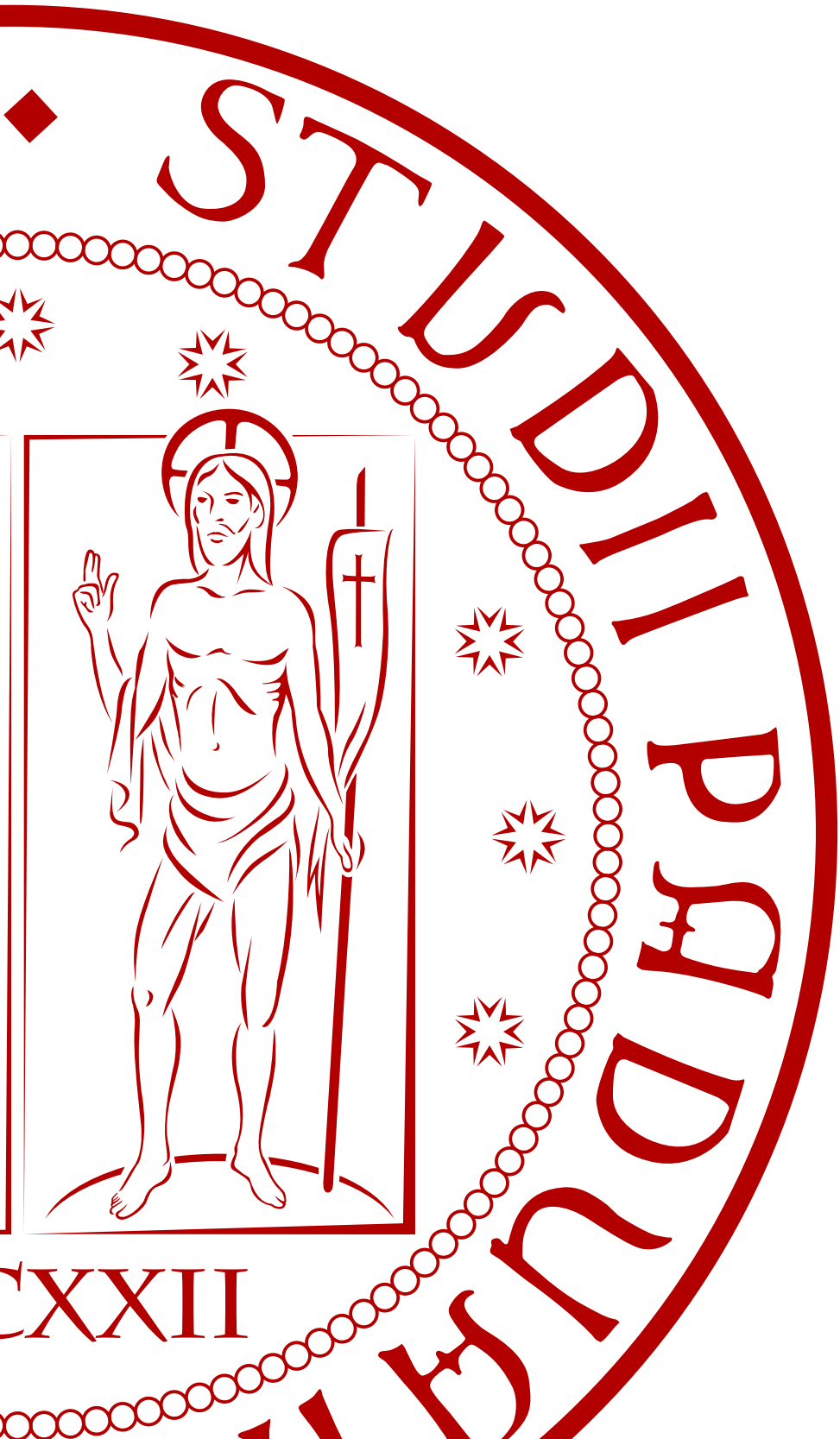
Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Generalized Moment Problems for Estimation of Spectral Densities and Quantum Channels



Ph.D. Candidate
Mattia Zorzi

Advisor
Prof. Augusto Ferrante

Ph.D. School in
Information Engineering
2013

Abstract

This thesis is concerned with two generalized moment problems arising in the estimation of stochastic models.

Firstly, we consider the THREE approach, introduced by Byrnes Georgiou and Lindquist, for estimating spectral densities. Here, the output covariance matrix of a known bank of filters is used to extract information on the input spectral density which needs to be estimated. The parametrization of the family of spectral densities matching the output covariance is a generalized moment problem. An estimate of the input spectral density is then chosen from this family. The choice criterium is based on the minimization of a suitable divergence index among spectral densities. After the introduction of the THREE-like paradigm, we present a multivariate extension of the Beta divergence for solving the problem. Afterward, we deal with the estimation of the output covariance of the filters bank given a finite-length data generated by the unknown input spectral density.

Secondly, we deal with the quantum process tomography. This problem consists in the estimation of a quantum channel which can be thought as the quantum equivalent of the *Markov* transition matrix in the classical setting. Here, a quantum system prepared in a known pure state is fed to the unknown channel. A measurement of an observable is performed on the output state. The set of the employed pure states and observables represents the experimental setting. Again, the parametrization of the family of quantum channels matching the measurements is a generalized moment problem. The choice criterium for the best estimate in this family is based on the maximization of maximum likelihood functionals. The corresponding estimate, however, may not be unique since the experimental setting is not "rich" enough in many cases of interest. We characterize the minimal experimental setting which guarantees the uniqueness of the estimate. Numerical simulation evidences that experimental settings richer than the minimal one do not lead to better performances.

Sommario

In questa tesi vengono presentati e analizzati due problemi dei momenti generalizzati che vengono utilizzati per la stima di modelli stocastici.

Inizieremo col considerare l'approccio THREE, introdotto da Byrnes Georgiou e Lindquist, per la stima di densità spettrali. In questo metodo la covarianza dell'uscita di un banco di filtri noto è utilizzata per estrarre informazione sulla densità spettrale da stimare del segnale all'ingresso del banco. La parametrizzazione della famiglia di densità spettrali compatibili con la covarianza di uscita è un problema dei momenti generalizzato. Una stima di questa densità spettrale è scelta in questa famiglia. Il criterio di tale scelta si basa sulla minimizzazione di un opportuno indice di divergenza tra densità spettrali. Dopo aver introdotto il paradigma di tipo THREE, presenteremo una estensione multivariata della Beta divergenza per risolvere questo problema. Successivamente, affronteremo il problema della stima della matrice di covarianza dell'uscita del banco di filtri avendo a disposizione una sequenza di dati generati dalla densità spettrale all'ingresso del banco.

Infine, tratteremo la tomografia di processi quantistici. Questo problema consiste nello stimare un canale quantistico che può essere pensato come l'equivalente della matrice di transizione di un processo *Markoviano* nel caso classico. Più precisamente, il canale quantistico da identificare è alimentato da un sistema quantistico preparato in uno stato puro noto. Il corrispondente stato all'uscita è successivamente soggetto alla misura di un osservabile. L'insieme di questi stati puri e osservabili caratterizza il *setting* sperimentale. Anche in questo caso, la parametrizzazione della famiglia di canali quantistici compatibili con le misure costituisce un problema dei momenti generalizzato. Il criterio di scelta della stima migliore in questa famiglia si basa sul principio a massima verosimiglianza. Tale stima può tuttavia non essere unica perché l'esperimento in molti casi non è sufficientemente "ricco". Individueremo il *setting* sperimentale minimo che garantisce l'unicità della stima. Le simulazioni numeriche evidenziano che *setting* sperimentali più ricchi di quello minimo non portano a migliori prestazioni.

