UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Rational Covariance Extension, Multivariate Spectral Estimation, and Related Moment Problems: Further Results and Applications

Ph.D. Candidate:
Bin ZHU

Advisor:
Prof. Giorgio PICCI

Coordinator:
Prof. Andrea NEVIANI

Ph.D. School in
Information Engineering

Department of Information Engineering
University of Padova
2018

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

University of Padova

Department of Information Engineering

Ph.D. Course in Information Engineering

Information and Communication Technologies

The Automatica Group

Rational Covariance Extension, Multivariate Spectral Estimation, and Related Moment Problems: Further Results and Applications

Coordinator: Prof. Andrea NEVIANI

Supervisor: Prof. Giorgio PICCI

Ph.D. Student: Bin ZHU

Acknowledgements

First I would like to thank my advisor Prof. Giorgio Picci, who is really an inventor of ideas when dealing with research problems. He is also very generous with his time, encouraging, and gives me plenty of freedom to work on problems that I would like to pursue. He is willing to help, whether in research or daily life. He is surely humorous and knows a load of funny gossips of people in academia. I enjoy talking with him a lot.

I am also grateful for the support from my family, especially my wife Bili 钟碧莉 and my little daughter Giulietta 朱纯熙 who was born in November 2017. They have made me a better human being in terms of interaction with people, in particular family members, although I am still learning and trying to improve. There have been difficult situations in the past year for us, even more so for Bili since she has the major responsibility of nourishing Giulietta while at the same time she has to continue her Ph.D. But the baby power of Giulietta can alway cheer us up and everything will hopefully get better and better.

I wish to thank also people at the Department of Information Engineering (DEI). I would like to mention particularly Profs. Augusto Ferrante and Mattia Zorzi, who have constantly offered their time and suggestions for my work. My office mates in Room 330 have together created a relaxing working environment, with social activities from time to time, sometimes (twice in my experience) in the house of Nicoletta. They are all nice people and ready to help, from theorem proving to Italian language consultation.

Special gratitude goes to my former advisor Prof. Anders Lindquist (who is also a reviewer of this thesis) at Shanghai Jiao Tong University. He was the one who introduced me to the problem of rational covariance extension and led me into academia. It was also because of his appreciation and encouragement that I came to Padova to pursue a doctorate.

I must acknowledge the extremely nice comments and many useful suggestions from the thesis reviewer Prof. Federico Ramponi.

Finally, I would like to thank China Scholarship Council (CSC) for the three-year financial support for my Ph.D. under file no. 201506230140.

Summary

This dissertation concerns the problem of spectral estimation subject to moment constraints. Its scalar counterpart is well-known under the name of rational covariance extension which has been extensively studied in past decades. The classical covariance extension problem can be reformulated as a truncated trigonometric moment problem, which in general admits infinitely many solutions. In order to achieve positivity and rationality, optimization with entropy-like functionals has been exploited in the literature to select one solution with a fixed zero structure. Thus spectral zeros serve as an additional degree of freedom and in this way a complete parametrization of rational solutions with bounded degree can be obtained.

New theoretical and numerical results are provided in this problem area of systems and control and are summarized in the following. First, a new algorithm for the scalar covariance extension problem formulated in terms of periodic ARMA models is given and its local convergence is demonstrated. The algorithm is formally extended for vector processes and applied to finite-interval model approximation and smoothing problems.

Secondly, a general existence result is established for a multivariate spectral estimation problem formulated in a parametric fashion. Efforts are also made to attack the difficult uniqueness question and some preliminary results are obtained. Moreover, well-posedness in a special case is studied throughly, based on which a numerical continuation solver is developed with a provable convergence property. In addition, it is shown that solution to the spectral estimation problem is generally not unique in another parametric family of rational spectra that is advocated in the literature.

Thirdly, the problem of image deblurring is formulated and solved in the framework of the multidimensional moment theory with a quadratic penalty as regularization.

Keywords: Rational covariance extension, ARMA modeling, multivariate spectral estimation, generalized moment problem, parametrization of rational spectra, well-posedness, numerical continuation method, convex optimization.

Sommario

Questa tesi riguarda il problema della stima spettrale soggetta a vincoli sui momenti. La sua controparte scalare è ben conosciuta sotto il nome di estensione razionale delle covarianze ed è stata ampiamente studiata negli ultimi decenni. Il classico problema di estensione delle covarianze può essere riformulato come un problema dei momenti trigonometrici troncato, che in generale ammette infinite soluzioni. Al fine di ottenere positività e razionalità, in letteratura è stata sfruttata l'ottimizzazione con funzionali entropici per selezionare una soluzione con una struttura degli zeri fissa. Così gli zeri spettrali fungono da grado di libertà addizionale e permettono di ottenere una parametrizzazione completa delle soluzioni razionali con grado limitato.

Nuovi risultati teorici e numerici sono forniti in questa branca della teoria dei sistemi e del controllo e sono riassunti di seguito. Innanzitutto si propone un nuovo algoritmo per il problema scalare dell'estensione delle covarianze formulato in termini di modelli ARMA periodici e se ne dimostra la convergenza locale. L'algoritmo è esteso formalmente ai processi vettoriali e applicato ai problemi di approssimazione dei modelli a intervallo finito e di livellamento.

In secondo luogo viene stabilito un risultato di esistenza generale per un problema di stima spettrale multivariata formulato in modo parametrico. Si fanno anche sforzi per attaccare la difficile questione dell'unicità e si ottengono alcuni risultati preliminari. Inoltre, in un caso speciale è studiata a fondo la buona posizione del problema, in base alla quale è sviluppato un risolutore a continuazione numerica con convergenza dimostrabile. Per di più, si dimostra che la soluzione al problema della stima spettrale in generale non è unica in un'altra famiglia parametrica di spettri razionali proposta in letteratura.

In terzo luogo, il problema del *deblurring* delle immagini è formulato e risolto nel quadro della teoria multidimensionale dei momenti con una regolarizzazione a penalità quadratica.

Parole chiave: Estensione razionale delle covarianze, modellazione ARMA, stima spettrale multivariata, problema dei momenti generalizzato, parametrizzazione di spettri razionali, buona posizione, metodo di continuazione numerica, ottimizzazione convessa.

Contents

1 Introduction

In this Ph.D. dissertation, we consider the problem of rational covariance extension and its multivariate generalization known as a parametric spectral estimation problem.

The (scalar) *rational covariance extension problem*, also called *partial stochastic realization*, has a long history and can be traced back to the formulation first given in (Kalman, 1982). The problem aims to find an infinite extension of a finite covariance sequence such that the resulting spectral density, i.e., the Fourier transform of the infinite sequence, is a rational function. It is an important problem in realization theory, system identification, and signal processing. The formulation is quite natural in the sense explained next.

Suppose that we have one finite-length simple path of a zero-mean discrete-time stationary process and we want to build a model for it. Important features of the process are clearly described by the second-order moments, i.e., covariances, which we shall estimate from the sample path and use as data. Typically we can only obtain a finite number of reliable estimates of the covariances because high-order estimates can be very noisy due to limited samples. To make full use of the estimated covariance data, we want to find a model for the underlying process such that the covariances match the data *exactly*. Such a problem admits a neat equivalent formulation as a *trigonometric moment problem*, which goes back to some classical theories, see e.g., (Grenander and Szegö, 1958; Akhiezer and Kreĭn, 1962; Kreĭn and Nudel'man, 1977).

However, classical moment theories do not take into account the rationality constraint

imposed on the candidate model, which is of practical interest for physical realizability. In other words, we need to restrict our attention to the class of linear models with rational spectra. A first choice of the model class is the so-called *autoregressive moving-average* (ARMA) models. Existence of a solution to the problem of rational covariance extension was first proven in (Georgiou, 1983, 1987a) when the numerator polynomial of the spectrum, or equivalently, the moving-average (MA) part of the ARMA model, is fixed. It was also conjectured that the corresponding denominator polynomial, or the equivalent autoregressive (AR) part, is unique. The conjecture was proved in (Byrnes, Lindquist, Gusev, and Matveev, 1995)[1], establishing a complete parametrization of solutions. Moreover, it was shown that the parametrization is smooth, and hence the solution can be tuned continuously. Including (Byrnes, Landau, and Lindquist, 1997; Byrnes and Lindquist, 1997), major arguments in these works were made using abstract technical tools from differential topology, modern nonlinear analysis, and differential geometry, notably the degree theory and the global inverse function theorem of Hadamard.

When the MA part is just a (discrete-time) white noise process, the problem reduces to AR modeling, for which there are quite some earlier works. In particular, we want to mention Yule-Walker equations to determine the AR coefficients (Yule, 1927; Walker, 1931), and the Levinson algorithm (Porat, 1994) to compute their solution. Generalization to vector AR processes was done in (Whittle, 1963). We would like to remark that AR modeling based on covariance matching is a linear problem. Another important work is (Burg, 1967), in which the principle of *maximum entropy* was first introduced for spectral estimation. In maximum entropy spectral analysis, one aims to find a stationary process that is *the most random or the least predictable* time series while being consistent with the given covariance data. Surprisingly, the unique solution (under the feasibility assumption) that maximizes the entropy subject to moment constraints is of AR type.

Inspired by Burg's maximum entropy method, later the theory of rational covariance extension was built into a more concrete form as the theory of optimization was incorporated. Specifically in (Byrnes, Gusev, and Lindquist, 1998), it was shown that each rational solution to the covariance extension problem can be obtained by minimizing a strictly convex functional over the cone of positive Laurent polynomials (of bounded degree). In (Byrnes, Gusev, and Lindquist, 2001b), it was further discovered that such a minimization problem is the dual of maximizing a *generalized entropy functional* subject to moment constraints. This optimization framework has led to a long list of results with various directions of generalizations:

- Further development and well-posedness (Georgiou, 2001; Enqvist, 2001; Byrnes,

[1] An earlier work (Byrnes and Lindquist, 1989) on the Kimura-Georgiou parametrization of modeling filters played an important role in this paper.

Enqvist, and Lindquist, 2001c, 2002; Byrnes, Fanizza, and Lindquist, 2005; Enqvist, 2006; Enqvist and Avventi, 2007),

- Analytic interpolation problem (Georgiou, 1987b, 1999; Byrnes, Georgiou, and Lindquist, 2001a; Blomqvist, Fanizza, and Nagamune, 2003a; Blomqvist, Lindquist, and Nagamune, 2003b; Byrnes, Georgiou, Lindquist, and Megretski, 2006; Takyar and Georgiou, 2010),

- Moment problem with general basis functions (Byrnes and Lindquist, 2003, 2006, 2008, 2009),

- Spectral estimation subject to a generalized moment constraint (Byrnes, Georgiou, and Lindquist, 2000; Georgiou, 2002b,a; Georgiou and Lindquist, 2003; Georgiou, 2005, 2006; Pavon and Ferrante, 2006; Ferrante, Pavon, and Ramponi, 2007, 2008; Enqvist and Karlsson, 2008; Ramponi, Ferrante, and Pavon, 2009, 2010; Ferrante, Pavon, and Zorzi, 2010; Avventi, 2011a; Ferrante, Ramponi, and Ticozzi, 2011; Ferrante, Pavon, and Zorzi, 2012b; Zorzi and Ferrante, 2012; Ferrante, Masiero, and Pavon, 2012a; Ning, Jiang, and Georgiou, 2013; Pavon and Ferrante, 2013; Zorzi, 2014b,a, 2015; Georgiou and Lindquist, 2017; Zhu and Baggio, 2017; Zhu, 2017; Baggio, 2018a; Zhu, 2018a,b),

- ARMA modeling and the circulant problem (Enqvist, 2004; Georgiou and Lindquist, 2008; Carli, Ferrante, Pavon, and Picci, 2010, 2011; Carli and Georgiou, 2011; Lindquist and Picci, 2013; Lindquist, Masiero, and Picci, 2013; Ringh and Karlsson, 2015; Lindquist and Picci, 2016; Picci and Zhu, 2017; Zhu and Picci, 2017),

- Multidimensional theory (Karlsson, Lindquist, and Ringh, 2016; Zhu and Lindquist, 2016; Ringh, Karlsson, and Lindquist, 2016, 2018).

Among them we shall elaborate more on the problem of *spectral estimation subject to a generalized moment constraint*. As indicated in the survey paper (Robinson, 1982), spectral estimation is an old problem that has its root deep in physics. In the paper (Byrnes et al., 2000), a new approach to spectral estimation (scalar version) was introduced by Byrnes, Georgiou, and Lindquist and then further developed in (Georgiou and Lindquist, 2003) in order to allow for an *a priori* information. This formulation, known under the name of *THREE-like spectral estimation*, has now become nearly standard and includes as special cases the aforementioned problems of *covariance extension* and *analytic interpolation*. The procedure to estimate the unknown spectrum of a zero-mean wide-sense stationary signal is described briefly as follows. First the signal is fed into a rational filter and the output is collected. Then the steady-state output covariance matrix is computed, and we want to find

the input spectrum that is consistent with such covariance data. It turns out that in this way, we are dealing with a generalized version of moment equations.

Similar to its classical counterpart (Grenander and Szegö, 1958; Akhiezer, 1965; Kreĭn and Nudel'man, 1977), the generalized moment problem has infinitely many solutions when a solution exists, unless in certain degenerate cases. Therefore, such a problem is not well-posed in the sense of Hadamard[2]. The mainstream approach today to promote uniqueness of the solution is built on calculus of variations and optimization theory. It has two main ingredients. One is the introduction of a prior spectral density function Ψ as additional data, which represents our "guess" of the desired solution Φ. The other is a cost functional $d(\cdot,\cdot)$, which is usually an entropy-like distance index (divergence) between two bounded and coercive spectral densities. Then one tries to solve the optimization problem of minimizing $d(\Phi,\Psi)$ subject to the (generalized) moment equation as a constraint. Still, it is not trivial to solve such an optimization problem. Indeed, although the dual problem is typically convex, the dual variable (i.e., the Lagrange multiplier) is a Hermitian matrix that lives in an *open, unbounded* domain and this usually gives rise to a number of numerical issues.

With reference to the scalar case, results produced through this optimization approach include (Georgiou and Lindquist, 2003; Pavon and Ferrante, 2006; Ferrante et al., 2007, 2011; Baggio, 2018a), in which the chosen distance index is the Kullback–Leibler divergence; (Enqvist and Karlsson, 2008), where the Itakura-Saito distance is used; and (Zorzi, 2014b), where a general family of divergences (the Alpha divergence family) is considered. In the multivariate case, the problem becomes much more challenging and its feasibility strongly depends on the selected distance. In particular, we mention the papers (Georgiou, 2006), where a multivariate extension of the Kullback–Leibler divergence, the quantum relative entropy, is considered; (Ferrante et al., 2008; Ramponi et al., 2009, 2010), which deal with a sensible generalization of the Hellinger distance; and (Ferrante et al., 2012a; Georgiou and Lindquist, 2017), where the selected distance index coincides with the multivariate Itakura–Saito distance. It is worth remarking that the latter two approaches lead to rational solutions with bounded McMillan degrees when the prior is rational. Finally, (Zorzi, 2014a) and (Zorzi, 2015) introduce two more general frameworks based on the notions of Beta and Tau divergence families, wherein the multivariate Kullback–Leibler divergence and Itakura–Saito distance can be recovered as limiting cases.

A key feature of the "THREE" approach is that parameter (i.e., the prior function Ψ) tuning is allowed in order to achieve high resolution in specified frequency bands.

Built upon the theme of rational covariance extension, multivariate spectral estimation, and related moment problems, with motivation from the open question of covariance-

[2]Recall that a problem is well-posed if 1) a solution exists; 2) the solution is unique; 3) the solution depends continuously on the data.

consistent vector ARMA modeling, the main body of this dissertation is divided into five chapters, whose contents are described next.

Chapter 2 concerns a variant of the rational covariance extension problem stated in terms of *periodic* ARMA models defined on a finite interval, which leads to a *circulant* matrix completion problem. When formulated directly in terms of the spectral density, the scalar problem was solved in (Lindquist and Picci, 2013) and extended to a special multivariate case in (Lindquist et al., 2013; Lindquist and Picci, 2016), which are in turn generalization of earlier work on reciprocal processes (Carli et al., 2011; Carli and Georgiou, 2011). In these papers, the circulant matrix completion problem is reformulated as optimization of a generalized entropy functional, with main tools being convex optimization and calculus of variations.

Our new finding about the alternative formulation, which we call the *ARMA covariance matching problem*, is that the problem amounts to solving a set of generalized Yule-Walker equations without involving the solution of a variational problem as in (Lindquist and Picci, 2013; Lindquist et al., 2013). Although the resulting equations turn out to be nonlinear, a natural iterative scheme is apparent from their structure which brings about a new algorithm that is proven to converge at least locally in the scalar case. The algorithm can be generalized to vector processes in a straightforward manner although a proof of convergence remains to be worked out. The chapter will be based on the following two works.

- **Zhu B. and Picci G.** Proof of local convergence of a new algorithm for covariance matching of periodic ARMA models. *IEEE Control Syst. Lett.*, 1(1):206–211, 2017.

- **Picci G. and Zhu B.** Approximation of vector processes by covariance matching with applications to smoothing. *IEEE Control Syst. Lett.*, 1(1):200–205, 2017.

Chapter 3 is about a multivariate spectral estimation problem considered in (Ferrante et al., 2010), where a parametric solution was proposed with the intention of generalizing the solution form in (Georgiou and Lindquist, 2003) of a corresponding scalar problem to the multivariate case. One such generalization taking the optimization approach was presented in (Avventi, 2011a) but the prior spectral density (which is part of the given data) was still kept as scalar. In contrast, the paper (Ferrante et al., 2010) aimed to develop a *bona fide* multivariate theory in the sense that a matrix-valued prior was incorporated. However, difficulties arose since a cost function that leads to the particular form of the parametric solution is not known (unless the prior is scalar). Actually in (Ferrante et al., 2010), the question of existence of a solution parameter in general was left open, and this is our main motivation here.

Main results in this chapter are the following three. First, we give a complete proof of

well-posedness of the problem given a scalar prior, thus complementing the results in (Avventi, 2011a). Secondly, we answer the open question in (Ferrante et al., 2010) affirmatively by showing that a solution to the parametric spectral estimation problem indeed exists given an arbitrary matrix-valued prior. Thirdly, we give some preliminary development to approach the much harder uniqueness question. This chapter is mostly based on the next two works.

- **Zhu B. and Baggio G.** On the existence of a solution to a spectral estimation problem *à la* Byrnes-Georgiou-Lindquist. To appear in IEEE Trans. Automat. Control, arXiv e-print: 1709.09012, 2017.

- **Zhu B.** On a parametric spectral estimation problem. Presented at the 18th IFAC Symposium on System Identification (SYSID 2018), arXiv e-print: 1712.07970, 2017.

Chapter 4 elaborates on two numerical solvers for the spectral estimation problem with a scalar prior, in which case results of well-posedness have been obtained in the previous chapter. Computations are carried out in the domain of spectral factors corresponding to a set of positive definite rational spectral densities due to the improvement of numerical conditioning. The first solver is a class of descent algorithms for minimizing a cost function that is only locally convex. As a consequence, only local convergence is guaranteed. Such an idea of optimization has been developed in (Avventi, 2011a), and the difference here is that our (optimization) variables do not contain redundant elements due to the diffeomorphic map of spectral factorization introduced in Section 3.5.

The second solver is an application of the continuation method (cf. (Allgower and Georg, 1990)) to numerically invert the moment map in a parametric form. Due to well-posedness, the desired solution parameter can be achieved by solving an ordinary differential equation given the initial condition. Instead of doing just numerical integration, a specialized algorithm in this context called "predictor-corrector" is implemented and a proof of convergence is worked out based on the argument of the famous Kantorovich theorem. Moreover, a crucial step in the Newton iterations, namely, computation of the inverse Jacobian is detailed utilizing a spectral factorization technique. This chapter is mainly based on the following paper.

- **Zhu B.** On the well-posedness of a parametric spectral estimation problem and its numerical solution. Conditionally accepted for publication in IEEE Trans. Automat. Control, arXiv e-print: 1802.09330, 2018a.

Chapter 5 is devoted to the multivariate spectral estimation problem with candidate solutions restricted to another family of matrix spectral densities. This alternative parametrization of spectral densities was proposed in the published paper (Georgiou, 2006), in which it is also claimed that uniqueness of the solution holds in that family. However, we show through

a numerical example that such a claim in general fails. An important point reflected by the numerical example is that the moment map in that alternative parametric form can have a critical point, namely, a point at which its Jacobian loses rank. Moreover, the critical point in the example is demonstrated to be a bifurcation point, implying that the moment map is not injective. Work in this chapter has been reported in the next paper.

- **Zhu B.** On Theorem 6 in "Relative Entropy and the Multivariable Multidimensional Moment Problem". Submitted to IEEE Trans. Inform. Theory, arXiv e-print: 1805.12060, 2018b.

Chapter 6 presents a new method of reconstructing an image that undergoes a spatially invariant blurring process and is corrupted by noise. The methodology is based on a theory of multidimensional moment problems with rationality constraints. This can be seen as generalized spectral estimation with a finiteness condition, which in turn can be considered a problem in system identification. With noise it becomes an ill-posed deconvolution problem and needs regularization. A Newton solver is developed, and the algorithm is tested on two images under different boundary conditions. These preliminary results show that the proposed method could be a viable alternative to regularized least squares for image deblurring, although more work is needed to perfect the method. The content of this chapter has been included in the conference paper below.

- **Zhu B. and Lindquist A.** An identification approach to image deblurring. In *Proc. 35th Chinese Control Conference (CCC 2016)*, pages 235–241. IEEE, 2016.

After the main body of this dissertation, there are three chapters of appendices. Appendix A collects some preliminaries on the spectral analysis of stationary periodic processes, and its connection with (block-)circulant matrices. These serve in Chapter 2. Appendix B gives some supplementary propositions and lemmas with some ancillary results on homogeneous polynomial equations, which play roles in the proofs of Chapter 3. Appendix C contains a standard procedure that takes the additive decomposition of a spectral density function to the corresponding outer factor, which is needed in Chapter 4.

List of Symbols

Some notations are common as $\mathbb{E}$ denotes mathematical expectation, $\mathbb{C}$ the complex plane, $\mathbb{Z}$ the set of integers, and $\mathbb{T}$ the unit circle $\{z : |z| = 1\}$. Those listed below are mainly used in Chapters 3, 4, and 5.

Sets:

- $\mathbb{D}$, the open complex unit disk $\{z \in \mathbb{C} : |z| < 1\}$. The unit circle $\mathbb{T} \equiv \partial\mathbb{D}$, where ∂ stands for the boundary.
- $\mathrm{GL}(n, \mathbb{C})$, group of $n \times n$ invertible complex matrices.
- $\mathfrak{H}_n$, the vector space of $n \times n$ Hermitian matrices.
- $\mathfrak{H}_{+,n}$, the subset of $\mathfrak{H}_n$ that contains positive definite matrices.
- $C(\mathbb{T}; \mathfrak{H}_m)$, the space of $\mathfrak{H}_m$-valued continuous functions on $\mathbb{T}$.
- $C_+(\mathbb{T})$, the set of continuous functions on $\mathbb{T}$ that take real and positive values, which is an open subset (under the metric topology) of $C(\mathbb{T}) \equiv C(\mathbb{T}; \mathfrak{H}_1)$.
- $\mathfrak{S}_m$, the family of $\mathfrak{H}_{+,m}$-valued functions defined on $\mathbb{T}$ that are bounded and coercive. More technically, for $\Psi \in \mathfrak{S}_m$, there exist real positive constants μ, M such that $\mu I \leq \Psi(e^{i\theta}) \leq MI$ for all $\theta \in (-\pi, \pi]$.

Linear algebra:

- $(\cdot)^*$, complex conjugate transpose. When considering a rational matrix-valued function with a state-space realization $G(z) = C(zI - A)^{-1}B + D$, $G^*(z) := B^*(z^{-1}I - A^*)^{-1}C^* + D^*$.
- $(\cdot)^{-*}$, shorthand for $[(\cdot)^{-1}]^*$.
- $\langle A, B \rangle := \mathrm{tr}(AB^*)$, matrix inner product for $A, B \in \mathbb{C}^{m \times n}$.

- $\|A\|_F := \sqrt{\langle A, A\rangle}$, the Frobenius norm.
- $\|x\|_2 := \sqrt{x^* x}$, the Euclidean 2-norm of $x \in \mathbb{C}^n$. The subscript is usually omitted and we simply write $\|\cdot\|$. When applied to a matrix $A \in \mathbb{C}^{m\times n}$ or more generally a multilinear function, $\|A\|$ means the induced 2-norm.

2

Periodic ARMA Modeling Based on Covariance Matching

2.1 Introduction

Traditionally the problem of ARMA modeling is usually cast in the framework of Maximum Likelihood (see e.g., (Rosenblatt, 1985)), or Minimum Prediction Error, which leads in general to a nonconvex optimization problem that may have many local minima. Uniqueness can only be proven asymptotically under *unverifiable* assumptions that the true model belongs to the model class and the data are ergodic. Finding the order of a best approximate model still seems to be to a large extent an unsolved problem.

In contrast, we consider in this chapter the ARMA modeling problem of periodic stationary processes by matching a finite number of (estimated) covariance lags. As we shall see later, this is in fact equivalent to the *circulant rational covariance extension problem* studied in (Lindquist and Picci, 2013; Lindquist et al., 2013; Lindquist and Picci, 2016), when the data are restricted to a finite interval. Under this framework, model order of the AR part is *a priori* fixed and equal to the number of available covariance lags. Then we know from previous works that for any fixed MA part, there is a unique AR part such that the resulting model has covariances matching the data.

The main contribution of this chapter is devising a new numerical algorithm to the

circulant rational covariance extension problem and proving its local convergence. We have found that the ARMA formulation leads to a generalization of the Yule-Walker equations whose structure naturally suggests an iterative solution. This idea seems to provide a viable alternative to the variational formulation of essentially the same problem. The convergence proof of the proposed iterative algorithm could be approached from the variational point of view by interpreting it as a quasi-Newton-type iteration. This idea requires a reformulation of the dual optimization problem in terms of the spectral factor, and then one can perform convex analysis (see e.g., (Boyd and Vandenberghe, 2004)) locally. In this chapter we take an alternative route and show local convergence using an elegant Lyapunov-type analysis of the algorithm interpreted as a nonlinear dynamical system. The analysis of convergence is carried out in the scalar case. A generalization to matrix covariance extension problem entails some extra complexity due to certain redundancy in the parameter and is yet to be completed.

As an application of the theory, periodic ARMA models can provide a useful finite-interval approximation of a stationary state space model, e.g., a Gauss-Markov model defined on the whole time line $\mathbb{Z}$, by matching a certain number of covariances lags of the original process. This finite-interval model can then be used to derive an easy-to-implement constant-coefficients algorithm for linear smoothing of data of finite duration.

The outline of this chapter is as follows: In Section 2.2, we review the representation of finite-interval scalar processes by periodic ARMA models and formulate the covariance matching problem. In Section 2.3, we approach the problem by deriving a set of nonlinear Yule-Walker equations. An iterative algorithm to compute the solution is described. The main results of this chapter are presented in Section 2.4, where we study in detail the local convergence of the algorithm viewed as a nonlinear dynamical system. The convergence is then proven via Lyapunov stability analysis. Later the algorithm is extended for the matrix covariance matching problem of vector ARMA models in Section 2.5, although a convergence proof is still absent. Finally in Section 2.6, we demonstrate an application of the theory to the finite-interval smoothing problem, where model approximation by covariance matching is exploited. A simplified numerical example is given for illustration.

2.2 Covariance Matching for Scalar Periodic ARMA Processes

Consider a discrete-time zero-mean second order stationary real process $y(t)$, defined on a finite interval $[-N+1, N]$ of the integer line $\mathbb{Z}$ and extended to all of $\mathbb{Z}$ as a periodic process with period $2N$ so that $y(t+2kN) = y(t)$, $k \in \mathbb{Z}$ almost surely. We shall write it as a random

vector

$$\mathbf{y} := \begin{bmatrix} y(t-N+1) & y(t-N+2) & \dots & y(t+N) \end{bmatrix}^\top \tag{2.1}$$

As shown in (Carli et al., 2011), in order for the random vector $\mathbf{y}$ to represent the restriction to $[-N+1, N]$ of a periodic process on $\mathbb{Z}$, the covariances $c_k := \mathbb{E}\, y(t+k)y(t)\,;\, k = 0, 1, \dots, N$, must form a *circulant matrix*, namely the matrix $\mathbf{\Sigma} := \mathbb{E}\,\mathbf{y}\mathbf{y}^\top$ must have the form

$$\begin{aligned} \mathbf{\Sigma} &= \begin{bmatrix} c_0 & c_1 & \cdots & c_N & c_{N-1} & \cdots & c_1 \\ c_1 & c_0 & \cdots & c_{N-1} & c_N & \cdots & c_2 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \vdots \\ c_{N-1} & c_{N-2} & \cdots & c_0 & \cdots & c_{N-1} & c_N \\ c_N & c_{N-1} & & \cdots & c_0 & \cdots & c_{N-1} \\ \vdots & \ddots & \ddots & \ddots & & \ddots & \vdots \\ c_1 & \cdots & c_N & c_{N-1} & \cdots & c_1 & c_0 \end{bmatrix} \\ &:= \mathrm{Circ}\{c_0, c_1, c_2, \dots, c_N, c_{N-1}, \dots, c_2, c_1\}, \end{aligned} \tag{2.2}$$

where we have the usual symmetry for the covariances $c_{-k} = c_k$. Circulant matrices will play a key role in the following. They are completely determined by their first column (or row). That is why we introduce the simplified notation $\mathrm{Circ}\{\cdot\}$. More references can be found in (Davis, 1979; Gray, 2006) .

Some important facts about the discrete Fourier transform (DFT), stationary periodic processes, and (block-)circulant matrices are collected in Appendix A. Here we recall some basics for the development. By stationarity y has a spectral representation

$$y(t) = \int_{-\pi}^{\pi} e^{it\theta} d\hat{y}(\theta), \quad \text{where} \quad \mathbb{E}\{|d\hat{y}(\theta)|^2\} = dF(\theta) \tag{2.3}$$

is the spectral distribution (see, e.g., (Lindquist and Picci, 2015, p. 74)), so that

$$c_k = \mathbb{E}\{y(t+k)y(t)\} = \int_{-\pi}^{\pi} e^{ik\theta} dF(\theta)\,. \tag{2.4}$$

As explained in (Lindquist and Picci, 2013), the support of the spectral distribution dF must be contained in the *discrete unit circle* $\mathbb{T}_{2N} := \{\zeta_{-N+1}, \zeta_{-N+2}, \dots, \zeta_N\}$, where

$$\zeta_k = e^{ik\pi/N}, \quad k = -N+1, \dots, N, \tag{2.5}$$

because of the periodicity condition. Moreover, one can represent the spectral distribution as $dF = \Phi\, d\nu$ where $d\nu$ is a uniform discrete measure supported on $\mathbb{T}_{2N}$ (cf. (A.4) for the

formula), and Φ is the DFT of the sequence $\{c_{-N+1},\dots,c_N\}$, called the *spectral density* of y,

$$\Phi(\zeta) := \sum_{k=-N+1}^{N} c_k\,\zeta^{-k}, \tag{2.6}$$

which is also known as the *symbol* of the circulant matrix $\mathbf{\Sigma}$. This is a nonnegative function of the discrete variable $\zeta \in \mathbb{T}_{2N}$ which is strictly positive if and only if the $2N \times 2N$ covariance matrix $\mathbf{\Sigma}$ is positive definite (see Appendix A, also (Carli and Georgiou, 2011, Proposition 2)), that is to say, the process is *full rank* which we shall assume all through this chapter.

2.2.1 Bilateral ARMA Model

The dynamics of a periodic process can be defined in terms of relations among its random variables in just one particular period, namely $[-N+1, N]$ which will be identified with the finite modular group $\mathbb{Z}_{2N}$. In this setting, the most general finitely parametrized analog of a stationary finite-dimensional linear model for a periodic process turns out to be a *bilateral ARMA model* of finite order n (Carli et al., 2011; Lindquist and Picci, 2013)

$$\sum_{k=-n}^{n} q_k\,y(t-k) = \sum_{k=-n}^{n} p_k\,e(t-k), \quad t \in \mathbb{Z}_{2N}, \tag{2.7}$$

with $\{q_k, p_k\}$ real parameters satisfying the symmetry $p_{-k} = p_k$ and $q_{-k} = q_k$ and $e(t)$ a periodic process, called the *conjugate process* of $y(t)$ (also called the *double-sided innovation* (Masani, 1960)). The conjugate process is *delta-correlated with* y in the sense that $\mathbb{E}y(t)e(s) = \delta_{t,s}$ where $\delta_{t,s} = 1$ for $t = s$ and zero otherwise. By periodicity, the model is associated to periodic boundary conditions at the end points

$$y(-N) = y(N),\dots,\ y(-N-n+1) = y(N-n+1), \tag{2.8}$$

which induce a circulant structure on the model (2.7). More explicitly, the bilateral model (2.7) can be rewritten as an equivalent circulant matrix equation

$$\mathbf{Q}\,\mathbf{y} = \mathbf{P}\,\mathbf{e}$$

where $\mathbf{Q}$ and $\mathbf{P}$ are symmetric positive semidefinite *circulants* with elements the coefficients $\{q_k\}$ and $\{p_k\}$ defined above, and the vector

$$\mathbf{e} := \begin{bmatrix} e(t-N+1) & e(t-N+2) & \dots & e(t+N) \end{bmatrix}^\top. \tag{2.9}$$

Moreover, $\mathbf{Q}$ and $\mathbf{P}$ are *banded* $2N \times 2N$ circulants of bandwidth n as illustrated below.

$$\mathbf{Q} = \begin{bmatrix} q_0 & q_1 & \cdots & q_n & 0 & \cdots & 0 & q_n & \cdots & q_1 \\ q_1 & q_0 & q_1 & \cdots & q_n & 0 & \cdots & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & & \vdots & \ddots & & & \ddots & q_n \\ q_n & q_{n-1} & \cdots & q_0 & q_1 & \cdots & q_n & 0 & \cdots & 0 \\ 0 & q_n & q_{n-1} & \cdots & q_0 & q_1 & & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & & \ddots & \ddots & & \ddots & 0 \\ 0 & 0 & \cdots & q_n & \cdots & q_1 & q_0 & q_1 & \cdots & q_n \\ q_n & 0 & \ddots & & \ddots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 & q_n & \cdots & q_1 & q_0 & q_1 \\ q_1 & \cdots & q_n & 0 & \cdots & 0 & q_n & \cdots & q_1 & q_0 \end{bmatrix}$$
$$:= \operatorname{Circ}\{q_0, q_1, \ldots, q_n, 0, \ldots, 0, q_n, \ldots, q_1\}$$

Because of the orthogonality of $\mathbf{y}$ to its conjugate process, assuming invertibility of $\mathbf{Q}$, the covariance matrix $\mathbf{\Sigma}$ of the stacked vector $\mathbf{y}$ (2.1) has the expression

$$\mathbf{\Sigma} = \mathbf{Q}^{-1}\mathbf{P}, \tag{2.10}$$

which is also a symmetric positive semidefinite circulant matrix. The special structure of the resulting covariance matrix expressed as a ratio of two circulant banded matrices, will later be exploited in Section 2.6 to solve the finite-interval smoothing problem in a stationary setting.

The expression (2.10) has an analog in terms of the spectral representation of the process y, due to the fact that the DFT is an *algebra homomorphism* mapping the circulant matrices of the same dimension to their symbols (cf. Section A.2, also (Lindquist and Picci, 2013, p. 2851)). More precisely, the representation (2.10) is equivalent to

$$\Phi(\zeta) = \frac{P(\zeta)}{Q(\zeta)} \tag{2.11}$$

where

$$Q(\zeta) := \sum_{k=-n}^{n} q_k \zeta^{-k}, \quad P(\zeta) := \sum_{k=-n}^{n} p_k \zeta^{-k} \tag{2.12}$$

are Laurent polynomials (with both positive and negative powers of the indeterminate) of degree n and positive semidefinite on $\mathbb{T}_{2N}$; they are also symbols of the circulants $\mathbf{Q}$ and $\mathbf{P}$. In other words, the spectral density $\Phi(\zeta)$ defined in (2.6) is now a *rational* function. We shall

require that

$$Q(\zeta) \neq 0, \ \forall\, \zeta \in \mathbb{T}_{2N}, \tag{2.13}$$

which is in turn equivalent to the nonsingularity of $\mathbf{Q}$ (Carli and Georgiou, 2011).

2.2.2 Unilateral ARMA Model and Problem Formulation

As described above, periodic processes can be conveniently seen as being defined on the finite group $\mathbb{Z}_{2N}$ made of the discrete interval $[-N+1, N]$ with arithmetic modulo $2N$. The bilateral model (2.7) has an equivalent *unilateral* representation of the form

$$\sum_{k=0}^{n} a_k y(t-k) = \sum_{k=0}^{n} b_k w(t-k), \qquad t \in \mathbb{Z}_{2N} \tag{2.14}$$

where $w(t)$ is a periodic white noise on $\mathbb{Z}_{2N}$ of unit variance and $\{a_k, b_k\}$ are real parameters. The equivalence is established through spectral factorization as we shall describe next. Moreover, we will formulate the covariance matching problem for the unilateral model since it is more useful for the recursive implementation of our algorithm.

Given the periodic boundary conditions (2.8), after introducing the vector notation

$$\mathbf{w} := \begin{bmatrix} w(-N+1) & w(-N+2) & \dots & w(N) \end{bmatrix}^\top$$

with $\mathbb{E}\{\mathbf{w}\mathbf{w}^\top\} = I_{2N}$ (identity), we obtain a compact circulant matrix representation of the model (2.14)

$$\mathbf{A}\mathbf{y} = \mathbf{B}\mathbf{w}, \tag{2.15}$$

where $\mathbf{A}$ and $\mathbf{B}$ are $2N \times 2N$ nonsingular *circulant lower-triangular*[1] matrices of bandwidth n denoted

$$\begin{aligned} \mathbf{A} &= \mathrm{Circ}\{a_0, a_1, \dots, a_n, 0, \dots, 0\}, \\ \mathbf{B} &= \mathrm{Circ}\{b_0,\ b_1, \dots, b_n, 0, \dots, 0\}. \end{aligned} \tag{2.16}$$

The symbols of $\mathbf{A}$ and $\mathbf{B}$ are the polynomials $a(\zeta),\ b(\zeta)$ in the indeterminate ζ defined in terms of the model coefficients as

$$a(\zeta) := \sum_{k=0}^{n} a_k\, \zeta^{-k}, \quad b(\zeta) := \sum_{k=0}^{n} b_k\, \zeta^{-k}, \tag{2.17}$$

[1] Notice that due to the circulant structure, the two matrices $\mathbf{A}$ and $\mathbf{B}$ are not really lower-triangular. They are called so because most of the upper-triangular entries are indeed zero since it is usually the case that $n \ll N$.

where the negative exponent in ζ^{-k} agrees with the interpretation of DFT as a k-steps delay operator in the frequency domain. In terms of the DFT, the model (2.14) can be rewritten as

$$a(\zeta)\hat{y}(\zeta) = b(\zeta)\hat{w}(\zeta), \qquad \zeta \in \mathbb{T}_{2N} \tag{2.18}$$

where

$$\hat{y}(\zeta) = \sum_{t=-N+1}^{N} y(t)\zeta^{-t}, \quad \hat{w}(\zeta) = \sum_{t=-N+1}^{N} w(t)\zeta^{-t}$$

are the DFT of the random vectors $\mathbf{y}$ and $\mathbf{w}$. The solution of (2.18) can formally be written as

$$\hat{y}(\zeta) = \frac{b(\zeta)}{a(\zeta)}\hat{w}(\zeta). \tag{2.19}$$

After plain calculation, it can be shown that the DFT $\hat{w}(\zeta_k)$ satisfies

$$\frac{1}{2N}\mathbb{E}\left[\hat{w}(\zeta_k)\overline{\hat{w}(\zeta_l)}\right] = \delta_{kl}. \tag{2.20}$$

From (2.19) it readily follows that the spectral density of $y(t)$ is

$$\Phi(\zeta) = \frac{1}{2N}\mathbb{E}\left[\hat{y}(\zeta)\overline{\hat{y}(\zeta)}\right] = \frac{b(\zeta)b(\zeta^{-1})}{a(\zeta)a(\zeta^{-1})} := \frac{P(\zeta)}{Q(\zeta)} \tag{2.21}$$

which is a rational function, i.e., quotient of two symmetric positive polynomials

$$P(\zeta) := b(\zeta)b(\zeta^{-1}), \qquad Q(\zeta) := a(\zeta)a(\zeta^{-1}). \tag{2.22}$$

We consider now the covariance matching problem for periodic unilateral ARMA processes.

Problem 2.2.1. Suppose that we are given the MA coefficients $\{b_k\,;\, k = 0, 1, 2, \dots, n\}$ of (2.14) and a partial covariance sequence $c_0, c_1, \dots, c_n$ with $n < N$, such that the Toeplitz matrix

$$\mathbf{T}_n = \begin{bmatrix} c_0 & c_1 & c_2 & \cdots & c_n \\ c_1 & c_0 & c_1 & \cdots & c_{n-1} \\ c_2 & c_1 & c_0 & \cdots & c_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_n & c_{n-1} & c_{n-2} & \cdots & c_0 \end{bmatrix}, \tag{2.23}$$

is positive definite. Determine the AR coefficients $\{a_k\}$ such that the first $n+1$ covariance lags of the periodic process $y(t)$ defined by (2.14) match the sequence $\{c_k\}$.

This problem, once stated in terms of the symmetric polynomials (2.22), is essentially the same moment problem discussed in (Lindquist and Picci, 2013, 2016), where it is proven

that for any fixed positive polynomial $P(\zeta)$, there is a unique solution $Q(\zeta)$. Actually the equivalence holds modulo a factorability condition which is stated in the next proposition.

