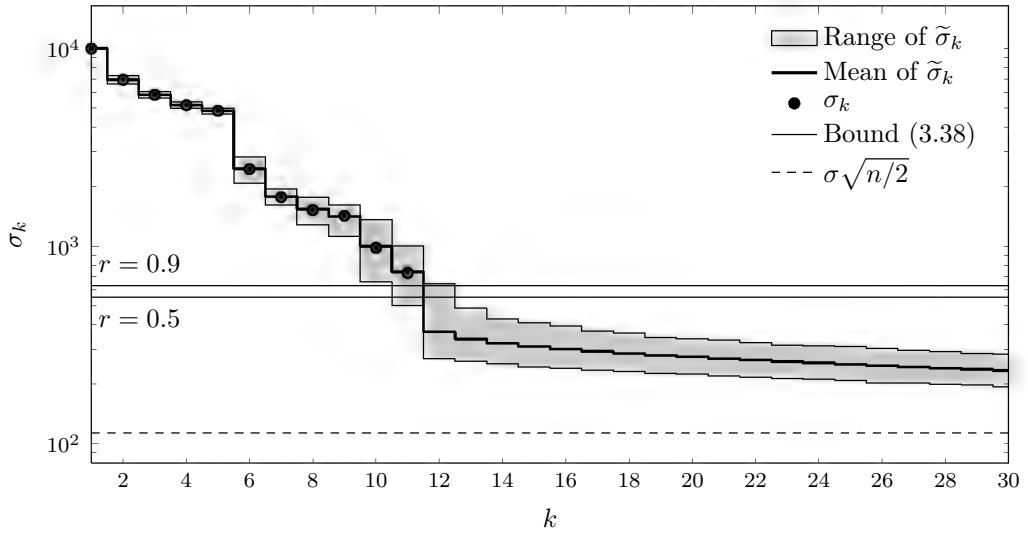


Figure 3.6 : Hankel singular values from a magnetic resonance spectroscopy example with noise. Using the example from [153, Table 1] with $p = 11$, $n = 256$, and noise $\sigma = 10$ for 10^5 trials, we show the maximum, minimum, and mean singular values $\widetilde{\sigma}_k$ of $\widetilde{\mathbf{H}}$. The bound from Corollary 3.1 (black line) successfully distinguishes a significant singular value $\widetilde{\sigma}_{11}$ from a noise singular value $\widetilde{\sigma}_{12}$ with high probability. However the approach of de Groen and de Moor (dashed line) does a poor job of bounding σ_{12}.

An alternative approach of de Groen and de Moor [39, §4] provides a less effective estimate. Their approach estimates a bound on the singular values associated noise by considering the eigenvalues of $\mathsf{E}[\widetilde{\mathbf{H}}^*\widetilde{\mathbf{H}}]$:

$$\mathsf{E}[\widetilde{\mathbf{H}}^*\widetilde{\mathbf{H}}] = \mathbf{H}^*\mathbf{H} + \mathsf{E}[\mathbf{E}^*\mathbf{H}] + \mathsf{E}[\mathbf{H}^*\mathbf{E}] + \mathsf{E}[\mathbf{E}^*\mathbf{E}] = \mathbf{H}^*\mathbf{H} + n\epsilon^2\mathbf{I},$$

where n is the number of rows in $\mathbf{H}$. As the singular values are eigenvalues of $\widetilde{\mathbf{H}}^*\widetilde{\mathbf{H}}$, we note $\widetilde{\sigma}_j^2 \approx \sigma_j^2 + n\epsilon^2$. This approach is performs poorly, as the estimate computes the eigenvalues of $\mathsf{E}[\widetilde{\mathbf{H}}^*\widetilde{\mathbf{H}}]$ rather than the eigenvalues of $\widetilde{\mathbf{H}}^*\widetilde{\mathbf{H}}$,

These results about the singular values of $\mathbf{H}$ also apply to systems theory. There, y_j corresponds to the jth Markov parameter of a discrete time linear system with system matrices $\mathbf{A} = \mathrm{diag}(\boldsymbol{\lambda})$ where $\boldsymbol{\lambda} = e^{\boldsymbol{\omega}}$, $\mathbf{C} = 1^\top$ (the vector of ones), and $\mathbf{B} = \mathrm{diag}(\mathbf{a})$. In this context $\mathbf{H}$ corresponds to the product of the controllability

and observability matrices [83, §2.2, eq. 49]. More sophisticated approaches for determining p using SVD are still being developed, often under the name of rank determination (after the rank of $\mathbf{A}$); see, e.g., [165].

3.5.2 Akaike Information Criterion

The advantage of the Hankel singular value approach is that we can estimate the number of exponentials p without determining $\boldsymbol{\omega}$ or $\mathbf{a}$. Model selection criteria provide another approach that allows us to compare models with different numbers of exponentials provided knowledge of maximum likelihood estimates of $\boldsymbol{\omega}$ and $\mathbf{a}$. Another approach is to use model selection criteria to compare models with different numbers of exponentials. One such model selection criteria is the *Akaike Information Criterion* (AIC) [26, §2.2] that ranks candidate models with q parameters by

$$\mathcal{A}(q) = 2q - 2\log\left(\max_{\boldsymbol{\theta}\in\mathbb{R}^q} \mathcal{L}_q(\boldsymbol{\theta})\right), \tag{3.39}$$

where $\mathcal{L}_q$ is the likelihood function associated with a q parameter model; cf., (A.10). The best model has a value of q that minimizes $\mathcal{A}(q)$. The AIC is an asymptotic result in the limit of $n \to \infty$ measurements; for a finite number of samples n, there is a second order correction, called the AICc due to Sugiura [148]

$$\mathcal{A}_c(q) = \mathcal{A}(q) + \frac{2q(q+1)}{n-q-1}. \tag{3.40}$$

For real measurements $\widetilde{\mathbf{y}}$ the there are $2p$ parameters ($\boldsymbol{\omega} \in \mathbb{R}^p$ and $\mathbf{a} \in \mathbb{R}^p$) and if $\widetilde{\mathbf{y}} = \mathbf{y} + \mathbf{g}$, where $\mathbf{g} \sim (\mathbf{0}, \boldsymbol{\Sigma})$ (real), the AIC is

$$\begin{aligned}\mathcal{A}(p) &= 2p - 2\log\left[\max_{\boldsymbol{\theta}\in\mathbb{R}^{2p}} \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}} e^{-(\widetilde{\mathbf{y}}-\mathbf{f}(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}^{-1}(\widetilde{\mathbf{y}}-\mathbf{f}(\boldsymbol{\theta}))/2}\right] \\ &= 2p + n\log 2\pi + \log\det\boldsymbol{\Sigma} + \min_{\boldsymbol{\theta}\in\mathbb{R}^{2p}} \|\widetilde{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Sigma}}.\end{aligned}$$

As we are interested in only the smallest value of $\mathcal{A}$, we neglect the terms that do not depend on p terms and compute

$$\underline{\mathcal{A}}(p) = 2p + \min_{\boldsymbol{\theta}\in\mathbb{R}^{2p}} \|\widetilde{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta})\|_{\boldsymbol{\Sigma}}^2. \tag{3.41}$$

If noise is independent and identically distributed, then $\boldsymbol{\Sigma} = \epsilon^2\mathbf{I}$ and

$$\underline{\mathcal{A}}(p) = 2p + \epsilon^{-2} \min_{\boldsymbol{\theta}\in\mathbb{R}^{2p}} \|\widetilde{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta})\|_2^2.$$

The correction for the AICc for real measurements is $4p(2p+1)/(n-2p-1)$.

For complex measurements there are $4p$ parameters and $2n$ measurements; similar results follow. The AIC is

$$\begin{aligned}\mathcal{A}(p) &= 4p - 2\log\left[\max_{\boldsymbol{\theta}\in\mathbb{C}^p} \frac{1}{\pi^n \det\boldsymbol{\Sigma}} e^{-(\widetilde{\mathbf{y}}-\mathbf{f}(\boldsymbol{\theta})^\top)\boldsymbol{\Sigma}^{-1}(\widetilde{\mathbf{y}}-\mathbf{f}(\boldsymbol{\theta}))}\right] \\ &= 4p + 2n\log\pi + 2\log\det\boldsymbol{\Sigma} + 2\min_{\boldsymbol{\theta}} \|\widetilde{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta})\|_{\boldsymbol{\Sigma}}^2.\end{aligned}$$

Neglecting the constant terms yields

$$\underline{\mathcal{A}}(p) = 4p + 2\min_{\boldsymbol{\theta}} \|\widetilde{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta})\|_{\boldsymbol{\Sigma}}^2. \tag{3.42}$$

This is a scaled version of (3.41), but the correction for the AICc for complex data is not a scalar multiple of the real case. The AICc correction for complex data is $8p(4p+1)/(2n-4p-1)$.

3.5.3 Small Amplitudes

A strategy often used by practitioners simply fits an excessive number of exponentials, and then removes the exponentials with amplitudes below some tolerance τ; see, e.g., [123, 126]. This avoids the complicated machinery of the singular values of the Hankel matrix $\mathbf{H}$ or the AIC, but provides only a crude heuristic. In this section we provide a truncation tolerance τ based on the AIC.

Suppose we have measurements $\widetilde{\mathbf{y}} = \mathbf{V}(\widehat{\boldsymbol{\omega}})\widehat{\mathbf{a}} + \epsilon\mathbf{g}$ where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\widehat{\boldsymbol{\omega}}, \widehat{\mathbf{a}} \in \mathbb{C}^p$ and we then fit $\widetilde{\mathbf{y}}$ with $p+p'$ exponentials. The AIC dictates that the $p+p'$ exponential is a better model if $\underline{\mathcal{A}}(p+p') - \underline{\mathcal{A}}(p) < 0$.

$$\begin{aligned}\underline{\mathcal{A}}(p+p') - \underline{\mathcal{A}}(p) &= 4p' + 2\epsilon^{-2} \min_{\boldsymbol{\omega},\mathbf{a}\in\mathbb{C}^{p+p'}} \|\widetilde{\mathbf{y}} - \mathbf{V}(\boldsymbol{\omega})\mathbf{a}\|^2 - 2\epsilon^{-2} \min_{\boldsymbol{\omega},\mathbf{a}\in\mathbb{C}^{p}} \|\widetilde{\mathbf{y}} - \mathbf{V}(\boldsymbol{\omega})\mathbf{a}\|^2 \\ &\leq 4p' + 2\epsilon^{-2} \min_{\boldsymbol{\omega},\mathbf{a}\in\mathbb{C}^{p'}} \|\mathbf{g} - \mathbf{V}(\boldsymbol{\omega})\mathbf{a}\|^2 - 2\epsilon^{-2}\|\mathbf{g}\|^2.\end{aligned}$$

Fixing the minimum of the p' sized problem as $\widetilde{\boldsymbol{\omega}}, \widetilde{\mathbf{a}}$, we invoke the reverse triangle inequality to find

$$\underline{\mathcal{A}}(p+p') - \underline{\mathcal{A}}(p) \leq 4p' - 2\epsilon^{-2}\|\mathbf{V}(\widetilde{\boldsymbol{\omega}})\widetilde{\mathbf{a}}\|^2.$$

Setting $p' = 1$, we find the $p+1$ exponentials are not justified according to the AIC if

$$|\widetilde{a}| < \frac{\sqrt{2}\epsilon}{\|\mathbf{V}(\widetilde{\omega})\|}. \tag{3.43}$$

If we require $\operatorname{Re}\omega_j = 0$, then $|\widetilde{a}|$ must be less than ϵ/n (we loose a factor of two since there are now half as many parameters). This yields the truncation tolerance $\tau = \epsilon/n$. For general $\widetilde{\omega}$, $\|\mathbf{V}(\widetilde{\omega})\|$ may be as small as one, hence we require $|\widetilde{a}|$ to be less than $\sqrt{2}\epsilon$, yielding the truncation tolerance $\tau = \sqrt{2}\epsilon$. A similar analysis would yield truncation tolerances where the error $\mathbf{g}$ is a result of rounding errors.

3.6 Peeling

Now we combine these elements – initial estimates of $\boldsymbol{\omega}$ in Section 3.4 and the estimation of the number of exponentials in Section 3.5 – into a single algorithm. Although each of these techniques is not new, the combination is, and allows us to use nonlinear least squares algorithms without initial estimates or a priori knowledge of p. This

Algorithm 3.2: Peeling for Exponential Fitting

Input : Noisy measurements $\widetilde{\mathbf{y}} \in \mathbb{C}^n$, noise estimate ϵ

Output: Exponential coefficients $\widetilde{\boldsymbol{\omega}}, \widetilde{\mathbf{a}} \in \mathbb{C}^p$

Initial estimate of p.

Compute the singular values $\widetilde{\sigma}_k$ of $\widetilde{\mathbf{H}}$ in decreasing order where $[\widetilde{\mathbf{H}}]_{j,k} = \widetilde{y}_{j+k}$;

$p_{\text{est}} \leftarrow \text{argmin}_k \, \widetilde{\sigma}_k > \epsilon\sqrt{-2n\log(1-\sqrt[n]{r})}$;

$p \leftarrow 0$;

$\mathbf{r} \leftarrow \widetilde{\mathbf{y}}$;

repeat

  $p \leftarrow p+1$;

  Estimate $\omega_p \leftarrow 2\pi i/n \, \text{argmax}_k \, |[\mathbf{F}_n^* \mathbf{r}]_k|$;

  Update $\boldsymbol{\omega} \leftarrow [\boldsymbol{\omega}, \omega_p]$;

  Apply a nonlinear least squares algorithm to refine $\boldsymbol{\omega}$ and recover $\mathbf{a}$;

  $\mathbf{r} \leftarrow \widetilde{\mathbf{y}} - \mathbf{V}(\boldsymbol{\omega})\mathbf{a}$;

until $p \geq p_{est}$ *and* $\underline{\mathcal{A}}_c(p-1) < \underline{\mathcal{A}}_c(p)$;

$\widetilde{\boldsymbol{\omega}} \leftarrow [\boldsymbol{\omega}]_{0:p-2}$;

$\widetilde{\mathbf{a}} \leftarrow [\mathbf{a}]_{0:p-2}$;

Apply a nonlinear least squares algorithm to refine $\widetilde{\boldsymbol{\omega}}$ and recover $\widetilde{\mathbf{a}}$.

technique is sketched in Algorithm 3.2: we sequentially add exponentials until the AICc dictates the fit is not sufficiently better to warrant another exponential. We use the largest entry in the residual to estimate the next value of ω_p following Section 3.4.3 and can further refine this estimate if necessary using the Filtered Matrix Pencil method. Figure 3.7 shows this approach in action.

This approach mimics a graphical approach for fitting real exponentials called *peeling* [122], prominent before the advent of computers [79, VI.B.1]. Peeling consists of fitting a line to a log plot of $\widetilde{\mathbf{y}}$, removing the fit, and then repeating the procedure until the signal vanishes. In the same way, our updated Peeling Algorithm 3.2 iteratively removes exponentials and uses the residual to estimate the next value of ω.

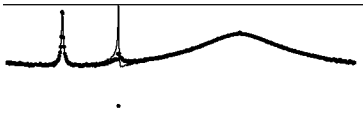
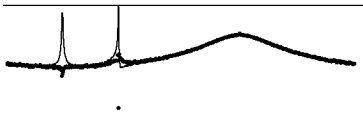
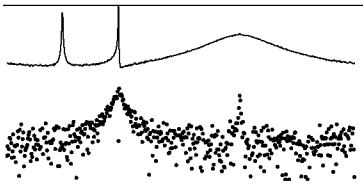
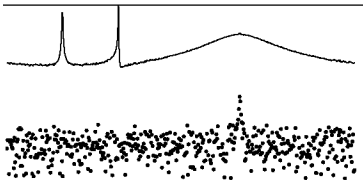
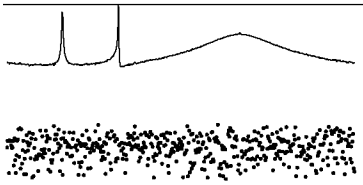
$\mathbf{F}^*\widetilde{\mathbf{r}}$	ω_0 a_0	ω_1 a_1	ω_2 a_2	ω_3 a_3	ω_4 a_4
			$0.001 + 2.010i$ $-0.725 + 0.069i$		
	$-0.001 + 1.000i$ $1.283 + 0.001i$		$0.001 + 2.010i$ $-0.726 + 0.078i$		
	$0.000 + 1.000i$ $1.000 + 0.008i$		$0.000 + 2.010i$ $-0.933 - 0.010i$	$-0.505 + 4.201i$ $50.739 - 0.107i$	
	$0.000 + 1.000i$ $0.999 + 0.001i$	$-0.100 + 2.002i$ $1.020 - 0.023i$	$-0.000 + 2.010i$ $-1.006 - 0.002i$	$-0.496 + 4.200i$ $50.078 + 0.063i$	
	$0.000 + 1.000i$ $0.999 + 0.001i$	$-0.098 + 2.002i$ $1.000 - 0.021i$	$-0.000 + 2.010i$ $-1.006 - 0.002i$	$-0.499 + 4.200i$ $50.031 + 0.062i$	$-0.010 + 4.201i$ $0.094 - 0.004i$
True solution	$0.000 + 1.000i$ $1.000 + 0.000i$	$-0.100 + 2.000i$ $1.000 + 0.000i$	$0.000 + 2.010i$ $-1.000 + 0.000i$	$-0.500 + 4.200i$ $50.000 + 0.000i$	$-0.010 + 4.200i$ $0.100 + 0.000i$

Figure 3.7 : Peeling with complex data. We recover five exponentials using the Peeling Algorithm 3.2, with $n = 512$. The black curve shows the original residual, $\mathbf{F}^*\widetilde{\mathbf{y}}$, and the dots $\mathbf{F}^*\mathbf{r}_k$ where $\mathbf{r}_k$ is the residual at the kth step.

We use the Hankel singular values to provide an initial estimate of p, since for early values of $k < p$, there are situations where $\mathcal{A}_c(p) < \mathcal{A}_c(k) < \mathcal{A}_c(k+1)$, leading our single look-ahead approach astray.

The updated peeling algorithm works well in some situations. For a randomly distributed exponentials exponential in noise, the numerical simulation in Table 3.1 illustrates that peeling frequently ensures convergence to the proper minimum. However, problems can emerge if exponentials are clustered. In Section 4.6.2, the Peeling Method produces poor estimates when multiple exponentials have similar imaginary components. Although the Peeling Method is not without its faults, it provides cheap estimates of initial conditions for large exponential fitting problems.

Chapter 4

Compression for Exponential Fitting

Solving large scale exponential fitting problems is expensive. Fitting p parameters $\boldsymbol{\theta}$ to n measurements $\mathbf{y}$ using nonlinear least squares requires $\mathcal{O}(np^2)$ operations per iteration that refines $\boldsymbol{\theta}$. Numerically stable Prony variants such as HSVD are even more expensive, requiring $\mathcal{O}(n^3)$ operations if implemented naively or $\mathcal{O}(pn \log n + np^2 + p^3)$ operations using fast inner-products and an Arnoldi based SVD [96]. Although the operation count grows only linearly in n for nonlinear least squares iterations, each iteration is too expensive for the millions or billions of measurements easily obtained with modern hardware. To combat this, practitioners throw away measurements to make each iteration feasible in one of three ways: discarding the last $n - m$ measurements (truncation), keeping every qth measurement – equivalent to reducing the sample rate – (decimation), or keeping a random subset of measurements (Incremental Gradient; see [50]). By discarding information in $\mathbf{y}$, the covariance of the parameters $\boldsymbol{\theta}$ increases. In this chapter, we build a new, fourth approach that sacrifices a minimum amount of information in $\mathbf{y}$ about $\boldsymbol{\theta}$ by compressing $\mathbf{y}$ onto a carefully chosen subspace of $\mathbb{C}^n$. The result is an exponential fitting algorithm that solves a nonlinear least squares problem over m compressed measurements in $\mathcal{O}(mp^2)$ operations per iteration plus an $\mathcal{O}(n \log n)$ starting cost. Most importantly, the number of compressed measurements m required to attain a fixed accuracy is *small* and *independent* of n; typically $m \approx 10p$ is sufficient for 90% accuracy. Although this algorithm is tailored to the structure of exponential fitting, the framework developed in Sections 4.1 and

4.2 can apply to any nonlinear least squares problem.

The key to this approach is building an efficient *compression space* $\mathcal{W} \subset \mathbb{C}^n$ from a set of *parent coordinates*. In comparison to the full problem

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \|\widetilde{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}}, \tag{4.1}$$

the compressed problem solves (4.1) with $\mathbf{y}$ restricted to the subspace $\mathcal{W} \subseteq \mathbb{C}^n$, spanned by the columns of $\mathbf{W} \in \mathbb{C}^{n \times m}$, a subset of the parent coordinates that form the columns of $\mathbf{U}$. The compressed estimate of $\widehat{\boldsymbol{\theta}}$, denoted $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}$. solves instead

$$\widetilde{\boldsymbol{\theta}}_{\mathcal{W}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \|\mathbf{W}^*\widetilde{\mathbf{y}} - \mathbf{W}^*\mathbf{f}(\boldsymbol{\theta})\|^2_{\widehat{\boldsymbol{\Gamma}}} \qquad \widehat{\boldsymbol{\Gamma}} = \left[(\mathbf{W}^*\mathbf{W})^{-1}\mathbf{W}^*\boldsymbol{\Gamma}^{-1}\mathbf{W}^*(\mathbf{W}^*\mathbf{W})^{-1}\right]^{-1}. \tag{4.2}$$

In the important case where the columns of $\mathbf{W}$ are orthonormal in the unweighted norm, then $\widehat{\boldsymbol{\Gamma}} = (\mathbf{W}^*\boldsymbol{\Gamma}^{-1}\mathbf{W})^{-1}$. In this framework decimation, truncation, and incremental gradient correspond to taking parent coordinates $\mathbf{U} = \mathbf{I}$, where decimation chooses columns $\mathcal{I} = 0\!:\!q\!:\!n-1$, truncation $\mathcal{I} = 0\!:\!m-1$, and Incremental Gradient a random column subset $\mathcal{I}$.

By discarding measurements, the covariance of $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}$ must be larger than the covariance of $\widetilde{\boldsymbol{\theta}}$. We quantify this in a single number called the *efficiency* of the subspace $\mathcal{W}$, defined as the limit of the respective covariances as noise vanishes uniformly ($\boldsymbol{\Gamma} \to \epsilon\boldsymbol{\Gamma}$)

$$\eta_{\widehat{\boldsymbol{\theta}}}(\mathcal{W}) := \lim_{\epsilon \to 0} \frac{\mathsf{Tr}\,\mathsf{Cov}\,\widetilde{\boldsymbol{\theta}}}{\mathsf{Tr}\,\mathsf{Cov}\,\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}} \in [0, 1]. \tag{4.3}$$

This generalizes Fisher's notion of efficiency to multiple variables [49, §4]. Following Fisher, we report $\eta_{\boldsymbol{\theta}}(\mathcal{W})$ as a percentage. Efficiency depends strongly on both the model function $\mathbf{f}$ and the value $\boldsymbol{\theta}$, as Figure 4.1 shows.

None of the compression subspaces generated by existing approaches are very efficient (see Figure 4.2). Better compression spaces can be generated by tailoring

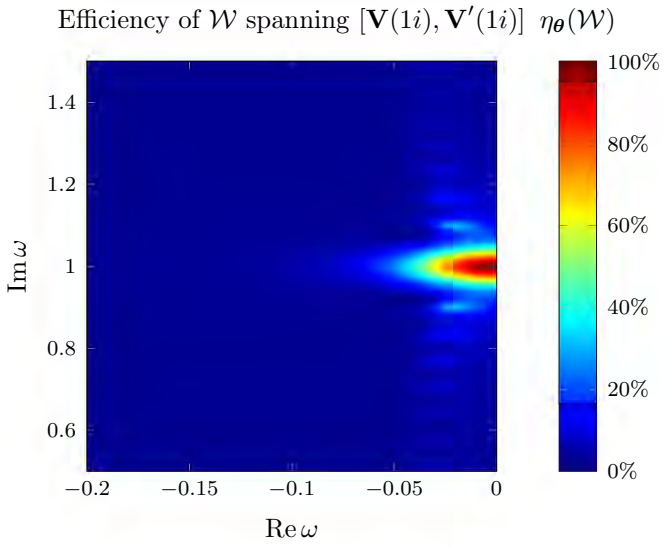


Figure 4.1 : Compression spaces are tailored to specific parameter values. Here we pick a subspace built to recover $\omega = 1i$ exactly and record the efficiency for neighboring parameter values. Efficiency drops off rapidly from $\omega = 1i$.

$\mathcal{W}$ to the current estimate of $\boldsymbol{\theta}$, and as subsequent iterations improve $\boldsymbol{\theta}$, changing $\mathcal{W}$ to match. For example, the ideal compression space for any problem (not just exponential fitting) contains the partial derivatives of $\mathbf{f}$ evaluated at $\widehat{\boldsymbol{\theta}}$. This effectively solves (4.2) using a QR-factorization of (4.1) – nothing novel. As we show in Lemma 4.3, this compression space has full efficiency ($\eta_{\boldsymbol{\theta}}(\mathcal{W}) = 1$) but efficiency falls off rapidly for $\boldsymbol{\theta}$ away from $\widehat{\boldsymbol{\theta}}$ as Figure 4.1 shows. However there are two other problems with this ideal compression space. These set conditions on compression spaces such that the compressed is computationally advantageous. The problem with the ideal subspace is as $\mathcal{W}$ is updated, recomputing $\mathbf{W}^*\widetilde{\mathbf{y}}$ takes $\mathcal{O}(nm)$ operations – a cost dependent on n we hope to avoid. To fix this, we impose the condition that $\mathbf{W}$ is chosen from a static set of parent coordinates $\mathbf{U}$ so that when we update $\mathcal{W}$, only a few additional inner products of $[\mathbf{U}]^*_{:,j}\widetilde{\mathbf{y}}$ need be computed, not m. The

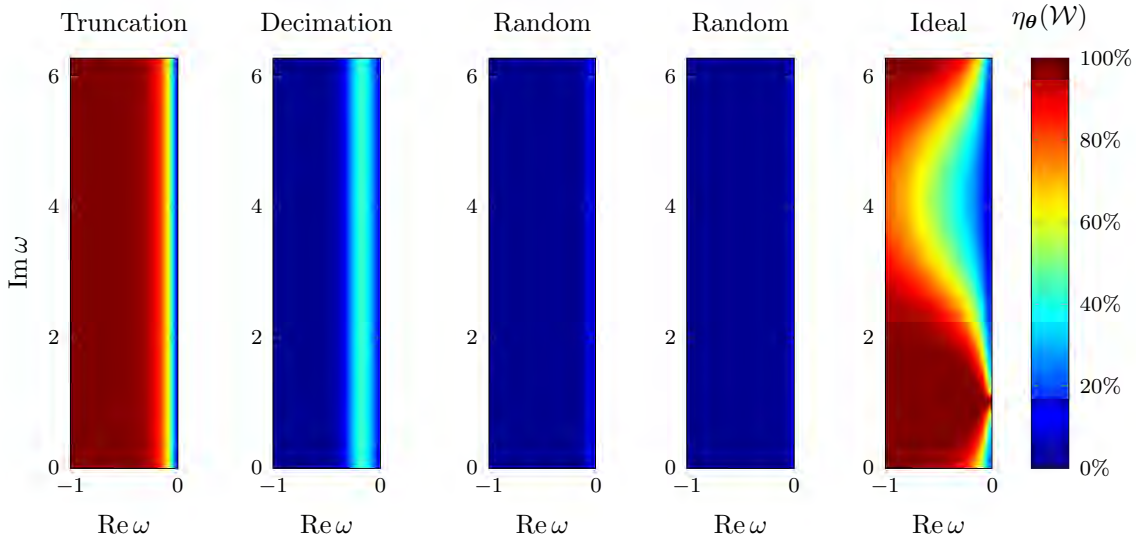


Figure 4.2 : Efficiency for truncation, decimation, incremental gradient, and ideal subspaces. In this example all subspaces are of dimension $m = 10$. Truncation works well for rapidly decaying signals, but poorly for undamped exponentials. Decimation works well for an intermediate decay, and for undamped signals has an efficiency of 10%. The two realizations of the random subspace from Incremental Gradient both provide poor subspaces that obtain a maximum expected efficiency of 10%. The ideal subspace for $\omega = [1i,\ 1i - 0.25,\ 1i - 0.5,\ 1i - 0.75,\ 1i - 1]$ and $\mathbf{a} = [1, 1, 1, 1, 1]$ provides very accurate estimates near $\operatorname{Im}\omega = 1$, but poorer estimates further away.

second problem with the ideal subspace, and most fatal, is that we need to compute $\mathbf{W}^*\mathbf{f}(\boldsymbol{\theta})$ and the derivative $\mathbf{W}^*\mathbf{F}(\boldsymbol{\theta})$ but in a way that avoids matrix multiplication requiring $\mathcal{O}(nm)$ and $\mathcal{O}(nmp)$ operations. One condition fixing this problem would be to require $\mathbf{U}^*\mathbf{f}(\boldsymbol{\theta})$ and $\mathbf{U}^*\mathbf{F}(\boldsymbol{\theta})$ can be computed row wise in closed form. These restrictive conditions make finding efficient parent coordinates difficult for an arbitrary problem. We may take $\mathbf{U} = \mathbf{I}$ for any problem, but it may not yield efficient compression spaces.

However, for the exponential fitting problem there is a class of parent coordinates satisfying these conditions. To satisfy the second condition, we choose parent coordinate built from blocks of the *Discrete Fourier Transform* (DFT) matrix $\mathbf{F}_b \in \mathbb{C}^{b\times b}$, where $[\mathbf{F}_b]_{j,k} = b^{-1/2}e^{-2\pi ijk/b}$,

$$\mathbf{U}(\mathbf{b}) = \operatorname{diag}(\mathbf{F}_{b_0}, \mathbf{F}_{b_1}, \cdots, \mathbf{F}_{b_{q-1}}) \text{ for } \mathbf{b} \in \mathbb{N}^q. \tag{4.4}$$

As required, we can compute $[\mathbf{U}^*\mathbf{f}]_j$ and $[\mathbf{U}^*\mathbf{F}]_{j,\cdot}$ in closed form using a generalization of the geometric sum formula in Theorem 4.1. Following the outline above, we could compute $\mathbf{U}^*\widetilde{\mathbf{y}}$ as columns of $\mathbf{U}$ are added to $\mathbf{W}$. However, it is faster to can compute $\mathbf{U}^*\widetilde{\mathbf{y}}$ at once using the Fast Fourier Transform (FFT), requiring fewer than $\mathcal{O}(qn\log n)$ operations.

This leaves two remaining challenges: choosing the columns of $\mathbf{U}(\mathbf{b})$ to form $\mathcal{W}$ and the picking block dimensions $\mathbf{b}$. We pick columns of $\mathbf{U}$ using greedy search maximizing the efficiency of $\mathcal{W}$. Section 4.1 motivates our generalization of efficiency and provides basic results. As computing efficiency is expensive, Section 4.2 derives fast updating formulas for the covariance and, thereby, the efficiency, Section 4.3 gives the closed form inner products used to compute $\mathbf{W}^*\mathbf{f}$ and $\mathbf{W}^*\mathbf{F}$ quickly. Section 4.4 details how to pick the initial set of indices for the greedy search and accelerates this search by restricting to a set of candidates. The choice of block dimensions

b is discussed in Section 4.5. Although we are unable to build an optimal set of blocks, we explore several options, finding that block sizes that grow geometrically (e.g., $b_j = 4^j$) work well. Other attempts to pick b_j using heuristics result in similar performance. Finally, Section 4.6 details how to use compression in conjunction with optimization and demonstrates that compression results in significant speed gains for large problems.

4.1 Compression Subspace Efficiency

It is natural to compare compression spaces by their covariances, but this alone cannot provide the ranking necessary for finding the best subspace. Recall from (3.20) the asymptotic covariance for $\widetilde{\boldsymbol{\theta}}$ is

$$\mathsf{Cov}[\widetilde{\boldsymbol{\theta}}] = \epsilon^2 \left(\mathbf{F}(\widehat{\boldsymbol{\theta}})^* \boldsymbol{\Gamma}^{-1} \mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{-1} + \mathcal{O}(\epsilon^3). \tag{4.5}$$

The same derivation applies to the compressed problem (4.2) (since it also is a non-linear least squares problem) yielding the asymptotic covariance

$$\mathsf{Cov}[\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}] = \epsilon^2 \left(\mathbf{F}(\widehat{\boldsymbol{\theta}})^* \mathbf{W}\mathbf{W}^* \boldsymbol{\Gamma}^{-1} \mathbf{W}\mathbf{W}^* \mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{-1} + \mathcal{O}(\epsilon^3) \text{ when } \mathbf{W}^*\mathbf{W} = \mathbf{I}. \tag{4.6}$$

Both asymptotic covariance matrices are, by construction, positive definite, and positive definite matrices have a *partial ordering*. For two positive definite matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \succ \mathbf{B}$ if $\mathbf{x}^*\mathbf{A}\mathbf{x} > \mathbf{x}^*\mathbf{B}\mathbf{x}$ for all $\mathbf{x}$. However, there are pairs of matrices that are not comparable – e.g., the inequality goes both ways: for some $\mathbf{x}$, $\mathbf{x}^*\mathbf{A}\mathbf{x} > \mathbf{x}^*\mathbf{B}\mathbf{x}$ and for some $\mathbf{y}$, $\mathbf{y}^*\mathbf{A}\mathbf{y} < \mathbf{y}^*\mathbf{B}\mathbf{y}$. Hence two compression subspaces $\mathcal{W}_1$ and $\mathcal{W}_2$ may not be comparable. Efficiency $\eta_{\widehat{\boldsymbol{\theta}}}$ resolves this problem of ordering by measuring the expected ℓ_2 error in $\widetilde{\boldsymbol{\theta}}$. This allows us to construct a preorder[1] of compression spaces where $\mathcal{W}_1 > \mathcal{W}_2$ if $\eta_{\widehat{\boldsymbol{\theta}}}(\mathcal{W}_1) > \eta_{\widehat{\boldsymbol{\theta}}}(\mathcal{W}_2)$.

[1]Preorders lack the requirement that $a \geq b$ and $a \leq b$ implies $a = b$.

Using our standard model where errors are of the form $\widetilde{\mathbf{y}} = \mathbf{y} + \mathbf{g} = \mathbf{f}(\widehat{\boldsymbol{\theta}}) + \mathbf{g}$ with $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})$, the expected ℓ_2 error in $\widetilde{\boldsymbol{\theta}}$ is

$$\mathsf{E}\left[\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\|_2^2\right] = \int_{\mathbb{C}^n} \left[\epsilon^2 \|\widehat{\mathbf{F}}\mathbf{g}\|^2 + \mathcal{O}(\epsilon^3)\right] \frac{e^{-\mathbf{g}^*\mathbf{g}/\epsilon^2}}{(\epsilon\pi)^n} \, \mathrm{d}\mathbf{g},$$

where $\widehat{\mathbf{F}}$ is the first order perturbation matrix $\widehat{\mathbf{F}} = (\mathbf{F}(\widehat{\boldsymbol{\theta}})^*\boldsymbol{\Gamma}^{-1}\mathbf{F}(\widehat{\boldsymbol{\theta}}))^{-1}\mathbf{F}(\widehat{\boldsymbol{\theta}})^*\boldsymbol{\Gamma}^{-1/2}$. If $\mathbf{F}(\widehat{\boldsymbol{\theta}})$ has full rank, we evaluate the first term by taking the SVD of $\widehat{\mathbf{F}}$, $\widehat{\mathbf{F}} = \widehat{\mathbf{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{V}}^*$, and by changing integration coordinates, $\widehat{\mathbf{g}} = \epsilon\widehat{\mathbf{V}}^*\mathbf{g}$, we have

$$\begin{aligned}\int_{\mathbb{C}^n} \|\widehat{\mathbf{F}}\mathbf{g}\|_2^2 \frac{e^{-\mathbf{g}^*\mathbf{g}/\epsilon^2}}{(\epsilon\pi)^n} \, \mathrm{d}\widehat{\mathbf{g}} &= \int_{\mathbb{C}^n} \widehat{\mathbf{g}}^*\widehat{\boldsymbol{\Sigma}}^2\widehat{\mathbf{g}} \frac{e^{-\widehat{\mathbf{g}}^*\widehat{\mathbf{g}}/\epsilon^2}}{(\epsilon\pi)^n} \, \mathrm{d}\mathbf{g} \\ &= \sum_{j=1}^{p} \sigma_j^2 \prod_{k=1}^{n} \int_{\mathbb{C}} |\widehat{g}_j|^2 \frac{e^{-|\widehat{g}_k|^2/\epsilon^2}}{\epsilon\pi} \, \mathrm{d}\widehat{g}_k = \frac{1}{2}\sum_{j=1}^{p} \sigma_j^2.\end{aligned}$$

This last expression has several equivalent formulations: the (scaled) square of the Frobenius norm of $\widehat{\mathbf{F}}$, the trace of the covariance matrix, or the sum of squares of singular values, where $\sigma_j(\mathbf{F})$ denotes the jth singular value (descending) of $\mathbf{F}$:

$$\lim_{\epsilon\to 0} \epsilon^{-2} \mathsf{E}\left[\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\|_2^2\right] = \frac{1}{2}\|\widehat{\mathbf{F}}\|_{\mathrm{F}}^2 \tag{4.7}$$

$$= \frac{1}{2} \mathsf{Tr}\left(\mathbf{F}(\widehat{\boldsymbol{\theta}})^*\boldsymbol{\Gamma}^{-1}\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{-1} \tag{4.8}$$

$$= \frac{1}{2}\sum_j \sigma_j\left(\boldsymbol{\Gamma}^{-1/2}\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{-2} \tag{4.9}$$

$$= \frac{1}{2}\lim_{\epsilon\to 0} \epsilon^{-2} \mathsf{Tr}\ \mathsf{Cov}\left[\widetilde{\boldsymbol{\theta}}\right]. \tag{4.10}$$

If $\mathbf{F}(\widehat{\boldsymbol{\theta}})$ does not have full rank, the expected ℓ_2 error is undefined. The same expected error result holds for the compressed problem:

$$\lim_{\epsilon\to 0} \epsilon^{-2} \mathsf{E}\left[\|\widetilde{\boldsymbol{\theta}}_{\mathcal{W}} - \widehat{\boldsymbol{\theta}}\|_2^2\right] = \frac{1}{2} \mathsf{Tr}\left(\mathbf{F}(\widehat{\boldsymbol{\theta}})^*\mathbf{W}\mathbf{W}^*\boldsymbol{\Gamma}^{-1}\mathbf{W}\mathbf{W}^*\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{-1} \tag{4.11}$$

$$= \frac{1}{2}\sum_j \sigma_j\left(\boldsymbol{\Gamma}^{-1/2}\mathbf{W}\mathbf{W}^*\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{-2} \tag{4.12}$$

$$= \frac{1}{2}\lim_{\epsilon\to 0} \epsilon^{-2} \mathsf{Tr}\ \mathsf{Cov}\left[\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}\right]. \tag{4.13}$$

In the definition of efficiency (4.3), we normalize by the full problem, since the Cramér-Rao lower bound guarantees that $\mathsf{Cov}\,\widetilde{\boldsymbol{\theta}}_{\mathcal{W}} \succeq \mathsf{Cov}\,\widetilde{\boldsymbol{\theta}}$. Then efficiency is bounded between zero and one:

$$\eta_{\boldsymbol{\theta}}(\mathcal{W}) := \lim_{\epsilon\to 0} \frac{\mathsf{Tr}\,\mathsf{Cov}\,\widetilde{\boldsymbol{\theta}}}{\mathsf{Tr}\,\mathsf{Cov}\,\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}} = \frac{\sum_j \sigma_j(\boldsymbol{\Gamma}^{-1/2}\mathbf{F}(\widehat{\boldsymbol{\theta}}))^{-2}}{\sum_j \sigma_j(\boldsymbol{\Gamma}^{-1/2}\mathbf{W}\mathbf{W}^*\mathbf{F}(\widehat{\boldsymbol{\theta}}))^{-2}} \in [0,1]. \tag{4.14}$$

We know that $\eta_{\boldsymbol{\theta}}(\mathcal{W}) \leq 1$ from the Cramér-Rao bound, but we can also see this immediately from linear algebra as well.