Contents

List of Symbols

$\mathbb{N}$	Set of natural numbers
$\mathbb{Z}$	Set of integer numbers
$\mathbb{R}$	Vector space of real numbers
$\mathbb{C}$	Vector space of complex numbers
$\mathbb{T}$	Unit circle
$\mathbb{N}_+$	Set of positive natural numbers
$\mathbb{R}_+$	Set of positive real numbers
$\mathbb{R}^{m\times n}$	Vector space of $m \times n$ matrices with entries in $\mathbb{R}$
$\mathbb{C}^{m\times n}$	Vector space of $m \times n$ matrices with entries in $\mathbb{C}$
$\mathcal{Q}_n$	Vector space of symmetric matrices of dimension n
$\mathcal{H}_n$	Vector space of Hermitian matrices of dimension n
$\mathcal{Q}_{n,+}$	Open cone of positive definite matrices in $\mathcal{Q}_n$
$\mathcal{H}_{n,+}$	Open cone of positive definite matrices in $\mathcal{H}_n$
$\overline{\mathcal{Q}}_{n,+}$	Closed cone of positive semi-definite matrices in $\mathcal{Q}_n$
$\overline{\mathcal{H}}_{n,+}$	Closed cone of positive semi-definite matrices in $\mathcal{H}_n$
$\otimes$	Kronecker product
X^*	Transposed and conjugated matrix of X
X^T	Transposed matrix of X
$\mathrm{tr}(X)$	Trace of the matrix X

Chapter 1

Generalized Moment Problems

1.1 Introduction

The term "moment problem" occurred for the first time in 1894 in the work about continued fractions by Stieltjes [60]. In particular, he considered the following setting: Let $\mu(u)$ be a non-decreasing function defined on the interval $[0, \infty)$ which represents a distribution of positive mass on $[0, \infty)$. Accordingly, the integrals

$$\int_0^\infty \mathrm{d}\mu(u), \quad \int_0^\infty u\mathrm{d}\mu(u), \quad \int_0^\infty u^2\mathrm{d}\mu(u)$$

represent the total mass on the line $[0, \infty)$, the static moment of μ, and the moment of inertia with respect to $u = 0$ of μ, respectively. Stieltjes assigned the name of generalized moment of order k to the integral

$$\int_0^\infty u^k\mathrm{d}\mu(u)$$

and he formalized the "moment problem" as follows: Given a certain sequence of numbers c_k with $k = 0, 1, \ldots$, find a non-decreasing function $\mu(u)$ $(u \geq 0)$ such that

$$c_k = \int_0^\infty u^k\mathrm{d}\mu(u), \quad k = 0, 1, \ldots$$

The usually named moment problem was introduced by Hamburger [41] in 1920 and it represents an extension of the above problem: Given a certain sequence of numbers c_k with $k = 0, 1, \ldots$, and $c_0 = 1$, find a non-decreasing function $\mu(u)$ such that

$$c_k = \int_{-\infty}^\infty u^k\mathrm{d}\mu(u), \quad k = 0, 1, \ldots$$

Since then, the moment problem, together with its modifications and generalizations (see for example [59],[2],[48]) has been employed in many issues arising in pure and applied mathematics, physics and engineering.

In the following Section we consider the generalized moment problem introduced by Georgiou in [37] and we provide some preparatory examples to the estimation issues successively considered.

1.2 Generalized Moment Problem

Let G_{left} and G_{right} be a $\mathbb{C}^{n_l \times m}$-valued function and a $\mathbb{C}^{m \times n_r}$-valued function, respectively, defined on $\mathcal{I}$. The multidimensional moment problem introduced by Georgiou, [37], may be stated as follows.

Problem 1.1. *Given $R \in \mathbb{C}^{n_l \times n_r}$, it is required to find a $\mathcal{H}_m$-valued measure $\mathrm{d}\mu$ on the support $\mathcal{I} \subseteq \mathbb{R}$ satisfying the constraint*

$$R = \int_{\mathcal{I}} G_{\text{left}}(\vartheta) \mathrm{d}\mu(\vartheta) G_{\text{right}}(\vartheta). \tag{1.1}$$

Then, it is possible to formulate the "discrete" version of Problem 1.1: Find a nonnegative $\mathcal{H}_m$-valued function μ on the support $\mathcal{I} \subseteq \mathbb{Z}$ such that

$$\sum_{k \in \mathcal{I}} G_{\text{left}}(k) \mu(k) G_{\text{right}}(k). \tag{1.2}$$

We now give some identification paradigms which are instances of the above generalized moment problem.

1.2.1 THREE-like spectral estimation

Let us consider an unknown zero mean, m-dimensional, $\mathbb{R}^m$-valued, purely non-deterministic, wide sense stationary process $y = \{y_k;\ k \in \mathbb{Z}\}$ with spectral density $\Omega(\mathrm{e}^{j\vartheta})$ defined on the unit circle $\mathbb{T}$. Assume that the *a priori* information on Ω is given by a *prior* spectral density $\Psi \in \mathbb{S}_+^m(\mathbb{T})$. Here, $\mathbb{S}_+^m(\mathbb{T})$ denotes the family of bounded and coercive $\mathbb{R}^{m \times m}$-valued spectral density functions on $\mathbb{T}$. Then, a finite-length data $y(1) \dots y(N)$ generated by y is observed. We want to find an estimate $\Phi \in \mathbb{S}_+^m(\mathbb{T})$ of Ω by exploiting Ψ and $y(1) \dots y(N)$. This spectral estimation task is accomplished by employing a THREE-like approach which hinges on the following four elements:

1. A *prior* spectral density $\Psi \in \mathbb{S}_+^m(\mathbb{T})$;

2. A rational filter to process the data

$$G(z) = (zI - A)^{-1}B, \tag{1.3}$$

where $A \in \mathbb{R}^{n\times n}$ is a stability matrix, $B \in \mathbb{R}^{n\times m}$ is full rank with $n > m$, and (A, B) is a reachable pair;

3. An estimate $\hat{\Sigma}$, based on the data $y(1)\dots y(N)$, of the steady state covariance $\Sigma = \Sigma^T > 0$ of the state x_k of the filter

$$x_{k+1} = Ax_k + By_k; \tag{1.4}$$

4. A divergence index $\mathcal{S}$ between two spectral densities.

An estimate $\Phi \in \mathbb{S}_+^m(\mathbb{T})$ of Ω, according to the THREE-like approach, is given by solving the following task.

Problem 1.2. *Given* $\Psi \in \mathbb{S}_+^m(\mathbb{T})$ *and* $\hat{\Sigma} > 0$,

$$\text{minimize } \mathcal{S}(\Phi\|\Psi) \text{ over the set } \left\{\Phi \in \mathbb{S}_+^m(\mathbb{T}) \mid \int G(\mathrm{e}^{j\vartheta})\Phi(\mathrm{e}^{j\vartheta})G(\mathrm{e}^{j\vartheta})^* \frac{\mathrm{d}\vartheta}{2\pi} = \hat{\Sigma}\right\}. \tag{1.5}$$

Remark 1.3. The THREE-like approach considered above may be extended to the complex case: y is $\mathbb{C}^m$-valued, $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times m}$ and Σ is a positive definite Hermitian matrix.

Note that Ψ is generally not consistent with $\hat{\Sigma}$, i.e.

$$\int G(\mathrm{e}^{j\vartheta})\Psi(\mathrm{e}^{j\vartheta})G(\mathrm{e}^{j\vartheta})^* \frac{\mathrm{d}\vartheta}{2\pi} \neq \hat{\Sigma}. \tag{1.6}$$

Hence, we have a spectrum approximation problem which is an instance of the previous generalized moment problem. Chapter 2 deals with Problem 1.2 and we will show how to solve it. Chapter 3 is devoted to a structured covariance estimation problem for finding an estimate of Σ from $y(1)\dots y(N)$. Thus, the considered THREE-like spectral estimation procedure consists in solving a structured covariance estimation problem and then in solving a spectrum approximation problem. The key feature for these estimators concerns the high resolution achievable in prescribed frequency bands, in particular with short data records. Significant applications to these methods can be found in H_∞ robust control [16],[39], biomedical engineering [52], and modeling and identification [13], [40], [45].

Finally, we want to stress that Problem 1.2 includes as special cases a variety of important problems. These are introduced below, where we assume $\hat{\Sigma} = \Sigma$ in order not to compromise the clarity of exposition.

Covariance extension problem: Let $y = \{y_k;\ k \in \mathbb{Z}\}$ be a scalar, $\mathbb{R}$-valued, zero-mean, stationary, purely non-deterministic, stochastic process. By choosing

$$A = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix},\ B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \tag{1.7}$$

we obtain $x_k = \begin{bmatrix} y_{k-n} & \dots & y_{k-1} \end{bmatrix}^T$, accordingly

$$\Sigma = \begin{bmatrix} r_0 & r_1 & \dots & r_{n-1} \\ r_1 & r_0 & \dots & r_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n-1} & r_{n-2} & \dots & r_0 \end{bmatrix} \tag{1.8}$$

where $r_l = E[y_{k+l}y_k]$ is the l-th covariance lag of y. In this case, Problem 1.2 consists in finding an extension $c_{n+1}, c_{n+2} \dots$ from the partial covariance sequence $c_0 \dots c_n$ such that

$$\Phi(\mathrm{e}^{j\vartheta}) = \sum_{k=-\infty}^{+\infty} c_k \mathrm{e}^{-j\vartheta k} \geq 0,\ \ \forall\, \mathrm{e}^{j\vartheta} \in \mathbb{T} \tag{1.9}$$

which is the classical covariance extension problem. We conclude that, the spectrum approximation problem above may be viewed as a generalized covariance extension problem [32], [19], [15] [35], [18], [17].