Proposition 2.2.2. *Assuming $N > n$, then the Laurent polynomial $P(\zeta) = \sum_{-n}^{n} p_k \zeta^{-k}$ admits a factorization $P(\zeta) = a(\zeta)a(\zeta^{-1})$ with $a(\zeta)$ as in (2.17), if and only if $P(z) = a(z)a(z^{-1})$ for $z \in \mathbb{T}$, i.e., the usual polynomial factorization holds with the same coefficients.*

Proof. Sufficiency is obvious. For the necessity, suppose $P(\zeta) = a(\zeta)a(\zeta^{-1})$ holds and define the Laurent polynomial in z

$$\check{P}(z) = \sum_{k=-n}^{n} \check{p}_k z^{-k} := a(z)a(z^{-1}).$$

Then the two polynomials $P(z)$ and $\check{P}(z)$ of order $2n$ coincide at $2N$ points, i.e., $\zeta_{-N+1}, \ldots, \zeta_N$. Since $N > n$, this implies that $P(z) = \check{P}(z)$ and hence admits a usual polynomial factorization. ∎

Positivity of a symmetric polynomial $P(\zeta)$ on $\mathbb{T}_{2N}$ does not necessarily imply the existence of *banded* spectral factors. For this to hold, nonnegativity of the extension $P(z)$ on the whole unit circle is necessary. Although such a requirement may seem restrictive, it is in fact satisfied if N is large enough, which we shall assume for the remaining part of this chapter. This point is demonstrated as follows.

Proposition 2.2.3. *Let $P(\zeta) = \sum_{-n}^{n} p_k \zeta^{-k}$ be positive on $\mathbb{T}_{2N}$. If N is large enough, the extension of $P(\zeta)$ to the unit circle $P(z)$, must be nonnegative for all $z \in \mathbb{T}$.*

Proof. Suppose that for some $z_0 \in \mathbb{T}$, $P(z_0) < 0$. Then there must exist an interval neighborhood $\mathscr{I}$ of z_0 in $\mathbb{T}$ having a positive Lebesgue measure such that $P(e^{i\theta}) < 0$ for any $e^{i\theta} \in \mathscr{I}$. But if N is large enough some $\zeta_k \in \mathbb{T}_{2N}$ must belong to this neighborhood and then $P(\zeta_k)$ must be negative which is a contradiction. ∎

2.3 An Iterative Algorithm Based on a Nonlinear Yule-Walker Equation

Let $\gamma := \{\gamma_k\,;\, k = -N+1, \ldots, N\}$ denote the inverse DFT of $\frac{b(\zeta)}{a(\zeta)}$. The time-domain version of (2.19) is a (circulant) convolution representation of $y(t)$ in terms of the input noise $w(t)$

$$y(t) = \sum_{s=-N+1}^{N} \gamma_{t-s} w(s), \quad t \in \mathbb{Z}_{2N}, \tag{2.24}$$

which can also be written in the matrix notation as

$$\mathbf{y} = \mathbf{\Gamma w}, \tag{2.25}$$

where $\mathbf{\Gamma} = \text{Circ}\{\gamma_0, \gamma_1, \ldots, \gamma_N, \gamma_{-N+1}, \ldots, \gamma_{-1}\}$ has the symbol

$$\Gamma(\zeta) := \sum_{t=-N+1}^{N} \gamma_t \, \zeta^{-t} = \frac{b(\zeta)}{a(\zeta)}. \tag{2.26}$$

In circulant matrix notation, from (2.15) we have

$$\mathbf{\Gamma} = \mathbf{A}^{-1}\mathbf{B}. \tag{2.27}$$

Now, multiplying the model equation (2.15) on both sides from the right with the transpose of (2.25) and taking expectations, we obtain an equation for the circulant covariance

$$\mathbf{A\Sigma} = \mathbf{B\Gamma}^{\top}. \tag{2.28}$$

Introduce the vector notation

$$\mathbf{a} = \begin{bmatrix} a_0 & \ldots & a_n \end{bmatrix}^{\top}, \quad \mathbf{b} = \begin{bmatrix} b_0 & \ldots & b_n \end{bmatrix}^{\top}$$

and denote the upper-left $(n+1) \times (n+1)$ submatrix of $\mathbf{\Sigma}$ by $\mathbf{\Sigma}_n$. Since $\mathbf{b}$ is fixed, the covariance matrix $\mathbf{\Sigma}$ is a function of $\mathbf{a}$ so it is appropriate to denote $\mathbf{\Sigma}_n$ by $\mathbf{\Sigma}_n(\mathbf{a})$. With these notations, our covariance matching equation can be written as

$$\mathbf{T}_n = \mathbf{\Sigma}_n(\mathbf{a}). \tag{2.29}$$

Our algorithm is based on a consequence of (2.29) which is obtained by a Yule-Walker-type calculation combining the model equation (2.14) with the one-sided representation (2.24). It is a nonlinear equation in the coefficient vector $\mathbf{a}$ of the polynomial $a(\zeta)$ having the form

$$\mathbf{T}_n\mathbf{a} = \mathbf{\Gamma}_n\mathbf{b}, \tag{2.30}$$

where $\mathbf{T}_n$ is the data matrix (2.23), and

$$\mathbf{\Gamma}_n = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_n \\ \gamma_{-1} & \gamma_0 & & \vdots \\ \vdots & & \ddots & \gamma_1 \\ \gamma_{-n} & \cdots & \gamma_{-1} & \gamma_0 \end{bmatrix}.$$

is the upper-left $(n+1)\times(n+1)$ block of the circulant impulse response matrix $\mathbf{\Gamma}$. For the same reason as in the case of $\mathbf{\Sigma}_n(\mathbf{a})$, we shall denote $\mathbf{\Gamma}_n$ by $\mathbf{\Gamma}_n(\mathbf{a})$. With this definition of $\Gamma_n(\mathbf{a})$, we have

$$\mathbf{\Gamma}_n(\mathbf{a})\mathbf{b} = \mathbf{\Sigma}_n(\mathbf{a})\mathbf{a}, \tag{2.31}$$

which obviously agrees with (2.29) and (2.30).

It is evident that any $\mathbf{a}$ solving (2.29) will be a solution to (2.30). On the other hand, the nonlinear equation (2.30) for $\mathbf{a}$ has in general several solutions, corresponding to different spectral factors $a(\zeta)$ of $Q(\zeta)$ obtained by flipping zeros about the unit circle. Among them we shall privilege the unique polynomial whose extension obtained by substituting ζ with $z\in\mathbb{C}$ is a Schur (minimum phase) polynomial. Here we slightly modify the definition of a Schur polynomial to accommodate the convention of the Fourier transform. Specifically, the set $\mathcal{S}_n$ of Schur polynomials of degree n contains those

$$p(z) = \sum_{k=0}^{n} p_k z^{-k}, \quad p_0 > 0$$

such that $p(z)$ has all its roots strictly inside the unit circle. We define also the set

$$\mathscr{S}_n := \left\{ \mathbf{p} = \begin{bmatrix} p_0 & \cdots & p_n \end{bmatrix} : \; p(z) \in \mathcal{S}_n \right\}$$

to distinguish the polynomials from their coefficients. Before attempting a solution to (2.30), let us define the set of vectors

$$\mathscr{A} := \left\{ \mathbf{a} \in \mathbb{R}^{n+1} \; : \; a(z) := \sum_{k=0}^{n} a_k z^{-k} \neq 0, \; \forall\, z \in \mathbb{T} \right\},$$

and notice that for any $\mathbf{a}\in\mathscr{A}$,

$$\mathbf{a}^\top \mathbf{\Sigma}_n(\mathbf{a})\mathbf{a} = \frac{1}{2N} \sum_{k=-N+1}^{N} P(\zeta_k) := m_P\,,$$

where m_P is a constant once the numerator $P(\zeta)$ of the spectral density $\Phi(\zeta)$ is fixed. Thus any solution to (2.30) must satisfy the constraint

$$\mathbf{a}^\top \mathbf{T}_n \mathbf{a} = m_P. \tag{2.32}$$

We call the model (2.14) *normalized* if the above constraint for the AR coefficients is satisfied. For any nonzero vector $\mathbf{a}$, the following map achieves the normalization

$$\mathbf{s}: \ \mathbf{a} \mapsto \sqrt{\frac{m_P}{\mathbf{a}^\top \mathbf{T}_n \mathbf{a}}}\mathbf{a}. \tag{2.33}$$

Now, consider the following iterative algorithm to solve numerically the nonlinear equation (2.30).

Algorithm 2.1 Fixed-point iteration with renormalization: scalar case

Initialize $\mathbf{a}^{(0)}$ e.g., as the output of the Levinson algorithm for the ordinary covariance extension
Set $k = 0$ and a threshold τ to decide convergence
repeat
  $\mathbf{a}^{(k+1)} := \mathbf{T}_n^{-1} \mathbf{\Gamma}_n(\mathbf{a}^{(k)})\mathbf{b}$
  Rescale $\mathbf{a}^{(k+1)} := \mathbf{s}(\mathbf{a}^{(k+1)})$
  Update $k := k + 1$
until $\|\mathbf{a}^{(k)} - \mathbf{a}^{(k-1)}\| \leq \tau$
return the last $\mathbf{a}^{(k)}$

The algorithm above has a connection with the variational approach stated in the next proposition, in which we shall introduce the objective function $\mathbb{J}_P$ from (Lindquist and Picci, 2013, Theorem 2).

Proposition 2.3.1. *Step 2 of Algorithm 2.1 can be interpreted as a quasi-Netwon step for the minimization of the function*

$$\mathbb{J}_P(\mathbf{a}) = \mathbf{a}^\top \mathbf{T}_n \mathbf{a} - \int_{-\pi}^{\pi} b(e^{i\theta}) b(e^{-i\theta}) \log[a(e^{i\theta}) a(e^{-i\theta})] \, d\nu. \tag{2.34}$$

Proof. We first compute the gradient

$$\nabla \mathbb{J}_P(\mathbf{a}) = 2\mathbf{T}_n \mathbf{a} - \int_{-\pi}^{\pi} b(e^{i\theta}) b(e^{-i\theta}) \left\{ \frac{1}{a(e^{i\theta})} \bar{\mathbf{u}}(e^{i\theta}) + \frac{1}{a(e^{-i\theta})} \mathbf{u}(e^{i\theta}) \right\} d\nu, \tag{2.35}$$

where for convenience we have introduced the column vectors

$$\begin{aligned}\mathbf{u}(z) &= \begin{bmatrix} 1 & z & \dots & z^n \end{bmatrix}^\top, \\ \bar{\mathbf{u}}(z) &= \begin{bmatrix} 1 & z^{-1} & \dots & z^{-n} \end{bmatrix}^\top.\end{aligned} \tag{2.36}$$

The left term of the integrand in (2.35) can be written as

$$\frac{b(e^{i\theta})}{a(e^{i\theta})}\bar{\mathbf{u}}(e^{i\theta})\mathbf{u}(e^{i\theta})^\top\mathbf{b},$$

so that this part of the integral becomes the sum

$$\frac{1}{2N}\sum_{j=-N+1}^{N}\frac{b(\zeta_j)}{a(\zeta_j)}\begin{bmatrix} 1 & \zeta_j & \dots & \zeta_j^n \\ \zeta_j^{-1} & 1 & & \vdots \\ \vdots & & \ddots & \zeta_j \\ \zeta_j^{-n} & \dots & \zeta_j^{-1} & 1 \end{bmatrix}\mathbf{b} = \mathbf{\Gamma}_n(\mathbf{a})\mathbf{b},$$

where the equality follows from the definition of γ. The computation involving the other term in the integral is similar, yielding in fact the same result, so that

$$\nabla\mathbb{J}_P(\mathbf{a}) = 2[\mathbf{T}_n\mathbf{a} - \mathbf{\Gamma}_n(\mathbf{a})\mathbf{b}]. \tag{2.37}$$

which, in force of (2.31) is equivalent to

$$\frac{1}{2}\nabla\mathbb{J}_P(\mathbf{a}) = [\mathbf{T}_n - \mathbf{\Sigma}_n(\mathbf{a})]\,\mathbf{a}. \tag{2.38}$$

The iteration in Step 2 of the algorithm can therefore be written as

$$\mathbf{T}_n[\mathbf{a}^{(k+1)} - \mathbf{a}^{(k)}] = \mathbf{\Gamma}_n(\mathbf{a}^{(k)})\mathbf{b} - \mathbf{T}_n\mathbf{a}^{(k)} = -\frac{1}{2}\nabla\mathbb{J}_P(\mathbf{a}^{(k)})$$

which is a quasi-Newton step

$$\mathbf{a}^{(k+1)} = \mathbf{a}^{(k)} - \frac{1}{2}\mathbf{T}_n^{-1}\nabla\mathbb{J}_P(\mathbf{a}^{(k)}). \tag{2.39}$$

■

The functional $\mathbb{J}_P$ parametrized in $Q(\zeta)$ is strictly convex (Lindquist and Picci, 2013, Theorem 2). Following the lines in (Enqvist, 2001, Propositions 4–7), one can show that (2.34) is in fact locally strictly convex about the normalized solution $\mathbf{a}$ of (2.29) in the set $\mathscr{S}_n$

since such a solution corresponds to the unique stable and minimum-phase spectral factor $Q(\zeta)$. The key argument is about a nonsingularity condition of a certain matrix $\mathbf{M}(\mathbf{a})$ which will be stated in the next proposition. Recall that a polynomial $a(z)$ is called *unmixing* if it has no reciprocal zeros. In particular, the Schur property implies unmixing.

Proposition 2.3.2. *The solution* $\hat{\mathbf{a}} \in \mathscr{S}_n$ *to the nonlinear equation* (2.30) *defines a periodic ARMA process* (2.14) *whose covariance matrix is a circulant extension of the data* $\mathbf{T}_n$. *Thus such* $\hat{\mathbf{a}}$ *is a minimizer of* (2.34).

Proof. Note that the term $\mathbf{T}_n\hat{\mathbf{a}}$ can be written as $\mathbf{M}(\hat{\mathbf{a}})\mathbf{c}$, where

$$\mathbf{M}(\mathbf{a}) = \begin{bmatrix} a_0 & a_1 & a_2 & \dots & a_n \\ a_1 & a_2 & \dots & a_n & 0 \\ a_2 & \dots & a_n & 0 & 0 \\ \vdots & \iddots & \iddots & & \vdots \\ a_n & 0 & \dots & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & a_0 & 0 & \dots & 0 \\ 0 & a_1 & a_0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & a_{n-1} & \dots & a_1 & a_0 \end{bmatrix}$$

is the so-called *Jury matrix* mentioned in (Demeure and Mullis, 1989) whose determinant is

$$a_0^{n+1} \prod_{j=1}^{n} \prod_{k=1}^{n} (1 - r_j r_k), \tag{2.40}$$

where r_j is the j-th root of the polynomial $a(z)$. Hence $\mathbf{M}(\hat{\mathbf{a}})$ is nonsingular if and only if $a(z)$ is unmixing, in particular if it is a Schur polynomial. Consider then the equation in the unknown $\mathbf{c}$

$$\mathbf{M}(\hat{\mathbf{a}})\mathbf{c} = \mathbf{\Gamma}_n(\hat{\mathbf{a}})\mathbf{b}. \tag{2.41}$$

with $\hat{\mathbf{a}}$ and the corresponding $\gamma = \gamma(\hat{\mathbf{a}})$ fixed. This is a linear equation which has a unique solution vector $\mathbf{c} = \begin{bmatrix} c_0 & \dots & c_n \end{bmatrix}^\top$, whose components are exactly the first $n+1$ covariance lags of the periodic ARMA process (2.14).

The other claim is implied by the covariance matching since by (Lindquist and Picci, 2013, Theorem 2), $\hat{Q}(\zeta) := \hat{a}(\zeta)\hat{a}(\zeta^{-1})$ is then the unique minimizer of the original dual functional parametrized in Q given $P(\zeta) := b(\zeta)b(\zeta^{-1})$. ∎

2.4 Proof of Local Convergence

Taking into account the normalization step (2.33), each iteration in our algorithm can be written as a composition of two maps

$$\mathbf{a}^{(k+1)} = \mathbf{g}(\mathbf{a}^{(k)}) := \mathbf{s}(\mathbf{f}(\mathbf{a}^{(k)})), \tag{2.42}$$

where

$$\mathbf{f}(\mathbf{a}^{(k)}) := \mathbf{T}_n^{-1}\boldsymbol{\Sigma}_n(\mathbf{a}^{(k)})\mathbf{a}^{(k)}. \tag{2.43}$$

The convergence of the iterative algorithm can be studied via the stability analysis of the system (2.42) around its equilibria and we shall do so through linearization. The following well-known theorem is mentioned in (Khalil, 2002, p. 194).

Theorem 2.4.1 (First method of Lyapunov for discrete-time autonomous systems)**.** *Let* $\mathbf{x}^* = 0$ *be an equilibrium of the discrete-time autonomous system*

$$\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k), \tag{2.44}$$

where $\mathbf{f} : \mathscr{D} \to \mathbb{R}^n$ *is continuously differentiable in a neighborhood of the origin* $\mathscr{D} \subset \mathbb{R}^n$*, and let* $J = [\partial \mathbf{f}/\partial \mathbf{x}_k]_{\mathbf{x}_k=\mathbf{0}}$ *be the Jacobian of the system evaluated at the origin. If all the eigenvalues of* J *are strictly less than one in absolute value, then the system is asymptotically stable about its zero equilibrium.*

Before stating the main theorem of this chapter, we need some lemmas in order to compute the Jacobian of the map $\mathbf{g}$ and study its eigenvalues.

Lemma 2.4.2. *Let*

$$\boldsymbol{\Xi}(\mathbf{a}) := \frac{d}{d\mathbf{a}}\left[-\boldsymbol{\Sigma}_n(\mathbf{a})\mathbf{a}\right], \tag{2.45}$$

and then the Hessian of $\mathbb{J}_P(\mathbf{a})$ *is the* $(n+1)\times(n+1)$ *matrix*

$$\nabla^2\mathbb{J}_P(\mathbf{a}) = 2(\mathbf{T}_n + \boldsymbol{\Xi}(\mathbf{a})) \tag{2.46}$$

where, with $\bar{\mathbf{u}}(e^{i\theta})$ *defined as in* (2.36)*,*

$$\boldsymbol{\Xi}(\mathbf{a}) = \int_{-\pi}^{\pi} \bar{\mathbf{u}}(e^{i\theta})\bar{\mathbf{u}}(e^{i\theta})^\top \frac{|b(e^{i\theta})|^2}{a(e^{i\theta})^2} d\nu. \tag{2.47}$$

Proof. Clearly (2.46) follows from (2.38). From the definition of γ_t, it is straightforward to

check that

$$\frac{\partial \gamma_t}{\partial a_j} = -\sum_{k=-N+1}^{N} \zeta_k^{t-j} \frac{b(\zeta_k)}{a(\zeta_k)^2} \frac{1}{2N}.$$

Then consider again the j-th entry of $\mathbf{\Gamma}_n(\mathbf{a})\mathbf{b}$, by interchanging the order of summation,

$$\frac{\partial}{\partial a_k}\left(\sum_{k=0}^{n} b_k \gamma_{-j+k}\right) = -\sum_{\ell=-N+1}^{N} \zeta_\ell^{-j-k} \frac{b(\zeta_\ell^{-1})b(\zeta_\ell)}{a(\zeta_\ell)^2} \frac{1}{2N},$$

which is the (j,k) element of the matrix

$$\frac{d}{d\mathbf{a}}[\mathbf{\Gamma}_n(\mathbf{a})\mathbf{b}].$$

With the help of (2.31), it is also straightforward to verify the matrix form (2.47). ∎

Lemma 2.4.3. *Consider a function defined by the ratio of two polynomials*

$$f(z) = \frac{\sum_{k=0}^{n} d_k z^k}{\sum_{k=0}^{n} a_k z^k},$$

where the denominator polynomial has all its zeros strictly outside the unit circle. Then if f takes real values on the unit circle, we must have $\mathbf{d} = \kappa \mathbf{a}$ for some $\kappa \in \mathbb{R}$, which in turn gives $f(z) = \kappa$ for any $z \in \mathbb{T}$.

Proof. Under the condition of the lemma, $f(z)$ must be holomorphic in some open region of the complex plane containing the closed unit disk $\overline{\mathbb{D}}$. For $z = x + iy$, we can write $f(z) = u(x,y) + iv(x,y)$. Then it is a fact that both u and v are *harmonic functions*, which follows directly from the Cauchy-Riemann equations. By the maximum/minimum principle for harmonic functions (Ahlfors, 1966, p. 164, Theorem 23), the function v can only achieve its maximum and minimum over $\overline{\mathbb{D}}$ at the boundary, i.e., the unit circle, where it is constantly zero. Hence, we have

$$v(z) = 0, \ \forall z \in \overline{\mathbb{D}}.$$

Furthermore, by the Cauchy-Riemann equations, the claim of the lemma follows. ∎

Lemma 2.4.4. *For any $\mathbf{a} \in S_n$, the matrix $\mathbf{\Xi}(\mathbf{a})$ satisfies the inequality*

$$|\mathbf{d}^\top \mathbf{\Xi}(\mathbf{a})\mathbf{d}| \le \mathbf{d}^\top \mathbf{\Sigma}_n(\mathbf{a})\mathbf{d} \tag{2.48}$$

which holds in a strict sense for all $\mathbf{d} = [d_0, \ldots, d_n]^\top \in \mathbb{R}^{n+1}$ except for the one dimensional subspace of vectors which are proportional to $\mathbf{a}$.

Proof. By (2.47) and the triangle inequality,

$$\begin{aligned}|\mathbf{d}^\top \Xi(\mathbf{a})\mathbf{d}| &= \left| \int_{-\pi}^{\pi} \mathbf{d}^\top \bar{\mathbf{u}}(e^{i\theta}) \frac{P(e^{i\theta})}{a(e^{i\theta})^2} \bar{\mathbf{u}}(e^{i\theta})^\top \mathbf{d}\, d\nu \right| \\ &= \left| \int_{-\pi}^{\pi} \frac{P(e^{i\theta})}{a(e^{i\theta})^2}\, d(e^{i\theta})^2\, d\nu \right| \\ &\leq \int_{-\pi}^{\pi} \frac{P(e^{i\theta})}{|a(e^{i\theta})|^2} |d(e^{i\theta})|^2\, d\nu = \mathbf{d}^\top \Sigma_n(\mathbf{a})\mathbf{d},\end{aligned}$$

where $d(z) := \sum_{k=0}^{n} d_k z^{-k}$. This proves the inequality. To prove the other statement, we first note that for any $\mathbf{a} \in \mathscr{A}$ it holds that

$$\Xi(\mathbf{a})\mathbf{a} = \Sigma_n(\mathbf{a})\mathbf{a} \tag{2.49}$$

which readily follows from the representation (2.47) of $\Xi(\mathbf{a})$ since $\bar{\mathbf{u}}(e^{i\theta})^\top \mathbf{a} = a(e^{i\theta})$. To show that (2.48) can hold with equality only for vectors $\mathbf{d} = \kappa\mathbf{a}$, $\kappa \in \mathbb{R}$, we argue as follows. For n complex numbers $z_1, \ldots, z_n$, the condition for the equality

$$|z_1 + \cdots + z_n| = |z_1| + \cdots + |z_n|$$

to hold is that for any j, $z_j = r_j z_0$ with real $r_j \geq 0$ (or all $r_i \leq 0$) and some common $z_0 \in \mathbb{C}$. Applying this to our case, i.e.,

$$z_j = \frac{P(\zeta_j)}{a(\zeta_j)^2} d(\zeta_j)^2, \quad j = -N+1, \ldots, N,$$

given N large enough, it amounts to requiring that the ratio of two polynomials $d(e^{i\theta})/a(e^{i\theta})$ takes real values for any $\theta \in [-\pi, \pi]$ since $d(\zeta_0)/a(\zeta_0)$ is real. By Lemma 2.4.3, this cannot happen in general unless $\mathbf{d}$ is proportional to $\mathbf{a}$. ∎

Theorem 2.4.5. *Algorithm 2.1 converges locally to the vector of AR coefficients $\hat{\mathbf{a}} \in \mathscr{S}_n$ that is a solution to (2.30).*

Proof. Take $\hat{\mathbf{a}}$ as the coefficient vector of the unique Schur polynomial that solves Problem 2.2.1. It is easy to check that $\hat{\mathbf{a}}$ is a fixed point of the function $\mathbf{g}(\cdot) = \mathbf{s}(\mathbf{f}(\cdot))$ since

$$\mathbf{s}(\hat{\mathbf{a}}) = \hat{\mathbf{a}}, \quad \mathbf{f}(\hat{\mathbf{a}}) = \hat{\mathbf{a}}.$$

Next use the representation $\mathbf{f}(\mathbf{a}) = \mathbf{a} - \frac{1}{2}\mathbf{T}_n^{-1} \nabla \mathbb{J}_P(\mathbf{a})$ from Proposition 2.3.1 to compute the

Jacobian matrix by the chain rule and evaluate it at $\hat{\mathbf{a}}$

$$\begin{aligned} J := \left.\frac{d\mathbf{g}}{d\mathbf{a}^{(k)}}\right|_{\mathbf{a}^{(k)}=\hat{\mathbf{a}}} &= \frac{1}{m_P}\hat{\mathbf{a}}\hat{\mathbf{a}}^\top \boldsymbol{\Xi}(\hat{\mathbf{a}}) - \mathbf{T}_n^{-1}\boldsymbol{\Xi}(\hat{\mathbf{a}}) \\ &= \frac{1}{m_P}\hat{\mathbf{a}}\hat{\mathbf{a}}^\top \mathbf{T}_n - \mathbf{T}_n^{-1}\boldsymbol{\Xi}(\hat{\mathbf{a}}) \end{aligned} \tag{2.50}$$

where we have also used Lemma 2.4.2 to get

$$\nabla \mathbf{f}(\mathbf{a}) = -\mathbf{T}_n^{-1}\boldsymbol{\Xi}(\mathbf{a}).$$

The last equality in (2.50) comes from (2.49).

In order to apply Theorem 2.4.1 to assert stability of the equilibrium $\hat{\mathbf{a}}$, we proceed to show that all eigenvalues of (2.50) have absolute value smaller than 1. To this end, there is a result in linear algebra stating that the spectral radius of a complex matrix A is less than 1 if and only if $A^k \to 0$ as $k \to \infty$, whose proof uses Jordan normal form (Horn and Johnson, 2013, p. 180). One can then show by induction that for $k = 1, 2, \ldots$

$$\begin{aligned} J^{2k} &= -\frac{1}{m_P}\hat{\mathbf{a}}\hat{\mathbf{a}}^\top \mathbf{T}_n + \left[\mathbf{T}_n^{-1}\boldsymbol{\Xi}(\hat{\mathbf{a}})\right]^{2k}, \\ J^{2k+1} &= \frac{1}{m_P}\hat{\mathbf{a}}\hat{\mathbf{a}}^\top \mathbf{T}_n - \left[\mathbf{T}_n^{-1}\boldsymbol{\Xi}(\hat{\mathbf{a}})\right]^{2k+1}. \end{aligned} \tag{2.51}$$

Therefore, it is equivalent to show that

$$\lim_{k\to\infty} \left[\mathbf{T}_n^{-1}\boldsymbol{\Xi}(\hat{\mathbf{a}})\right]^k = \frac{1}{m_P}\hat{\mathbf{a}}\hat{\mathbf{a}}^\top \mathbf{T}_n. \tag{2.52}$$

Naturally, we consider the eigenvalue problem

$$\mathbf{T}_n^{-1}\boldsymbol{\Xi}(\hat{\mathbf{a}})\mathbf{v} = \lambda\mathbf{v}, \quad \mathbf{v} \neq 0.$$

It is equivalent to the generalized eigenvalue problem of the ordered pair $(\boldsymbol{\Xi}(\hat{\mathbf{a}}), \mathbf{T}_n)$

$$\boldsymbol{\Xi}(\hat{\mathbf{a}})\mathbf{v} = \lambda\mathbf{T}_n\mathbf{v}.$$

From (Parlett, 1998, Theorem 15.3.3, p. 345), in the present case where $\boldsymbol{\Xi}(\hat{\mathbf{a}})$ and $\mathbf{T}_n$ are symmetric with $\mathbf{T}_n$ positive definite, all the eigenvalues λ are real and it is guaranteed that there exists a basis of generalized eigenvectors. Moreover, eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ with distinct eigenvalues are $\mathbf{T}_n$-orthogonal ($\mathbf{v}_1^\top \mathbf{T}_n \mathbf{v}_2 = 0$). The eigenvalues can be expressed in terms of

the eigenvectors as Rayleigh quotients

$$\lambda = \frac{\mathbf{v}^\top \Xi(\hat{\mathbf{a}})\mathbf{v}}{\mathbf{v}^\top \mathbf{T}_n \mathbf{v}}. \tag{2.53}$$

By Lemma 2.4.4, we must have $|\lambda| \leq 1$. Furthermore, there is exactly one (generalized) eigenvalue equal to 1 since we have taken the corresponding $\hat{a}(z) \in \mathcal{S}_n$.

We are now ready to show (2.52). First notice that

$$\Xi(\hat{\mathbf{a}})\mathbf{P} = \mathbf{T}_n \mathbf{P}\mathbf{D}, \tag{2.54}$$

where the columns of $\mathbf{P}$ are the eigenvectors and $\mathbf{D}$ is the diagonal matrix of eigenvalues. The $\mathbf{T}_n$-orthogonal relation can be written as

$$\mathbf{P}^\top \mathbf{T}_n \mathbf{P} = \mathbf{I}, \tag{2.55}$$

where the eigenvectors are normalized with the $\mathbf{T}_n$-norm. Specifically, the eigenvector corresponding to the eigenvalue 1 is $\frac{1}{\sqrt{m_P}}\hat{\mathbf{a}}$. Hence,

$$\lim_{k\to\infty} \left[\mathbf{T}_n^{-1}\Xi(\hat{\mathbf{a}})\right]^k = \lim_{k\to\infty} \mathbf{P}\mathbf{D}^k\mathbf{P}^\top\mathbf{T}_n = \frac{1}{m_P}\hat{\mathbf{a}}\hat{\mathbf{a}}^\top\mathbf{T}_n, \tag{2.56}$$

which concludes the proof of local convergence. ■

2.5 Generalization to Vector Processes

The nonlinear Yule-Walker equation (2.30) and Algorithm 2.1 can be easily generalized to their counterparts for vector-valued processes. Unfortunately, the convergence proof in the vector case seems quite difficult and the linearization argument does not seem to work in a straightforward manner due to some redundant parametrization in the matrix coefficients. So far, from our numerical simulations, the algorithm works well when the MA part is restricted to be scalar, which is the case in the next section for the application to finite-interval smoothing. In general when the MA coefficients are matrix-valued, the algorithm can converge to different values of AR coefficients. Hence some work remains to be done in order to remedy this. We will use some overload of notations whose meanings are clear from the context.

Let us begin with an m-dimensional zero-mean stationary periodic process $y(t)$ that is

described by a forward unilateral ARMA representation

$$\sum_{k=0}^{n} A_k\, y(t-k) = \sum_{k=0}^{n} B_k\, w(t-k), \quad t \in \mathbb{Z}_{2N}, \tag{2.57}$$

where $w(t)$ is an m-dimensional periodic normalized white noise, i.e., $\mathbb{E}\, w(t)w(t)^\top = I_m$. With this convention the model is also called *normalized*. Naturally it must also be associated to periodic boundary conditions (2.8) at the end points. With the stacked vector notation

$$\mathbf{y} = \begin{bmatrix} y(-N+1) \\ \vdots \\ y(N) \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} e(-N+1) \\ \vdots \\ e(N) \end{bmatrix} \in \mathbb{R}^{2mN}, \tag{2.58}$$

the model (2.57) can be written compactly as a linear equation

$$\mathbf{A}\mathbf{y} = \mathbf{B}\mathbf{w}, \tag{2.59}$$

in which $\mathbf{A}$ and $\mathbf{B}$ are $2mN \times 2mN$ *block-circulant* matrices of bandwidth n, that is

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} A_0 & 0 & \cdots & 0 & A_n & \cdots & A_1 \\ A_1 & A_0 & 0 & \cdots & 0 & \cdots & A_2 \\ \vdots & \vdots & \ddots & & \vdots & & \vdots \\ A_n & A_{n-1} & \cdots & A_0 & 0 & \cdots & 0 \\ 0 & A_n & A_{n-1} & & & & \\ \vdots & \vdots & \ddots & \ddots & & \ddots & \vdots \\ 0 & 0 & \cdots & A_n & \cdots & A_1 & A_0 \end{bmatrix} \\ &:= \mathrm{Circ}\{A_0, A_1, \ldots, A_n, 0, \ldots, 0\} \end{aligned} \tag{2.60}$$

and similarly

$$\mathbf{B} = \mathrm{Circ}\{B_0, B_1, \ldots, B_n, 0, \ldots, 0\}. \tag{2.61}$$

We require the condition that $\mathbf{A}$ is invertible. Define also the matrix polynomials

$$A(\zeta) := \sum_{k=0}^{n} A_k\, \zeta^{-k}, \quad B(\zeta) := \sum_{k=0}^{n} B_k\, \zeta^{-k}, \tag{2.62}$$

which are symbols of the block circulants $\mathbf{A}$ and $\mathbf{B}$. Then similar spectral analysis of the vector ARMA model (2.57) can be carried out as those in Section 2.2. Specifically, the solution of

(2.57) can formally be written in the Fourier and time domains as

$$\hat{y}(\zeta) = A(\zeta)^{-1}B(\zeta)\hat{w}(\zeta) \iff \mathbf{y} = \mathbf{A}^{-1}\mathbf{B}\mathbf{w}. \tag{2.63}$$

The discrete transfer function $W(\zeta) := A(\zeta)^{-1}B(\zeta)$ has the inverse Fourier transform

$$W_t := \sum_{k=-N+1}^{N} \zeta_k^t A(\zeta_k)^{-1}B(\zeta_k)\frac{1}{2N}, \quad t \in \mathbb{Z}_{2N} \tag{2.64}$$

called the *impulse response* of the system, which yields a convolution representation of the process

$$y(t) = \sum_{s=-N+1}^{N} W_{t-s}\, w(s), \quad t \in \mathbb{Z}_{2N}. \tag{2.65}$$

This is certainly equivalent to the block-circulant matrix equation in (2.63). The spectral density of the process (2.57) has the form of a bilateral matrix fraction

$$\begin{aligned} \Phi(\zeta) &= W(\zeta)W(\zeta^{-1})^\top \\ &= A(\zeta)^{-1}B(\zeta)B(\zeta^{-1})^\top A(\zeta^{-1})^{-\top}, \qquad \zeta \in \mathbb{T}_{2N} \end{aligned} \tag{2.66}$$

which has an isomorphic counterpart in terms of block circulants

$$\mathbf{\Sigma} = \mathbf{W}\mathbf{W}^\top = \mathbf{A}^{-1}\mathbf{B}\mathbf{B}^\top\mathbf{A}^{-\top}, \tag{2.67}$$

where $\mathbf{\Sigma} := \mathbb{E}\,\mathbf{y}\mathbf{y}^\top$ is the (block-circulant) covariance matrix, and $\mathbf{W}$ is the block circulant having symbol $W(\zeta)$, namely

$$\mathbf{W} = \mathrm{Circ}\{W_0, W_1, \ldots, W_N, W_{-N+1}, \ldots, W_{-1}\}. \tag{2.68}$$

Later on we shall need to express the covariance matrix $\mathbf{\Sigma}$ as a single matrix fraction of the type (2.10). This operation is easy when $B(z)$ is a scalar polynomial times the identity matrix, i.e., $B(z) = b(z)I_m$, in which case $W(\zeta)$ admits a fraction representation of the type $A(\zeta)^{-1}b(\zeta)I_m$ so that

$$\Phi(\zeta) = \left[A(\zeta^{-1})^\top A(\zeta)\right]^{-1} b(\zeta)b(\zeta^{-1})$$

and one may take $Q(\zeta) = A(\zeta^{-1})^\top A(\zeta)$ and $P(\zeta) = b(\zeta)b(\zeta^{-1})I_m$.

2.5.1 Problem of Matrix Covariance Matching

We are now ready to formulate the *ARMA covariance matching problem* for vector processes.

Problem 2.5.1. Suppose we are given $n+1$ matrix MA parameters $\{B_k\}$ in the unilateral ARMA model (2.57) and $n+1$ real $m \times m$ matrices $\Sigma_0, \Sigma_1, \ldots, \Sigma_n$, such that the block-Toeplitz matrix

$$\mathbf{T}_n = \begin{bmatrix} \Sigma_0 & \Sigma_1^\top & \Sigma_2^\top & \cdots & \Sigma_n^\top \\ \Sigma_1 & \Sigma_0 & \Sigma_1^\top & \cdots & \Sigma_{n-1}^\top \\ \Sigma_2 & \Sigma_1 & \Sigma_0 & \cdots & \Sigma_{n-2}^\top \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Sigma_n & \Sigma_{n-1} & \Sigma_{n-2} & \cdots & \Sigma_0 \end{bmatrix}, \qquad n \in \mathbb{Z}_+ \tag{2.69}$$

is positive definte. We want to determine the matrix parameters $\{A_k,\ k = 0, 1, \ldots, n\}$ such that the first $n+1$ covariance matrices of the periodic process $y(t)$ match the sequence $\{\Sigma_k\}$.

When $B_1 = B_2 = \ldots = B_n = 0$, that is, the MA part is trivial, this is essentially the *Covariance Extension Problem for n-Reciprocal Processes* discussed and solved in (Carli et al., 2011). The solution of this problem when the MA part is scalar, stated in terms of symmetric polynomials $Q(\zeta)$, $P(\zeta)$ under the criterion of generalized maximum entropy, is discussed in (Lindquist et al., 2013). Here for reasons that will be apparent in a moment our unknowns will instead be the coefficients of some distinguished spectral factor of the denominator $Q(\zeta)$. An important point is that to get a unique solution $Q(\zeta)$ one needs to fix the numerator polynomial $P(\zeta)$. Clearly to fix the MA part is equivalent to fixing $P(\zeta)$ and this explains the problem formulation with fixed MA parameters $\{B_k\}$.

Observe that Problem 2.5.1 is just asking that the submatrix made of the upper-left $(n+1) \times (n+1)$ blocks extracted from the block circulant covariance matrix $\mathbf{\Sigma}$ in (2.67) should match the Toeplitz data (2.69). To make this precise we need to fix some notations. Let $A, B \in \mathbb{R}^{m \times m(n+1)}$ denote the AR and MA matrix coefficients of the vector ARMA model (2.57), i.e.,

$$A = \begin{bmatrix} A_0 & A_1 & \ldots & A_n \end{bmatrix}, \quad B = \begin{bmatrix} B_0 & B_1 & \ldots & B_n \end{bmatrix},$$

and let us use the symbol $\mathbf{\Sigma}_n(A)$ to denote the upper-left $(n+1) \times (n+1)$ block-submatrix of $\mathbf{\Sigma}$ written as a function of the unknown AR parameters since B is fixed. The same notation applies to the matrix $\mathbf{W}$ (2.68), namely

$$\mathbf{W}_n(A) := \begin{bmatrix} W_0 & W_{-1} & \ldots & W_{-n} \\ W_1 & W_0 & \ldots & W_{-n+1} \\ \vdots & \vdots & \ddots & \vdots \\ W_n & W_{n-1} & \ldots & W_0 \end{bmatrix}.$$

Moreover, $\mathbf{T}_n^\perp$ is the matrix obtained by taking the transpose of each block in $\mathbf{T}_n$. Obviously, the operation $(\cdot)^\perp$ applies to any block matrix with square blocks of the same size. The

covariance matching condition leads to the equation

$$\mathbf{T}_n = \mathbf{\Sigma}_n(A)\,. \tag{2.70}$$

Proposition 2.5.2. *The covariance matching condition* (2.70) *implies the equation*

$$A\mathbf{T}_n^{\perp} = B\mathbf{W}_n^{\perp}(A). \tag{2.71}$$

Proof. Equation (2.71) can be proven directly by a Yule-Walker-type calculation. Specifically, combining the model equation (2.57) with the convolution representation (2.65) and using the relation $\mathbb{E}\,w(t-k)y(t-j)^{\top} = W_{k-j}^{\top}$, we easily obtain

$$\sum_{k=0}^{n} A_k \Sigma_{j-k} = \sum_{k=0}^{n} B_k\, W_{k-j}^{\top}, \quad j = 0, 1, \ldots, n, \tag{2.72}$$

which is equivalent to (2.71). ∎

With the MA coefficients B and the covariance data $\mathbf{T}_n$ fixed, (2.71) is a nonlinear equation in the unknown A. The counterpart of the constraint (2.32) for the AR coefficient matrix A solving the matching problem is

$$A\mathbf{T}_n^{\perp}A^{\top} = \sum_{k=0}^{n} B_k B_k^{\top} = \frac{1}{2N} \sum_{k=-N+1}^{N} B(\zeta_k)B(\zeta_k)^{*} := M_P\,, \tag{2.73}$$

where M_P is a constant matrix, and the second equality comes from a matrix DFT analog of the Parseval Formula (A.5) (see Appendix A). The first equality can be seen as a part of the circulant identity $\mathbf{A\Sigma A}^{\top} = \mathbf{BB}^{\top}$.

The structure of (2.71) suggests a natural iterative scheme for the solution, namely:

$$A^{(k+1)} = B\mathbf{W}_n^{\perp}(A^{(k)})(\mathbf{T}_n^{\perp})^{-1}, \tag{2.74}$$

with $A^{(0)}$ initialized e.g., with the output of the Levinson-Whittle algorithm (Whittle, 1963) for the data $\{\Sigma_k\,;\, k = 0, 1, \ldots, n\}$. After each iteration, the new $A^{(k+1)}$ does not necessarily satisfy the normalization constraint (2.73) and thus needs to be scaled by a suitable map

$$S:\ A \mapsto K^{-1}A,$$

for some $m \times m$ nonsingular matrix K. This amounts to solving for K the matrix equation

$$A\mathbf{T}_n^{\perp}A^{\top} = KM_PK^{\top}\,, \tag{2.75}$$

which can be done by various methods such as Cholesky factorization. The corresponding algorithm in the present context is summarized below.