Lemma 4.1.

$$\lim_{\epsilon\to 0} \epsilon^{-2}\,\mathsf{Cov}\left[\widetilde{\boldsymbol{\theta}}\right] \preceq \lim_{\epsilon\to 0} \epsilon^{-2}\,\mathsf{Cov}\left[\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}\right]. \tag{4.15}$$

Proof. Abbreviating $\mathbf{F}(\widehat{\boldsymbol{\theta}}) = \mathbf{F}$, since $\mathbf{W}\mathbf{W}^*$ is an orthogonal projector onto $\mathcal{W}$ the covariance of the full (4.5) and compressed (4.6) problems obey

$$\mathbf{x}^*\mathbf{F}^*\mathbf{W}\mathbf{W}^*\boldsymbol{\Gamma}^{-1}\mathbf{W}\mathbf{W}^*\mathbf{F}\mathbf{x} \leq \mathbf{x}^*\mathbf{F}^*\boldsymbol{\Gamma}^{-1}\mathbf{F}\mathbf{x}.$$

Hence $\mathbf{F}^*\mathbf{W}\mathbf{W}^*\boldsymbol{\Gamma}^{-1}\mathbf{W}\mathbf{W}^*\mathbf{F} \preceq \mathbf{F}^*\boldsymbol{\Gamma}^{-1}\mathbf{F}$. Then by [68, Cor. 7.7.4], $(\mathbf{F}^*\boldsymbol{\Gamma}^{-1}\mathbf{F})^{-1} \preceq (\mathbf{F}^*\mathbf{W}\mathbf{W}^*\boldsymbol{\Gamma}^{-1}\mathbf{W}\mathbf{W}^*\mathbf{F})^{-1}$. □

We also know that if one compression subspace $\mathcal{W}_1$ is a subspace of $\mathcal{W}_2$, then the efficiency of $\mathcal{W}_1$ is bounded above by the efficiency of $\mathcal{W}_2$.

Lemma 4.2. *If $\mathcal{W}_1 \subseteq \mathcal{W}_2$, then $\eta_{\boldsymbol{\theta}}(\mathcal{W}_1) \leq \eta_{\boldsymbol{\theta}}(\mathcal{W}_2)$.*

Proof. If $\mathbf{W}_1$ is an orthogonal basis for $\mathcal{W}_1$ and $\mathbf{W}_2$ for $\mathcal{W}_2$, then $\mathbf{W}_1\mathbf{W}_1^* \preceq \mathbf{W}_2\mathbf{W}_2^*$. Consequently,

$$\begin{aligned}
\mathbf{F}^*\mathbf{W}_1\mathbf{W}_1^*\boldsymbol{\Gamma}^{-1}\mathbf{W}_1\mathbf{W}_1^*\mathbf{F} &\preceq \mathbf{F}^*\mathbf{W}_2\mathbf{W}_2^*\boldsymbol{\Gamma}^{-1}\mathbf{W}_2\mathbf{W}_2^*\mathbf{F} \\
(\mathbf{F}^*\mathbf{W}_1\mathbf{W}_1^*\boldsymbol{\Gamma}^{-1}\mathbf{W}_1\mathbf{W}_1^*\mathbf{F})^{-1} &\succeq (\mathbf{F}^*\mathbf{W}_2\mathbf{W}_2^*\boldsymbol{\Gamma}^{-1}\mathbf{W}_2\mathbf{W}_2^*\mathbf{F})^{-1} \\
\frac{\mathsf{Tr}(\mathbf{F}^*\boldsymbol{\Gamma}^{-1}\mathbf{F})^{-1}}{\mathsf{Tr}(\mathbf{F}^*\mathbf{W}_1\mathbf{W}_1^*\boldsymbol{\Gamma}^{-1}\mathbf{W}_1\mathbf{W}_1^*\mathbf{F})^{-1}} &\leq \frac{\mathsf{Tr}(\mathbf{F}^*\boldsymbol{\Gamma}^{-1}\mathbf{F})^{-1}}{\mathsf{Tr}(\mathbf{F}^*\mathbf{W}_2\mathbf{W}_2^*\boldsymbol{\Gamma}^{-1}\mathbf{W}_2\mathbf{W}_2^*\mathbf{F})^{-1}}.
\end{aligned}$$

□

Finally, we have a necessary and sufficient condition for $\mathcal{W}$ to have full efficiency.

Lemma 4.3. *If and only if $\partial/\partial\theta_j f(\boldsymbol{\theta}) \in \mathcal{W}$ for all j, then $\eta_{\widehat{\boldsymbol{\theta}}}(\mathcal{W}) = 1$.*

Proof. Suppose $\mathbf{W}$ is an orthogonal basis for $\mathcal{W}$ and $\mathbf{F} = \mathbf{F}(\widehat{\boldsymbol{\theta}})$. If $\partial/\partial\theta_j f(\boldsymbol{\theta}) \in \mathcal{W}$ for all j, then $\mathbf{W}\mathbf{W}^*\mathbf{F} = \mathbf{F}$, and hence $\mathbf{F}^*\mathbf{W}\mathbf{W}^*\boldsymbol{\Gamma}\mathbf{W}\mathbf{W}^*\mathbf{F} = \mathbf{F}^*\boldsymbol{\Gamma}\mathbf{F}$, thus $\eta_{\widehat{\boldsymbol{\theta}}}(\mathcal{W}) = 1$. Alternatively, if $\eta_{\widehat{\boldsymbol{\theta}}}(\mathcal{W}) = 1$, then $\mathsf{Tr}(\mathbf{F}^*\boldsymbol{\Gamma}\mathbf{F})^{-1} = \mathsf{Tr}(\mathbf{F}^*\mathbf{W}\mathbf{W}^*\boldsymbol{\Gamma}\mathbf{W}\mathbf{W}^*\mathbf{F})^{-1}$. By Lemma 4.1, then $\mathbf{F}^*\boldsymbol{\Gamma}\mathbf{F} = \mathbf{F}^*\mathbf{W}\mathbf{W}^*\boldsymbol{\Gamma}\mathbf{W}\mathbf{W}^*\mathbf{F}$ and the range of $\mathbf{W}$ must contain the range of $\mathbf{F}$. □

We can also define efficiency with respect to the frequencies $\boldsymbol{\omega}$ by considering only the covariance with respect to $\boldsymbol{\omega}$. If we order $\boldsymbol{\theta} = [\boldsymbol{\omega}, \mathbf{a}]$, then $\mathsf{Cov}\,\widetilde{\boldsymbol{\omega}} := [\mathsf{Cov}\,\widetilde{\boldsymbol{\theta}}]_{0:p-1,0:p-1}$, and correspondingly

$$\eta_{\widehat{\boldsymbol{\omega}}}(\mathcal{W}) = \lim_{\epsilon\to 0} \frac{\sum_{j=0}^{p-1}[\mathsf{Cov}\,\widetilde{\boldsymbol{\theta}}]_{j,j}}{\sum_{j=0}^{p-1}[\mathsf{Cov}\,\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}]_{j,j}}. \tag{4.16}$$

We can avoid computing the covariance of $\widetilde{\boldsymbol{\theta}}$ by noting the upper left block of $\mathsf{Cov}\,\widetilde{\boldsymbol{\theta}}$ is given by the Variable Projection asymptotic covariance (3.30),

$$\lim_{\epsilon\to 0} \epsilon^{-2}\,\mathsf{Cov}[\widetilde{\boldsymbol{\theta}}]_{0:p-1,0:p-1} = \left(\mathbf{P}(\widehat{\boldsymbol{\omega}})\boldsymbol{\mathcal{V}}(\widehat{\boldsymbol{\omega}})\,\bar{\times}_2\,\widehat{\mathbf{a}}\right)^{+}\mathbf{P}(\widehat{\boldsymbol{\omega}})\left(\mathbf{P}(\widehat{\boldsymbol{\omega}})\boldsymbol{\mathcal{V}}(\widehat{\boldsymbol{\omega}})\,\bar{\times}_2\,\widehat{\mathbf{a}}\right)^{+*}. \tag{4.17}$$

We can do the same for the compressed problem. Recall, the compressed problem solves

$$\widetilde{\boldsymbol{\omega}}_{\mathcal{W}} = \operatorname*{argmin}_{\boldsymbol{\omega}} \min_{\mathbf{a}} \|\mathbf{W}^*\widetilde{\mathbf{y}} - \mathbf{W}^*\mathbf{V}(\boldsymbol{\omega})\mathbf{a}\|_2^2. \tag{4.18}$$

The Variable Projector for this compressed problem is

$$\mathbf{P}_{\mathcal{W}}(\boldsymbol{\omega}) = \mathbf{I} - (\mathbf{W}^*\mathbf{V}(\boldsymbol{\omega}))(\mathbf{W}^*\mathbf{V}(\boldsymbol{\omega}))^*, \tag{4.19}$$

and consequently the first order perturbation of $\widetilde{\boldsymbol{\omega}}_{\mathcal{W}}$ follows the same derivation as (3.29):

$$\boldsymbol{\omega}_{\mathcal{W}}^{(1)} = -(\mathbf{P}_{\mathcal{W}}(\widehat{\boldsymbol{\omega}})\mathbf{W}^*\boldsymbol{\mathcal{V}}(\widehat{\boldsymbol{\omega}})\,\bar{\times}_2\,\widehat{\mathbf{a}})^{+}\mathbf{P}_{\mathcal{W}}(\widehat{\boldsymbol{\omega}})\mathbf{g}. \tag{4.20}$$

Consequently, we have

$$\begin{aligned}\lim_{\epsilon\to 0}\epsilon^{-2}\,\mathsf{Cov}[\widetilde{\boldsymbol{\omega}}_{\mathcal{W}}] &= (\boldsymbol{\omega}_{\mathcal{W}}^{(1)})(\boldsymbol{\omega}_{\mathcal{W}}^{(1)})^{*} \\ &= \left(\mathbf{P}_{\mathcal{W}}(\widehat{\boldsymbol{\omega}})\mathbf{W}^{*}\boldsymbol{\mathcal{V}}(\widehat{\boldsymbol{\omega}})\,\bar{\times}_2\,\widehat{\mathbf{a}}\right)^{+}\mathbf{P}_{\mathcal{W}}(\widehat{\boldsymbol{\omega}})\left(\mathbf{P}_{\mathcal{W}}(\widehat{\boldsymbol{\omega}})\mathbf{W}^{*}\boldsymbol{\mathcal{V}}(\widehat{\boldsymbol{\omega}})\,\bar{\times}_2\,\widehat{\mathbf{a}}\right)^{+*}.\end{aligned} \tag{4.21}$$

The same options for computing the trace are available as before. The SVD provides an efficient and comparatively stable approach, so we take

$$\eta_{\widehat{\boldsymbol{\omega}}}(\mathcal{W}) = \frac{\sum_{j=1}^{p}\sigma_j(\mathbf{P}(\widehat{\boldsymbol{\omega}})\boldsymbol{\mathcal{V}}(\widehat{\boldsymbol{\omega}})\,\bar{\times}_2\,\widehat{\mathbf{a}})^{-2}}{\sum_{j=1}^{p}\sigma_j\left(\mathbf{P}_{\mathcal{W}}(\widehat{\boldsymbol{\omega}})\mathbf{W}^{*}\boldsymbol{\mathcal{V}}(\widehat{\boldsymbol{\omega}})\,\bar{\times}_2\,\widehat{\mathbf{a}}\right)^{-2}}. \tag{4.22}$$

4.2 Updating Subspace Efficiency

With efficiency defining the ordering of compression subspaces, we now seek the best subspace for a given set of parameters. In the absence of a better approach, finding the best subspace of dimension m requires computing the efficiency of $\binom{n}{m}$ possible subspaces. This is easily more expensive than solving the full optimization problem (i.e., optimizing without compression). Rather than finding the best possible subspace, we pick a good enough subspace by greedily adding columns from the parent coordinates to the current compression subspace. Even this greedy approach is too expensive if we consider all of the $n - m$ possible columns. Rather, we consider columns from a small set of candidate columns $\mathcal{C}$. Then the dominant cost becomes computing the SVD of $\mathbf{W}^{*}\mathbf{F}(\widehat{\boldsymbol{\theta}})$ to compute the covariance trace of $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}$, since we need not compute the covariance trace of $\widetilde{\boldsymbol{\theta}}$, as it is constant for all candidate subspaces. In this section, we develop rank-1 update formulas for these singular values when we append the column i to the current set of columns $\mathcal{I}$. Although these formulas can be unstable, they provide insight into how to pick the initial set of columns $\mathcal{I}$ and how to form the candidate set $\mathcal{C}$ for any problem. Later, Section 4.4 specializes these results for the exponential fitting problem.

For simplicity, we will assume that noise is white ($\boldsymbol{\Gamma} = \gamma\mathbf{I}$), parent coordinates $\mathbf{U}$ are orthonormal, and $\mathbf{U}$ has already been applied to $\mathbf{f}(\boldsymbol{\theta})$ (e.g., $\mathbf{f}(\boldsymbol{\theta}) \leftarrow \mathbf{U}^*\mathbf{f}(\boldsymbol{\theta})$). Thus instead of considering what columns of $\mathbf{U}$ to add to $\mathbf{W}$, we consider which rows of $\mathbf{f}(\boldsymbol{\theta})$ to include. We abbreviate the rows $\mathcal{I} \subseteq \mathcal{N}$ of $\mathbf{f}(\boldsymbol{\theta})$ and its derivative $\mathbf{F}(\boldsymbol{\theta})$ by

$$\mathbf{f}_{\mathcal{I}} = [\mathbf{f}(\widehat{\boldsymbol{\theta}})]_{\mathcal{I}} \qquad\qquad \mathbf{F}_{\mathcal{I}} = [\mathbf{F}(\widehat{\boldsymbol{\theta}})]_{\mathcal{I},\cdot}.$$

Further, as γ appears in both the numerator and denominator of efficiency, we can ignore it, assuming $\gamma = 1$. We develop four update formulas: one for the singular values, and one for the covariance for both the general problem ($\boldsymbol{\theta} = [\boldsymbol{\omega}, \mathbf{a}]$) and the separable problem ($\boldsymbol{\omega}$ alone). These formulas use the Sherman-Morrison formula, but can easily be generalized to allow multiple rows to be appended simultaneously using the Woodbury identity.

4.2.1 Updating the Singular Values for the General Problem

At step $m-1$ of a greedy search, we have the index set $\mathcal{I}_{m-1}$ and the SVD of $\mathbf{F}_{\mathcal{I}_{m-1}}$. We then seek the row $i \notin \mathcal{I}_{m-1}$ minimizing the covariance of $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}_m}$, where $\mathcal{W}_m = \text{span}\mathbf{U}_{\cdot,\mathcal{I}_m}$ and $\mathcal{I}_m = \mathcal{I}_{m-1} \cup i$. Under this section's assumptions, we can simplify (4.12):

$$\sum_j \sigma_j \left(\boldsymbol{\Gamma}^{-1/2}\mathbf{W}_m\mathbf{W}_m^*\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{-2} = \sum_j \sigma_j \left(\mathbf{W}_m^*\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{-2} = \sum_j \sigma_j(\mathbf{F}_{\mathcal{I}_m})^{-2}.$$

We update the SVD of $\mathbf{F}_{\mathcal{I}_{m-1}}$ to that of $\mathbf{F}_{\mathcal{I}_m}$, noting

$$\mathbf{F}_{\mathcal{I}_m} = \begin{bmatrix} \mathbf{F}_{\mathcal{I}_{m-1}} \\ \mathbf{F}_i \end{bmatrix} = \begin{bmatrix} \mathbf{F}_{\mathcal{I}_{m-1}} \\ \mathbf{0} \end{bmatrix} + \mathbf{e}_m\mathbf{F}_i. \tag{4.23}$$

The singular values of $\mathbf{F}_{\mathcal{I}_m}$ are then a rank-1 perturbation of the original singular values

$$\mathbf{F}_{\mathcal{I}_m}^*\mathbf{F}_{\mathcal{I}_m} = \mathbf{F}_{\mathcal{I}_{m-1}}^*\mathbf{F}_{\mathcal{I}_{m-1}} + \mathbf{F}_i^*\mathbf{F}_i = \mathbf{V}\left[\boldsymbol{\Sigma}^2 + (\mathbf{F}_i\mathbf{V})^*(\mathbf{F}_i\mathbf{V})\right]\mathbf{V}^* \tag{4.24}$$

that can be computed using the secular equation (`dlaed4` in LAPACK [6]; see [56, §8.5.3] for details). This requires only one $\mathcal{O}(mp^2)$ operation to compute the SVD of $\mathbf{F}_{\mathcal{I}_{m-1}}$ and $\mathcal{O}(p)$ operations per candidate i, rather than $\mathcal{O}(mp^2)$ operations to compute the SVD of $\mathbf{F}_{\mathcal{I}_m}$ for each candidate.

4.2.2 Updating the Covariance for the General Problem

Alternatively we could update the covariance using the Sherman-Morrison formula. Starting from (4.11), we note that since $\mathbf{W}_m$ is orthonormal and $\mathbf{\Gamma} = \mathbf{I}$,

$$\mathsf{Tr}\left(\mathbf{F}(\widehat{\boldsymbol{\theta}})^*\mathbf{W}_m\mathbf{W}_m^*\mathbf{\Gamma}^{-1}\mathbf{W}_m\mathbf{W}_m^*\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{-1} = \mathsf{Tr}\left(\mathbf{F}(\widehat{\boldsymbol{\theta}})^*\mathbf{W}_m\mathbf{W}_m^*\mathbf{F}(\widehat{\boldsymbol{\theta}})\right)^{-1} = \mathsf{Tr}(\mathbf{F}_{\mathcal{I}_m}^*\mathbf{F}_{\mathcal{I}_m})^{-1}.$$

Splitting $\mathbf{F}_{\mathcal{I}_m}$ as in (4.23), we apply the Sherman-Morrison inverse update formula [56, eq. (2.1.4)] to the result above, yielding

$$\begin{aligned}\mathsf{Tr}\left[(\mathbf{F}_{\mathcal{I}_m}^*\mathbf{F}_{\mathcal{I}_m})^{-1}\right] &= \mathsf{Tr}\left[(\mathbf{F}_{\mathcal{I}_{m-1}}^*\mathbf{F}_{\mathcal{I}_{m-1}})^{-1} - \frac{\mathbf{c}_i\mathbf{c}_i^*}{1+\beta_i}\right] \\ &= \mathsf{Tr}(\mathbf{F}_{\mathcal{I}_{m-1}}^*\mathbf{F}_{\mathcal{I}_{m-1}})^{-1} - \frac{\mathbf{c}_i^*\mathbf{c}_i}{1+\beta_i},\end{aligned} \tag{4.25}$$

where $\mathbf{c}_i = (\mathbf{F}_{\mathcal{I}_{m-1}}^*\mathbf{F}_{\mathcal{I}_{m-1}})^{-1}\mathbf{F}_i^*$ and $\beta_i = \mathbf{c}_i^*\mathbf{F}_i^*$. As the first term is constant, the best i has the largest *drop*

$$\Delta_g(i) := \frac{\mathbf{c}_i^*\mathbf{c}_i}{1+\beta_i}. \tag{4.26}$$

When this is a one dimensional problem, we can further simply our choice of the new index i. Then $(\mathbf{F}_{\mathcal{I}_{m-1}}^*\mathbf{F}_{\mathcal{I}_{m-1}})^{-1}$ is a scalar γ and the best $\mathbf{F}_i$ minimizes the drop

$$\Delta_g(i) = -\frac{\gamma^2\|\mathbf{F}_i\|^2}{1+\gamma\|\mathbf{F}_i\|^2}. \tag{4.27}$$

As this function monotonically decreases in $\|\mathbf{F}_i\|^2$ regardless of γ, the largest drop corresponds to the largest row in $\mathbf{F}(\boldsymbol{\theta})$ not already in $\mathcal{I}$.

4.2.3 Updating the Singular Values for the Separable Problem

Similar update formulas are available when considering $\widetilde{\boldsymbol{\omega}}$ as well. We do so by updating the singular values of $\mathbf{P}_{\mathcal{W}}(\widehat{\boldsymbol{\omega}})\mathbf{W}^*\boldsymbol{\mathcal{V}}(\widehat{\boldsymbol{\omega}})\,\bar{\times}_2\,\widehat{\mathbf{a}}$ through the eigenvalues of $\mathbf{S}_{\mathcal{I}_m}$:

$$\mathbf{S}_{\mathcal{I}_m} = \mathbf{A}_{\mathcal{I}_m}^* \mathbf{P}_{\mathcal{I}_m} \mathbf{A}_{\mathcal{I}_m}, \tag{4.28}$$

where $\mathbf{A}_{\mathcal{I}_m} := [\boldsymbol{\mathcal{V}}(\boldsymbol{\omega})]_{\mathcal{I}_m,\cdot,\cdot}\,\bar{\times}_2\,\mathbf{a}$ and $\mathbf{P}_{\mathcal{I}_m} := \mathbf{P}^{\perp}_{[\mathbf{V}(\boldsymbol{\omega})]_{\mathcal{I}_m,\cdot}}$, the orthogonal projector onto the subspace perpendicular to the range of $[\mathbf{V}]_{\mathcal{I}_m,\cdot}$. As with the general problem, adding a new index to $\mathcal{I}_{m-1}$ results in a rank-1 update to $\mathbf{S}_{\mathcal{I}_{m-1}}$ to form $\mathbf{S}_{\mathcal{I}_m}$.

If i is the new index added to $\mathcal{I}_{m-1}$ to form $\mathcal{I}_m$, we write $\mathbf{A}_{\mathcal{I}_m}$ and $\mathbf{V}_{\mathcal{I}_m} = [\mathbf{V}(\widehat{\boldsymbol{\omega}})]_{\mathcal{I}_m,\cdot}$ as updates of $\mathcal{I}_{m-1}$:

$$\mathbf{V}_{\mathcal{I}_m} = \mathbf{V}_{\mathcal{I}_{m-1}} + \mathbf{e}_m \mathbf{V}_i \qquad\qquad \mathbf{A}_{\mathcal{I}_m} = \mathbf{A}_{\mathcal{I}_{m-1}} + \mathbf{e}_m \mathbf{A}_i \ ,$$

where both $\mathbf{V}_{\mathcal{I}_{m-1}}$ and $\mathbf{A}_{\mathcal{I}_{m-1}}$ are padded with a row of zeros at the bottom. For brevity, $\mathbf{V}$ and $\mathbf{A}$ without subscripts will denote $\mathbf{V}_{\mathcal{I}_{m-1}}$ and $\mathbf{A}_{\mathcal{I}_{m-1}}$. Starting with the orthogonal projector

$$\mathbf{P}_{\mathcal{I}_m} = \mathbf{I} - \mathbf{V}_{\mathcal{I}_m}(\mathbf{V}_{\mathcal{I}_m}^* \mathbf{V}_{\mathcal{I}_m})^{-1}\mathbf{V}_{\mathcal{I}_m}^* \ ,$$

we expand the middle term $\mathbf{B} = \mathbf{V}^*\mathbf{V}$ using the Sherman-Morrison update,

$$\widehat{\mathbf{B}} := (\mathbf{V}_{\mathcal{I}_m}^* \mathbf{V}_{\mathcal{I}_m})^{-1} = (\mathbf{B} + \mathbf{V}_i^* \mathbf{V}_i)^{-1} = \mathbf{B}^{-1} - \frac{\mathbf{b}_i \mathbf{b}_i^*}{1+\alpha},$$

where $\mathbf{b}_i = \mathbf{B}^{-1}\mathbf{V}_i^*$ and $\alpha_i = \mathbf{b}_i^*\mathbf{V}_i^* \in \mathbb{R}$. We then write the orthogonal projector onto $\mathbf{V}_{\mathcal{I}_m}$ as

$$\mathbf{V}_{\mathcal{I}_m}\widehat{\mathbf{B}}^{-1}\mathbf{V}_{\mathcal{I}_m}^* = \mathbf{V}\mathbf{B}^{-1}\mathbf{V}^* + \frac{\alpha_i \mathbf{e}_m\mathbf{e}_m^* + \mathbf{e}_m\mathbf{b}_i^*\mathbf{V}^* + \mathbf{V}\mathbf{b}_i\mathbf{e}_m^* - \mathbf{V}\mathbf{b}_i\mathbf{b}_i^*\mathbf{V}^*}{1+\alpha}.$$

When $\mathbf{A}_{\mathcal{I}_m}$ is applied to both sides, most cross terms vanish due to orthogonality of

the update to both $\mathbf{A}$ and $\mathbf{V}$:

$$\begin{aligned}\mathbf{A}^*_{\mathcal{I}_m}\mathbf{V}_{\mathcal{I}_m}\widehat{\mathbf{B}}^{-1}\mathbf{V}^*_{\mathcal{I}_m}\mathbf{A}_{\mathcal{I}_m} =&\mathbf{A}^*\mathbf{V}\mathbf{B}^{-1}\mathbf{V}^*\mathbf{A}\\ &+\frac{\alpha_i\mathbf{A}_i^*\mathbf{A}_i+\mathbf{A}_i^*\mathbf{b}_i^*\mathbf{V}^*\mathbf{A}+\mathbf{A}^*\mathbf{V}\mathbf{b}_i\mathbf{A}_i-\mathbf{A}^*\mathbf{V}\mathbf{b}_i\mathbf{b}_i^*\mathbf{V}^*\mathbf{A}}{1+\alpha}.\end{aligned}$$

Defining $\mathbf{d}_i=\mathbf{A}^*\mathbf{V}\mathbf{b}_i$, we expand

$$\begin{aligned}\mathbf{S}_{\mathcal{I}_m}&=\mathbf{A}^*_{\mathcal{I}_m}\mathbf{A}_{\mathcal{I}_m}-\mathbf{A}^*_{\mathcal{I}_m}\mathbf{V}_{\mathcal{I}_m}\widehat{\mathbf{B}}^{-1}\mathbf{V}^*_{\mathcal{I}_m}\mathbf{A}_{\mathcal{I}_m}\\ &=\mathbf{A}^*\mathbf{A}+\mathbf{A}_i^*\mathbf{A}_i-\mathbf{A}^*\mathbf{V}\mathbf{B}^{-1}\mathbf{V}^*\mathbf{A}-\frac{\alpha_i\mathbf{A}_i^*\mathbf{A}_i+\mathbf{A}_i^*\mathbf{d}^*+\mathbf{d}_i\mathbf{A}_i-\mathbf{d}_i\mathbf{d}_i^*}{1+\alpha}.\end{aligned}$$

Further simplification reveals the rank-one update to $\mathbf{S}_{\mathcal{I}_{m-1}}$:

$$\mathbf{S}_{\mathcal{I}_m}=\mathbf{S}_{\mathcal{I}_{m-1}}+\frac{1}{1+\alpha}(\mathbf{A}_i^*-\mathbf{d}_i)(\mathbf{A}_i^*-\mathbf{d}_i)^*. \tag{4.29}$$

We can now update the eigenvalues of $\mathbf{S}_{\mathcal{I}_m}$ using the secular equation and hence the singular values needed by (4.22) to compute efficiency.

4.2.4 Updating the Covariance for the Separable Problem

Rather than updating the singular values, alternatively, we can invoke the Sherman-Morrison formula and update $\mathbf{S}^{-1}_{\mathcal{I}_{m-1}}$ to $\mathbf{S}^{-1}_{\mathcal{I}_m}$ by

$$\mathbf{S}^{-1}_{\mathcal{I}_m}=\mathbf{S}^{-1}_{\mathcal{I}_{m-1}}-\frac{1}{1+\alpha_i}\,\frac{\mathbf{c}_i\mathbf{c}_i^*}{1+\beta_i(1+\alpha_i)}, \tag{4.30}$$

where $\mathbf{c}_i=\mathbf{S}^{-1}(\mathbf{A}_i^*-\mathbf{d}_i)$ and $\beta=\mathbf{c}_i^*(\mathbf{A}_i^*-\mathbf{d}_i)$. From here, the updated trace is

$$\mathsf{Tr}\,\mathbf{S}^{-1}_{\mathcal{I}_m}=\mathsf{Tr}\,\mathbf{S}^{-1}_{\mathcal{I}_{m-1}}-\frac{1}{1+\alpha_i}\,\frac{\mathbf{c}_i^*\mathbf{c}_i}{1+\beta_i(1+\alpha_i)}, \tag{4.31}$$

with analogous drop

$$\Delta_s(i):=\frac{1}{1+\alpha_i}\,\frac{\mathbf{c}_i^*\mathbf{c}_i}{1+\beta_i(1+\alpha_i)}. \tag{4.32}$$

Unfortunately, two applications of the Sherman-Morrison formula can make this formulation sensitive to round off errors; see Figure 4.4 showing that Δ_s choses the wrong best column for the exponential fitting problem.

When both $\boldsymbol{\omega}$ and $\mathbf{a}$ are scalars, we can further simply Δ_s, resulting in the one dimension drop Δ_{s1},

$$\Delta_{s1}(i) := \frac{-1}{1+\eta^{-2}|\mathbf{V}_i|^2}\,\frac{\sigma^{-2}|a|^2|\mathbf{V}'_i-\chi\eta^{-2}\mathbf{V}_i|^2}{1+\sigma^{-1}|a|^2|\mathbf{V}'_i-\chi\eta^{-2}\mathbf{V}_i|^2(1+\eta^{-2}|\mathbf{V}_i|^2)}; \\ \text{where } \sigma=\mathbf{S}_{\mathcal{I}_{m-1}},\ \eta=\|\mathbf{V}_{\mathcal{I}_{m-1}}\|_2,\ \chi=\mathbf{V}^*_{\mathcal{I}_{m-1}}\mathbf{V}'_{\mathcal{I}_{m-1}}. \tag{4.33}$$

Certain limits of this formulation yield the same heuristics as before in (4.26). In the limit of $\eta\to 0$, we recover

$$\lim_{\eta\to 0}\Delta_{s1}(i) = -\eta^{-2}\frac{\sigma^{-1}}{|\mathbf{V}_i|^4}. \tag{4.34}$$

Hence, if $\mathbf{V}_{\mathcal{I}_{m-1}}$ is small, the next index should attempt to increase this value by maximizing $|\mathbf{V}_i|$. In the opposite limit $\eta\to\infty$,

$$\lim_{\eta\to\infty}\Delta_{s1}(i) = -\frac{\sigma^{-2}|\mathbf{V}'_i|^2}{1+\sigma^{-1}|\mathbf{V}'_i|^2}.$$

Just like (4.26), this a monotonically decreasing function in $|\mathbf{V}'_i|$, so our best choice is to maximize $\mathbf{V}'_i$.

The drop Δ_{s1} is sensitive to errors in $\mathbf{S}_{\mathcal{I}_{m-1}}$. If we make the approximation $|\mathbf{V}_i|/\eta = 0$, assuming individual entries $\mathbf{V}_i$ are small compared to $\|\mathbf{V}_{\mathcal{I}_{m-1}}\|$, many terms drop out. The remainder is a monotonic function of $|\mathbf{V}'_i-\chi\eta^{-2}\mathbf{V}_i|^2$,

$$\Delta_{s1}(i)\approx\Delta_{s2}(i) := \frac{-\sigma^{-2}|a|^2|\mathbf{V}'_i-\chi\eta^{-2}\mathbf{V}_i|^2}{1+\sigma^{-1}|a|^2|\mathbf{V}'_i-\chi\eta^{-2}\mathbf{V}_i|^2}. \tag{4.35}$$

Then the *approximate* best new index solves

$$\min_i\left|\mathbf{V}'_i-\frac{\mathbf{V}'^*_{\mathcal{I}_{m-1}}\mathbf{V}_{\mathcal{I}_{m-1}}}{\mathbf{V}^*_{\mathcal{I}_{m-1}}\mathbf{V}_{\mathcal{I}_{m-1}}}\mathbf{V}_i\right|^2. \tag{4.36}$$

As the approximation $|\mathbf{V}_i|/\eta = 0$ is only valid when there are many entries in $\mathcal{I}_{m-1}$, this heuristic performs poorly for small subspaces. Consequently, building compression spaces using this heuristic can yield significantly malformed subspaces.

4.3 Closed Form Inner Products for Block Fourier Matrices

We now turn to developing specialized results for the exponential fitting problem. The first component is the inner product of the parent coordinates $\mathbf{U}$ with the model function $\mathbf{f}(\boldsymbol{\theta})$ and the derivative $\mathbf{F}(\boldsymbol{\theta})$, where $\mathbf{U}$ is a block Fourier matrix (4.4). To compute $\mathbf{U}^*\mathbf{f}(\boldsymbol{\theta})$ and $\mathbf{U}^*\mathbf{F}(\boldsymbol{\theta})$ for exponential fitting, we use the formulations from Variable Projection (3.14):

$$\mathbf{f}(\boldsymbol{\theta}) = \mathbf{V}(\boldsymbol{\omega})\mathbf{a} \quad \text{and} \quad [\mathbf{F}(\boldsymbol{\theta})]_{\cdot,k} = \frac{\partial}{\partial \omega_k}\mathbf{V}(\omega_k)a_k = \mathbf{V}'(\omega_k)a_k,$$

where $[\mathbf{V}(\boldsymbol{\omega})]_{j,k} = e^{j\omega_k}$. So to compute the inner products $\mathbf{U}^*\mathbf{f}(\boldsymbol{\theta})$ and $\mathbf{U}^*\mathbf{F}(\boldsymbol{\theta})$ in closed form, we need only compute $\mathbf{U}^*\mathbf{V}(\omega_k)$ and $\mathbf{U}^*\mathbf{V}'(\omega_k)$ in closed form.

4.3.1 $\mathbf{U}^*\mathbf{V}(\omega)$

Computing $\mathbf{U}^*\mathbf{V}(\omega)$ is a simple application of the geometric sum formula. We split $\mathbf{V}(\omega)$ into q chunks $\mathbf{v}_\ell(\omega)$ each corresponding to a Fourier block of dimension b_j:

$$\mathbf{U}^*\mathbf{V}(\omega) = \begin{bmatrix} \mathbf{F}^*_{b_0} & & & \\ & \mathbf{F}^*_{b_1} & & \\ & & \ddots & \\ & & & \mathbf{F}^*_{b_{q-1}} \end{bmatrix} \begin{bmatrix} \mathbf{v}_0(\omega) \\ \mathbf{v}_1(\omega) \\ \vdots \\ \mathbf{v}_{q-1}(\omega) \end{bmatrix} = \begin{bmatrix} \mathbf{F}^*_{b_0}\mathbf{v}_0(\omega) \\ \mathbf{F}^*_{b_1}\mathbf{v}_1(\omega) \\ \vdots \\ \mathbf{F}^*_{b_{q-1}}\mathbf{v}_{q-1}(\omega) \end{bmatrix}.$$

Each row of $\mathbf{F}^*_{b_\ell}\mathbf{v}_\ell(\omega)$ is a geometric sum

$$\left[\mathbf{F}^*_{b_\ell}\mathbf{v}_\ell(\omega)\right]_k = \sum_{j=0}^{b_\ell-1} \frac{e^{-2\pi ijk/b_\ell}}{\sqrt{b_\ell}} e^{(s_\ell+j)\omega} = \begin{cases} \dfrac{e^{s_\ell\omega}}{\sqrt{b_\ell}} \dfrac{1-e^{b_\ell\omega-2\pi ik}}{1-e^{\omega-2\pi ik/b_\ell}}, & \omega \neq 2\pi ik/b_\ell; \\ \sqrt{b_\ell}e^{s_\ell\omega}, & \omega = 2\pi ik/b_\ell, \end{cases} \tag{4.37}$$

where $s_\ell = \sum_{j<\ell} b_j$ is the start of each block (with $s_0 = 0$) and $0 \leq k \leq b_\ell - 1$.

4.3.2 $\mathbf{U}^*\mathbf{V}'(\omega)$

Computing the inner product with the derivative $\mathbf{V}'(\omega)$ requires more work. Splitting this vector into analogous blocks $\mathbf{v}'_\ell(\omega)$, we seek a closed form expression for the sum

$$\left[\mathbf{F}^*_{b_\ell}\mathbf{v}'_\ell(\omega)\right]_k = \sum_{j=0}^{b_\ell-1} \frac{e^{-2\pi ijk/b_\ell}}{\sqrt{b_\ell}}(s_\ell + j)e^{(s_\ell+j)\omega}, \quad 0 \le k \le b_\ell - 1. \tag{4.38}$$

This could be computed using the *polylogarithm*

$$\mathrm{Li}_n(z) = \sum_{k=1}^{\infty} \frac{z^k}{k^n}; \tag{4.39}$$

see, e.g. [99, eq. (7.1)] or using Sterling or Eulerian numbers following the results of Wood [163]. These expressions are unwieldy so we develop a generalization of the finite geometric sum formula provided by Theorem 4.1.