Nevanlinna-Pick interpolation problem: Let us consider the function

$$f(z) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \Phi(\mathrm{e}^{j\vartheta}) \frac{z + \mathrm{e}^{-\mathrm{j}\theta}}{z - \mathrm{e}^{-\mathrm{j}\theta}} \mathrm{d}\theta, \quad \Phi(\mathrm{e}^{j\vartheta}) \in \mathbb{S}_+^1(\mathbb{T}) \tag{1.10}$$

which is a positive real function, i.e. $f(z)$ is analytic in $|z| > 1$ with positive real part in $|z| > 1$. Moreover, it can be shown that f admits a series representation

$$f(z) = \frac{1}{2} c_0 + c_1 z^{-1} + c_2 z^{-2} + c_3 z^{-3} + \dots \tag{1.11}$$

The Nevanlinna-Pick interpolation problem consists in finding a positive real function interpolating given values $\{w_1, \dots, w_n\},\ w_l \in \mathbb{C}$, at given points

$\{p_1^{-1}, p_2^{-1}, \ldots, p_n^{-1}\}$ lying in the domain of analyticity. Here, $n \geq 3$. We now want to show that this problem can be recovered by Problem 1.2 by choosing

$$A = \begin{bmatrix} p_1 & 0 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & p_n \end{bmatrix}, \; B = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}. \tag{1.12}$$

In this case, $G(z) = \begin{bmatrix} G_1(z) & \ldots & G_n(z) \end{bmatrix}^T$ with $G_l(z) = \frac{1}{z-p_l}$, $l = 1 \ldots n$. Let $x^l = \{x_k^l;\ k \in \mathbb{Z}\}$ be the stationary process obtained as the output of the filter $G_l(z)$ when driven by y. Then

$$\begin{aligned} x_k^l &= p_l x_{k-1}^l + y_{k-1} \\ &= y_{k-1} + p_l y_{k-2} + p_l^2 y_{k-3} + \ldots \end{aligned} \tag{1.13}$$

and so we have

$$\begin{aligned} E[x_k^l \bar{x}_k^q] &= E[(y_{k-1} + p_l y_{k-2} + p_l^2 y_{k-3} + \ldots) \\ &\quad \times (\bar{y}_{k-1} + \bar{p}_q \bar{y}_{k-2} + \bar{p}_q^2 \bar{y}_{k-3} + \ldots)] \\ &= c_0(1 + p_l\bar{p}_q + (p_l\bar{p}_q)^2 + \ldots) \\ &\quad + c_1 p_l(1 + p_l\bar{p}_q + (p_i\bar{p}_q)^2 + \ldots) + \bar{c}_1\bar{p}_q(1 + p_l\bar{p}_q + (p_l\bar{p}_q)^2 + \ldots) \\ &\quad + c_2 p_l^2(1 + p_l\bar{p}_q + (p_l\bar{p}_q)^2 + \ldots) + \bar{c}_2\bar{p}_q^2(1 + p_l\bar{p}_q + (p_l\bar{p}_q)^2 + \ldots) \\ &= (c_0 + c_1 p_l + \bar{c}_1\bar{p}_q + c_2 p_l^2 + \bar{c}_2\bar{p}_q^2 + \ldots)\frac{1}{1 - p_l\bar{p}_q} \\ &= [(\frac{1}{2}c_0 + c_1 p_l + c_2 p_l^2 + \ldots) + (\frac{1}{2}c_0 + \bar{c}_1\bar{p}_q + \bar{c}_2\bar{p}_q^2 + \ldots)]\frac{1}{1 - p_l\bar{p}_q} \\ &= \frac{f(p_l^{-1}) + \overline{f(p_q^{-1})}}{1 - p_l\bar{p}_q}. \end{aligned} \tag{1.14}$$

Therefore the values of the positive real function $f(z)$ at the points $\{p_1^{-1}, p_2^{-1}, \ldots, p_n^{-1}\}$ can be expressed in terms of the covariance of the output $x_k = \begin{bmatrix} x_k^1 & \ldots & x_k^n \end{bmatrix}^T$ of the filters bank as in (1.14). Accordingly, by choosing the filters bank as in (1.12) and setting

$$\Sigma = E[x_k x_k^*] = \begin{bmatrix} \frac{w_1+\bar{w}_1}{1-p_1\bar{p}_1} & \frac{w_1+\bar{w}_2}{1-p_1\bar{p}_2} & \cdots & \frac{w_1+\bar{w}_n}{1-p_1\bar{p}_n} \\ \frac{w_2+\bar{w}_1}{1-p_2\bar{p}_1} & \frac{w_2+\bar{w}_2}{1-p_2\bar{p}_2} & \cdots & \frac{w_2+\bar{w}_n}{1-p_2\bar{p}_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{w_n+\bar{w}_1}{1-p_n\bar{p}_1} & \frac{w_n+\bar{w}_2}{1-p_n\bar{p}_2} & \cdots & \frac{w_n+\bar{w}_n}{1-p_n\bar{p}_n} \end{bmatrix}, \tag{1.15}$$

Problem 1.2 solves the Nevanlinna-Pick interpolation problem.

1.2.2 Quantum process tomography

Consider a d-level quantum system, [53], with associated Hilbert space $\mathbb{H}$ isomorphic to $\mathbb{C}^d$. The *state* of the system is described by a density operator, namely by a positive semidefinite, unit-trace matrix

$$\rho \in \mathfrak{D}(\mathbb{H}) = \{\rho \in \mathcal{H}_d \,|\rho \geq 0,\, \mathrm{tr}(\rho) = 1\},$$

which plays the role of probability distribution in classical probability. Measurable quantities or *observables* are associated with Hermitian matrices $X = \sum_l x_l \Pi_l$, where $\{\Pi_l\}$ is the associated spectral family of orthogonal projection and the spectrum $\{x_l\}$, $x_l \in \mathbb{R}$ represents the possible outcomes. The probability of observing the l-th outcome can be computed as $p_\rho(\Pi_l) = \mathrm{tr}(\Pi_l \rho)$. A quantum channel (in Schrödinger's picture) is a map $\mathcal{E} : \mathfrak{D}(\mathbb{H}) \to \mathfrak{D}(\mathbb{H})$. It is well known [47],[53] that a physically admissible quantum channel must be linear and completely positive (CP) and trace preserving (TP), i.e. ε maps elements of $\mathfrak{D}(\mathbb{H})$ in $\mathfrak{D}(\mathbb{H})$. We will see in Chapter 4 that such a ε is completely described by a positive semi-definite matrix χ satisfying the constraint

$$\sum_{m,n=1}^{d^2} e_m^* \chi e_n F_n^* F_m = I, \tag{1.16}$$

where $\{e_i\}$ is the canonical basis of $\mathbb{R}^{d^2}$ and $\{F_i\}$ is a suitable basis for $\mathcal{H}_d$.

Next, consider the following setting: A quantum system prepared in a known state ρ is fed to an unknown channel ε. The system in the output state $\varepsilon(\rho)$ is then subjected to a projective measurement of an observable Π. We will show that the probability of observing the l-th outcome is

$$p_\rho(\Pi) = \mathrm{tr}(\chi(\Pi \otimes \rho^T)) = \sum_{m,n=1}^{d^2} \alpha_{mn} e_n^* \chi e_m, \tag{1.17}$$

where

$$\Pi \otimes \rho^T = \sum_{m,n=1}^{d^2} \alpha_{mn} e_m e_n^*, \;\; \text{with } \alpha_{mn} \in \mathbb{C}. \tag{1.18}$$

Assume that the experiment is repeated with a series of known input states $\{\rho_k\}_{k=1}^L$, and to each trial the same observables $\{\Pi_j\}_{j=1}^M$ are measured N times obtaining a series of outcomes $\{x_l^{jk}\}$. We consider the sampled frequencies to be our data, namely

$$f_{jk} = \frac{1}{N} \sum_{l=1}^{N} x_l^{jk}. \tag{1.19}$$

The channel identification problem (or as it is referred to in the physics literature, the *quantum process tomography* problem [54],[53],[50]) consists in finding a positive semi-definite matrix $\hat{\chi}$ (which represents a CPTP quantum channel $\hat{\varepsilon}$) satisfying the following constraints

$$\begin{aligned} f_{jk} &= \sum_{m,n=1}^{d^2} \alpha_{jkmn} e_n^* \hat{\chi} e_m \\ I &= \sum_{m,n=1}^{d^2} e_m^* \hat{\chi} e_n F_n^* F_m. \end{aligned} \tag{1.20}$$

Again, the above problem is an instance of Problem 1.1. We will analyze it in detail in Chapter 4.