Algorithm 2.2 Fixed-point iteration with renormalization: vector case

Initialize $A^{(0)}$
Set $k = 0$ and a threshold τ to decide convergence
repeat
  Do the iteration (2.74)
  Solve (2.75) with $A \equiv A^{(k+1)}$ for the scaling matrix K
  Set $A^{(k+1)} := K^{-1}A^{(k+1)}$
  Update $k := k+1$
until $\|A^{(k)} - A^{(k-1)}\| \le \tau$
return the last $A^{(k)}$

2.6 Smoothing of Stationary Linear Systems with Boundary Constraints

Consider the following problem. We have a wide-sense stationary zero-mean vector signal $x(t)$ observed on the finite interval $[-N+1, N]$, the observation channel being described by the linear equation

$$y(t) = Cx(t) + v(t), \qquad t \in [-N+1, N] \tag{2.76}$$

where $v(t)$ is a stationary white noise with a known variance matrix $R = R^\top > 0$, independent of $x(t)$. We want to compute the smoothed estimate $\hat{x}(t)$ given a finite chunk of observations,

$$\hat{x}(t) := \mathbb{E}\{x(t) \mid y(s),\ s \in [-N+1, N]\}, \quad t \in [-N+1, N]. \tag{2.77}$$

The right-hand side of (2.77) is the orthogonal projection onto the Hilbert space of random variables spanned by the components of $\{y(s),\ s \in [-N+1, N]\}$. We shall assume that the process $x(t)$ has a (stationary) *periodic extension* to the whole integer line $\mathbb{Z}$. Equivalently $x(t)$ can be imagined to be the restriction to the interval $[-N+1, N]$ of a periodic stationary process defined on $\mathbb{Z}$. There are estimates in the literature (Dembo, Mallows, and Shepp, 1989) for how large should this N be. For short, we shall call $x(t)$ a *periodic process* and think of it as being defined on the finite modular group $\mathbb{Z}_{2N}$. Even if we do not care about the extension, which we are never going to see, this apparently innocent assumption (which is obviously always legitimate for deterministic signals on finite intervals) has important consequences. In particular, the covariance matrix of the finite string $\{x(t);\ t \in [-N+1, N]\}$ must be *block-circulant* and automatically, $x(t)$ is associated to *periodic boundary conditions*

at the extremes.

We shall first discuss the problem of finding a periodic model (2.57) defined on the finite discrete interval $[-N+1, N]$ which approximates in a suitable sense a given Gauss-Markov stationary model defined on $\mathbb{Z}$. Consider a stationary signal $x(t)$ given as the output of the state-space model

$$\begin{cases} \xi(t+1) & = F\xi(t) + Hu(t) \\ x(t) & = G\xi(t) + Ju(t) \end{cases} \tag{2.78}$$

where $u(t)$ is a normalized white noise. We assume without loss of generality that the Lyapunov equation $\Pi = F\Pi F^\top + HH^\top$ for the variance matrix of $\xi(t)$ has a unique positive definite solution. Define $D := F\Pi G^\top + HJ^\top$ and let

$$\begin{cases} \Sigma_0 & := G\Pi G^\top + JJ^\top \\ \Sigma_k & := GF^{k-1}D\,; \qquad k = 1, \dots, n \end{cases} \tag{2.79}$$

be the string of the first $n+1$ output covariance matrices. One needs to provide a set of $n+1$ MA coefficients, or equivalently a positive polynomial $P(\zeta)$ to fix the zero dynamics of the system. One possible way to do this is to approximate in the DFT domain the numerator polynomial of the model (2.78) or use estimates of its *cepstral coefficients* (see (Byrnes et al., 2001c; Enqvist, 2004; Lindquist and Picci, 2013)). Form with the data $\{\Sigma_k\}$ the block-Toeplitz matrix (2.69); then the periodic ARMA approximation can be computed by running the covariance matching algorithm described in the previous section.

With the identified approximate model (2.57) at hand, one can now proceed to compute the solution of the smoothing problem. The procedure is inspired by that for reciprocal processes described in (Levy, Frezza, and Krener, 1990, Section VI). Write the observation equation (2.76) in vector notation as

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{v},$$

where $\mathbf{C} = \operatorname{diag}\{C, \dots, C\}$. Then use the standard one-shot solution for the minimum variance Bayesian estimate $\hat{\mathbf{x}}$ (Lindquist and Picci, 2015, p. 29) to get the relation

$$(\mathbf{\Sigma}^{-1} + \mathbf{C}^\top \mathbf{R}^{-1}\mathbf{C})\hat{\mathbf{x}} = \mathbf{C}^\top \mathbf{R}^{-1}\mathbf{y}, \tag{2.80}$$

Substituting (2.10) into the above equation, the matrix on the left-hand side becomes

$$\mathbf{P}^{-1}\mathbf{Q} + \mathbf{C}^\top \mathbf{R}^{-1}\mathbf{C} = \mathbf{P}^{-1}(\mathbf{Q} + \mathbf{P}\mathbf{C}^\top \mathbf{R}^{-1}\mathbf{C}). \tag{2.81}$$

Then define

$$\hat{\mathbf{Q}} := \mathbf{Q} + \mathbf{P}\mathbf{C}^\top \mathbf{R}^{-1}\mathbf{C}, \tag{2.82}$$

which is a positive-definite block-circulant since $\mathbf{C}^\top\mathbf{R}^{-1}\mathbf{C}$ is a block-diagonal matrix with positive-semidefinite blocks and the symbol of $\mathbf{P}$ is essentially scalar. In fact, $\hat{\mathbf{Q}}$ is bilaterally banded of bandwidth n since such are both summands on the right-hand side of (2.82). Then (2.80) is equivalent to

$$\hat{\mathbf{Q}}\hat{\mathbf{x}} = \mathbf{P}\mathbf{C}^\top\mathbf{R}^{-1}\mathbf{y} := \hat{\mathbf{y}}. \tag{2.83}$$

In order to carry out a two-sweep smoothing procedure in the style of the Rauch-Striebel-Tung smoother (Rauch, Striebel, and Tung, 1965), we first perform a banded matrix factorization $\hat{\mathbf{Q}} = \hat{\mathbf{A}}\hat{\mathbf{A}}^\top$, where

$$\hat{\mathbf{A}} = \mathrm{Circ}\{\hat{A}_0, \hat{A}_1, \ldots, \hat{A}_n, 0, \ldots, 0\}. \tag{2.84}$$

As discussed in Proposition 2.2.3, such a factorization is possible if N is taken large enough, and it can be computed in the spectral domain by standard matrix polynomial factorization algorithms (see e.g., (Rissanen, 1973)). Then given $\hat{\mathbf{A}}$ and $\hat{\mathbf{y}}$, to compute the solution to (2.83) we first perform a forward sweep described by

$$\hat{\mathbf{A}}\mathbf{z} = \hat{\mathbf{y}}, \tag{2.85}$$

and then a backward sweep

$$\hat{\mathbf{A}}^\top\hat{\mathbf{x}} = \mathbf{z}. \tag{2.86}$$

The two sweeps can be implemented by a forward and a backward recursive algorithm described by unilateral AR models. To this end we need to attach to them explicit boundary values $\hat{x}(-N+1), \hat{x}(-N+2), \ldots, \hat{x}(-N+n)$ and $\hat{x}(N-n+1), \ldots, \hat{x}(N)$ extracted from the process $\hat{x}(t)$, which we assume are given. Due to the banded block-circulant structure of $\hat{\mathbf{A}}$ exactly like (2.60), the first equation of the forward sweep can be written as

$$\hat{A}_0 z(-N+1) = -\sum_{i=1}^{n} \hat{A}_i z(N-i+1) + \hat{y}(-N+1), \tag{2.87}$$

which needs to be initialized with the boundary values $z(N-n+1)$, $z(N-n+2)$, $\ldots, z(N)$. These values can be obtained by solving for $\mathbf{z}$ the last n block equations in the backward sweep (2.86) since only the boundary values at two ends of $\hat{\mathbf{x}}$ are involved there due to the banded block-circulant structure of $\hat{\mathbf{A}}^\top$.

The forward sweep starts by computing the boundary values $z(N-n+1), \ldots, z(N)$ as described above. After these n endpoints of $\mathbf{z}$ are available, the recursion for $\mathbf{z}$ can be

implemented by the scheme

$$z(t) = \hat{A}_0^{-1}\left[\hat{y}(t) - \sum_{i=1}^{n} \hat{A}_i z(t-i)\right], \quad t \in [-N+1, N-n]. \tag{2.88}$$

One should notice that in this notation, we impose implicitly that $z(-N) = z(N), \ldots, z(-N-n+1) = z(N-n+1)$. The backward sweep then proceeds by

$$\hat{x}(t) = \hat{A}_0^{-\top}\left[z(t) - \sum_{i=1}^{n} \hat{A}_i^{\top} \hat{x}(t+i)\right], \; t \in [-N+n+1, N-n], \tag{2.89}$$

which is initialized with the known terminal boundary values $\hat{x}(N-n+1), \hat{x}(N-n+2), \ldots, \hat{x}(N)$.

There is also a dual factorization which will lead to a backward-forward sequence of sweeps but we shall not insist on this point.

2.6.1 A Numerical Example

Just to show the feasibility of the method, we shall discuss a toy example. For this particular example we have chosen scalar MA coefficients resulting in obvious computational advantage for the smoothing algorithm. We should stress that this example is not meant to reflect any realistic situation. Referring to model (2.78), fix the matrices

$$F = \begin{bmatrix} 0.9 & -0.3 \\ 0.3 & 0.9 \end{bmatrix}, \quad G = \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix} \tag{2.90}$$

and H, J equal to identity. For the observation process (2.76), we take

$$C = \begin{bmatrix} 1 & 1 \end{bmatrix}.$$

The eigenvalues of A are $0.9 \pm 0.3i$ with modulus 0.9487.

To compute the smoothed process (2.77), we first build a periodic ARMA model of order $n = 1$ to approximately describe the process $x(t)$ on a finite interval by matching the first two steady-state covariances

$$\Sigma_0 = \begin{bmatrix} 51 & 10 \\ 10 & 11 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 46 & 17 \\ 4 & 9 \end{bmatrix}$$

computed with the formulae (2.79). The period of interest is set as $2N = 50$ and the MA parameters are chosen (quite arbitrarily) as $b_0 = 0.4893$, $b_1 = 0.3377$. The unilateral ARMA

model looks like

$$A_0 x(t) + A_1 x(t-1) = b_0 w(t) + b_1 w(t-1), \tag{2.91}$$

and the AR parameters

$$A_0 = \begin{bmatrix} 0.3725 & 0 \\ 0.1324 & 0.3341 \end{bmatrix}, \; A_1 = \begin{bmatrix} -0.2571 & -0.3579 \\ -0.0739 & -0.3659 \end{bmatrix}$$

are computed with Algorithm 2.2. The resulting poles of (2.91), i.e., roots of the equation $\det A(z) = 0$ are $0.7022 \pm 0.2240i$ of modulus 0.7371.

Given the approximate model (2.91) and the observation process (2.76), the two-sweep smoothing algorithm described in the previous part can be implemented. The two components of the smoothed process $\hat{x}(t)$ are shown in Figures 2.1 and 2.2. The effect of smoothing is appreciable.

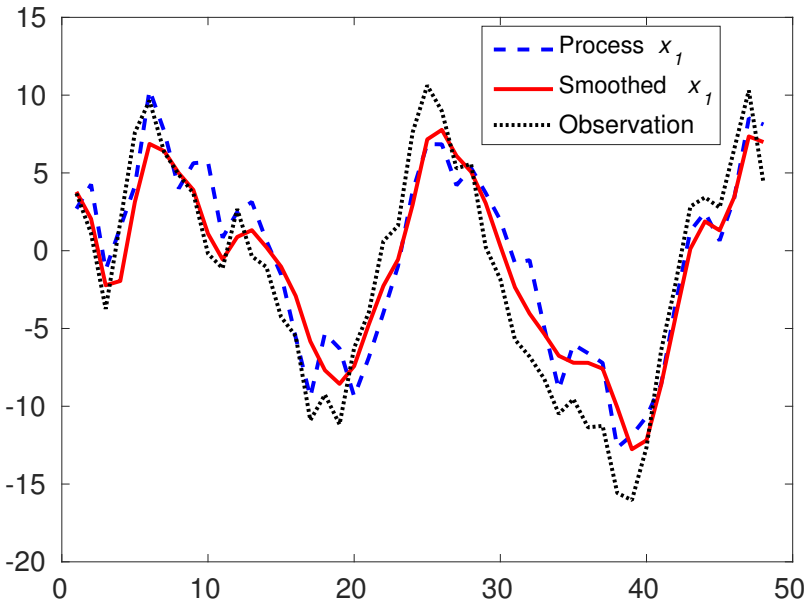

Figure 2.1: Result of smoothing for x_1

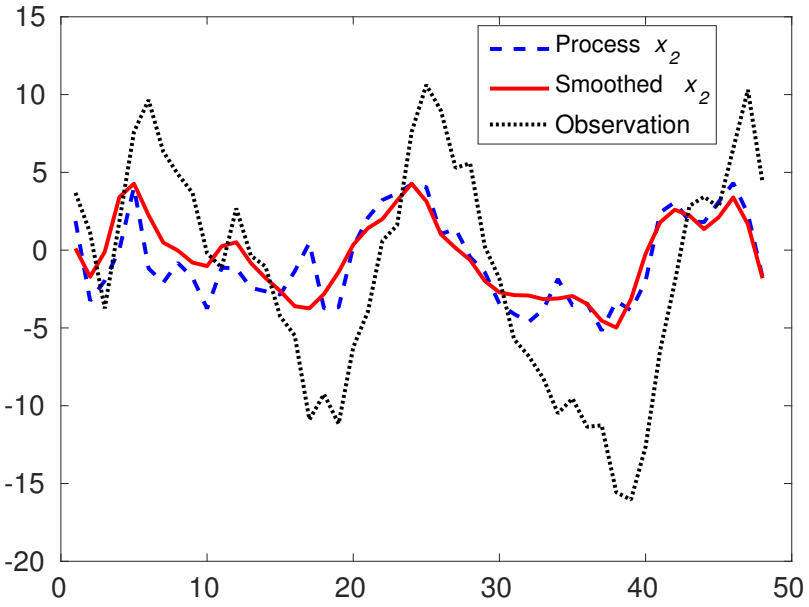

Figure 2.2: Result of smoothing for x_2

2.7 Conclusion

We have proved local convergence of a new iterative algorithm to solve the covariance matching problem for scalar periodic ARMA models. The algorithm can be seen as a nonlinear dynamical system that is asymptotically stable about its equilibrium. From our simulations *global* convergence to a solution of (2.29) appears to be true but a proof is left to future work.

The algorithm has been extended to deal with the covariance matching problem for vector ARMA processes on a finite interval. As an application, this algorithm is used to construct an approximate stationary periodic model for an original underlying process based on matching a finite number of covariance lags. With the approximate model, we have indicated how to solve a class of finite-interval smoothing problems subjected to boundary constraints. However, some open questions remain in the vector case such as the effectiveness of Algorithm 2.2 when MA parameters are matrix-valued as well as the issue of convergence.

3

Further Results on a Parametric Multivariate Spectral Estimation Problem

3.1 Introduction

This chapter concerns a multivariate spectral estimation problem subject to a (generalized) moment constraint. As explained in Chapter 1, the common approach in the literature is to find one particular solution to the moment equation that extremizes a certain criterion. Here we shall take a different route following the lines in (Ferrante et al., 2010). The idea is to restrict the candidate solution to a parametric family of spectral densities, in which each spectral density function is uniquely determined by a finite-dimensional parameter. Two fundamental questions arise in this formulation. First, does a solution of the moment equation exist in such a family? If so, the second question is whether such a solution is unique (in the predefined family).

Before addressing these questions, we wish to point out that the particular parametric form of matrix spectral densities introduced in (Ferrante et al., 2010) generalizes the scalar solution in (Georgiou and Lindquist, 2003) of a constrained Kullback-Leibler spectrum approximation problem. The optimization approach in the latter can be extended to the multivariate case provided that the given prior spectral density is still kept as scalar, as reported in (Avventi, 2011a). However, it is quite unnatural to use a scalar prior for matrix spectral densities, and

out of this consideration, the parametrization proposed in (Ferrante et al., 2010) includes a matrix-valued prior. Moreover, such a parametric problem is closely related to ARMA modeling of vector-valued stationary processes subject to covariance matching, which is a multivariate generalization of the rational covariance extension problem.

In order to answer the questions of existence and uniqueness, in (Ferrante et al., 2010) a moment map between two finite-dimensional spaces (of the same dimension) was defined sending a parameter to generalized moments. Then the map was studied in light of a Hadamard-type global inverse function theorem (Byrnes and Lindquist, 2007). However, the result in (Ferrante et al., 2010) was not satisfactory because the authors only showed that a solution exists when the prior spectral density has a very special structure. In fact, this is the motivation behind the current chapter. As a continuation of the work in (Ferrante et al., 2010), some new developments will be reported here.

We show first that the parametric spectral estimation problem is well-posed given a scalar prior, which implies existence and uniqueness of the solution (identifiability), thus a complement to the results in (Avventi, 2011a). Moreover, we prove that the unique solution parameter depends also continuously on the prior function under a suitable metric topology. One important technical tool for the proofs of well-posedness is the global inverse function theorem of Hadamard that can be found e.g., in (Gordon, 1972). We wish to point out that continuous dependence of the solution on data does not seem to have attracted much attention in multivariate formulations of the spectral estimation problem, unlike the scalar rational covariance extension problem (Byrnes et al., 1995, 1997, 1998, 2001b, 2002), where such continuity argument is actually part of the results of well-posedness. We mention (Ramponi et al., 2010), where continuous dependence of the solution on the covariance matrix has been shown in the context of optimization with the Hellinger distance.

Then we present an existence result for the parametric problem under *any* fixed matrix-valued prior that is bounded and coercive. The important special case of covariance extension is addressed in connection with vector ARMA modeling. The main machinery behind our existence proof is the *topological degree theory* from nonlinear analysis. As a historical remark, Georgiou was the first to apply the degree theory to rational covariance extension (Georgiou, 1983, 1987a,b) to show existence of a solution, and it was further developed by Byrnes, Lindquist, and coworkers (Byrnes et al., 1995) to prove the uniqueness and well-posedness. These theories were established before the discovery of the cost function in the optimization framework (Byrnes et al., 1998, 2001b,a), which was later called generalized entropy criterion.

Later in this chapter, we try to approach the question of uniqueness thanks to the introduction of a diffeomorphic spectral factorization. Specifically, we show well-posedness when

the prior is scalar times a constant positive definite matrix. This is still a preliminary result and much more work is needed to deal with the general case.

The outline of this chapter is as follows. In Section 3.2, we review the problem formulation and in particular, give a parametric family of matrix spectral densities. Section 3.3 contains results on the well-posedness of the problem when the prior has a special structure of scalar times identity. One of our main results, existence of a solution under an arbitrary matrix prior, is presented in Section 3.4. A part of the degree theory is reviewed in order to carry out our proof. A spectral factorization problem is discussed in Section 3.5, whose result will be useful for the development in Section 3.6, where some preliminary results on the uniqueness of the solution are provided. In the end, we conclude with some open questions.

3.2 Parametric Formulation of a Multivariate Spectral Estimation Problem

Consider a linear system with a state-space representation

$$x(t+1) = Ax(t) + By(t), \tag{3.1}$$

where $A \in \mathbb{C}^{n\times n}$ is Schur stable, i.e., has all its eigenvalues in $\mathbb{D}$, $B \in \mathbb{C}^{n\times m}$ is of full column rank ($n \geq m$). Moreover, the pair (A,B) is assumed to be *reachable*. The input process $y(t)$ is zero-mean wide-sense stationary with an unknown spectral density matrix $\Phi(z)$. The transfer function of (3.1) is just

$$G(z) = (zI - A)^{-1}B, \tag{3.2}$$

which can be interpreted as a bank of filters. An estimate of the steady-state covariance matrix $\Sigma := \mathbb{E}\, x(t)x(t)^*$ of the state vector $x(t)$ is assumed to be known. (For the problem of estimating covariance matrices in this setting, we refer to (Zorzi and Ferrante, 2012; Ferrante et al., 2012b; Ning et al., 2013)). Hence we have

$$\int G\Phi G^* = \Sigma, \tag{3.3}$$

where the function is integrated on $\mathbb{T}$ with respect to the normalized Lebesgue measure $\frac{d\theta}{2\pi}$. This notation will be adopted throughout in the present and later chapters.

Given the matrix $\Sigma \in \mathfrak{H}_{+,n}$, we want to estimate the spectral density Φ such that the generalized moment constraint (3.3) is satisfied. For example, consider the following choice

of the matrix pair (A, B):

$$A = \begin{bmatrix} 0 & I_m & 0 & \cdots & 0 \\ 0 & 0 & I_m & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & I_m \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ I_m \end{bmatrix}. \tag{3.4}$$

Here each block in A or B is of $m \times m$ and A is a $(p+1) \times (p+1)$ block matrix while B is a $(p+1)$-block column vector. It is easy to verify that in this case

$$G(z) = (zI - A)^{-1}B = \begin{bmatrix} z^{-p-1}I_m \\ z^{-p}I_m \\ \vdots \\ z^{-1}I_m \end{bmatrix}, \tag{3.5}$$

Symbolically, the steady state vector

$$x(t) = G(z)y(t) = \begin{bmatrix} y(t-p-1) \\ \vdots \\ y(t-2) \\ y(t-1) \end{bmatrix}, \tag{3.6}$$

and the covariance matrix Σ has a block-Toeplitz structure, i.e.,

$$\Sigma = \begin{bmatrix} \Sigma_0 & \Sigma_1^* & \Sigma_2^* & \cdots & \Sigma_p^* \\ \Sigma_1 & \Sigma_0 & \Sigma_1^* & \cdots & \Sigma_{p-1}^* \\ \Sigma_2 & \Sigma_1 & \Sigma_0 & \cdots & \Sigma_{p-2}^* \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \Sigma_p & \Sigma_{p-1} & \cdots & \Sigma_1 & \Sigma_0 \end{bmatrix}, \tag{3.7}$$

where $\Sigma_k := \mathbb{E}\, y(t+k)y(t)^* \in \mathbb{C}^{m\times m}$ with a slight abuse of notation.[1] In fact, the constraint (3.3) is equivalent to the set of *moment equations*

$$\int_{-\pi}^{\pi} e^{ik\theta}\Phi(e^{i\theta})\frac{d\theta}{2\pi} = \Sigma_k, \quad k = 0, 1, \ldots, p. \tag{3.8}$$

[1] The largest subscript here is p while in Section 2.5 it is n.

To find a spectral density Φ satisfying (3.8) is the classical *covariance extension problem* (Grenander and Szegö, 1958).

In general, existence of $\Phi \in \mathfrak{S}_m$ satisfying (3.3) is not trivial. Such feasibility problem was addressed in (Georgiou, 2002b,a) (see also (Ferrante et al., 2010, 2012b; Zorzi and Ferrante, 2012; Ferrante et al., 2007, 2008; Ramponi et al., 2009, 2010; Ferrante et al., 2012a)). In order for $\Sigma > 0$ to be a state covariance, the Lyapunov-like equation

$$\Sigma - A\Sigma A^* = BH + H^*B^*$$

in the unknown $H \in \mathbb{C}^{m\times n}$ has to be solvable or an equivalent rank condition must hold

$$\operatorname{rank}\begin{bmatrix} \Sigma - A\Sigma A^* & B \\ B^* & 0 \end{bmatrix} = \operatorname{rank}\begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix}.$$

Here we shall take the feasibility as a standing assumption, which is also expressible in terms of the linear operator defined below

$$\begin{aligned} \Gamma\colon C(\mathbb{T};\mathfrak{H}_m) &\to \mathfrak{H}_n \\ \Phi &\mapsto \int G\Phi G^*. \end{aligned} \tag{3.9}$$

More precisely, we assume that the (positive definite) covariance matrix $\Sigma \in \operatorname{Range}\Gamma$. Various properties of the set $\operatorname{Range}\Gamma$ are elaborated in e.g., (Ferrante et al., 2012b, Section III). In particular, by Proposition 3.1 of that paper, $\operatorname{Range}\Gamma \subset \mathfrak{H}_n$ is a linear space with real dimension $m(2n-m)$.

The parametric formulation of the spectral estimation problem starts by defining the set of parameters

$$\mathscr{L}_+ := \{\Lambda \in \mathfrak{H}_n \ :\ G^*(z)\Lambda G(z) > 0,\ \forall z \in \mathbb{T}\}, \tag{3.10}$$

which obviously contains all the Hermitian positive definite matrices, since $G(z)$ is of full column rank for any $z \in \mathbb{T}$ which readily follows from the problem setup. By the continuous dependence of eigenvalues on the matrix entries, one can verify that $\mathscr{L}_+$ is an open subset of $\mathfrak{H}_n$. For $\Lambda \in \mathscr{L}_+$, take W_Λ as the unique stable and minimum phase (right) spectral factor of $G^*\Lambda G$ (Ferrante et al., 2010, Lemma 11.4.1), i.e.,

$$G^*\Lambda G = W_\Lambda^* W_\Lambda. \tag{3.11}$$

To avoid any redundancy in the parametrization, we have to define the set $\mathscr{L}_+^\Gamma := \mathscr{L}_+ \cap \operatorname{Range}\Gamma$. This is due to a simple geometric result. More precisely, the adjoint map of Γ in

(3.9) is given by (cf. (Ferrante et al., 2010))

$$\begin{aligned} \Gamma^* : \mathfrak{H}_n &\to C(\mathbb{T}; \mathfrak{H}_m) \\ X &\mapsto G^* X G, \end{aligned} \tag{3.12}$$

and we have the relation

$$(\operatorname{Range}\Gamma)^\perp = \operatorname{Ker}\Gamma^* = \{ X \in \mathfrak{H}_n \,:\, G^*(z) X G(z) = 0,\ \forall z \in \mathbb{T} \}. \tag{3.13}$$

Hence for any $\Lambda \in \mathscr{L}_+$, we have the orthogonal decomposition

$$\Lambda = \Lambda^\Gamma + \Lambda^\perp$$

with $\Lambda^\Gamma \in \operatorname{Range}\Gamma$ and $\Lambda^\perp$ in the orthogonal complement. In view of (3.13), the part $\Lambda^\perp$ does not contribute to the function value of $G^*\Lambda G$ on the unit circle, and we simply have

$$\mathscr{L}_+^\Gamma = \Pi_{\operatorname{Range}\Gamma} \mathscr{L}_+,$$

where $\Pi_{\operatorname{Range}\Gamma}$ denotes the orthogonal projection operator onto the linear space $\operatorname{Range}\Gamma$.

Now suppose that we are given $\Psi \in \mathfrak{S}_m$, which represents an *a priori* information that we have on the desired solution to (3.3). We can then define a parametric family of spectral densities

$$\mathscr{D} := \left\{ \Phi_\Lambda = W_\Lambda^{-1} \Psi W_\Lambda^{-*} \,:\, \Lambda \in \mathscr{L}_+^\Gamma \right\}. \tag{3.14}$$

We have the map

$$\Lambda \mapsto W_\Lambda \mapsto W_\Lambda^{-1} \Psi W_\Lambda^{-*}$$

from the parameter $\Lambda \in \mathscr{L}_+^\Gamma$ to the density function Φ_Λ.

Remark 3.2.1. In the scalar case, the form of spectral densities in the family (3.14) reduces to

$$\Phi_\Lambda = \frac{\Psi}{G^* \Lambda G},$$

which is precisely the solution (4.3) in (Georgiou and Lindquist, 2003) of a constrained optimization problem in terms of the Lagrange multiplier Λ. An alternative matricial parametrization has been proposed in (Georgiou, 2006) and will be revisited in Chapter 5.

Our problem is formulated as follows.

Problem 3.2.2. Given the filter bank $G(z)$ in (3.2), the prior $\Psi \in \mathfrak{S}_m$, and a positive definite matrix $\Sigma \in \operatorname{Range}\Gamma$, find a spectral density Φ_Λ in the parametric family $\mathscr{D}$ defined in (3.14)

such that

$$\int G\Phi_\Lambda G^* = \Sigma. \tag{3.15}$$

The above problem has an equivalent formulation. Define $\operatorname{Range}_+ \Gamma := \operatorname{Range}\Gamma \cap \mathfrak{H}_{+,n}$ where $\mathfrak{H}_{+,n}$ is the open set of $n \times n$ Hermitian positive definite matrices. Consider the map

$$\begin{aligned} \omega &: \mathscr{L}_+^\Gamma \to \operatorname{Range}_+ \Gamma \\ \Lambda &\mapsto \int G W_\Lambda^{-1} \Psi W_\Lambda^{-*} G^*. \end{aligned} \tag{3.16}$$

Then Problem 3.2.2 is asking: what is the preimage of a given $\Sigma \in \operatorname{Range}_+ \Gamma$ under the map ω? As will be clear in Section 3.4, this is a continuous map between open subsets of the linear space $\operatorname{Range}\Gamma$. Naturally, a first question is existence of a solution, i.e., *surjectivity* of the map ω. If so, the second question is on the (much harder) *injectivity*, in other words, uniqueness of the solution parameter to the above problem. In the next section, we will show that ω is indeed a *bijection* if the prior Ψ is scalar. The general case where Ψ is arbitrarily matrix-valued will be treated later.

Remark 3.2.3. The codomain of the map deserves a bit of justification. By definition we have $\omega(\Lambda) \in \operatorname{Range}\Gamma$. To see the fact that $\omega(\Lambda)$ is also positive definite, notice that there exists a real number $\mu > 0$ such that $\Phi_\Lambda \geq \mu I$ since $\Phi_\Lambda = W_\Lambda^{-1}\Psi W_\Lambda^{-*}$ is coercive. Then we have $\omega(\Lambda) \geq \mu \int GG^*$, whose right hand side is strictly positive following from the reachability of (A, B) (see Proposition 3.5.1).

3.3 Well-Posedness Given a Scalar Prior

In the case of a scalar prior, in which we take $\Psi(z) = \psi(z) I_m$ where the scalar-valued function $\psi(z) \in \mathfrak{S}_1$, the map ω reduces to

$$\begin{aligned} \tilde{\omega} &: \mathscr{L}_+^\Gamma \to \operatorname{Range}_+ \Gamma \\ \Lambda &\mapsto \int \psi G (G^* \Lambda G)^{-1} G^*, \end{aligned} \tag{3.17}$$

and the family of spectral densities becomes

$$\tilde{\mathscr{D}} := \{ \Phi_\Lambda = \psi (G^* \Lambda G)^{-1} \, : \, \Lambda \in \mathscr{L}_+^\Gamma \}. \tag{3.18}$$

In this section, we shall assume the prior density ψ to be also *continuous*. This assumption is not strictly necessary, but it does not entail much loss of generality and we make it here for

the sake of simplicity.

According to (Ferrante et al., 2010), a solution to Problem 3.2.2 under a scalar prior exists and is unique. We shall next show that given a continuous prior ψ, the map $\tilde{\omega}$ is a C^1 *diffeomorphism*[2] between $\mathscr{L}_+^\Gamma$ and $\mathrm{Range}_+\,\Gamma$, which in particular, means that the solution Λ depends continuously on the covariance data Σ, and thus the problem is well-posed in the sense of Hadamard. The proof is an application of the global inverse function theorem of Hadamard that appears e.g., in (Gordon, 1972) (see also (Krantz and Parks, 2013, p. 127)).

Theorem 3.3.1 (Hadamard)**.** *Let M_1 and M_2 be connected, oriented, boundary-less n-dimensional manifolds of class C^1, and suppose that M_2 is simply connected. Then a C^1 map $f : M_1 \to M_2$ is a diffeomorphism if and only if f is proper*[3] *and the Jacobian determinant of f never vanishes.*

Conditions on the domain and codomain of $\tilde{\omega}$ can be verified easily. In fact, the set $\mathscr{L}_+^\Gamma = \mathscr{L}_+ \cap \mathrm{Range}\,\Gamma$ is easily seen to be open and path-connected since both $\mathscr{L}_+$ and $\mathrm{Range}\,\Gamma$ are such. The simple connectedness of $\mathrm{Range}_+\,\Gamma$ follows from its convexity (see Proposition B.1.1 in Appendix B). The fact that $\tilde{\omega}$ is of class C^1 is also reported there (cf. Lemma B.1.3). Moreover, properness of the more general map ω has been proven in (Ferrante et al., 2010, Theorem 11.4.1). Therefore, it is only left to check the Jacobian of $\tilde{\omega}$. The next result can be viewed as an interpretation of (Ferrante et al., 2010, Theorem 11.4.2). Here and in the sequel, we shall introduce the notation $\Phi(z;\Lambda)$ to denote a spectral density function that depends on the parameter Λ, and use it interchangeably with $\Phi_\Lambda(z)$.

Proposition 3.3.2. *The Jacobian determinant of $\tilde{\omega}$ never vanishes in $\mathscr{L}_+^\Gamma$, and hence the map $\tilde{\omega}$ is a diffeomorphism.*

Proof. From Lemma B.1.3, the differential of $\tilde{\omega}$ at $\Lambda \in \mathscr{L}_+^\Gamma$ is given by (B.3) such that $\delta\Lambda \in \mathrm{Range}\,\Gamma$. Our target is to show that

$$\delta\tilde{\omega}(\Lambda;\delta\Lambda) = 0 \implies \delta\Lambda = 0.$$

To this end, first notice that the middle part of the integrand in (B.3) is just the differential of the spectral density $\Phi_\Lambda = \psi(G^*\Lambda G)^{-1}$ w.r.t. Λ:

$$\delta\Phi(z;\Lambda;\delta\Lambda) := -\psi(G^*\Lambda G)^{-1}(G^*\delta\Lambda G)(G^*\Lambda G)^{-1}.$$

Then the condition $\delta\tilde{\omega}(\Lambda;\delta\Lambda) = 0$ means that

$$\delta\Phi(z;\Lambda;\delta\Lambda) \in \mathrm{Ker}\,\Gamma = (\mathrm{Range}\,\Gamma^*)^\perp,$$

[2] The word "diffeomorphism" in the sequel should always be understood in the C^1 sense. Hence the attributive C^1 will be omitted.

[3] Recall that f is called proper if the preimage of every compact set in M_2 is compact in M_1.

which in view of (3.12), reads

$$\begin{aligned}\langle G^*XG, \delta\Phi(z;\Lambda;\delta\Lambda)\rangle &= \operatorname{tr}\int G^*XG\,\delta\Phi(z;\Lambda;\delta\Lambda)\\ &= 0, \quad \forall X \in \mathfrak{H}_n.\end{aligned}$$

In particular, following (Ferrante et al., 2010, Equations 11.44–11.45), choosing $X = \delta\Lambda$ will lead to

$$G^*\delta\Lambda G \equiv 0, \quad \forall z \in \mathbb{T},$$

which by (3.13), implies that $\delta\Lambda \in (\operatorname{Range}\Gamma)^{\perp}$. Since at the same time $\delta\Lambda \in \operatorname{Range}\Gamma$, it is necessary that $\delta\Lambda = 0$. The rest is just an application of Theorem 3.3.1.

■

Remark 3.3.3. The unique solution in $\tilde{\mathscr{D}}$ to the spectral estimation problem has an interesting characterization in terms of an optimization problem studied in (Avventi, 2011a). In fact, the equation $\tilde{\omega}(\Lambda) = \Sigma$ is equivalent to the stationarity condition of the dual function introduced in that paper. We will get back to this point in more details in the next chapter.

3.3.1 Continuity with Respect to the Prior Function

In this subsection, we shall show that the unique solution in the family $\tilde{\mathscr{D}}$ to the parametric spectral estimation problem depends also continuously on the prior function ψ. The idea is to study the moment map with the prior function incorporated. Due to the regularity of the moment map, we can view the solution parameter as an implicit functional of the prior, and then invoke the Banach space version of the implicit function theorem to prove continuity.

Consider the following map

$$\begin{aligned}\mathbf{f}:\ D = C_+(\mathbb{T}) \times \mathscr{L}_+^{\Gamma} &\to \operatorname{Range}_+\Gamma\\ (\psi,\Lambda) &\mapsto \int G\psi(G^*\Lambda G)^{-1}G^*.\end{aligned} \tag{3.19}$$

Given $\Sigma \in \operatorname{Range}_+\Gamma$, we aim to solve the equation

$$\mathbf{f}(\psi,\Lambda) = \Sigma. \tag{3.20}$$

For a fixed $\psi \in C_+(\mathbb{T})$, we have

$$\tilde{\omega}(\,\cdot\,) = \mathbf{f}(\psi,\,\cdot\,):\ \mathscr{L}_+^{\Gamma} \to \operatorname{Range}_+\Gamma \tag{3.21}$$

a section of the map $\mathbf{f}$. Since $\tilde{\omega}$ is a diffeomorphism by Proposition 3.3.2, we know that the solution map

$$s : (\psi, \Sigma) \mapsto \Lambda$$

is well defined, and for a fixed ψ, the map $s(\psi, \cdot) : \mathrm{Range}_+ \Gamma \to \mathscr{L}_+^\Gamma$ is continuous. We shall next show the well-posedness in the other respect, namely, continuity of the map

$$s(\cdot, \Sigma) : C_+(\mathbb{T}) \to \mathscr{L}_+^\Gamma \tag{3.22}$$

when Σ is held fixed. Note that continuity here is to be understood in the metric space setting. Clearly, it is equivalent to consider solving the functional equation (3.20) for Λ in terms of ψ when its right-hand side is fixed, which naturally falls in to the scope of the implicit function theorem.

We first show that $\mathbf{f}$ is of class C^1 on its domain D. According to (Lang, 1999, Proposition 3.5, p. 10), it is equivalent to show that the two partial derivatives of $\mathbf{f}$ exist and are continuous in D. More precisely, the partials evaluated at a point are understood as linear operators between two underlying vector spaces

$$\begin{aligned} \mathbf{f}_1' &: D \to L(C(\mathbb{T}), \mathrm{Range}\,\Gamma), \\ \mathbf{f}_2' &: D \to L(\mathrm{Range}\,\Gamma, \mathrm{Range}\,\Gamma). \end{aligned} \tag{3.23}$$

The symbol $L(X, Y)$ denotes the vector space of continuous linear operators between two Banach spaces X and Y, which is itself a Banach space. We need some lemmas. Notice that convergence of a sequence of continuous functions on a fixed interval $[a, b] \subset \mathbb{R}$ will always be understood in the max-norm

$$\|f\| := \max_{t \in [a,b]} |f(t)|. \tag{3.24}$$

For $m \times n$ matrix valued continuous functions in one variable, define the norm as

$$\|M\| := \max_{t \in [a,b]} \|M(t)\|_F \tag{3.25}$$

It is easy to verify that convergence in the norm (3.25) is equivalent to element-wise convergence in the max-norm (3.24).

Lemma 3.3.4. *For an $n \times p$ matrix continuous function $M(\theta)$ on $[-\pi, \pi]$, the inequality holds for the Frobenius norm*

$$\left\| \int M(\theta) \right\|_F \leq \sqrt{np} \int \|M(\theta)\|_F. \tag{3.26}$$

Proof. Let $m_{jk}(\theta)$ be the (j,k) element of $M(\theta)$. Then we have

$$\begin{aligned}\left\|\int M(\theta)\right\|_F^2 &= \sum_{j,k}\left|\int m_{jk}(\theta)\right|^2 \le np\max_{j,k}\left|\int m_{jk}(\theta)\right|^2 \\ &\le np\max_{j,k}\left(\int |m_{jk}(\theta)|\right)^2 \\ &\le np\left(\int \|M(\theta)\|_F\right)^2\end{aligned} \tag{3.27}$$

where the third inequality holds because $|m_{jk}(\theta)| \le \|M(\theta)\|_F$ for any j,k. ∎

Lemma 3.3.5. *If a sequence $\{\Lambda_k\} \subset \mathscr{L}_+^\Gamma$ converges to $\Lambda \in \mathscr{L}_+^\Gamma$, then the sequence of functions $\{(G^*\Lambda_k G)^{-1}\}$ converges to $(G^*\Lambda G)^{-1}$ in the norm (3.25).*

Proof. From Lemma B.1.2, there exists $\mu > 0$ such that for any k and $\theta \in [-\pi,\pi]$, $G^*\Lambda_k G \ge \mu I$. Hence we have

$$\begin{aligned}&\|(G^*\Lambda_k G)^{-1} - (G^*\Lambda G)^{-1}\|_F \\ =&\|(G^*\Lambda_k G)^{-1}G^*(\Lambda-\Lambda_k)G(G^*\Lambda G)^{-1}\|_F \\ \le&\kappa^2\mu^{-2}G_{\max}\|\Lambda_k - \Lambda\|_F \to 0,\end{aligned} \tag{3.28}$$

where the constant $G_{\max}$ defined in (B.1), and we have used submultiplicativity of the Frobenius norm and norm equivalence $\|\cdot\|_F \le \kappa\|\cdot\|_2$. ∎

Proposition 3.3.6. *The map* $\mathbf{f}$ *in* (3.19) *is of class* C^1.

Proof. Consider the partial derivative w.r.t. the first argument. Due to linearity, one has

$$\begin{aligned}\mathbf{f}_1'(\psi,\Lambda):\ &C(\mathbb{T}) \to \operatorname{Range}\Gamma \\ &\delta\psi \mapsto \int G\delta\psi(G^*\Lambda G)^{-1}G^*.\end{aligned} \tag{3.29}$$

Clearly, the operator does not depend on ψ. Let the sequence $\{(\psi_k,\Lambda_k)\}_{k\ge1} \subset D$ converge in the product topology to $(\psi,\Lambda) \in D$, that is, $\psi_k \to \psi$ in the max-norm and $\Lambda_k \to \Lambda$ in any matrix norm. We need to show that

$$\mathbf{f}_1'(\psi_k,\Lambda_k) \to \mathbf{f}_1'(\psi,\Lambda).$$

in the operator norm. Indeed, we have

$$\begin{aligned}
&\|\mathbf{f}_1'(\psi_k,\Lambda_k)-\mathbf{f}_1'(\psi,\Lambda)\| \\
&= \sup_{\|\delta\psi\|=1}\left\|\int G\delta\psi\left[(G^*\Lambda_k G)^{-1}-(G^*\Lambda G)^{-1}\right]G^*\right\|_F \\
&\le nG_{\max}\left\|(G^*\Lambda_k G)^{-1}-(G^*\Lambda G)^{-1}\right\| \to 0.
\end{aligned} \tag{3.30}$$

where we have used the inequality (3.26) and Lemma 3.3.5.