Theorem 4.1. *Given $n_1, n_2 \in \mathbb{Z}$ with $n_1 < n_2$, β a nonnegative integer, and $\omega \in \mathbb{C}$,*

$$\sum_{k=n_1}^{n_2-1} k^\beta e^{\omega k} = \begin{cases} \displaystyle\sum_{k=0}^{\beta} \frac{\chi_{n_1}(\beta,k)e^{(n_1+k)\omega} - \chi_{n_2}(\beta,k)e^{(n_2+k)\omega}}{(1-e^\omega)^{k+1}}, & \omega \neq 0; \\ \dfrac{B_{\beta+1}(n_2) - B_{\beta+1}(n_1)}{\beta+1}, & \omega = 0, \end{cases} \tag{4.40}$$

where the function $\chi_n : \mathbb{N} \times \mathbb{N} \to \mathbb{Z}$ is defined by the recurrence

$$\begin{aligned} \chi_n(\ell+1,k) &= (n+k)\chi_n(\ell,k) + k\chi_n(\ell,k-1), \quad \ell,k \ge 0 \\ \chi_n(0,k) &= \delta_{k,0}, \end{aligned} \tag{4.41}$$

and $B_{\beta+1}$ denotes the $\beta+1$ Bernoulli polynomial.

Proof. The $\omega = 0$ case follows directly from an existing result [111, eq. (24.4.9)].

When $\omega \neq 0$, we first extract derivatives and apply the geometric sum formula

$$\sum_{k=n_1}^{n_2-1} k^\beta e^{\omega k} = \frac{\partial^\beta}{\partial\omega^\beta} \sum_{k=n_1}^{n_2-1} e^{\omega k} = \frac{\partial^\beta}{\partial\omega^\beta} \frac{e^{\omega n_1} - e^{\omega n_2}}{1-e^\omega}. \tag{4.42}$$

We then apply these derivatives, by induction showing

$$\frac{\partial^\beta}{\partial\omega^\beta}\frac{e^{n\omega}}{1-e^\omega} = \sum_{k=0}^{\beta}\chi_n(\beta,k)\frac{e^{(n+k)\omega}}{(1-e^\omega)^{k+1}}. \tag{4.43}$$

The base case $\beta = 0$ holds as $\chi_n(0,0) = 1$. Taking the derivative, we show the inductive step

$$\begin{aligned}\frac{\partial}{\partial\omega}\sum_{k=0}^{\beta}\chi_n(\beta,k)\frac{e^{(n+k)\omega}}{(1-e^\omega)^{k+1}} &= \sum_{k=0}^{\beta+1}\left[\chi_n(\beta,k)(n+k)+\chi_n(\beta,k-1)k\right]\frac{e^{(n+k)\omega}}{(1-e^\omega)^{k+1}}\\ &= \sum_{k=0}^{\beta+1}\chi_n(\beta+1,k)\frac{e^{(n+k)\omega}}{(1-e^\omega)^{k+1}}.\end{aligned}$$

Subtracting the n_2 case from the n_1 case gives the first case of (4.40). □

Using this theorem, each row of $\mathbf{F}^*_{b_\ell}\mathbf{v}'_\ell(\omega)$ can then be evaluated by

$$\left[\mathbf{F}^*_{b_\ell}\mathbf{v}'_\ell(\omega)\right]_k = \begin{cases}\dfrac{e^{s_\ell\omega}}{\sqrt{b_\ell}}\left[\dfrac{s_\ell(1-e^{b_\ell\phi})-b_\ell e^{b_\ell\phi}}{1-e^\phi}+e^\phi\dfrac{1-e^{b_\ell\phi}}{(1-e^\phi)^2}\right], & \omega\neq 2\pi ik/b_\ell;\\ \dfrac{b_\ell^2+2b_\ell s_\ell-b_\ell}{2\sqrt{b_\ell}}, & \omega = 2\pi ik/b_\ell,\end{cases} \tag{4.44}$$

where $\phi = \omega - 2\pi ik/b_\ell$.

4.4 Subspace Selection for Exponential Fitting

With closed form expressions for $\mathbf{U}^*_{\cdot,i}\mathbf{V}(\omega)$ and $\mathbf{U}^*_{\cdot,i}\mathbf{V}'(\omega)$, the next question is: Which columns of $\mathbf{U}$ should form $\mathbf{W}$? Our basic approach is a greedy search maximizing the efficiency of $\mathcal{W}$ given by either $\eta_{\widehat{\boldsymbol{\theta}}}(\mathcal{W})$ or $\eta_{\widehat{\boldsymbol{\omega}}}(\mathcal{W})$. Often the greedy search will result in the optimal subspace for a fixed dimension, but there are rare exceptions, as Figure 4.3 shows. Even if the compression subspace is slightly suboptimal, the alternative is a computationally infeasible search over the $\binom{n}{m}$ possible subspaces of dimension m. The basic outline of this greedy search is given in Algorithm 4.1, but three details must be explained. First, as efficiency is undefined when $m < 2p$, we need an initial

Algorithm 4.1: Greedy subspace selection for exponential fitting.

Input : Parameters $\boldsymbol{\omega}, \mathbf{a} \in \mathbb{C}^p$, block dimensions $\mathbf{b}$, desired subspace dimension m or efficiency $\widehat{\eta}$, and current index set $\mathcal{I}$ (optional).

Output: Index set $\mathcal{I}$

$\mathcal{C} \leftarrow$ initial columns for each ω_j;
$\mathcal{I} \leftarrow \mathcal{I} \cup \mathcal{C}$;
while $|\mathcal{I}| < 2p$ **do**
    $\mathcal{C} \leftarrow$ candidate columns based on $\mathcal{I}$;
    for $k = 1, \ldots, p$ **do**
        $\mathcal{I} \leftarrow \mathcal{I} \cup \operatorname{argmax}_{i \in \mathcal{C}}$ eq. (4.36) ;
while $|\mathcal{I}| < m$ *or* $\eta_\omega(\mathbf{U}_{\cdot,\mathcal{I}}) < \widehat{\eta}$ **do**
    $\mathcal{C} \leftarrow$ candidate columns based on $\mathcal{I}$;
    $\mathcal{I} \leftarrow \mathcal{I} \cup \operatorname{argmax}_{i \in \mathcal{C}} q(\mathcal{I}, i, \boldsymbol{\omega}, \mathbf{a})$ for some quality function q ;

set of columns to start the greedy optimization. Second, for the purposes of speed, we need to restrict the set of candidate columns from the $n - m$ that make up the complement of $\mathcal{I}$ to a smaller set of candidates $\mathcal{C}$. And third, computing the efficiency of every candidate in $\mathcal{C}$ can be too expensive, so we briefly discuss the update formulas and heuristics developed in Section 4.2 as applied to exponential fitting. For clarity, we index the columns of $\mathbf{U}$ by pairs (j, k) referring to the block j and the kth column in block; i.e., $[\mathbf{F}_{b_j}^* \mathbf{v}_j(\omega)]_k$.

4.4.1 Initial Subspace

Efficiency is undefined when the compression subspace has fewer columns than there are parameters. Recalling the heuristic in which there is one parameter (4.26), good subspaces should contain rows with large entries of $\mathbf{U}^* \mathbf{F}(\boldsymbol{\theta})$. For the exponential fitting problem, these rows correspond to large rows of $\mathbf{F}_{b_j}^* \mathbf{v}_j(\omega)$ and $\mathbf{F}_{b_j}^* \mathbf{v}_j'(\omega)$. By knowing ω, we know approximately which blocks and the entry in the blocks that will

be largest; these should be in the initial column set. For both $\mathbf{F}_{b_j}^* \mathbf{v}_j(\omega)$ and $\mathbf{F}_{b_j}^* \mathbf{v}_j'(\omega)$, the largest row in each block corresponds to the column in $\mathbf{F}_{b_j}$ with frequency closest to $\operatorname{Im}\omega$; i.e., column (j,k), where

$$k = \lceil \operatorname{Im} \frac{\omega b_j}{2\pi} \bmod b_j - 1/2 \rceil.$$

The block with the largest component mostly depends on the sign of $\operatorname{Re}\omega$. If $\operatorname{Re}\omega < 0$, the first block contains the largest entry of $\mathbf{V}(\omega)$ and the block containing the $\lfloor -1/\operatorname{Re}\omega \rfloor$ entry contains the largest entry from $\mathbf{V}'(\omega)$. This does not necessarily mean that these rows of $\mathbf{U}^*\mathbf{V}(\omega)$ and $\mathbf{U}^*\mathbf{V}'(\omega)$ are largest – this depends on b_j and $\operatorname{Im}\omega$ as well (see the $\mathbf{G}_2$, $\omega = 2i$ case in Figure 4.3, there, the first block is not chosen) – but this is often close enough. Similarly, when $\omega \geq 0$, both $\mathbf{V}(\omega)$ and $\mathbf{V}'(\omega)$ are the largest in the last block. To generate the initial set of columns, we repeat this procedure for every ω_k, appending these indices to the set $\mathcal{I}$.

If this procedure does not produce a sufficiently large subspace, we continue to append columns according the one-dimensional heuristic for minimizing the covariance of ω in (4.36). This may not produce an optimal set of starting columns, as (4.36) is valid only in the limit that $\mathbf{W}^*\mathbf{V}(\omega)$ is large; however it is better than randomly picking from the candidate set.

4.4.2 Candidate Set Selection

We build the candidate set using the same heuristics as for the initial set: we want to include large rows of $\mathbf{U}^*\mathbf{V}(\omega)$ and $\mathbf{U}^*\mathbf{V}'(\omega)$. Looking at the closed form expressions for these quantities, (4.37) and (4.44), if we allow k to be a real number, both $[\mathbf{F}_{b_\ell}^* \mathbf{v}_\ell(\omega)]_k$ and $[\mathbf{F}_{b_\ell}^* \mathbf{v}_\ell'(\omega)]_k$ are continuous functions of k (modulo b_j). Thus, we should add new indices inside the block, adjacent to those already in $\mathcal{I}$ modulo b_j;

e.g.,

$$(j, k+1 \bmod b_j) \cup (j, k-1 \bmod b_j).$$

Similarly, if we allow s_j to be a continuous variable, then $[\mathbf{F}_{b_\ell}^* \mathbf{v}_\ell(\omega)]_k$ and $[\mathbf{F}_{b_\ell}^* \mathbf{v}'_\ell(\omega)]_k$ are continuous functions of s_j. Hence, we should add blocks to the candidate set that are adjacent to the blocks in $\mathcal{I}$. As we desire large rows inside these blocks, we should enter these new blocks with k corresponding to the nearest Fourier frequency to $\operatorname{Im} \omega$; e.g.,

$$(j \pm 1, k_\pm) \quad k_\pm \approx \operatorname{Im} \frac{\omega}{2\pi b_{j\pm 1}}.$$

This candidate selection heuristic is summarized in Algorithm 4.2. Figure 4.3 illustrates that this heuristic is obeyed by the optimal index sets $\mathcal{I}_m$ of dimension m for several different parameter values and parent coordinates. In this example, $2^{20} - 21$ possible subspaces were evaluated, and in every case the next index came from the candidate set. Only once were the subspaces not nested, an assumption of our greedy search, meaning the greedy search would have returned a suboptimal index set. However, this would be corrected on the next iteration.

4.4.3 Quality Measures

In building compression spaces for exponential fitting we have two choices for measuring efficiency: either using all the parameters $\boldsymbol{\theta} = [\boldsymbol{\omega}^\top, \mathbf{a}^\top]^\top$ by $\eta_{\widehat{\boldsymbol{\theta}}}$ or only frequencies $\boldsymbol{\omega}$ by $\eta_{\widehat{\boldsymbol{\omega}}}$. These yield similar subspaces, but using all parameters results in more accurate estimates of $\mathbf{a}$. However, computing $\eta_{\boldsymbol{\theta}}$ and $\eta_{\boldsymbol{\omega}}$ is expensive, requiring $\mathcal{O}(np^2 + p^3)$ operations. Thus we construct quality functions q that mimic the behavior of $\eta_{\boldsymbol{\theta}}$ and $\eta_{\boldsymbol{\omega}}$ but are easier to evaluate.

The easiest simplification sacrifices nothing. We note that the covariance of the full problem is constant with respect to $\mathcal{I}$, so maximizing $\eta_{\boldsymbol{\theta}}$ and $\eta_{\boldsymbol{\omega}}$ is equivalent to

Algorithm 4.2: Candidate selection algorithm for exponential fitting.

Input : $\mathcal{I}$ and exponential parameters $\boldsymbol{\omega}$

Output: Candidate set $\mathcal{C}$

$\mathcal{C} \leftarrow \emptyset$;

for $(j, k) \in \mathcal{I}$ **do**

  for $\omega \in \boldsymbol{\omega}$ **do**

    Add adjacent columns inside the current Fourier block:

    $\mathcal{C} \leftarrow \mathcal{C} \cup (j, k - 1 \bmod b_j) \cup (j, k + 1 \bmod b_j)$;

    Add columns with closest frequency in adjacent blocks:

    $\mathcal{C} \leftarrow \mathcal{C} \cup (j + 1, \lceil \operatorname{Im} \omega / 2\pi b_{j+1} \bmod b_{j+1} - 1/2 \rceil)$;

    $\mathcal{C} \leftarrow \mathcal{C} \cup (j - 1, \lceil \operatorname{Im} \omega / 2\pi b_{j-1} \bmod b_{j-1} - 1/2 \rceil)$;

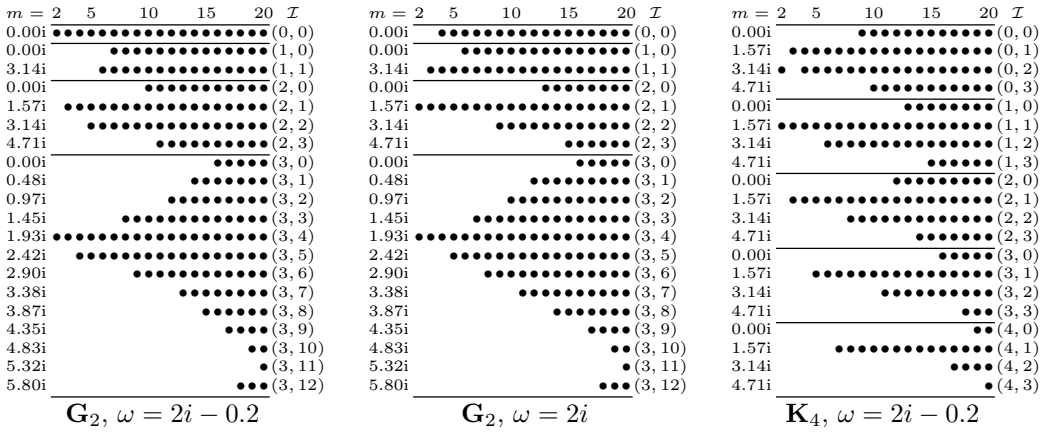


Figure 4.3 : Optimal columns from various parent coordinates computed by a combinatorial search. Along the top runs the dimension of the compression subspace; on the left, the frequency of the column in the Fourier block; on the right, the coordinates of the column; dots show which columns were selected for a given compression subspace dimension m. The left diagram shows a typical case where all the heuristics hold. The center diagram shows a case where the first block is not necessarily picked. The right diagram shows one of the rare cases where successive subspaces are not nested, i.e., $\mathcal{I}_m \not\subset \mathcal{I}_{m+1}$. Here $\mathcal{I}_2 \not\subset I_3$.

maximizing

$$q_g(\mathcal{I}, i, \boldsymbol{\omega}, \mathbf{a}) = -\sum_{j=1}^{p} \sigma_j(\mathbf{W}^* \mathbf{F}([\boldsymbol{\omega}, \mathbf{a}]))^{-2}, \qquad \mathbf{W} = \mathbf{U}_{\cdot, \mathcal{I} \cup \{i\}}, \tag{4.45}$$

$$q_s(\mathcal{I}, i, \boldsymbol{\omega}, \mathbf{a}) = -\sum_{j=1}^{p} \sigma_j \left(\mathbf{P}_{\mathcal{W}}(\boldsymbol{\omega}) \mathbf{W}^* \boldsymbol{\mathcal{V}}(\boldsymbol{\omega}) \bar{\times}_2 \mathbf{a}\right)^{-2}. \tag{4.46}$$

This avoids the $\mathcal{O}(np^2 + p^3)$ cost of computing the full covariance, and only costs $\mathcal{O}(mp^2 + p^3)$ operations per candidate. If we still wanted to evaluate $\eta_{\boldsymbol{\theta}}$ or $\eta_{\boldsymbol{\omega}}$, we could compute the full covariance using Theorem 4.1 to provide a closed form expression for entries of $\mathbf{F}(\boldsymbol{\theta})^* \mathbf{F}(\boldsymbol{\theta})$; e.g.,

$$\mathbf{F}(\boldsymbol{\theta})^* \mathbf{F}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{C} \end{bmatrix}, \quad \text{where}$$

$$[\mathbf{A}]_{j,k} = \overline{a_j} a_k \mathbf{V}'(\omega_j)^* \mathbf{V}'(\omega_k) = \overline{a_j} a_k \sum_{\ell=0}^{n-1} \ell^2 e^{(\overline{\omega_j} + \omega_k)\ell}$$

$$[\mathbf{B}]_{j,k} = \overline{a_j} \mathbf{V}'(\omega_j)^* \mathbf{V}(\omega_k) = \overline{a_j} \sum_{\ell=0}^{n-1} \ell e^{(\overline{\omega_j} + \omega_k)\ell}$$

$$[\mathbf{C}]_{j,k} = \mathbf{V}(\omega_j)^* \mathbf{V}(\omega_k) = \sum_{\ell=0}^{n-1} e^{(\overline{\omega_j} + \omega_k)\ell}.$$

Hence we could form $\mathsf{Tr}(\mathbf{F}(\boldsymbol{\theta})^* \mathbf{F}(\boldsymbol{\theta}))^{-1}$ in $\mathcal{O}(p^3)$ operations, but this form requires care when $\overline{\omega_j} + \omega_k$ is close to zero, due to rounding errors.

Other quality functions are much faster to evaluate than q_g and q_s. For the efficiency of $\boldsymbol{\theta} = [\boldsymbol{\omega}^\top, \mathbf{a}^\top]^\top$, we could compute q_g using the singular value update formula (4.24). Despite this being the fastest asymptotic approach, requiring only $\mathcal{O}(p)$ operations per candidate, the overhead of the MATLAB implementation makes this approach prohibitively expensive. Alternatively, we can use the drop (4.26):

$$q_{g1}(\mathcal{I}, i, \boldsymbol{\omega}, \mathbf{a}) = \frac{\mathbf{c}_i^* \mathbf{c}_i}{1 + \beta_i}, \tag{4.47}$$

	Code	Description	Operations
0.	$\mathbf{Q}, \mathbf{R} \leftarrow [\mathbf{U}_{\mathcal{I}_m}^* \mathbf{F}(\boldsymbol{\theta})]$	QR factorization update	$\mathcal{O}(mp)$
1.	$\mathbf{F}_i \leftarrow \mathbf{U}_i^* \mathbf{F}(\boldsymbol{\theta})$	candidate row	$\mathcal{O}(p)$
2.	$\mathbf{c}_i \leftarrow \mathbf{R}^{-1} \left(\mathbf{R}^{-*} \mathbf{F}_i^* \right)$		$\mathcal{O}(p^2)$
3.	$\beta_i \leftarrow \mathrm{Re}\left(\mathbf{c}_i^* \mathbf{F}_i^* \right)$		$\mathcal{O}(p)$
4.	$q_{g1}(\mathcal{I}_m, i, \boldsymbol{\omega}, \mathbf{a}) \leftarrow (1+\beta_i)^{-1} \sum_{j=0}^{2p-1} \lvert[\mathbf{c}_i]_j\rvert$	Covariance of $\widetilde{\boldsymbol{\omega}}$ and $\widetilde{\mathbf{a}}$	$\mathcal{O}(p)$
5.	$q_{s1}(\mathcal{I}_m, i, \boldsymbol{\omega}, \mathbf{a}) \leftarrow (1+\beta_i)^{-1} \sum_{j=0}^{p-1} \lvert[\mathbf{c}_i]_j\rvert$	Covariance of $\widetilde{\boldsymbol{\omega}}$	$\mathcal{O}(p)$
		initialization cost:	$\mathcal{O}(mp)$
		per candidate cost:	$\mathcal{O}(p^2)$

Algorithm 4.3: Computing the covariance drop for both $\boldsymbol{\theta} = [\boldsymbol{\omega}^\top, \mathbf{a}^\top]^\top$ and $\boldsymbol{\omega}$ alone. Step 0 updates the QR factorization from the existing QR factorization of $\mathbf{U}_{\mathcal{I}_{m-1}} \mathbf{F}(\boldsymbol{\theta})$ once [56, §12.5.3]. The remaining steps are repeated for each candidate columns $i \in \mathcal{C}$ using either line 4 or line 5 to compute the covariance drop.

where $\mathbf{c}_i = \mathbf{R}^{-1}\mathbf{R}^{-*}\mathbf{F}_i^*$, $\beta_i = \mathbf{c}_i^*\mathbf{F}_i^*$, and the QR Factorization $\mathbf{QR} = \mathbf{F}_{\mathcal{I}_{m-1}}$. This approach requires $\mathcal{O}(p^2)$ operations per candidate, plus a one-time cost of $\mathcal{O}(mp^2)$ operations for the QR Factorization. When m is small, it may be advantageous to update the QR Factorization rather than computing the QR factorization of $\mathbf{F}_{\mathcal{I}_{m-1}}$ afresh, as this requires $\mathcal{O}(mp)$ operations but $\mathcal{O}(m^2)$ storage; see Algorithm 4.3 and [56, § 12.5.3]. Both of these approaches appear to be stable; each provides the right next index in the numerical experiments of Figure 4.4. It is tempting to add only the largest row of $\mathbf{F}$ following (4.27), resulting in the quality function

$$q_{g2}(\mathcal{I}, i, \boldsymbol{\omega}, \mathbf{a}) = -\|\mathbf{F}_i\|_\infty. \tag{4.48}$$

Unfortunately, this quality function almost always picks the wrong index, as Figure 4.4 illustrates.

Quality functions considering only the covariance of $\boldsymbol{\omega}$ follow as similar pattern to those based on the covariance of $\boldsymbol{\theta}$, except numerical stability is a larger concern. The singular value update formula (4.29) appears stable, but again the overhead is

prohibitive in MATLAB. We can use the drop (4.32) to provide a quality function:

$$q_{s1}(\mathcal{I}, i, \boldsymbol{\omega}, \mathbf{a}) = \frac{1}{1+\alpha_i} \frac{\mathbf{c}_i^* \mathbf{c}_i}{1+\beta_i(1+\alpha_i)}. \tag{4.49}$$

Unfortunately this approach is often unstable, as Figure 4.4 illustrates. The Sherman-Morrison formula is often ill-conditioned, and the two nested applications to arrive at (4.32) exacerbate this problem in comparison to q_{g1}. A stable alternative is to use (4.47). The upper left block in $\mathsf{Cov}[\widetilde{\boldsymbol{\theta}}]$ is perturbed by $\widehat{\mathbf{c}}_i \widehat{\mathbf{c}}_i^*$ where $\widehat{\mathbf{c}}_i = [\mathbf{c}_i]_{0:p-1}$, hence the trace of $\mathsf{Cov}[\widetilde{\boldsymbol{\omega}}]$ is modified by $\widehat{\mathbf{c}}_i^* \widehat{\mathbf{c}}_i$. This leads to the quality function

$$q_{sg}(\mathcal{I}, i, \boldsymbol{\omega}, \mathbf{a}) = \frac{\widehat{\mathbf{c}}_i^* \widehat{\mathbf{c}}_i}{1+\beta_i}. \tag{4.50}$$

This strategy works well; in the numerical experiments of Figure 4.4, this quality function gives the right subspace in every instance. Finally, we consider the one parameter approximation from (4.36):

$$q_{s2}(i) = \max_j \left| \mathbf{U}_{\cdot,i}^* \mathbf{V}(\omega_j) - \mathbf{U}_{\cdot,i}^* \mathbf{V}'(\omega_j) \frac{\chi_j}{\kappa_j} \right|, \tag{4.51}$$

where $\chi_j = (\mathbf{U}_{\cdot,\mathcal{I}}^* \mathbf{V}(\omega_j))^* (\mathbf{U}_{\cdot,\mathcal{I}} \mathbf{V}'(\omega_j))$ and $\kappa_j = \|\mathbf{U}_{\cdot,\mathcal{I}}^* \mathbf{V}(\omega_j)\|_2^2$ are precomputed and shared between candidate indices. For the same reasons as q_{g2}, this is not an appealing quality function.

In conclusion, we recommend using q_{g1} to maximize the efficiency of $\boldsymbol{\theta}$ and q_{sg} to maximize the efficiency of $\boldsymbol{\omega}$. If the overhead of invoking `dlaed4` to solve the seculare equation can be reduced, updating the singular values would be the best approach: the asymptotic cost per candidate is only $\mathcal{O}(p)$ for the SVD update approach instead of $\mathcal{O}(p^2)$ for the drop approach.

4.5 Block Dimensions for Exponential Fitting

Our choice of block dimensions greatly affects the efficiency of the generated compression subspaces. Ideally, we would like to choose block dimensions that maximize

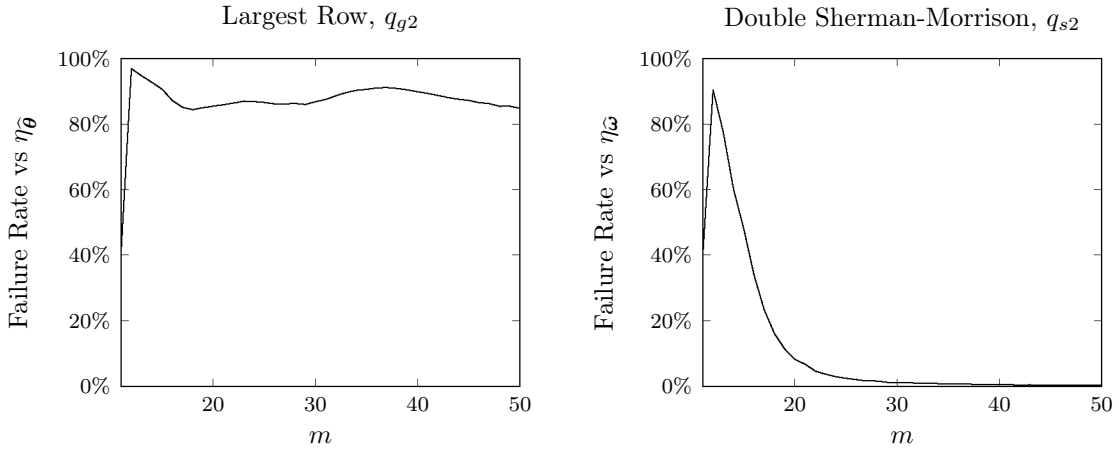


Figure 4.4 : Failure rate for fast updates and heuristics. A method fails if, starting from a subspace of dimension $m-1$, the m dimensionial subspace generated by the heuristic does not match subspace constructed by greedily maximizing $\eta_{\widehat{\boldsymbol{\theta}}}$ or $\eta_{\widehat{\boldsymbol{\omega}}}$. In this example, five exponentials were randomly chosen with ω_j distributed uniformly in the range $[-0.1, 0] \times [0, 2\pi]i$, $|a_j| = 1$ with a uniform random phase, and $n = 100$. In 3×10^4 trials, q_{g1} and q_{sg} matched the $\eta_{\widehat{\boldsymbol{\theta}}}$ and $\eta_{\widehat{\boldsymbol{\omega}}}$ generated subspace exactly. The double Sherman-Morrison update (4.31) improves as the norm of $\mathbf{S}_{\mathcal{I}_{m-1}}$ improves, whereas choosing the largest row frequently results in the wrong subspace.

the worst case efficiency for all possible parameters for a fixed subspace dimension. For simplicity, we consider the single exponential case, seeking $\mathbf{b}$ that solves[2]

$$\max_{\mathbf{b}} \min_{\substack{\boldsymbol{\theta}=[\omega,a]^\top \\ \omega \in \mathbb{C}^-}} \max_{|\mathcal{I}|=m} \eta_{\boldsymbol{\theta}}([\mathrm{diag}(\mathbf{F}_{b_0}, \mathbf{F}_{b_1}, \ldots)]_{\cdot,\mathcal{I}}) \quad \text{such that} \quad \sum_j b_j = n. \tag{4.52}$$

Solving (4.52) is difficult: for each value of ω, we must solve a combinatorial problem for $\mathcal{I}$. Even if we use a greedy search to pick $\mathcal{I}$, efficiency is a highly oscillating function of ω (see Figure 4.9), and even if we could solve this problem, we would need a solution for each n and m. Rather, we use this problem as a guide for informing heuristics for picking block dimensions. These heuristics are based on intuition gained from several examples: the existing, fixed compression subspaces of truncation and

[2]We may take $a = 1$ as efficiency for the single exponential case is independent of a.

decimation (Section 4.5.1) and new sets of parent coordinates that use *fixed blocks* $b_j = r$ (Section 4.5.2) and *geometric blocks* ($b_j = r^j$) (Section 4.5.3). Using intuition from these examples, we build block dimensions that solve an approximation of (4.52) (Section 4.5.4). We term these block dimensions *rationally chosen blocks*, since they built to solve (4.52) rather than some preconcived rule. These rationally constructed subspaces turn out to have comparable performance to the much simpler geometric blocks. Attempts to solve (4.52) starting with block dimensions provided by these heuristics provides only marginally better subspaces (Section 4.5.5). For simplicity, we recommend using geometric blocks with a growth rate of $r = 4$, i.e., $\mathbf{G}_4(n)$. However, if $\boldsymbol{\omega}$ is restricted to a small subset of the complex plane, there may be better parent coordinates than $\mathbf{G}_4(n)$.

4.5.1 Truncation and Decimation

An existing compression approach is to either truncate the measurements, removing the last $n - m$ entries in $\widetilde{\mathbf{y}}$, or decimate the measurements, keeping only every rth measurement. Both correspond to taking parent coordinates with n size 1 blocks. However, there is no tailoring of subspaces specific to $\boldsymbol{\omega}$; these coordinates are fixed from the outset. This avoids the cost of constructing a subspace, but yields inefficient subspaces in comparison to tailored compression spaces. As Figure 4.5 illustrates, truncation works well when decay is strong as there is little information in the tail of the signal; decimation works well when decay is weak as a longer base line allows more accurate frequency measurements. However, decimation risks aliasing exponentials with high frequencies, making them appear with a lower frequency.

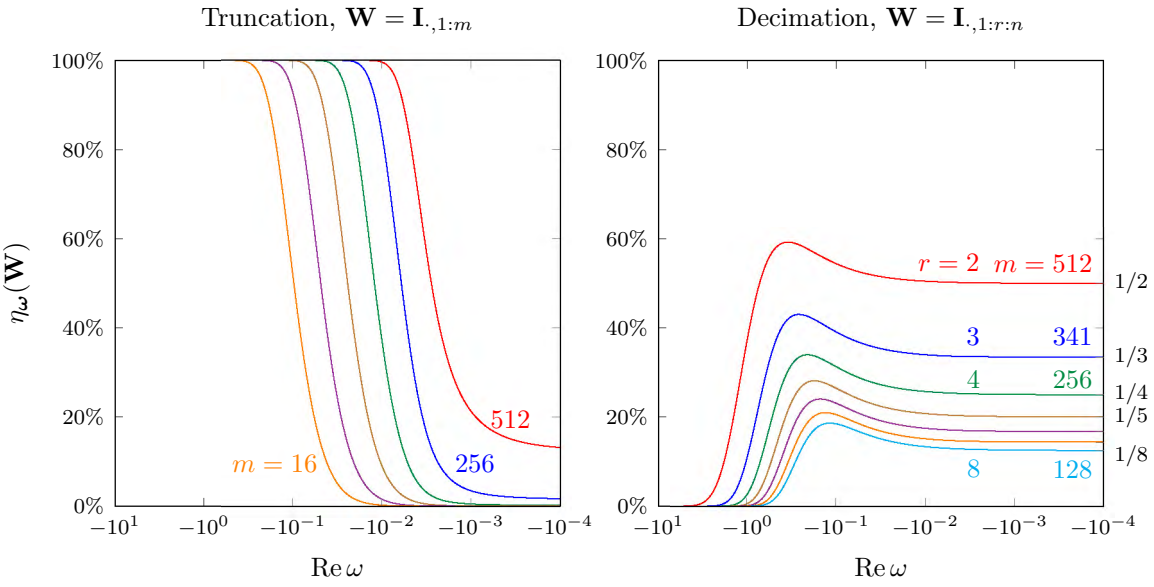


Figure 4.5 : The efficiency of truncating and decimating measurements. The left plot shows efficiency for various lengths of the truncated $\widetilde{\mathbf{y}}$ in powers of two; the right plot show the same for decimation rates two through eight. In these examples $n = 1024$, and since there is only a single exponential, performance is independent of $\mathrm{Im}\,(\omega)$.

4.5.2 Fixed Blocks

Inspired by the short time Fourier transform, the first set of novel block dimensions splits $\mathbf{U}$ into equal sized Fourier blocks, giving parent coordinates

$$\mathbf{K}_r(n) = \operatorname{diag}(\mathbf{F}_r, \mathbf{F}_r, \ldots, \mathbf{F}_r, \mathbf{F}_{r \bmod n}). \tag{4.53}$$

There are two important limits: $r = n$ yields the Fourier transform ($\mathbf{K}_n(n) = \mathbf{F}_n$) and $r = 1$ yields the identity matrix ($\mathbf{K}_1(n) = \mathbf{I}$).

As Figure 4.6 shows, this approach does not yield efficient subspaces for all amounts of damping. When $r = n$, good subspaces are generated for lightly damped exponentials ($-1 \ll \operatorname{Re}\omega < 0$), but poor subspaces are generated for strongly damped exponentials ($\operatorname{Re}\omega \ll 0$). Conversely, taking $r = 1$ has the opposite behavior: it generates good subspaces for strong damping but bad subspaces for weak damping. The middle way, taking $r \approx \sqrt{n}$ does well for intermediate values, but scales poorly as $n \to \infty$, as Figure 4.7 shows.

4.5.3 Geometric Blocks

To better choose blocks, consider the two limits for the real part of ω. If $\operatorname{Re}\omega \ll 0$, then most of the energy in $\mathbf{V}(\omega)$ and $\mathbf{V}'(\omega)$ is concentrated in the first few entries. As these vectors decay rapidly, they are strongly non-periodic and, correspondingly, their energy will be spread through out all the entires of $\mathbf{F}^*_{b_0}\mathbf{v}_0(\omega)$. So all the entries of the first block will likely need to be in the index set $\mathcal{I}$; hence a large first block would be inefficient. Conversely, if damping is light ($-1 \ll \operatorname{Re}\omega < 0$), then $\mathbf{V}(\omega)$ is well approximated by a few columns of $\mathbf{F}^*_{b_j}\mathbf{v}_j(\omega)$. The larger the block, the better the resolution, as $\operatorname{Im}\omega$ will be closer to a Fourier frequency $\phi_k = e^{2\pi i k/b_j}$, and the larger the block, the larger the biggest entry in $\mathbf{F}^*_{b_j}\mathbf{v}_j(\omega)$. As weak damping implies $\mathbf{V}(\omega)$

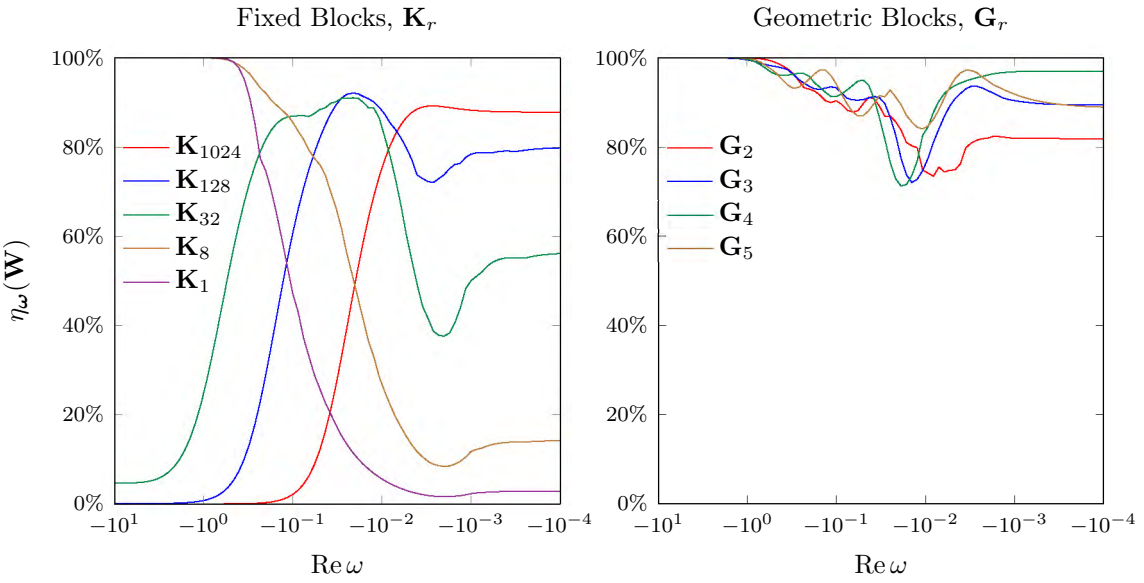


Figure 4.6 : Efficiency for the best $m = 10$ columns chosen using Algorithm 4.1 from fixed and geometric block parent coordinates, where $n = 1024$ and $\operatorname{Im}\omega = 1$.

will still be large at the end of $\mathbf{V}(\omega)$, it does not matter where large Fourier blocks are placed. Hence, we should place small blocks at the beginning to resolve strong decay, and large Fourier blocks at the end to resolve weak damping (see Figure 4.8 for a visualization).