Chapter 2

Spectrum approximation problem

2.1 Introduction

We consider the spectrum approximation problem introduced in Section 1.2.1 (i.e. Problem 1.2) which constitutes part of the THREE-like procedure for estimating multivariate spectral densities. This problem "chooses" as estimate of the input spectral density Ω the spectral density which minimizes a divergence index $\mathcal{S}$, with respect to an *a priori* spectral density Ψ, over the family of spectral densities $\Phi \in \mathbb{S}_+^m(\mathbb{T})$ matching the estimated output covariance matrix through the constraint

$$\int G\Phi G^* = \hat{\Sigma}. \tag{2.1}$$

Throughout this Chapter, integration takes place on $(0, 2\pi]$ with respect to the normalized Lebesgue measure $\mathrm{d}\vartheta/2\pi$:

$$\int f := \int_0^{2\pi} f(\mathrm{e}^{j\vartheta}) \frac{\mathrm{d}\vartheta}{2\pi}. \tag{2.2}$$

Moreover, we assume that $\hat{\Sigma}$ is given and such that Problem 1.2 is feasible, i.e. there exists $\Phi \in \mathbb{S}_+^m(\mathbb{T})$ satisfying constraint (2.1). In Chapter 3 we will show how $\hat{\Sigma}$ should be computed from the finite-length data $y(1) \dots y(N)$. We only anticipate that a necessary and sufficient condition which guarantees the feasibility is $\hat{\Sigma} \in \text{Range } \Gamma \cap \mathcal{Q}_{n,+}$, where Range Γ is the range of the linear operator defined in (3.5)[1]. Once we have $\hat{\Sigma}$, we can replace G with $\overline{G} = \hat{\Sigma}^{-\frac{1}{2}} G$

[1] In Chapter 3, the general case is considered where $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times m}$ and the process y is complex valued too. Thus, in the case we are dealing with, the codomain of Γ is $\mathcal{Q}_n$.

and (A, B) with $(\overline{A} = \hat{\Sigma}^{-\frac{1}{2}} A \hat{\Sigma}^{\frac{1}{2}}, \overline{B} = \hat{\Sigma}^{-\frac{1}{2}} B)$. Thus, the constraint may be rewritten as $\int \bar{G}\Phi\bar{G}^* = I$. Accordingly, from now on we consider the following equivalent formulation.

Problem 2.1. *Given* $\Psi \in \mathbb{S}_+^m(\mathbb{T})$ *and* $G(z) = (zI - A)^{-1}B$ *such that* $I \in$ Range Γ*,*

$$\text{minimize } \mathcal{S}(\Phi\|\Psi) \text{ over the set} \\ \left\{ \Phi \in \mathbb{S}_+^m(\mathbb{T}) \mid \int G\Phi G^* = I \right\}. \tag{2.3}$$

The most delicate issue in the above problem deals with the choice of the divergence index $\mathcal{S}$. In fact, the corresponding solution to the spectrum approximation problem (that heavily depends on the divergence index) must be computable and possibly with bounded *McMillan* degree. Accordingly, it is important to have many different indexes available in such a way to choose the most appropriate index in relation to the specific application. Note that, a divergence index among spectral densities in $\mathbb{S}_+^m(\mathbb{T})$ must satisfy the following basic property for all $\Phi, \Psi \in \mathbb{S}_+^m(\mathbb{T})$:

$$\begin{aligned} &\mathcal{S}(\Phi\|\Psi) \geq 0 \\ &\mathcal{S}(\Phi\|\Psi) = 0 \text{ if and only if } \Phi = \Psi. \end{aligned} \tag{2.4}$$

The Chapter is structured as follows. We start by introducing the divergence indexes for multivariate spectral densities suggested in the literature. We then introduce the Beta divergence family and we consider the corresponding spectrum approximation problem, [66]. We will show that it is possible to characterize a family of solutions to Problem 2.1 with bounded *McMillan* degree. Moreover, its limit coincides to the solution obtained by using the *Kullback-Leibler* divergence.

2.2 Divergences indexes employed in the literature

The THREE estimator, introduced by Byrnes Georgiou and Lindquist in [14], only works when the process y is scalar. Moreover, its solution corresponds to the maximum entropy scalar spectrum satisfying the constraint in (2.3) which can be expressed in closed form (see [35]) as

$$\hat{\Phi}_{\text{THREE}} = \left[G^* B (B^T B)^{-1} B^T G \right]^{-1}. \tag{2.5}$$

In [38], this setting was generalized by considering Problem 2.1, where the new "ingredient" is the possibility of considering prior information encoded in an *a priori* spectral density Ψ. Here, the *Kullback-Leibler* divergence for coercive spectra with the same zeroth moment was considered:

$$\mathcal{S}_{\mathrm{KL0}}(\Psi\|\Phi) = \int \Psi \log\left(\frac{\Psi}{\Phi}\right), \tag{2.6}$$

and they showed that the unique solution to Problem 2.1 is

$$\Phi_{\mathrm{PRIOR}} = \frac{\Psi}{G^*\Lambda G}, \tag{2.7}$$

where $\Lambda \in \mathcal{Q}_n$. Note that Φ_{PRIOR} is rational when Ψ is a rational spectral density.

In [37], a *Kullback-Leibler* divergence for multivariate spectral densities with the same trace of the zeroth-moment has been introduced

$$\mathcal{S}_{KL0}(\Psi\|\Phi) = \int \mathrm{tr}[\Psi(\log(\Psi) - \log(\Phi))] \tag{2.8}$$

where $\log(\cdot)$, whose definition will be given in Section 2.3, is the matrix logarithm. This divergence is inspired by the *Umegaki-von Neumann*'s relative entropy [53] of statistical quantum mechanics. It turns out that it is possible to compute the solution to Problem 2.1 with (2.8) only with $\Psi = I$. In this case we obtain (2.5), [35]. On the other hand, the solution to Problem 2.1 with

$$\mathcal{S}_{KL0}(\Phi\|\Psi) = \int \mathrm{tr}[\Phi(\log(\Phi) - \log(\Psi))] \tag{2.9}$$

is computable, [37]. Note that, (2.9) may be readily extended to the nonequal-trace case, see [20] for the scalar case,

$$\mathcal{S}_{\mathrm{KL}}(\Phi\|\Psi) = \int \mathrm{tr}[\Phi(\log(\Phi) - \log(\Psi)) - \Phi + \Psi] \tag{2.10}$$

and $\mathcal{S}_{KL0}(\Phi\|\Psi) = \mathcal{S}_{\mathrm{KL}}(\Phi\|\Psi)$ when $\int \mathrm{tr}\Phi = \int \mathrm{tr}\Psi$. Moreover, $\mathcal{S}_{\mathrm{KL}}$ satisfies property (2.4). The corresponding spectrum approximation problem consists in minimizing $\mathcal{S}_{\mathrm{KL}}(\Phi\|\Psi)$ over $\left\{\Phi \in \mathbb{S}_+^m(\mathbb{T}) \mid \int G\Phi G^* = I\right\}$ which is a constrained convex optimization problem. Its *Lagrangian* is

$$\begin{aligned} L_{\mathrm{KL}}(\Phi,\Lambda) &= \mathcal{S}_{\mathrm{KL}}(\Phi\|\Psi) + \left\langle \int G\Phi G^* - I, \Lambda \right\rangle \\ &= \mathrm{tr}[\int \Phi(\log(\Phi) - \log(\Psi)) - \Phi + \Psi + G^*\Lambda G\Phi] - \mathrm{tr}(\Lambda) \end{aligned}$$