For the partial derivative of $\mathbf{f}$ w.r.t. the second argument, we have

$$\begin{aligned}
&\mathbf{f}_2'(\psi,\Lambda):\ \operatorname{Range}\Gamma\to\operatorname{Range}\Gamma \\
&\qquad\qquad \delta\Lambda\mapsto -\int G\psi(G^*\Lambda G)^{-1}(G^*\delta\Lambda G)(G^*\Lambda G)^{-1}G^*.
\end{aligned} \tag{3.31}$$

To ease the notation, set $\Phi(\psi,\Lambda)=\psi(G^*\Lambda G)^{-1}$ and

$$\delta\Phi(\psi,\Lambda;\delta\Lambda)=\Phi(\psi,\Lambda)(G^*\delta\Lambda G)(G^*\Lambda G)^{-1}.$$

Through similar computation, we arrive at

$$\begin{aligned}
&\|\mathbf{f}_2'(\psi_k,\Lambda_k)-\mathbf{f}_2'(\psi,\Lambda)\| \\
&\le \sup_{\|\delta\Lambda\|=1} nG_{\max}\|\delta\Phi(\psi_k,\Lambda_k;\delta\Lambda)-\delta\Phi(\psi,\Lambda;\delta\Lambda)\| \to 0.
\end{aligned} \tag{3.32}$$

The limit tends to 0 because the part

$$\begin{aligned}
&\sup_{\|\delta\Lambda\|=1}\|\delta\Phi(\psi_k,\Lambda_k;\delta\Lambda)-\delta\Phi(\psi,\Lambda;\delta\Lambda)\| \\
&= \max_{\substack{\|\delta\Lambda\|=1,\\ \theta\in[-\pi,\pi]}}\|\delta\Phi(\psi_k,\Lambda_k;\delta\Lambda)-\delta\Phi(\psi,\Lambda;\delta\Lambda)\|_F \\
&= \max_{\substack{\|\delta\Lambda\|=1,\\ \theta\in[-\pi,\pi]}}\Big\|\delta\Phi(\psi_k,\Lambda_k;\delta\Lambda)-\Phi(\psi_k,\Lambda_k)(G^*\delta\Lambda G)(G^*\Lambda G)^{-1} \\
&\qquad\qquad +\Phi(\psi_k,\Lambda_k)(G^*\delta\Lambda G)(G^*\Lambda G)^{-1}-\delta\Phi(\psi,\Lambda;\delta\Lambda)\Big\|_F \\
&\le \max_{\substack{\|\delta\Lambda\|=1,\\ \theta\in[-\pi,\pi]}}\big(\|\Phi(\psi_k,\Lambda_k)\|_F\|(G^*\Lambda_k G)^{-1}-(G^*\Lambda G)^{-1}\|_F \\
&\qquad\qquad +\|\Phi(\psi_k,\Lambda_k)-\Phi(\psi,\Lambda)\|_F\|(G^*\Lambda G)^{-1}\|_F\big)\|G^*\delta\Lambda G\|_F \\
&\le \kappa\mu^{-1}G_{\max}\big(K_\psi\|(G^*\Lambda_k G)^{-1}-(G^*\Lambda G)^{-1}\|+\|\Phi(\psi_k,\Lambda_k)-\Phi(\psi,\Lambda)\|\big)
\end{aligned} \tag{3.33}$$

Note that $\|\psi_k\|\le K_\psi$ for some $K_\psi>0$ uniformly in k because $\psi_k\to\psi$. Also, $\Phi(\psi_k,\Lambda_k)\to$

$\Phi(\psi, \Lambda)$ is a simple consequence of Lemma 3.3.5 and the fact

$$f_k g_k \to f\, g \text{ if } f_k \to f,\ g_k \to g.$$

■

We are now in a place to state the main result of this subsection.

Theorem 3.3.7. *For a fixed* $\Sigma \in \mathrm{Range}_+ \Gamma$*, the implicit function* $s(\,\cdot\,, \Sigma)$ *in* (3.22) *is of class* C^1.

Proof. The assertion follows directly from the Banach space version of the implicit function theorem (see, e.g., (Lang, 1999, Theorem 5.9, p. 19)), because restrictions of $s(\,\cdot\,, \Sigma)$ must coincide with those locally defined, continuously differentiable implicit functions, which exist around every $\psi \in C_+(\mathbb{T})$ following from Proposition 3.3.6 and the fact that the partial $\mathbf{f}_2'(\psi, \Lambda)$ is a vector space isomorphism everywhere in D. ■

3.4 Existence of a Solution Given a Matrix Prior

In this section, we tackle the nontrivial existence question of a solution to Problem 3.2.2. Unfortunately, attempts of extending the optimization framework in (Avventi, 2011a) in order to accommodate a matrix-valued prior Ψ turn out to be problematic, as pointed out in (Georgiou, 2006; Ferrante et al., 2010). In other words, there does not seem to exist a suitable cost function whose minimizer will lead to a solution of form (3.14). Therefore, we have to take a different route.

The proof of existence of a solution to Problem 3.2.2 in general relies on the notion of topological degree of a continuous map. The degree theory forms an important part of differential topology and is closely related to fixed-point theory[4] (cf. (Outerelo and Ruiz, 2009, Chapter I) for a rather informative historical account). In particular, the degree theory is a powerful tool to prove existence of a solution to a system of nonlinear equations. There are several versions of the theory for different types of maps. Although the maps that we consider in this chapter are between open subsets of the Euclidean space, we shall use the more general degree theory for continuous maps between smooth, connected, boundary-less manifolds. Some main points of the theory are reviewed below.

3.4.1 A Short Review of the Degree Theory

We mainly follow the lines of (Outerelo and Ruiz, 2009, Chapter III). Suppose that $U, V \subset \mathbb{R}^n$ are open and connected, and $f : U \to V$ is a *proper* C^1 function. Our major concern is

[4]The proof of Theorem 3.3.1 also relies on the degree theory (cf. (Gordon, 1972)).

solvability of the equation

$$f(x) = y. \tag{3.34}$$

A point $y \in V$ is called a regular value of f if either

(i) for any $x \in f^{-1}(y)$, $\det f'(x) \neq 0$ or

(ii) $f^{-1}(y)$ is empty.

Here $f^{-1}(y)$ denotes the preimage of y under f, i.e., the set

$$\{x \in U : f(x) = y\},$$

and $f'(x)$ denotes the Jacobian matrix of f evaluated at x. Let y be a regular value of type (i), and the degree of f at y is defined as

$$\deg(f, y) := \sum_{f(x)=y} \operatorname{sign} \det f'(x), \tag{3.35}$$

where the sign function

$$\operatorname{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases}$$

and not defined at 0.

Throughout this chapter, properness will be a crucial property of our function. Since f is proper, one can show that the preimage $f^{-1}(y)$ is finite following the classical inverse function theorem, and hence the sum above is well defined. For regular values of type (ii), we set $\deg(f, y) = 0$. Moreover, the set of regular values is dense in V by Sard–Brown Theorem (Outerelo and Ruiz, 2009, p. 63). Further properties of the degree related to our problem are listed below:

- The degree of f at y does not depend on the choice of regular value. Therefore, we can define the degree of f as

  $$\deg(f) = \deg(f, y)$$

  for any regular value y.

- If $\deg(f) \neq 0$, then for any $y \in V$, there exists $x \in U$ such that $f(x) = y$, that is, the map f is surjective. A proof of this fact can be found e.g., in (Byrnes et al., 1995, p. 1849).

- Homotopy invariance. If $H: U \times [0,1] \to V$, $(x,t) \mapsto y$ is jointly continuous in (x,t) and proper, then $\deg(H_t, y)$ is defined and independent of $t \in [0,1]$. Here $H_t: U \to V$ is defined by $H_t(x) = H(x,t)$.

One important point of the theory is that degree can be defined for continuous functions through approximation by smooth functions(Outerelo and Ruiz, 2009, Proposition and Definition 3.1, p. 111), and (3.35) is just a way of computing it in the special case of C^1 (Schwartz, 1969, Remark p. 71). In particular, the homotopy invariance of the degree holds in the continuous case (Outerelo and Ruiz, 2009, Proposition 3.4, p. 112).

3.4.2 Proof of Existence

Take $\Psi = I$ the identity matrix, and the map ω reduces to

$$\begin{aligned} \tilde{\omega}_{\mathbf{1}}: \mathscr{L}_+^\Gamma &\to \text{Range}_+ \Gamma \\ \Lambda &\mapsto \int G(G^*\Lambda G)^{-1}G^*. \end{aligned} \tag{3.36}$$

The fact that $\tilde{\omega}_{\mathbf{1}}$ is C^1 follows from Lemma B.1.3. We will need the next lemma before proving our main theorem of this section.

Lemma 3.4.1. *The map*

$$\begin{aligned} H: \mathscr{L}_+^\Gamma \times [0,1] &\to \text{Range}_+ \Gamma \\ (\Lambda, t) &\mapsto \int G\Phi_{\Lambda,t}G^* \end{aligned} \tag{3.37}$$

is a proper continuous homotopy between ω and $\tilde{\omega}_{\mathbf{1}}$, where

$$\Phi_{\Lambda,t} := W_\Lambda^{-1}[\, t\Psi + (1-t)I \,]W_\Lambda^{-*}. \tag{3.38}$$

Proof. By definition we need to show two things, namely that H is jointly continuous in Λ and t and that H is proper. In order to prove joint continuity, we first notice that the spectral factor $W_\Lambda(z)$ can be written as (Ferrante et al., 2010, Lemma 11.4.1)

$$W_\Lambda(z) = L_\Lambda^{-*}B^*P_\Lambda A(zI-A)^{-1}B + L_\Lambda, \tag{3.39}$$

where P_Λ is the unique stabilizing solution of the following Discrete-time Algebraic Riccati Equation (DARE)

$$\Pi = A^*\Pi A - A^*\Pi B(B^*\Pi B)^{-1}B^*\Pi A + \Lambda, \tag{3.40}$$

and L_Λ is the right Cholesky factor of $B^*P_\Lambda B > 0$, i.e.,

$$B^*P_\Lambda B = L_\Lambda^* L_\Lambda \tag{3.41}$$

with L_Λ being lower triangular having real and positive diagonal entries. Next, let us introduce a change of variables by letting

$$C_\Lambda := L_\Lambda^{-*} B^* P_\Lambda. \tag{3.42}$$

Then it is not difficult to recover the relation $L_\Lambda = C_\Lambda B$. In this way, the spectral factor (3.39) can be rewritten as

$$\begin{aligned} W_\Lambda(z) &= C_\Lambda A(zI - A)^{-1}B + C_\Lambda B \\ &= zC_\Lambda G\,, \end{aligned} \tag{3.43}$$

where the second equality holds because of the identity $A(zI-A)^{-1}+I = z(zI-A)^{-1}$. According to (Avventi, 2011a, Theorem A.5.5), the dependence of the $m \times n$ matrix C_Λ defined above on $\Lambda \in \mathscr{L}_+^\Gamma$ turns out to be a homeomorphism. From this fact it follows that $W_\Lambda(e^{i\theta})$ depends continuously on $\Lambda \in \mathscr{L}_+^\Gamma$, for all $\theta \in [-\pi, \pi]$. Consider now

$$\Phi_{\Lambda,t}(e^{i\theta}) = W_\Lambda^{-1}(e^{i\theta})[\, t\Psi(e^{i\theta}) + (1-t)I\,]W_\Lambda^{-*}(e^{i\theta}).$$

As a linear combination in $t \in [0,1]$ of continuous functions of Λ, $\Phi_{\Lambda,t}(e^{i\theta})$ is jointly continuous w.r.t. $t \in [0,1]$ and $\Lambda \in \mathscr{L}_+^\Gamma$, for all $\theta \in [-\pi, \pi]$.

Next we need to show the continuity together with the integral. Consider any sequence $\{(\Lambda_k, t_k)\}_{k\geq 1} \subset \mathscr{L}_+^\Gamma \times [0,1]$ such that $\lim_{k\to\infty} t_k = \bar{t} \in [0,1]$ and $\lim_{k\to\infty} \Lambda_k = \bar{\Lambda} \in \mathscr{L}_+^\Gamma$. By Lemma B.1.2, there exists $\mu > 0$ such that $G^*\Lambda_k G \geq \mu I,\ \forall k$. Therefore, it holds that

$$\begin{aligned} G\Phi_{\Lambda_k,t_k}G^* &\leq \kappa G(G^*\Lambda_k G)^{-1}G^* \\ &\leq \kappa\mu^{-1}GG^*, \quad \forall\, k \geq 1, \end{aligned}$$

where κ is a positive real number such that

$$t\Psi(e^{i\theta}) + (1-t)I \leq \kappa I,\ \forall\, t \in [0,1],\ \theta \in [-\pi, \pi].$$

Such κ exists since Ψ is bounded. The rest argument is similar to that in the proof of Lemma B.1.3. Given the joint continuity of $\Phi_{\Lambda,t}$ in Λ and t, one can show that the following limit

holds

$$\lim_{k\to\infty}\int G\Phi_{\Lambda_k,t_k}G^* = \int \lim_{k\to\infty} G\Phi_{\Lambda_k,t_k}G^* = \int G\Phi_{\bar{\Lambda},\bar{t}}G^*.$$

This proves joint continuity of H in t and Λ.

Once we have joint continuity, the properness is not difficult to prove. In fact, let $K \subset \mathrm{Range}_+\,\Gamma$ be a compact subset, and we next show that the set

$$H^{-1}(K) := \{\,(\Lambda,t) \in \mathscr{L}_+^\Gamma \times [0,1] : H(\Lambda,t) \in K\,\}$$

is compact. The argument is essentially the same as the proof of Theorem 11.4.1 of (Ferrante et al., 2010). Since our setting is finite-dimensional, a set being compact is equivalent to being closed and bounded. If $H^{-1}(K)$ is unbounded, one can then find a sequence $\{(\Lambda_k,t_k)\} \subset H^{-1}(K)$ such that $\|(\Lambda_k,t_k)\| \to \infty$ as $k\to\infty$, which necessarily implies $\|\Lambda_k\| \to \infty$. However, in this case $H(\Lambda_k,t_k)$ will tend to be singular, which contradicts the premise of K being compact. This proves the boundedness.

To prove the closedness, if a sequence $\{(\Lambda_k,t_k)\}$ in $H^{-1}(K)$ converges to (Λ,t), then Λ cannot be on the boundary of $\mathscr{L}_+$, otherwise $\|H(\Lambda_k,t_k)\| \to \infty$, which again contradicts the compactness of K. To see the latter fact, notice that

$$\begin{aligned} H(\Lambda_k,t_k) &= \int G\Phi_{\Lambda_k,t_k}G^* \\ &= \int GW_{\Lambda_k}^{-1}[\,t_k\Psi + (1-t_k)I\,]W_{\Lambda_k}^{-*}G^* \\ &\geq \kappa_{\min}\int G(G^*\Lambda_k G)^{-1}G^*, \end{aligned}$$

where the constant $\kappa_{\min}$ is such that $t_k\Psi(e^{i\theta})+(1-t_k)I \geq \kappa_{\min}$ for all θ and k. Such constant exists since Ψ is coercive. Now if $\{\Lambda_k\}$ approaches $\partial\mathscr{L}_+$, then $G^*(e^{i\theta})\Lambda_k G(e^{i\theta})$ tends to be singular for some θ. Since G has rank m on $\mathbb{T}$, this in turn implies that $\|H(\Lambda_k,t_k)\| \to \infty$ as $k\to\infty$ (cf. Lemma B.1.4 in Appendix B for more details on this point). Therefore, by the joint continuity of H, $(\Lambda,t) \in H^{-1}(K)$. This concludes the proof of properness. ∎

Theorem 3.4.2. *The map ω is surjective.*

Proof. Given the second listed property of the degree, the claim follows directly if we can show that

$$\deg(\omega) \neq 0.$$

We notice first that ω is proper by Theorem 11.4.1 from (Ferrante et al., 2010), and thus the degree is well defined. By Lemma 3.4.1 and the homotopy invariance of the degree,

$$\deg(\omega) = \deg(\tilde{\omega}_1).$$

As a consequence of Sard–Brown theorem (Outerelo and Ruiz, 2009, p. 63), the codomain $\mathrm{Range}_+ \Gamma$ must contain a regular value of $\tilde{\omega}_1$ since it has positive $\mathrm{Range}\,\Gamma$-Lebesgue measure. By Lemma B.1.3, the C^1 degree (3.35) of $\tilde{\omega}_1$ at a regular value is well-defined. Meanwhile, from Proposition 3.3.2, we know that $\tilde{\omega}_1$ is a diffeomorphism. Therefore, we must have

$$\deg(\tilde{\omega}_1) \neq 0,$$

and this concludes the proof. ∎

3.4.3 The Special Case of Covariance Extension

Given $\Lambda \in \mathscr{L}_+$ and $G(z)$ in (3.5), $G^*\Lambda G$ is now a matrix Laurent polynomial that takes positive definite values on the unit circle. Let us take

$$Q(z) := \sum_{k=-p}^{p} Q_k z^k \equiv G^*\Lambda G, \quad Q_{-k} = Q_k^* \in \mathbb{C}^{m\times m}. \tag{3.44}$$

Then according e.g. to (Baggio and Ferrante, 2016), $Q(z)$ admits a spectral factorization

$$Q(z) = D^*(z)D(z), \tag{3.45}$$

where $D(z) = \sum_{k=0}^{p} D_k z^{-k}$ is a $m \times m$ matrix polynomial (with negative powers) and the scalar polynomial $\det D(z)$ has all its roots strictly inside the unit circle. We shall call such $D(z)$ Schur.[5] Therefore, the outer spectral factor in (3.11) is just

$$W_\Lambda(z) \equiv D(z). \tag{3.46}$$

We have the following corollary of Theorem 3.4.2.

Corollary 3.4.3. *Given a finite $m \times m$ matrix covariance sequence $\Sigma_0, \Sigma_1, \ldots, \Sigma_p$, for any $\Psi \in \mathfrak{S}_m$, there exists a Schur polynomial $D(z)$ of degree p such that the spectral density*

$$\Phi := D^{-1}\Psi D^{-*} \tag{3.47}$$

[5] Moreover, one can make such spectral factor unique if the constant matrix coefficient D_0 is required to be lower triangular with real and positive diagonal elements.

satisfies the moment equations (3.8). *The polynomial* $D(z)$ *is a right Schur spectral factor of* $G^*\Lambda G$ *for some* $\Lambda \in \mathscr{L}_+^\Gamma$.

In particular, when taking $\Psi(z) = N(z)N^*(z)$ with $N(z) = \sum_{k=0}^q N_k z^{-k}$, $N_k \in \mathbb{C}^{m\times m}$, which is the spectral density of a moving-average process, the spectral density Φ in (3.47) corresponds to an m-dimensional vector ARMA (p,q) process

$$\sum_{k=0}^{p} D_k\, y(t-k) = \sum_{k=0}^{q} N_k\, w(t-k), \quad t \in \mathbb{Z}, \tag{3.48}$$

and we recover one of the main results of (Georgiou, 1983, Section V) under a more general setting.

3.5 A Diffeomorphic Spectral Factorization

Given the existence result in the previous section, we would like to approach the question of uniqueness. Quite naturally, we want to extend the analysis in Proposition 3.3.2 to the more general map ω. However, a difficulty then arises as it will entail the differentiation of the spectral factor W_Λ in (3.11) w.r.t. the parameter Λ. Such a difficulty can be bypassed by the change of variable introduced in (Avventi, 2011a) and to be reviewed and further developed next.

Following the lines in the proof of Lemma 3.4.1, given a $\Lambda \in \mathscr{L}_+^\Gamma$, the right outer spectral factor of $G^*\Lambda G$ can be written as $W_\Lambda(z) = zCG$ (3.43), where the matrix $C := L^{-*}B^*P$ (3.42) is defined in terms of the stabilizing solution of the DARE (3.40). Notice that here we have dropped the subscript Λ for the variables to ease the notation. In view of this, the factorization (3.11) can then be rewritten as

$$G^*\Lambda G = G^*C^*CG, \quad \forall z \in \mathbb{T}. \tag{3.49}$$

This relation has also been expressed in (Ferrante et al., 2010, Equation 11.29). In the sequel, we shall also call the $m \times n$ matrix C a "spectral factor".

As reported in (Avventi, 2011a, Section A.5.5), it is possible to build a *homeomorphic* factorization by carefully choosing the set where the factor C lives. More precisely, let the set $\mathscr{C}_+ \subset \mathbb{C}^{m\times n}$ contain those matrices C that satisfy the following two conditions:

- CB is lower triangular with real and positive diagonal entries,
- $A - B(CB)^{-1}CA$ has eigenvalues strictly inside the unit circle.

Define the map

$$\begin{aligned} h : \mathscr{L}_+^\Gamma &\to \mathscr{C}_+ \\ \Lambda &\mapsto C \text{ via (3.42).} \end{aligned} \tag{3.50}$$

Then according to (Avventi, 2011a, Theorem A.5.5), the map h of spectral factorization is a homeomorphism. We shall next strengthen this result by showing that the map h is in fact a diffeomorphism using Theorem 3.3.1.

3.5.1 Characterization of the Diffeomorphism

We are going to apply Theorem 3.3.1 to the inverse of h

$$\begin{aligned} h^{-1} : \mathscr{C}_+ &\to \mathscr{L}_+^\Gamma \\ C &\mapsto \Lambda := \Pi_{\operatorname{Range}\Gamma}(C^*C). \end{aligned} \tag{3.51}$$

Those technical requirements on the domain and codomain of h^{-1} can be verified without difficulty. The set $\mathscr{C}_+$ is an open subset of the linear space

$$\mathfrak{C} := \left\{ C \in \mathbb{C}^{m\times n} \,:\, CB \text{ is lower triangular with real diagonal entries} \right\},$$

whose real dimension coincides with $\operatorname{Range}\Gamma$ (cf. (Avventi, 2011a)). The fact that $\mathscr{C}_+$ is also path-connected is a consequence of h being a homeomorphism. Furthermore, the set $\mathscr{L}_+^\Gamma$ is simply connected due to Proposition B.1.1. The map h^{-1} is actually smooth (hence of course C^1) because it is a composition of the quadratic map $C \mapsto C^*C$ and the projection $\Pi_{\operatorname{Range}\Gamma}$, both of which are smooth. The fact that h^{-1} is proper has also been reported in (Avventi, 2011a). An alternative proof of such properness *independent of optimization* is also given in Appendix B (Proposition B.1.5). Therefore, it remains to investigate the Jacobian of h^{-1}. In order to carry out explicit computation, it is necessary to choose bases for the two linear spaces $\mathfrak{C}$ and $\operatorname{Range}\Gamma$.

Let $M := m(2n-m)$, and let $\{\Lambda_1, \Lambda_2, \dots, \Lambda_M\}$ and $\{C_1, \dots, C_M\}$ be orthonormal bases of $\operatorname{Range}\Gamma$ and $\mathfrak{C}$, respectively. Then one can parametrize $\Lambda \in \mathscr{L}_+^\Gamma$ and $C \in \mathscr{C}_+$ as

$$\begin{aligned} \Lambda(x) &= x_1\Lambda_1 + x_2\Lambda_2 + \cdots + x_M\Lambda_M, \\ C(y) &= y_1C_1 + y_2C_2 + \cdots + y_MC_M, \end{aligned} \tag{3.52}$$

for some $x_j, y_j \in \mathbb{R}$, $j = 1, \dots, M$. The map h^{-1} can then be expressed coordinate-wisely as

$$x_j = \langle \Lambda_j, C(y)^*C(y) \rangle. \tag{3.53}$$

Then the partial derivatives can be computed as

$$\frac{\partial x_j}{\partial y_k} = \langle \Lambda_j, C_k^* C(y) + C^*(y) C_k \rangle, \tag{3.54}$$

which is the (j,k) element of the Jacobian matrix denoted as $J_{h^{-1}}(y)$. We need some ancillary results in order to show that h^{-1} has everywhere nonvanishing Jacobian.

Proposition 3.5.1. *If $v \in \mathbb{C}^n$ is such that $v^* G(z) = 0$ for all $z \in \mathbb{T}$, then $v = 0$.*

Proof. The condition that $v^* G(z) = 0$ for all $z \in \mathbb{T}$ implies that

$$v^* \int G G^* v = 0.$$

Under our problem setting stated in Section 3.2, we have $\int GG^* > 0$ and thus the assertion of the proposition follows. To see the fact of positive definiteness, note first that the following expansion holds

$$\begin{aligned} G(z) &= (zI - A)^{-1} B \\ &= z^{-1} \sum_{k=0}^{\infty} z^{-k} A^k B, \quad \text{for } |z| \geq 1, \end{aligned} \tag{3.55}$$

since A is stable. Then by the Parseval identity, we have

$$\int GG^* = \sum_{k=0}^{\infty} A^k B B^* (A^*)^k = RR^*,$$

where $R = [B, AB, \ldots, A^k B, \ldots]$. The above is the unique solution of the discrete-time Lyapunov equation

$$X - AXA^* = BB^*. \tag{3.56}$$

Since (A,B) is by assumption reachable, R is of full row rank, and therefore $\int GG^* > 0$. ∎

Proposition 3.5.2. *Given $C \in \mathscr{C}_+$, the rational matrix equation in the unknown $V \in \mathbb{C}^{m \times n}$*

$$G^*(C^* V + V^* C) G = 0, \quad \forall z \in \mathbb{T} \tag{3.57}$$

has the general solution

$$V = QC \tag{3.58}$$

where $Q \in \mathbb{C}^{m \times m}$ is an arbitrary constant skew-Hermitian matrix. If one further requires $V \in \mathfrak{C}$, then (3.57) has only the trivial solution $V = 0$.

Proof. The equation (3.57) is equivalent to

$$z^*G^*(C^*V + V^*C)Gz = 0, \quad \forall z \in \mathbb{T}. \tag{3.59}$$

Let

$$\begin{aligned} zCG(z) &= zC(zI - A)^{-1}B \\ &= \frac{P_C(z)}{z^{-n}\det(zI - A)}, \end{aligned}$$

where $P_C(z) := z^{-n+1}C\ \mathrm{adj}(zI - A)B$ and $\mathrm{adj}(\cdot)$ denotes the adjugate matrix. Obviously, $P_C(z)$ is a matrix polynomial in the indeterminate z^{-1}, which is intended to conform to the engineering convention. From (3.55), we have

$$\lim_{z\to\infty} zCG = CB = \lim_{z\to\infty} P_C(z),$$

where the second equality holds since $\lim_{z\to\infty} z^{-n}\det(zI - A) = 1$. Moreover, the scalar polynomial $\det P_C(z)$ has all its roots inside $\mathbb{D}$, which can be seen from (3.43) as zCG is minimum phase, i.e., admits a stable inverse.

Define similarly $P_V(z) := z^{-n+1}V\ \mathrm{adj}(zI - A)B$. Then one can reduce (3.59) to the matrix polynomial equation

$$P_C^*(z)P_V(z) + P_V^*(z)P_C(z) = 0, \quad \forall z \in \mathbb{T}, \tag{3.60}$$

in which we have

$$P_C^*(0) = \left[\lim_{z\to\infty} P_C(z)\right]^* = (CB)^*$$

nonsingular because $C \in \mathscr{C}_+$. By the identity theorem for holomorphic functions, if the above equation holds on $\mathbb{T}$, then it holds for any $z \in \mathbb{C}$ except for 0 (and ∞). Hence the restriction $z \in \mathbb{T}$ can be removed here. Since P_C^* is anti-stable and $P_C^*(0)$ nonsingular, according to Theorem B.2.10, the general solution of (3.60) is

$$P_V = QP_C,$$

where $Q \in \mathbb{C}^{m\times m}$ is an arbitrary constant skew-Hermitian matrix. This in turn implies that

$$VG(z) = QCG(z), \quad \forall z \in \mathbb{T}, \tag{3.61}$$

which in view of Proposition 3.5.1, further implies that $V = QC$.

To prove the remaining part of the claim, just apply the power series expansion (3.55) to (3.61), and notice that all the Fourier coefficients on the two sides of (3.61) must coincide.

This in particular means that

$$VB = QCB.$$

Since we have $C \in \mathscr{C}_+$ and $V \in \mathfrak{C}$ in addition, both VB and CB are lower triangular and the latter is invertible. Therefore Q turns out to be also lower triangular and at the same time skew-Hermitian, which necessarily means that Q is equal to 0 and so is V. ∎

Theorem 3.5.3. *The Jacobian determinant of h^{-1} never vanishes in $\mathscr{C}_+$, and hence the map h in (3.50) is a diffeomorphism.*

Proof. Suppose $v \in \mathbb{R}^M$ is such that $J_{h^{-1}}(y)v = 0$. We need to show that $v = 0$. To this end, notice from (3.54) that equivalently we have for $j = 1, 2, \ldots, M$,

$$\begin{aligned} 0 &= \sum_{k=1}^{M} v_k \langle \Lambda_j, C_k^* C(y) + C^*(y) C_k \rangle \\ &= \langle \Lambda_j, C^*(v)C(y) + C^*(y)C(v) \rangle, \end{aligned}$$

which implies that

$$C^*(v)C(y) + C^*(y)C(v) \perp \operatorname{Range} \Gamma.$$

In view of (3.13), this in turn means

$$G^*(z)\,[C^*(v)C(y) + C^*(y)C(v)]\,G(z) = 0, \quad \forall z \in \mathbb{T}.$$

By Proposition 3.5.2, the only solution is $v = 0$. Thus Theorem 3.3.1 is applicable and this completes the proof. ∎

3.6 Preliminary Results on the Uniqueness Question

Let us return to the map ω defined in (3.16). We shall use the result obtained in the previous section to approach the question of uniqueness of the solution to Problem 3.2.2. Similar to the case of scalar prior, we shall make the assumption $\Psi \in C_{+,m}(\mathbb{T})$, which will facilitate reasoning. Given the relation (3.43), the spectral density Φ_Λ can be reparametrized in C as

$$\Phi_\Lambda \equiv \Phi_C := (CG)^{-1} \Psi (CG)^{-*}. \tag{3.62}$$

In this way, the map ω can be expressed as a composition

$$\omega = \tau \circ h : \ \omega(\Lambda) = \tau(h(\Lambda)), \tag{3.63}$$

with h in (3.50) and

$$\begin{aligned} \tau : \mathscr{C}_+ &\to \text{Range}_+ \Gamma \\ C &\mapsto \int G\Phi_C G^*. \end{aligned} \tag{3.64}$$

Since h has been proved to be a diffeomorphism, we can restrict our attention to the map τ due to the next simple result.

Proposition 3.6.1. *Let X, Y, Z be open subsets of $\mathbb{R}^n$. Suppose we have functions $f : X \to Y$, $g : Y \to Z$ and f is a diffeomorphism between X and Y. Define the composite function*

$$h = g \circ f : X \to Z. \tag{3.65}$$

Then h is a diffeomorphism between X and Z if and only if g is a diffeomorphism between Y and Z.

Proof. The "if" part is trivial since a composition of two diffeomorphisms is again a diffeomorphism. To see the converse, for $y \in Y$, let $x = f^{-1}(y) \in X$ and put it into (3.65) as an argument of h. Then one gets

$$g = h \circ f^{-1},$$

which is again a composition of two diffeomorphisms. ∎

Since properness of the map ω has already be proven, it remains to show that ω is continuously differentiable and has everywhere nonvanishing Jacobian. In view of the relation (3.63) and the previous proposition, it will be sufficient and necessary that the map τ possesses such two properties. We need the next lemma before proving the continuous differentiability.

Proposition 3.6.2. *The map τ in (3.64) is of class C^1.*

Proof. We can proceed by mimicking the proof of Lemma B.1.3, although the argument here is slightly more general. First compute the differential of $\Phi(z; C)$ w.r.t. $C \in \mathscr{C}_+$ as

$$\delta\Phi(z; C; \delta C) = -(CG)^{-1}\delta C G\Phi_C - \Phi_C G^* \delta C^* (CG)^{-*}, \tag{3.66}$$

which is easily seen to be continuous in C and $\theta \in [-\pi, \pi]$ for a fixed $\delta C \in \mathfrak{C}$. This means that we can take the differential of the map τ inside the integral in (3.64)

$$\delta\tau(C; \delta C) = \int G\delta\Phi(e^{i\theta}; C; \delta C)G^*. \tag{3.67}$$

Next we show that the above differential is continuous in C for a fixed δC. To this end, suppose we have a sequence $\{C_k\}_{k\geq 1} \subset \mathscr{C}_+$ that converges to some $\bar{C} \in \mathscr{C}_+$ as $k \to \infty$. Due to the relation (3.49), we have for each k

$$G^*\Lambda_k G = G^* C_k^* C_k G, \quad \forall z \in \mathbb{T}, \tag{3.68}$$

where $\Lambda_k = h^{-1}(C_k) \in \mathscr{L}_+^\Gamma$. Since h is a diffeomorphism by Theorem 3.5.3, we have

$$\lim_{k\to\infty} \Lambda_k = \bar{\Lambda} := h^{-1}(\bar{C}).$$

Let $\lambda_{\min,k}(\theta)$ be the smallest eigenvalue of $G^*(e^{i\theta})\Lambda_k G(e^{i\theta})$, and $\sigma_{\min,k}(\theta)$ be the smallest singular value of $C_k G(e^{i\theta})$. In view of (3.68), we have

$$\lambda_{\min,k}(\theta) = \sigma^2_{\min,k}(\theta)$$

By Lemma B.1.2, there exist a real number $\mu > 0$ such that

$$\lambda_{\min,k}(\theta) \geq \mu \implies \sigma_{\min,k}(\theta) \geq \sqrt{\mu}, \quad \forall k, \theta.$$

Then we have

$$\begin{aligned}
\|\delta\Phi(e^{i\theta}; C_k; \delta C)\|_2 &\leq 2\|(C_k G)^{-1}\delta C G \Phi_{C_k}\|_2 \\
&\leq 2\|(C_k G)^{-1}\|_2^3 \|\delta C G\|_2 \|\Psi\|_2 \\
&\leq \frac{2}{\sigma^3_{\min,k}(\theta)} \|\delta C G\|_F \|\Psi\|_F \leq K,
\end{aligned}$$

where the constant

$$K = \frac{2}{\mu^{3/2}} \max_\theta \|\delta C G(e^{i\theta})\|_F \max_\theta \|\Psi(e^{i\theta})\|_F.$$

We can now bound the integrand in (3.67). For any $\theta \in [-\pi, \pi]$ and $k \geq 1$, we have

$$\begin{aligned}
\left|\left[G\delta\Phi(e^{i\theta}; C_k; \delta C)G^*\right]_{j\ell}\right| &\leq \|G\delta\Phi(e^{i\theta}; C_k; \delta C)G^*\|_F \\
&\leq \kappa \|G\delta\Phi(e^{i\theta}; C_k; \delta C)G^*\|_2 \\
&\leq \kappa K \|G\|_2^2 \leq \kappa K G_{\max},
\end{aligned}$$

where κ is a constant for norm equivalence and $G_{\max}$ in (B.1). The last step is an application of Lebesgue's dominated convergence theorem to conclude

$$\lim_{k\to\infty} \delta\tau(C_k; \delta C) = \delta\tau(\bar{C}; \delta C),$$

which completes the proof. ∎

We are now left with the task of investigating whether the Jacobian of τ vanishes nowhere, which can be approached via the differential (3.67). However, the trick of orthogonality in the proof of Proposition 3.3.2 does not apply in a straightforward manner to the general map ω. The next result is due to (Baggio, 2018b). It is a small improvement over Proposition 3.3.2 and complements (Ferrante et al., 2010, Theorem 11.4.3), one of the main results in that paper. The proof uses essentially the same technique as in the scalar case.

Proposition 3.6.3. *If the prior $\Psi = \psi M$ with $\psi \in \mathfrak{S}_1$ and $M \in \mathfrak{H}_{+,m}$, then the Jacobian determinant of τ vanishes nowhere in $\mathscr{C}_+$, and hence the map ω is a diffeomorphism.*

Proof. First observe that we can rewrite the differential (3.66) as

$$\delta\Phi(z; C; \delta C) = -\Phi_C G^* \delta(C^*\Psi^{-1}C) G \Phi_C, \tag{3.69}$$

where $\delta(C^*\Psi^{-1}C)$ denotes the differential of $C^*\Psi^{-1}C$ w.r.t. C, i.e.,

$$\delta(C^*\Psi^{-1}C) = C^*\Psi^{-1}\delta C + \delta C^*\Psi^{-1}C.$$

Fix $C \in \mathscr{C}_+$ and let $\delta\tau(C; \delta C) = 0$ for some $\delta C \in \mathfrak{C}$. In view of (3.67), this implies that

$$\delta\Phi(z; C; \delta C) \in \ker \Gamma = (\operatorname{Range} \Gamma^*)^\perp,$$

which in view of (3.12), means

$$\begin{aligned} \langle G^* X G, \delta\Phi(z; C; \delta C)\rangle &= \operatorname{tr} \int G^* X G \, \delta\Phi(e^{i\theta}; C; \delta C) \\ &= -\operatorname{tr} \int G^* X G \, \Phi_C G^* \delta(C^*\Psi^{-1}C) G \Phi_C \\ &= 0, \quad \forall X \in \mathfrak{H}_n. \end{aligned} \tag{3.70}$$

When $\Psi = \psi M$, choosing $X = \delta(C^*M^{-1}C)$ in (3.70) will lead to the relation

$$\left\| \psi^{1/2} \left(\Phi_C^{1/2}\right)^* G^* \delta(C^*M^{-1}C) G \Phi_C^{1/2} \right\|_{L_2}^2 = 0,$$

where $\|\chi\|_{L_2}^2 := \operatorname{tr} \int \chi\chi^*$ and $\Phi_C^{1/2}$ is a spectral factor of Φ_C. By the same reasoning as in Proposition 3.3.2, this implies $G^*\delta(C^*M^{-1}C)G \equiv 0$ on the unit circle, or more explicitly,

$$G^*(C^*M^{-1}\delta C + \delta C^*M^{-1}C)G = 0, \quad \forall z \in \mathbb{T}.$$

In view of Proposition 3.5.2, the general solution to the above equation is

$$\delta C = MQC, \tag{3.71}$$

where Q is an $m \times m$ skew-Hermitian matrix. Equation (3.71) in particular implies

$$\delta CB = MQCB, \tag{3.72}$$

where δCB and CB lower triangular matrices with real diagonal entries and CB is invertible. Introduce also the Cholesky factorization $M = L_M L_M^*$. Then we have

$$\begin{aligned} Q &= M^{-1}\delta CB(CB)^{-1} \\ &= L_M^{-*}V, \end{aligned} \tag{3.73}$$

where $V := L_M^{-1}\delta CB(CB)^{-1}$ is lower triangular. The matrix Q being skew-Hermitian gives the relation

$$L_M^{-*}V + V^*L_M^{-1} = 0 \Longleftrightarrow VL_M + L_M^*V^* = 0.$$

We see that VL_M is skew-Hermitian and at the same time lower triangular, which means $VL_M = 0$. Therefore we must have $V = 0$ and $Q = 0$ by (3.73), which in turn implies $\delta C = 0$. ∎

3.7 Concluding Remarks

We have studied the multivariate spectral estimation problem formulated in a parametric fashion first introduced in (Ferrante et al., 2010). We have shown that the problem is well-posed with respect to the covariance data if the chosen prior is scalar. Moreover, the unique solution parameter depends also continuously on the scalar prior function while the covariance matrix is held fixed. Thus we have provided a complete proof of well-posedness in this special case, which will facilitate the design of numerical algorithms in the next chapter.

For the general case with an arbitrary matrix prior density, we have shown that the parametric spectral estimation problem admits a solution, and that well-posedness holds when the prior has the structure of a scalar density function times a constant positive definite matrix. The uniqueness question in general is still open and we hope to answer it in a future work.

Another research direction concerns the computation of a solution to the problem. To accomplish this task, in (Ferrante et al., 2010) the following matricial fixed-point iteration

was introduced

$$\Lambda_{k+1} = \int \Lambda_k^{1/2} G(W_{\Lambda_k}^{-1} \Psi W_{\Lambda_k}^{-*}) G^* \Lambda_k^{1/2}, \tag{3.74}$$

where the initialization is set to $\Lambda_0 = \frac{1}{n} I$. Iteration (3.74) can be seen as a multivariate generalization of the scalar algorithm proposed in (Pavon and Ferrante, 2006) for the Kullback–Leibler estimation of spectral densities. The latter algorithm has proved to be extremely efficient and numerically robust, and its convergence properties have been thoroughly investigated in (Ferrante et al., 2007, 2011; Baggio, 2018a). The extension of these convergence results to the multivariate case will be another subject of future investigation.

4

Numerical Solvers for the Spectral Estimation Problem

4.1 Introduction

This chapter concerns the numerical solution to the parametric spectral estimation problem. We consider only the case with a scalar prior in which results of well-posedness have been established in Section 3.3. There are two different ways to do the computation. One is to solve an optimization problem which was reported in (Avventi, 2011a), and the other is to numerically invert the moment map directly using a continuation method. We shall discuss the optimization problem briefly and focus more on the continuation method.

As for the optimization part, what interests us here is to do the optimization in the domain of spectral factors, namely, to perform a bijective change of optimization variables using the relation of spectral factorization studied in Section 3.5. This idea has been partially pursued in (Avventi, 2011a). However, the spectral factor introduced in that thesis for the optimization lives in a linear space which has a larger dimension than Range Γ. Hence there is redundancy in the chosen variables for optimization in (Avventi, 2011a), which stimulates the development here. In contrast, our optimization problem (after a change of variables) is still well-posed although convexity holds only locally. Convergence of descent algorithms is revisited.

The rest of this chapter is devoted to another numerical solver to compute the unique solution to the parametric spectral estimation problem using a continuation method which is a quite standard tool from nonlinear analysis. The idea is to solve a family of moment equations parametrized by one real variable, and to trace the solution curve from a known starting point based on the continuity result in Subsection 3.3.1. An algorithm called "predictor-corrector" (see (Allgower and Georg, 1990)) is adapted for the current problem. Thanks to the well-posedness of the problem, convergence of the algorithm is ensured when the step length is sufficiently small. The proof, inspired by (Enqvist, 2001), is built upon the Kantorovich theorem for the convergence of Newton iterations to solve nonlinear equations.

In both cases, we do computation in the domain of spectral factors as introduced in Section 3.5 due to the improvement of conditioning, as noted in (Enqvist, 2001; Avventi, 2011a), especially when the solution lies near the boundary of the feasible set.

The chapter is organized as follows. Following the result on spectral factorization in Section 3.5, Section 4.2 concerns the optimization problem studied in (Avventi, 2011a) reparametrized in terms of the spectral factor. We show that only local convergence can be guaranteed when using descent algorithms due to loss of convexity. Section 4.3 contains a numerical continuation solver to compute the solution parameter of the well-posed parametric problem (Section 3.3) again in the domain of spectral factors. Convergence of the proposed algorithm is investigated in detail. Moreover, a key computational step concerning the inverse Jacobian is elaborated.