In between, inspired by the exponentials present, we increase the block dimensions geometrically:

$$\mathbf{G}_r(n) = \operatorname{diag}(\mathbf{F}_1, \mathbf{F}_r, \mathbf{F}_{r^2}, \ldots, \mathbf{F}_{r^{q-1}}, \mathbf{F}_s), \qquad s = n - \frac{r^q - 1}{r - 1}, \quad s > r^q. \tag{4.54}$$

Parent coordinates from this family generate good subspaces for a variety of decay rates, as Figure 4.6 illustrates for a fixed value of $\operatorname{Im}\omega$. As Figure 4.9 shows performance does vary with $\operatorname{Im}\omega$, but the effect diminishes with increasing subspace dimension. In comparison to fixed blocks, geometric blocks gain performance rapidly as subspace dimension grows, essentially independent of n, as Figure 4.7 illustrates.

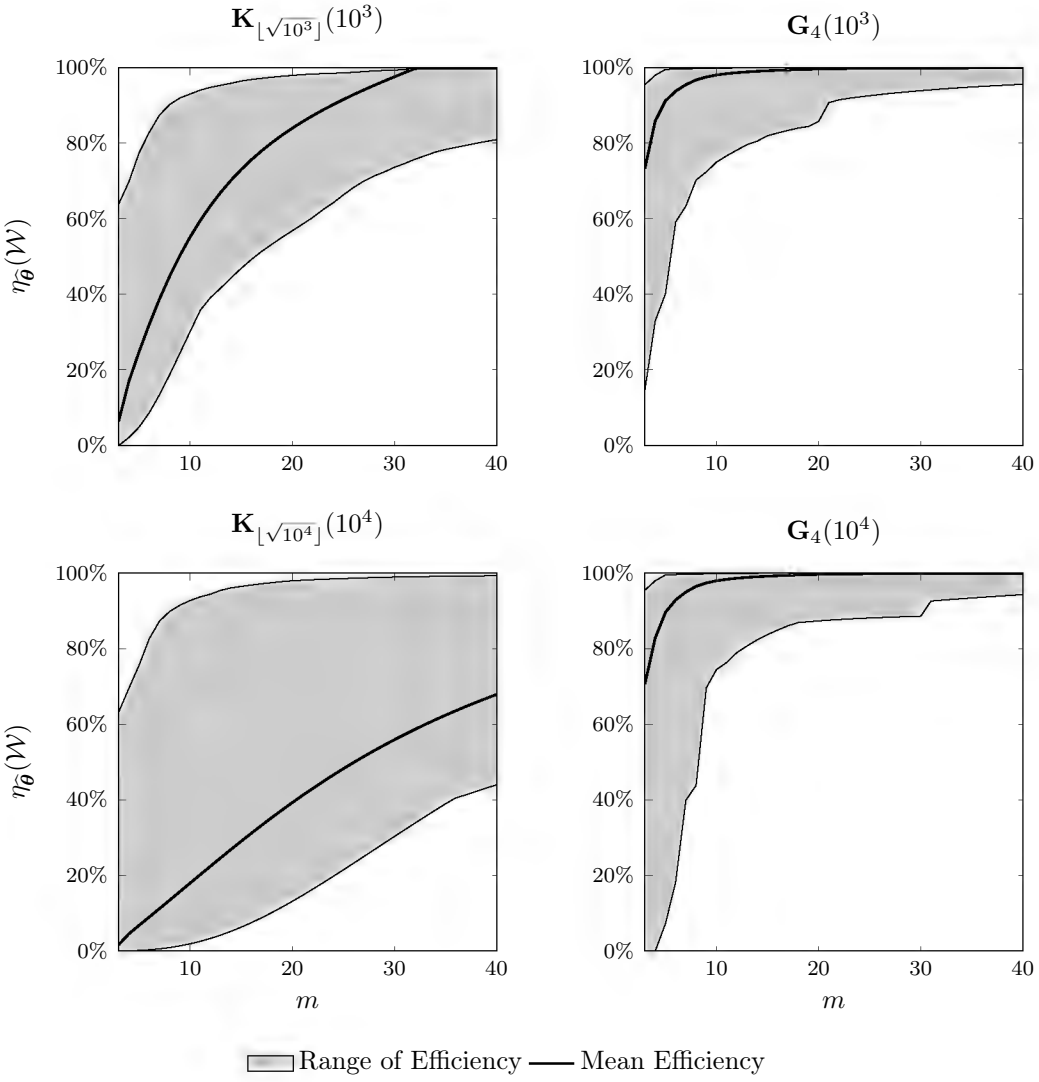


Figure 4.7 : Upper and lower bounds on the efficiency subspaces generated from parent coordinates of fixed Fourier blocks and geometric blocks. As the second row shows, fixed Fourier blocks lose efficiency as n increases, whereas geometric blocks maintain similar performance. In this example we use a single random exponential uniformly distributed in $[-1, 0] \times [0, 2\pi)i$ for 10,000 trials.

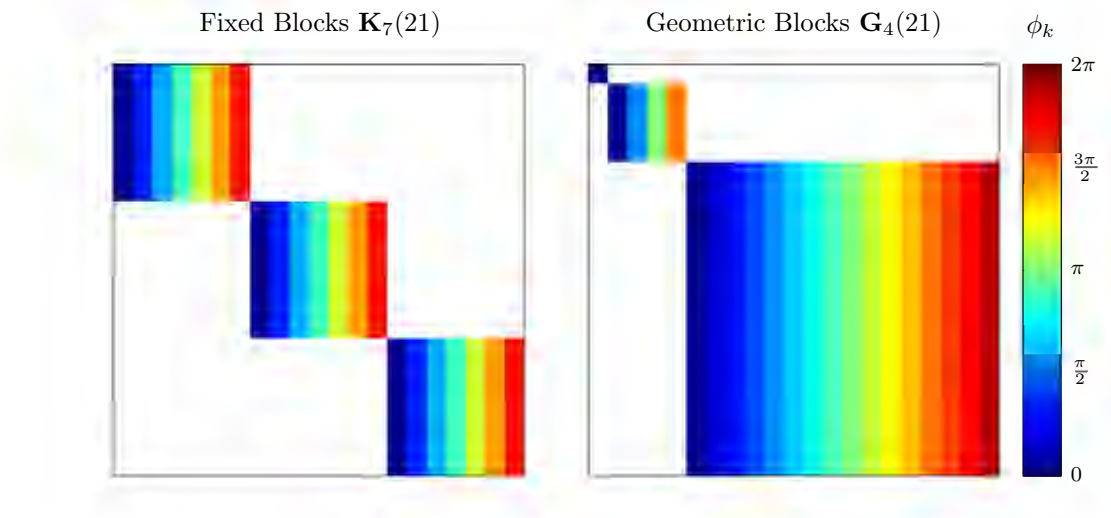


Figure 4.8 : A visualization of block Fourier matrices. Each colored square is a Fourier block and columns in each block are colored by their Fourier frequency $\phi_k = 2\pi ik/b_j$. Geometric blocks provide finer resolution of $\operatorname{Im}\omega$ than fixed sized blocks.

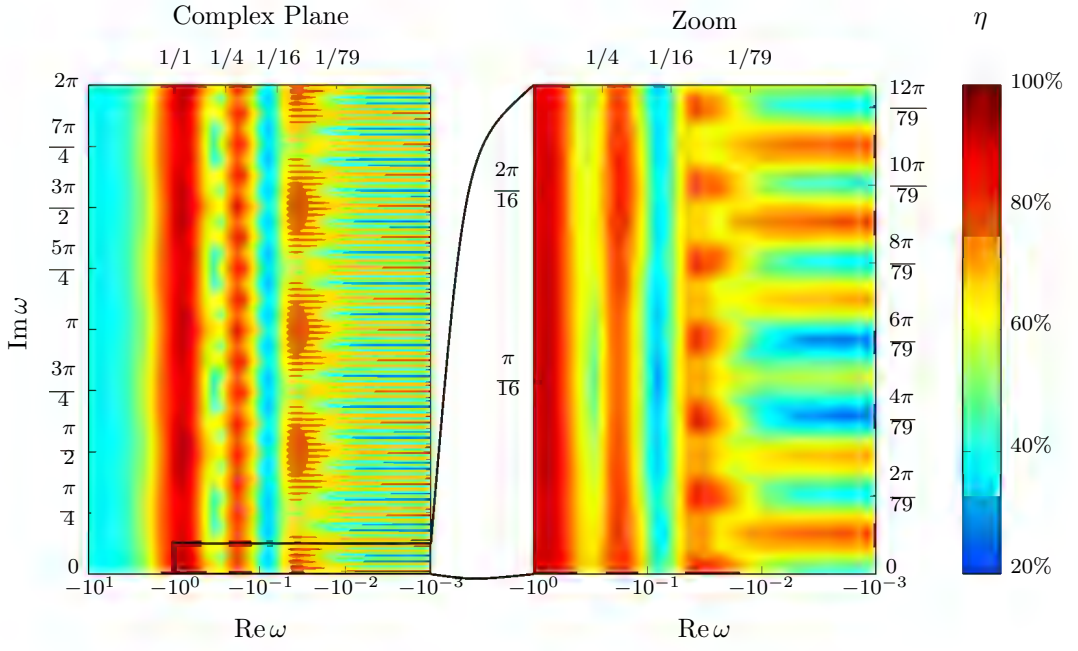


Figure 4.9 : Efficiency of the best three dimensional subspace from $\mathbf{G}_4(100) = \operatorname{diag}(\mathbf{F}_1, \mathbf{F}_4, \mathbf{F}_{16}, \mathbf{F}_{79})$ as a function of ω. Efficiency oscillates with $\operatorname{Im}\omega$, with periods corresponding to block dimensions, and reaches maxima near the reciprocals of the block dimensions.

4.5.4 Rationally Chosen Blocks

The principles that spawned geometric blocks did not specify how blocks should grow in size, only that they should. Rather than arbitrarily having blocks grow geometrically, we prefer a rational basis for choosing block dimensions b_j. As finding block dimensions maximizing the worst case efficiency (4.52) is difficult, we choose block dimensions that maximize a proxy problem we can solve – that of maximizing the largest row of $\mathbf{U}^*\mathbf{V}(\omega)$.

We start with the case where ω is real and seek

$$\max_{\mathbf{b}} \max_{i} |[\mathbf{U}(\mathbf{b})^*\mathbf{V}(\omega)]_i| \tag{4.55}$$

for a fixed $\omega \in \mathbb{R}^-$. The largest entry in each block corresponds to the zero frequency part of the Fourier transform, which, for the jth block, is

$$p_{\text{best}}(\omega, s_j, b_j) := \max_i |\mathbf{F}_{b_j}^* \mathbf{v}_j(\omega)| = \sum_{k=s}^{s_j+b_j-1} \frac{e^{k\omega}}{\sqrt{b_j}} = \frac{e^{s_j\omega}}{\sqrt{b_j}} \frac{1-e^{b_j\omega}}{1-e^{\omega}}, \quad \omega \neq 0. \tag{4.56}$$

In the limit that $\omega \to -\infty$, we know that the row with the largest block will be the first block. Then $\max_i |[\mathbf{U}(\mathbf{b})^*\mathbf{V}(\omega)]_i| = p_{\text{best}}(\omega, s_0, b_0) = b_0^{-1/2}$. Maximizing this, we choose $b_0 = 1$, and hence, $s_1 = 1$. Next we consider $p_{\text{best}}(\omega, s_1, b_1)$ for multiple values of b_1, as illustrated on the left of Figure 4.10. Each choice of $p_{\text{best}}(\omega, s_1, b_1)$ is monotonically increasing in ω, but intersects $p_{\text{best}}(\omega, s_0, b_0)$ at a different values of ω. Later blocks will not be able to surpass the value of $p_{\text{best}}(\omega, s_1, b_1)$ for small ω, so we choose b_1 such that $p_{\text{best}}(\omega, s_1, b_1) \geq p_{\text{best}}(\omega, s_0, b_0)$ for the smallest possible ω; in this case, $b_1 = 4$. We then repeat this procedure to find b_2, b_3, etc., resulting in the *best blocks*

$$\begin{aligned} \mathbf{b}_{\text{best}} = 1,\ 4,\ 17,\ 71,\ 298,\ 1250,\ 5242,\ 21984,\ 92198,\ 386668, \\ 1621642,\ 6800984,\ 28522561,\ 119620719,\ 501676621, \ldots. \end{aligned} \tag{4.57}$$

The largest of the resulting curves of $p_{\text{best}}(\omega, s_j, b_j)$ is $\max_i |[\mathbf{U}(\mathbf{b}_{\text{best}})^* \mathbf{V}(\omega)]_i|$ and approximates $\|\mathbf{V}(\omega)\|_2$, as shown in the right side of Figure 4.10.

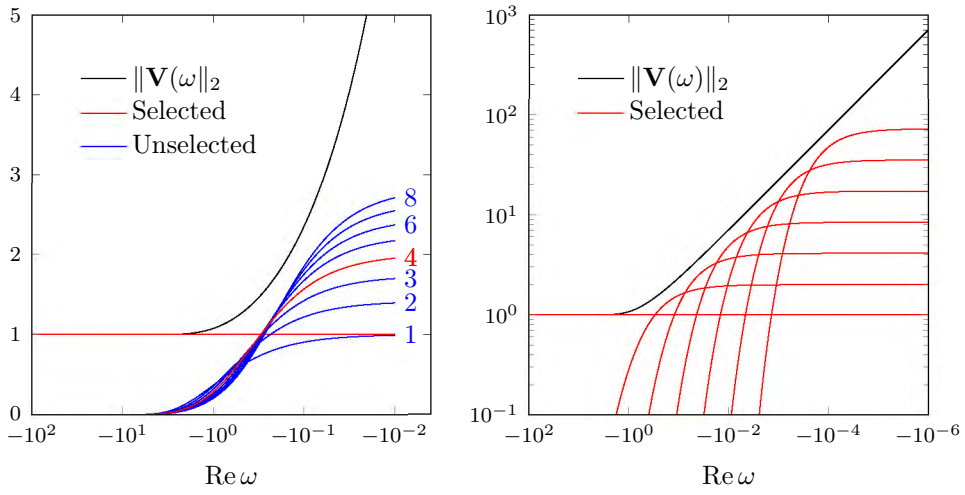


Figure 4.10 : Picking the block dimensions for the 'best' block parent coordinates. On the left, the horizontal red line shows the partial sum of the first block $p_{\text{best}}(\omega, 1, 0)$. The other red and blue lines show $p_{\text{best}}(\omega, 1, b_1)$ for various b_1 values ($b_1 = 1, 2, \ldots, 8$) indicated by the number to the right of the curve. The choice intersecting $p_{\text{best}}(\omega, 1, 0)$ earliest is $b_1 = 4$ (in red). The value for $p_{\text{best}}(\omega, s_j, b_j)$ is bounded above by $\|\mathbf{V}(\omega)\|_2$ (for $n = \infty$), shown in black. On the right, we see a sequence of p_{best} resulting from optimal choices of s_j and b_j. The maximum of these curves approximates $\|\mathbf{V}(\omega)\|_2$.

We call these the *best* blocks because restricting to $\omega \in \mathbb{R}^-$ is a best case scenario. For any Fourier block, 0 is Fourier frequency and $\|\mathbf{F}_{b_j}^* \mathbf{v}_j(\alpha + i\beta)\|_\infty$ attains one of its b_j maxima over β when $\beta = 0$ (the remaining being the b_j roots of unity). In contrast, the worst case scenario is when β is precisely between two Fourier frequencies. We can define the largest entry for the worst case imaginary part of ω, in an analogous way to p_{best}:

$$p_{\text{worst}}(\omega, s_j, b_j) := \left| \sum_{k=s}^{s_j+b_j-1} \frac{e^{k\omega + i\pi/b_j}}{\sqrt{b_j}} \right|. \tag{4.58}$$

Following the same procedure to generate $\mathbf{b}_{\text{best}}$, using p_{worst} instead, results in the

worst blocks

$$\mathbf{b}_{\text{worst}} = 1,\ 11,\ 60,\ 319,\ 1279,\ 5119,\ 20479,\ 81919,\ 327679,\ 1310719,\\ 5242879,\ 20971519,\ 83886079,\ 335544319,\ 1342176956, \ldots. \tag{4.59}$$

The best and worst blocks were generated by maximizing $\mathbf{U}(\mathbf{b})^*\mathbf{V}(\omega)$, but we could also consider the derivative $\mathbf{U}(\mathbf{b})^*\mathbf{V}'(\omega)$. The analogous functions to p_{best} and p_{worst} are not monotonic for the first two blocks. Defining the first two entries using the same values as in $\mathbf{b}_{\text{best}}$ and $\mathbf{b}_{\text{worst}}$ results in the sequences

$$\mathbf{b}'_{\text{best}} = 1,\ 4,\ 7,\ 18,\ 45,\ 113,\ 285,\ 717,\ 1806,\ 4548,\ 11452,\ 28837,\ 72615, \ldots;$$

$$\mathbf{b}'_{\text{worst}} = 1,\ 11,\ 27,\ 84,\ 265,\ 838,\ 2650,\ 8381,\ 26506,\ 83828, \ldots.$$

None of the sequences in this section correspond to known integer sequences, but all grow approximately geometrically with a growth coefficient around four. Figure 4.11 illustrates the performance for each of set of block dimensions. If n is not the sum of the block dimensions, the last block is enlarged (as with geometric blocks). For example, if $n = 100$ then $b_{\text{best}}(100) = [1,\ 4,\ 17,\ 78]$. In this example, each block sequence has similar worst case performance, but best blocks has a larger mean performance, similar to $\mathbf{G}_4$.

There are many alternative ways to construct sets of block dimensions. For example, we could work from $\omega = 0$ towards $-\infty$ (customized to a particular value of n) or we could balance performance over a range rather than seeking the block dimension with the earliest intersect. However, as these procedures only consider the largest row, and not a larger subspace, none of these procedures are likely to yield significantly better compression subspaces.

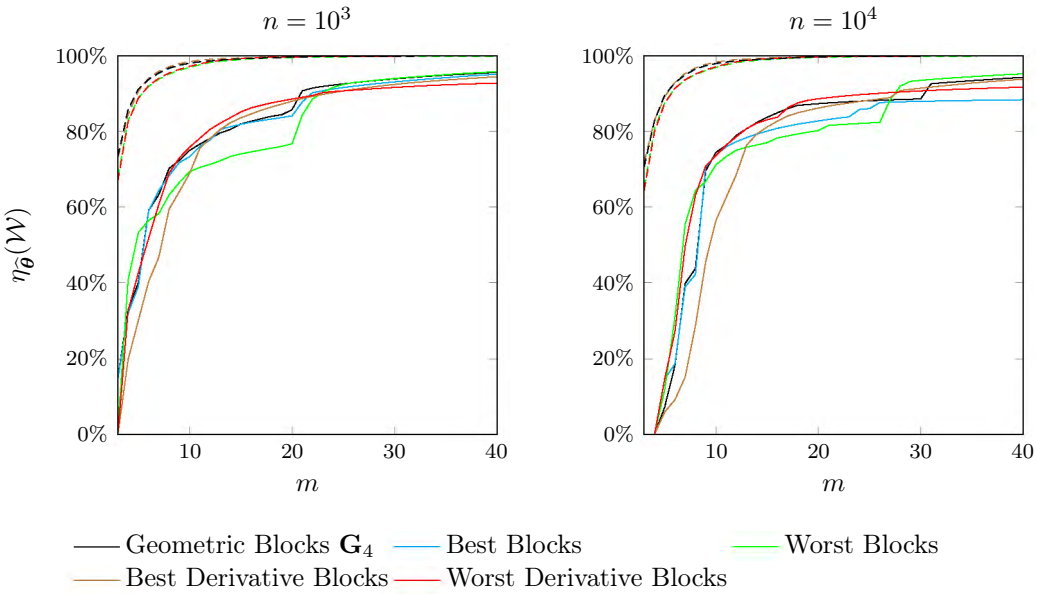
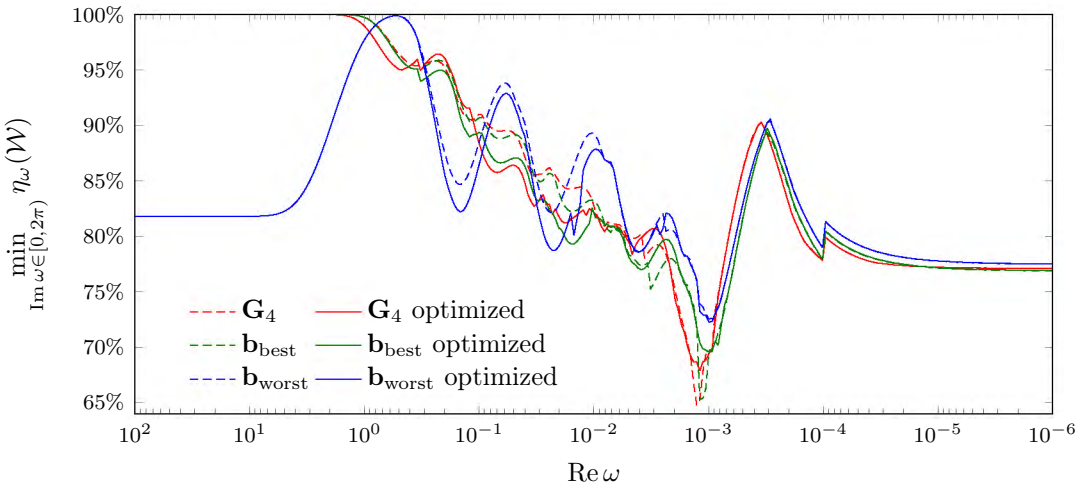


Figure 4.11 : Comparison of tailored compression subspace efficiency for geometric blocks, best blocks, and worst blocks. As in Figure 4.7, solid curves show the worst case subspace efficiency for a randomly generated $\omega \in [-1, 0] \times [0, 2\pi)i$ as a function of subspace dimension m; dashed curves show the mean efficiency. The worst case performance for each of these sets of blocks are about the same, with geometric blocks and best blocks having similar mean performance.



$\mathbf{G}_4(10^4)$	1	4	16	64	256	1024	8635
optimized	1	3	17	64	256	1024	8635
$\mathbf{b}_{\mathrm{best}}(10^4)$	1	4	17	71	298	1250	8359
optimized	1	4	17	70	298	1251	8359
$\mathbf{b}_{\mathrm{worst}}(10^4)$	1	11	60	319	1279	8330	
optimized	1	11	60	320	1279	8329	

Figure 4.12 : Optimized sets of block dimensions maximizing (4.52). The optimized block dimensions improve the worst case performance by about 5% for $\mathbf{G}_4$ and $\mathbf{b}_{\mathrm{best}}$, but decrease efficiency for other values of $\mathrm{Re}\,\omega$. In contrast, the performance of $\mathbf{b}_{\mathrm{worst}}$ only mildly improves, but also decreases efficiency for other values of $\mathrm{Re}\,\omega$. Blocks that increased in size are marked in red, those that decreased in blue. In this example, $n = 10,000$.

4.5.5 Optimizing Block Dimensions

Optimizing block dimensions is difficult. Minimizing the worst-case subspace efficiency requires first finding it, and as Figure 4.9 shows, efficiency of the best subspace is a highly oscillatory function of $\mathrm{Im}\,\omega$. Fortunately, the oscillations have periods corresponding to the block dimensions, allowing us to estimate the worst-case ω. Taking uniform random samples on a line with a fixed real part, we fit a cosine series with periods corresponding to the block dimensions. The minimum of this series is then used as the starting value for an optimization iteration (MATLAB's `fminsearch`), that

returns the minimum efficiency ω for a given $\operatorname{Re}\omega$. This optimization is then repeated over a grid of $\operatorname{Re}\omega$ to find the worst-case ω in the left half plane. The results are fed to an outer optimization routine that adjusts the size of the blocks by ± 1. The block dimensions optimized by this procedure are shown in Figure 4.12. The block dimensions do not move far from their initial values, each initial set of blocks shifting a single entry between two blocks in the optimized blocks. The resulting parent coordinates have slightly better worst-case performance, but decrease performance for other $\operatorname{Re}\omega$.

Due to the cost of computing optimum block dimensions and the minimal improvement, we recommend using a precomputed set of block dimensions. As the best and worst blocks do not do substantially better than $\mathbf{G}_4(n)$, we recommend using $\mathbf{G}_4(n)$ for all problems.

4.6 Linking Compression and Optimization

Compression can introduce many spurious local minima away from where the compression space is tailored. This effect is especially dramatic when small subspaces are drawn from the geometric block parent coordinates, as illustrated in Figure 4.13. To prevent convergence to a spurious minimum, as $\boldsymbol{\theta}$ moves during optimization the compression space must remain efficient.

We propose two approaches to maintain an efficient compression space as $\boldsymbol{\theta}$ is optimized. The faster approach is to precompute an efficient compression space based on prior knowledge of $\widehat{\boldsymbol{\theta}}$. If we know $\widehat{\boldsymbol{\theta}} \in \mathcal{T} \subset \mathbb{C}^p$, we can build a fixed subspace $\mathcal{W}$ where $\eta_{\boldsymbol{\theta}}(\mathcal{W}) > \widehat{\eta}$ for all $\boldsymbol{\theta} \in \mathcal{T}$. Then if the initial estimate is sufficiently close, iterations of $\boldsymbol{\theta}$ should remain in $\mathcal{T}$. To build this subspace, we can sample $\boldsymbol{\theta} \in \mathcal{T}$ uniformly, and for each sample append indices maximizing $\eta_{\boldsymbol{\theta}}(\mathcal{W})$, as described in

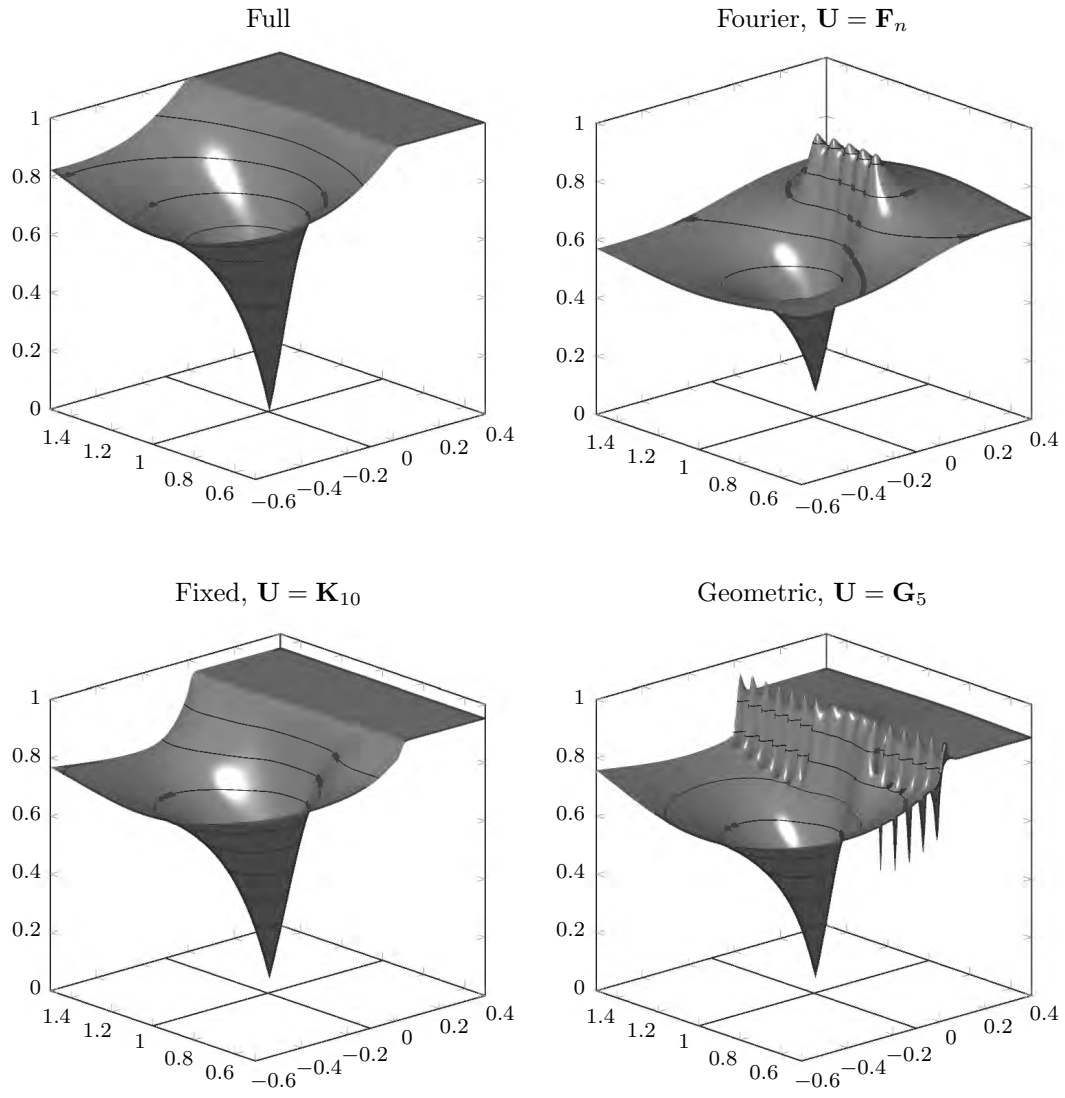


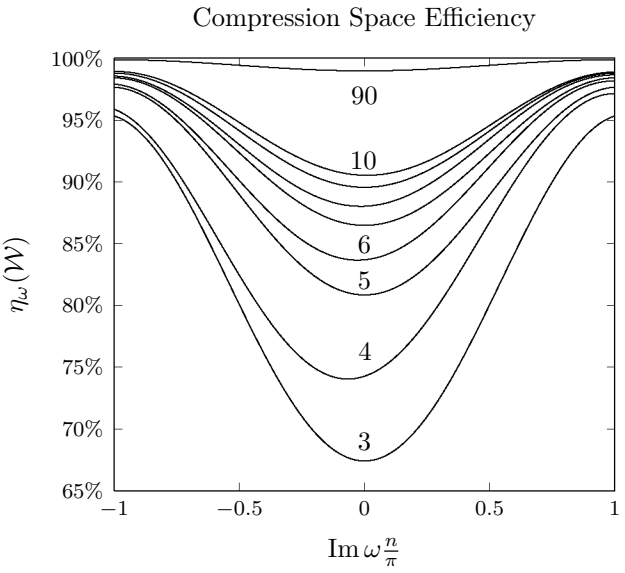
Figure 4.13 : Shape of the compressed residual $\|\mathbf{W}^*[\mathbf{f}(\omega)-\mathbf{f}(1i-0.1)]\|_2$ for different parent coordinates. In this example $\omega = 1i-0.1$, $n = 100$, and subspaces of dimension six were built.

Section 4.4, until the threshold is reached. However, this precomputed subspace approach may not be viable if $\mathcal{T}$ is unknown or $\mathcal{T}$ is large, as the dimension of $\mathcal{W}$ may be on the order of n. Alternatively, we can refine the compression space alongside the optimization of $\boldsymbol{\theta}$, interspersing updates to $\mathcal{W}$ between iterations of $\boldsymbol{\theta}$. The following two sections give an example of each approach.

4.6.1 Precomputed Compression Subspaces

In this example, we consider the canonical signal processing problem of determining the frequency $\omega \in i\mathbb{R}$ of a single undamped exponential $y_k = a^{\omega k}$ from n measurements. This problem appears in many real-time applications (e.g., Doppler radar) and has been a topic of continued interest in electrical engineering since the 1970s [131]. The simplest approach uses the largest entry of the Fourier transform $k = \text{argmax}_j \, |[\mathbf{F}^* \mathbf{y}]_j|$, estimating $\widetilde{\omega} = 2\pi i k/n$ (see discussion in Section 3.4.3). The resolution of this approach is limited to $\pm \pi i/n$, leading many authors to seek high resolution estimates using interpolation (e.g., [81, 128]), an iterative procedure (e.g., [2, 106]), or a combination of the two (e.g., [107]). The goal is a fast algorithm for computing ω in real time that can be constructed either in hardware or low level-software. In the following example, we show how compression can provide a more accurate and faster estimate of ω. Moreover, unlike previous algorithms, compression allows us to prescribe the accuracy of the estimator arbitrarily.

For parent coordinates, we choose the Fourier matrix $\mathbf{F}$, as ω is purely imaginary in this application. Additionally, due to the shift invariance of the Fourier transform, if we find a set of parent coordinates for $\omega = 0$, then shifting all the coordinates by $k \bmod n$ yields a set of coordinates tailored to $2\pi i k/n$; i.e., if $\mathcal{I}_0$ is tailored to $\omega = 0$,



m	$\min_{\omega\in[-\frac{\pi}{n},\frac{\pi}{n}]i} \eta_\omega(\mathcal{W})$	$\mathcal{I}$
1	·	0
2	36.8%	1
3	67.4%	1023
4	74.0%	2
5	80.9%	1022
6	83.7%	3
7	86.5%	1021
8	88.0%	4
9	90.0%	1020
10	90.5%	1019
⋮	⋮	⋮
90	99.0%	979

Figure 4.14 : Fixed compression space for undamped exponentials. The figure on the left the efficiency for compression spaces of increasing dimension between bins of the Fourier transform ($[-\pi/n, \pi/n]$) for subspaces of increasing dimension. The table on the right shows the minimum efficiency of these subspaces for a given dimension m, and the near optimal nested index set $\mathcal{I}$ is read down the right column.

then

$$\mathcal{I}_k = \{j + k \bmod n : \forall j \in \mathcal{I}_0\} \tag{4.60}$$

is tailored to $2\pi ik/n$. To estimate ω, we perform the three steps summarized in Algorithm 4.4. First, we find the largest value in the Fourier transform, $k = \operatorname{argmax}_j |[\mathbf{F}^*\widetilde{\mathbf{y}}]|_j$, yielding the initial estimate $\widetilde{\omega} = 2\pi ik/n$. Next, we use the interpolation algorithm of Quinn [128] to improve the estimate of $\widetilde{\omega}$, requiring $\mathcal{O}(1)$ operations. Finally, we optimize $\widetilde{\omega}$ on $\mathcal{W}_k = \operatorname{span} \mathbf{F}_{\cdot,\mathcal{I}_k}$ using Variable Projection with Kaufman's simplification. As we see in Figure 4.15, starting from the nearest Fourier frequency, Gauss-Newton requires three iterations to converge. However, starting from Quinn's interpolation estimate reduces this requirement to a single iteration.

Compression requires fewer operations than other high efficiency algorithms, as

Algorithm 4.4: Estimating a single undamped exponential with a fixed subspace

Input : Measurements $\mathbf{y} \in \mathbb{C}^n$, compression space $\mathcal{I}$ for $\omega = 0$

Output: Frequency ω, amplitude a

$\mathbf{w} \leftarrow \mathbf{F}^*\mathbf{y}$;

$k \leftarrow \operatorname{argmax}_j |[\mathbf{w}]_j|$;

$\mathcal{I} \leftarrow \{j + k \bmod n : j \in \mathcal{I}\}$;

$\mathbf{w} \leftarrow [\mathbf{w}]_{\mathcal{I}}$;

$\omega \leftarrow 2\pi ik/n$;

Use Quinn's interpolation to improve estimate of ω;

Gauss-Newton loop using VARPRO functional.

for $j = 1, 2, \ldots, N$ **do**

  $\mathbf{V} \leftarrow [\mathbf{F}^*\mathbf{V}(\omega)]_{\mathcal{I}}$;

  $\dot{\mathbf{V}} \leftarrow [\mathbf{F}^*\dot{\mathbf{V}}(\omega)]_{\mathcal{I}}$;

  $\alpha \leftarrow \mathbf{V}^*\mathbf{V}$;

  $a \leftarrow \alpha^{-1}(\mathbf{V}^*\mathbf{w})$;

  $\mathbf{r} \leftarrow \mathbf{w} - \mathbf{V}a$;

  Kaufman's simplification of the VARPRO Jacobian.

  $\mathbf{J} \leftarrow a[\dot{\mathbf{V}} - \alpha^{-1}\mathbf{V}(\mathbf{V}^*\dot{\mathbf{V}})]$;

  $x \leftarrow (\mathbf{J}^*\mathbf{r})/(\mathbf{J}^*\mathbf{J})$;

  $\omega \leftarrow \omega + \operatorname{Im}(x)$;

shown in Figure 4.15. Each algorithm begins with an $\mathcal{O}(n \log n)$ fast Fourier transform to compute the nearest Fourier frequency. Then fixed point methods, such as those proposed by Aboutanios and Mulgrew [2] and Minhas and Gaydecki [107] require $2n$ operations per iteration, with Aboutanios and Mulgrew's algorithm requiring two iterations and Minhas and Gaydecki's algorithm only one. However starting from Quinn's interpolation estimate, an $\mathcal{O}(1)$ operation, our compressed approach requires only $\mathcal{O}(8m)$ operations. Hence compression is the fastest algorithm, requiring only ≈ 720 flops beyond the Fourier transform to estimate ω to higher than 99% efficiency when $n = 1024$, whereas Minhas and Gaydecki's algorithm requires ≈ 2048 flops.

Moreover, compression gives a simple mechanism to generate estimators with a prescribed efficiency. Further improvements may be possible with other parent coordinates (e.g., $\mathbf{G}_4$) and explicit consideration of the covariance of a purely imaginary ω versus a complex ω in constructing the compression space. Additionally, some applications also seek decay rates (i.e., $\omega \in \mathbb{C}^-$), e.g., [1, 16]. Compression can also be applied here in a similar fashion to build fast estimators.

4.6.2 Dynamically Computed Compression Subspaces

If building a fixed compression space is impractical, an alternative is to build $\mathcal{W}$ while optimizing $\boldsymbol{\theta}$, at a penalty of an additional $\mathcal{O}(p^2)$ operations per candidate to update the compression space. In practice, this additional cost is small, as the following example shows.