where $\Lambda \in \mathcal{Q}_n$ is the *Lagrange* multiplier. It is easy to see that $L_{\mathrm{KL}}(\cdot, \Lambda)$ is strictly convex over $\mathbb{S}_+^m(\mathbb{T})$. Thus, its unique minimum point is given by annihilating its first directional derivative for each $\delta\Phi \in \mathrm{L}_\infty^{m\times m}(\mathbb{T})$:

$$\delta L_{\mathrm{KL}}(\Phi, \Lambda; \delta\Phi) = \mathrm{tr} \int [\log(\Phi) - \log(\Psi) + G^*\Lambda G]\delta\Phi \tag{2.11}$$

where we exploited the expression for the differential of the logarithm matrix, see Appendix A. Thus, the minimum point for $L_{\mathrm{KL}}(\cdot, \Lambda)$ is

$$\Phi_{\mathrm{KL}}(\Lambda) := \mathrm{e}^{\log(\Psi) - G^*\Lambda G} \tag{2.12}$$

and

$$L_{\mathrm{KL}}(\Phi_{\mathrm{KL}}(\Lambda), \Lambda) \leq L_{\mathrm{KL}}(\Phi, \Lambda), \ \forall\ \Phi \in \mathbb{S}_+^m(\mathbb{T}). \tag{2.13}$$

Following the same lines in [37], it is possible to prove that there exists Λ° such that $\Phi_{\mathrm{KL}}(\Lambda^\circ) \in \mathbb{S}_+^m(\mathbb{T})$ and $\int G\Phi_{\mathrm{KL}}(\Lambda^\circ)G^* = I$. Accordingly, (2.13) implies that $\Phi_{\mathrm{KL}}(\Lambda^\circ)$ is the unique solution to Problem 2.1 with $\mathcal{S}_{\mathrm{KL}}$. The resulting solution is however not rational, even when $\Psi = I$.

A multivariate extension of the *Itakura-Saito* distance has been recently presented by Ferrante *et al.*, [26]:

$$\mathcal{S}_{\mathrm{IS}}(\Phi\|\Psi) = \int \mathrm{tr}[\log(\Psi) - \log(\Phi) + \Phi\Psi^{-1} - I], \tag{2.14}$$

which has an interpretation in terms of relative entropy rate among processes. They have shown that the corresponding solution to Problem 2.1 always admits a unique solution

$$\Phi_{\mathrm{IS}}(\Lambda^\circ) := [\Psi^{-1} + G^*\Lambda^\circ G]^{-1}, \tag{2.15}$$

where Λ° is given by solving the corresponding dual problem. Note that, Φ_{IS} has bounded *McMillan* degree when Ψ is rational.

We will show in Section 2.3 that the divergence indexes (2.10) and (2.14) belong to the same multivariate Beta divergence family.

Remark 2.2. We mention that there exists another multivariate distance, called *Hellinger* distance, which gives a rational solution to Problem 2.1, [27].

2.3 Beta divergence family

The Beta divergence family for scalar spectral densities was firstly introduced in [4]. Then, it has been widely used in many applications: Robust principal

component analysis and clustering [51], robust independent component analysis [49], and robust nonnegative matrix and tensor factorization [22], [21]. In what follows we will adopt the same notation employed in [20]. First of all, we need to introduce the following function

$$\begin{aligned} \log_c : \ & \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R} \\ & (x,y) \mapsto \begin{cases} \frac{1}{1-c}\left[\left(\frac{x}{y}\right)^{1-c} - 1\right], & c \in \mathbb{R} \setminus \{1\} \\ \log(x) - \log(y), & c = 1 \end{cases} \end{aligned} \tag{2.16}$$

which is referred to as *generalized logarithm discrepancy* function throughout the Chapter. Notice that $\log_c$ is a continuous function of real variable c and $\log_c(x,y) = 0$ if and only if $x = y$. The (asymmetric) Beta divergence between two scalar spectral densities $\Phi, \Psi \in \mathbb{S}^1_+(\mathbb{T})$ is defined by

$$\begin{aligned} \mathcal{S}_\beta(\Phi\|\Psi) &:= -\frac{1}{\beta}\int [\Phi^\beta \log_{\frac{1}{\beta}}\left(\Psi^\beta, \Phi^\beta\right) + \Phi^\beta - \Psi^\beta] \\ &= \int [\frac{1}{\beta-1}(\Phi^\beta - \Phi\Psi^{\beta-1}) - \frac{1}{\beta}(\Phi^\beta - \Psi^\beta)] \end{aligned} \tag{2.17}$$

where the parameter β is a real number. For $\beta = 0$ and $\beta = 1$, it is defined by continuity in the following way

$$\begin{aligned} \lim_{\beta\to 0} \mathcal{S}_\beta(\Phi\|\Psi) &= \mathcal{S}_{\mathrm{IS}}(\Phi\|\Psi) \\ \lim_{\beta\to 1} \mathcal{S}_\beta(\Phi\|\Psi) &= \mathcal{S}_{\mathrm{KL}}(\Phi\|\Psi), \end{aligned} \tag{2.18}$$

where $\mathcal{S}_{\mathrm{IS}}$ and $\mathcal{S}_{\mathrm{KL}}$ are the scalar versions of (2.14) and (2.10), respectively. Moreover, the Beta divergence is a continuous function of real variable β in the whole range including singularities. Thus, it smoothly connects the *Itakura-Saito* distance with the *Kullback-Leibler* divergence. Since $\mathcal{S}_\beta$ is a divergence index, property (2.4) is satisfied. Finally, $\mathcal{S}_\beta$ is always strictly convex in the first argument, but is often not in the second argument.

We are now ready to extend the Beta divergence family to multivariate spectral densities. Likewise to the scalar case, we start by introducing the *generalized multivariate logarithm discrepancy*. To this aim, recall that the exponentiation of a positive definite matrix X to an arbitrary real number c, is defined as $X^c := U\mathrm{diag}(d_1^c, \ldots, d_m^c)U^T$ where $X := U\mathrm{diag}(d_1, \ldots, d_m)U^T$ is the usual spectral decomposition with U orthogonal, i.e. $UU^T = I$, and $\mathrm{diag}(d_1, \ldots, d_m) > 0$ diagonal matrix.[2] The *generalized logarithm discrepancy*

[2]It is also possible to take the exponentiation of positive semidefinite matrices when $c \neq 0$.

in the multivariate case is defined as follows

$$\begin{aligned}\log_c : \quad & \mathcal{Q}_{m,+} \times \mathcal{Q}_{m,+} \to \mathcal{Q}_m \\ & (X,Y) \mapsto \begin{cases} \frac{1}{1-c}(X^{1-c}Y^{c-1} - I), & c \in \mathbb{R} \setminus \{1\} \\ \log(X) - \log(Y), & c = 1 \end{cases}\end{aligned} \tag{2.19}$$

where $\log(X) = U\mathrm{diag}(\log(d_1), \ldots, \log(d_m))U^T$ is the matrix logarithm of X.

Proposition 2.3. *The* generalized multivariate logarithm discrepancy *is a continuous function of real variable c in the whole range. Moreover,* $\log_c(X,Y) = 0$ *if and only if* $X = Y$.