4.2 Optimization in the Domain of Spectral Factors

As mentioned briefly in Remark 3.3.3, the unique solution in $\tilde{\mathscr{D}}$ (3.18) to the spectral estimation problem studied in the previous chapter has an interesting characterization that it solves the following optimization problem

$$\underset{\Phi\in\mathfrak{S}_m}{\text{minimize}} \quad \mathbb{S}(\psi||\Phi) := \int \psi \log\det(\psi\Phi^{-1}) \quad \text{subject to (3.3)} \tag{4.1}$$

where $\mathbb{S}(\psi||\Phi)$ is the multivariate counterpart of the Kullback-Leibler divergence between spectral densities studied in (Georgiou and Lindquist, 2003). The parameter Λ appears in the dual problem

$$\underset{\Lambda\in\mathscr{L}_+^\Gamma}{\text{minimize}} \quad \mathbb{J}_\psi(\Lambda) := \langle\Lambda,\Sigma\rangle - \int \psi \log\det(G^*\Lambda G). \tag{4.2}$$

It can be shown that the dual problem is strictly convex, and convergence of some numerical algorithms to solve the problem has been reported in (Avventi, 2011a).

Next we shall pursue the idea of reparametrizing the cost function in terms of the spectral factor as expressed in the relation (3.49). Given orthonormal bases (3.52), we can rewrite the cost as a function of the coordinates. More precisely, let us define

$$f(x) := \mathbb{J}_\psi(\Lambda(x)), \tag{4.3a}$$

$$g(y) := (f \circ h^{-1})(y) = \mathbb{J}_\psi(C^*(y)C(y)), \tag{4.3b}$$

where the map h^{-1} is understood in terms of the coordinates (3.53), and the second equality comes from the fact that the part of C^*C that is orthogonal to $\operatorname{Range}\Gamma$ plays no role in the evaluation of the function $\mathbb{J}_\psi$. The dual problem (4.2) can then be reformulated in terms of the spectral factor C as

$$\underset{C(y)\in\mathscr{C}_+}{\text{minimize}} \quad g(y). \tag{4.4}$$

This is a nonconvex problem since neither the feasibility set $\mathscr{C}_+$ nor the function g is convex. However, it still possesses a number of desired properties as stated below.

Proposition 4.2.1. *The function $g(y)$ has a unique stationary point $\hat{y}$ such that $C(\hat{y}) \in \mathscr{C}_+$, which is the unique solution of the optimization problem (4.4), and it is related to the unique stationary point $\hat{x}$ of $f(x)$ via the spectral factorization $\hat{y} = h(\hat{x})$. Moreover, the function $g(y)$ is strictly convex in a neighborhood of $\hat{y}$.*

Proof. One can proceed essentially in the same way as Propositions 4–7 in (Enqvist, 2001). For simplicity, let us rename the unique minimizer of (4.2) as $\hat{\Lambda} = \Lambda(\hat{x})$. Then $\hat{C} = C(\hat{y}) := h(\hat{\Lambda})$ is necessarily a minimizer of (4.4). If there is another $\tilde{C} \in \mathscr{C}_+$ with the same minimum value, then it must happen that $\hat{\Lambda} = h^{-1}(\tilde{C})$, which implies $\hat{C} = \tilde{C}$. This proves the uniqueness of the solution to (4.4).

From the relation (4.3b), we have

$$\nabla g(y) = J_{h^{-1}}(y)^\top \nabla f(h^{-1}(y)). \tag{4.5}$$

By Theorem 3.5.3, the Jacobian matrix of h^{-1} vanishes nowhere in $\mathscr{C}_+$. Therefore, y is a stationary point of g if and only if $h^{-1}(y)$ is a stationary point of f. By the same reasoning as above, the stationary point of g is unique and solves the optimization problem (4.4).

To see the last assertion, let us proceed to compute the Hessian of g

$$\nabla^2 g(y) = J_{h^{-1}}(y)^\top \nabla^2 f(h^{-1}(y)) J_{h^{-1}}(y) + \left(\frac{d}{dy} J_{h^{-1}}(y)^\top\right) \nabla f(h^{-1}(y))$$

where $\nabla^2 f(\cdot)$ is the Hessian matrix of f, and $\frac{d}{dy} J_{h^{-1}}(y)^\top$ denotes the "array" of second-order partials of h^{-1}. Evaluating at $y = \hat{y}$ coordinate of the minimizer $\hat{C}$, we must have $\nabla^2 g(\hat{y}) > 0$ since $\nabla^2 f(\cdot) > 0$, $J_{h^{-1}}(\cdot)$ is nonsingular, and $\nabla f(h^{-1}(\hat{y})) = 0$ due to stationarity. Because of the continuity of the Hessian, there exists $\delta_1 > 0$ such that $\nabla^2 g(y)$ is strictly positive definite in the closed ball

$$\overline{B}(\hat{y}, \delta_1) := \{y \in \mathbb{R}^M \,:\, \|y - \hat{y}\| \leq \delta_1\}. \tag{4.6}$$

■

4.2.1 Local Convergence of Descent Algorithms

A class of iterative algorithms known as descent algorithms produce a sequence of points $\{y_k\}_{k\geq 0}$ such that

$$y_{k+1} = y_k + \alpha_k p_k, \tag{4.7}$$

where p_k is the descent direction given by

$$B_k p_k = -\nabla g(y_k), \tag{4.8}$$

with $B_k > 0$, and $\alpha_k > 0$ is the step length.

Convergence of descent algorithms for the problem (4.4) has been studied in (Avventi, 2011a, Subsection A.5.4). However, the proof relies on Propositions A.4.1 and A.5.4 in the same paper whose assertions will in general fail if B_k in (4.8) is an arbitrary positive definite matrix. The reason is due to (Horn and Johnson, 2013, Theorem 4.5.15(a), p. 286), which states that the product of two Hermitian positive definite matrices is again Hermitian positive definite if and only if they commute. Nonetheless, at least local convergence can be guaranteed.

Proposition 4.2.2. *Consider the optimization problem* (4.4)*. Let the initial guess y_0 be close enough to the stationary point $\hat{y}$ of g, and let $\{y_k\}$ be generated by* (4.7) *with the direction p_k given by* (4.8)*. Suppose that*

$$0 < \beta\_I \leq B_k \leq \beta_+ I, \quad \forall k,$$

for some $\beta_-, \beta_+ \in \mathbb{R}$. Then one can determine the step length α_k such that the descent algorithm converges to $\hat{y}$.

Proof. Since the problem is locally strictly convex by Proposition 4.2.1, tools from convex analysis can be utilized in the neighborhood $\overline{B}(\hat{y}, \delta_1)$ in (4.6). The proof is an adaption of the procedure in (Boyd and Vandenberghe, 2004, p. 468). We shall specify $\delta > 0$ for y_0 to live in $B(\hat{y}, \delta)$ and determine a constant step length to achieve convergence.

First we can make $\delta \le \delta_1$ so that y_0 is in the strictly convex region. Let us recall some basic facts derived from convexity. Due to locally strict convexity of g, we have (cf. (Boyd and Vandenberghe, 2004, Subsection 9.1.2)) for some constants $\mu_-, \mu_+ \in \mathbb{R}$ and $\forall y \in \overline{B}(\hat{y}, \delta_1)$

$$0 < \mu_- I \le \nabla^2 g(y) \le \mu_+ I,$$

$$\|y - \hat{y}\| \le \sqrt{\frac{2}{\mu_-}(g(y) - g(\hat{y}))}, \tag{4.9}$$

$$\|\nabla g(y)\|^2 \ge 2\mu_-(g(y) - g(\hat{y})). \tag{4.10}$$

Moreover, we have the Mean Value Inequality

$$\|\nabla g(y)\| \le \mu_+ \|y - \hat{y}\|. \tag{4.11}$$

Let us determine first the step length α_0 such that

(i) $y_1 \in \overline{B}(\hat{y}, \delta_1)$;

(ii) $g(y_1) < g(y_0) + \eta\alpha_0 \nabla g(y_0)^\top p_0$ for some constant $\eta \in (0, 1)$.

By choosing the step length small enough, the point (i) above can be guaranteed. More precisely, choosing

$$\alpha_0 \le \beta_- \frac{\delta_1 - \|y_0 - \hat{y}\|}{\|\nabla g(y_0)\|}$$

will be sufficient, since it implies

$$\begin{aligned} \|y_1 - \hat{y}\| &\le \|y_0 - \hat{y}\| + \alpha_0 \|B_0^{-1} \nabla g(y_0)\| \\ &\le \|y_0 - \hat{y}\| + \frac{\alpha_0}{\beta_-} \|\nabla g(y_0)\| \le \delta_1. \end{aligned}$$

By the continuity of g, there exists $\delta_2 > 0$ such that the right-hand side of (4.9) can be made less than $\delta_1/2$ if $\|y - \hat{y}\| < \delta_2$. If we let $\|y_0 - \hat{y}\| < \delta := \min\{\delta_1, \delta_2\}$, by (4.9) we shall have $\|y_0 - \hat{y}\| < \delta_1/2$, which further implies

$$\beta_- \frac{\delta_1 - \|y_0 - \hat{y}\|}{\|\nabla g(y_0)\|} \ge \beta_- \frac{\delta_1 - \|y_0 - \hat{y}\|}{\mu_+ \|y_0 - \hat{y}\|} \ge \frac{\beta_-}{\mu_+},$$

where we have used (4.11), and we can thus choose $\alpha_0 \le \beta_-/\mu_+$.

The point (ii) is also called sufficient descent in the literature of optimization. To achieve

this, consider the second-order Taylor expansion for $y_1 \in \overline{B}(\hat{y}, \delta_1)$

$$\begin{aligned} g(y_1) &= g(y_0) + \alpha_0 \nabla g(y_0)^\top p_0 + \frac{1}{2}\alpha_0^2 p_0^\top \nabla^2 g(y_0 + \xi_0 p_0) p_0 \\ &\leq g(y_0) + \|\nabla g(y_0)\|^2 \left(-\frac{\alpha_0}{\beta_+} + \frac{\alpha_0^2}{2\beta_-^2}\mu_+ \right) \end{aligned} \tag{4.12}$$

where $0 < \xi_0 < \alpha_0$. It is then easy to verify that any $\alpha_0 < \frac{2\beta_-(\beta_- - \beta_+\eta)}{\mu_+\beta_+}$ with $\eta < \beta_-/\beta_+$ will satisfy the condition of sufficient descent, since it readily implies that the constant in the second line of (4.12) satisfies

$$-\frac{\alpha_0}{\beta_+} + \frac{\alpha_0^2}{2\beta_-^2}\mu_+ < -\frac{\eta\alpha_0}{\beta_-},$$

and at the same time, we have

$$-\frac{\eta\alpha_0}{\beta_-}\|\nabla g(y_0)\|^2 \leq \eta\alpha_0 \nabla g(y_0)^\top p_0.$$

In particular, we can choose a constant step length, e.g., $\alpha_0 = \frac{\beta_-(\beta_- - \beta_+\eta)}{\mu_+\beta_+}$.

Finally, we can set

$$\alpha_0 = \min\left\{ \frac{\beta_-}{\mu_+}, \frac{\beta_-(\beta_- - \beta_+\eta)}{\mu_+\beta_+} \right\}.$$

Due to the descent in objective function value, by (4.9) we must have again $\|y_1 - \hat{y}\| \leq \delta_1/2$ and the above reasoning holds also for future iterates. In this way, we obtain a sequence of iterates $\{y_k\} \subset \overline{B}(\hat{y}, \delta_1)$ with sufficient descent

$$g(y_{k+1}) < g(y_k) + \eta\alpha_k \nabla g(y_k)^\top p_k \tag{4.13}$$

for some constant $\eta < \beta_-/\beta_+$ at each step, with $\alpha_k \equiv \alpha_0$. Subtracting $g(\hat{y})$ on both sides of (4.13), we have

$$\begin{aligned} g(y_{k+1}) - g(\hat{y}) &< g(y_k) - g(\hat{y}) + \eta\alpha_k \nabla g(y_k)^\top p_k \\ &\leq g(y_k) - g(\hat{y}) - \frac{\eta\alpha_k}{\beta_+}\|\nabla g(y_k)\|^2 \\ &\leq \left(1 - \frac{2\eta\alpha_k\mu_-}{\beta_+}\right)(g(y_k) - g(\hat{y})) \end{aligned} \tag{4.14}$$

where the third inequality follows from (4.10). Apply the inequality recursively and we

obtain the relation

$$g(y_k) - g(\hat{y}) < \left(1 - \frac{2\eta\alpha_k\mu_-}{\beta_+}\right)^k (g(y_0) - g(\hat{y})).$$

Letting $\eta < \min\{\beta_-/\beta_+, 1/2\}$, the constant $1 - \frac{2\eta\alpha_k\mu_-}{\beta_+}$ is less than one and we have at least linear convergence locally. ■

4.3 A Continuation Solver for the Parametric Problem

In this section, we consider the problem of numerically solving the generalized moment equation in the parametric form directly without referring to the cost function in (4.2). Again we shall exploit the spectral factorization studied in Section 3.5 and reformulate Problem 3.2.2 in terms of the spectral factor of $G^*\Lambda G$ for $\Lambda \in \mathscr{L}_+^\Gamma$. Though it may appear slightly more complicated, this reformulation is preferred from a numerical viewpoint, as the Jacobian of the new map corresponding to (3.19) will have a smaller condition number when the solution is close to the boundary of the feasible set. This point has been illustrated in (Enqvist, 2001; Avventi, 2011a) (see also later in Subsection 4.3.2).

First let us introduce the moment map $\mathbf{g} : C_+(\mathbb{T}) \times \mathscr{C}_+ \to \text{Range}_+ \Gamma$ parametrized in the new variable C as

$$\mathbf{g}(\psi, C) := \mathbf{f}(\psi, h^{-1}(C)) = \int G\psi(G^*C^*CG)^{-1}G^*, \tag{4.15}$$

and the sectioned map when $\psi \in C_+(\mathbb{T})$ is held fixed

$$\tilde{\tau} := \tilde{\omega} \circ h^{-1} : \mathscr{C}_+ \to \text{Range}_+ \Gamma, \tag{4.16}$$

where $\tilde{\omega}$ has been defined in (3.21). A corresponding problem is formulated as follows.

Problem 4.3.1. Given the filter bank $G(z)$ in (3.2), the matrix $\Sigma \in \text{Range}_+ \Gamma$, and an arbitrary $\psi \in C_+(\mathbb{T})$, find the parameter $C \in \mathscr{C}_+$ such that

$$\tilde{\tau}(C) = \Sigma. \tag{4.17}$$

The next corollary is an immediate consequence of Theorems 3.3.7 and 3.5.3 and is stated without proof.

Corollary 4.3.2. *The map $\tilde{\tau}$ in (4.16) is a diffeomorphism. Moreover, if we fix the matrix Σ and allow the prior ψ to vary, then the solution map*

$$h \circ s(\,\cdot\,, \Sigma) : C_+(\mathbb{T}) \to \mathscr{C}_+ \tag{4.18}$$

is of class C^1.

Therefore, Problem 4.3.1 is also well-posed exactly like Problem 3.2.2. However, unlike the case of the map $\tilde{\omega}$ as mentioned in Remark 3.3.3, the equation (4.17) is different from the stationarity condition of the function $\mathbb{J}_\psi \circ h^{-1}$ for a fixed ψ. Therefore, we cannot deal with Problem 4.3.1 in the way of optimization. Nonetheless, it can be solved using a numerical continuation method as will be detailed next. We shall work with coordinates as explained in Subsection 3.5.1 whenever convenient. With reference to the coordinate representations of $\Lambda \in \mathscr{L}_+^\Gamma$ and $C \in \mathscr{C}_+$ in (3.52), we shall then introduce some abuse of notation and make no distinction between the variable and its coordinates. For example, $\mathbf{f}(\psi, x)$ is understood as $\mathbf{f}(\psi, \Lambda(x))$ defined in the previous chapter and similarly, $\tilde{\tau}(y)$ means $\tilde{\tau}(C(y))$.

Instead of dealing with one particular equation (4.17), a continuation method (cf. (Allgower and Georg, 1990)) aims to solve a family of equations related via a homotopy, i.e., a continuous deformation. In our context, there are two ways to construct different homotopies. One is to deform the covariance data Σ and study the equation (4.17) for a fixed ψ.[1] Such an argument has been used extensively in (Georgiou, 2005, 2006). Here we shall adopt an alternative, that is, deforming the prior function ψ while keeping the covariance matrix fixed, which can be seen as a multivariate generalization of the argument in (Enqvist, 2001, Section 4). An advantage to do so is that we can obtain a family of matrix spectral densities that are consistent with the covariance data.

The set $C_+(\mathbb{T})$ is easily seen to be convex. One can then connect ψ with the constant function $\mathbf{1}$ (taking value 1 on $\mathbb{T}$) via the line segment

$$p(t) = (1-t)\mathbf{1} + t\psi, \quad t \in U = [0,1], \tag{4.19}$$

and construct a convex homotopy $U \times \mathscr{C}_+ \to \mathrm{Range}_+ \Gamma$ given by

$$(t, y) \mapsto \mathbf{g}(p(t), y). \tag{4.20}$$

Now let the covariance matrix $\Sigma \in \mathrm{Range}_+ \Gamma$ be fixed whose coordinate vector is x_Σ, and consider the family of equations

$$\mathbf{g}(p(t), y) = x_\Sigma \tag{4.21}$$

[1] This will be reviewed in Chapter 5.

parametrized by $t \in U$. By Corollary 4.3.2, we will have a continuously differentiable solution path in the set $\mathscr{C}_+$

$$y(t) = h(s(p(t), x_\Sigma)). \tag{4.22}$$

Moreover, differentiating (4.21) on both sides w.r.t t, one gets

$$\mathbf{g}_1'(p(t), y(t); p'(t)) + \mathbf{g}_2'(p(t), y(t); y'(t)) = 0,$$

where $p'(t) \equiv \psi - \mathbf{1}$ independent of t, and the partial derivatives are given by

$$\mathbf{g}_1'(\psi, y) = \mathbf{f}_1'(\psi, h^{-1}(y)), \tag{4.23a}$$
$$\mathbf{g}_2'(\psi, y) = \mathbf{f}_2'(\psi, h^{-1}(y)) J_{h^{-1}}(y). \tag{4.23b}$$

The symbol $J_{h^{-1}}(y)$ means the Jacobian matrix of h^{-1} evaluated at y. Hence the path $y(t)$ is a solution to the initial value problem (IVP)

$$\begin{cases} y'(t) = -\left[\, \mathbf{g}_2'(p(t), y(t)) \,\right]^{-1} \mathbf{g}_1'(p(t), y(t); p'(t)) \\ y(0) = y^{(0)} \end{cases} \tag{4.24}$$

Notice that the partial $\mathbf{g}_2'$ is a finite-dimensional Jacobian matrix which is invertible everywhere in its domain of definition since both terms on the right hand side of (4.23b) are nonsingular (cf. Chapter 3). From classical results on the uniqueness of solution to an ODE, we know that the IVP formulation and (4.21) are in fact equivalent.

The initial value $y^{(0)}$ corresponds to $\psi = \mathbf{1}$, and it is the spectral factor of the so-called maximum entropy solution, i.e., solution to the problem

$$\underset{\Phi \in \mathfrak{S}_m}{\text{maximize}} \int \log\det\Phi \quad \text{subject to (3.3)}. \tag{4.25}$$

As has been worked out in (Georgiou, 2002a), the above optimization problem has a unique solution $\Phi = (G^* \Lambda G)^{-1}$ with

$$\Lambda = \Sigma^{-1} B (B^* \Sigma^{-1} B)^{-1} B^* \Sigma^{-1},$$

from which the corresponding spectral factor C can be computed as

$$C = L^{-*} B^* \Sigma^{-1}, \tag{4.26}$$

where L is the right Cholesky factor of $B^* \Sigma^{-1} B$. According to (Georgiou, 2002a), such C is indeed in the set $\mathscr{C}_+$, i.e., CB lower triangular and the closed-loop matrix is stable.

At this stage, any numerical ODE solver can in principle be used to solve the IVP and obtain the desired solution $y(1)$ corresponding to a particular prior ψ. However, this IVP is special in the sense that for a fixed t, $y(t)$ is a solution to a finite-dimensional nonlinear system of equations, for which there are numerical methods (such as Newton's method) that exhibit rapid local convergence properties, while a general-purpose ODE solver does not take this into account. Out of such consideration, a method called "predictor-corrector" is recommended in (Allgower and Georg, 1990) to solve the IVP, which is reviewed next.

Suppose for some $t \in U$ we have got a solution $y(t)$ and we aim to solve (4.21) at $t + \delta t$ where δt is a chosen step length. The predictor step is just numerical integration of the differential equation in (4.24) using e.g., the Euler method

$$z(t + \delta t) := y(t) + v(t)\delta t, \tag{4.27}$$

where $v(t) := -\left[\, \mathbf{g}_2'(p(t), y(t)) \,\right]^{-1} \mathbf{g}_1'(p(t), y(t); p'(t))$. The corrector step is accomplished by the Newton's method to solve (4.21) initialized at the predictor $z(t + \delta t)$. If the new solution $y(t + \delta t)$ can be attained in this way, one can repeat such a procedure until reaching $t = 1$. The algorithm is summarized in the table.

Algorithm 4.1 Predictor-Corrector

Let $k = 0$, $t = 0$, and $y^{(0)}$ initialized as in (4.26)
Choose a sufficiently small step length δt
while $t \leq 1$ **do**
  Predictor: $z^{(k+1)} = y^{(k)} + v(t)\delta t$ the Euler step (4.27)
  Corrector: solve (4.21) at $t + \delta t$ for $y^{(k+1)}$ initiated at $z^{(k+1)}$ using Newton's method
  Update $t := \min\{1, t + \delta t\}$, $k := k + 1$
end while
return The last $y^{(k)}$ corresponding to $t = 1$

4.3.1 Convergence Analysis

We are now left to determine the step length δt so that the corrector step can converge and the algorithm can return the target solution $y(1)$ in a finite number of steps. We show next that one can choose a uniformly constant step length δt such that the predictor $z^{(k)}$ will be close enough to the solution $y^{(k)}$ for the Newton's method to converge locally.[2] We shall need the next famous Kantorovich theorem which can be found in (Ortega and Rheinboldt, 2000, p. 421).

[2]Notice that convergence results in (Allgower and Georg, 1990) under some general assumptions do not apply here directly.

Theorem 4.3.3 (Kantorovich). *Assume that $f : D \subset \mathbb{R}^n \to \mathbb{R}^n$ is differentiable on a convex set $D_0 \subset D$ and that*

$$\|f'(x) - f'(y)\| \le \gamma \|x - y\|, \quad \forall\, x, y \in D_0$$

for some $\gamma > 0$. Suppose that there exists an $x^{(0)} \in D_0$ such that $\alpha = \beta\gamma\eta \le 1/2$, for some $\beta, \eta > 0$ meeting

$$\beta \ge \|f'(x^{(0)})^{-1}\|, \quad \eta \ge \|f'(x^{(0)})^{-1} f(x^{(0)})\|.$$

Set

$$t^* = (\beta\gamma)^{-1}\left[\, 1 - (1-2\alpha)^{1/2} \,\right], \tag{4.28a}$$

$$t^{**} = (\beta\gamma)^{-1}\left[\, 1 + (1-2\alpha)^{1/2} \,\right], \tag{4.28b}$$

and assume that the closed ball $\overline{B}(x^{(0)}, t^*)$ is contained in D_0. Then the Newton iterates

$$x^{(k+1)} = x^{(k)} - f'(x^{(k)})^{-1} f(x^{(k)}), \quad k = 0, 1, \ldots$$

are well-defined, remain in $\overline{B}(x^{(0)}, t^*)$, and converge to a solution x of $f(x) = 0$ which is unique in $\overline{B}(x^{(0)}, t^{**}) \cap D_0$.

In order to apply the above theorem, we need to take care of the locally Lipschitz property. To this end, we shall first introduce a compact set in which we can take extrema of various norms.

Lemma 4.3.4. *There exists a compact set $K \subset \mathscr{C}_+$ that contains the solution path $\{y(t) : t \in U\}$ given by (4.22) in its interior.*

Proof. We know from previous reasoning that the solution path is contained in the open set $\mathscr{C}_+$. By continuity, the set $\{y(t)\}$ is easily seen to be compact, i.e., closed and bounded, and thus admits a compact neighborhood $K \subset \mathscr{C}_+$. Such a neighborhood K can be constructed explicitly as follows. Let $B(y(t)) \subset \mathscr{C}_+$ be an open ball centered at $y(t)$ such that its closure is also contained in $\mathscr{C}_+$. Then the set $\bigcup_{t\in U} B(y(t))$ is an open cover of $\{y(t)\}$, which by compactness, has a finite subcover

$$\bigcup_{k=1}^{n} B(y(t_k))$$

whose closure can be taken as K. ∎

Lemma 4.3.5. *For a fixed $t \in U$, the derivative $\mathbf{g}_2'(p(t), y) \in L(\mathfrak{C}, \operatorname{Range}\Gamma)$ is locally Lipschitz continuous in y in any convex subset of the compact set K constructed in Lemma 4.3.4, where $p(t)$ is the line segment given in (4.19). Moreover, the Lipschitz constant can be made independent of t.*

Proof. It is a well known fact a continuously differentiable function is locally Lipschitz. Hence we need to check the continuity of the second-order derivative following from (4.23b)

$$\begin{aligned}\mathbf{g}''_{22}(\psi, y; \delta y_1, \delta y_2) =& \mathbf{f}''_{22}(\psi, h^{-1}(y); J_{h^{-1}}(y)\delta y_2, J_{h^{-1}}(y)\delta y_1) \\ &+ \mathbf{f}'_2(\psi, h^{-1}(y))\frac{d}{dy}J_{h^{-1}}(y)(\delta y_2, \delta y_1).\end{aligned} \tag{4.29}$$

Here $\frac{d}{dy}J_{h^{-1}}(y)$ is the second order derivative of h^{-1} evaluated at y which is viewed as a bilinear function $\mathfrak{C} \times \mathfrak{C} \to \operatorname{Range}\Gamma$. It is continuous in y since the function (3.51) is actually smooth (of class C^∞). Albeit more tedious computation, one can also show that the second-order partial $\mathbf{f}''_{22}(\psi, x)$ is continuous following the lines in Subsection 3.3.1. Therefore, for fixed $\delta y_1, \delta y_2$, the differential (4.29) is continuous in (ψ, y). Consider any convex subset $D_0 \subset K$ with $y_1, y_2 \in D_0$. By the mean value theorem we have

$$\begin{aligned}&\|\mathbf{g}'_2(\psi, y_2) - \mathbf{g}'_2(\psi, y_1)\| \\ =&\left\| \int_0^1 \mathbf{g}''_{22}(\psi, y_1 + \xi(y_2 - y_1))d\xi(y_2 - y_1) \right\| \\ \le& \max_{y\in K} \|\mathbf{g}''_{22}(\psi, y)\| \|y_2 - y_1\|.\end{aligned} \tag{4.30}$$

Now let us replace ψ with $p(t)$. The local Lipschitz constant can be taken as

$$\gamma := \max_{t\in U,\ y\in K} \|\mathbf{g}''_{22}(p(t), y)\|. \tag{4.31}$$

■

Based on precedent lemmas, our main result in this section is stated as follows.

Theorem 4.3.6. *Algorithm 4.1 returns a solution to (4.21) for $t = 1$ in a finite number of steps.*

Proof. At each step, our task is to solve the equation (4.21) for $t + \delta t$ from the initial point $z(t+\delta t) = y(t) + v(t)\delta t$ given in (4.27). The idea is to work in the compact set K introduced in Lemma 4.3.4. The boundary of K is denoted by ∂K which is also compact.

First, we show that the predictor $z(t + \delta t)$ will always stay in K as long as the step length δt is sufficiently small. Define

$$c_1 := \min_{t\in U} d(y(t), \partial K), \tag{4.32}$$

$$c_2 := \max_{t\in U} \|v(t)\|, \tag{4.33}$$

where $d(x, A) := \min_{y\in A} d(x, y)$ is the distance function from a point x to a set A. Note that

$c_1 > 0$ because all the points $\{y(t)\}$ are in the interior of K. Then we see that the condition

$$\delta t < \frac{c_1}{c_2} := \delta t_1 \tag{4.34}$$

is sufficient since in this way

$$\|v(t)\delta t\| \le c_2 \delta t < c_1 \le d(y(t), \partial K), \ \forall t \in U,$$

which implies that $z(t+\delta t) \in K$. The reason is that one can always go from $y(t)$ in the direction of $v(t)$ until the boundary of K is hit. Hence we are safe to take the step length $\delta t = \delta t_1/2$.

Secondly, we want to apply the Kantorovich Theorem to ensure convergence of the corrector step, i.e., the Newton iterates. The function $\psi = p(t+\delta t)$ is held fixed in the corrector step. The uniform Lipschitz constant γ has been given in Lemma 4.3.5, and there are two remaining points:

(i) We need to take care of the constraint $\alpha = \beta\gamma\eta \le 1/2$. Clearly, we can simply take $\beta = \|\mathbf{g}_2'(\psi, y_{\text{in}}^{(0)})^{-1}\|$ and

$$\eta = \left\| \mathbf{g}_2'(\psi, y_{\text{in}}^{(0)})^{-1} \left(\mathbf{g}(\psi, y_{\text{in}}^{(0)}) - x_\Sigma \right) \right\|,$$

where $y_{\text{in}}^{(0)} = z(t+\delta t)$ is the initialized inner-loop variable. Define

$$c_3 := \max_{y \in K, t \in U} \|\mathbf{g}_2'(p(t), y)^{-1}\|, \tag{4.35}$$

$$c_4 := \max_{y \in K, t \in U} \|\mathbf{g}_{22}''(p(t), y)\|, \tag{4.36}$$

and obviously we have $\beta \le c_3$, $\eta \le c_3 \|\mathbf{g}(\psi, y_{\text{in}}^{(0)}) - x_\Sigma\|$. Hence a sufficient condition is

$$\|\mathbf{g}(\psi, y_{\text{in}}^{(0)}) - x_\Sigma\| \le \frac{1}{2c_3^2\gamma},$$

and we need an estimate of the left hand side. The Taylor expansion of $\mathbf{g}$ in its second argument is

$$\mathbf{g}(\psi, y(t) + v(t)\delta t) = \mathbf{g}(\psi, y(t)) + \delta t \mathbf{g}_2'(\psi, y(t))v(t) + \frac{\delta t^2}{2}\mathcal{B}[v(t), v(t)], \tag{4.37}$$

where $\mathcal{B}$ is the bilinear function determined by the second order partials. Due to linearity

and the identity $\psi = p(t+\delta t) = p(t) + \delta t p'(t)$, the first term

$$\begin{aligned} \mathbf{g}(p(t+\delta t), y(t)) &= \mathbf{g}(p(t), y(t)) + \delta t \mathbf{g}_1'(p(t), y(t); p'(t)) \\ &= x_\Sigma + \delta t \mathbf{g}_1'(p(t), y(t); p'(t)). \end{aligned} \tag{4.38}$$

The matrix in the second term[3]

$$\mathbf{g}_2'(p(t+\delta t), y(t)) = \mathbf{g}_2'(p(t), y(t)) + \delta t \mathbf{g}_2'(p'(t), y(t)) \tag{4.39}$$

Substituting these two expressions into (4.37), we obtain a cancellation due to the definition of $v(t)$ after (4.27) and we have

$$\mathbf{g}(\psi, y(t) + v(t)\delta t) - x_\Sigma = \delta t^2 \mathbf{g}_2'(p'(t), y(t)) v(t) + \frac{\delta t^2}{2} \mathcal{B}[v(t), v(t)] \tag{4.40}$$

whose norm is less than $\delta t^2 (c_5 c_2 + \frac{1}{2} c_2^2 c_4)$, where

$$c_5 := \max_{y \in K} \|\mathbf{g}_2'(p'(t), y)\|.$$

We end up having the sufficient condition

$$\delta t^2 (c_5 c_2 + \frac{1}{2} c_2^2 c_4) \leq \frac{1}{2 c_3^2 \gamma} \implies \delta t \leq \delta t_2.$$

(ii) We need to insure that the closed ball $\overline{B}(y_{\text{in}}^{(0)}, t^*)$ is also contained in K. Clearly, we only need to make $t^* \leq \min_{t \in U} d(p(t) + v(t)\delta t_1/2, \partial K) =: r_2$, where δt_1 is the uniform step determined in (4.34). This can be done since by its definition (4.28a), t^* tends to 0 when the step length $\delta t \to 0$. A sufficient condition is

$$1 - \sqrt{1 - 2\alpha} \leq c_6 \gamma r_2 \iff \alpha \leq \frac{1}{2}(1 - (1 - c_6 \gamma r_2)^2),$$

provided that $1 - c_6 \gamma r_2 > 0$, where

$$c_6 := \min_{y \in K, t \in U} \|\mathbf{g}_2'(p(t), y)^{-1}\|.$$

With the bound for β and η in the previous point, a more sufficient condition is

$$\delta t^2 \gamma c_3^2 (c_5 c_2 + \frac{1}{2} c_2^2 c_4) \leq \frac{1}{2}(1 - (1 - c_6 \gamma r_2)^2),$$

[3] Attention: $\mathbf{g}_2'(p'(t), y(t))$ is an abuse of notation because $p'(t) = \psi - \mathbf{1}$ may not be in the domain of the functional. It should be understood as substituting ψ with $p'(t)$ in the expression of $\mathbf{g}_2'(\psi, y(t))$.

which implies $\delta t \leq \delta t_3$ (constant).

At last we can just take $\delta t := \min\{\delta t_1/2, \delta t_2, \delta t_3\}$. In this way, the Kantorovich theorem is applicable to ensure local convergence in each inner loop. The reasoning above is independent of t and hence the step length is uniform. This concludes the proof. ∎

4.3.2 Computation of the Inverse Jacobian

The coordinate thinking is suitable for theoretical reasoning. However, when implementing the algorithm, it is better to work with matrices directly. In this section, we present a matricial linear solver adapted from (Ramponi et al., 2009) (see also (Ferrante et al., 2012a)). Here we shall assume ψ is rational and admits a factorization $\psi = \sigma\sigma^*$ where σ is outer rational and hence realizable. A crucial step in the implementation of the numerical algorithm is the computation of the Newton direction $\mathbf{g}_2'(\psi, y)^{-1}\mathbf{g}(\psi, y)$, which amounts to solving the linear equation in V given C and ψ

$$\mathbf{g}_2'(\psi, C; V) = \mathbf{g}(\psi, C) \tag{4.41}$$

where

$$\mathbf{g}(\psi, C) = \int G\psi(G^*C^*CG)^{-1}G^* \tag{4.42a}$$

$$\mathbf{g}_2'(\psi, C; V) = -\int G\psi(G^*C^*CG)^{-1}G^*(V^*C + C^*V)G(G^*C^*CG)^{-1}G^* \tag{4.42b}$$

$$= -\int G\psi(CG)^{-1}\left[(G^*C^*)^{-1}G^*V^* + VG(CG)^{-1}\right](G^*C^*)^{-1}G^* \tag{4.42c}$$

The cancellation of one factor CG from (4.42b) to (4.42c) is precisely why the condition number of the Jacobian $\mathbf{g}_2'$ is smaller than that of $\mathbf{f}_2'$ in (3.31) when C tends to the boundary of $\mathscr{C}_+$, i.e., when $CG(e^{i\theta})$ tends to be singular for some θ. This point is illustrated in the next example.

Example 4.3.7 (Reduction of the condition number of the Jacobian under the C parametrization)**.** First we need to find a matrix representation of the Jacobian $g_2'(\psi, C)$ which is a linear operator from $\mathfrak{C}$ to Range Γ. Fix the orthonormal bases of the two vector spaces as in (3.52). Then the (j,k) element of the *real* $M \times M$ Jacobian matrix corresponding to $g_2'(\psi, C)$ is just

$$\langle \mathbf{\Lambda}_j, g_2'(\psi, C; \mathbf{C}_k)\rangle. \tag{4.43}$$

Similarly, the matrix representation of the Jacobian (3.31) is

$$\langle \mathbf{\Lambda}_j, f_2'(\psi, \Lambda; \mathbf{\Lambda}_k) \rangle, \quad j,k = 1, \ldots, M. \tag{4.44}$$

Next let us fix $\psi \in C_+(\mathbb{T})$ and $C \in \mathscr{C}_+$, evaluate explicitly the Jacobian matrix (4.43), and compute its condition number. The same computation is done for the Jacobian matrix (4.44) evaluated at $(\psi, h^{-1}(C))$ where h^{-1} has been defined in (**??**). Comparison is made in such a way because taking $\Lambda = h^{-1}(C)$ will lead to the same spectrum in the moment map f as that in g due to the spectral factorization (3.49).

Our example is about the problem of matrix covariance extension mentioned in Section **??** with (A, B) matrices given in (**??**) and the filter $G(z)$ in (3.5). We set the dimension $m = 2$, the maximal covariance lag $p = 1$, and we have $n = m(p+1) = 4$.

Let us treat the problem for real processes. Then Range Γ is the $M = 7$-dimensional vector space of symmetric block-Toeplitz matrices of the form

$$\begin{bmatrix} \Lambda_0 & \Lambda_1^\top \\ \Lambda_1 & \Lambda_0 \end{bmatrix}, \tag{4.45}$$

where Λ_0, Λ_1 are 2×2 blocks. An orthonormal basis of Range Γ can be determined from the matrix pairs

$$\begin{aligned} (\Lambda_0, \Lambda_1) \in \{\mathbf{0}\} \times \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\} \\ \bigcup \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\} \times \{\mathbf{0}\} \end{aligned} \tag{4.46}$$

after normalization. Here the bold symbol $\mathbf{0}$ denotes the 2×2 zero matrix.

On the other hand, the vector space $\mathfrak{C}$ contains matrices of the shape

$$\begin{bmatrix} C_1 & C_0 \end{bmatrix},$$

where C_1, C_0 are also 2×2 blocks and C_0 is lower triangular. An orthonormal basis of $\mathfrak{C}$ can be determined from the standard basis of $\mathbb{R}^{m \times n}$ which is made up of matrices E_{jk} whose elements are all zero except that on (j,k) position it is one. A basis of $\mathfrak{C}$ is obtained by excluding those E_{jk} which constitute the (strict) upper triangular part of C_0. Notice that given $C \in \mathfrak{C}$, zCG is a matrix polynomial of degree $-p$.

The prior is chosen as a positive Laurent polynomial $\psi(z) = b(z)b(z^{-1})$ where the polynomial $b(z) = 1 - z^{-1} + 0.89z^{-2}$ has roots $0.5 \pm 0.8i$ with a modulus 0.9434. We choose

the parameter

$$C = \begin{bmatrix} 0.5 & 0.65 & 1 & 0 \\ -2.2615 & -1 & 2 & 1 \end{bmatrix}, \tag{4.47}$$

which belongs to the set $\mathscr{C}_+$, because the roots of $\det zCG$ are $0.9 \pm 0.4i$ with a modulus 0.9849.

Integrals such as (4.42c) are approximated with Riemann sums

$$\int F(\theta) \approx \frac{\Delta\theta}{2\pi} \sum_k F(\theta_k),$$

where $\{\theta_k\}$ are equidistant points on the interval $(-\pi, \pi]$ and the subinterval length $\Delta\theta = 10^{-4}$. The resulting condition number of (4.43) is 2.4674×10^5 while that of (4.44) is 3.8187×10^8.

In order to invert the Jacobian g_2' at a given "point" (ψ, C) without doing numerical integration, we first need to fix an orthonormal basis $\{C_1, \ldots, C_M\}$ of $\mathfrak{C}$ such that $C_1 = C/\|C\|$.[4] Then one can obtain a basis $\{V_1, \ldots, V_M\}$ of $\mathfrak{C}$ such that for $k = 1, \ldots, M$

$$G^*(z)(V_k^* C + C^* V_k)G(z) > 0, \quad \forall z \in \mathbb{T}$$

by setting $V_k = C_k + \alpha_k C$ for some $\alpha_k \geq 0$. The procedure for solving (4.41) is described as follows:

1) Compute $Y = \mathbf{g}(\psi, C)$ and $Y_k = \mathbf{g}_2'(\psi, C; V_k)$.

2) Find α_k such that $Y = \sum \alpha_k Y_k$.

3) Set $V = \sum \alpha_k V_k$.

In order to obtain the coordinates α_k in Step 2, one needs to solve a linear system of equations whose the coefficient matrix is consisted of inner products $\langle Y_k, Y_j \rangle$. The matrix is invertible because $\{Y_k\}$ are linearly independent, which is a consequence of the Jacobian $\mathbf{g}_2'(\psi, C)$ being nonsingular.

The difficult part is Step 1 where we need to compute the integrals $\mathbf{g}(\psi, C)$ and $\mathbf{g}_2'(\psi, C; V_k)$. Since we want to avoid numerical integration, we shall need some techniques from spectral factorization. Evaluation of the former integral was essentially done in the proof of (Ferrante et al., 2010, Theorem 11.4.3). More precisely, we have the expression

$$G(zCG)^{-1} = (zI - \Pi)^{-1} B (CB)^{-1},$$

[4] This is always possible by adding C into any set of basis matrices and performing Gram-Schmidt orthonormalization starting from C.

where $\Pi := A - B(CB)^{-1}CA$ is the closed-loop matrix which is stable. With a state-space realization (A_1, B_1, C_1) of the stable proper transfer function $\sigma G(zCG)^{-1}$, one then solves a discrete-time Lyapunov equation for R

$$R - A_1 R A_1^* = B_1 B_1^*.$$

Finally the integral $\mathbf{g}(\psi, C) = C_1 R C_1^*$.

The integral $\mathbf{g}_2'(\psi, C; V_k)$ can be computed similarly. The only difference is that we need to compute a left outer factor $W(z)$ of

$$Z^*(z) + Z(z) > 0 \text{ on } \mathbb{T}$$

where

$$\begin{aligned} Z(z) &= zVG(zCG)^{-1} \\ &= Vz(zI - \Pi)^{-1}B(CB)^{-1} \\ &= V\Pi(zI - \Pi)^{-1}B(CB)^{-1} + VB(CB)^{-1} \end{aligned} \tag{4.48}$$

The factorization involves solving a DARE for the unique stabilizing solution, in terms of which the factor can be expressed. Such a procedure is standard (cf. Appendix C for details). Once we have the factor $W(z)$, a realization of the transfer function $\sigma G(zCG)^{-1}W$ can be obtained, and we can just proceed in the same way as computing $\mathbf{g}(\psi, C)$.