One situation requiring a dynamic compression space is the blind exponential fitting problem: given only measurements $\widetilde{\mathbf{y}}$ and the number of exponentials p, this problem seeks the exponential parameters $\boldsymbol{\omega}$ and $\mathbf{a}$. Figure 4.16 shows a numerical experiment that compares four variants of the Peeling Method (Section 3.6) to Kung's

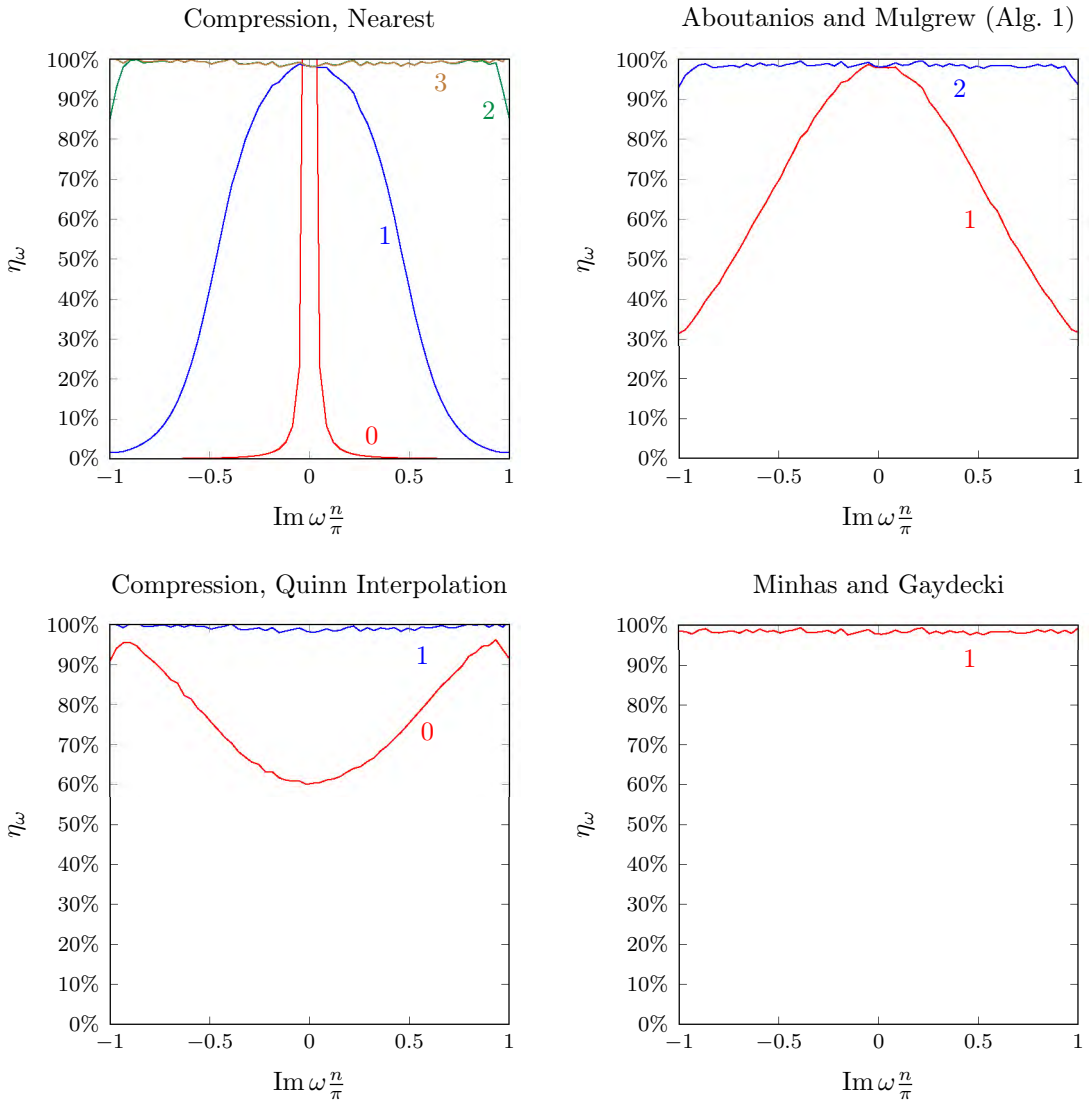


Figure 4.15 : The statistical efficiency for several high efficiency single undamped exponential fitting algorithms. Each plot shows estimates of the efficiency of each estimator for different values of ω from 10^5 trials, where $n = 1024$ and $\epsilon = 1$ (corresponding to a signal-to-noise ratio of 0 dB). The left side shows a new algorithm based on compression, and the right side shows two existing algorithms. The top left shows iterations of VARPRO with Kaufman's simplification, starting from the nearest Fourier frequency, and the number indicates how many iterations were taken. The bottom left shows the same, but starting from Quinn's interpolation estimate of ω. In these examples an $m = 90$ subspace was used, with a minimum efficiency of 99%. The right plots show two existing algorithms, where, again, numbers show how many iterations were taken. Compression starting from Quinn's interpolation estimate requires $\approx 65\%$ fewer operations beyond the FFT than the current leader, Minhas and Gaydecki.

Method, representative of the numerically stable variants of Prony's Method, and Variable Projection using the true values $\widehat{\boldsymbol{\omega}}$ as initial estimates. As Kung's Method solves a dense SVD of size $n/2$, the computational cost grows cubically in the number of measurements n. In contrast, the cost Variable Projection with the true initial estimates grows linearly in n for the full problem, whereas the cost is approximately constant for the compressed problem. The four variants of the Peeling Method use either the Orthogonalized Matrix Pencil Method (OMPM) (Section 2.8) or the largest entry in the Fourier transform of the residual (Section 3.4.3) and either solve the compressed problem or the full problem at each iteration. One hopes the Orthogonalized Matrix Pencil Method will provide better initial estimates of ω, requiring fewer optimization iterations to converge. However, the additional computation required to use the OMPM and the occasional poor initial estimates of ω make this approach unviable. In contrast, the largest entry in the Fourier transform of the residual is faster and provides better estimates. This experiment also shows that compression is only computationally advantageous for large problems, say $n > 800$. Although in this example, the compressed problems are of dimension $m \approx 10p$, the smaller compressed matrices $\mathbf{W}^*\mathbf{V}(\boldsymbol{\omega})$ and $\mathbf{W}^*\dot{\mathbf{V}}(\boldsymbol{\omega})$ cost more to compute. The large increase in wall clock time between $n = 10^4$ to $n = 4 \cdot 10^4$ and beyond $n = 10^5$ for the Peeling Method using the largest entries in the Fourier transform of the residual and compression correspond to where the initial values led to a spurious minima. The additional computation time is a result of solving the full problem with the estimates from Peeling. In most cases, the cost of this additional step is small, since the estimates of $\boldsymbol{\omega}$ are approximately local minima. However, when Peeling converges to a spurious minima, many steps of the full problem are required until the convergence criteria are met. This combination of the Peeling Method with compression provides a promising,

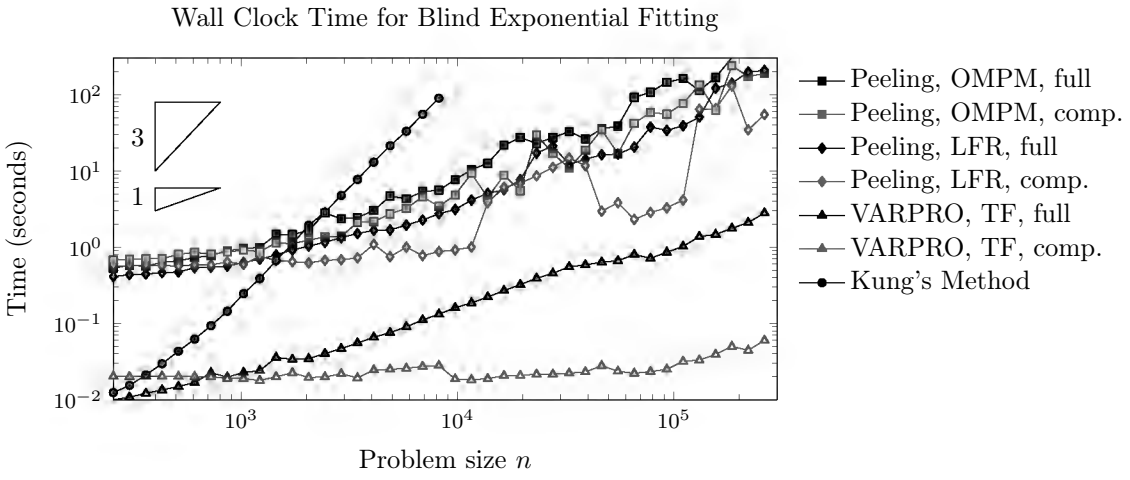


Figure 4.16 : Wall clock time for several blind exponential fitting algorithms. We compare four variants of the Peeling Method using initial conditions provided by the Orthogonalized Matrix Pencil Method (OMPM) or by the largest entry of the Fourier transform of the residual (LFR) and that solve either the full or compressed problem. For comparison, we also include the wall clock times of Kung's Method and Variable Projection (VARPRO) using the true values of $\boldsymbol{\omega}$ as initial estimates (TF). The measurements $\widetilde{\mathbf{y}}$ were generated from the $p = 11$ model from [153, Table 1] adding complex noise of standard deviation 1, Reported times were averaged over ten trials of each method implemented in MATLAB 7.14 running in Linux on a Intel i3-2120 CPU with a 3 MiB L1 cache and 3781 MB RAM.

fast solution to the blind exponential fitting problem with superior performance to existing methods such as Kung's Method ($\mathcal{O}(n \log n)$ versus $\mathcal{O}(n^3)$).

We can guarantee that the compressed estimates converge provided the compression space always grows in dimension. At the kth step, compression provides an approximate gradient $\widetilde{\mathbf{g}}_k$ of the true gradient $\widehat{\mathbf{g}}_k$ with error $\mathbf{e}_k$, where[3]

$$\widehat{\mathbf{g}}_k = \widetilde{\mathbf{g}}_k + \mathbf{e}_k = \mathbf{e}_k + \sum_{j \in \mathcal{I}} \nabla_{\mathbf{x}} [\mathbf{U}^* \mathbf{f}(\boldsymbol{\theta})]_j \big|_{\boldsymbol{\theta} = \boldsymbol{\theta}_k} . \tag{4.61}$$

If we used a gradient descent method with step size $\|\mathbf{H}(\boldsymbol{\theta}_k)\|_2^{-1}$, successive iterates of $\boldsymbol{\theta}$ have weak linear convergence by [50, Thm. 2.2] provided $\|\mathbf{e}_k\|_2^2 \leq B_k$ and $|\mathcal{I}|$ grows such that $\lim_{k \to \infty} B_{k+1}/B_k \leq 1$ (e.g., $|\mathcal{I}|$ grows linearly in k). Our box-constrained

trust region method will generally have better performance, since these gradient descent steps are included in the search space. Further, if full efficiency estimates are required, we can solve increasingly large compressed problems until they converge, and then use only a few final iterations of the full problem to refine $\boldsymbol{\omega}$.

[3]We write the compressed gradient in this form to emphasize the connection to incremental gradient methods, cf. [50, §3, first equation]. We omit the normalization by $|\mathcal{I}|$ since the selected rows are not typical.

Chapter 5

Damping in Vibrating Strings

Our investigation of the exponential fitting algorithm described in the previous chapters was motivated by the desire to approximate spectral properties of a damped, vibrating string. A simple description of the transverse motion of a thin elastic string subjected to viscous damping is given by

$$\begin{aligned} u_{tt}(s,t) &= c^2 u_{ss}(s,t) - 2\delta(s) u_t(s,t) \\ u(0,t) = u(\pi,t) = 0, \quad u(s,0) &= u_0(s), \quad u_t(s,t) = u_1(s), \end{aligned} \tag{5.1}$$

where $u(s,t)$ is the transverse displacement, s is the position along the rest configuration string of length π, and t is time; see, e.g. [147, § 131]. After separation of variables, for constant δ

$$u(s,t) = \sum_{n \neq 0} a_n e^{\omega_n t} \sin(ns), \tag{5.2}$$

the eigenvalues ω_n of the differential operator associated with (5.1) are

$$\omega_{\pm n} = -\delta \pm \sqrt{\delta^2 - c^2 n^2}, \quad n = 1, 2, \ldots. \tag{5.3}$$

A challenging class of modern problems seeks information about the underlying differential equation from knowledge of the spectrum. One such inverse eigenvalue problem recovers $\delta(s)$ from knowledge of $\{\omega_n\}_{n \neq 0}$. Cox and Embree have recently provided an algorithm to perform this reconstruction for an even damping function $\delta(s)$ [37]. We would like to ask: can we perform this reconstruction using experimental data? This is not without precedent: recently Cox, Embree, and Hokanson were

able to determine the masses and locations of beads on a string from eigenvalue estimates [38], realizing in the laboratory the classical inverse spectral theory for discrete self-adjoint systems [52]. However, reconstructing the damping function is challenging. Rather than simply requiring $\operatorname{Im}\omega$, as the (undamped) beaded string does, for damped problems $\operatorname{Re}\omega$ is needed as well. Further, the reconstruction algorithm of Cox and Embree hinges on specific asymptotic eigenvalue estimates that can

In this chapter, we investigate how well the vibrations of a metal piano wire match the theoretical model of a viscously damped string. First we describe the experimental apparatus, examine the background effects, and calibrate the detectors. Next, we build a robust variant of compressed exponential fitting to estimate $\operatorname{Re}\omega$. Finally, we compare these eigenvalue estimates to those predicted by string models of increasing accuracy, with the goal of determining the appropriate model of a vibrating string. With better knowledge of the physical model, future researchers should be able to recover $\delta(s)$ from experimental measurements of $\boldsymbol{\omega}$ by modifying the Cox-Embree algorithm to apply to a more physically realistic non self-adjoint system.

5.1 Experimental Apparatus

In our experiments we seek to recover the eigenvalues of the linear operator that describes the motion of a string, parameterized by the pressure of the surrounding air. We are particularly interested in how the pressure affects the real parts of these eigenvalues, which describe the asymptotic rate at which energy decays in the string; see, e.g., [36]. This objective differs from previous experimental work which observed chaos and other nonlinear effects in vibrating strings; see, e.g., [62, 110]. As a result, our apparatus and procedure differs from these previous experiments, which drove the string continuously and measured the equilibrium amplitude of vibration. Although

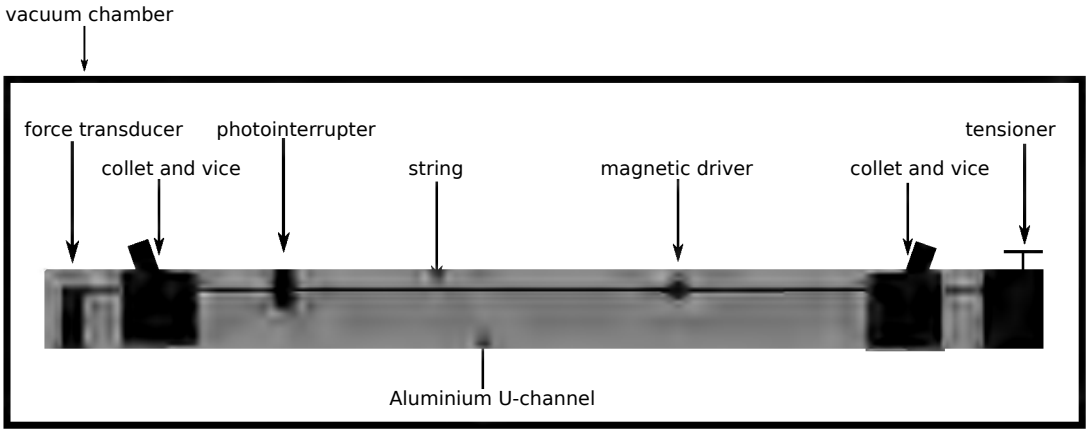


Figure 5.1 : Cutaway diagram showing the vacuum chamber and string apparatus.

the amplitude at equilibrium does correspond to the real part of the eigenvalues, nonlinear effects easily influence this value. Instead, we measure free response motion of the string, with initial conditions provided by driving the string at a frequency near resonance. The resulting time series is, to first approximation, a sum of exponentials plus noise, where the exponential coefficients correspond to the desired eigenvalues.

Our apparatus derived from the set-up described in [38], originally built with assistance from Sean Hardesty based on a design by Mr. Fan Tao of J. D'Addarío & Co. The electronics were built by Jeffrey Bridge, and Stan Dodds provided assistance in the assembly of the vacuum chamber. As illustrated in Figure 5.1, our apparatus consists of a steel string that is connected at one end to a force transducer and, at the other end, wound along a spindle to apply tension. In between the string passes through two collets, mounted in collet vises, clamping the string to enforce Dirichlet boundary conditions. Between the collets, the string passes through a photodetector

Table 5.1 : Material properties of the string

Property	Value
Standard	ASTM A228
Description	Plated Steel Music Wire
Manufacturer	Mapes Piano String Company
Diameter	0.015 in
Length	$43\,^{3}/_{16}$ in
Density	$7850\,\mathrm{kg/m^3}$
Yield Strength	2500 MPa to 2880 MPa
Young's Modulus	210 GPa
Poisson's Ratio	0.313
Shear Modulus	80.0 GPa
Composition	C (0.7–1.0%); Fe (97.8–99%); Mn (0.20–0.60%); P ($\leq$ 0.025%); Si (0.10–0.30%); S ($\leq$ 0.030%)

measuring the displacement of the string (described in §5.1.4) and an electromagnetic drive coil (described in §5.1.3). Rather than being placed on an optical table as in [38], the string and its supports are mounted on an aluminum U-channel to fit inside the 16 in diameter PVC pressure vessel. The remainder of this section provides further detail on each component of this experiment.

5.1.1 String

For our string we use a 0.015 in diameter steel music wire (ASTM A228), whose properties are listed in Table 5.1. This alloy and tempering of steel is often used in metal music instrument strings, either by itself or wound with a soft metal wire to increase mass without increasing bending stiffness.

This steel wire has many advantages from an experimental standpoint. First, we can place the string under a high tension, typically from 100 N to 200 N. Increased tension decreases the amplitude of vibrations, given a constant energy, which de-

creases errors due to the linearization in Section 5.3.1. Second, although tension is high, we are still well within the elastic limit: at 200 N, the string is under a stress of 1.75 GPa, well below the yield strength of 2.5 GPa. The string would need to displace 4 cm from equilibrium at the midpoint in the fundamental mode to exceed this yield strength – a far larger oscillation than experimentally observed. Consequently, we can assume that the stress-strain relationship is linear to high accuracy. Third, non-metal string materials like Nylon tend to creep, increasing in length while under tension over a period of hours [105, §1.3]. Steel does not exhibit this effect unless heated to a large fraction of its melting point. Finally, steel strings can be driven electromagnetically, as we describe in Section 5.1.3.

5.1.2 String Mounting

The string is mounted on a 5 ft long, 9 in wide, and 4 in tall aluminum U-channel with 1/2 in side walls and a 3/8 in thick base. The U-channel then slides inside a 16 in diameter schedule 40 PVC pipe, which is used as a pressure vessel; the U-channel is supported by the walls of the pipe. The ends of the pipe are sealed with a 3/8 in thick aluminum plates with a 16 in O-ring making a pressure seal against the PVC pipe; the plates are held in place with non-pressure rated PVC end caps. The minimum pressure this chamber reached was 0.006 bar.

Two holes were drilled in the side of the pressure vessel. One was connected via a hose to a vacuum pump. The other contained an pass through for the electrical wires carrying power and returning data from the string and was sealed using epoxy.

5.1.3 Excitation and Driver

With the string sealed inside the vacuum chamber, we must excite motion within the closed chamber. There are three typical approaches: add a mechanical plucker or shaker (e.g., [63]), place a portion of the string in a magnetic field and run a current through the string (e.g., [62]), or use an electromagnetic coil to periodically attract the string (e.g., [42]). The mechanical approach requires either transmitting the stimulation through the boundary conditions or connecting the interior of the string to the shaker; both introduce perturbations to the wave equation for which it is difficult to account for. Passing a current through the string leads to resistive heating that causes a time-dependent increase in the effective length of the string and, correspondingly, decreases the frequencies of the eigenvalues [110, §3]. In previous studies, this was not a significant effect, as the string is conductively cooled by the surrounding air [62, p. 1551]. However at low pressures the string is only cooled by radiation and convection through the endpoints – these may not be sufficient to keep the string cool. Hence, we drive the string with an electromagnet.

The electromagnetic drive coil consists of a rectangular soft iron core and 180 turns of magnet wire, pictured in Figure 5.2. Although the iron core limits high frequency transmission, it increases the effect of the electromagnet on the string. The closed loop contains the magnetic flux, and one end is beveled to a create gradient of magnetic flux across the gap. When the coil is active, it induces a dipole moment in the wire which in turn feels a force proportional to gradient of the magnetic field. However, as the steel string moves in the direction of the gradient, regardless of the polarity of the field, we drive the electromagnet with a half sine wave: $\min(a\sin(\omega t), 0)$.

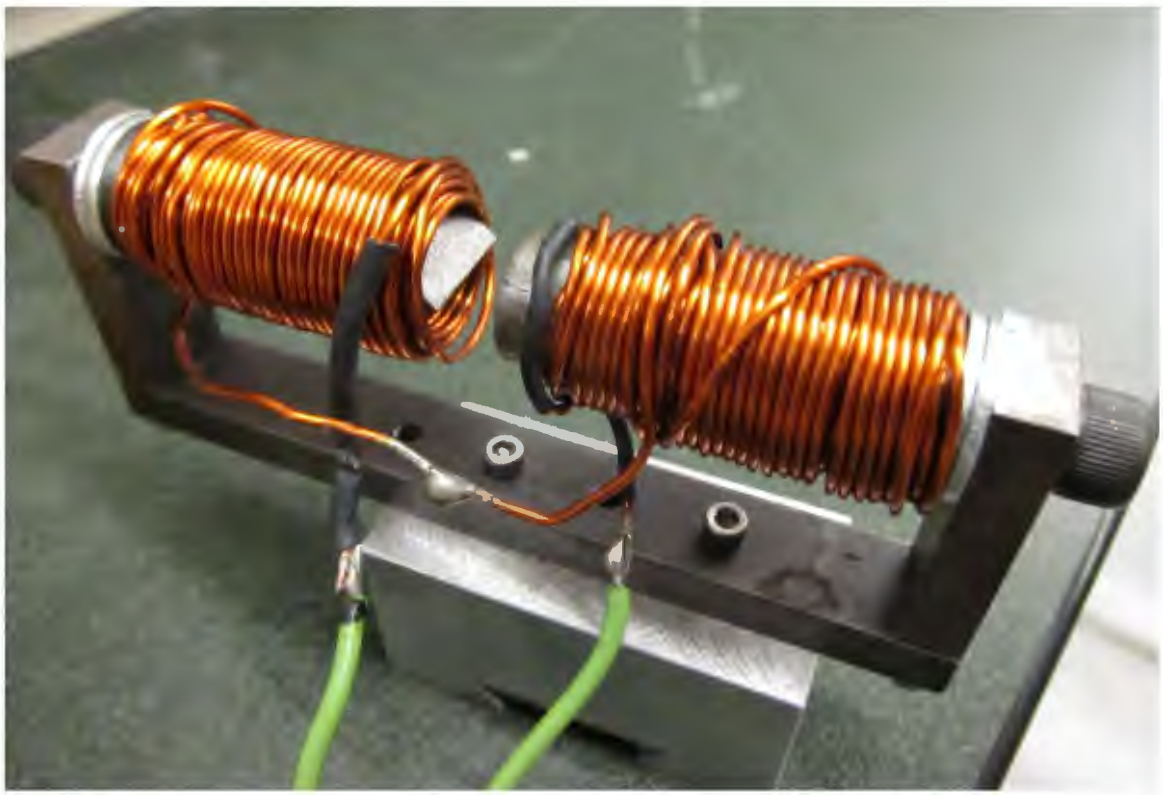


Figure 5.2 : The electromagnetic drive coil.

5.1.4 Photodetector

We record the position of the string by measuring how much the string occults a beam of light between an LED and a phototransistor mounted on either side of the string. This resembles the detector described in [62, Fig. 2], but we use a polynomial model of the detector to better infer the actual position of the string, rather than assuming the response is linear. In our device, both the LED and phototransitors are part of a single Fairchild QVE00034 package. Two of these LED/phototransistor pairs are mounted in a plastic block at right angles, allowing us to measure the motion of the string in the two transverse planes of motion, pictured in Figure 5.3. The attached printed circuit board (designed by Jeffrey Bridge) converts the current response of the diodes to light to a voltage, which is then read by a National Instruments USB-6251 data acquisition board (DAQ). Unless otherwise specified, all experiments record voltages at 10^5 samples per second in both axes. There is a 5×10^{-6} s lag between the time when the x and y channels are read that we consider negligible. The whole assembly is mounted on a 2-axis micrometer stand, allowing us to position the detector in an optimal location to measure the motion of the string – near where both channels read 0 V.

Background

To determine the background noise in the photodetector, we conduct two experiments: one with the string starting from rest and another starting from rest with a cloth placed over a part of the string to damp vibrations. The additional damping removes vibrations induced by the environment. Figure 5.4 shows the discrete Fourier transform for both measurements. Without the cloth present, some peaks grow larger by a factor of 1.5. This indicates that although there is some transmis-

Figure 5.3 : The photodetector apparatus. A pair of phototransistors are embedded in the white block on the right and moved by the black micrometer stand on the far right. The printed circuit board amplifies and converts the current output of the phototransistor to a voltage.

sion of energy from the environment, the size of these peaks (0.07 V) indicates the amount of mechanical energy transmitted is minimal.

However, several large peaks remain. These peaks in the Fourier transform are large in comparison to the background noise: however this noise is orders of magnitude smaller than the signal from physical vibrations of the string. Multiples of 1000 Hz (and, to a lesser extent, 500 Hz) appear in both axes. Due to its purity and frequency, this is likely background electrical noise; this signal remains despite changing the sample rate of the data acquisition card. Another set of peaks center near 20 Hz. We attributed these to the resonances of table on which the vacuum chamber is mounted. Finally, there is broad hump around 14 kHz in the X-axis and 12 kHz in the Y-axis. We attribute this to the 12.6 kHz fundamental of the pier on which the photodetector is mounted (a 4 in × 1 in × 1 in block of aluminum)[1]. The splitting of this resonance is due to the additional moment of inertia from the photodetector assembly in one axis, but not the other.

Least squares methods assume the residual is normally distributed. However, the data acquisition card samples a finite number of voltages due to its inherent 16-bit resolution. To be justified in using least squares methods, the background should be approximately normally distributed. Examining a histogram of the voltage readings in Figure 5.5, we see that a normal distribution describes the tails well. However, the peak of both distributions does not fit the measurements as well; this may be due to the background noise that has not been removed.

[1]It is well known that the fundamental of a cantilevered beam is approximately $\frac{3.515}{L^2}\sqrt{EI/(\rho A)}$ where E is the elastic modulus, I the moment of inertia, L the length, and ρA the linear density.

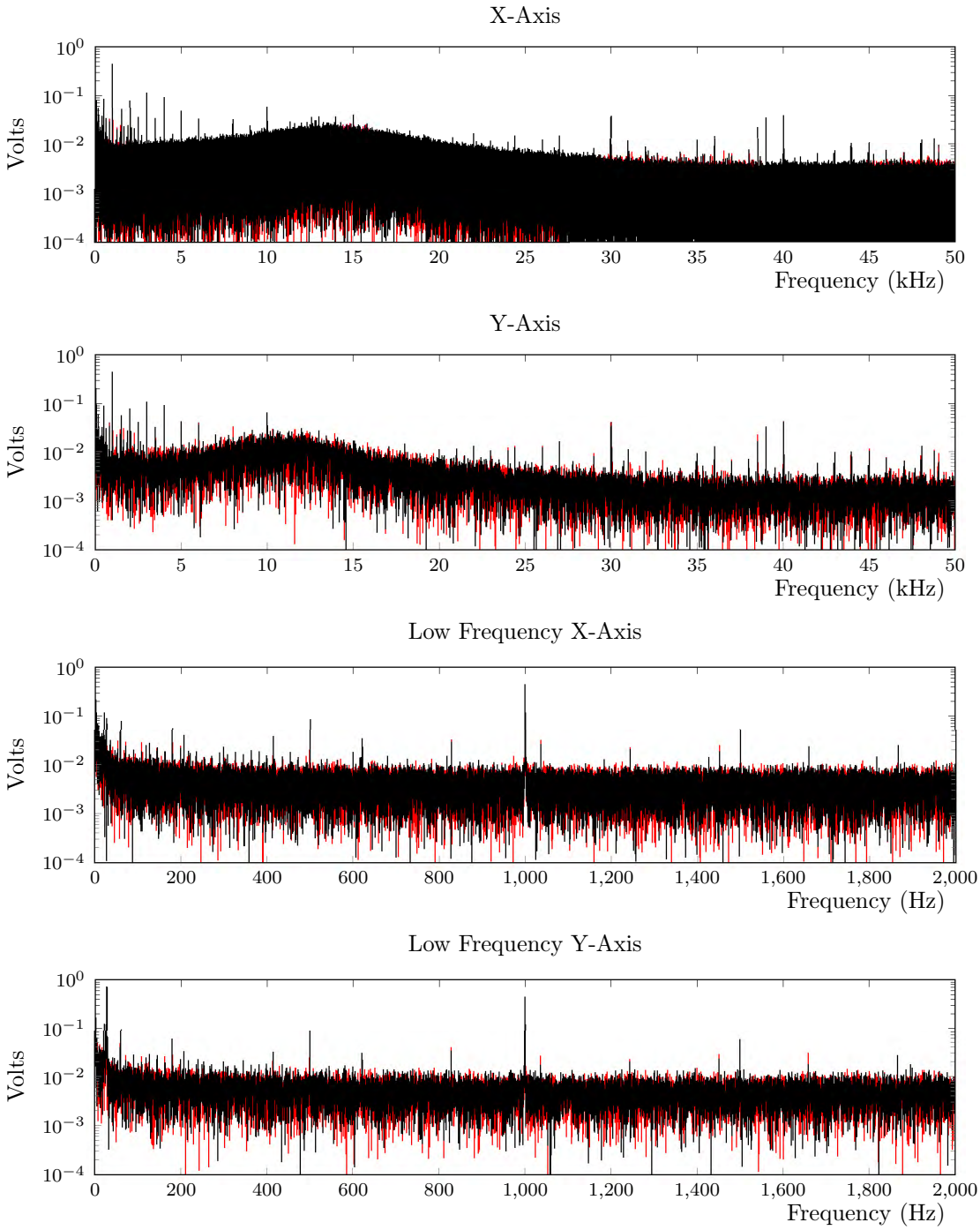


Figure 5.4 : Fourier transform of string at rest. The top two plots show $\mathbf{F}^*\widetilde{\mathbf{y}}$ for a large range of frequencies; the bottom two show the low frequency response. The black curves show the spectrum when a cloth was placed on the string; the red curve shows the string when this extra damping was absent.

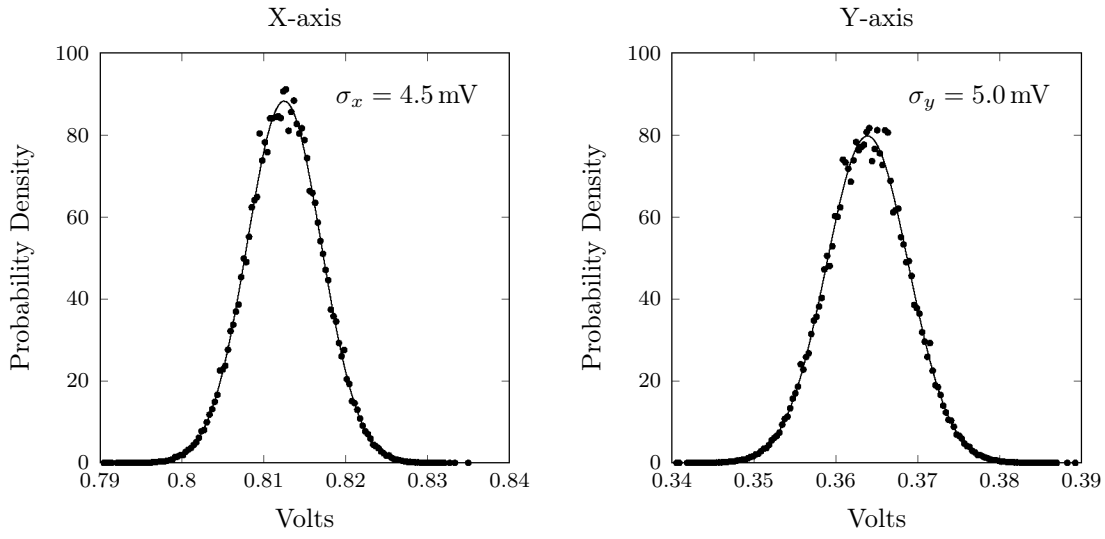


Figure 5.5 : The experimental probability density function of background signal. Each dot shows the probability of the corresponding discrete 16-bit voltage reading, and the curve shows the normal distribution fit, with the standard deviation shown in the upper right hand corner. The different mean voltage in each axis is due to the location of the string when these measurements were taken.

Calibration

To convert voltage readings from the data acquisition card into position measurements, we construct a polynomial model of the detector. Moving the photodetector relative to the string with the micrometer stand, we record the voltages in both axes as a function of position, resulting in Figure 5.6. After examining multiple models, we conclude that position is best described by a fourth-order polynomial in one axis plus a linear contribution from the other:

$$\begin{aligned} v_x(x,y) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 y + \beta_6 xy \\ v_y(x,y) &= \gamma_0 + \gamma_1 y + \gamma_2 y^2 + \gamma_3 y^3 + \gamma_4 y^4 + \gamma_5 x + \gamma_6 xy. \end{aligned} \tag{5.4}$$

The least squares fits for these models are shown in Figure 5.6, and result in a residual norm of less than 3% of the original signal. (Further increases of polynomial order

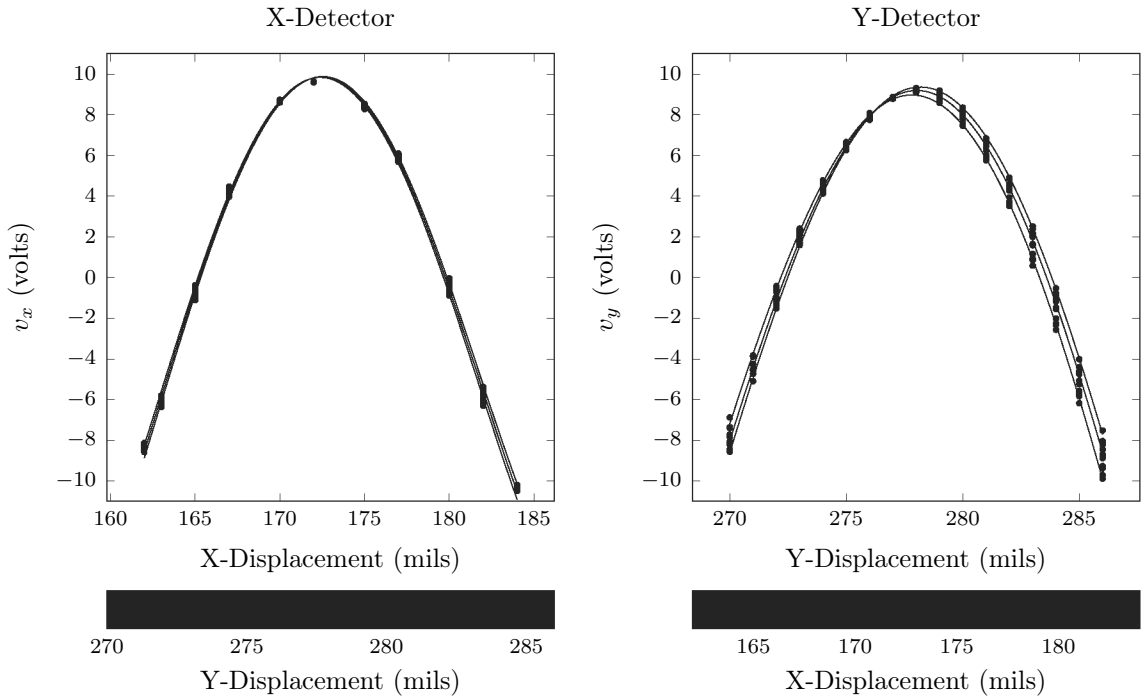


Figure 5.6 : Measurements of the displacement-voltage relationship. Dots show the measured voltage as a function of displacement in one axis, with colors denoting the displacement in the orthogonal axis. The colored curves show the corresponding polynomial fits (functions of both x and y).

yield negligible improvements.) To invert the position-voltage map given by (5.4), we use Newton's Method, assuming the string remains in a proscribed quadrant with an initial estimate corresponding to 0 V in both axes.

Combining these polynomial fits with the noise model given in Figure 5.5 yields a standard deviation for the absolute displacement of the string that depends on the location of the string in the detector. For the range the string occupied during the experiments described in this chapter (-2 V to 2 V in both axes), the maximum standard deviation assocated with displacement is 0.98 µm (0.039 mils).

5.2 Data Analysis

After removing the effects of the detector, we have a single complex vector $\mathbf{y} \in \mathbb{C}^n$, where $\operatorname{Re}\mathbf{y}$ and $\operatorname{Im}\mathbf{y}$ are the displacements in the X and Y axes. Although an ideal string has eigenvalues whose frequencies are approximately linearly separated, a real string can have pairs of eigenvalues that are approximately linearly separated, one for each axis of motion; see Section 5.3.6.

To recover a single eigenvalue instead of a pair (or more), and to remove the effects of any nonlinearities or poor initial estimates, we apply the following algorithm to estimate $\boldsymbol{\omega}$. First, we compute $\boldsymbol{\omega}$ using compressed exponential fitting, starting from a list of initial estimates and enforcing that eigenvalues must come in conjugate pairs (since motion in each axis is real). Next, we partition $\mathbf{V}(\boldsymbol{\omega})$ and $\mathbf{y}$ into length q segments using rows qj to $q(j+1)-1$, yielding $\mathbf{V}_j$ and $\mathbf{y}_j$, and compute

$$\mathbf{a}_j = \mathbf{V}(\boldsymbol{\omega})^{+}\mathbf{y}_j. \tag{5.5}$$

Finally, we determine the decay rate γ_k of the kth eigenvalue by fitting a single exponential to $e^{\operatorname{Re}(\omega_k)qj}\|[\mathbf{a}_j]_{k,k'}\|_2$, where k and k' are the indices of conjugate pairs

($\omega_k = \overline{\omega_{k'}}$). Then ω_k is updated to $\gamma_k + \operatorname{Im}\omega_k$ and the process is repeated until $\boldsymbol{\omega}$ converges. This ensures that $\operatorname{Re}\omega_k$ is robust to slight changes in $\operatorname{Im}\omega_k$.