Proof. By definition X^{1-c} and Y^{c-1} are continuous function of real variable c. Thus, the function $\log_c(X,Y)$ of real variable c is continuous in $\mathbb{R} \setminus \{1\}$. It remains to prove that $\log_c$ is continuous in $c = 1$. This is equivalent to show that $\lim_{c\to 1} \log_c(X,Y) = \log(X) - \log(Y)$. Let $X = U\mathrm{diag}(d_1, \ldots, d_m)U^T$, then

$$\frac{1}{1-c}(X^{1-c} - I) = U\mathrm{diag}(\frac{d_1^{1-c} - 1}{1-c}, \ldots, \frac{d_m^{1-c} - 1}{1-c})U^T. \tag{2.20}$$

Taking the limit for $c \to 1$, we get

$$\begin{aligned}&\lim_{c\to 1} \frac{1}{1-c}(X^{1-c} - I) \\ &\quad = U\mathrm{diag}(\lim_{c\to 1} \frac{d_1^{1-c} - 1}{1-c}, \ldots, \lim_{c\to 1} \frac{d_m^{1-c} - 1}{1-c})U^T \\ &\quad = U\mathrm{diag}(\log(d_1), \ldots, \log(d_m))U^T = \log(X).\end{aligned} \tag{2.21}$$

Accordingly,

$$\begin{aligned}&\lim_{c\to 1} \log_c(X,Y) \\ &\quad = \lim_{c\to 1}[\frac{1}{1-c}(X^{1-c} - I) - \frac{1}{1-c}(Y^{1-c} - I)]Y^{c-1} \\ &\quad = \lim_{c\to 1}[\frac{1}{1-c}(X^{1-c} - I)] - \lim_{c\to 1}[\frac{1}{1-c}(Y^{1-c} - I)] \\ &\quad = \log(X) - \log(Y)\end{aligned} \tag{2.22}$$

which proves that $\log_c$ is continuous in $c = 1$. Concerning the last statement, it is straightforward that $X = Y$ implies $\log_c(X,Y) = 0$. On the contrary, $\log_c(X,Y) = 0$, with $c \neq 1$, implies $X^{1-c}Y^{c-1} = I$ which is equivalent to $X^{1-c} = Y^{1-c}$, since $X, Y \in \mathcal{Q}_{m,+}$. Thus, $X = Y$. We get the same conclusion for $c = 1$ by exploiting similar argumentations. ■

The exponentiation of a spectral density $\Phi(\mathrm{e}^{j\vartheta}) \in \mathbb{S}_+^m(\mathbb{T})$ to an arbitrary real number c is punctually defined by exploiting the previous spectral decomposition:

$$\Phi(\mathrm{e}^{j\vartheta})^c = U(\mathrm{e}^{j\vartheta})\mathrm{diag}(d_1(\mathrm{e}^{j\vartheta})^c, \ldots, d_m(\mathrm{e}^{j\vartheta})^c)U(\mathrm{e}^{j\vartheta})^T$$

where $\Phi(\mathrm{e}^{j\vartheta}) = U(\mathrm{e}^{j\vartheta})\mathrm{diag}(d_1(\mathrm{e}^{j\vartheta}), \ldots, d_m(\mathrm{e}^{j\vartheta}))U(\mathrm{e}^{j\vartheta})^T$ with $U(\mathrm{e}^{j\vartheta}) \in \mathrm{L}_\infty^{m\times m}(\mathbb{T})$ such that $U(\mathrm{e}^{j\vartheta})U(\mathrm{e}^{j\vartheta})^T = I$. Observe that Φ^c belongs to $\mathbb{S}_+^m(\mathbb{T})$. We are now ready to introduce the multivariate (asymmetric) Beta divergence among $\Phi, \Psi \in \mathbb{S}_+^m(\mathbb{T})$:

$$\begin{aligned} \mathcal{S}_\beta(\Phi\|\Psi) : &= -\frac{1}{\beta}\int \mathrm{tr}[\Phi^\beta \log_{\frac{1}{\beta}}\left(\Psi^\beta, \Phi^\beta\right) + \Phi^\beta - \Psi^\beta] \\ &= \int \mathrm{tr}[\frac{1}{\beta-1}(\Phi^\beta - \Phi\Psi^{\beta-1}) - \frac{1}{\beta}(\Phi^\beta - \Psi^\beta)] \end{aligned} \tag{2.23}$$

where $\beta \in \mathbb{R}\backslash\{0,1\}$. Similarly to the scalar case, we can extend by continuity the definition of Beta divergence for $\beta = 0$ and $\beta = 1$.

Proposition 2.4. *The following limits hold:*

$$\lim_{\beta\to 0} \mathcal{S}_\beta(\Phi\|\Psi) = \mathcal{S}_{\mathrm{IS}}(\Phi\|\Psi) \tag{2.24}$$

$$\lim_{\beta\to 1} \mathcal{S}_\beta(\Phi\|\Psi) = \mathcal{S}_{\mathrm{KL}}(\Phi\|\Psi). \tag{2.25}$$

Proof. Since Φ and Ψ belong to $\mathbb{S}_+^m(\mathbb{T})$, i.e. Φ and Ψ are coercive and bounded, it is possible to show by standard argumentations that the integrand function of (2.23) uniformly converges on $\mathbb{T}$ for $\beta \to 0$ and $\beta \to 1$. Hence, it is allowed to pass the limits, for $\beta \to 0$ and $\beta \to 1$, under the integral sign. Taking into account the first limit, we get

$$\begin{aligned} &\lim_{\beta\to 0} \mathcal{S}_\beta(\Phi\|\Psi) \\ &= \lim_{\beta\to 0} \int \mathrm{tr}[\frac{1}{\beta-1}(\Phi^\beta - \Phi\Psi^{\beta-1}) - \frac{1}{\beta}(\Phi^\beta - \Psi^\beta)] \\ &= \int \mathrm{tr}\left\{-I + \Phi\Psi^{-1} - \lim_{\beta\to 0}\frac{1}{\beta}[(\Phi^\beta - I) - (\Psi^\beta - I)]\right\} \\ &= \int \mathrm{tr}[-I + \Phi\Psi^{-1} - \log(\Phi) + \log(\Psi)] \\ &= \mathcal{S}_{\mathrm{IS}}(\Phi\|\Psi) \end{aligned} \tag{2.26}$$

where we exploited (2.21). For the second limit, we obtain

$$\begin{aligned}
&\lim_{\beta\to 1}\mathcal{S}_\beta(\Phi\|\Psi)\\
&=\lim_{\beta\to 1}\left\{-\frac{1}{\beta}\int \mathrm{tr}[\Phi^\beta \log_{\frac{1}{\beta}}\left(\Psi^\beta,\Phi^\beta\right)+\Phi^\beta-\Psi^\beta]\right\}\\
&=-\int \mathrm{tr}[\Phi\lim_{\beta\to 1}\log_{\frac{1}{\beta}}\left(\Psi^\beta,\Phi^\beta\right)+\Phi-\Psi]\\
&=-\int \mathrm{tr}[\Phi\lim_{\beta\to 1}\log_{2-\beta}\left(\Psi,\Phi\right)+\Phi-\Psi]\\
&=\int \mathrm{tr}[\Phi\left(\log(\Phi)-\log(\Psi)\right)+\Psi-\Phi]\\
&=\mathcal{S}_{\mathrm{KL}}(\Phi\|\Psi)
\end{aligned}\tag{2.27}$$

where we exploited (2.22). ■

In view of Proposition 2.3 and Proposition 2.4, we conclude that the multivariate Beta divergence is a continuous function of real variable β in the whole range including singularities and it smoothly connects the multivariate *Itakura-Saito* distance with the multivariate *Kullback-Leibler* divergence.

Remark 2.5. For $\beta = 2$, the Beta divergence corresponds, up to a constant scalar factor, to the standard squared *Euclidean* distance (L_2-norm)

$$\mathcal{S}_{\mathrm{L2}}(\Phi\|\Psi)=\int\langle\Phi-\Psi,\Phi-\Psi\rangle \tag{2.28}$$

where $\langle X,Y\rangle=\mathrm{tr}(XY)$ is the usual scalar product in $\mathcal{Q}_m$.

Finally, we show that the multivariate Beta divergence satisfies condition (2.4).

Proposition 2.6. *Given $\Phi,\Psi\in\mathbb{S}^m_+(\mathbb{T})$, the following facts hold:*

1. *$\mathcal{S}_\beta(\cdot\|\Psi)$ is strictly convex over $\mathbb{S}^m_+(\mathbb{T})$,*
2. *$\mathcal{S}_\beta(\Phi\|\Psi)\geq 0$ and equality holds if and only if $\Psi=\Phi$.*

Proof. In order to prove the statements, we need of the following first variations of the maps $X\mapsto\mathrm{tr}(X^c)$ and $X\mapsto\mathrm{tr}(X^cY)$, respectively (further details may be found in Appendix A):

$$\begin{aligned}
&\delta(\mathrm{tr}[X^c];\delta X)=c\mathrm{tr}[X^{c-1}\delta X]\\
&\delta(\mathrm{tr}(X^cY);\delta X)=\mathrm{tr}[O_{X,c}(\delta X)Y],
\end{aligned}\tag{2.29}$$

where $Y \in \mathcal{Q}_m$ and the map $O_{X,c}$ is defined in (A.4).
1) The first variation of $\mathcal{S}_\beta(\Phi\|\Psi)$, with respect to Φ, in direction $\delta\Phi \in \mathrm{L}_\infty^{m\times m}(\mathbb{T})$ is

$$\delta(\mathcal{S}_\beta(\Phi\|\Psi);\delta\Phi) = \frac{1}{\beta-1}\int_0^{2\pi} \mathrm{tr}[(\Phi^{\beta-1}-\Psi^{\beta-1})\delta\Phi]\frac{\mathrm{d}\vartheta}{2\pi}. \tag{2.30}$$