Example 4.3.8. Let us continue Example 4.3.7 with the prior ψ and the parameter C given. We begin the simulation by computing the covariance matrix $\Sigma = g(\psi, C)$ with the formula given in (4.15). Then the maximum entropy solution can be obtained with (4.26), which is our initial value of Algorithm 4.1. The step length is set as $\delta t = 0.1$ which is quite large but sufficient for convergence in this particular example. Our target parameter $C^{(1)}$ is certainly equal to the given C in (4.47). The simulation result is shown in Figures 4.1 and 4.2, where the coordinates of the solution parameter are plotted against the variable $t \in [0, 1]$. One can see that the solution curves are smooth.

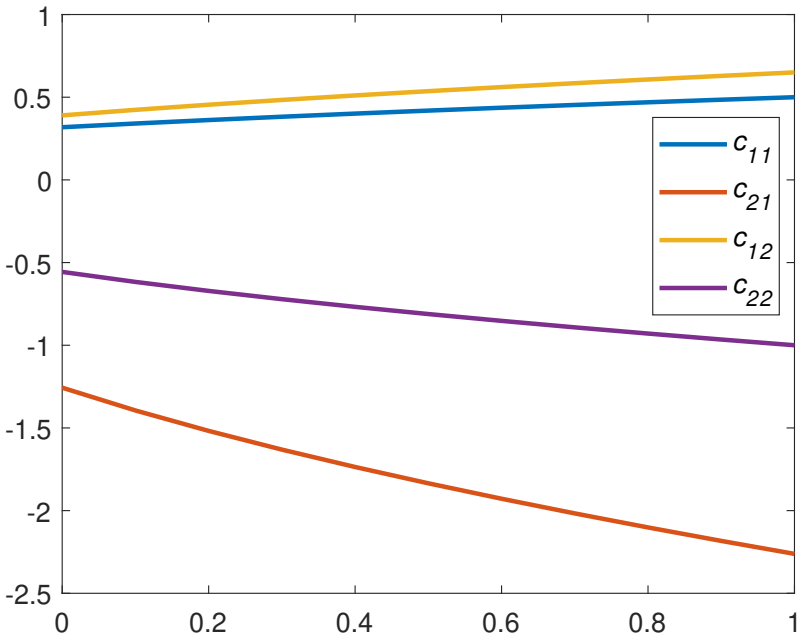

Figure 4.1: Solution parameter (coordinates) against the variable t.

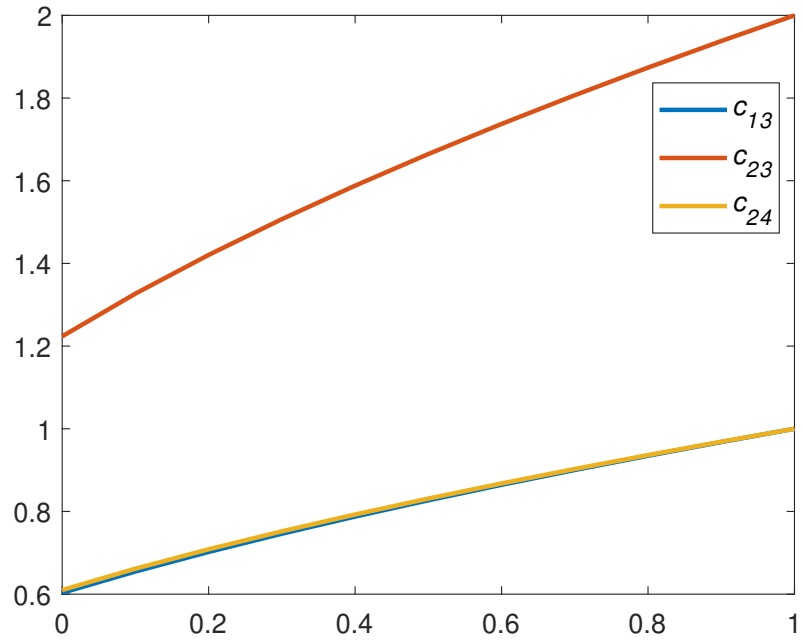

Figure 4.2: Solution parameter (coordinates) against the variable t.

4.4 Conclusions

We have indicated how to numerically solve the spectral estimation problem given a scalar prior. Although the problem is well-posed, in practice the Jacobian of the moment map may become ill-conditioned when the parameter goes near the boundary. Such a numerical issue can be alleviated if we carry out computations in the domain of spectral factors.

The optimization approach to the spectral estimation problem has been well studied in (Avventi, 2011a). When reparametrized in terms of the spectral factor, the cost function becomes only locally convex, and descent algorithms are proven to converge locally. The second numerical solver is built upon the continuation method to trace a family of solutions parametrized by a real variable living on the unit interval. Although the resulting algorithm seems more complicated than the optimization approach, it serves as a viable alternative.

At last, we hope to generalize the results in this chapter to the open problem left in Chapter 3 when the prior function is matrix-valued. One technical difficulty, namely, uniqueness of the solution in that more general case remains to be tackled. One can expect that once well-posedness is established, the numerical continuation procedure to find the solution will be in a sense standard.

5

On an Alternative Parametrization of Matricial Spectral Densities

5.1 Introduction

In this chapter, we consider the same multivariate spectral estimation problem as in Chapter 3, although in a different family of matrix spectral densities. As described in Chapters 1 and 3, in the "THREE" approach to the problem of spectral estimation, the steady-state covariance matrix of the output process of a rational filter is used as data for the reconstruction of the input spectrum, which naturally admits a formulation as a generalized moment problem. Due to the typical ill-posedness of moment problems (Grenander and Szegö, 1958; Kreĭn and Nudel'man, 1977), *entropy*(-like) functionals are then exploited as optimization criteria to promote uniqueness of the solution. More specifically, one tries to find the input spectrum consistent with the output covariance matrix that maximizes some entropy or minimizes some distance index to an *a priori* spectral density.

Different choices of cost functionals lead to different forms of solutions, especially in the multivariate case (cf. (Georgiou, 2002a, 2006; Ferrante et al., 2008; Ramponi et al., 2009, 2010; Avventi, 2011a; Ferrante et al., 2012a; Zorzi, 2014a, 2015; Georgiou and Lindquist, 2017)). Among them (Georgiou, 2006) is an important work utilizing the following relative

entropy as the optimization criterion

$$\mathbb{S}(\Phi|\Psi) = \int_{\mathbb{T}} \mathrm{tr}\left[\Phi\left(\log\Phi - \log\Psi\right)\right]$$

which in turn, draws inspiration from quantum mechanics. Here Ψ is the known prior and $\mathbb{T}$ stands for the unit circle. Minimization of $\mathbb{S}(\Phi|\Psi)$ with respect to Φ subject to the generalized moment constraint can be worked out explicitly leading to an exponential-type spectral density. Such a solution can also be recovered as a limit case of a family of solutions based on the multivariate Beta divergence discussed in (Zorzi, 2014a). Difficulty arises in the other direction, namely, minimization of $\mathbb{S}(\Psi|\cdot)$ with respect to the second argument. As reported in (Georgiou, 2006), variational analysis and duality reasoning hit an obstruction in the middle because the functional dependence of the optimal primal on the dual variable cannot be described in a closed form (see also (Ferrante et al., 2008)). As a response to this difficulty, Theorem 6 of (Georgiou, 2006) suggests to "forgo an explicit form for the entropy functional and start instead with a computable Jacobian". In other words, a parametric form of the spectral density has been proposed, which possibly does not correspond to any cost functional. Although the statement of that theorem looks rather exciting, it is extremely nontrivial and its validity remains elusive as a rigorous proof is absent. In this chapter, we are motivated to address this issue. We shall only consider the first half of (Georgiou, 2006, Theorem 6) concerning *rational* solutions to the spectral estimation problem.

The continuation argument is used extensively in the proofs of (Georgiou, 2006) which follows the previous work (Georgiou, 2005) in the scalar case by the same author. As will be reviewed later in Section 5.4, in order for the argument to be effective, the Jacobian of the parametric moment map is required to vanish nowhere in the feasible set, which is fulfilled when the prior is taken as $\Psi = \psi I$, namely a scalar spectral density function times the identity matrix. In this chapter, we show through a numerical example that the requirement of everywhere nonvanishing Jacobian is not met in general by the moment map in question when the prior is nontrivial, contrary to what is claimed in (Georgiou, 2006, Section IV). Furthermore, a critical point of the moment map is computed in the example and demonstrated to be a bifurcation point. In consequence, the parametric solution to the spectral estimation problem considered in (Georgiou, 2006, Section IV) is generally not unique.

This chapter is organized as follows. In Section 5.2, we review the parametric form of the moment map introduced in (Georgiou, 2006) that will be the central object of investigation in this chapter. We give a numerical example in Section 5.3 where a critical point of the moment map is detected and computed. In Section 5.4, we apply a part of the bifurcation

theory and carry out some further computation which allows us to conclude that the afore obtained critical point is in fact a bifurcation point. Finally, we make some remarks on an alternative parametrization of rational spectral densities.

5.2 Problem Review

One of the problems considered in (Georgiou, 2006) is about finding a matrix spectral density function in a particular parametric family that satisfies a (generalized) moment constraint. The problem setup is the same as in Chapter 3, only that the family of spectral densities considered is different from (3.14). We shall first review the problem and restate one of the main results of (Georgiou, 2006). Some notational differences are highlighted below to avoid confusion.

- The symbol used in (Georgiou, 2006) for the linear operator Γ defined in (3.9) is L.
- Ψ is a bounded and coercive $m \times m$ spectral density function. It admits a (unique) left outer factor W_Ψ, namely, $\Psi = W_\Psi W_\Psi^*$. The notations used in (Georgiou, 2006) for Ψ and its factor are σ and $\sigma^{1/2}$, respectively.
- $$\kappa : \Lambda \mapsto \int G W_\Psi (G^* \Lambda G)^{-1} W_\Psi^* G^* \tag{5.1}$$
  is a map from $\mathscr{L}_+^\Gamma$ to $\mathrm{Range}_+ \Gamma$. This is the map h_σ defined in (Georgiou, 2006, Section IV). The domain and codomain of κ coincides with our ω map defined in (3.16), and they are denoted with $\mathcal{K}_+^{\mathrm{dual}}$ and $\mathrm{int}(\mathcal{K})$ in (Georgiou, 2006), respectively. Moreover, the argument Λ is lower cased in (Georgiou, 2006).

Theorem 6 of (Georgiou, 2006) states that the map κ is a bijection given any bounded and coercive prior Ψ. In other words, given any positive definite matrix $\Sigma \in \mathrm{Range}\,\Gamma$, there exists a unique parameter $\Lambda \in \mathscr{L}_+^\Gamma$ such that the spectral density

$$\Phi = W_\Psi (G^* \Lambda G)^{-1} W_\Psi^* \tag{5.2}$$

solves the generalized moment equation $\Gamma(\Phi) = \Sigma$. A key argument in that paper is that the Jacobian $\nabla\kappa(\Lambda) : \mathrm{Range}\,\Gamma \to \mathrm{Range}\,\Gamma$ is invertible for any $\Lambda \in \mathscr{L}_+^\Gamma$. We will provide a two-dimensional ($m = 2$) numerical counterexample in the next section to this argument showing that the Jacobian of κ can be singular at one point.

5.3 Singular Jacobian of the Moment Map

The Jacobian of the moment map κ, i.e., its Fréchet derivative, is a linear operator from Range Γ to itself:

$$\nabla\kappa(\Lambda):\ \delta\Lambda \mapsto -\int GW_\Psi\Gamma^*(\Lambda)^{-1}\Gamma^*(\delta\Lambda)\Gamma^*(\Lambda)^{-1}W_\Psi^*G^*, \tag{5.3}$$

where $\Gamma^*: X \mapsto G^*XG$ is the adjoint operator of Γ in (3.9) from $\mathfrak{H}_n$ to $C(\mathbb{T};\mathfrak{H}_m)$, and $\Gamma^*(\Lambda)^{-1}$ is understood as $(G^*\Lambda G)^{-1}$.

As mentioned in the Introduction, the claim that $\nabla\kappa(\Lambda)$ vanishes nowhere in $\mathscr{L}_+^\Gamma$ is true in the special case when the prior $\Psi = \psi I$ with ψ a scalar spectral density. Details can be found in Chapter 3 (see also (Georgiou, 2006) itself and (Ferrante et al., 2010)). An important observation is that the Jacobian in that case is a self-adjoint operator, and in fact, it is equal to the negative Hessian of the cost function (4.2). Therefore, the reasoning of nonvanishing Jacobian is built upon the definiteness of the quadratic form $\langle\delta\Lambda, \nabla\kappa(\Lambda)(\delta\Lambda)\rangle$, where the standard inner product over $\mathfrak{H}_n$ is defined as $\langle A, B\rangle := \mathrm{tr}(AB)$. Such reasoning fails in general when Ψ is arbitrarily (but fixed) matrix-valued because the self-adjoint property is lost. One can simply verify that the adjoint operator $\nabla\kappa(\Lambda)^*$: Range $\Gamma \to$ Range Γ of the Jacobian (5.3) is given by

$$\delta\Lambda \mapsto -\int G\Gamma^*(\Lambda)^{-1}W_\Psi^*\Gamma^*(\delta\Lambda)W_\Psi\Gamma^*(\Lambda)^{-1}G^*,$$

which is different from $\nabla\kappa(\Lambda)$.

In the sequel, we want to evaluate numerically the Jacobian determinant. Before that, we will have to build a matrix representation of the linear operator $\nabla\kappa(\Lambda)$.

5.3.1 Matrix Representation of the Jacobian

The Jacobian (5.3) is a linear map from a finite dimensional vector space to itself. It admits a matrix representation if we fix an orthonormal basis of Range Γ, say $\{\mathbf{\Lambda}_k\}_{k=1}^M$, where $M = m(2n-m)$ in the complex case (cf. (Ferrante et al., 2012b, Proposition 3.1) for the dimension). More precisely, the (j,k) element of the *real* $M \times M$ Jacobian matrix $\mathbf{J}_\kappa(\Lambda)$ is

$$\langle\mathbf{\Lambda}_j, \nabla\kappa(\Lambda)(\mathbf{\Lambda}_k)\rangle. \tag{5.4}$$

The domain of the map κ, namely the set $\mathscr{L}_+^\Gamma$, is convex, which is in particular path-connected. We have the next simple proposition.

Proposition 5.3.1. *Consider a C^1 map $f: D \subset \mathbb{R}^n \to \mathbb{R}^n$ such that D is path-connected. If its*

Jacobian $\nabla f : D \to \mathbb{R}^{n\times n}$ *is everywhere nonsingular, then its determinant* $\det \nabla f(\cdot)$ *does not change sign over* D.

Proof. Suppose the contrary, i.e., there exist two points $x_1, x_2 \in D$ such that $\det \nabla f(x_1) > 0$ and $\det \nabla f(x_2) < 0$. By the assumption of path-connectedness, there exists a continuous function $p : [0,1] \to D$ such that $p(0) = x_1$ and $p(1) = x_2$. Since f is C^1, the real-valued function $\det \nabla f(p(\cdot))$ is continuous. By the intermediate value theorem it must be zero for some $t \in (0,1)$. ∎

Therefore, if a sign change of the Jacobian determinant is detected, the Jacobian of the map under consideration cannot be everywhere nonsingular. This is the idea behind our numerical example.

5.3.2 A Numerical Example

Here we consider the problem of matrix covariance extension of dimension $m = 2$ with maximal covariance lag $p = 1$, the (probably) simplest nontrivial case. We have $n = m(p+1) = 4$. The matrix pair (A, B) of the filter bank $G(z)$ is given by

$$A = \begin{bmatrix} 0 & I_2 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ I_2 \end{bmatrix}, \quad \text{with} \quad G(z) = \begin{bmatrix} z^{-2} I_2 \\ z^{-1} I_2 \end{bmatrix}.$$

Let us work in the real case, in which an orthogonal but unnormalized basis of the linear space $\operatorname{Range}\Gamma$ can be determined from (4.45) and (4.46) in Subsection 4.3.2. Normalization of the basis matrices is necessary to compute the quantity (5.4) correctly.

The prior is taken as $\Psi = KGG^*K^*$, a matrix Laurent polynomial, for

$$K = \begin{bmatrix} -0.22 & -1.23 & 2.22 & 0 \\ -1.11 & -0.96 & 1.14 & 2.49 \end{bmatrix}.$$

The polynomial zKG is Schur, with determinantal roots $0.5868, -0.3558$, and thus the outer factor $W_\Psi \equiv zKG$ in this example.

The argument Λ lives in the open set $\mathscr{L}_+^\Gamma$. In practice, it is better to start with a factor of the form zCG with $C \in \mathbb{C}^{m\times n}$. Then we can form the function

$$G^*\Lambda G := G^*C^*CG. \tag{5.5}$$

Notice that if we assign the elements of C with Gaussian or uniformly distributed random numbers, it is unlikely that the polynomial $\det zCG$ has roots on the unit circle. From (5.5)

we have the relation that Λ is equal to the projection of C^*C onto the subspace RangeΓ. Details of the spectral factorization (5.5) can be found in Section 3.5.

We have picked two C matrices with corresponding Λ matrices and the determinantal roots of zCG reported below:

$$C^{(0)} = \begin{bmatrix} -1.08 & -0.57 & 2.45 & 0 \\ 0.84 & -0.08 & 1.01 & 0.78 \end{bmatrix}$$

corresponds to the blocks of $\Lambda^{(0)}$

$$\Lambda_0^{(0)} = \begin{bmatrix} 4.4473 & 0.6681 \\ 0.6681 & 0.4698 \end{bmatrix}, \ \Lambda_1^{(0)} = \begin{bmatrix} -1.7976 & -1.4773 \\ 0.6552 & -0.0624 \end{bmatrix}$$

with the roots of $\det zC^{(0)}G$ at $0.1211 \pm 0.5302i$ (modulus 0.5438).

$$C^{(1)} = \begin{bmatrix} 0.63 & 0.67 & 1.45 & 0 \\ 1.68 & -0.61 & 1.04 & 2 \end{bmatrix}$$

corresponds to the blocks of $\Lambda^{(1)}$

$$\Lambda_0^{(1)} = \begin{bmatrix} 3.2017 & 0.7387 \\ 0.7387 & 2.4105 \end{bmatrix}, \ \Lambda_1^{(1)} = \begin{bmatrix} 2.6607 & 0.3371 \\ 3.3600 & -1.2200 \end{bmatrix}$$

with the roots of $\det zC^{(1)}G$ at $0.7791, -0.6683$.

The integral in (5.3) is approximated with the Riemann sum in Matlab:

$$\int F(\theta) \approx \frac{\Delta\theta}{2\pi} \sum_k F(\theta_k),$$

where $\{\theta_k\}$ are equidistant points on the interval $(-\pi, \pi]$ and the "step length" $\Delta\theta = 10^{-4}$. With the normalized basis obtained from (4.46), the Jacobian matrix can be computed explicitly as in (5.4) and its determinant can be evaluated. We have the numerical result $\det \mathbf{J}_\kappa(\Lambda^{(k)}) = 10.6871, -326.6439$ for $k = 0, 1$, respectively.

Computation of the above example has also been implemented in Mathematica in order to evaluate the integrals symbolically given the numerical values of Λ. The result is consistent with the numerical computation in Matlab, i.e., a sign change of the Jacobian determinant has been detected.

Further, the critical point Λ^c can be computed using the bisection method on the real-

valued function $\det \mathbf{J}_\kappa(\Lambda^{(t)})$ where

$$\Lambda^{(t)} = (1-t)\Lambda^{(0)} + t\Lambda^{(1)},\ t \in [0,1] \tag{5.6}$$

is the line segment between $\Lambda^{(k)},\ k = 0, 1$. We have the blocks

$$\Lambda_0^{\mathrm{c}} = \begin{bmatrix} 4.3901 & 0.6713 \\ 0.6713 & 0.5589 \end{bmatrix},\ \Lambda_1^{\mathrm{c}} = \begin{bmatrix} -1.5930 & -1.3940 \\ 0.7793 & -0.1155 \end{bmatrix},$$

with the corresponding $t^{\mathrm{c}} = 0.0459$, $\det \mathbf{J}_\kappa(\Lambda^{\mathrm{c}}) = -5.4964 \times 10^{-14}$, and the two smallest singular values of $\mathbf{J}_\kappa(\Lambda^{\mathrm{c}})$ are 1.1053×10^{-16} and 0.0573. Hence the Jacobian matrix of κ loses exactly rank 1 at Λ^{c}.

5.4 Characterization of the Critical Point

The quest for nowhere vanishing Jacobian is motivated by the use of continuation methods to solve the nonlinear equation $\kappa(\Lambda) = \Sigma$ for the parameter Λ. The idea is briefly reviewed in the next proposition when the map under consideration is a diffeomorphism (cf. (Allgower and Georg, 1990) for more general settings).

Proposition 5.4.1. *Assume for simplicity that D, E are open and convex subsets of $\mathbb{R}^n$. Let $f : D \to E$ be a C^1 diffeomorphism. Then for $y \in E$, the solution $x = f^{-1}(y)$ can be found by solving the initial value problem*

$$\begin{cases} \dot{x}(t) = [\,\nabla f\,(x(t))\,]^{-1}\,(y - y_0) \\ x(0) = x_0 \end{cases} \tag{5.7}$$

and evaluating $x := x(1)$. The initial value $x_0 \in D$ is arbitrary and $y_0 := f(x_0)$.

Proof. By the assumption of convexity, the line segment

$$p(t) = ty + (1-t)y_0,\ t \in [0,1]$$

is inside E. It is easy to verify that the solution curve $x(t) := f^{-1}(p(t))$ satisfy the IVP (5.7). In fact, the differential equation comes from differentiating the two sides of $f(x(t)) = p(t)$ w.r.t. t and inverting the Jacobian $\nabla f\,(x(t))$. Due to the assumption that f is a diffeomorphism, the solution curve $x(t)$ is indeed continuously differentiable and the Jacobian of f is everywhere invertible in D. ∎

The precise terminal point $x(1)$ can be obtained using a predictor-corrector algorithm

(see Section 4.3 and more generally (Allgower and Georg, 1990)). If one is satisfied enough with an approximate solution, then a general-purpose ODE solver can be used to numerically integrate (5.7). Of course, the map f in the above proposition being a diffeomorphism is a sufficient condition for the continuation method to return a *unique* solution. This is indeed the case for our κ map when the prior takes the special form $\Psi = \psi I$ as mentioned previously (cf. Section 3.3). However, in the presence of a singular Jacobian, it can happen that the solution curve to the IVP branches out at a critical point, and several terminal points exist. On the other hand, a numerical ODE solver diverges in that case because the norm of the derivative tends to infinity near the critical point.

Next we shall demonstrate numerically that the critical point computed in Section 5.3 is a bifurcation point. To this end, it is customary to define the augmented map

$$\mathcal{H}(\Lambda, t) := \kappa(\Lambda) - p(t)$$

from $\mathscr{L}_+^\Gamma \times [0,1] \to \operatorname{Range}\Gamma$, where $p(t) := \kappa\left(\Lambda^{(t)}\right)$ is a smooth curve in $\operatorname{Range}_+\Gamma$ with $\Lambda^{(t)}$ in (5.6). Under this convention, the curve $\left(\Lambda^{(t)}, t\right)$ parametrized by t is in the zero set $\mathcal{H}^{-1}(0)$. When a basis of $\operatorname{Range}\Gamma$ is fixed as in the previous section, the map $\mathcal{H}$ can be identified as a function H from a subset of $\mathbb{R}^{M+1}$ to $\mathbb{R}^M$, whose coordinates have the expression

$$H_j : (\Lambda, t) \mapsto \langle \mathbf{\Lambda}_j, \mathcal{H}(\Lambda, t)\rangle, \quad j = 1, \dots, M, \tag{5.8}$$

where $\Lambda = \sum_k x_k \mathbf{\Lambda}_k$ with $x \in \mathbb{R}^M$ the coordinate vector. Explicit calls of the coordinate x will be avoided subsequently in order to ease the notation.

The matrix representation of the augmented Jacobian $\mathbf{J}_\mathcal{H} \in \mathbb{R}^{M\times(M+1)}$ can be described in terms of the following vector with each entry in $\operatorname{Range}\Gamma$:

$$\nabla H(\Lambda, t) = \left[\begin{array}{ccc|c} \nabla\kappa(\Lambda)(\mathbf{\Lambda}_1) & \cdots & \nabla\kappa(\Lambda)(\mathbf{\Lambda}_M) & -\dot{p}(t) \end{array}\right],$$

where $\dot{p}(t) = \nabla\kappa\left(\Lambda^{(t)}\right)\left(\Lambda^{(1)} - \Lambda^{(0)}\right)$. Then the (j,k) element of the augmented Jacobian

$$[\mathbf{J}_\mathcal{H}(\Lambda, t)]_{jk} = \langle \mathbf{\Lambda}_j, [\nabla H(\Lambda, t)]_k\rangle.$$

Notice that the last column of $\mathbf{J}_\mathcal{H}(\Lambda^c, t^c)$ does not increase the rank due to the relation $\Lambda^{(t^c)} = \Lambda^c$. Hence we have

$$\operatorname{rank}\mathbf{J}_\mathcal{H}(\Lambda^c, t^c) = M - 1, \quad \dim\operatorname{Ker}\mathbf{J}_\mathcal{H}(\Lambda^c, t^c) = 2.$$

Let us introduce the Lyapunov-Schmidt reduction in our finite dimensional context:

$$\begin{aligned}
&\mathbb{R}^{M+1} = D_1 \oplus D_2, \quad \mathbb{R}^M = E_1 \oplus E_2, \text{ where} \\
&D_1 := \operatorname{Ker} \mathbf{J}_{\mathcal{H}}(\Lambda^{\mathrm{c}}, t^{\mathrm{c}}), \quad D_2 := D_1^{\perp}, \\
&E_2 := \operatorname{Range} \mathbf{J}_{\mathcal{H}}(\Lambda^{\mathrm{c}}, t^{\mathrm{c}}), \quad E_1 := E_2^{\perp}.
\end{aligned}$$

The above subspaces can be made more precise by performing SVD to the Jacobian matrix of H at $(\Lambda^{\mathrm{c}}, t^{\mathrm{c}})$, namely

$$\begin{aligned}
\mathbf{J}_{\mathcal{H}}(\Lambda^{\mathrm{c}}, t^{\mathrm{c}}) &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top} \\
&= \begin{bmatrix} \mathbf{u}_{1:M-1} & \mathbf{u}_M \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_{M-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_{1:M-1}^{\top} \\ \mathbf{v}_{M:M+1}^{\top} \end{bmatrix} \\
&:= \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_{M-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^{\top} \\ \mathbf{V}_2^{\top} \end{bmatrix},
\end{aligned} \tag{5.9}$$

where $\mathbf{\Sigma}_{M-1}$ is the (square) diagonal matrix containing all the nonzero singular values, $\mathbf{u}, \mathbf{v}$ are columns of the orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$, respectively, and the notation $\mathbf{u}_{j:k}$ denotes the matrix obtained by putting together the columns $\mathbf{u}_j, \mathbf{u}_{j+1}, \ldots, \mathbf{u}_k$. It is then elementary to verify that

$$\begin{aligned}
D_1 &= \operatorname{Range} \mathbf{V}_2, \quad D_2 = \operatorname{Range} \mathbf{V}_1, \\
E_2 &= \operatorname{Range} \mathbf{U}_1, \quad E_1 = \operatorname{Range} \mathbf{U}_2.
\end{aligned}$$

We can then partition H w.r.t. the new bases determined by the singular vectors. Specifically, let us define

$$\tilde{H}(y) = \begin{bmatrix} \tilde{H}_1(y_1, y_2) \\ \tilde{H}_2(y_1, y_2) \end{bmatrix} := \begin{bmatrix} \mathbf{U}_2^{\top} \\ \mathbf{U}_1^{\top} \end{bmatrix} H(\mathbf{V}_2 y_1 + \mathbf{V}_1 y_2),$$

where $y = (y_1, y_2) \in \mathbb{R}^2 \times \mathbb{R}^{M-1}$ are coordinates of the argument vector (Λ, t) in (5.8) under the new basis. The Jacobian of $\tilde{H}$ is computed as

$$\begin{aligned}
\nabla \tilde{H}(y) &= \begin{bmatrix} \nabla_1 \tilde{H}_1(y_1, y_2) & \nabla_2 \tilde{H}_1(y_1, y_2) \\ \nabla_1 \tilde{H}_2(y_1, y_2) & \nabla_2 \tilde{H}_2(y_1, y_2) \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{U}_2^{\top} \\ \mathbf{U}_1^{\top} \end{bmatrix} \nabla H(\mathbf{V}_2 y_1 + \mathbf{V}_1 y_2) \begin{bmatrix} \mathbf{V}_2 & \mathbf{V}_1 \end{bmatrix},
\end{aligned}$$

where $\nabla_k \tilde{H}_j$ denotes the Jacobian matrix of $\tilde{H}_j$ w.r.t. the variable y_k. It is then straightforward

to check that

$$\nabla\tilde{H}(y^{\mathrm{c}}) = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{M-1} \end{bmatrix},$$

where y^{c} is the coordinate of $(\Lambda^{\mathrm{c}}, t^{\mathrm{c}})$ and $\mathbf{\Sigma}_{M-1}$ nonsingular. Since we have $\tilde{H}_2(y_1^{\mathrm{c}}, y_2^{\mathrm{c}}) = 0$, the implicit function theorem can be applied to assert that locally around y^{c}

$$\tilde{H}_2(y_1, y_2) = 0 \iff y_2 = \varphi(y_1)$$

for some smooth function φ. Substituting y_2 with the above local functional dependence on y_1 into the equation $\tilde{H}_1(y_1, y_2) = 0$, we obtain that equivalently,

$$b(y_1) := \tilde{H}_1(y_1, \varphi(y_1)) = 0,$$

which is called the *bifurcation equation* at the critical point y^{c} of $\tilde{H}$. Notice that b is a real-valued function defined on some subset of $\mathbb{R}^2$. According to (Allgower and Georg, 1990, Definition 8.1.11), if the Hessian matrix $\nabla^2 b(y_1^{\mathrm{c}})$ has two eigenvalues of distinct signs, then y^{c} is a *simple* bifurcation point of the equation $\tilde{H}(y) = 0$.

Following the derivation in (Allgower and Georg, 1990, pp. 77-78), we have the equality

$$\nabla^2 b(y_1^{\mathrm{c}}) = \nabla_1^2 \tilde{H}_1(y^{\mathrm{c}}).$$

We now need a computable expression for the Hessian matrix. Its operator form is easily obtained

$$\begin{aligned} \nabla_1^2 \tilde{H}_1(y) : (\delta y_{1,1}, \delta y_{1,2}) \\ \mapsto \mathbf{U}_2^\top \nabla^2 H(\mathbf{V}_2 y_1 + \mathbf{V}_1 y_2)(\mathbf{V}_2 \delta y_{1,1}, \mathbf{V}_2 \delta y_{1,2}) \end{aligned}$$

as a bilinear map from $\mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$, whose matrix representation follows immediately

$$\nabla_1^2 \tilde{H}_1(y) = \mathbf{V}_2^\top \left[\sum_j u_{jM} \nabla^2 H_j(\mathbf{V}_2 y_1 + \mathbf{V}_1 y_2) \right] \mathbf{V}_2, \tag{5.10}$$

where $\mathbf{U}_2 \equiv \mathbf{u}_M$ is the last left singular vector in (5.9), and $\nabla^2 H_j$ is the Hessian of the component function in (5.8).

Therefore, computation is ultimately reduced to evaluating the 3-d array of second-order partials $\nabla^2 H$ under the standard (to be normalized) basis introduced in (4.46). Define the

symmetric matrix with Range Γ-valued entries

$$\nabla^2 H(\Lambda, t) := \left[\begin{array}{ccc|c} \nabla^2\kappa(\Lambda)(\mathbf{\Lambda}_1, \mathbf{\Lambda}_1) & \cdots & \nabla^2\kappa(\Lambda)(\mathbf{\Lambda}_1, \mathbf{\Lambda}_M) & 0 \\ \vdots & \ddots & \vdots & \vdots \\ \nabla^2\kappa(\Lambda)(\mathbf{\Lambda}_M, \mathbf{\Lambda}_1) & \cdots & \nabla^2\kappa(\Lambda)(\mathbf{\Lambda}_M, \mathbf{\Lambda}_M) & 0 \\ \hline 0 & \cdots & 0 & -\ddot{p}(t) \end{array}\right], \tag{5.11}$$

where

$$\nabla^2\kappa(\Lambda)(\delta\Lambda_1, \delta\Lambda_2) = \int F + F^*$$

is the second-order differential of κ with

$$F := GW_\Psi\Gamma^*(\Lambda)^{-1}\Gamma^*(\delta\Lambda_2)\Gamma^*(\Lambda)^{-1}\Gamma^*(\delta\Lambda_1)\Gamma^*(\Lambda)^{-1}W_\Psi^*G^*$$

and

$$\ddot{p}(t) = \nabla^2\kappa(\Lambda^{(t)})(\Lambda^{(1)} - \Lambda^{(0)}, \Lambda^{(1)} - \Lambda^{(0)}).$$

The Hessian matrix of the component function results from taking element-wise inner product with (5.11), i.e.,

$$[\nabla^2 H_j(\Lambda, t)]_{k\ell} = \langle \mathbf{\Lambda}_j, [\nabla^2 H(\Lambda, t)]_{k\ell} \rangle, \ k, \ell = 1, \ldots, M+1.$$

Continuing our numerical example in the previous section, the Hessian matrix $\nabla^2 b(y_1^c)$ is computed according to the formula (5.10) and its two eigenvalues are $-0.3226, 0.0239$. Therefore, we confirm that y^c, or equivalently (Λ^c, t^c), is a bifurcation point. Following the very definition of a bifurcation point (Allgower and Georg, 1990, p. 76), the original map κ in (5.1) is not injective.

Remark 5.4.2. The sole purpose of the computation above is to show that the Hessian matrix $\nabla^2 b(y_1^c)$ is nonsingular, which according to (Allgower and Georg, 1990, p. 78) is generic. In this case, the Hessian cannot have two eigenvalues of the same sign, since otherwise (Λ^c, t^c) would be an isolated zero point of $\mathcal{H}$ which cannot be reached through curve tracing. This is a consequence of a celebrated theorem of Morse (Allgower and Georg, 1990, Lemma 8.1.10).

5.5 Concluding Remarks

Although only nonvanishing Jacobian is emphasized in (Georgiou, 2006), properness[1] is another important property of the moment map, as it is closely related to the question of surjectivity (cf. Section 3.4, also (Zhu and Baggio, 2017)). The argument on properness has been made implicitly when the prior is taken to be $\Psi = I$, as can be seen in the second column of (Georgiou, 2006, p. 1060), the part proving that the solution to the IVP can be "continued" until $t = 1$. However, in the general case of a matrix-valued prior, a proof of the κ map being proper does not seem obvious.

It is also worth pointing out that the solution form (5.2) to the moment problem plays a major role in (Takyar and Georgiou, 2010), where the factor of Ψ is taken as $W_\Psi = I + KG$ for some $K \in \mathbb{C}^{m\times n}$, which is certainly matrix-valued, i.e., not scalar times identity.

At last, we wish to point out that the problem of real interest to us is how to parametrize (possibly) all rational solutions of "minimal degree" to the moment equation $\Gamma(\Phi) = \Sigma$ in the matrix case, since the scalar counterpart has been well solved in (Byrnes et al., 1995, 1998, 2001b) in the case of covariance extension. Out of such motive, we would like to mention again the parametrization of rational spectral densities discussed in Chapter 3, where the "denominator" $G^*\Lambda G$ is factored instead of breaking the prior down into factors as in (5.2). The moment map τ has been given in (4.16), which has been shown to be *surjective* (Theorem 3.4.2). Moreover, the derivative of (4.16) can be written down explicitly as in (3.67), and a singular Jacobian has so far not been detected numerically, which suggests that there is still hope for uniqueness in this alternative parametrization.

Acknowledgment

The author would like to thank Dr. Giacomo Baggio for implementing the numerical example in Mathematica.

[1] Recall that a map between two topological spaces is called proper if the preimage of every compact set in the codomain is compact in the domain.

6

Application of the Multidimensional Moment Theory to Image Deblurring

6.1 Introduction

Image deblurring is a deconvolution problem in two dimensions. It is well known that the problem of deconvolution is ill-posed (Commenges, 1984; Bertero and Boccacci, 1998; Chan and Shen, 2005), and hence regularization is crucial. The deblurring problem is often formulated as a regularized least squares problem, such as Tikhonov regularization, which has a closed form solution. Other regularization methods include those exploiting partial derivatives (Hansen, Nagy, and O'leary, 2006), total-variation deblurring (Chan, Golub, and Mulet, 1999; Vogel, 2002), or penalized maximum likelihood (Hanke, Nagy, and Vogel, 2000).

Blurring a two-dimensional image $\Phi(x)$, $x \in K \subset \mathbb{R}^2$, can be modeled as a convolution integral

$$b(x) = \int_K \kappa(x-y)\Phi(y)dy, \tag{6.1}$$

where κ is a kernel function, called the point spread function (PSF). Deblurring amounts to the deconvolution of (6.1), i.e., to recover the original image Φ from the blurred image b.

If the blurred image is observed at discrete points $x_1, x_2, \ldots, x_n$ like pixels, then (6.1)

becomes a generalized two-dimensional moment problem

$$c_k = \int_K \alpha_k(x)\Phi(x)dx, \quad k = 1, 2, \dots, n, \tag{6.2}$$

where $c_k := b(x_k)$ and $\alpha_k(x) := \kappa(x_k - x)$, $k = 1, 2, \dots, n$. Here $\alpha_1, \alpha_2, \dots, \alpha_n$ are called *basis functions*. Reconstructing Φ from $c_1, c_2, \dots, c_n$ is an inverse problem, which may or may not have a solution. If it does, it will in general have infinitely many. To achieve compression of data, we impose the rationality constraint

$$\Phi(x) = \frac{P(x)}{Q(x)}, \tag{6.3}$$

where P and Q are nonnegative functions formed by linear combinations of the basis functions. This can be seen as a (generalized) two-dimensional spectral estimation problem with a finiteness condition, and hence as a two-dimensional identification problem (Lindquist and Picci, 2015).[1] If (6.2) does not have a solution, which is the usual case, a regularized approximate solution need to be determined.

The one-dimensional moment problem with rationality constraint has been studied intensively during the last decades and a review is given in Chapter 1. More recently, these results were generalized to the multidimensional case (Karlsson et al., 2016) with applications to spectral estimation and image compression (Ringh et al., 2016). Related results can be found in (Georgiou, 2006). It turns out that the early papers (McClellan and Lang, 1982; Lang and McClellan, 1982, 1983) contain results that are equivalent to some major results in (Karlsson et al., 2016; Ringh et al., 2016), but the basic idea of smooth parametrization is missing there.

In this chapter, we apply the method of the moment problem with rationality constraint to image deblurring with the help of regularization. The chapter is organized as follows. In Section 6.2, we briefly introduce the main result of the theory of multidimensional moment problem and in Section 6.3 regularized approximate solutions are determined for the case that the estimated moments contain errors. We consider the optimization problem for image deblurring in the framework of the multidimensional moment problem in Section 6.4, and a Newton solver is developed. Finally, some implementation details of the proposed method are given in Section 6.5 along with two reconstructed images. These results are preliminary, and better methods to tune the solutions will be developed in future work.

[1] In fact, general basis functions, rather than trigonometric ones, are also used in system identification (Wahlberg, 1991).

6.2 The Multidimensional Moment Problem

We start by reviewing some results in (Karlsson et al., 2016). Let $\mathfrak{P}_+$ be the positive cone of vectors $p := (p_1, p_2, \ldots, p_n)$ such that

$$P(x) = \sum_{k=1}^{n} p_k \alpha_k(x) > 0 \quad \text{for all } x \in K, \tag{6.4}$$

and let $\bar{\mathfrak{P}}_+$ be the closure of $\mathfrak{P}_+$ and $\partial\mathfrak{P}_+ := \bar{\mathfrak{P}}_+ \backslash \mathfrak{P}_+$ its boundary. Then, given a set of real numbers $c_1, c_2, \ldots, c_n$, and linearly independent functions $\alpha_1, \alpha_2, \ldots, \alpha_n$ defined on a compact subset $K \subset \mathbb{R}^d$, consider the problem to find solutions Φ to the moment condition (6.2) of the rational form (6.3), where $p, q \in \mathfrak{P}_+$. Here of course q is the vector of coefficients of Q. Next define the open dual cone $\mathfrak{C}_+$ of vectors $c := (c_1, c_2, \ldots, c_n)$, i.e.,

$$\mathfrak{C}_+ = \left\{ c \,:\, \langle c, p \rangle = \sum_{k=1}^{n} c_k p_k > 0, \quad \forall p \in \bar{\mathfrak{P}}_+ \backslash \{0\} \right\}. \tag{6.5}$$

If the cone $\mathfrak{P}_+$ is nonempty and has the property

$$\int_K \frac{1}{Q} dx = \infty \text{ for all } q \in \partial\mathfrak{P}_+, \tag{6.6}$$

it follows from (Karlsson et al., 2016, Corollary 3.5) that the moment equations

$$c_k = \int_K \alpha_k \frac{P}{Q} dx, \quad k = 1, 2, \ldots, n. \tag{6.7}$$

have a unique solution $q \in \mathfrak{P}_+$ for each $(c, p) \in \mathfrak{C}_+ \times \mathfrak{P}_+$. Moreover, the solution can be obtained by minimizing the strictly convex functional

$$\mathbb{J}_p^c(q) = \langle c, q \rangle - \int_K P \log Q dx, \tag{6.8}$$

over all $q \in \mathfrak{P}_+$. This is the dual of the optimization problem to maximize an entropy-like functional

$$\mathbb{I}_p(\Phi) = \int_K P(x) \log \Phi(x) dx \tag{6.9}$$

over all $\Phi \in \mathfrak{F}_+$ satisfying

$$\int_K \alpha_k(x) \Phi(x) dx = c_k, \quad k = 1, 2, \ldots, n, \tag{6.10}$$

where $\mathfrak{F}_+$ is the class of positive functions in $L_1(K)$.

We note that maximizing (6.9) is equivalent to minimizing the Kullback-Leibler pseudo-distance given P

$$\mathbb{D}(P\|\Phi) = \int_K P(x) \log \frac{P(x)}{\Phi(x)} dx. \tag{6.11}$$

In fact,

$$\mathbb{D}(P\|\Phi) = \int_K P(x) \log P(x) dx - \mathbb{I}_p(\Phi). \tag{6.12}$$

From (Karlsson et al., 2016, Theorem 3.4) we have that the map sending $q \in \mathfrak{P}_+$ to $c \in \mathfrak{C}_+$ is a diffeomorphism, so the problem as stated above is well-posed.