5.3 Matching Reality to Physical Models

In this section we derive models of increasing fidelity to our experimental measurements. Our approach combines on Lagrangian mechanics, following Rayleigh [147], with the rigor of Antman's derivation [7]. This Lagrangian approach uses Hamilton's Principle of Stationary Action [35, §IV.10] to derive the equations of motion from the kinetic and potential energy in the system. As a result, we can easily incorporate additional effects, like bending stiffness, without deriving a new constitutive relation.

5.3.1 First Order Model

Suppose we have a string of rest length ℓ_0 stretched between two supports positioned length ℓ apart. Without loss of generality, we assume that the line between the two supports is $\{s\mathbf{e}_1\}_{s=0}^{\ell_0}$, and that $\mathbf{e}_2$ and $\mathbf{e}_3$ are the remaining orthogonal coordinates for a right-handed Cartesian coordinate system. Then the displacement of the string is written $\mathbf{u}(s,t) \in \mathbb{R}^3$ where $s \in [0, \ell_0]$ is the position along the string in the rest configuration at time $t \in [0, \infty)$. Further, we assume that each end of the string is subject to Dirichlet boundary conditions: $\mathbf{u}(0,t) = 0$ and $\mathbf{u}(\ell_0, t) = 0$.

In the Lagrangian Mechanics, the equations of motion are derived from the kinetic and potential (strain) energy in the system. Kinetic energy in the string is proportional to the square of velocity:

$$T(t) = \frac{1}{2}\int_0^{\ell_0} (\rho A)(s)\|\mathbf{u}_t(s,t)\|_2^2 \,\mathrm{d}s, \tag{5.6}$$

where (ρA) is the product of density and cross-sectional area (cf. [7, Ch. 2, eq. (10.9)]).

Assuming the string is *elastic*, the force per unit length of the string $\mathbf{n}(s,t)$ is proportional to the stretch $\nu(s,t) := \|\mathbf{u}_s(s,t)\|_2$ and tangent to the string

$$\mathbf{n}(s,t) = \widehat{N}(\nu,s)\frac{\mathbf{u}_s(s,t)}{\|\mathbf{u}_s(s,t)\|_2} \tag{5.7}$$

(cf. [7, Ch. 2, eq. (2.10b), (2.11)]). The potential energy in the string comes from integrating this force $\mathbf{n}(s,t)$ per unit length from the reference configuration $\nu = 1$ to the current state

$$W(\nu,s) := \int_1^\nu \widehat{N}(\beta,s)\,\mathrm{d}\beta \tag{5.8}$$

(cf. [7, Ch. 2, eq. (10.10)]). Then integrating this energy density over the length of the string gives the total potential energy

$$V(t) := \int_0^{\ell_0} W(\nu(s,t),s)\,\mathrm{d}s; \tag{5.9}$$

(cf. [7, Ch. 2, eq. (10.11)]).

For elastic materials, like the steel string used in these experiments, the stress-strain relationship is linear provided the tension is below the yield strength. Our steel string appears to stay in the linear regime provided stretch is less than 1.01; hence,

$$\widehat{N}(\nu,s) = (EA)(s)(\nu-1) \quad \text{for } \nu \in (0.99, 1.01), \tag{5.10}$$

where $(EA)(s)$ is the product of the Young's modulus and cross-sectional area. After integrating (5.8), the potential energy is[2]

$$V(t) = \frac{1}{2}\int_0^{\ell_0} (EA)(s)(\|\mathbf{u}_s(s,t)\|_2^2 - 2\|\mathbf{u}_s(s,t)\|_2 + 1)\,\mathrm{d}s. \tag{5.11}$$

[2]Authors frequently omit the $\|\mathbf{u}_s(s,t)\|$ term, as, in the limit of small vibration, this term only adds a scaling in the resulting differential equation. Rayleigh omits this term in *The Theory of Sound* [147, §128, eq. (3)]. Antman does the same by taking $\widehat{N}(\nu,s) = (EA)(s)\nu$ [7, Ch. 2, eq. (7.1)], violating the assumption that the rest state is natural, $\widehat{N}(1,s) = 0$.

The motion $\mathbf{u}(s,t)$ is a stationary point of the action functional

$$\int_{t_0}^{t_1} L(\mathbf{u}, \mathbf{u}_t, t)\,\mathrm{d}t, \qquad \text{where} \qquad L = T - V \tag{5.12}$$

with respect to $\mathbf{u}$ [35, § IV.10]. The terms in the kinetic and potential energy with respect to $\|\mathbf{u}_t\|_2^2$ and $\|\mathbf{u}_s\|_2^2$ follow directly, and the constant term is negligible. However, the $\|\mathbf{u}_s\|_2$ term must be expanded in terms of the series

$$\mathbf{u}(s,t) = \mathbf{u}^{(0)}(s,t) + \epsilon\mathbf{u}^{(1)}(s,t) + \epsilon^2\mathbf{u}^{(2)}(s,t) + \cdots. \tag{5.13}$$

Here, the zeroth order term corresponds to the string at rest. Under the assumption that the string is uniform, $\mathbf{u}^{(0)}(s,t) = s\ell/\ell_0\mathbf{e}_1$. Defining $\widehat{\nu} := \ell/\ell_0$ and $\widetilde{\mathbf{u}} = \mathbf{u} - \mathbf{u}^{(0)}$, we expand the $\|\mathbf{u}_s\|_2$ term as

$$\begin{aligned}
\|\mathbf{u}_s\|_2 &= \widehat{\nu}\sqrt{1 + 2\widehat{\nu}^{-1}\mathbf{e}_1^\top\widetilde{\mathbf{u}}_s + \widehat{\nu}^{-2}\widetilde{\mathbf{u}}_s^\top\widetilde{\mathbf{u}}_s} \\
&= \widehat{\nu}\left(1 + \frac{1}{2}(\widehat{\nu}^{-1}2\mathbf{e}_1^\top\widetilde{\mathbf{u}}_s + \widehat{\nu}^{-2}\widetilde{\mathbf{u}}_s^\top\widetilde{\mathbf{u}}_s) - \frac{1}{8}(2\widehat{\nu}^{-1}\mathbf{e}_1^\top\widetilde{\mathbf{u}}_s + \widehat{\nu}^{-2}\widetilde{\mathbf{u}}_s^\top\widetilde{\mathbf{u}}_s)^2\right) + \mathcal{O}(\|\widetilde{\mathbf{u}}_s\|_2^3) \\
&= \widehat{\nu} + \mathbf{e}_1^\top\widetilde{\mathbf{u}}_s + \frac{1}{2\widehat{\nu}}\left([\widetilde{\mathbf{u}}_s]_2^2 + [\widetilde{\mathbf{u}}_s]_3^2\right) + \mathcal{O}(\|\widetilde{\mathbf{u}}_s\|_2^3).
\end{aligned}$$

Upon integration, the $[\widetilde{\mathbf{u}}_s]_1$ term vanishes as a result of the Dirichlet boundary conditions, leaving

$$\int_0^{\ell_0} \|\mathbf{u}_s(s,t)\|_2\,\mathrm{d}s = \ell + \frac{\ell_0}{2\ell}\int_0^{\ell_0} [\widetilde{\mathbf{u}}_s]_2^2 + [\widetilde{\mathbf{u}}_s]_2^3\,\mathrm{d}s + \mathcal{O}(\|\widetilde{\mathbf{u}}_s\|_2^3).$$

This leaves the Lagrangian, omitting constants and terms of order ϵ^2 and above,

$$L(t) = \frac{1}{2}\int_0^{\ell_0} (\rho A)(s)\|\mathbf{u}_t^{(1)}(s,t)\|_2^2 - (EA)(s)\left[[\mathbf{u}_s]_1^2 + \frac{\ell-\ell_0}{\ell}[\mathbf{u}_s^{(1)}]_2^2 + \frac{\ell-\ell_0}{\ell}[\mathbf{u}_s^{(1)}]_3^2\right]\mathrm{d}s. \tag{5.14}$$

If ρA and EA are independent of s, then the solution is given by [35, § IV.10]:

$$\begin{aligned}
\rho A[\mathbf{u}_{tt}^{(1)}]_1 &= EA[\mathbf{u}_{ss}^{(1)}]_1 \\
\rho A[\mathbf{u}_{tt}^{(1)}]_2 &= \frac{EA(\ell-\ell_0)}{\ell}[\mathbf{u}_{ss}^{(1)}]_2 \\
\rho A[\mathbf{u}_{tt}^{(1)}]_3 &= \frac{EA(\ell-\ell_0)}{\ell}[\mathbf{u}_{ss}^{(1)}]_3.
\end{aligned} \tag{5.15}$$

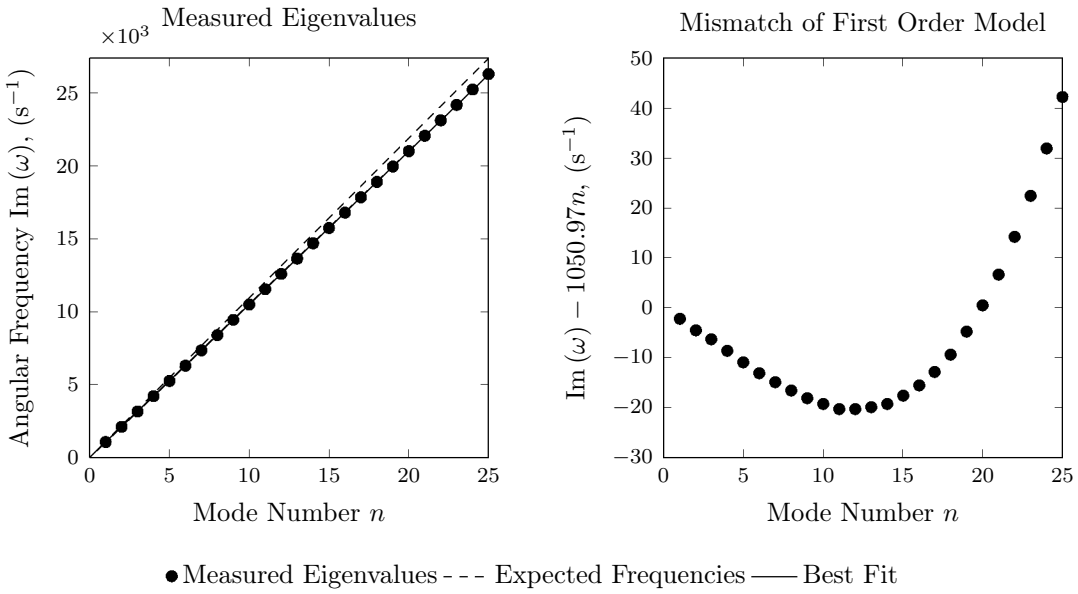


Figure 5.7 : The imaginary part of the measured eigenvalues at between 7×10^{-3} and 9×10^{-3}atm, compared to the first order model. The right plot shows the mismatch between the best fit, corresponding to a fundamental of $1050.97\,\mathrm{s}^{-1}$.

This is the familiar wave equation which, via separation of variables, has solutions

$$\mathbf{u}^{(1)}(s,t) = \sum_{k=1}^{3} \sum_{n \neq 0} a_{n,k} e^{\omega_{n,k} t} \sin(n\pi s/\ell_0)\mathbf{e}_k, \tag{5.16}$$

where

$$\omega_{n,1} = i\frac{n\pi}{\ell_0}\sqrt{\frac{E}{\rho}} \tag{5.17}$$

$$\omega_{n,2} = \omega_{n,3} = i\frac{n\pi}{\ell_0}\sqrt{\frac{E(\ell-\ell_0)}{(\rho\ell)}} = in\pi\sqrt{\frac{\tau}{\rho A \ell \ell_0}}. \tag{5.18}$$

Here τ is the measured tension. Using the material properties of the string, the fundamental ($n = 1$) for longitudinal waves $\omega_{n,1}$ is $14.8\,\mathrm{ks}^{-1}$. For the two transverse waves, the tension in the string varied between 133 N and 131 N corresponding to a fundamental of between $1096\,\mathrm{s}^{-1}$ to $1104\,\mathrm{s}^{-1}$.

As Figure 5.7 shows, the imaginary components of the transverse eigenvalues (frequencies) correspond to the predicted fundamental. We attribute the mismatch at

higher frequencies to a slow creep of the wire, causing a gradual decrease in tension that cannot be measured. After fitting a new fundamental based on these experiments, the residual shows a quadratic trend, indicating the presence of another effect: resistance to bending.

5.3.2 Energy in Bending

The shift of frequencies seen in Figure 5.7 is a well known effect due to *stiffness* [147, § 188]. As the string has a finite size, bending causes compression and expansion around the central axis of the string, increasing the potential energy in the string. According to the Euler-Bernoulli beam theory, the energy due to bending is

$$V_{\text{bend}} = \frac{1}{2}\int_0^{\ell_0} (EI_2)(s)[\mathbf{u}_{ss}]_2^2 + (EI_3)(s)[\mathbf{u}_{ss}]_3^2 \,\mathrm{d}s \tag{5.19}$$

where EI_2 and EI_3 are the product of Young's Modulus and the second moment of area in each axis. This is an incomplete model, as it omits twisting and other motions; see Antman for a full model [7, Ch. 8]. Unlike these fuller models, we can treat stiffness without introducing additional degrees of freedom.

Under the assumption that EI_2 and EI_3 are constant, the stationary solution $\mathbf{u}$ of the action functional solves [35, § IV.10]

$$\begin{aligned} \rho A[\mathbf{u}_{tt}^{(1)}]_2 &= \frac{EA(\ell-\ell_0)}{\ell}[\mathbf{u}_{ss}^{(1)}]_2 - EI_2[\mathbf{u}_{ssss}^{(1)}]_2 \\ \rho A[\mathbf{u}_{tt}^{(1)}]_3 &= \frac{EA(\ell-\ell_0)}{\ell}[\mathbf{u}_{ss}^{(1)}]_3 - EI_3[\mathbf{u}_{ssss}^{(1)}]_3. \end{aligned} \tag{5.20}$$

The higher-order term necessitates two additional boundary conditions to specify a unique solution.

If the additional boundary conditions are *hinged*, $\mathbf{u}_{ss}(0,t) = \mathbf{u}_{ss}(\ell_0,t) = 0$, the same separation of variables as (5.16) may be applied, yielding eigenvalues

$$\omega_{n,k} = in\sqrt{\frac{\tau\pi^2}{\rho A\ell\ell_0} + \frac{EI_k}{\rho A}\left(\frac{\pi}{\ell_0}\right)^4 n^2} \qquad k = 2,3. \tag{5.21}$$

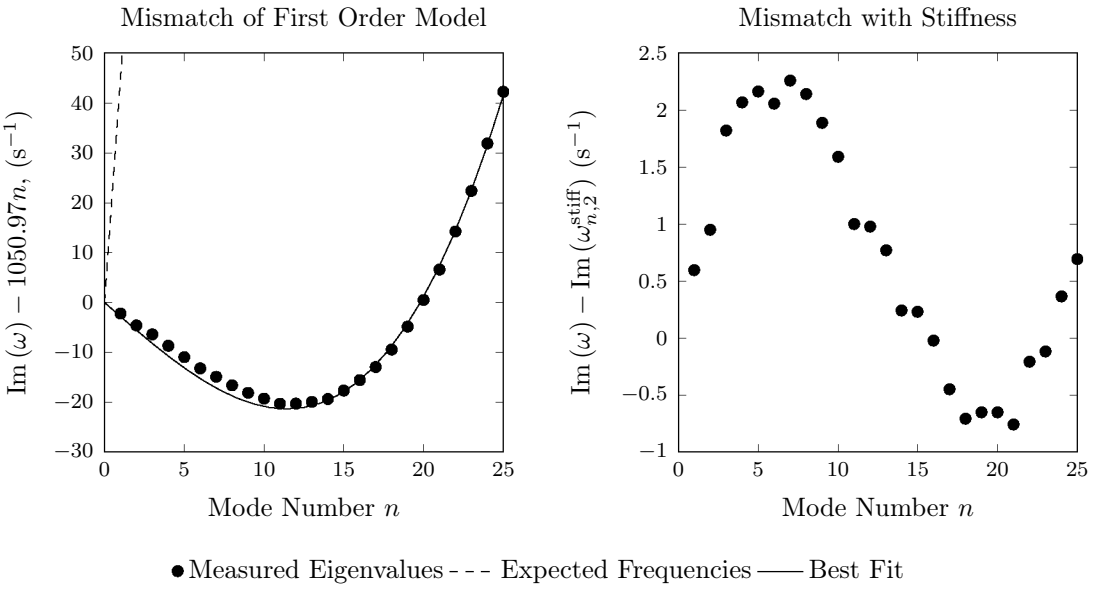


Figure 5.8 : Stiffness corrections to eigenvalue frequencies. Using the linear model in Figure 5.7, the left plot shows the residual from the linear fit using a fundamental of $1050.97\,\mathrm{s}^{-1}$ and the curve shows the best fit (5.23). The right plot shows the residual between (5.23) and the measured eigenvalues.

The second moment of area for a circular cylinder about a perpendicular axis is $I = \pi r^4/4$, where r is the radius of the cylinder. Using the measured values of these parameters, we have

$$\omega_{n,2} \approx i\sqrt{(1\,201\,000\,\mathrm{s}^{-1})n^2 + (16.327\,\mathrm{s}^{-2})n^4}, \tag{5.22}$$

This compares favorably to the least squares fit in Figure 5.8:

$$\omega_{n,2}^{\mathrm{stiff}} \approx i\sqrt{(1\,098\,661\,\mathrm{s}^{-1})n^2 + (15.001\,\mathrm{s}^{-2})n^4}. \tag{5.23}$$

Even if the boundary conditions are not hinged, we can treat the stiffness term as a perturbation, since $EI/\ell_0^2 = 3.6 \times 10^{-4}\,\mathrm{N}$ is much smaller than $\tau = 131\,\mathrm{N}$. Then the separation of variables in (5.16) corresponds to the unperturbed system and the perturbed eigenvalues are given by (5.21).

Table 5.2 : Properties of Air

Parameter	Symbol	Approximate Value
Density	ρ_a	$1.2\,\mathrm{kg/m^3}$
Viscosity	μ	$18.6\,\mu\mathrm{Pa\,s}$
Kinematic viscosity	$\nu = \frac{\mu}{\rho_a}$	$1.55 \times 10^{-5}\,\mathrm{m^2/s}$

An alternative boundary condition enforces zero slope at each end, termed *clamped* endpoints, $\mathbf{u}_s(0,t) = \mathbf{u}_s(\ell_0,t) = 0$. The spacial component in the separation of variables then depends on the root of a transcendental equation [147, § 189]. As we are unsure what boundary condition the collets enforce, this may account for the pattern in the residuals in Figure 5.8; however this effect is on the order of $5\,\mathrm{s}^{-1}$, the same order of magnitude as the unaccounted for frequency shift due to damping in Section 5.3.4, so an identification is unwarranted.

5.3.3 Air Damping

The experimental results just discussed were performed in our pressure vessel at 0.008 atm. We now study the effect of pressure on the recovered spectrum. As Figure 5.9 shows, the decay rate of the eigenvalues depends strongly on the pressure. The primary mechanism coupling of the string to the surrounding air is the Naiver-Stokes equations. An additional effect is slight change in thermal conduction efficiency from the string into the surrounding air (see Section 5.3.5).

Fluid flow is characterized by the Reynolds number [13, §4.7, p. 245],

$$R_e = \frac{\rho v d}{\mu} = \frac{v d}{\nu} \tag{5.24}$$

where ρ_a is the density of the fluid, v, the mean velocity, d, the characteristic length, μ the dynamic viscosity of the fluid, and $\nu = \mu/\rho$ the kinematic viscosity; approximate

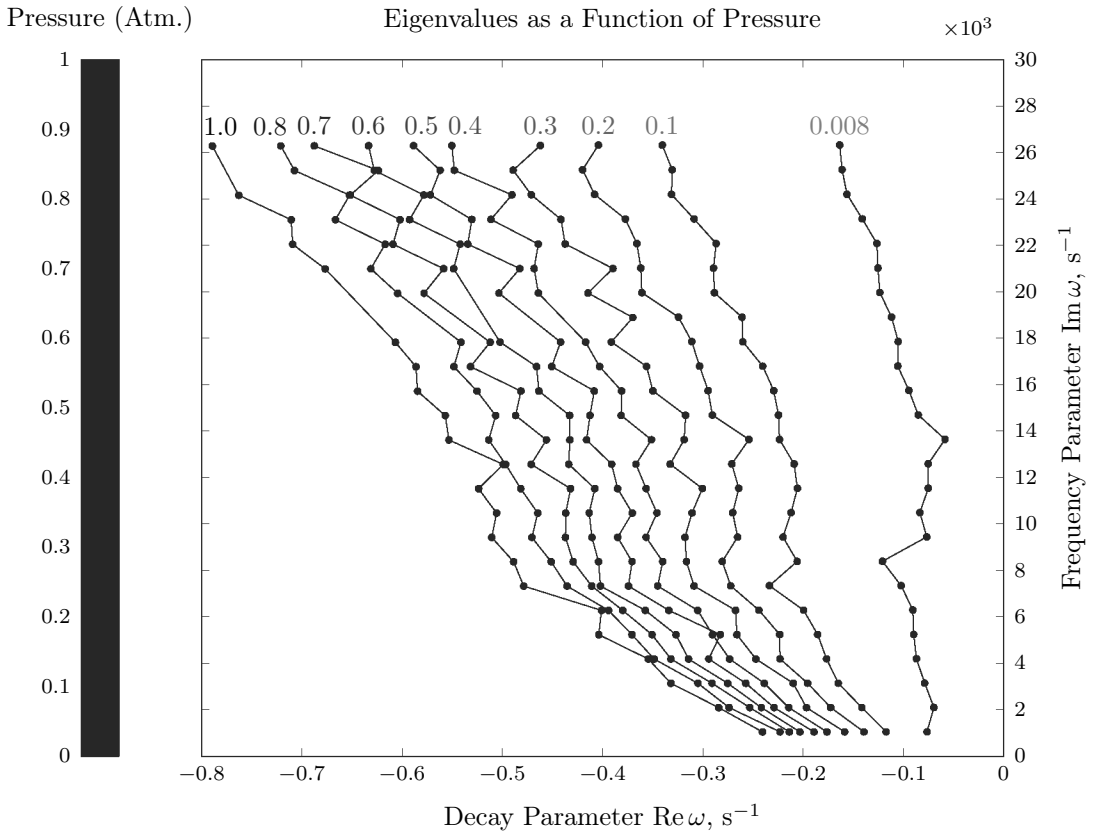


Figure 5.9 : Eigenvalues of a vibrating string as a function of atmospheric pressure. These eigenvalues were estimated using the decay-robust exponential fitting algorithm given in Section 5.2, from measurements described in Section 5.1.

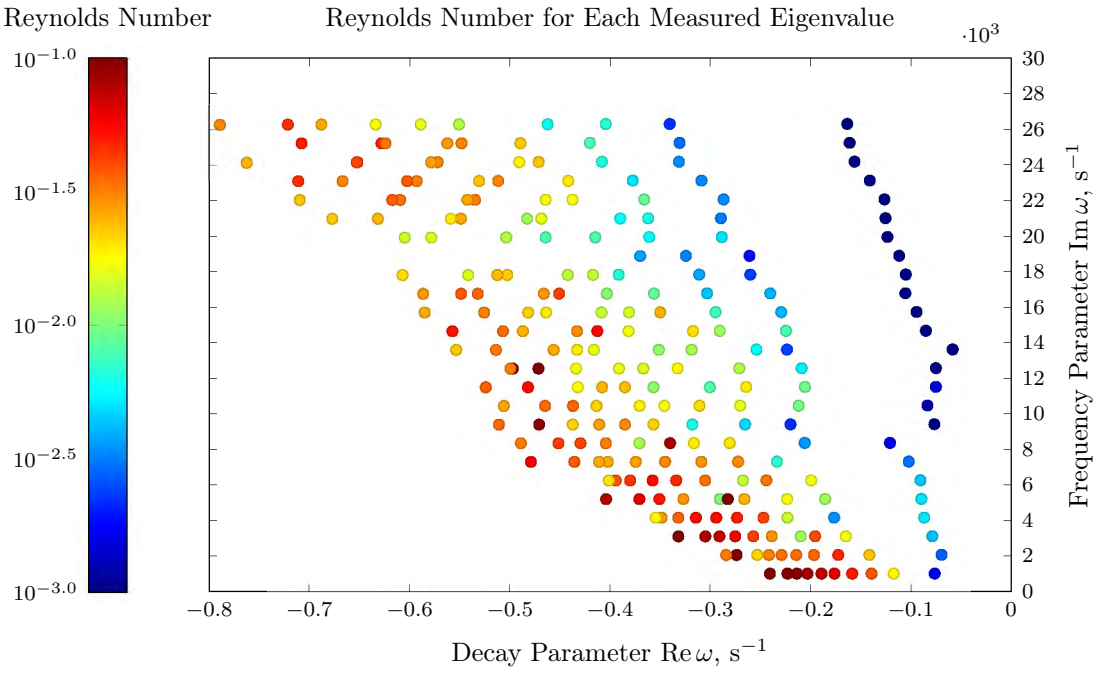


Figure 5.10 : Reynolds number

values are given in Table 5.2. For a vibrating string, the characteristic length is the diameter $d = 0.015\,\text{in}$ and the characteristic velocity for the kth mode is the maximum velocity $a\text{Im}\,\omega$ under the assumption $[\mathbf{u}(s,t)]_k = \text{Re}\,[ae^{\omega t}]\sin(n\pi s)$ for the transverse modes $k = 2, 3$.

Using these parameters, the Reynolds number for the string at pressure P is

$$R_e = (24\,\text{s/m})\,a\text{Im}\,\omega\frac{P}{1\,\text{atm}}. \tag{5.25}$$

The effects of frequency and amplitude on the Reynolds number generally cancel: high frequency modes tend to have smaller vibrations, as seen in Figure 5.10.

The quasi-steady approach models the string as moving at a constant velocity. Although there is no steady state solution to the Stokes equations around an infinite cylinder due to Stokes' paradox, Oseen's equations resolve this paradox and yield a

drag per unit length of

$$D = 2\pi\mu v C = 2\pi\mu v \frac{2}{\log(7.4/R_e)} = \frac{4\pi\mu v}{\log(7.4) - \log(\rho d/\mu) - \log(v)}, \tag{5.26}$$

[13, eq. (4.10.13), (4.10.15)]. Applying this drag formula to a vibrating string is complicated by the nonlinear dependence on v, both directly and through R_e. Decreasing velocity decreases the drag coefficient C, so we may bound the energy lost due to air with C corresponding to the maximum Reynolds number.

Drag following (5.26) adds an external force opposing the motion of the string:

$$\mathbf{f}_D(\mathbf{u}_t) = [0, \delta[\mathbf{u}_t]_2, \delta[\mathbf{u}_t]_3]^\top, \quad \text{where } \delta = \frac{4\pi\mu}{\log(7.4/R_e)}. \tag{5.27}$$

This adds an external forcing term to (5.20) in both transverse axes [35, § IV.10]

$$\begin{aligned} \rho A[\mathbf{u}_{tt}^{(1)}]_1 &= EA[\mathbf{u}_{ss}^{(1)}]_1 \\ \rho A[\mathbf{u}_{tt}^{(1)}]_k &= \frac{EA(\ell - \ell_0)}{\ell}[\mathbf{u}_{ss}^{(1)}]_k - EI_3[\mathbf{u}_{ssss}^{(1)}]_k + \delta[\mathbf{u}_t^{(1)}]_k, \quad k = 2, 3. \end{aligned} \tag{5.28}$$

Solving these equations using separation of variables, the eigenvalues of the two transverse modes are

$$\omega_{n,k} = -\frac{\delta}{\rho A} + in\sqrt{\frac{\tau\pi^2}{\rho A \ell \ell_0} + \frac{EI_k}{\rho A}\left(\frac{\pi}{\ell_0}\right)^4 n^2 - \frac{\delta}{\rho A}} \qquad , k = 2, 3. \tag{5.29}$$

For our string, $R_e = 0.3$ is an appropriate upper bound on the Reynolds number, giving $\delta/(\rho A) \approx 0.14\,\mathrm{s}^{-1}$.

Although this quasi-steady flow drag estimate gives the right order of magnitude, changes in pressure only coupled through the changing Reynolds number (5.25). Rather than using the quasi-steady estimate, Lin [101] directly couples the wave equation to a solution of the Stokes equation in the limit of either small Reynolds number ($R_e \ll 1$) or small oscillation $a \ll b$, where a is the amplitude of oscillation and

$b = d/2$ the radius of the cylinder (string). This second case applies to this experiment; the largest amplitude vibration is 8 µm, much smaller than the 381 µm diameter of the string. Then, following [101, eq. (6)], the force opposing motion is

$$\frac{f_D(\tau)}{\rho_a \nu} = \pi \dot{v}(\tau) + \frac{16}{\pi} \int_0^\tau \int_0^\infty \dot{v}(\lambda) \frac{e^{-x^2(\tau-\lambda)}}{x(J_0^2(x) + Y_0^2(x))} \, \mathrm{d}x \, \mathrm{d}\lambda \tag{5.30}$$

where $\tau(b^2/\nu) = t$, $\dot{v}$ denotes differentiation with respect to τ, and J_0 and Y_0 are Bessel functions of the first and second kind. Coupling this damping model to the motion of the string results in an integro-differential equation [101, eq. (10)]. However, since there is no known solution, we treat this force as a perturbation to $\mathbf{u}^{(1)}$, estimating the drag coefficient while assuming motion is restricted to a plane. Treating each term separately in (5.16), we have $v_n(s,t) = \omega_{n,2} e^{\omega_{n,2} t}$ and $\dot{v}_n(t) = \omega_{n,2}^2 (b^2/\nu)$. Changing the order of integration yields

$$\begin{aligned} \int_0^\tau \int_0^\infty \dot{v}(\lambda) \frac{e^{-x^2(\tau-\lambda)}}{x(J_0^2(x) + Y_0^2(x))} \, \mathrm{d}x \, \mathrm{d}\lambda &= \frac{b^2}{\nu} \int_0^\infty \frac{\omega_{n,2}^2 e^{-x^2\tau}}{x(J_0^2(x) + Y_0^2(x))} \int_0^\tau e^{(\omega_{n,2} + x^2)\lambda} \, \mathrm{d}\lambda \, \mathrm{d}x \\ &= \frac{b^2}{\nu} e^{\omega_{n,2}\tau} \int_0^\infty \frac{\omega_{n,2}^2 (1 - e^{-x^2\tau})}{x(J_0^2(x) + Y_0^2(x))(\omega_{n,2} + x^2)} \, \mathrm{d}x. \end{aligned}$$

Taking the limit $\tau \to \infty$ removes the time dependent term, leaving

$$\Phi(\omega_{n,2}) := \int_0^\infty \frac{\omega_{n,2}}{x(J_0^2(x) + Y_0^2(x))(\omega_{n,2} + x^2)} \, \mathrm{d}x, \tag{5.31}$$

with force per mode:

$$\frac{f_D(\tau)}{\rho_a b^2} \approx \pi \omega_{n,2}^2 e^{\omega_{n,2} t} + \omega_{n,2} e^{\omega_{n,2} t} \frac{16}{\pi} \Phi(\omega_{n,2}). \tag{5.32}$$

Adding this forcing term to (5.20) yields the following equation for the eigenvalues:

$$\rho A \omega_{n,2}^2 = -\frac{\tau \ell_0}{\ell} (n\pi/\ell_0)^2 - EI_2 (n\pi/\ell_0)^4 - (\rho_a b^2 \pi) \omega_{n,2}^2 - \omega_{n,2} \frac{16 \rho_a b^2}{\pi} \Phi(\omega_{n,2}). \tag{5.33}$$

Figure 5.11 shows the predicted damping versus pressure. While this model is off by roughly a factor of two from measurements at one atmosphere and underestimates

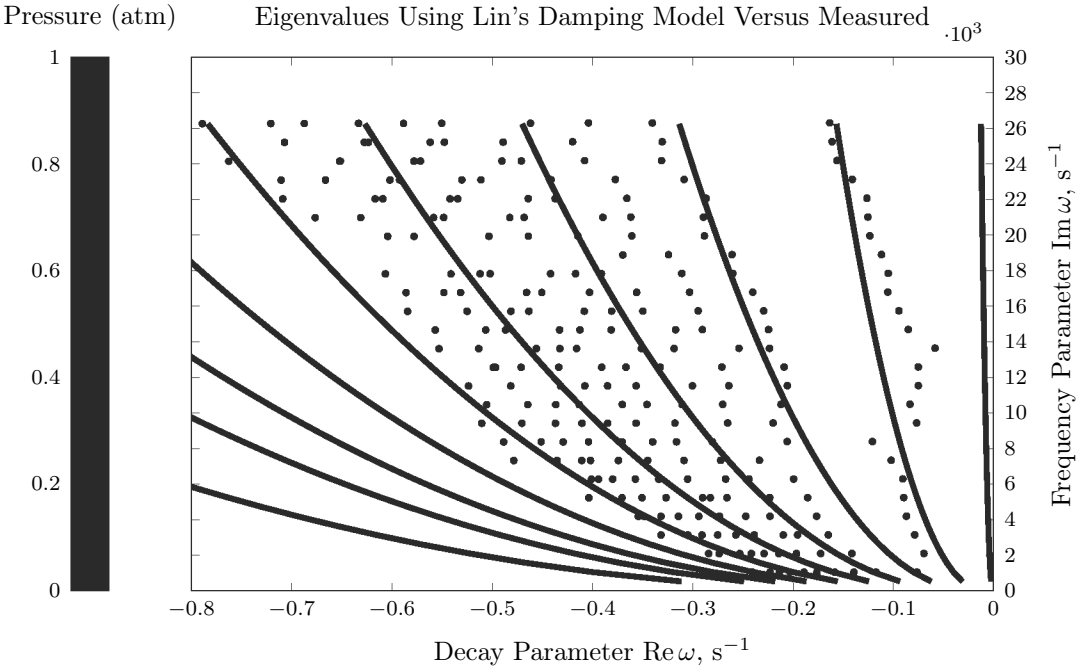


Figure 5.11 : Expected eigenvalues using Lin's damping model. Dots show the measured eigenvalues while curves show the expected decay rate using Lin's damping model. Lin's model appears accurate to within a factor of two of the true damping for pressures near one atmosphere, but underestimates damping at low pressures.

damping at 0.008 atm, it qualitatively captures the effects of pressure and frequency on the damping that we observe in our physical experiments.

5.3.4 Frequency Shift Due to Damping

Two effects conspire to decrease the frequencies: increased damping and the added mass captured by Lin's model. The solutions for the eigenvalues in (5.33) are

$$\omega_{n,2} = -\frac{8\rho_a}{\pi^2(\rho+\rho_a)}\Phi(\omega_{n,2}) + ni\sqrt{\frac{\tau\pi^2}{\ell\ell_0 A(\rho+\rho_a)} + \frac{EI\pi^4}{\ell_0^4 A(\rho+\rho_a)}n^2 - \frac{8\rho_a}{\pi^2(\rho+\rho_a)}\Phi(\omega_{n,2})}. \tag{5.34}$$

This formula predicts a shift of at most $0.07\,s^{-1}$ for the fundamental ($n = 1$); however Figure 5.12 shows a shift of $3\,s^{-1}$. The mismatch in the shift grows with increasing

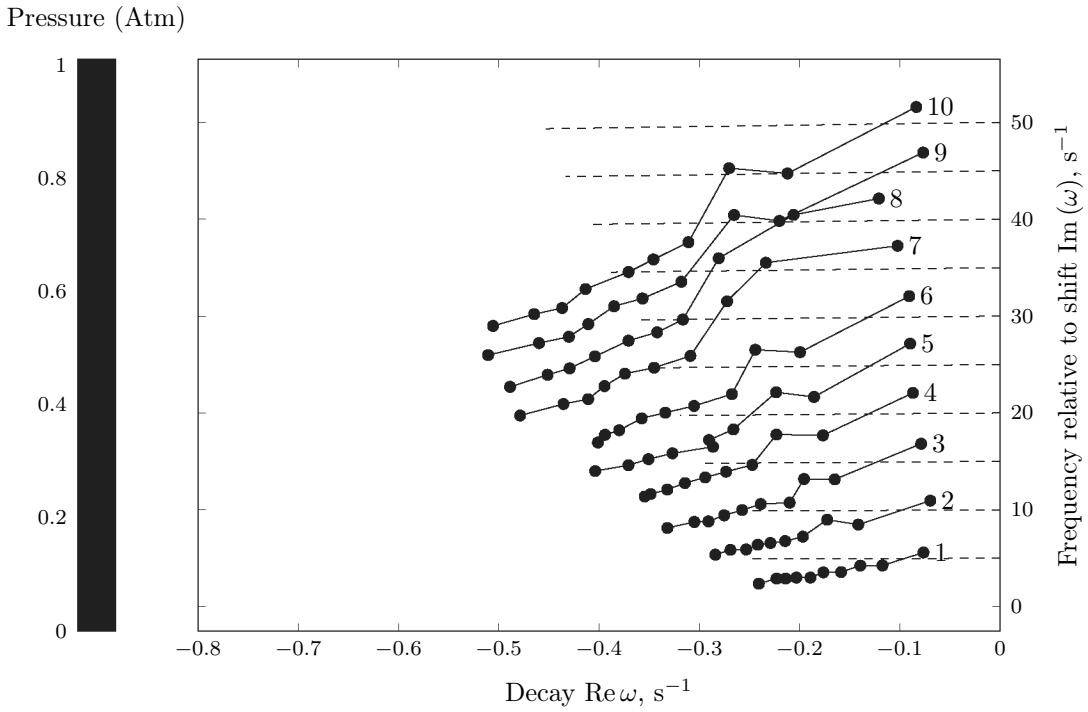


Figure 5.12 : Dependence of frequency on damping and pressure. The dots show the real and imaginary parts of ω relative to an arbitrary imaginary shift as a function of mode number. The dashed curves show the predictions of Lin's damping model with tuned parameters. Frequency shifts with changing pressure are an order of magnitude larger than predicted by Lin's model.

n, but the dependence on damping remains linear throughout. This indicates there may be additional, neglected physics relevant to the eigenvalues.