The second variation in direction $\delta\Phi$ is

$$\begin{aligned}
\delta^2(\mathcal{S}_\beta(\Phi\|\Psi);\delta\Phi) &= \frac{1}{\beta-1}\int_0^{2\pi} \mathrm{tr}[O_{\Phi,\beta-1}(\delta\Phi)\delta\Phi]\frac{\mathrm{d}\vartheta}{2\pi} \\
&= \int_0^{2\pi} \mathrm{tr}\left[\int_0^1 \Phi^{(\beta-1)(1-\tau)}\int_0^\infty (\Phi+tI)^{-1}\delta\Phi\right. \\
&\qquad \left.\times\,(\Phi+tI)^{-1}\mathrm{d}t\Phi^{(\beta-1)\tau}\mathrm{d}\tau\delta\Phi\right]\frac{\mathrm{d}\vartheta}{2\pi} \\
&= \int_0^{2\pi}\int_0^1\int_0^\infty \mathrm{tr}[\Phi^{(\beta-1)(1-\tau)}(\Phi+tI)^{-1}\delta\Phi \\
&\qquad \times(\Phi+tI)^{-1}\Phi^{(\beta-1)\tau}\delta\Phi]\mathrm{d}t\mathrm{d}\tau\frac{\mathrm{d}\vartheta}{2\pi}.
\end{aligned}$$

By the cyclic property of the trace and since $\Phi^{(\beta-1)\tau}$ and $(\Phi+tI)^{-1}$ commute, we get

$$\delta^2(\mathcal{S}_\beta(\Phi\|\Psi);\delta\Phi) = \int_0^{2\pi}\int_0^1\int_0^\infty f_{t,\tau}(\Phi,\delta\Phi)\mathrm{d}t\mathrm{d}\tau\frac{\mathrm{d}\vartheta}{2\pi} \tag{2.31}$$

where

$$\begin{aligned}
f_{t,\tau}(X,\Delta) = \mathrm{tr}[&X^{\frac{(\beta-1)\tau}{2}}(X+tI)^{-\frac{1}{2}}\Delta(X+tI)^{-\frac{1}{2}} \\
&\times X^{(\beta-1)(1-\tau)}(X+tI)^{-\frac{1}{2}}\Delta(X+tI)^{-\frac{1}{2}}X^{\frac{(\beta-1)\tau}{2}}]
\end{aligned} \tag{2.32}$$

with $X \in \mathcal{Q}_{m,+}$, $\Delta \in \mathcal{Q}_m$, $t \in [0,\infty)$ and $\tau \in [0,1]$. Thus, $f_{t,\tau}(X,\Delta) \geq 0$ and $f_{t,\tau}(X,\Delta) = 0$ if and only if $\Delta = 0$. We conclude that integral (2.31), i.e. the second variation of $\mathcal{S}_\beta(\cdot\|\Psi)$, is positive for $\delta\Phi \neq 0$. Accordingly, $\mathcal{S}_\beta(\cdot\|\Psi)$ is strictly convex over the convex set $\mathbb{S}_+^m(\mathbb{T})$.
2) As a consequence of the previous statement, the minimum point is unique and it is given by annihilating (2.30) for each $\delta\Phi \in \mathrm{L}_\infty^{m\times m}(\mathbb{T})$. Since $\Phi^{\beta-1} - \Psi^{\beta-1} \in \mathrm{L}_\infty^{m\times m}(\mathbb{T})$, it follows that the minimum point satisfies the condition $\Phi^{\beta-1} = \Psi^{\beta-1}$. Accordingly, $\Phi = \Psi$. Finally it is sufficient to observe that $\mathcal{S}_\beta(\Psi\|\Psi) = 0$. ■

Note that $\mathcal{S}_\beta(\Phi\|\cdot)$ is not convex on $\mathbb{S}_+^m(\mathbb{T})$ (not even in the scalar case).

2.4 Spectrum approximation problem

Since the Beta divergence is well-defined for $\beta \in \mathbb{R}$, we choose $\beta = -\frac{1}{\nu}+1$ with $\nu \in \mathbb{Z} \setminus \{0\}$. As we will see, this choice guarantees that the corresponding solution to Problem 2.1 is rational for a suitable choice of Ψ. In order to simplify the notation we define $\mathcal{S}_\nu(\Phi\|\Psi) := \mathcal{S}_\beta(\Phi\|\Psi)$ with $\beta = -\frac{1}{\nu}+1$. We have to minimize $\mathcal{S}_\nu(\Phi\|\Psi)$ over $\left\{\Phi \in \mathbb{S}_+^m(\mathbb{T}) \mid \int G\Phi G^* = I\right\}$. Since it is a constrained convex optimization problem, we consider the corresponding *Lagrange* functional

$$\begin{aligned} & L_\nu(\Phi,\Lambda) \\ & \quad = \mathcal{S}_\nu(\Phi\|\Psi) + \frac{\nu}{1-\nu}\int \mathrm{tr}(\Psi^{\frac{\nu-1}{\nu}}) + \left\langle \int G\Phi G^* - I, \Lambda \right\rangle \\ & \quad = \int \mathrm{tr}\left[-\nu(\Phi^{\frac{\nu-1}{\nu}} - \Phi\Psi^{-\frac{1}{\nu}}) + \frac{\nu}{1-\nu}\Phi^{\frac{\nu-1}{\nu}} + G^*\Lambda G\Phi\right] - \mathrm{tr}(\Lambda) \end{aligned} \quad (2.33)$$

where we exploited the fact that the term $\int \mathrm{tr}(\Psi^{\frac{\nu-1}{\nu}})$ plays no role in the optimization problem. Note that, the *Lagrange* multiplier $\Lambda \in \mathcal{Q}_n$ can be uniquely decomposed as $\Lambda = \Lambda_\Gamma + \Lambda_\perp$ where $\Lambda_\Gamma \in \text{Range } \Gamma$, $\Lambda_\perp \in [\text{Range } \Gamma]^\perp$. Since $\Lambda_\perp$ is such that $G^*(\mathrm{e}^{j\vartheta})\Lambda_\perp G(\mathrm{e}^{j\vartheta}) \equiv 0$ and $\mathrm{tr}(\Lambda_\perp) = \langle \Lambda_\perp, I\rangle = 0$ (see [56, Section III]), it does not affect the *Lagrangian*, i.e. $L_\nu(\Phi,\Lambda) = L_\nu(\Phi,\Lambda_\Gamma)$. Accordingly we can impose from now on that $\Lambda \in \text{Range } \Gamma$.

Consider now the unconstrained minimization problem $\min_\Phi \{L_\nu(\Phi,\Lambda) \mid \Phi \in \mathbb{S}_+^m(\mathbb{T})\}$. Since $L_\nu(\cdot,\Lambda)$ is strictly convex over $\mathbb{S}_+^m(\mathbb{T})$, its unique minimum point Φ_ν is given by annihilating its first variation in each direction $\delta\Phi \in \mathrm{L}_\infty^{m\times m}(\mathbb{T})$:

$$\delta L_\nu(\Phi,\Lambda;\delta\Phi) = \int \mathrm{tr}\left[\left(\nu(\Psi^{-\frac{1}{\nu}} - \Phi^{-\frac{1}{\nu}}) + G^*\Lambda G\right)\delta\Phi\right] \quad (2.34)$$

where we exploited (2.29). Note that, $\nu(\Psi^{-\frac{1}{\nu}} - \Phi^{-\frac{1}{\nu}}) + G^*\Lambda G \in \mathrm{L}_\infty^{m\times m}(\mathbb{T})$. Thus, (2.34) is zero $\forall\Phi \in \mathrm{L}_\infty^{m\times m}(\mathbb{T})$ if and only if

$$\Phi^{-\frac{1}{\nu}} = \Psi^{-\frac{1}{\nu}} + \frac{1}{\nu}G^*\Lambda G. \quad (2.35)$$

Since $\Phi^{-\frac{1}{\nu}} \in \mathbb{S}_+^m(\mathbb{T})$, the set of the admissible *Lagrange* multipliers is

$$\mathcal{L}_\nu := \left\{\Lambda \in \mathcal{Q}_n \mid \Psi^{-\frac{1}{\nu}} + \frac{1}{\nu}G^*\Lambda G > 0 \text{ on } \mathbb{T}\right\}. \quad (2.36)$$