6.3 Regularized Approximation

In practice, the moments are often estimated from a finite number of data, for example, the ergodic estimates for covariance lags, and they may not belong to the dual cone $\mathfrak{C}_+$ so that no solution exists. The problem may be ill-posed also for other reasons. When the data sequence is short, the estimates may contain large errors. Therefore, it is reasonable to match the estimated moments only approximately by allowing an error $d := (d_1, d_2, \dots, d_n)$ in the moment equations so that

$$c_k - \int_K \alpha_k \Phi dx = d_k, \quad k = 1, 2, \dots, n. \tag{6.13}$$

Then the problem is modified to minimize

$$\frac{1}{2}\|d\|^2 + \lambda \mathbb{D}(P\|\Phi), \tag{6.14}$$

subject to (6.13) for some suitable $\lambda > 0$. Here $\lambda\mathbb{D}(P\|\Phi)$ is a regularization term which makes the solution smooth. In view of (6.12), this problem can be reformulated as the problem to maximize

$$\mathbb{I}(\Phi, d) = \int_K P(x) \log \Phi(x) dx - \frac{1}{2\lambda}\|d\|^2 \tag{6.15}$$

subject to (6.13) over all Φ and d. Regularization problems of this type have been considered in (Enqvist and Avventi, 2007) and (Avventi, 2011b, Paper B). Also see (Ringh et al., 2018), where similar results are given.

We assume that the condition (6.6) holds. Modifying the idea of (Enqvist and Avventi, 2007) and (Avventi, 2011b, Paper B) to the setting of (Karlsson et al., 2016), we form the

Lagrangian

$$\begin{aligned} L(\Phi, d, q) &= \mathbb{I}(\Phi, d) + \sum_{k=1}^{n} q_k \left(c_k - \int_K \alpha_k \Phi dx - d_k \right) \\ &= \int_K P \log \Phi dx - \int_K Q\Phi dx - \frac{1}{2\lambda} d^\top d + \langle c - d, q \rangle \end{aligned} \tag{6.16}$$

with the directional derivative

$$\delta L(\Phi, d, q; \delta\Phi, \delta d) = \int_K \left(\frac{P}{\Phi} - Q \right) \delta\Phi dx - (\lambda^{-1} d + q)^\top \delta d. \tag{6.17}$$

For stationarity we require that

$$\Phi = \frac{P}{Q} \quad \text{and} \quad d = -\lambda q, \tag{6.18}$$

which inserted into $L(\Phi, d, q)$ yields the dual functional

$$\varphi(q) = \mathbb{J}_p(q) + \int_K P(\log P - 1) dx, \tag{6.19}$$

where the last term is constant and

$$\mathbb{J}_p(q) = \frac{\lambda}{2} \langle q, q \rangle + \langle c, q \rangle - \int_K P \log Q dx. \tag{6.20}$$

Setting the gradient of $\mathbb{J}_p$ equal to zero, we obtain the moment equations with errors

$$\int_K \alpha_k \frac{P}{Q} dx = c_k + \lambda q_k, \quad k = 1, 2, \ldots, n. \tag{6.21}$$

The regularization parameter λ controls how much error/noise is allowed in the solution. By choosing λ small, the error in the moment equation becomes small. In practice, however, it may be difficult for the algorithm to converge if λ is chosen too small.

We need to show that (6.21) actually has a solution, which will follow if (6.20) has an interior minimum. It is easy to see that (6.20) is strictly convex.

Lemma 6.3.1. *The functional* (6.20) *has compact sublevel sets* $\mathbb{J}_p^{-1}(-\infty, r]$, $r \in \mathbb{R}$.

Proof. The sublevel set $\mathbb{J}_p^{-1}(-\infty, r]$ is closed, so it remains to prove that it is bounded, i.e.,

$\alpha = \|Q\|_\infty$ is bounded. Set $Q = \alpha\tilde{Q}$, where $\tilde{Q}(x) \le 1$. Then we have

$$\mathbb{J}_p(q) = \frac{\lambda}{2}\langle\tilde{q},\tilde{q}\rangle\alpha^2 + \langle c,\tilde{q}\rangle\alpha - \int_K P dx \log\alpha$$
$$- \int_K P\log\tilde{Q}dx \ge a_0\alpha^2 + a_1\alpha - a_2\log\alpha,$$

where $a_0 := \lambda\langle\tilde{q},\tilde{q}\rangle/2 > 0$, $a_1 := \langle c,\tilde{q}\rangle$ and $a_2 := \int_K P dx > 0$. Hence, if $q \in \mathbb{J}_p^{-1}(-\infty, r]$,

$$a_0\alpha^2 + a_1\alpha - a_2\log\alpha \le r.$$

Comparing quadratic and logarithmic growth we see that α is bounded from above. Since $\log\alpha \to -\infty$ as $\alpha \to 0$, it is also bounded away from zero. ∎

Consequently, by strict convexity, (6.20) has a unique minimum. We have to rule out that this minimum is on the boundary of $\mathfrak{P}_+$. In other words, we need to establish that the minimal point is an interior point so that it satisfies the stationary condition (6.21).

Lemma 6.3.2. *The minimum point of $\mathbb{J}_p$ does not lie on the boundary.*

Proof. We proceed along the lines of (Byrnes et al., 1998, p. 662). Let $q \in \mathfrak{P}_+$ be arbitrary, and let q_0 be on the boundary. Set $\delta q = q - q_0$ and define $q_\mu = q_0 + \mu\delta q$. Since $q_\mu = \mu q + (1-\mu)q_0$ and $\mathfrak{P}_+$ is convex, it belongs to $\mathfrak{P}_+$ for all $\mu \in (0,1]$. Next, calculate the directional derivative

$$\delta\mathbb{J}_p(q_\mu,\delta q) = \lambda\langle q_\mu,\delta q\rangle + \langle c,\delta q\rangle - \int_K \frac{P}{Q_\mu}\delta Q dx$$
$$= \langle c + \lambda q_\mu,\delta q\rangle - \int_K R_\mu dx, \text{ where } R_\mu := \frac{P}{Q_\mu}\delta Q.$$

Since

$$\frac{dR_\mu}{d\mu} = -P\frac{(Q-Q_0)^2}{Q_\mu^2} \le 0,$$

R_μ is monotonically decreasing and converges to $R_0 = P(Q-Q_0)/Q_0$ as $\mu \to 0$. However, by condition (6.6), R_0 is not integrable, and hence $\delta\mathbb{J}(q_\mu,\delta q) \to -\infty$ as $\mu \to 0$. ∎

6.4 Application to Image Deblurring

We now return to the convolution equation (6.1) introduced in Section 6.1, where κ is the point spread function (PSF), Φ is original image and b is the blurred image. Then setting

$c_k := b(x_k)$ and $\alpha_k(x) := \kappa(x_k - x)$, we obtain the moment equations (6.2). We want to recover the object Φ from the blurred image b given the PSF κ.

After discretization, the blurring process is described by a linear transform plus some additive noise, i.e.,

$$\mathbf{b} = A\mathbf{x} + \eta. \tag{6.22}$$

Here we have introduced the bold lower-case letters $\mathbf{b}$ and $\mathbf{x}$ to denote the vectorized discretization of the bivariate functions $b(x)$ and $\Phi(x)$, respectively. The blurring matrix A is determined by the PSF and the boundary condition depending on our assumptions of how the picture would be continued outside the image (Ng, Chan, and Tang, 1999; Nagy, Palmer, and Perrone, 2004; Hansen et al., 2006).

As pointed out in (Bertero and Boccacci, 1998), the continuous inverse problem (6.1) is ill-posed. Although the problem may become well-posed after discretization, the blurring matrix A is typically ill-conditioned. Due to the presence of the noise term η, the directly inverted solution is very often not visually meaningful. Therefore, regularization must be introduced as a way to add more information (e.g., smoothness, edge enhancement, etc.) on the desired solution.

Note that each row of the blurring matrix A is the discrete analogue to the basis function α_k in the formulation of the moment problem. As already mentioned, A is nonsingular although rather close to being singular, and hence its rows are linearly independent. Therefore, linear combination of the basis functions becomes matrix-vector multiplication

$$\mathbf{q} := \mathrm{vec}(Q) = A^\top q, \tag{6.23}$$

where the matrix Q here is the discretization of the function $Q(x)$, and "vec" denotes the vectorizing operation for the matrix. Due to the fact that the blurring matrix A is highly structured (Vogel, 2002; Hansen et al., 2006), evaluation of the multiplication can be obtained efficiently with 2-dimensional fast Fourier transform (FFT) or discrete cosine transform (DCT), depending on the boundary condition.

6.4.1 The Optimization Problem

Using the vectorized notation as in (6.22) and (6.23), the discretized objective functional corresponding to (6.8) can be written as

$$\mathbb{J}_p(q) = \mathbf{b}^\top q - \mathbf{p}^\top \log(A^\top q), \tag{6.24}$$

where $\mathbf{p}$ here is the discretized prior function P. The vector-valued log function denotes taking logarithm for each entry of the vector. The reconstructed image

$$\hat{\mathbf{x}} = \mathbf{p}./(A^\top q^*), \tag{6.25}$$

where q^* is the optimal solution that minimizes (6.24) and the operation "./" means element-wise division.

Consider the vector-valued log function first. For a matrix $A \in \mathbb{R}^{n\times n}$ and a vector $x \in \mathbb{R}^n$,

$$(\log A^\top x)_i = \log(a_i^\top x),$$

where a_i is the i-th column of A. The elements of the first order derivative (Jacobian) of $\log A^\top x$ are given by

$$\left[\frac{d\log(A^\top x)}{dx}\right]_{ji} = \frac{\partial \log(a_j^\top x)}{\partial x_i} = \frac{a_{ij}}{a_j^\top x},$$

that is, the j-th row of the Jacobian matrix is $a_j^\top/(a_j^\top x)$, so we have

$$\frac{d\log(A^\top x)}{dx} = D_1(x)A^\top,$$

where $D_1(x) := \operatorname{diag}(1/a_j^\top x)$. Consequently,

$$\begin{aligned}\frac{d}{d\tau}\mathbb{J}_p(q+\tau v)\Big|_{\tau=0} &= \mathbf{b}^\top v - \mathbf{p}^\top D_1(q)A^\top v \\ &= \langle \mathbf{b} - AD_1(q)\mathbf{p}, v\rangle,\end{aligned}$$

and therefore the gradient of $\mathbb{J}_p$ is given by

$$\nabla \mathbb{J}_p(q) = \mathbf{b} - AD_1(q)\mathbf{p}. \tag{6.26}$$

Similarly, for the computation of the Hessian, we form the following

$$\begin{aligned}\frac{\partial^2}{\partial\tau\partial\xi}\mathbb{J}_p(q+\tau v+\xi w)\Big|_{\tau,\xi=0} &= \frac{\partial}{\partial\xi}\left[\mathbf{b}^\top v - \mathbf{p}^\top D_1(y)A^\top v\right]\Big|_{\tau,\xi=0} \\ &= \mathbf{p}^\top \operatorname{diag}\left[\frac{a_j^\top w}{(a_j^\top q)^2}\right]A^\top v,\end{aligned}$$

where $y = q + \tau v + \xi w$. We can rewrite

$$\mathbf{p}^\top \operatorname{diag}\left[\frac{a_j^\top w}{(a_j^\top q)^2}\right] = w^\top A D_2(\mathbf{p}, q)$$

in the last term, where $D_2(\mathbf{p}, q) := \operatorname{diag}(\mathbf{p}_j/(a_j^\top q)^2)$. We then have

$$\frac{\partial^2}{\partial\tau\partial\xi}\mathbb{J}_p(q+\tau v+\xi w)\bigg|_{\tau,\xi=0} = w^\top A D_2(\mathbf{p}, q) A^\top v.$$

Therefore, the formula for Hessian is

$$\nabla^2 \mathbb{J}_p(q) = A D_2(\mathbf{p}, q) A^\top. \tag{6.27}$$

6.4.2 Choice of the Prior P

Recall that the primal problem to maximize (6.9) subject to (6.10) is equivalent to minimizing the Kullback-Leibler divergence (6.11) subject to the same moment equations. Although the Kullback-Leibler divergence is not a metric, it can be used as a pseudo-distance. In $\mathbb{D}(P\|\Phi)$ the function P could be regarded as a prior, and we want the Φ to be "as close as possible" to P in this sense. The choice of P considerably affects the quality of the solution. Choosing $P \equiv 1$ corresponds to no prior information, and the solution is referred to as the *maximum entropy* solution (Lindquist and Picci, 2015). It is also demonstrated in the literature that the maximum entropy solution is often unsatisfactory. In the setting of image deblurring, the blurred image itself should serve as better prior information.

6.5 Numerical Examples

For the image deblurring problem in the presence of noise we solve the regularized optimization problem to minimize

$$\min_{\mathbf{q}>0} \quad \mathbb{J}_p(q) = \mathbf{b}^\top q - \mathbf{p}^\top \log(A^\top q) + \frac{\lambda}{2}\|q\|^2. \tag{6.28}$$

The gradient (6.26) and Hessian (6.27) are modified a bit as

$$\nabla \mathbb{J}_p(q) = \mathbf{b} - A D_1(q)\mathbf{p} + \lambda q, \tag{6.29}$$

$$\nabla^2 \mathbb{J}_p(q) = A D_2(\mathbf{p}, q) A^\top + \lambda I. \tag{6.30}$$

Newton's method (Boyd and Vandenberghe, 2004) is used to solve the optimization problem (6.28).

Two images are chosen for the numerical test. One is the famous Lena with a resolution 256×256 and the other shows a part of the moon surface with a resolution 512×512. The blur type on the test images is out-of-focus and the PSF array is given below with radius $r = 15$:

$$\kappa_{ij} = \begin{cases} 1/(\pi r^2) & \text{if } (i-k)^2 + (j-l)^2 \leq r^2, \\ 0 & \text{elsewhere,} \end{cases} \tag{6.31}$$

where (k,l) is the center of the PSF array. Moreover, a periodic boundary condition is assumed for the Lena image, while a reflexive boundary condition is chosen for the reconstruction of the moon image. The intensity of the noise is characterized by the signal-to-noise ratio (SNR), which is set as 40dB in the test.

The central part of Newton's method is to solve the linear system of equations

$$\nabla^2 \mathbb{J}_p(q)\Delta q = \nabla \mathbb{J}_p(q),$$

for the Newton direction Δq, and we use the conjugate gradient (CG) method (Saad, 1996; Greenbaum, 1997) to solve it iteratively. In each CG iteration, multiplication with the Hessian is evaluated with 4 two-dimensional FFTs/inverse FFTs (or DCTs), which makes this linear solver the major computational cost of the algorithm. To enforce the positivity constraint on $\mathbf{q} = \mathrm{vec}(Q)$ we restrict the step length τ of the line search in the Newton direction. In fact, we have in the Newton iteration

$$q_+ = q - \tau \Delta q,$$

and therefore

$$\mathbf{q}_+ = A^\top q_+ = A^\top q - \tau A^\top \Delta q = \mathbf{q} - \tau \Delta \mathbf{q},$$

where $\Delta \mathbf{q} := A^\top \Delta q$. The maximum step length is taken as

$$\tau_{\max} = \min\{\mathbf{q}_i / \Delta \mathbf{q}_i | \Delta \mathbf{q}_i > 0\}.$$

With the constraint $0 < \tau < \tau_{\max}$, various line search methods can be used.

The original image and the corresponding blurred one is depicted in Fig. 6.1 for the Lena image and in Fig. 6.4 for the moon image. The reconstructed images are shown in Figure 6.3 and Figure 6.5, respectively. For comparison we also compute the classical Tikhonov reconstruction in Figure 6.2, where the regularization parameter is chosen with generalized cross-validation (GCV).

In Figure 6.3 we see that choosing the blurred image $\mathbf{b}$ as the prior indeed improves the

Figure 6.1: Lena: original sharp image and the blurred one

Figure 6.2: Reconstructed image with the Tikhonov method

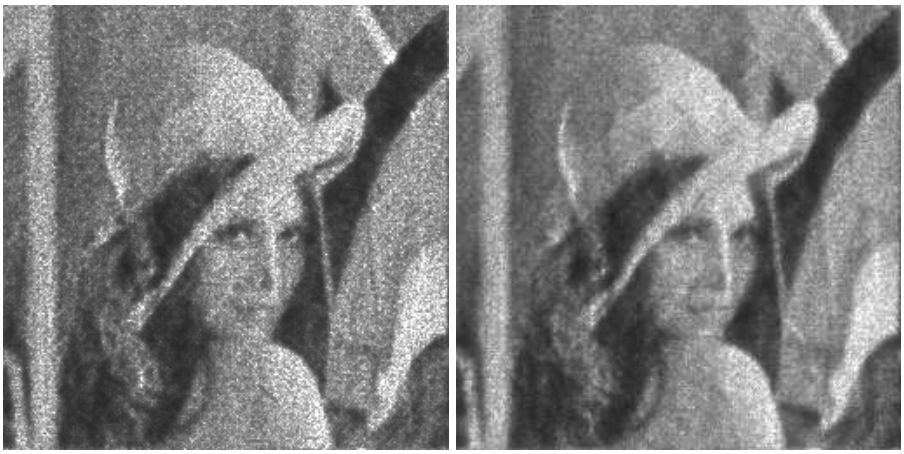

Figure 6.3: Reconstructed images by solutions of the moment problem, $\mathbf{p} = \mathbf{1}$, $\lambda = 12$ (left), and $\mathbf{p} = \mathbf{b}$, $\lambda = 0.11$ (right).

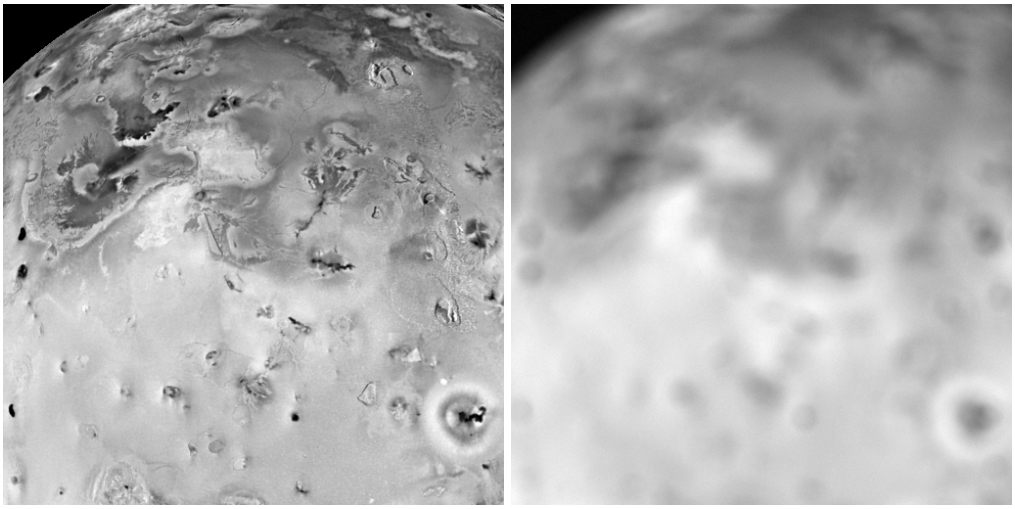

Figure 6.4: Moon: original sharp image and the blurred one

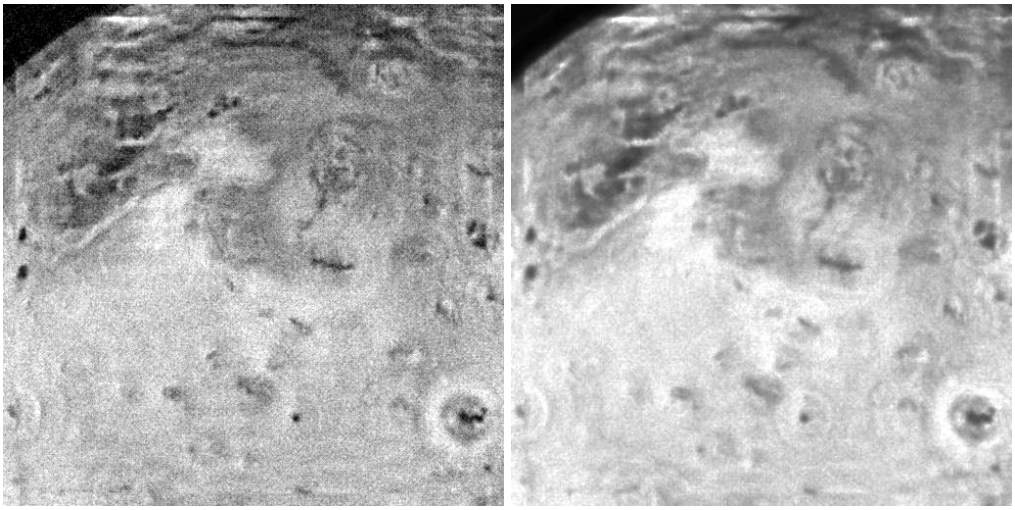

Figure 6.5: Reconstructed images, Tikhonov method (left), and solution of the moment problem with $\mathbf{p} = \mathbf{b}$, $\lambda = 0.4$ (right).

reconstruction. Moreover, the solution of the regularized moment problem looks smoother compared with Tikhonov reconstruction without losing many details. In fact, some reconstruction artifacts are less pronounced. This can also be observed from Fig. 6.5. However, some work remains to perfect this method.

6.6 Concluding Remarks

In this chapter, we have dealt with the image deblurring problem with a spatially invariant blurring operator. We formulate the problem as a multidimensional moment problem and

utilize the generalized entropy as an optimization criterion to select a solution. Due to typical ill-posedness of deconvolution problems, a quadratic regularization term is added to the cost function and the resulting optimization problem is solved with Newton's method. The reconstructed images given in the examples are comparable with those produced by the classical Tikhonov regularization.

There are a few open questions for future research. The first one is whether exchanging $\|d\|^2$ in (6.14) for a more general positive definite form $d^\top W d$ giving different weights to the error components could improve the reconstruction. Moreover, an obvious downside is that the number of basis functions is very high. One could investigate whether including a sparsity promoting regularization term in the cost function could improve numerics. Further, instead of using the blurred image as a prior, one could try to modify the procedure in the style of (Byrnes et al., 2001c, 2002) to use estimated logarithmic moments. How to actually construct such estimates is however an open question.

7 Conclusions and Outlook

In this Ph.D. dissertation, we have mainly tackled problems of ARMA modeling (Chapter 2) and multivariate spectral estimation (Chapters 3, 4, and 5). Both are formulated as moment problems and we try to find a solution in a particular family of spectra parametrized by a finite dimensional variable. Although the methodologies here may appear different from the mainstream approach of optimization[1], they are deeply connected. In fact, the form of candidate solutions in the parametric family introduced in Chapter 3 takes inspiration from the paper (Georgiou and Lindquist, 2003) where optimization was done with the (scalar) Kullback-Leibler divergence. A nontrivial point, or the intrinsic difficulty in the multivariate counterpart of the theory when a matrix prior comes into view is due to the noncommutativity of matrix algebra and apparent absence of an optimization criterion (for our particular formulation). Interesting open problems remain for future investigation.

Main contributions of this dissertation are listed below.

Contributions

- A new algorithm for covariance matching of scalar periodic ARMA models is developed with guaranteed local convergence.

[1]The optimization approach is touched upon in Section 4.2, and Chapter 6 which treats image deblurring using the multidimensional moment theory.

- A procedure to solve the finite-interval smoothing problem for vector-valued stationary processes is described based on model approximation subject to covariance matching.
- An existence result is established for the multivariate spectral estimation problem with solution restricted to a parametric family of spectral densities.
- A possible approach to the uniqueness question in the parametric spectral estimation problem is described together with some preliminary results through the introduction of a diffeomorphic spectral factorization.
- A complete well-posedness result for the spectral estimation problem is given in the special case of a scalar prior function, and numerical solvers to this problem and their convergence properties are studied.
- An counterexample is provided to show that uniqueness of the solution to the multivariate spectral estimation problem in general does not hold in an alternative parametric family of matrix spectral densities that has been proposed in the literature.
- An application of the multidimensional moment theory to image deblurring is made, and a Newton solver for the optimization problem is built up.

Outlook

Several important problems remain open in this field of research, with different emphases on theoretical or numerical aspects. Some have already been mentioned in concluding sections of previous chapters, and they are relisted here as a summary.

- Convergence of the fixed-point-type iterative algorithm for covariance matching of vector ARMA models in Section 2.5.
- Uniqueness of the solution to the parametric spectral estimation problem and its well-posedness in the presence of a matrix prior Ψ (Section 3.4).
- Convergence of the fixed-point iteration introduced in (Ferrante et al., 2010) to solve the parametric spectral estimation problem that is also mentioned in the concluding section of Chapter 3.
- A sparse promoting method for the reconstruction of images in Chapter 6 that may reduce the number of basis functions in use.

One would also like to find some nice applications of the spectral estimation theory. Some pioneers in this direction are exploring noninvasive temperature estimation (Amini, Ebbini, and Georgiou, 2005), and inverse problems in fluid flow control (Zare, Jovanović, and Georgiou, 2017).

A

Appendix for Chapter 2

A.1 Harmonic Analysis in $\mathbb{Z}_{2N}$ and Stationary Periodic Processes

Let $\zeta_1 := e^{i\Delta}$ where $\Delta = \pi/N$, be the primitive $2N$-th root of unity and define the discrete variable ζ taking the $2N$ values $\zeta_k := \zeta_1^k = e^{i\Delta k}$; $k = -N+1, \ldots, 0, \ldots, N$ running counter-clockwise on the unit circle $\mathbb{T}$. The set of $2N$ points $\{\zeta_k\}$ will be called the *discrete unit circle*, denoted by $\mathbb{T}_{2N}$. In particular, we have $\zeta_{-k} = \overline{\zeta_k}$ (complex conjugate). The DFT $\mathcal{F}$ maps a (possibly random) finite support $\mathbb{C}^m$ signal $g = \{\, g(t) : t = -N+1, \ldots, N \,\}$, into a complex vector sequence

$$\hat{g}(\zeta_k) := \sum_{t=-N+1}^{N} g(t)\zeta_k^{-t}, \quad k = -N+1, -N+2, \ldots, N; \tag{A.1}$$

and the signal g can be recovered from its DFT $\hat{g}$ by the formula

$$g(t) = \frac{1}{2N} \sum_{k=-N+1}^{N} \zeta_k^t \hat{g}(\zeta_k), \quad t = -N+1, -N+2, \ldots, N, \tag{A.2}$$

which can also be written as a Stieltjes integral

$$g(t) = \int_{-\pi}^{\pi} e^{it\theta} \hat{g}(e^{i\theta}) d\,\nu(\theta), \tag{A.3}$$

where ν is a step function with steps $\frac{1}{2N}$ at each ζ_k, i.e.,

$$d\,\nu(\theta) = \sum_{k=-N+1}^{N} \delta(e^{i\theta} - \zeta_k) \frac{d\theta}{2N}. \tag{A.4}$$

With $\hat{h}$ being the DFT of $h(t)$, we have

$$\sum_{t=-N+1}^{N} g(t)h(t)^* = \frac{1}{2N} \sum_{k=-N+1}^{N} \hat{g}(\zeta_k)\hat{h}(\zeta_k)^* = \int_{-\pi}^{\pi} \hat{g}(e^{i\theta})\hat{h}(e^{i\theta})^* d\,\nu, \tag{A.5}$$

which is *Parseval's Formula* for DFT.

The DFT (A.1) can also be written in the matrix form

$$\hat{\mathbf{g}} = \mathbf{F}\mathbf{g}, \tag{A.6}$$

where

$$\begin{aligned} \hat{\mathbf{g}} &:= \left[\hat{g}(\zeta_{-N+1})^\top, \hat{g}(\zeta_{-N+2})^\top, \ldots, \hat{g}(\zeta_N)^\top\right]^\top, \\ \mathbf{g} &:= \left[g(-N+1)^\top, g(-N+2)^\top, \ldots, g(N)^\top\right]^\top, \end{aligned} \tag{A.7}$$

and $\mathbf{F}$ is the nonsingular $2mN \times 2mN$ block Vandermonde matrix

$$\mathbf{F} = \begin{bmatrix} \zeta_{-N+1}^{N-1} I_m & \zeta_{-N+1}^{N-2} I_m & \cdots & \zeta_{-N+1}^{-N} I_m \\ \vdots & \vdots & \cdots & \vdots \\ \zeta_0^{N-1} I_m & \zeta_0^{N-2} I_m & \cdots & \zeta_0^{-N} I_m \\ \vdots & \vdots & \cdots & \vdots \\ \zeta_N^{N-1} I_m & \zeta_N^{N-2} I_m & \cdots & \zeta_N^{-N} I_m \end{bmatrix}. \tag{A.8}$$

Likewise, it follows from (A.2) that

$$\mathbf{g} = \frac{1}{2N} \mathbf{F}^* \hat{\mathbf{g}}, \tag{A.9}$$

i.e., $\mathcal{F}^{-1}$ corresponds to $\frac{1}{2N}\mathbf{F}^*$, and consequently, $\mathbf{F}\mathbf{F}^* = 2N\,\mathbf{I}$.

Next consider a zero-mean stationary m-dimensional process $y(t)$ defined on $\mathbb{Z}_{2N}$, i.e., a

stationary process defined on a finite interval $[-N+1, N]$ of the integer line $\mathbb{Z}$ and extended to all of $\mathbb{Z}$ as a periodic stationary process with period $2N$. Let $C_{-N+1}, C_{-N+2}, \dots, C_N$ be the $m \times m$ covariance lags $C_k := \mathbb{E}\{y(t+k)y(t)^*\}$, and define its DFT

$$\Phi(\zeta_j) := \sum_{k=-N+1}^{N} C_k \zeta_j^{-k}, \qquad j = -N+1, \dots, N, \tag{A.10}$$

which is a Hermitian matrix-valued function of ζ. Then, as seen from (A.2) and (A.3),

$$C_k = \frac{1}{2N} \sum_{j=-N+1}^{N} \zeta_j^k \Phi(\zeta_j) = \int_{-\pi}^{\pi} e^{ik\theta} \Phi(e^{i\theta}) d\nu, \quad k = -N+1, \dots, N. \tag{A.11}$$

The $m \times m$ matrix function Φ is the *spectral density* of the vector process y. In fact, let

$$\hat{y}(\zeta_k) := \sum_{t=-N+1}^{N} y(t) \zeta_k^{-t}, \quad k = -N+1, \dots, N, \tag{A.12}$$

be the DFT of the process y. Since $\frac{1}{2N} \sum_{t=-N+1}^{N} (\zeta_k \zeta_\ell^*)^t = \delta_{k\ell}$, the random variables (A.12) are uncorrelated, and

$$\frac{1}{2N} \mathbb{E}\{\hat{y}(\zeta_k) \hat{y}(\zeta_\ell)^*\} = \Phi(\zeta_k) \delta_{k\ell}, \tag{A.13}$$

from which we see that Φ is positive-semidefinite over $\mathbb{T}_{2N}$. The inverse DFT also yields a spectral representation of y analogous to the usual one valid for stationary processes on $\mathbb{Z}$ (Lindquist and Picci, 2015), namely

$$y(t) = \frac{1}{2N} \sum_{k=-N+1}^{N} \zeta_k^t \hat{y}(\zeta_k) = \int_{-\pi}^{\pi} e^{it\theta} d\hat{y}(\theta), \tag{A.14}$$

where $d\hat{y} := \hat{y}(e^{i\theta}) d\nu$ is an orthogonal random measure supported on $\mathbb{T}_{2N}$.

A.2 Block-Circulant Matrices

Block-circulant matrices are block-Toeplitz matrices with a special circulant structure

$$\mathrm{Circ}\{\Lambda_0, \Lambda_1, \Lambda_2, \dots, \Lambda_\nu\} = \begin{bmatrix} \Lambda_0 & \Lambda_\nu & \Lambda_{\nu-1} & \cdots & \Lambda_1 \\ \Lambda_1 & \Lambda_0 & \Lambda_\nu & \cdots & \Lambda_2 \\ \Lambda_2 & \Lambda_1 & \Lambda_0 & \cdots & \Lambda_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Lambda_\nu & \Lambda_{\nu-1} & \Lambda_{\nu-2} & \cdots & \Lambda_0 \end{bmatrix}, \tag{A.15}$$

where the block columns (or equivalently, block rows) are shifted cyclically, and where $\Lambda_0, \Lambda_1, \ldots, \Lambda_\nu$ here are taken to be complex square matrices of the same size. A good survey of Toeplitz and circulant matrices can be found in (Gray, 2006).

In the multivariable circulant rational covariance extension problem, we consider *Hermitian* block-circulant matrices

$$\mathbf{M} := \mathrm{Circ}\{M_0, M_1, M_2, \ldots, M_N, M_{N-1}^*, \ldots, M_2^*, M_1^*\}, \tag{A.16}$$

with each $M_k \in \mathbb{C}^{m\times m}$. The materials below are mostly contained in (Lindquist and Picci, 2016). The matrix $\mathbf{M}$ admits a representation of the form

$$\mathbf{M} = \sum_{k=-N+1}^{N} (S^{-k} \otimes M_k), \quad M_{-k} = M_k^* \tag{A.17}$$

where $\otimes$ is the Kronecker product and S is the nonsingular $2N \times 2N$ cyclic shift matrix

$$S := \begin{bmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{A.18}$$

The $m \times m$ Laurent polynomial

$$M(\zeta) = \sum_{k=-N+1}^{N} M_k \zeta^{-k}, \quad M_{-k} = M_k^* \tag{A.19}$$

is called the *symbol* of $\mathbf{M}$. Let $\mathbf{S} = S \otimes I_m$ be the $2mN \times 2mN$ block cyclic shift matrix. Clearly we have $\mathbf{S}^{2N} = \mathbf{S}^0 = I_{2mN}$, and

$$\mathbf{S}^{k+2N} = \mathbf{S}^k, \qquad \mathbf{S}^{2N-k} = \mathbf{S}^{-k} = (\mathbf{S}^k)^\top. \tag{A.20}$$

Moreover,

$$\mathbf{S}\mathbf{M}\mathbf{S}^* = \mathbf{M} \tag{A.21}$$

is both necessary and sufficient for $\mathbf{M}$ to be block-circulant. With the stacked vector $\mathbf{g}$

introduced in (A.7), we have

$$[\mathbf{Sg}]_k = \mathbf{g}_{k+1}, \quad k \in \mathbb{Z}_{2N}. \tag{A.22}$$

Then, in view of (A.1), $\zeta\mathcal{F}(\mathbf{g})(\zeta) = \mathcal{F}(\mathbf{Sg})(\zeta)$, from which we have

$$\mathcal{F}(\mathbf{Mg})(\zeta) = M(\zeta)\mathcal{F}(\mathbf{g})(\zeta), \tag{A.23}$$

where the matrix function $M(\zeta)$ is the symbol (A.19) of the block-circulant matrix $\mathbf{M}$. An important property of block-circulant block matrices is that they can be block-diagonalized by the DFT. More precisely, it follows from (A.23) that

$$\mathbf{M} = \frac{1}{2N}\mathbf{F}^*\mathrm{diag}\big(M(\zeta_{-N+1}), \ldots, M(\zeta_{-1}), M(\zeta_0), M(\zeta_1), \ldots, M(\zeta_N)\big)\mathbf{F}, \tag{A.24}$$

where "diag" denotes block diagonal. If $\mathbf{M}$ is invertible, the inverse is

$$\mathbf{M}^{-1} = \frac{1}{2N}\mathbf{F}^*\mathrm{diag}\big(M(\zeta_{-N+1})^{-1}, \ldots, M(\zeta_N)^{-1}\big)\mathbf{F}. \tag{A.25}$$

Moreover, since

$$\mathbf{S} = \frac{1}{2N}\mathbf{F}^*\mathrm{diag}\big(\zeta_{-N+1}I_m, \ldots, \zeta_N I_m\big)\mathbf{F} \quad \text{and} \quad \mathbf{S}^* = \frac{1}{2N}\mathbf{F}^*\mathrm{diag}\big(\zeta_{-N+1}^{-1}I_m, \ldots, \zeta_N^{-1} I_m\big)\mathbf{F},$$

we have

$$\mathbf{S}\mathbf{M}^{-1}\mathbf{S}^* = \mathbf{M}^{-1}.$$

Consequently, $\mathbf{M}^{-1}$ is also a block-circulant matrix with symbol $M(\zeta)^{-1}$. In view of the properties (A.17) and (A.20), quotients of symbols are themselves Laurent polynomials of degree at most N and hence symbols. More generally, if $\mathbf{A}$ and $\mathbf{B}$ are block-circulant matrices of the same dimension with symbols $A(\zeta)$ and $B(\zeta)$ respectively, then $\mathbf{AB}$ and $\mathbf{A}+\mathbf{B}$ are block-circulant matrices with symbols $A(\zeta)B(\zeta)$ and $A(\zeta)+B(\zeta)$, respectively. In fact, block-circulant matrices of a fixed dimension form an algebra, and the DFT is an *algebra homomorphism* from the set of block-circulant matrices onto the set of matrix Laurent polynomials of degree at most N in the variable $\zeta \in \mathbb{T}_{2N}$.

B

Appendix for Chapter 3

B.1 Supplementary Propositions and Lemmas

Proposition B.1.1. *A convex set $A \subset \mathbb{R}^n$ is simply connected.*

Proof. By definition (Krantz and Parks, 2013, p. 127), we need to show that: whenever $f : [0,1] \to A$ is a closed curve, i.e., f is continuous with $f(0) = f(1) = x \in A$, there exists a continuous function $F : [0,1] \times [0,1] \to A$ such that

(i) $F(t,0) = f(t)$, for all $t \in [0,1]$,

(ii) $F(0,u) = F(1,u) = x$, for all $u \in [0,1]$, and

(iii) $F(t,1) = x$, for all $t \in [0,1]$.

One can easily verify that $F(t,u) := (1-u)f(t) + ux$ is the desired function. ∎

Lemma B.1.2. *Let a sequence $\{\Lambda_k\}_{k\geq 1} \subset \mathscr{L}_+$ converge to some $\bar{\Lambda} \in \mathscr{L}_+$. Then there exists a real number $\mu > 0$ such that*

$$G^*(e^{i\theta})\Lambda_k G(e^{i\theta}) \geq \mu I, \quad \forall k, \theta.$$

Proof. Note that the sequence of matrix-valued functions $\{G^*\Lambda_k G\}_{k\geq 1}$ is such that

$$G^*(e^{i\theta})\Lambda_k G(e^{i\theta}) > 0, \quad \forall \theta \in [-\pi, \pi] \text{ and } k.$$

Since the eigenvalues of a continuous matrix-valued function

$$F : [a, b] \to \mathbb{C}^{n\times n},\ \theta \mapsto F(\theta)$$

depend continuously on θ (Bhatia, 2013, Corollary VI.1.6), we have that $G^*\Lambda_k G \geq \mu_k I$ where

$$\mu_k := \min_{\theta} \lambda_{\min}\left(G^*(e^{i\theta})\Lambda_k G(e^{i\theta})\right) > 0$$

and $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue. To prove the lemma, it suffices to show that the real number $\mu > 0$ exists such that $\mu_k \geq \mu$ for all k.

For a Hermitian matrix A, we have

$$\lambda_{\min}(A) = \min_{\|x\|_2=1} x^*Ax,$$

which is a special case of the min-max theorem. Hence, we see that

$$\mu_k = \min_{\substack{\theta\in[-\pi,\pi]\\ \|x\|_2=1}} F_k(\theta, x)$$

where $F_k(\theta, x) = x^*G^*(e^{i\theta})\Lambda_k G(e^{i\theta})x$, with $x \in \mathbb{C}^m$. Obviously the real-valued function F_k is continuous in θ and x. We claim that the sequence of functions $\{F_k(\theta, x)\}_{k\geq 1}$ converges uniformly to $\bar{F}(\theta, x) := x^*G^*(e^{i\theta})\bar{\Lambda}G(e^{i\theta})x$. To see this, let us compute

$$\begin{aligned}
|F_k(\theta, x) - \bar{F}(\theta, x)| &= |x^*G^*(e^{i\theta})(\Lambda_k - \bar{\Lambda})G(e^{i\theta})x| \\
&\leq \|G(e^{i\theta})\|_2^2 \|\Lambda_k - \bar{\Lambda}\|_2 \\
&\leq \|G(e^{i\theta})\|_F^2 \|\Lambda_k - \bar{\Lambda}\|_2 \\
&\leq G_{\max}\|\Lambda_k - \bar{\Lambda}\|_2,
\end{aligned}$$

where

$$G_{\max} := \max_{\theta\in[-\pi,\pi]} \operatorname{tr}\left\{G(e^{i\theta})G^*(e^{i\theta})\right\}. \tag{B.1}$$

The above maximum exists since the function on the right is continuous. The second inequality follows from the fact that $\|\cdot\|_2 \leq \|\cdot\|_F$. Therefore, the uniform convergence of F_k to $\bar{F}$ indeed holds.

Next, define similarly

$$\bar{\mu} := \min_{\substack{\theta\in[-\pi,\pi]\\ \|x\|_2=1}} \bar{F}(\theta,x) > 0.$$

By the definition of uniform convergence, there exists a natural number N such that for all $k > N$, we have

$$|F_k(\theta,x) - \bar{F}(\theta,x)| < \frac{\bar{\mu}}{2}, \quad \forall\, \theta, x,$$

which implies

$$F_k(\theta,x) > \bar{F}(\theta,x) - \frac{\bar{\mu}}{2} \geq \frac{\bar{\mu}}{2},$$

and we can just take $\mu = \min\{\mu_1,\ldots,\mu_N,\bar{\mu}/2\} > 0$. ∎

Lemma B.1.3. *The map $\tilde{\omega}$ is continuously differentiable.*

Proof. The map

$$\mathrm{GL}(n,\mathbb{C}) \to \mathrm{GL}(n,\mathbb{C}) : X \mapsto X^{-1} \tag{B.2}$$

is smooth, which follows from Cramer's rule in linear algebra. Hence, the function $F(z;\Lambda) := \psi G(G^*\Lambda G)^{-1}G^*$ inside the integral of (3.17) is also smooth in Λ. Moreover, since G is a rational function, all the partial derivatives of $F(e^{i\theta};\Lambda)$ with respect to Λ are continuous in θ (and Λ). Then by Leibniz's rule for differentiation under the integral sign, partial derivatives of $\tilde{\omega}$ of all orders exist.