5.3.5 Thermal Effects

In addition to damping from the surrounding fluid, there are also internal mechanisms of damping. One hypothesis is that, in analogy to gasses, the local temperature in a solid increases and decreases with compression and expansion [168]. This leads to damping as the heat diffuses through the solid and through contact with the surrounding medium. To first order, this affects only the longitudinal modes of

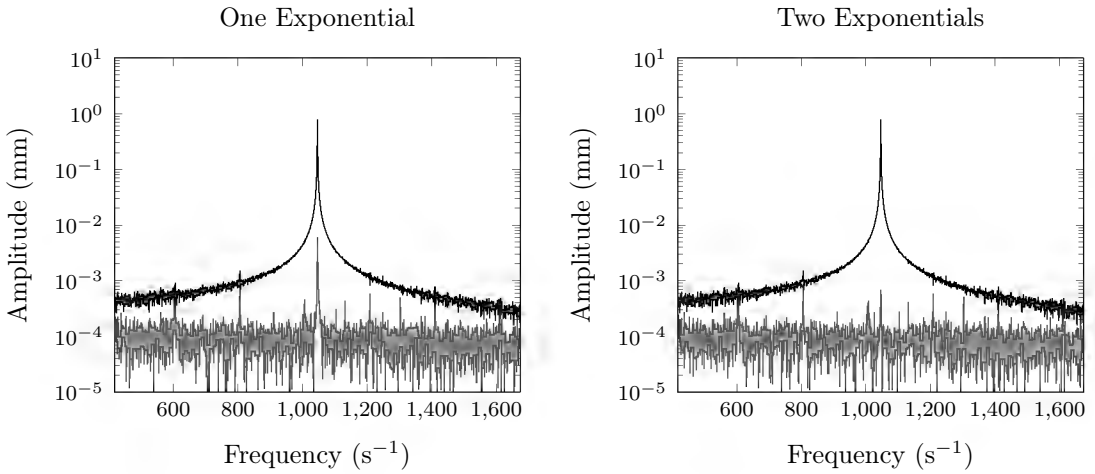
vibration, since transverse motion only induces an infinitesimal compression proportional to $EI\mathbf{u}_{ssss}(s,t)$. Figure 5.11 may show evidence of this at low pressure, as the eigenvalues have much stronger decay than predicted by Lin's model, but this also may be due to the approximations in Lin's model. For further discussion of thermal effects, see Antman [7, §12.14].

5.3.6 Eigenvalue Splitting and Nonlinear Effects

Thus far, we have treated the two transverse axes of motion $[\mathbf{u}^{(1)}]_2$ and $[\mathbf{u}^{(1)}]_3$ as identical. This yields two identical eigenvalues, described by the preceding estimates of $\omega_{n,2}$ and $\omega_{n,3}$. However, slight difference in the material properties in each axis can split the duplicate eigenvalues into two distinct, but close, eigenvalues. This effect is difficult to resolve. Figure 5.13 shows one example where this frequency splitting is successfully resolved. However, for other modes this effect is not well resolved, as illustrated in Figure 5.14. This may be due to either convergence to a spurious minima by the exponential fitting routine or an additional (nonlinear) term perturbing the solution $\mathbf{u}(s,t)$.

5.3.7 Further Refinement

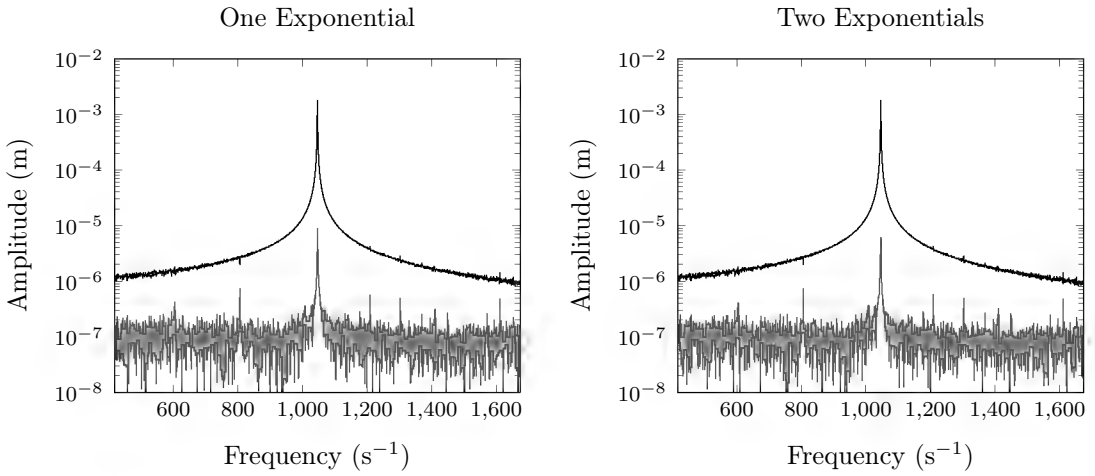
There are many additional effects that may change the eigenvalues and these may explain the mismatch between theory and measurements. For example, the decay rates predicted by Lin's damping model are too large at atmospheric pressure and too low at vacuum. This may be due to the second term in the series expansion of the Navier-Stokes equations, which exposes a four-fold symmetric boundary layer around the string that is confirmed by experiments [67]. Alternatively, damping at low pressure may be dominated by internal friction mechanisms. Further affecting the



$\operatorname{Re}\omega$	$\operatorname{Im}\omega$	$\lvert a\rvert$
-0.197703	1,045.993	$1.983 \cdot 10^{-6}$
-0.197703	627,272.537	$2.036 \cdot 10^{-6}$

$\operatorname{Re}\omega$	$\operatorname{Im}\omega$	$\lvert a\rvert$
-0.314306	1,045.934	$4.223 \cdot 10^{-7}$
-0.205932	1,046.013	$1.975 \cdot 10^{-6}$
-0.314306	627,272.596	$1.156 \cdot 10^{-6}$
-0.205932	627,272.518	$1.171 \cdot 10^{-6}$

Figure 5.13 : Resolved eigenvalue splitting. The top row shows the Fourier transform of $\mathbf{y}$ and the residual (gray). The table below shows the fitted parameters of pairs of conjugate exponentials. This example shows the $n = 10$ mode at 0.008 atm. Although the units of the vertical axis are millimeters, the amplitude of a peak corresponding to a unit sine wave has amplitude $\sqrt{n} = 10^3$; hence the amplitude of the mode shown here is $\approx$1 µm.



$\mathrm{Re}\,\omega$	$\mathrm{Im}\,\omega$	$\lvert a\rvert$
-0.225299	1,045.884	$5.684 \cdot 10^{-6}$
-0.225299	627,272.647	$5.633 \cdot 10^{-6}$

$\mathrm{Re}\,\omega$	$\mathrm{Im}\,\omega$	$\lvert a\rvert$
-0.084469	1,046.493	$2.944 \cdot 10^{-8}$
-0.229081	1,045.885	$5.739 \cdot 10^{-6}$
-0.084469	627,272.038	$1.91 \cdot 10^{-7}$
-0.229081	627,272.646	$5.629 \cdot 10^{-6}$

Figure 5.14 : Unresolved eigenvalues. Identical to Figure 5.13, except this figure shows the fit for the fundamental $n = 1$ at 1 atm. In this example, the mode is not clearly resolved into two exponentials, either due to poor initial estimates or a physical effect.

damping parameters, the diameter of the string may change in response to elongation in a process called *necking*.

The one unexplained phenomenon is the unexpected frequency shift with changing pressure and decay rate described in Section 5.3.4. Capturing this effect requires a more sophisticated beam model, introducing $\mathbf{u}_{tttt}$ terms corresponding to the kinetic energy of the wire with respect to rotations about the transverse axes, as well as additional potential energy due to shearing. These models include Timoshenko's beam model (e.g., [32, §7.7], experiments [133]) and Antman's rod theory [7, Ch. 8].

5.4 Conclusion

Using the compressed exponential fitting routines developed in Chapter 4, we were to obtain accurate estimates of the eigenvalues of a string from measurements of its evolution. Each of the 2,894 experiments took one million measurements of string displacement. Compression made it possible to quickly and efficiency analyse this data, thus illustrating how compression can provide insight into other mechanical systems, such as those applications described in [102].

Chapter 6

Conclusions and Future Work

In this thesis, we have focused on two classes of exponential fitting algorithms: those similar to Prony's Method that recover the complex frequencies $\boldsymbol{\omega}$ through an implicit autoregressive model, and those that use an optimization approach maximizing the likelihood of $\boldsymbol{\omega}$. Chapter 2 summarized existing theoretical results for many variants of Prony's Method, provided new asymptotic estimates of the bias and covariance in Prony's Method, and extended these results to Kung's Method. We also developed new variants of Prony's Method, including the Maximum Likelihood Prony Method, the Filtered Matrix Pencil Method, and the Orthogonalized Matrix Pencil Method. In the maximum likelihood framework, we provided a new technique for compressing general nonlinear least squares problems onto small subspaces. We generalized Fischer's efficiency for multi-parameter problems and used this to construct subspaces that capture the salient features of the data. A closed-form inner product between the model function $\mathbf{f}(\boldsymbol{\theta})$ and the basis for the compression space $\mathbf{W}$ allows optimization steps to be independent of the number of measurements. For the exponential fitting problem, compression spaces built from columns of block diagonal Fourier parent coordinates satisfy both qualities: the inner product can be implmented in closed-form inner product and provides an efficient subspace that could be constructed for every $\boldsymbol{\omega}$.

Exponential fitting, although an important problem in many fields, is a simple example of a nonlinear least squares problem. Compression may yield improved

performance for other, more challenging nonlinear least squares problems. In this final chapter, we discuss remaining open problems involving Prony's Method and examine how compression can be applied to other areas of research.

6.1 Further Extensions of Prony's Method

Chapter 2 focused on the statistical and numerical properties of many variations of Prony's Method and built several new algorithms. This investigation pointed to several remaining extensions and improvements of Prony's Method we discuss here.

6.1.1 Numerically Stable Maximum Likelihood Prony Methods

As illustrated in Figure 2.3, many maximum likelihood Prony methods are numerically unstable when fitting more than 20 exponentials. This instability can be fixed by including extraneous exponentials, as shown in Figure 2.7, but these extraneous exponentials are a nuisance to remove. Numerically stable Prony variants, such as Kung's Method, implicitly remove these spurious exponentials, but fail to provide maximum-likelihood estimates. An ideal variant of Prony's Method would provide both maximum likelihood estimates while implicitly filtering the spurious exponentials.

6.1.2 Improving Performance

The computational cost for many numerically stable Prony-type methods is dominated by the cost of computing the singular value decomposition. Most implementations developed in the 1980s use classical dense algorithms for this operation, but two new classes of algorithms exist: Krylov subspace methods and randomized matrix factorizations.

The Implicitly Restarted Arnoldi Method (IRAM) [142] in ARPACK [98], a Krylov subspace method, computes the rank-k SVD of an $n \times n$ matrix using multiple iterations of the Arnoldi cycle, each requiring k matrix-vector products and $\mathcal{O}(nk^2 + k^3)$ additional operations. Combined with fast Hankel-vector products [113, §3.4] that require only $\mathcal{O}(kn \log n)$ operations, IRAM can provide the rank-k SVD much faster than classical dense SVD algorithms that require $\mathcal{O}(n^3)$ operations. The combination IRAM and the fast Hankel-vector products with Kung's Method was discussed in a 2005 thesis by Laudadio [96, Ch. 3].

Randomized matrix factorizations are a recent alternative for large problems [61]. Although these methods are still in their infancy, they may provide significant speed improvements for Prony-type algorithms applied to huge problems ($n > 10^6$) for which even Krylov methods are infeasible. These fast random factorizations deliver left singluar vectors whose accuracy satsifies probabilistic error bounds. The effect of the resulting approximation on Kung's Method or HTLS should be explored before these two algorithms are combined.

6.1.3 Restarted Matrix Pencil Method

The Orthogonalized Matrix Pencil Method (Algorithm 2.15) developed in this thesis can estimate $\boldsymbol{\omega}$ near specified frequencies to high precision, provided the frequencies are well separated — a feature not shared by other variants of Prony's Method. However, when the frequencies to recover are clustered, this technique fails to converge. Inspired by the Implicitly Restarted Arnoldi Method [142], we would like to restart the Orthogonalized Matrix Pencil Method using a new compression space that utilizes information about the current estimates, both wanted and unwanted, to filter unwanted exponentials in future iterations. Ideally when this method converges,

the left and right compression spaces remove the influence of exponentials outside the desired region, yielding near maximum-likelihood estimates of the desired parameters.

6.2 Improving Compression for Exponential Fitting

Chapter 4 demonstrates the application of compression using block diagonal Fourier parent coordinates in conjunction with conventional nonlinear least squares methods. However, there are many possible compression spaces and alternative optimization techniques that could yield improvements for the exponential fitting problem.

6.2.1 New Compression Parent Coordinates

Chapter 4 exclusively used parent coordinates from the class of block diagonal Fourier matrices

$$\mathbf{U} = \text{diag}(\mathbf{F}_{b_0}, \mathbf{F}_{b_1}, \mathbf{F}_{b_2}, \ldots). \tag{6.1}$$

We used these parent coordinates, since we could construct simple, closed form expressions for $[\mathbf{U}]_{\cdot,i}^{*}\mathbf{V}(\omega)$ and $[\mathbf{U}]_{\cdot,i}^{*}\mathbf{V}'(\omega)$. However, there may be alternative parent coordinates for this problem. Wavelet transforms are an appealing choice due to their recursive structure and the existence of fast transforms; see, e.g., [146]. For example, the unnormalized Haar wavelet transform matrix $\mathbf{H}_k$ is defined recursively:

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \qquad \mathbf{H}_{2^\ell} = \begin{bmatrix} \mathbf{H}_{2^{\ell-1}} \otimes [1 \quad 1] \\ \mathbf{I}_{2^{\ell-1}} \otimes [1 \quad -1] \end{bmatrix}. \tag{6.2}$$

This and other wavelet transforms may have sufficient structure to both have closed-form inner products for $[\mathbf{U}]_{\cdot,i}^{*}\mathbf{V}(\omega)$ and $[\mathbf{U}]_{\cdot,i}^{*}\mathbf{V}'(\omega)$ and yield efficient subspaces.

6.2.2 Alternative Trust Region Quadratic Models

The trust-region optimization method used in Chapters 3 and 4 used a quadratic model of the residual,

$$\|\widetilde{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta} + \mathbf{p})\| \approx \phi(\boldsymbol{\theta}) + \mathbf{g}(\boldsymbol{\theta})^*\mathbf{p} + \frac{1}{2}\mathbf{p}^*\mathbf{H}(\boldsymbol{\theta})\mathbf{p}, \tag{6.3}$$

to construct the next iterate of $\boldsymbol{\theta}$ by minimizing the right hand size over $\mathbf{p}$ subject to the constraint $\|\mathbf{p}\| \leq \Delta$. This is called the trust region subproblem, solved by (3.11). The choice of a quadratic model is motivated by simplicity and generality. Taylor's theorem implies that for smooth functions and a fixed error, there exists an region on which the quadratic model does not exceed the specified error. However, there may be more appropriate models for a particular residual.

Inspired by Bunch, Nielsen, and Sorensen's [25] rational-Newton method for computing the roots of the secular equation in symmetric eigenvalue computations, we too might seek a rational model of the residual. If we consider the residual under the Fourier transform,

$$[\mathbf{F}^*\widetilde{\mathbf{y}} - \mathbf{F}^*\mathbf{V}(\boldsymbol{\omega})\mathbf{a}]_k = [\mathbf{F}^*\widetilde{\mathbf{y}}]_k - \frac{1}{\sqrt{n}}\sum_{j=0}^{p-1}\frac{1 - e^{n\omega_j - 2\pi ik}}{1 - e^{\omega_j - 2\pi ik/n}}a_j, \tag{6.4}$$

each entry of the residual is the square of a constant minus a sum of rational functions of the form $(1-\lambda^n)/(1-\lambda)$. A better model function would generate better iterates of $\boldsymbol{\omega}$, reducing the computational cost by decreasing the number of iterations necessary.

6.2.3 Block Coordinate Descent

When there are many exponentials, the dominant cost for optimization is the $\mathcal{O}(p^3)$ cost for solving a dense, $p \times p$ linear system to find the minimum of the trust region subproblem. One approach to reduce this cost is *block coordinate descent*, also known

as the Gauss-Seidel-Newton Method [115, Ch. 7, eq. (28)]. This technique splits the nonlinear variables $\boldsymbol{\omega}$ into (perhaps disjoint) sets $\boldsymbol{\omega}_k$ and optimizes these subsets sequentially. Compression adds separate compression spaces for each $\boldsymbol{\omega}_k$, $\mathcal{W}_k$. In this variant, each iteration would update $\boldsymbol{\omega}_k$, solving the exponential fitting problem with the remaining variables fixed; i.e.,

$$\boldsymbol{\omega}_k \leftarrow \operatorname*{argmin}_{\boldsymbol{\omega}} \min_{\mathbf{a}} \left\| \mathbf{W}_k \left(\widetilde{\mathbf{y}} - \mathbf{V}(\boldsymbol{\omega})\mathbf{a} - \sum_{j \neq k} \mathbf{V}(\boldsymbol{\omega}_j)\mathbf{a}_j \right) \right\|_2 . \tag{6.5}$$

Convergence is typically linear [112], and this approach is beneficial if the computational savings of splitting the domain outweigh the convergence penalty incurred.

Recently, Nesterov developed a block coordinate descent using a compression space formed from randomly chosen columns of the identity matrix, and showed that for some problems convergence is quadratic [112]. Compression provides an alternative interpretation of these results in a statistical context, and could inform better compression spaces.

6.3 Applying Compression to System Identification Problems

As discussed in Section 1.3, exponential fitting is equivalent to finding the eigenvalues $\boldsymbol{\omega}$ of the matrix $\mathbf{A}$ from samples of the impulse response

$$y(t) = \mathbf{C}e^{t\mathbf{A}}\mathbf{x}_0. \tag{6.6}$$

As compression is effective for the exponential fitting problem, we ask: can compression be applied to other system identification problems? In this section, we discuss several such problems and in two cases show how the block Fourier matrices provide compression subspaces with closed-form inner products with the model function $\mathbf{f}$ via Theorem 4.1.

6.3.1 Multiple Output Impulse Response System Identification

Throughout this thesis, for simplicity, we restricted ourselves to the case where $y(t)$ is a scalar. The general case where $\mathbf{y}(t)$ is a vector corresponds to a *multiple output* system, with $\mathbf{C} \in \mathbb{C}^{m \times p}$. Assuming that noise in the measurements is independent and normally distributed with uniform covariance, we can pose this as the nonlinear least squares problem

$$\min_{\boldsymbol{\omega} \in \mathbb{C}^p, \mathbf{a}_\ell \in \mathbb{C}^p} \sum_{\ell=0}^{m-1} \left\| [\mathbf{Y}]_{\cdot,\ell} - \mathbf{V}(\boldsymbol{\omega}) \mathbf{a}_\ell \right\|_2 \quad \text{where} \quad [\mathbf{Y}]_{\cdot,j} = [\mathbf{y}(t_j)]_\ell. \tag{6.7}$$

Using the n-mode matrix product notation of Bader and Kolda [90, §2.5], we can rewrite (6.7) without the sum

$$\min_{\boldsymbol{\omega} \in \mathbb{C}^p, \mathbf{A} \in \mathbb{C}^{p \times m}} \left\| \mathbf{Y} - \mathbf{V}(\boldsymbol{\omega}) \times_2 \mathbf{A} \right\|_{\mathrm{F}}, \qquad [\mathbf{X} \times_2 \mathbf{A}]_{j,k} := \sum_\ell [\mathbf{X}]_{j,\ell} [\mathbf{A}]_{\ell,k}. \tag{6.8}$$

The structure of this nonlinear least squares problem can be exploited to reduce storage and operation counts using the ideas of Kaufman, Sylvester, and Wright [87]. Compression for this problem follows the same outline as the exponential fitting problem. The results in Section 4.3 provide closed-form inner products for block Fourier matrices with $\mathbf{V}(\boldsymbol{\omega})$. The only complication is selecting columns from the parent coordinates where a modified heuristic is necessary. The new heuristic should follow the procedure in Section 4.4, replacing $\mathbf{Y} - \mathbf{V}(\boldsymbol{\omega}) \times_2 \mathbf{A}$ with the vectorized version $\operatorname{vec}(\mathbf{Y}) - (\mathbf{I}_m \otimes \mathbf{V}(\boldsymbol{\omega})) \operatorname{vec}(\mathbf{A})$, where $\otimes$ is the Kronecker product.

6.3.2 Frequency Domain System Identification

Whereas the time domain system identification determines the eigenvalues $\boldsymbol{\omega}$ of $\mathbf{A}$ from samples of $\mathbf{y}(t)$, the frequency domain system identification determines $\boldsymbol{\omega}$ from samples of the *transfer function*,

$$\mathbf{H}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}. \tag{6.9}$$

In general, the transfer function of a finite dimension system is written as the ratio of two polynomials,

$$\mathbf{H}(s) = \frac{\sum_{j=0}^{p} \boldsymbol{\beta}_j s^j}{s^p + \sum_{j=0}^{p-1} \alpha_j s^j}, \tag{6.10}$$

which, if $\mathbf{A}$ is diagonalizable, can be written more simply in pole-residue form:

$$\mathbf{H}(s) = \sum_{j=0}^{p-1} \frac{[\boldsymbol{\mathcal{A}}]_{\cdot,\cdot,j}}{s - \omega_j}. \tag{6.11}$$

This special case of the rational approximation problem is the frequency domain counterpart to the exponential fitting problem. The Loewner (Löwner) framework developed by Anderson and Antoulas [5] plays the role of Prony's Method in the frequency domain; see also [89, 97]. The Loewner framework uses a generalized eigenvalue problem to reveal the signal poles $\boldsymbol{\omega}$, paralleling the matrix pencil variant of Prony's Method. However, unlike Prony's Method, numerical experiments suggest that when the number of measurements matches the number of parameters, the Loewner framework produces maximum likelihood estimates. Further, as the number of measurements increases, performance decreases slightly (i.e., $\approx 95\%$) in a similar manner to Kung's Method.

The associated maximum likelihood formulation for $\boldsymbol{\omega}$ and $\boldsymbol{\mathcal{A}}$ when the noise in the measurements $[\boldsymbol{\mathcal{H}}]_{\cdot,\cdot,k} = \mathbf{H}(s_k)$ is independently and identically normally distributed, yields the following nonlinear least squares problem for the poles $\boldsymbol{\omega}$ and residues $\boldsymbol{\mathcal{A}}$:

$$\min_{\boldsymbol{\omega},\boldsymbol{\mathcal{A}}} \|\boldsymbol{\mathcal{H}} - \boldsymbol{\mathcal{A}} \times_3 (\mathbf{R}(\boldsymbol{\omega})^\top)\|_{\mathrm{F}}, \quad \text{where} \quad [\mathbf{R}(\boldsymbol{\omega})]_{j,k} = (s_j - \omega_k)^{-1}. \tag{6.12}$$

Here, the n-mode matrix product is

$$[\boldsymbol{\mathcal{A}} \times_3 \mathbf{X}]_{i_1,i_2,i_3} = \sum_j [\boldsymbol{\mathcal{A}}]_{i_1,i_2,j} [\mathbf{X}]_{j,i_3}.$$

Compression can be applied to the rational approximation problem; however, at this point, we know of no parent coordinates that satisfy the closed-form property other

than the identity matrix. However, the identity matrix parent coordinates is likely efficient, since entries of $\mathbf{R}(\boldsymbol{\omega})^\top$ are only large when s_j is near some ω_k.

6.3.3 System Identification with Known Inputs

Finally, we can use compression to find a realization of a system subject to a known input. For an arbitrary deterministic system subject to a known input $\mathbf{u}(t)$, the output is

$$\mathbf{y}(t) = \mathbf{D}\mathbf{u}(t) + e^{t\mathbf{A}}\mathbf{x}_0 + \int_0^t \mathbf{C}e^{(t-s)\mathbf{A}}\mathbf{B}\mathbf{u}(s)\,\mathrm{d}s. \tag{6.13}$$

For notational simplicity, we include the initial condition term in the input $\mathbf{u}(s)$ by replacing $\mathbf{u}(s)$ with $\mathbf{u}(s) + \delta_+(t)\mathbf{x}_0$ where $\delta_+(t)$ is the right-sided Dirac delta function. If we again make the assumption that $\mathbf{A}$ is diagonalizable with eigenvalues $\boldsymbol{\omega}$, the solution for $\mathbf{y}(t)$ is the convolution of $e^{\omega t}$ with $\mathbf{u}(t)$:

$$\mathbf{y}(t) = \mathbf{D}\mathbf{u}(t) + \sum_{j=0}^{p-1} [\mathbf{C}\mathbf{V}]_{\cdot,j}[\mathbf{V}^{-1}\mathbf{B}]_{j,\cdot} \int_0^t e^{\omega_j(t-s)}\mathbf{u}(s)\,\mathrm{d}s. \tag{6.14}$$

Sampling $\mathbf{u}(t)$ and $\mathbf{y}(t)$ into matrices $[\mathbf{U}]_{j,k} = [\mathbf{u}(k\delta)]_j$ and $[\mathbf{Y}]_{j,k} = [\mathbf{y}(k\delta)]_j$, we seek eigenvalues $\boldsymbol{\omega}$ and linear parameters $\mathbf{D}$ and $\boldsymbol{\mathcal{X}}$ solving

$$\begin{aligned} &\min_{\boldsymbol{\omega}\in\mathbb{C}^p, \mathbf{D}\in\mathbb{C}^{q\times q}, \boldsymbol{\mathcal{X}}\in\mathbb{C}^{m\times q\times p}} \left\| \mathbf{Y} - (\mathbf{D}\mathbf{U}^\top)^\top - \sum_{\ell=0}^{p-1} ([\boldsymbol{\mathcal{X}}]_{\cdot,\cdot,\ell}\mathbf{V}_{\mathbf{u}}(\omega_\ell)^\top)^\top \right\|_{\mathrm{F}} \\ &\text{where } [\mathbf{V}_{\mathbf{u}}(\omega)]_{j,k} = \int_0^{\delta j} e^{(\delta j - s)\omega}[\mathbf{u}(s)]_k\,\mathrm{d}s. \end{aligned} \tag{6.15}$$

As with the previous problems, we can apply the insights of [87] to reduce the computational burden of this problem.

Although $\mathbf{V}_{\mathbf{u}}(\omega_\ell)$ is more complicated that $\mathbf{V}(\omega)$ for the exponential fitting problem, we can apply the closed form inner product results of Section 4.3 by approximating $\mathbf{u}(t)$ in a basis for which $\mathbf{V}_{\mathbf{u}}(\omega_\ell)$ does have a closed form expression. One such

basis contains polynomials, exponentials, and the Heaviside step function $\Theta(t)$:

$$u_{q,\phi,\tau}(t) = t^{q_k}\ e^{\phi_k t}\Theta(t-\tau_k). \tag{6.16}$$

The convolution of this element with an exponential is then

$$\int_0^t e^{\omega(t-s)}u_{q,\phi,\tau}(s)\,\mathrm{d}s = \begin{cases} e^{\omega t}\ e^{(\phi-\omega)s}\left(\displaystyle\sum_{\ell=0}^{q}\frac{(-1)^\ell\ q^{\underline{\ell}}\ s^{q-\ell}}{(\phi-\omega)^\ell}\right)\Bigg|_{s=\min(0,\tau)}^{t}, & \phi\neq\omega; \\ e^{\omega t}\ \dfrac{s^{q+1}}{q+1}\Bigg|_{s=\min(0,\tau)}^{t}, & \phi=\omega; \end{cases} \tag{6.17}$$

where $q^{\underline{\ell}}$ is the falling factorial power [59, eq. (2.43)], $q^{\underline{\ell}} = q(q-1)\dots(q-\ell+1)$. Then, since each of these expressions in (6.17) is the product of an exponential and a polynomial, Theorem 2.5 provides a formula for a Fourier matrix times $\mathbf{V}_{u_{q,\phi,\tau}}(\omega)$.

The set of functions $\{u_{q,\phi,\tau}\}(t)\}_{q\in\mathbb{Z}^+,\phi\in\mathbb{C},\tau\mathbb{R}^+}$ is quite general; it includes polynomials, piecewise polynomials, sines, and cosines. As such, a sum of these functions can approximate any continuous function to arbitrary accuracy. However, the challenge is finding a compact representation of $\mathbf{u}$ in these functions; both additional terms and high polynomial orders increase the number of operations required to compute $\mathbf{V}_{\mathbf{u}}$. Further, the geometric block parent coordinates that proved so successful for the exponential fitting problem will provide poor subspaces here. For example, if $\mathbf{u}$ consists of a single step halfway through the measurements, then $\mathbf{G}_4$ will not efficiently capture rapidly decaying eigenvalues expressed at this transition, as the step occurs in the last and largest block. Instead, new compressions spaces are necessary, likely built on wavelet bases, and these compression spaces may need knowledge of $\mathbf{u}$ to be efficient.

6.4 Applying Compression to Other Nonlinear Least Squares Problems

In the previous section, we saw how compression can be extended from the exponential fitting problem to several problems in system identification. We ask: can compression be applied to other problems as well? The modern approach to parameter identification problems with vast numbers of measurements is to randomly sample subset of these measurements [50]. In settings where this current approach is applied, compression could replace random sampling to choose a set of measurements that contain significant information about the desired parameters. This could accelerate convergence, even if the compression spaces correspond keeping rows of the original data. Combined with Nesterov's block coordinate descent, compression might play an invaluable role in solving large scale, many parameter models, such as seismic inversion [8].

Appendix A

Statistics

Throughout this thesis we use statistics to measure the efficiency of our recovered parameters. Our main concern is with the expected value (to check for bias) and covariance of $\widetilde{\boldsymbol{\lambda}}$ or $\widetilde{\boldsymbol{\omega}}$. Most results employed are standard, but the setting is slightly different. Statistics generally treats real random variables, but here we use complex random variables of a particular type that correspond to the complex ℓ_2 norm. Most asymptotic statistical results are in the limit of a large number of measurements; here, we assume the number of measurements is fixed, but instead the covariance of the noise decreases uniformly. By treating this limit, we expose biases that emerge when only a few measurements are used.

Most of the definitions follow Schreier and Scharf, wherein more information about complex random variables can be found [137].

A.1 Definitions

For simplicity, we restrict ourselves to continuous random variables $\mathbf{z}$ that map a sample space Ω to $\mathbb{C}^n$; $\mathbf{z} : \Omega \to \mathbb{C}^n$. For each event we assign a probability measure p describing the probability of $\mathbf{z}$ in some measurable set $\mathcal{A} \subseteq \mathbb{C}^n$

$$\mathsf{P}[\mathbf{z} \in \mathcal{A}] := \int_{\mathcal{A}} p(\mathbf{z})\, \mathrm{d}\mathbf{z}; \qquad \mathsf{P}[\mathbf{z} \in \mathbb{C}^n] = 1. \tag{A.1}$$

(More general, measure based formulations are available; see, e.g., [124].) Using p we can define the expectation of $\mathbf{z}$,

$$\mathsf{E}[\mathbf{z}] := \int_{\Omega} \mathbf{z}\, p(\mathbf{z})\, \mathrm{d}\mathbf{z}. \tag{A.2}$$

For real random vectors $\mathbf{x} : \Omega \to \mathbb{R}^n$, covariance is defined as the outer product $\mathsf{E}[\mathbf{x}\mathbf{x}^\top]$. For complex random vectors, we need to consider both the covariance of $\mathbf{z}$ and its conjugate $\overline{\mathbf{z}}$ with each other. Taking inspiration from Wirtinger, we treat the augmented vector

$$\underline{\mathbf{z}} = \begin{bmatrix} \mathbf{z} \\ \overline{\mathbf{z}} \end{bmatrix}.$$

We can then consider the covariance of $\underline{\mathbf{z}}$

$$\mathsf{Cov}[\underline{\mathbf{z}}] = \begin{bmatrix} \mathsf{E}[\mathbf{z}\mathbf{z}^*] & \mathsf{E}[\mathbf{z}\mathbf{z}^\top] \\ \mathsf{E}[\overline{\mathbf{z}}\mathbf{z}^*] & \mathsf{E}[\overline{\mathbf{z}}\mathbf{z}^\top] \end{bmatrix} = \begin{bmatrix} \mathsf{E}[\mathbf{z}\mathbf{z}^*] & \mathsf{E}[\mathbf{z}\mathbf{z}^\top] \\ \overline{\mathsf{E}[\mathbf{z}\mathbf{z}^\top]} & \overline{\mathsf{E}[\mathbf{z}\mathbf{z}^*]} \end{bmatrix}. \tag{A.3}$$

This matrix is similar to the real covariance associated with $\operatorname{Re}\mathbf{z}$ and $\operatorname{Im}\mathbf{z}$ treated as separate random variables [137, eq. 2.4]. When $\mathsf{E}[\mathbf{z}\mathbf{z}^\top] = \mathbf{0}$, $\mathbf{z}$ is a *proper random variable* and the covariance is

$$\mathsf{Cov}[\mathbf{z}] = \mathsf{E}[\mathbf{z}\mathbf{z}^*]. \tag{A.4}$$

A sufficient condition for $\mathbf{z}$ to be proper is $\operatorname{Re}\mathbf{z}$ and $\operatorname{Im}\mathbf{z}$ are independent and identically distributed.

A.2 Complex Gaussian Random Vectors

Most of this thesis considers Gaussian random vectors. We can define these in generality, for $\mathbf{z} : \Omega \to \mathbb{C}^n$

$$p(\mathbf{z}) = \frac{1}{\pi^n \sqrt{\det \underline{\boldsymbol{\Sigma}}}} e^{-1/2(\underline{\mathbf{z}}-\underline{\boldsymbol{\mu}})^* \underline{\boldsymbol{\Sigma}}^{-1} (\underline{\mathbf{z}}-\underline{\boldsymbol{\mu}})} \tag{A.5}$$

has mean $\boldsymbol{\mu}$ and augmented covariance $\underline{\boldsymbol{\Sigma}}$. Notice $\underline{\boldsymbol{\Sigma}}$ can still encode a nonzero covariance between $\mathbf{z}$ and $\overline{\mathbf{z}}$.

Whereas a *proper* (or circular) complex Gaussian random vector, $\mathbf{z}$ and $\overline{\mathbf{z}}$ are uncorrelated. This makes the probabilty density function a function of only $\mathbf{z}$;

$$p(\mathbf{z}) = \frac{1}{\pi^n \det \Sigma} e^{-(\mathbf{z}-\boldsymbol{\mu})^* \boldsymbol{\Sigma}^{-1} (\mathbf{z}-\boldsymbol{\mu})}. \tag{A.6}$$

This convient formulation is related to the inner-product on $\mathbb{C}^n$. We can define the inner product $\langle \mathbf{u}, \mathbf{v} \rangle_{\boldsymbol{\Sigma}} = \mathbf{u}^* \boldsymbol{\Sigma}^{-1} \mathbf{v}$ with corresponding norm $\|\mathbf{u}\|^2_{\boldsymbol{\Sigma}} = \mathbf{u}^* \boldsymbol{\Sigma}^{-1} \mathbf{u}$, allowing the probability density function to be rewritten

$$p(\mathbf{z}) = \frac{1}{\pi^n \det(\boldsymbol{\Gamma})} e^{-\|\mathbf{z}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Gamma}}}. \tag{A.7}$$

Notationally, we say $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma})$.

For the majority of this thesis, we will consider noise that is a proper Gaussian random vector $\mathbf{g} : \Omega \to \mathbb{C}^n$ with covariance $\boldsymbol{\Sigma} = \sigma \mathbf{I}$ resulting in the probability density function

$$p(\mathbf{g}) = \frac{1}{\pi^n \sigma^n} e^{-\|\mathbf{g}\|^2_2 / \sigma}. \tag{A.8}$$

A.3 Estimators

Suppose we measure $\widetilde{\mathbf{y}} \in \mathbb{C}^n$ that combines of a vector that depends on a set of parameters $\widehat{\boldsymbol{\theta}} \in \mathbb{C}^p$ through $\mathbf{f} : \mathbb{C}^p \to \mathbb{C}^n$ and random vector $\mathbf{g}$;

$$\widetilde{\mathbf{y}} = \mathbf{y} + \mathbf{g} = \mathbf{f}(\widehat{\boldsymbol{\theta}}) + \mathbf{g}. \tag{A.9}$$

(Here we denote vectors perturbed by random noise with hats, e.g., $\widetilde{\mathbf{y}}$). Our goal is to build function, termed an *estimator* in statistical parlance, that takes $\widetilde{\mathbf{y}}$ and returns an estimate of $\widehat{\boldsymbol{\theta}}$. We call this function $\mathbf{t}(\widetilde{\mathbf{y}}) = \widetilde{\boldsymbol{\theta}}$.

One way to construct an estimator is to choose $\mathbf{t}$ that maximizes the likelihood of $\widetilde{\boldsymbol{\theta}}$ explains the data. If $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})$, the likelihood function is:

$$\mathcal{L}(\boldsymbol{\theta}) := p(\widetilde{\mathbf{y}}|\boldsymbol{\theta}) = p(\widetilde{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta})) = \frac{1}{\pi^n \det(\boldsymbol{\Gamma})} e^{-\|\widetilde{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta})\|_{\boldsymbol{\Gamma}}^2}. \tag{A.10}$$

The maximum likelihood estimator chooses $\widetilde{\boldsymbol{\theta}}$ maximizing $\mathcal{L}$. After taking the logarithm of (A.10), maximizing $\mathcal{L}$ is equivalent to minimizing the norm

$$\mathbf{t}(\widetilde{\mathbf{y}}) := \operatorname*{argmin}_{\boldsymbol{\theta}} \|\widetilde{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta})\|_{\boldsymbol{\Gamma}}^2 = \widetilde{\boldsymbol{\theta}}. \tag{A.11}$$

We have arrived at a weighted least squares problem.

Often times we have that $\mathbf{g} : \Omega \to \mathbb{R}^n$ rather than $\mathbb{C}^n$. In this case, we still minimize the same norm, but the normalization on the probability density function p is different. Observe

Other approaches for exponential fitting qualify as estimators for $\widehat{\boldsymbol{\theta}}$. The key question we to ask about these algorithms is do they exhibit a bias, e.g., $\mathsf{E}[\mathbf{t}(\widetilde{\mathbf{y}})] \neq \widehat{\boldsymbol{\theta}}$? Do they have minimum covariance? If so, $\mathbf{t}$ is called a *minimum variance unbiased estimator* (MVUE) [57, Def. 1.3.3]. As almost all examples are nonlinear functions of $\widetilde{\mathbf{y}}$, it is difficult to compute the expected value

$$\mathsf{E}[\widetilde{\boldsymbol{\theta}}] = \int_{\mathbb{C}^n} \mathbf{t}(\mathbf{y} + \mathbf{g}) p(\mathbf{g}) \, \mathrm{d}\mathbf{g}$$

and the covariance

$$\mathsf{Cov}[\widetilde{\boldsymbol{\theta}}] = \int_{\mathbb{C}^n} \left[\mathbf{t}(\mathbf{y} + \mathbf{g}) - \widehat{\boldsymbol{\theta}}\right] \left[\mathbf{t}(\mathbf{y} + \mathbf{g}) - \widehat{\boldsymbol{\theta}}\right]^* p(\mathbf{g}) \, \mathrm{d}\mathbf{g}$$

in closed form. Instead we satify ourselves with two approaches: estimating the expected value and covariance through many random trials and considering the asymptotic limit of small noise.

A.4 Asymptotics

In the limit of small noise, we make the assumption that $\mathbf{g}$ is small, so we can approximate $\mathbf{t}$ by its Taylor expansion about $\mathbf{y}$ to first order

$$\mathbf{t}(\mathbf{y}+\mathbf{g}) = \mathbf{t}(\mathbf{y}) + \underline{\mathbf{T}\mathbf{g}} + \mathcal{O}(\|\mathbf{g}\|^2) = \mathbf{t}(\mathbf{y}) + \mathbf{T}\mathbf{g} + \mathbf{T}'\overline{\mathbf{g}} + \mathcal{O}(\|\mathbf{g}\|^2).$$

For maximum likelihood estimators of $\widehat{\boldsymbol{\theta}}$, we have $\mathbf{t}(\mathbf{y}) = \widehat{\boldsymbol{\theta}}$ and $\mathbf{T} = \mathbf{F}(\widehat{\boldsymbol{\theta}})^+$ and $[\mathbf{F}(\boldsymbol{\theta})]_{j,k} = \partial f_k(\boldsymbol{\theta})/\partial\theta_j$ and $\mathbf{T}' = \mathbf{0}$ if $\mathbf{f}$ is analytic. Similar $\mathbf{T}' = \mathbf{0}$ for Prony type methods as roots of a polynomial are analytic functions of their parameters if their are no repeated roots.[1]

Using the Taylor expansion (provided $\mathbf{t}$ is analytic), we can evaluate these integrals in the uniform limit of $\sigma \to 0$ where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$. This is equivalent to simply scaling $\mathbf{g} \to \sigma\mathbf{g}$ keeping the same probability density function p as before (i.e., without σ). Hence, evaluating the integrals for expectation

$$\lim_{\sigma\to 0} \mathsf{E}[\widetilde{\boldsymbol{\theta}}] = \lim_{\sigma\to 0} \int_{\mathbb{C}^n} (\mathbf{t}(\mathbf{y}) + \sigma\mathbf{T}\mathbf{g}) p(\mathbf{g})\, \mathrm{d}\mathbf{g} = \mathbf{t}(\mathbf{y}).$$

Similarly for the covariance, if $\mathbf{t}(\mathbf{y}) = \widehat{\boldsymbol{\theta}}$,

$$\lim_{\sigma\to 0} \mathsf{Cov}[\widetilde{\boldsymbol{\theta}}] = \lim_{\sigma\to 0} \int_{\mathbb{C}^n} \sigma^2 \mathbf{T}\mathbf{g}\mathbf{g}^*\mathbf{T} p(\mathbf{g})\, \mathrm{d}\mathbf{g} = \lim_{\sigma\to 0} \sigma^2 \mathbf{T}\boldsymbol{\Sigma}\mathbf{T}^*.$$

Statistician's often work in the limit of a large number of measurments $n \to \infty$ allowing them to invoke the Central Limit Theorem and there by prove their estimators are 'asymptotically optimal.' With exponential fitting, this limit is not appropreate. If $\operatorname{Re}\omega_j < 0$, as $n \to \infty$ we recover less and less information about ω_j. We could be careful, keeping $n\boldsymbol{\omega}$ constant (essentially increasing the sampling rate of our measurements). Instead, throughout this paper we work in the limit of small

[1] through Kato; Knopp II, ch.5

noise $\|\mathbf{\Gamma}\| \to 0$ and instead invoke perturbation results to establish that in the limit of small noise, our estimators are asymptotically optimial.

A.5 Fisher Information and the Cramér-Rao Bound

Key to most of this thesis is the question: how does our algorithm for estimating $\boldsymbol{\omega}$ perform as a function of noise in the measurements $\mathbf{y}$. This is different than the (equally important) question of numerical stability.

$$\mathcal{I}(\boldsymbol{\theta}) = \mathsf{E}_{\mathbf{g}}\left\{\mathbf{s}(\boldsymbol{\theta},\mathbf{g})\mathbf{s}(\boldsymbol{\theta},\mathbf{g})^*\right\} \qquad [\mathbf{s}(\boldsymbol{\theta},\mathbf{g})]_j = \overline{\frac{\partial}{\partial\theta_j}\log p(\mathbf{g};\boldsymbol{\theta})}. \tag{A.12}$$

[137, eq. 6.49]

In the case where we have additive noise, and $\mathbf{f}$ is analytic with first derivative $\mathbf{F}$, then

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbf{F}(\boldsymbol{\theta})^*\mathbf{\Sigma}^{-1}\mathbf{F}(\boldsymbol{\theta}). \tag{A.13}$$

The Cramér-Rao Bound states

$$\mathsf{Cov}[\widetilde{\boldsymbol{\theta}}] \geq \mathcal{I}(\boldsymbol{\theta})^{-1} \tag{A.14}$$

for any unbiased estimator [137, eq. 6.51]

A.6 Principle of Invariance

Principle of Invariance [167]. This yields two important results in this thesis. Maximum likelihood estiamtes of exponential parameters yield maximum likelihood estimates of parameters derived from the exponential parameters. This means that rather than building specialized parameter estimators, we can use an exponential fitting algorithm followed by parameter recovery.

Second, that when Prony's method recovers polynomial coefficients whose roots are the expoential parameters if we recover the coefficients in a maximum likelihood sense then the exponential parameters are also recovered in a maximum likelihood sense (at least asymptoticially).

Appendix B

Wirtinger Derivatives and Complex Optimization

At several points in this thesis we compute gradients with respect to real functions of complex variables. Some concern should be present because these functions are not analytic.

The basic idea is we parameterize with respect to two derivatives:

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial \mathrm{Re}\, z} - i\frac{\partial}{\partial \mathrm{Im}\, z}\right) \quad \text{and} \quad \frac{\partial}{\partial \overline{z}} = \frac{1}{2}\left(\frac{\partial}{\partial \mathrm{Re}\, z} + i\frac{\partial}{\partial \mathrm{Im}\, z}\right). \tag{B.1}$$

When a function is analytic, $\frac{\partial f}{\partial z}$ corresponds to the regular complex derivative and $\frac{\partial f}{\partial \overline{z}} = 0$ since f satifies the Cauchy-Riemann equations.

B.1 Least Squares

Consider the problem

$$\min_{\mathbf{x}\in\mathbb{C}^p} f(\mathbf{x}) \qquad f(\mathbf{x}) := \mathbf{r}(\mathbf{x})^*\mathbf{r}(\mathbf{x}), \quad \mathbf{r} : \mathbb{C}^p \to \mathbb{C}^n, \text{ analytic} \tag{B.2}$$

This is the typical least squares problem. As f is not analytic, we regard it as a function of both $\mathbf{x}$ and $\overline{\mathbf{x}}$. Taking the Taylor expansion of f and applying the Chain Rule, we have

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \begin{bmatrix} \dfrac{\partial f}{\partial \mathbf{x}}(\mathbf{x}_0) & \dfrac{\partial f}{\partial \overline{\mathbf{x}}}(\mathbf{x}_0) \end{bmatrix} \begin{bmatrix} \mathbf{x} - \mathbf{x}_0 \\ \overline{\mathbf{x}} - \overline{\mathbf{x}_0} \end{bmatrix} + \mathcal{O}(\|\mathbf{x} - \mathbf{x}_0\|^2).$$

Applying the Chain Rule again,

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{x}} &= \frac{\partial f}{\partial \mathbf{r}}\frac{\partial \mathbf{r}}{\partial \mathbf{x}} + \frac{\partial f}{\partial \overline{\mathbf{r}}}\frac{\partial \overline{\mathbf{r}}}{\partial \mathbf{x}};\\ \frac{\partial f}{\partial \overline{\mathbf{x}}} &= \frac{\partial f}{\partial \mathbf{r}}\frac{\partial \mathbf{r}}{\partial \overline{\mathbf{x}}} + \frac{\partial f}{\partial \overline{\mathbf{r}}}\frac{\partial \overline{\mathbf{r}}}{\partial \overline{\mathbf{x}}}.\end{aligned}$$

Consulting [137, §A2.2] we have

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{r}} &= \mathbf{r}^*, & \frac{\partial \mathbf{r}}{\partial \mathbf{x}} &= \overline{\frac{\partial \overline{\mathbf{r}}}{\partial \overline{\mathbf{x}}}}\\ \frac{\partial f}{\partial \overline{\mathbf{r}}} &= \mathbf{r}^\top, & \frac{\partial \overline{\mathbf{r}}}{\partial \mathbf{x}} &= \overline{\left(\frac{\partial \mathbf{r}}{\partial \overline{\mathbf{x}}}\right)} = \mathbf{0},\end{aligned}$$

and $[\frac{\partial \mathbf{r}}{\partial \mathbf{x}}]_{j,k} = \frac{\partial r_k}{\partial x_j}$. Hence

$$\begin{aligned}f(\mathbf{x}) &= f(\mathbf{x}_0) + \mathbf{r}^*(\mathbf{x}_0)\frac{\partial \mathbf{r}}{\partial \mathbf{x}}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) + \mathbf{r}^\top(\mathbf{x}_0)\overline{\left(\frac{\partial \mathbf{r}}{\partial \mathbf{x}}(\mathbf{x}_0)\right)(\mathbf{x}-\mathbf{x}_0)} + \mathcal{O}(\|\mathbf{x}-\mathbf{x}_0\|^2)\\ &= f(\mathbf{x}_0) + 2\mathrm{Re}\left[\mathbf{r}^*(\mathbf{x}_0)\frac{\partial \mathbf{r}}{\partial \mathbf{x}}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)\right] + \mathcal{O}(\|\mathbf{x}-\mathbf{x}_0\|^2).\end{aligned}$$

From here, we note as [137] does, that to satisfy the first order optimiality conditions we can ignore the real component and simply find $\mathbf{x}$ solving

$$\mathbf{r}^*(\mathbf{x})\frac{\partial \mathbf{r}}{\partial \mathbf{x}}(\mathbf{x}) = \mathbf{r}^*(\mathbf{x})\mathbf{J}(\mathbf{x}) = \mathbf{0}. \tag{B.3}$$

We can go one step further and write the next term in the Taylor expansion as

well

$$
\begin{aligned}
f(\mathbf{x}) =& f(\mathbf{x}_0) + \mathbf{r}^*(\mathbf{x}_0)\mathbf{J}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) + \mathbf{r}^\top(\mathbf{x}_0)\overline{\mathbf{J}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)} \\
&+ \sum_k \mathbf{r}(\mathbf{x})^* \frac{\partial \mathbf{J}}{\partial x_k}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)(x_k - x_{0k}) + (\mathbf{x}-\mathbf{x}_0)^\top \mathbf{J}(\mathbf{x}_0)^\top \overline{\mathbf{J}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)} \\
&+ \sum_k \mathbf{r}(\mathbf{x})^\top \frac{\partial \mathbf{J}}{\partial \overline{x_k}}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)(\overline{x_k - x_{0k}}) + (\mathbf{x}-\mathbf{x}_0)^* \mathbf{J}(\mathbf{x}_0)^* \mathbf{J}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) \\
&+ \mathcal{O}(\|\mathbf{x}-\mathbf{x}_0\|^3) \\
=& f(\mathbf{x}_0) + 2\mathrm{Re}\ [\mathbf{r}(\mathbf{x}_0)^* \mathbf{J}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)] \\
&+ \mathrm{Re}\ \left[\sum_k \mathbf{r}(\mathbf{x})^* \frac{\partial \mathbf{J}}{\partial x_k}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)(x_k - x_{0k}) + (\mathbf{x}-\mathbf{x}_0)^* \mathbf{J}(\mathbf{x}_0)^* \mathbf{J}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) \right] \\
&+ \mathcal{O}(\|\mathbf{x}-\mathbf{x}_0\|^3)
\end{aligned}
$$

From here, we can still apply our typical Hessian approximation strategy, neglecting the second derivative of f.

Bibliography

[1] Elias Aboutanios. Estimation of the frequency and decay factor of a decaying exponential in noise. *IEEE Transactions on Signal Processing*, 58(2):501–509, February 2010.

[2] Elias Aboutanios and Bernard Mulgrew. Iterative frequency estimation by interpolation on Fourier coefficients. *IEEE Transactions on Signal Processing*, 53(4):1237–1242, April 2005.

[3] Moritz Allmaras, Wolfgang Bangerth, Jean Marie Linhard, Javier Polanco, Fang Wang, Kainan Wang, Jennifer Webster, and Sarah Zedler. Estimating parameters in physical models through Bayesian inversion: A complete example. *SIAM Review*, 55(1):149–167, 2013.

[4] G. Ammar, W. Dayawansa, and C. Martin. Exponential interpolation: Theory and numerical algorithms. *Applied Mathematics and Computation*, 41:189–232, 1991.

[5] B. D. O. Anderson and A. C. Antoulas. Rational interpolation and state-variable realizations. *Linear Algebra and its Applications*, 138:479–509, 1990.

[6] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999.

[7] S. S. Antman. *Nonlinear Problems in Elasticity*. Springer, New York, second edition, 2005.

[8] Aleksandr Aravkin, Michael P. Friedlander, Felix J. Hermann, and Tristan van Leeuwen. Robust inversion, dimensionality reduction, and randomized sampling. *Mathematical Programming*, 134(1):101–125, 2012.

[9] Jon R. Auton and Michael L. Van Blaricum. Investigation of procedures for automatic resonance extraction from noisy transient electromagnetics data. Technical Report 79, General Research Corporation, Santa Barbara, CA, 1981.

[10] Arthur J. Barabell. Improving the resolution performance of eigenstructure-based direction-finding algorithms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 8, pages 336–339, 1983.

[11] H. Barkhuijsen, R. de Beer, W. M. M. J. Bovée, and D. van Ormondt. Retrieval of frequencies, amplitudes, damping factors, and phases from time-domain signals using a linear least squares procedure. *Journal of Magnetic Resonance*, 61:465–481, 1985.

[12] H. Barkhuijsen, R. de Beer, and D. van Ormondt. Improved algorithm for noniterative time-domain model fitting to exponentially damped magnetic resonance signals. *Journal of Magnetic Resonsance*, 73:553–557, 1987.

[13] G. K. Batchelor. *An Introduction to Fluid Dynamics*. Cambridge University Press, Cambridge, 1967.

[14] Dmitry Batenkov and Yosef Yomdin. On the accuracy of solving confluent Prony systems. *SIAM Journal on Matrix Analysis and Applications*, 73(1):134–154, 2013.

[15] Emanuele Berti, Vitor Cardoso, José A. González, and Ulrich Sperhake. Mining information from binary black hole mergers: A comparison of estimation methods for complex exponentials in noise. *Physical Review D*, 75, 2007.

[16] Matteo Bertocco, Carlo Offelli, and Dario Petri. Analysis of damped sinusodial signals via a frequency-domain interpolation algorithm. *IEEE Transactions on Instrumentation and Measurement*, 43(2):245–250, 1994.

[17] Philip R. Bevington and D. Keith Robinson. *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill, Boston, 2003.

[18] Gregory Beylkin and Lucas Monzón. On approximation of functions by exponential sums. *Applications of Computational Harmonic Analysis*, 19:17–48, 2005.

[19] LH Bezerra and FSV Bazan. Eigenvalue locations of generalized companion predictor matrices. *SIAM Journal on Matrix Analysis and Applications*, 19(4):886–897, October 1998.

[20] Åke Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, PA, 1996.

[21] Albrecht Böttcher and Bernd Silbermann. *Introduction to Large Truncated Toeplitz Matrices*. Springer, New York, 1998.

[22] Gregory Boutry, Michael Elad, Gene H. Golub, and Peyman Milanfar. The generalized eigenvalue problem for nonsquare pencils using a minimal perturbation approach. *SIAM Journal on Matrix Analysis and Applications*, 27(2):582–601, 2005.

[23] Yoram Bresler and Albert Macovski. Exact maximum likelihood parameter estimation of superimposed exponential signals in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(5):1081–1089, 1986.

[24] G. Larry Bretthorst, William C. Hutton, Joel R. Garbow, and Joseph J. H. Ackerman. Exponential parameter estimation (in NMR) using Bayesian probability theory. *Concepts in Magnetic Resonance Part A*, 27(2):55–63, 2005.

[25] James R. Bunch, Christopher P. Nielsen, and Danny C. Sorensen. Rank-one modification of the symmetric eigenproblem. *Numericshe Mathematik*, 31:31–48, 1978.

[26] Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference*. Springer, New York, 2002.

[27] Stephen L. Campbell and C. D. Meyer, Jr. *Generalized Inverses of Linear Transformations*. Pitman, Boston, 1979.

[28] David G. Cantor and John W. Evans. On approximation by positive sums of powers. *SIAM Journal on Applied Mathematics*, 18(2):380–388, 1970.

[29] Hua Chen, Sabine Van Huffel, and Joos Vandewalle. Bandpass prefiltering for exponential data fitting with known frequency region of interest. *Signal Processing*, 48:135–154, 1996.

[30] Moody T. Chu and Gene H. Golub. *Inverse Eigenvalue Problems: Theory, Algorithms, and Applications*. Oxford University Press, Oxford, 2005.

[31] Thomas F. Coleman and Yuying Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6(2):418–445, May 1996.

[32] A. R. Collar and A. Simpson. *Matrices and Engineering Dynamics*. Ellis Horwood Limited, Chichester, West Sussex, England, 1987.

[33] R. Cornell. A method for fitting linear combinations of exponentials. *Biometrics*, 18:104–113, 1962.

[34] R. G. Cornell. A new estimation procedure for linear combinations of exponentials. Technical Report ORNL 2120, Oak Ridge National Laboratory, 1956.

[35] R. Courant and D. Hilbert. *Methods of Mathematical Physics*, volume 1. Interscience, New York, 1953.

[36] Steven Cox and Enrique Zuazua. The rate at which energy decays in a damped string. *Communications in Partial Differential Equations*, 19:212–243, 1994.

[37] Steven J. Cox and Mark Embree. Reconstructing an even damping from a single spectrum. *Inverse Problems*, 27:1–18, 2011.

[38] Steven J. Cox, Mark Embree, and Jeffrey M. Hokanson. One can hear the composition of a string: experiments with an inverse eigenvalue problem. *SIAM Review*, 54(1):157–178, 2012.

[39] P. de Groen and B. de Moor. The fit of a sum of exponentials to noisy data. *Journal of Computational and Applied Mathematics*, 20:175–187, 1987.

[40] J. Dennis, Jr. and J. Moré. Quasi-Newton methods, motivation and theory. *SIAM Review*, 19(1):46–89, 1977.

[41] Joseph J. DiStefano, III and Elliot M. Landaw. Multiexponential, multicompartmental, and noncompartmental modeling. I. methodological limitations and physiological interpretations. *American Journal of Physiology*, 246:R651–R664, 1984.

[42] Stan Dodds. Standing waves. Unpublished lab manual for Physics 231, Rice University, 2013.

[43] D. G. Dudly. Fitting noisy data with a complex exponential series. Technical report, Lawrence Livermore Laboratory, 1977.

[44] Michael Elad, Peyman Milanfar, and Gene H. Golub. Shape from moments – an estimation theory prespective. *IEEE Transactions on Signal Processing*, 52(7):1814–1829, 2004.

[45] Charles L. Epstein and John Schotland. The bad truth about Laplace's transform. *SIAM Review*, 50(3):504–520, 2008.

[46] D. J. Ewins. *Modal Testing: Theory and Practice*. Research Studies Press LTD., Letchworth, Hertfordshire, England, 1984.

[47] Richard William Farebrother. *Fitting Linear Relationships: A History of the Calculus of Observations 1750-1900*. Springer, New York, 1999.

[48] Marie Farge. Wavelet transforms and their applications to turbulence. *Annual Review of Fluid Mechanics*, 24:395–458, 1992.

[49] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, 222:309–368, 1922.

[50] Michael P. Friedlander and Mark Schmidt. Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing*, 34(3):A1380–A1405, 2012.

[51] D. Gale and L. S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

[52] F. R. Gantmacher and M. G. Krein. *Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems*. AMS Chelsea, Providence, RI, revised edition, 2002.

[53] Donald G. Gardner. Resolution of multi-component exponential decay curves using Fourier transforms. *Annals of the New York Academy of Sciences*, 108:195–203, 1963.

[54] Donald G. Gardner, Jeanne C. Gardner, George Laush, and W. Wayne Meinke. Method for the analysis of multicomponent exponential decay curves. *The Journal of Chemical Physics*, 31(4):978–986, 1959.

[55] Gene H. Golub and Victor Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10(2):413–432, April 1973.

[56] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[57] Graham C. Goodwin and Robert L. Payne. *Dynamic System Identification: Experiment Design and Data Analysis*. Academic Press, New York, 1977.

[58] William B. Gragg and Anders Lindquist. On the partial realization problem. *Linear Algebra and its Applications*, 50:277–319, 1983.

[59] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics: a foundation for computer science*. Addison-Wesley, Reading, Massachusetts, 1989.

[60] Ulf Grenander and Gabor Szegö. *Toepliz forms and their applications*. University of California Press, Berkeley, 1958.

[61] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[62] Roger J. Hanson, James M. Anderson, and H. Kent Macomber. Measurements of nonlinear effects in a driven vibrating wire. *Journal of the Acoustical Society of America*, 96(3):1549–1556, 1994.

[63] Umer Hassan, Zubair Usman, and Muhammad Sabieh Anwar. Video-based spatial portraits of a nonlinear vibrating string. *American Journal of Physics*, 80(10):862–869, 20012.

[64] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, 2002.

[65] F. B. Hildebrand. *Introduction to Numerical Analysis*. McGraw-Hill, New York, 1956.

[66] Kenneth Holmström and Jöran Petersson. A review of the parameter estimation problem of fitting positive exponential sums to empirical data. *Applied mathematics and computation*, 126:31–61, 2002.

[67] J. Holtsmark, I. Johnsen, T. Sikkeland, and S. Skavlem. Boundary layer flow near a cylindrical obstacle in an oscillating, incompressible fluid. *Journal of the Acoustical Society of America*, 26(1):26–39, January 1954.

[68] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

[69] A. S. Householder. On Prony's method of fitting exponential decay curves and multiple-hit survival curves. Technical Report ORNL 445, Oak Ridge National Laboratory, 1949.

[70] V. E. Howle and Lloyd N. Trefethen. Eigenvalues and musical instruments. *Journal of Computational and Applied Mathematics*, 135:23–40, 2001.

[71] Yingbo Hua. *On Techniques for Estimating Parameters of Exponentially Damped/Undamped Sinusoids in Noise*. PhD thesis, Syracuse University, 1988.

[72] Yingbo Hua and Tapan K. Sarkar. Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(5):814–824, 1990.

[73] Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H. Lui. The empirical mode decomposition and the Hilbert spectrum for nonlinear non-stationary time series analysis. *Proceedings of the Royal Society, Series A*, 454:903–995, 1998.

[74] G. E. Hudson. Trend analysis of physical data. *American Journal of Physics*, 21(5):362–367, May 1953.

[75] Sabine Van Huffel, Hua Chen, Caroline Decanniere, and Paul Van Hecke. Algorithm for time-domain NMR data fitting based on total least squares. *Journal of Magnetic Resonance, Series A*, 110:228–237, 1994.

[76] Irvin Isenberg. Robust estimation in pulse fluorometry: A study of the method of moments and least squares. *Biophysical Journal*, 43:141–148, August 1983.

[77] Irvin Isenberg and Robert D. Dyson. The analysis of fluorescence decay by a method of moments. *Biophysical Journal*, 9:1337–1350, 1969.

[78] Irvin Isenberg, Robert D. Dyson, and Richard Hanson. Studies on the analysis of fluorescence decay data by the method of moments. *Biophysical Journal*, 13, 1973.

[79] Andrei A. Istratov and Oleg F. Vyvenko. Exponential analysis in physical phenomena. *Rev. Sci. Instrum.*, 70:1233–1257, 1999.

[80] Sachin K. Jain and S. N. Singh. Harmonics estimation in emerging power system: key issues and challenges. *Electric Power Systems Research*, 81:1754–1766, 2011.

[81] Vijay K. Jain, William L. Collins, Jr., and David C. Davis. High-accuracy analog measurements via interpolated FFT. *IEEE Transactions on Instrumentation and Measurement*, IM-28(2):113–122, June 1979.

[82] M. H. Kahn, M. S. Mackisack, M. R. Osborne, and G. K. Smyth. On the consistency of Prony's method and related algorithms. *Journal of Computational and Graphical Statistics*, 1(4):329–349, 1992.

[83] Thomas Kailath. *Linear Systems*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1980.

[84] David W. Kammler and Robert J. McGlinn. A bibliography for approximation with exponential sums. *Journal of Computational and Applied Mathematics*, 4(2):167 – 173, 1978.

[85] Tohru Katayama. *Subspace Methods for System Identification*. Springer, London, 2005.

[86] Linda Kaufman. A variable projection method for solving separable nonlinear least squares problems. *BIT Numerical Mathematics*, 15:49–57, 1975. 10.1007/BF01932995.

[87] Linda Kaufman, Garrett S. Sylvester, and Margaret H. Wright. Structured linear least-squares problems in system identification and separable nonlinear data fitting. *SIAM Journal on Optimization*, (4):847–871, 1994.

[88] G. R. Keepin, T. F. Wimett, and R. K. Zeigler. Delayed neutrons from fissionable isotopes of Uranium, Plutonium and Thorium. *Physical Review*, 107(4):1044–1049, August 1957.

[89] Luc Knockaert. A simple and accurate algorithm for barycentric rational interpolation. *IEEE Signal Processing Letters*, 15:154–157, 2008.

[90] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51:455–500, August 2009.

[91] Ramdas Kumaresan. On the zeros of the linear prediction-error filter for deterministic signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 31(1):217–220, 1983.

[92] Ramdas Kumaresan and Donald W. Tufts. Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 30(6):833–840, December 1982.

[93] Ramdas Kumaresan and Donald W. Tufts. Estimating the angles of arrival of multiple plane waves. *IEEE Transactions on Aerospace and Electronic Systems*, 19(1), 1983.

[94] S. Y. Kung, K. S. Arun, and Bhaskar D. V. Rao. State-space and singular-value decomposition-based approximation methods for the harmonic retrieval problem. *Journal of the Optical Society of America*, 83(12):1799–1811, 1983.

[95] Cornelius Lanczos. Prentice-Hall, Englewood Cliffs, NY.

[96] Teresa Laudadio. *Subspace-based quantification of magnetic resonance spectroscopy data using biochemical prior knowledge.* PhD thesis, Katholieke Universiteit Leuven, 2005.

[97] Sanda Lefteriu and Athanasios C. Antoulas. A new approach to modeling multiport systems from frequency-domain data. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 29(1):14–27, 2010.

[98] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK User's Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods.* SIAM, Philadelphia, 1998.

[99] Leonard Lewin. *Polylogarithms and associated functions.* North-Holland Publishing Co., New York, 1981.

[100] Hai-Ping Lin. Direct and inverse methods on free vibration analysis of simply supported beams with a crack. *Engineering Structures*, 26:427–436, 2004.

[101] S. P. Lin. Damped vibration of a string. *Journal of Fluid Mechanics*, 72(4):787–797, 1975.

[102] N. M. M. Maia and J. M. M. Silva. Modal analysis identification techniques. *Philosophical Transactions of the Royal Society A*, 359:29–40, 2001.

[103] V. A. Mandelshtam. FDM: the filter diagonalization method for data processing in NMR experiments. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 38:159–196, 2001.

[104] S. Lawrence Marple, Jr. *Digital Spectral Analysis with Applications*. Prentice-Hall, Englewood Cliffs, New Jersey, 1987.

[105] L. W. McKeen. *The Effect of Creep and Other Time Related Factors on Plastics and Elastomers*. Elsevier Science, Oxford, 2009.

[106] Shahab Minhas and Elias Aboutanios. Estimation of the frequency of a complex exponential. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3693–3696, 2008.

[107] Shahab Faiz Minhas and Patrick Gaydecki. A non-iterative approach to frequency estimation of complex exponential in noise by interpolation of fourier coefficients. In *18th International Conference on Digital Signal Processing*, pages 1–6, 2013.

[108] Sharmishtha Mitra and Amit Mitra. A genetic algorithms based technique for computing the nonlinear least squares estimates of the parameters of sum of exponentials model. *Expert Systems with Applications*, 39:6370–6379, 2012.

[109] P. Mohanty and D. J. Rixen. A modified Ibrahim time domain algorithm for operational modal analysis including harmonic excitation. *Journal of Sound and Vibration*, 275(1-2):375 – 390, 2004.

[110] Timothy C. Molteno and Nicholas B. Tufillaro. An experimental investigation into the dynamics of a string. *American Journal of Physics*, 72(9):1157–1169, 2004.

[111] National Institute of Standards and Technology. *Digital Library of Mathematical Functions*, May 2010.

[112] Yu. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–262, 2012.

[113] Michael K. Ng. *Iterative Methods for Toeplitz Systems*. Oxford University Press, New York, 2004.

[114] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, 2006.

[115] James M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York, 1970.

[116] M. R. Osborne. Some special nonlinear least squares problems. *SIAM Journal on Numerical Analysis*, 12(4):571–592, 1975.

[117] M. R. Osborne and G. K. Smyth. A modified Prony algorithm for fitting functions defined by difference equations. *SIAM Journal on Scientific and Statistical Computing*, 12(2):362–382, March 1991.

[118] H. Ouibrahim. Prony, Pisarenko, and the matrix pencil: a unified presentation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(1):133 –134, January 1989.

[119] Lewis Pakula. Asymptotic zero distribution of orthogonal polynomials in sinusoidal frequency estimation. *IEEE Transactions on Information Theory*, 33(4):569–576, 1987.

[120] Haesun Park, Sabine Van Huffel, and Lars Eldén. Fast algorithms for exponential data modeling. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 25–28, 1994.

[121] D. H. Parsons. Biological problems involving sums of exponential functions of time: A mathematical analysis that reduces experimental time. *Mathematical Biosciences*, 2:123–128, 1968.

[122] W Perl. A method for curve-fitting by exponential functions. *International Journal of Applied Radiation and Isotopes*, 8:211–222, 1960.

[123] Thomas Peter, Daniel Potts, and Manfred Tasche. Nonlinear approximation by sums of exponentials and translates. *SIAM Journal on Scientific Computing*, 33(4):1920–1947, 2011.

[124] Paul E. Pfeiffer. *Probability for Applications*. Springer-Verlag, New York, 1990.

[125] V. F. Pisarenko. The retrieval of harmonics from a covariance function. *Geophysical Journal of the Royal Astronomical Society*, 33(3):347–366, 1973.

[126] Daniel Potts and Manfred Tasche. Parameter estimation for exponential sums by approximate Prony method. *Signal Processing*, 90:1631–1642, 2010.

[127] R. Prony. Essai expérimental et analytique sur les lois de la dilatabilité et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, à différentes températures. *J. de l'Ecole Polytechnique*, 1:24–76, 1795. English translation in [151, App. A].

[128] Barry G. Quinn. Estimating frequency by interpolation using Fourier coefficients. *IEEE Transactions on Signal Processing*, 42(5):1264–1268, May 1994.

[129] Bhaskar D. Rao. Perturbation analysis of an SVD-based linear prediction method for estimating the frequencies of multiple sinusoids. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7):1026–1035, 1988.

[130] S. O. Rice. Statistical properties of random noise currents. *Bell System Technical Journal*, 24:46–156, 1945.

[131] D. C. Rife and G. A. Vincent. Use of the discrete Fourier transform in the measurement of frequencies and levels of tones. *The Bell System Technical Journal*, pages 197–228, February 1970.

[132] Severine Rosat, Toshio Fukushima, Tadahiro Sato, and Yoshiaki Tamura. Application of a non-linear damped harmonic analysis method to the normal modes of the Earth. *Journal of Geodynamics*, 45:63–71, 2008.

[133] H. E. Rosinger and I. G. Ritchie. On Timoshenko's correction for shear in vibrating isotropic beams. *Journal of Physics D: Applied Physics*, 10, 1977.

[134] Richard Roy and Thomas Kailath. ESPRIT–estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7):984–995, 1989.

[135] Axel Ruhe and Per Ake Wedin. Algorithms for separable nonlinear least squares problems. *SIAM Review*, 22(3):318–337, 1980.

[136] Ralph Schmidt. Multiple emitter location and singal parameter estimation. In *Proceedings of the RADC Spectrum Estimation Workshop*, pages 243–258, 1979.

[137] Peter J. Schreier and Louis L. Scharf. *Statistical Signal Processing of Complex-Valued Data: Theory of Improper and Noncircular Signals*. Cambridge University Press, Cambridge, 2010.

[138] H. D. Scolnik. *On the solution of nonlinear least squares problems*. PhD thesis, University of Zurich, 1970.

[139] William Simon. A method of exponential separation applicable to small computers. *Physics in Medicine and Biology*, 15(2):355–360, 1970.

[140] M. W. Simpson-Morgan. A note on fitting multiexponential functions of time to experimental data. *Mathematical Biosciences*, 5:195–199, 1969.

[141] D. C. Sorensen. Newton's method with a model trust region modification. *SIAM Journal on Optimization*, 19(2):409–426, 1982.

[142] D. C. Sorensen. Implicit application of polynomial filters in a k-step Arnoldi method. *SIAM Journal on Matrix Analysis and Applications*, 13(1):357–385, 1992.

[143] G. W. Stewart and Ji Guang Sun. *Matrix Perturbation Theory*. Academic Press, Boston, 1990.

[144] Petre Stoica and Randolph Moses. *Introduction to Spectral Analysis*. Prentice Hall, Upper Saddle River, NJ, 1997.

[145] Petre Stoica, Randolph L. Moses, Benjamin Friedlander, and Torsten Söderström. Maximum likehihood estimation of the parameters of multiple sinusoids from noisy measurements. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3), March 1989.

[146] Gilbert Strang and Truong Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge, Wellesley, MA, 1996.

[147] John William Strutt. *The Theory of Sound*. Macmillan, London, 1877.

[148] N. Sugiura. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods*, A7:13–26, 1978.

[149] Kim-Chuan Toh and Lloyd N. Trefethen. Pseudozeros of polynomials and pseudospectra of companion matrices. *Numerische Mathematik*, 68:403–425, 1994.

[150] Alle-Jan Van Der Veen, Ed F. Deprettere, and A. Lee Swindlehurst. Subspace-based signal analysis using singular value decomposition. *Proceedings of the IEEE*, 81(9):1277–1308, 1993.

[151] David Vandevoorde. *A fast exponential decomposition algorithm and its applications to structured matrices*. PhD thesis, Rensselaer Polytechnic Institute, 1996.

[152] Leentie Vanhamme, Tomas Sudin, Paul Van Hecke, and Sabine Van Huffel. MR spectroscopy quantitation: a review of time-domain methods. *NMR in Biomedicine*, 14:233–246, 2001.

[153] Leentje Vanhamme, Aad van den Boogaart, and Sabine Van Huffel. Improved method for accurate and efficient quantification of MRS data with use of prior knowledge. *Journal of Magnetic Resonance*, 129:35–43, 1997.

[154] Michel Verhaegen and Vincent Verdult. *Filtering and System Identification*. Cambridge, 2007.

[155] Jan P. B. Vreeberg. Prony analysis of flat-spin motion from accelerometer data. *IEEE Transactions on Aerospace and Electronic Systems*, 48(2), 2012.

[156] Gilbert Walker. On periodicity in series of related terms. *Philosophical Transactions of the Royal Society of London, Series A*, 131(818):518–532, June 1931.

[157] L. Weiss and R. N. McDonough. Prony's method, z-transforms, and Padé approximation. *SIAM Review*, 5(2), April 1963.

[158] Eric W. Weisstein. *The CRC concise encyclopedia of mathematics*. CRC Press, Boca Raton, Florida, 1998.

[159] E. T. Whittaker and G. Robinson. *The Calculus of Observations*. Blackie & Son Limited, London, 1924.

[160] J. H. Wilkinson. The evaluation of zeros of ill-conditioned polynomials, part I. *Numerische Mathematik*, 1(1):150–166, 1959.

[161] J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford, 1965.

[162] W. J. Wiscombe and J. W. Evans. Exponential-sum fitting of radiative transmission functions. *Journal of Computational Physics*, 24(4):416 – 444, 1977.

[163] David Wood. The computation of polylogarithms. Technical Report 15-92*, Computing Laboratory, University of Kent, Canterbury, UK, June 1992.

[164] Thomas G. Wright and Lloyd N. Trefethen. Pseudospectra of rectangular matrices. *IMA Journal on Numerical Analysis*, 22:501–519, 2002.

[165] Shiyong Yang and Hongwei Li. Estimating the number of harmonics using enhanced matrix. *IEEE Signal Processing Letters*, 14(2):137 – 140, 2007.

[166] G. Udny Yule. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London, Series A*, 226:267–298, 1927.

[167] Peter W. Zehna. Invariance of maximum likelihood estimators. *Annals of Mathematical Statistics*, 37:744, 1966.

[168] Clarence Zener. Internal friction in solids: I. Theory of internal friction in reeds. *Physical Review*, 52:230–235, 1937.