Next, we show that the first order partial derivatives are continuous. From (Brookes, 2005), the differential of the map (B.2) at X is given by

$$\mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n} : V \mapsto -X^{-1}VX^{-1}.$$

Using this fact, the differential of $\tilde{\omega}$ at $\Lambda \in \mathscr{L}_+^\Gamma$ is

$$\delta\tilde{\omega}(\Lambda;\delta\Lambda) = -\int \psi G(G^*\Lambda G)^{-1}(G^*\delta\Lambda G)(G^*\Lambda G)^{-1}G^* \tag{B.3}$$

such that $\delta\Lambda \in \mathrm{Range}\,\Gamma$. The integrand in (B.3) with the minus sign is just $\delta F(e^{i\theta};\Lambda;\delta\Lambda)$, the differential of $F(e^{i\theta};\Lambda)$ w.r.t. the parameter Λ. For a fixed $\delta\Lambda$, one can see that the differential $\delta\tilde{\omega}(\Lambda;\delta\Lambda)$ is continuous in Λ. To see this fact, let a sequence $\{\Lambda_k\}_{k\geq 1} \subset \mathscr{L}_+^\Gamma$ converge to some $\bar{\Lambda} \in \mathscr{L}_+^\Gamma$ as $k \to \infty$. By Lemma B.1.2, we know that there exits $\mu > 0$ such that $G^*(e^{i\theta})\Lambda_k G(e^{i\theta}) \geq \mu I$ where for all k, θ. On the other hand, since $\delta\Lambda$ is fixed, it must hold that $G^*\delta\Lambda G \leq MI$, where

$$M := \max_\theta \rho\left(G^*(e^{i\theta})\delta\Lambda G(e^{i\theta})\right).$$

Here $\rho(\cdot)$ denotes the spectral radius of a matrix. Therefore, we have

$$\|\delta F(e^{i\theta};\Lambda_k;\delta\Lambda)\|_2 \le K_\psi M\mu^{-2}\|G(e^{i\theta})\|_2^2, \quad \forall k,\ \theta$$

where $K_\psi = \max_\theta |\psi|$. Moreover,

$$\begin{aligned}\left|[\delta F(e^{i\theta};\Lambda_k;\delta\Lambda)]_{j\ell}\right| &\le \left\|\delta F(e^{i\theta};\Lambda_k;\delta\Lambda)\right\|_F \\ &\le \kappa \left\|\delta F(e^{i\theta};\Lambda_k;\delta\Lambda)\right\|_2 \\ &\le \kappa K_\psi M\mu^{-2}G_{\max}, \quad \forall\, k\ge 1,\ \theta,\ \forall\, j,\ell,\end{aligned}$$

with $G_{\max}$ in (B.1) and κ the constant for norm equivalence. Hence, by Lebesgue's dominated convergence theorem, we have

$$\lim_{k\to\infty} \delta\tilde{\omega}(\Lambda_k;\delta\Lambda) = \int \lim_{k\to\infty} \delta F(e^{i\theta};\Lambda_k;\delta\Lambda) = \delta\tilde{\omega}(\bar{\Lambda};\delta\Lambda).$$

Partial derivatives can then be recovered by the operation $\langle \delta\Lambda_1, \delta\tilde{\omega}(\Lambda;\delta\Lambda_2)\rangle$ by choosing $\delta\Lambda_k,\ k=1,2$ to be orthonormal basis matrices of $\operatorname{Range}\Gamma$. In this way, one can see that every partial derivative of $\tilde{\omega}$ is continuous in Λ. ∎

Lemma B.1.4. *If a sequence $\{\Lambda_k\}_{k\ge1} \subset \mathscr{L}_+^\Gamma$ converges to some $\bar{\Lambda} \in \partial\mathscr{L}_+^\Gamma$, then*

$$\operatorname{tr}\int G(G^*\Lambda_k G)^{-1}G^* \to \infty \text{ as } k\to\infty.$$

Proof. The condition $\bar{\Lambda}\in\partial\mathscr{L}_+^\Gamma$ means that $G^*(e^{i\theta})\bar{\Lambda}G(e^{i\theta})$ is singular for some θ_0. To ease the notation, let us call the integrand $f_k := G(G^*\Lambda_k G)^{-1}G^*$. Notice first that for a fixed k, the quantity $\operatorname{tr} f_k$ is strictly positive since G of full column rank on $\mathbb{T}$. To see this, for any $\theta\in[-\pi,\pi]$, one can pick m rows of $G(e^{i\theta})$ to form a nonsingular submatrix $G_{\text{sub}}(e^{i\theta})$. Then we must have

$$\operatorname{tr} f_k(e^{i\theta}) \ge \operatorname{tr} G_{\text{sub}}(e^{i\theta})\left(G^*(e^{i\theta})\Lambda_k G(e^{i\theta})\right)^{-1} G_{\text{sub}}^*(e^{i\theta}) > 0$$

due to positive definiteness. We can thus restrict ourselves to a subinterval $[a,b]$ containing θ_0 in which the chosen $G_{\text{sub}}(e^{i\theta})$ remains nonsingular. Clearly the largest eigenvalue of $G_{\text{sub}}(G^*\Lambda_k G)^{-1}G_{\text{sub}}^*$ at $e^{i\theta_0}$ tends to infinity as $k\to\infty$. Hence its trace (equal to the sum of eigenvalues) is a rational function that tends to have a pole at $e^{i\theta_0}$. The claim of the lemma follows since the trace of the integral in question diverges in $[a,b]$. ∎

Proposition B.1.5. *The map h^{-1} defined in (3.51) is proper.*

Proof. Let V be a compact set in $\mathscr{L}_+^\Gamma$, we need to show that its preimage U under h^{-1} is again compact, which is equivalent to being closed and bounded in this finite dimensional setting.

Suppose first that there exists a sequence $\{C_j\}_{j\geq 1} \subset U$ such that $\|C_j\| \to \infty$. Apparently, there exists a positive definite matrix $Q \in \operatorname{Range}\Gamma$. Let $\Lambda_j := h^{-1}(C_j)$, and we have

$$\langle Q, \Lambda_j \rangle = \langle Q, C_j C_j^* \rangle = \operatorname{tr}(C_j^* Q C_j) \to \infty,$$

which means that $\{\Lambda_j\}_{j\geq 1} \subset V$ is unbounded, a contradiction. Hence U must be bounded.

Next suppose that there exists a sequence $\{C_j\}_{j\geq 1} \subset U$ tending to some limit $\bar{C} \in \partial\mathscr{C}_+$. There are two cases according to the definition of $\mathscr{C}_+$:

(i) $\bar{C}B$ has at least one diagonal entry equal to zero;

(ii) The closed-loop matrix $\bar{Z} := A - B(\bar{C}B)^{-1}\bar{C}A$ has at least one eigenvalue on the unit circle.

Clearly, we can let $\bar{\Lambda} := h^{-1}(\bar{C})$ since h^{-1} is well defined for any $m \times n$ matrix. By continuity, we have $\Lambda_j := h^{-1}(C_j) \to \bar{\Lambda}$. Since the sequence $\{\Lambda_j\}_{j\geq 1}$ is in the compact set V, it holds that $\bar{\Lambda} \in V$, namely, $G^*\bar{\Lambda}G > 0$ on $\mathbb{T}$.

Let us treat case (ii) first. Since $\bar{Z}$ is the state matrix of $(z\bar{C}G)^{-1}$, the condition means that $(z\bar{C}G)^{-1}$ has a pole on the unit circle. Consequently, $(G^*\bar{\Lambda}G)^{-1} = (z\bar{C}G)^{-1}(z\bar{C}G)^{-*}$ has an eigenvalue tending to infinity at that pole, which in turn means that $G^*\bar{\Lambda}G$ is singular there, a contradiction.

For case (i), notice that for $C \in \mathscr{C}_+$ we have the expression

$$(zCG)^{-1} = (CB)^{-1} - (CB)^{-1}CA(zI - Z)^{-1}B(CB)^{-1},$$

which is obtained by applying the matrix inversion lemma to (3.43). Since the corresponding Z_j is stable, it follows that

$$(C_jB)^{-1} = \int (zC_jG)^{-1}.$$

As illustrated in the proof of Proposition 3.6.2, one can argue that the smallest singular value of $zC_jG(e^{i\theta})$ is bounded from below by a positive constant for any j and θ due to compactness of V. It then follows that $\|(C_jB)^{-1}\|$ is bounded uniformly in j, which is again a contradiction.

Therefore, for any convergent sequence $\{C_j\} \subset U$, its limit must stay in $\mathscr{C}_+$. Then closedness of U follows easily from the continuity of h^{-1}. ∎

B.2 Homogeneous Polynomial Equations

Results here are adapted from (Ježek, 1983, 1986) to the case of polynomials with complex coefficients. The adaption is straightforward and requires only notational changes most of the time. Let us review some preliminaries first.

The expression in the indeterminate z

$$p(z) = \sum_{k=m}^{n} p_k z^k, \quad p_k \in \mathbb{C},\ m, n \in \mathbb{Z} \tag{B.4}$$

is called a Laurent polynomial, abbreviated as LP. We adopt the convention that $n \geq m$ otherwise p is the zero LP. The set of LP's, denoted as $\mathbb{C}[z, z^{-1}]$, has a ring structure under the usual addition and multiplication. A given LP (B.4) is a polynomial if and only if $m \geq 0$. Polynomials form a subring denoted with $\mathbb{C}[z]$. It is worth pointing out that a unit[1] of $\mathbb{C}[z, z^{-1}]$ has the form cz^n, where $c \in \mathbb{C}$ is nonzero and $n \in \mathbb{Z}$, which is different from the ring of polynomials. Two elements p and p' in $\mathbb{C}[z, z^{-1}]$ are called "associated" if there exists a relation $p' = up$ for some unit u. Clearly every LP is associated to a polynomial.

Despite certain differences, the ring of LP's shares many properties with that of polynomials. In particular, we have the following.

Theorem B.2.1. *The ring of Laurent polynomials $\mathbb{C}[z, z^{-1}]$ is a principal ideal domain.*

Proof. Given an ideal I of $\mathbb{C}[z, z^{-1}]$, we need to show that it is principal, i.e., it can be generated by a single element. To this end, consider the set $I \cap \mathbb{C}[z]$. It is straightforward to verify that this is an ideal of $\mathbb{C}[z]$. According to (MacLane and Birkhoff, 1999, Theorem 20, p. 115), the polynomial ring $\mathbb{C}[z]$ is a principal ideal domain. Hence $I \cap \mathbb{C}[z] = (d)_{\mathbb{C}[z]}$ for some polynomial d, where the subscript signifies the ring where the ideal lives. Now for any $p \in I$, there exists a unit u such that $p' = up$ is a polynomial and thus $p' \in I \cap \mathbb{C}[z]$. Consequently, we have $p = u^{-1}p' = u^{-1}qd$ for some $q \in \mathbb{C}[z]$, which implies $I \subset (d)_{\mathbb{C}[z,z^{-1}]}$. The opposite inclusion is by definition of an ideal. Therefore, we must have $I = (d)_{\mathbb{C}[z,z^{-1}]}$ and this completes the proof. ∎

One of the consequences of the above theorem is that every element in $\mathbb{C}[z, z^{-1}]$ admits a unique prime factorization (MacLane and Birkhoff, 1999, Theorem 24, p. 118). However, in $\mathbb{C}[z, z^{-1}]$ only factors $(z - z_j)$ with $z_j \neq 0$ are considered primes (cf. (MacLane and Birkhoff, 1999, p. 111) for more details on algebraic terms).

[1] Recall that a unit is an element that admits a multiplicative inverse.

Define the conjugate polynomial of (B.4) as

$$p^*(z) := \sum_{k=m}^{n} \overline{p_k} z^{-k}. \tag{B.5}$$

It is easy to verify that this operation of conjugation has the properties

$$(p+q)^* = p^* + q^*, \quad (pq)^* = p^* q^*, \quad p^{**} = p.$$

The next lemma concerns the greatest common divisor (gcd) between a polynomial and its conjugate.

Lemma B.2.2. *For every polynomial p, $g := \gcd(p, p^*)$ can be selected so that it satisfies one of the two following conditions:*

(i) $g^* = g$;

(ii) $g^* = z^{-1} g$.

Proof. One can proceed along the same lines as in the proof of (Ježek, 1983, Lemma 2). First one can show that

$$h = \gcd(a, b) \implies h^* = \gcd(a^*, b^*).$$

Then let $h = \gcd(p, p^*)$, it follows from above that $h^* = \gcd(p, p^*)$. Hence h and h^* are associated:

$$h^* = uh, \tag{B.6}$$

where u is a unit. Moreover, one can see by substitution that $h^* = uu^* h^*$ and thus u is in fact unitary, i.e., $u^{-1} = u^*$. In the ring of LP's, unitaries have the form $e^{i\theta} z^n$ with $\theta \in (-\pi, \pi]$ and $n \in \mathbb{Z}$. Next, one can construct a new gcd g such that $h = vg$ with the required property, where $v = \rho e^{i\varphi} z^m$ is a unit for some real $\rho > 0$. Substituting into (B.6), we have

$$v^* g^* = uvg \implies g^* = u \frac{v}{v^*} g = e^{i(\theta + 2\varphi)} z^{n+2m} g.$$

By choosing $\varphi = -\frac{1}{2}\theta$, the product uv/v^* can obviously be made equal to either 1 or z^{-1}. ∎

Theories developed in (Ježek, 1983) focus on the ring $\mathbb{R}[z]$ of real polynomials. However, we want to emphasize that realness of polynomial coefficients plays no role in the proofs. As one of our interests here, the next result applies exactly to the case of complex polynomials.

Theorem B.2.3 ((Ježek, 1983)). *Let $a, b \in \mathbb{C}[z]$ and consider the equation in two unknown Laurent polynomials x and y*

$$ax^* + b^*y = 0.$$

Let $g = \gcd(a, b^*)$ such that $a_0 = a/g$, $b_0 = b/g^*$ are polynomials (which is always possible). The general solution in $\mathbb{C}[z, z^{-1}]$ is

$$x = b_0 q, \quad y = -a_0 q^*,$$

where q is an arbitrary LP.

Proof. Cf. the proof of (Ježek, 1983, Theorem 1) that uses Theorem B.2.1. ∎

We are interested in the case where $a = b$ is *unmixing* in the sense defined next.

Definition B.2.4. A polynomial p is called unmixing if it does not have two roots z_1, z_2 such that $z_1\overline{z_2} = 1$, or equivalently, $\gcd(p, p^*) = 1$.

Remark B.2.5. Stable and anti-stable polynomials in the discrete-time sense certainly possess the unmixing property.

Corollary B.2.6. *Let a be an unmixing polynomial such that the constant term $a_0 \neq 0$. Then the polynomial equation in two unknowns*

$$ax^* + a^*y = 0$$

has the general solution in $\mathbb{C}[z]$

$$x = aq, \quad y = -a\overline{q},$$

where q is an arbitrary complex number.

Proof. The form of the solution follows directly from the previous theorem. Since we are looking for polynomial solutions and $a_0 \neq 0$, it is necessary that the LP q in Theorem B.2.3 reduces to a constant. ∎

Next we give results on symmetric polynomial equations.

Theorem B.2.7 ((Ježek, 1983)). *Given $a \in \mathbb{C}[z]$, consider the equation*

$$ax^* + a^*x = 0. \tag{B.7}$$

Choose $g = \gcd(a, a^*)$ to satisfy one of the conditions in Lemma B.2.2. Then $a_0 = a/g$ is a polynomial and the general solution in $\mathbb{C}[z, z^{-1}]$ is:

(i) *for* $g^* = g$, $x = a_0(r - r^*)$;

(ii) *for* $g^* = z^{-1}g$, $x = a_0(zr - r^*)$,

where r *is an arbitrary polynomial.*

Proof. Follow the proof of (Ježek, 1983, Theorem 3(a,c)) with reference to Theorem B.2.1. ■

Corollary B.2.8. *Let* a *be an unmixing polynomial such that* $a_0 \neq 0$. *Then the homogeneous polynomial equation* (B.7) *has the general solution* $x = i\kappa a$ *in* $\mathbb{C}[z]$, *where* κ *is an arbitrary real number.*

Proof. The situation here falls in to the case (i) of Theorem B.2.7 since a is unmixing. The solution form simplifies because of the extra constraint of being a polynomial plus $a_0 \neq 0$. ■

We are now ready to deal with symmetric matrix polynomial equations. For a matrix polynomial

$$P(z) = \sum_{k=0}^{n} P_k z^k, \quad P_k \in \mathbb{C}^{m\times m},\ n \geq 0, \tag{B.8}$$

the definition of its conjugate polynomial extends naturally

$$P^*(z) = \sum_{k=0}^{n} P_k^* z^{-k}. \tag{B.9}$$

Notice that here we do not consider general Laurent polynomials with both positive and negative powers.

Definition B.2.9. A matrix polynomial P is called unmixing if its determinantal polynomial $\det P$ is unmixing in the sense of Definition B.2.4.

The next result interests us in particular.

Theorem B.2.10. *The homogeneous matrix polynomial equation*

$$AX^* + XA^* = 0 \tag{B.10}$$

with A *unmixing and* $\det A_0 \neq 0$, *has the general solution*

$$X = AQ \tag{B.11}$$

where Q *is an arbitrary constant skew-Hermitian matrix.*

Proof. One can proceed as the proof of (Ježek, 1986, Theorem MP1). Denote $\tilde{A} = \operatorname{adj} A$, $a = \det A$. Multiply (B.10) on both sides from left by $\tilde{A}$ and from right by $\tilde{A}^*$, and we get

$$aX^*\tilde{A}^* + \tilde{A}Xa^* = 0.$$

Define a new unknown $\tilde{X} := \tilde{A}X$, and we have the relation

$$a\tilde{X}^* + \tilde{X}a^* = 0.$$

The above equation can be solved elementwisely. For diagonal entries we have

$$a\tilde{x}_{jj}^* + \tilde{x}_{jj}a^* = 0.$$

Since a is unmixing and $a_0 = a(0) = \det A(0) = \det A_0 \neq 0$, we can apply Corollary B.2.8 to conclude that the general solution is $\tilde{x}_{jj} = i\kappa a$ for an arbitrary real κ. For non-diagonal entries, we have

$$a\tilde{x}_{jk}^* + \tilde{x}_{kj}a^* = 0,$$

whose general solution, according to Corollary B.2.6, is

$$\tilde{x}_{jk} = aq, \quad \tilde{x}_{kj} = -a\overline{q},$$

where q is an arbitrary complex number. One can write compactly $\tilde{X} = aQ$ with Q skew-Hermitian. Finally we can recover the unknown $X = \tilde{A}^{-1}\tilde{X} = AQ$, which is desired. ∎

C

Appendix for Chapter 4

C.1 From Additive Decomposition to Spectral Factorization

Let $Z(z) = H(zI - F)^{-1}G + J$ with $F \in \mathbb{C}^{n\times n}$ stable, $G \in \mathbb{C}^{n\times m}$ $H \in \mathbb{C}^{m\times n}$, and $J \in \mathbb{C}^{m\times m}$. Suppose that $\Phi(z) = Z(z) + Z^*(z) > 0$ for all $z \in \mathbb{T}$. Set $R := J + J^* > 0$. Then one can write

$$\Phi(z) = \begin{bmatrix} H(zI-F)^{-1} & I \end{bmatrix} \begin{bmatrix} 0 & G \\ G^* & R \end{bmatrix} \begin{bmatrix} (z^{-1}I - F^*)^{-1}H^* \\ I \end{bmatrix},$$

which adding to the identity that holds for any Hermitian P

$$0 \equiv \begin{bmatrix} H(zI-F)^{-1} & I \end{bmatrix} \begin{bmatrix} FPF^* - P & FPH^* \\ HPF^* & HPH^* \end{bmatrix} \begin{bmatrix} (z^{-1}I - F^*)^{-1}H^* \\ I \end{bmatrix}$$

yields

$$\Phi(z) = \begin{bmatrix} H(zI-F)^{-1} & I \end{bmatrix} \begin{bmatrix} FPF^* - P & G + FPH^* \\ G^* + HPF^* & R + HPH^* \end{bmatrix} \begin{bmatrix} (z^{-1}I - F^*)^{-1}H^* \\ I \end{bmatrix}.$$

Consequently, if P is the unique stabilizing solution of the DARE

$$P = FPF^* - (G + FPH^*)(R + HPH^*)^{-1}(G^* + HPF^*)$$

such that $R+HPH^* > 0$, then one obtains the factorization

$$\begin{bmatrix} FPF^*-P & G+FPH^* \\ G^*+HPF^* & R+HPH^* \end{bmatrix} = \begin{bmatrix} G+FPH^* \\ R+HPH^* \end{bmatrix} (R+HPH^*)^{-1} \begin{bmatrix} G^*+HPF^* & R+HPH^* \end{bmatrix}.$$

Taking L as the Cholesky factor of $R+HPH^*(=LL^*)$, one gets a left outer factor of $\Phi(z)$ in this way

$$\begin{aligned} W(z) &= \begin{bmatrix} H(zI-F)^{-1} & I \end{bmatrix} \begin{bmatrix} G+FPH^* \\ R+HPH^* \end{bmatrix} L^{-*} \\ &= H(zI-F)^{-1}(G+FPH^*)L^{-*}+L. \end{aligned}$$

References

Ahlfors L. V. *Complex Analysis: An Introduction to the Theory of Analytic Functions of One Complex Variable*. McGraw-Hill, second edition, 1966.

Akhiezer N. I. and Kreĭn M. G. *Some Questions in the Theory of Moments*, volume 2 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, Rhode Island, 1962.

Akhiezer N. I. *The Classical Moment Problem and Some Related Questions in Analysis*. Oliver & Boyd, Edinburgh, 1965.

Allgower E. L. and Georg K. *Numerical Continuation Methods: An Introduction*, volume 13 of *Springer Series in Computational Mathematics*. Springer-Verlag, 1990.

Amini A. N., Ebbini E. S., and Georgiou T. T. Noninvasive estimation of tissue temperature via high-resolution spectral analysis techniques. *IEEE Trans. Biomed. Eng.*, 52(2):221–228, 2005.

Avventi E. Fast, globally converging algorithms for spectral moments problems. In *the Ph.D. thesis "Spectral Moment Problems: Generalizations, Implementation and Tuning"*, pages 11–41. KTH Royal Institute of Technology, Stockholm, 2011a.

Avventi E. *Spectral Moment Problems: Generalizations, Implementation and Tuning*. PhD thesis, KTH Royal Institute of Technology, Stockholm, 2011b.

Baggio G. Further results on the convergence of the Pavon-Ferrante algorithm for spectral estimation. *IEEE Trans. Automat. Control*, 2018a.

Baggio G. Private communication, 2018b.

Baggio G. and Ferrante A. On the factorization of rational discrete-time spectral densities. *IEEE Trans. Automat. Control*, 61(4):969–981, 2016.

Bertero M. and Boccacci P. *Introduction to Inverse Problems in Imaging*. CRC press, 1998.

Bhatia R. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag New York, 2013.

Blomqvist A., Fanizza G., and Nagamune R. Computation of bounded degree Nevanlinna-Pick interpolants by solving nonlinear equations. In *Proc. 42nd IEEE Conference on Decision and Control (CDC 2003)*, volume 5, pages 4511–4516. IEEE, 2003a.

Blomqvist A., Lindquist A., and Nagamune R. Matrix-valued Nevanlinna–Pick interpolation with complexity constraint: An optimization approach. *IEEE Trans. Automat. Control*, 48 (12):2172–2190, 2003b.

Boyd S. and Vandenberghe L. *Convex Optimization*. Cambridge University Press, 2004.

Brookes M. *The matrix reference manual*. Imperial College London, 2005.

Burg J. P. Maximum entropy spectral analysis. In *Proc. 37th Annual International Meeting of the Society of Exploration Geophysicists, Oklahoma City, OK*, 1967.

Byrnes C. I., Georgiou T. T., and Lindquist A. A new approach to spectral estimation: A tunable high-resolution spectral estimator. *IEEE Trans. Signal Process.*, 48(11):3189–3205, 2000.

Byrnes C. I., Enqvist P., and Lindquist A. Identifiability and well-posedness of shaping-filter parameterizations: A global analysis approach. *SIAM J. Control Optim.*, 41(1):23–59, 2002.

Byrnes C. I., Fanizza G., and Lindquist A. A homotopy continuation solution of the covariance extension equation. In *New Directions and Applications in Control Theory*, pages 27–42. Springer, 2005.

Byrnes C. I., Georgiou T. T., and Lindquist A. A generalized entropy criterion for Nevanlinna–Pick interpolation with degree constraint. *IEEE Trans. Automat. Control*, 46(6):822–839, 2001a.

Byrnes C. I., Georgiou T. T., Lindquist A., and Megretski A. Generalized interpolation in H^∞ with a complexity constraint. *Trans. Amer. Math. Soc.*, 358(3):965–987, 2006.

Byrnes C. I., Gusev S. V., and Lindquist A. A convex optimization approach to the rational covariance extension problem. *SIAM J. Control Optim.*, 37(1):211–229, 1998.

Byrnes C. I., Gusev S. V., and Lindquist A. From finite covariance windows to modeling filters: A convex optimization approach. *SIAM Rev.*, 43(4):645–675, 2001b.

Byrnes C. I., Landau H. J., and Lindquist A. On the well-posedness of the rational covariance extension problem. In *Current and Future Directions in Applied Mathematics*, pages 83–108. Springer, 1997.

Byrnes C. I. and Lindquist A. Geometry of the Kimura-Georgiou parametrization of modelling filters. *Internat. J. Control*, 50(6):2301–2312, 1989.

Byrnes C. I. and Lindquist A. On the partial stochastic realization problem. *IEEE Trans. Automat. Control*, 42(8):1049–1070, 1997.

Byrnes C. I. and Lindquist A. A convex optimization approach to generalized moment problems. In *Control and Modeling of Complex Systems*, pages 3–21. Springer, 2003.

Byrnes C. I. and Lindquist A. The generalized moment problem with complexity constraint. *Integral Equations Operator Theory*, 56(2):163–180, 2006.

Byrnes C. I. and Lindquist A. Interior point solutions of variational problems and global inverse function theorems. *Internat. J. Robust Nonlinear Control*, 17(5-6):463–481, 2007.

Byrnes C. I. and Lindquist A. Important moments in systems and control. *SIAM J. Control Optim.*, 47(5):2458–2469, 2008.

Byrnes C. I. and Lindquist A. The moment problem for rational measures: convexity in the spirit of Krein. In *Modern Analysis and Applications*, pages 157–169. Springer, 2009.

Byrnes C. I., Lindquist A., Gusev S. V., and Matveev A. S. A complete parameterization of all positive rational extensions of a covariance sequence. *IEEE Trans. Automat. Control*, 40 (11):1841–1857, 1995.

Byrnes C., Enqvist P., and Lindquist A. Cepstral coefficients, covariance lags, and pole-zero models for finite data strings. *IEEE Trans. Signal Process.*, 49(4):677–693, 2001c.

Carli F. P., Ferrante A., Pavon M., and Picci G. A maximum entropy solution of the covariance selection problem for reciprocal processes. In *Three Decades of Progress in Control Sciences*, pages 77–93. Springer, 2010.

Carli F. P., Ferrante A., Pavon M., and Picci G. A maximum entropy solution of the covariance extension problem for reciprocal processes. *IEEE Trans. Automat. Control*, 56(9):1999–2012, 2011.

Carli F. P. and Georgiou T. T. On the covariance completion problem under a circulant structure. *IEEE Trans. Automat. Control*, 56(4):918–922, 2011.

Chan T. F., Golub G. H., and Mulet P. A nonlinear primal-dual method for total variation-based image restoration. *SIAM J. Sci. Comput.*, 20(6):1964–1977, 1999.

Chan T. F. and Shen J. J. *Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods*, volume 94 of *Other Titles in Applied Mathematics*. SIAM, Philadelphia, 2005.

Commenges D. The deconvolution problem: fast algorithms including the preconditioned conjugate-gradient to compute a MAP estimator. *IEEE Trans. Automat. Control*, 29(3): 229–243, 1984.

Davis P. J. *Circulant Matrices*. John Wiley&Sons, New York, 1979.

Dembo A., Mallows C. L., and Shepp L. A. Embedding nonnegative definite Toeplitz matrices in nonnegative definite circulant matrices, with application to covariance estimation. *IEEE Trans. Inform. Theory*, 35(6):1206–1212, 1989.

Demeure C. and Mullis C. T. The Euclid algorithm and the fast computation of cross-covariance and autocovariance sequences. *IEEE Trans. Acoust. Speech Signal Process.*, 37 (4):545–552, 1989.

Enqvist P. A homotopy approach to rational covariance extension with degree constraint. *Int. J. Appl. Math. Comput. Sci.*, 11:1173–1201, 2001.

Enqvist P. A convex optimization approach to ARMA (n, m) model design from covariance and cepstral data. *SIAM J. Control Optim.*, 43(3):1011–1036, 2004.

Enqvist P. On the simultaneous realization problem—Markov-parameter and covariance interpolation. *Signal Process.*, 86(10):3043–3054, 2006.

Enqvist P. and Avventi E. Approximative covariance interpolation with a quadratic penalty. In *Proc. 46th IEEE Conference on Decision and Control (CDC 2007)*, pages 4275–4280. IEEE, 2007.

Enqvist P. and Karlsson J. Minimal Itakura-Saito distance and covariance interpolation. In *Proc. 47th IEEE Conference on Decision and Control (CDC 2008)*, pages 137–142. IEEE, 2008.

Ferrante A., Pavon M., and Ramponi F. Further results on the Byrnes–Georgiou–Lindquist generalized moment problem. In *Modeling, Estimation and Control*, pages 73–83. Springer Berlin Heidelberg, 2007.

Ferrante A., Ramponi F., and Ticozzi F. On the convergence of an efficient algorithm for Kullback–Leibler approximation of spectral densities. *IEEE Trans. Automat. Control*, 56(3): 506–515, 2011.

Ferrante A., Masiero C., and Pavon M. Time and spectral domain relative entropy: A new approach to multivariate spectral estimation. *IEEE Trans. Automat. Control*, 57(10): 2561–2575, 2012a.

Ferrante A., Pavon M., and Ramponi F. Hellinger versus Kullback–Leibler multivariable spectrum approximation. *IEEE Trans. Automat. Control*, 53(4):954–967, 2008.

Ferrante A., Pavon M., and Zorzi M. Application of a global inverse function theorem of Byrnes and Lindquist to a multivariable moment problem with complexity constraint. In

Three Decades of Progress in Control Sciences, pages 153–167. Springer Berlin Heidelberg, 2010.

Ferrante A., Pavon M., and Zorzi M. A maximum entropy enhancement for a family of high-resolution spectral estimators. *IEEE Trans. Automat. Control*, 57(2):318–329, 2012b.

Georgiou T. T. *Partial Realization of Covariance Sequences*. PhD thesis, University of Florida, Gainesville, 1983.

Georgiou T. T. Realization of power spectra from partial covariance sequences. *IEEE Trans. Acoust. Speech Signal Process.*, 35(4):438–449, 1987a.

Georgiou T. T. and Lindquist A. Kullback–Leibler approximation of spectral density functions. *IEEE Trans. Inform. Theory*, 49(11):2910–2917, 2003.

Georgiou T. T. A topological approach to Nevanlinna–Pick interpolation. *SIAM J. Math. Anal.*, 18(5):1248–1260, 1987b.

Georgiou T. T. The interpolation problem with a degree constraint. *IEEE Trans. Automat. Control*, 44(3):631–635, 1999.

Georgiou T. T. Spectral estimation via selective harmonic amplification. *IEEE Trans. Automat. Control*, 46(1):29–42, 2001.

Georgiou T. T. Spectral analysis based on the state covariance: the maximum entropy spectrum and linear fractional parametrization. *IEEE Trans. Automat. Control*, 47(11): 1811–1823, 2002a.

Georgiou T. T. The structure of state covariances and its relation to the power spectrum of the input. *IEEE Trans. Automat. Control*, 47(7):1056–1066, 2002b.

Georgiou T. T. Solution of the general moment problem via a one-parameter imbedding. *IEEE Trans. Automat. Control*, 50(6):811–826, 2005.

Georgiou T. T. Relative entropy and the multivariable multidimensional moment problem. *IEEE Trans. Inform. Theory*, 52(3):1052–1066, 2006.

Georgiou T. T. and Lindquist A. A convex optimization approach to ARMA modeling. *IEEE Trans. Automat. Control*, 53(5):1108–1119, 2008.

Georgiou T. T. and Lindquist A. Likelihood analysis of power spectra and generalized moment problems. *IEEE Trans. Automat. Control*, 62(9):4580–4592, 2017.

Gordon W. B. On the diffeomorphisms of Euclidean space. *Amer. Math. Monthly*, 79(7): 755–759, 1972.

Gray R. M. Toeplitz and circulant matrices: A review. *Found. Trends Commun. Inform. Theory*, 2(3):155–239, 2006.

Greenbaum A. *Iterative Methods for Solving Linear Systems*, volume 17 of *Frontiers in Applied Mathematics*. SIAM, 1997.

Grenander U. and Szegö G. *Toeplitz Forms and Their Applications*. California Monographs in Mathematical Sciences. University of California Press, 1958.

Hanke M., Nagy J. G., and Vogel C. Quasi-Newton approach to nonnegative image restorations. *Linear Algebra Appl.*, 316(1-3):223–236, 2000.

Hansen P. C., Nagy J. G., and O'leary D. P. *Deblurring Images: Matrices, Spectra, and Filtering*, volume 3 of *Fundamentals of Algorithms*. SIAM, Philadelphia, 2006.

Horn R. A. and Johnson C. R. *Matrix Analysis*. Cambridge University Press, second edition, 2013.

Ježek J. Conjugated and symmetric polynomial equations II: Discrete-time systems. *Kybernetika (Prague)*, 19(3):196–211, 1983.

Ježek J. Symmetric matrix polynomial equations. *Kybernetika (Prague)*, 22(1):19–30, 1986.

Kalman R. E. Realization of covariance sequences. In *Toeplitz Centennial*, pages 331–342. Springer, 1982.

Karlsson J., Lindquist A., and Ringh A. The multidimensional moment problem with complexity constraint. *Integral Equations Operator Theory*, 84(3):395–418, 2016.

Khalil H. K. *Nonlinear Systems*. Prentice-Hall, New Jersey, third edition, 2002.

Krantz S. G. and Parks H. R. *The Implicit Function Theorem: History, Theory, and Applications*. Modern Birkhäuser Classics. Birkhäuser Basel, 2013.

Kreĭn M. G. and Nudel'man A. A. *The Markov Moment Problem and Extremal Problems*, volume 50 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, Rhode Island, 1977.

Lang S. and McClellan J. Multidimensional MEM spectral estimation. *IEEE Trans. Acoust. Speech Signal Process.*, 30(6):880–887, 1982.

Lang S. and McClellan J. Spectral estimation for sensor arrays. *IEEE Trans. Acoust. Speech Signal Process.*, 31(2):349–358, 1983.

Lang S. *Fundamentals of Differential Geometry*, volume 191 of *Graduate Texts in Mathematics*. Springer-Verlag New York, Inc., 1999.

Levy B. C., Frezza R., and Krener A. J. Modeling and estimation of discrete-time Gaussian reciprocal processes. *IEEE Trans. Automat. Control*, 35(9):1013–1023, 1990.

Lindquist A., Masiero C., and Picci G. On the multivariate circulant rational covariance extension problem. In *Proc. 52nd IEEE Conference on Decision and Control (CDC 2013)*, pages 7155–7161. IEEE, 2013.

Lindquist A. and Picci G. *Linear Stochastic Systems: A Geometric Approach to Modeling, Estimation and Identification*, volume 1 of *Series in Contemporary Mathematics*. Springer-Verlag Berlin Heidelberg, 2015.

Lindquist A. and Picci G. Modeling of stationary periodic time series by ARMA representations. In *Optimization and Its Applications in Control and Data Sciences*, pages 281–314. Springer, 2016.

Lindquist A. G. and Picci G. The circulant rational covariance extension problem: The complete solution. *IEEE Trans. Automat. Control*, 58(11):2848–2861, 2013.

MacLane S. and Birkhoff G. *Algebra*. AMS Chelsea Publishing, American Mathematical Society, Providence, Rhode Island, 3rd edition, 1999.

Masani P. The prediction theory of multivariate stochastic processes, III. *Acta Math.*, 104 (1-2):141–162, 1960.

McClellan J. and Lang S. Multi-dimensional MEM spectral estimation. In *Proc. Institute of Acoustics "Spectral Analysis and its Use in Underwater Acoustics": Underwater Acoustics Group Conference*, pages 10.1–10.8, Imperial College, London, 1982.

Nagy J. G., Palmer K., and Perrone L. Iterative methods for image deblurring: a Matlab object-oriented approach. *Numer. Algorithms*, 36(1):73–93, 2004.

Ng M. K., Chan R. H., and Tang W. C. A fast algorithm for deblurring models with Neumann boundary conditions. *SIAM J. Sci. Comput.*, 21(3):851–866, 1999.

Ning L., Jiang X., and Georgiou T. On the geometry of covariance matrices. *IEEE Signal Process. Lett.*, 20(8):787–790, 2013.

Ortega J. M. and Rheinboldt W. C. *Iterative Solution of Nonlinear Equations in Several Variables*, volume 30 of *SIAM's Classics in Applied Mathematics*. SIAM, 2000.

Outerelo E. and Ruiz J. M. *Mapping Degree Theory*, volume 108 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island, and Real Sociedad Matemática Española, Madrid, Spain, 2009.

Parlett B. N. *The Symmetric Eigenvalue Problem*, volume 20 of *Classics in Applied Mathematics*. SIAM, Philadelphia, reprinted edition, 1998.

Pavon M. and Ferrante A. On the Georgiou–Lindquist approach to constrained Kullback–Leibler approximation of spectral densities. *IEEE Trans. Automat. Control*, 51(4):639–644, 2006.

Pavon M. and Ferrante A. On the geometry of maximum entropy problems. *SIAM Rev.*, 55 (3):415–439, 2013.

Picci G. and Zhu B. Approximation of vector processes by covariance matching with applications to smoothing. *IEEE Control Syst. Lett.*, 1(1):200–205, 2017.

Porat B. *Digital Processing of Random Signals: Theory and Methods*. Prentice Hall, Inc., 1994.

Ramponi F., Ferrante A., and Pavon M. A globally convergent matricial algorithm for multivariate spectral estimation. *IEEE Trans. Automat. Control*, 54(10):2376–2388, 2009.

Ramponi F., Ferrante A., and Pavon M. On the well-posedness of multivariate spectrum approximation and convergence of high-resolution spectral estimators. *Systems Control Lett.*, 59(3):167–172, 2010.

Rauch H. E., Striebel C., and Tung F. Maximum likelihood estimates of linear dynamic systems. *AIAA J.*, 3(8):1445–1450, 1965.

Ringh A. and Karlsson J. A fast solver for the circulant rational covariance extension problem. In *Proc. 14th European Control Conference (ECC 2015)*, pages 727–733. IEEE, 2015.

Ringh A., Karlsson J., and Lindquist A. Multidimensional rational covariance extension with applications to spectral estimation and image compression. *SIAM J. Control Optim.*, 54(4):1950–1982, 2016.

Ringh A., Karlsson J., and Lindquist A. Multidimensional rational covariance extension with approximate covariance matching. *SIAM J. Control Optim.*, 56(2):913–944, 2018.

Rissanen J. Algorithms for triangular decomposition of block Hankel and Toeplitz matrices with application to factoring positive matrix polynomials. *Math. Comp.*, 27(121):147–154, 1973.

Robinson E. A. A historical perspective of spectrum estimation. *Proceedings of the IEEE*, 70 (9):885–907, 1982.

Rosenblatt M. *Stationary Sequences and Random Fields*. Birkhäuser, Boston, 1985.

Saad Y. *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company, 1996.

Schwartz J. T. *Nonlinear Functional Analysis*. CRC Press, 1969.

Takyar M. S. and Georgiou T. T. Analytic interpolation with a degree constraint for matrix-valued functions. *IEEE Trans. Automat. Control*, 55(5):1075–1088, 2010.

Vogel C. R. *Computational Methods for Inverse Problems*, volume 23 of *Frontiers in Applied Mathematics*. SIAM, Philadelphia, 2002.

Wahlberg B. System identification using Laguerre models. *IEEE Trans. Automat. Control*, 36 (5):551–562, 1991.

Walker G. T. On periodicity in series of related terms. *Proc. R. Soc. Lond. A*, 131(818): 518–532, 1931.

Whittle P. On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika*, 50(1-2):129–134, 1963.

Yule G. U. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Phil. Trans. R. Soc. Lond. A*, 226(636-646):267–298, 1927.

Zare A., Jovanović M. R., and Georgiou T. T. Colour of turbulence. *J. Fluid Mech.*, 812: 636–680, 2017.

Zhu B. On a parametric spectral estimation problem. Presented at the 18th IFAC Symposium on System Identification (SYSID 2018), arXiv e-print: 1712.07970, 2017.

Zhu B. On the well-posedness of a parametric spectral estimation problem and its numerical solution. Conditionally accepted for publication in IEEE Trans. Automat. Control, arXiv e-print: 1802.09330, 2018a.

Zhu B. On Theorem 6 in "Relative Entropy and the Multivariable Multidimensional Moment Problem". Submitted to IEEE Trans. Inform. Theory, arXiv e-print: 1805.12060, 2018b.

Zhu B. and Baggio G. On the existence of a solution to a spectral estimation problem *à la* Byrnes-Georgiou-Lindquist. To appear in IEEE Trans. Automat. Control, arXiv e-print: 1709.09012, 2017.

Zhu B. and Lindquist A. An identification approach to image deblurring. In *Proc. 35th Chinese Control Conference (CCC 2016)*, pages 235–241. IEEE, 2016.

Zhu B. and Picci G. Proof of local convergence of a new algorithm for covariance matching of periodic ARMA models. *IEEE Control Syst. Lett.*, 1(1):206–211, 2017.

Zorzi M. A new family of high-resolution multivariate spectral estimators. *IEEE Trans. Automat. Control*, 59(4):892–904, 2014a.

Zorzi M. Rational approximations of spectral densities based on the Alpha divergence. *Math. Control Signals Systems*, 26(2):259–278, 2014b.

Zorzi M. Multivariate spectral estimation based on the concept of optimal prediction. *IEEE Trans. Automat. Control*, 60(6):1647–1652, 2015.

Zorzi M. and Ferrante A. On the estimation of structured covariance matrices. *Automatica J. IFAC*, 48(9):2145–2151, 2012.