Therefore, the natural set for Λ is

$$\mathcal{L}_\nu^\Gamma = \mathcal{L}_\nu \cap \text{Range } \Gamma. \quad (2.37)$$

In conclusion, the unique minimum point of the *Lagrange* functional has the form

$$\Phi_\nu(\Lambda) := [\Psi^{-\frac{1}{\nu}} + \frac{1}{\nu} G^* \Lambda G]^{-\nu}. \tag{2.38}$$

Assuming that $\Psi^{\frac{1}{\nu}}$ is rational in $\mathrm{e}^{j\vartheta}$, there always exists a unique (up to a right-multiplication by a constant orthogonal matrix) stable and minimum phase spectral factor W such that $\Psi(\mathrm{e}^{j\vartheta})^{\frac{1}{\nu}} = W(\mathrm{e}^{j\vartheta})W(\mathrm{e}^{j\vartheta})^*$. By defining $G_1(\mathrm{e}^{j\vartheta}) = \frac{1}{\sqrt{\nu}} G(\mathrm{e}^{j\vartheta})W(\mathrm{e}^{j\vartheta})$, we obtain an equivalent form of (2.38):

$$\Phi_\nu(\Lambda) = [W(I + G_1^* \Lambda G_1)^{-1} W^*]^\nu. \tag{2.39}$$

Corollary 2.7. *Assume that $\Psi^{\frac{1}{\nu}}$ has bounded* McMillan *degree. Then, Φ_ν is rational in $\mathrm{e}^{j\vartheta}$ with* McMillan *degree less than or equal to $|\nu|(\deg[\Psi^{\frac{1}{\nu}}] + 2n)$. Moreover, among all the spectral densities Φ_ν with $\nu \in \mathbb{Z} \setminus \{0\}$, the spectral densities with the smallest upper bound on the McMillan degree correspond to the* Itakura-Saito *and the squared* Euclidean *distance.*

Proof. In view of (2.38) and (2.39), $\deg[\Phi_\nu] \leq |\nu|(\deg[\Psi^{\frac{1}{\nu}}] + 2n)$ where n is the *McMillan* degree of $G(z)$. Since $\nu \in \mathbb{Z} \setminus \{0\}$, the spectral densities with the smallest upper bound on the *McMillan* degree are attained for $\nu = \pm 1$, i.e. $\beta = 0$ and $\beta = 2$, which are the optimal forms related to $\mathcal{S}_{\mathrm{IS}}(\Phi\|\Psi)$ and $\mathcal{S}_{\mathrm{L2}}(\Phi\|\Psi)$, respectively. Note that, $\Phi_1(\Lambda) = [\Psi^{-1} + G^*\Lambda G]^{-1}$, which is the same optimal form found in [26] for the multivariate *Itakura-Saito* distance, and $\Phi_{-1}(\Lambda) = \Psi - G^*\Lambda G$. ■

Corollary 2.8. *As $\nu \to \pm\infty$, Φ_ν tends to the spectral density (2.12) corresponding to the* Kullback-Leibler *divergence.*

Proof. We know that that the optimal form obtained by using the *Kullback-Leibler* divergence is $\Phi_{\mathrm{KL}}(\Lambda) = \mathrm{e}^{\log(\Psi) - G^*\Lambda G}$. We want to show that $\Phi_\nu \to \Phi_{\mathrm{KL}}$ as $\nu \to \pm\infty$. Let us consider the function $F(\lambda) := \log(\Psi^{-\lambda} + \lambda G^*\Lambda G)$ with $\lambda \in \mathbb{R}$ such that $\Psi^{-\lambda} + \lambda G^*\Lambda G > 0$ on $\mathbb{T}$. Its first order *Taylor* expansion with respect to $\lambda = 0$ is $\Psi^{-\lambda} + \lambda G^*\Lambda G - I$. Accordingly,

$$\begin{aligned} &\lim_{\nu\to\pm\infty} \nu \log(\Psi^{-\frac{1}{\nu}} + \frac{1}{\nu} G^*\Lambda G) \\ &\quad = \lim_{\nu\to\pm\infty} \frac{\Psi^{-\frac{1}{\nu}} - I}{\nu^{-1}} + G^*\Lambda G \\ &\quad = -\log(\Psi) + G^*\Lambda G \end{aligned} \tag{2.40}$$

where we exploited (2.21) and the previous *Taylor* expansion. Finally,

$$\begin{aligned}\lim_{\nu\to\pm\infty}\Phi_\nu(\Lambda) &= \lim_{\nu\to\pm\infty}\mathrm{e}^{\log[(\Psi^{-\frac{1}{\nu}}+\frac{1}{\nu}G^*\Lambda G)^{-\nu}]}\\ &= \lim_{\nu\to\pm\infty}\mathrm{e}^{-\nu\log(\Psi^{-\frac{1}{\nu}}+\frac{1}{\nu}G^*\Lambda G)}\\ &= \mathrm{e}^{-\lim_{\nu\to\pm\infty}\nu\log(\Psi^{-\frac{1}{\nu}}+\frac{1}{\nu}G^*\Lambda G)}\\ &= \mathrm{e}^{\log(\Psi)-G^*\Lambda G}=\Phi_{\mathrm{KL}}(\Lambda). \end{aligned} \tag{2.41}$$

■

In this section we showed that $\Phi_\nu(\Lambda)$ is the unique minimum point of $L_\nu(\cdot,\Lambda)$, namely

$$L_\nu(\Phi_\nu(\Lambda),\Lambda)\le L_\nu(\Phi,\Lambda),\ \ \forall\Phi\in\mathbb{S}_+^m(\mathbb{T}). \tag{2.42}$$

Hence, if we produce $\Lambda^\circ\in\mathcal{L}_\nu^\Gamma$ satisfying constraint in (2.3), inequality (2.42) implies

$$\mathcal{S}_\nu(\Phi_\nu(\Lambda^\circ)\|\Psi)\le\mathcal{S}_\nu(\Phi\|\Psi),\ \ \forall\Phi\in\mathbb{S}_+^m(\mathbb{T})\text{ s.t. }\int G\Phi G^*=I \tag{2.43}$$

namely such a $\Phi_\nu(\Lambda^\circ)$ is the unique solution to Problem 2.1 with $\mathcal{S}_\nu$. The following step consists in showing the existence of such a Λ° by exploiting the duality theory.

2.5 Dual problem

Here, we do not deal with the case $\nu=1$, since the existence of the solution to the dual problem was already showed in [26]. We start by considering the case $\nu\in\mathbb{N}_+\setminus\{1\}$. The dual problem consists in maximizing the functional

$$\begin{aligned}\inf_\Phi L_\nu(\Phi,\Lambda) &= L_\nu(\Phi_\nu,\Lambda)\\ &= \frac{\nu}{1-\nu}\int\mathrm{tr}[(\Psi^{-\frac{1}{\nu}}+\frac{1}{\nu}G^*\Lambda G)^{1-\nu}]-\mathrm{tr}(\Lambda)\end{aligned} \tag{2.44}$$

which is equivalent to minimize the following functional hereafter referred to as *dual functional*:

$$J_\nu(\Lambda)=-\frac{\nu}{1-\nu}\int\mathrm{tr}[(\Psi^{-\frac{1}{\nu}}+\frac{1}{\nu}G^*\Lambda G)^{1-\nu}]+\mathrm{tr}(\Lambda). \tag{2.45}$$

Theorem 2.9. *The dual functional J_ν belongs to $\mathcal{C}^\infty(\mathcal{L}_\nu^\Gamma)$ and it is strictly convex over $\mathcal{L}_\nu^\Gamma$.*

Proof. In view of (2.29), the first variation of $J_\nu(\Lambda)$ in direction $\delta\Lambda_1 \in \mathcal{Q}_n$ is

$$\begin{aligned}
&\delta J_\nu(\Lambda;\delta\Lambda_1)\\
&\quad= -\int \mathrm{tr}[(\Psi^{-\frac{1}{\nu}} + \frac{1}{\nu}G^*\Lambda G)^{-\nu}G^*\delta\Lambda_1 G] + \mathrm{tr}(\delta\Lambda_1)\\
&\quad= -\int \mathrm{tr}\left[\left(W(I+G_1^*\Lambda G_1)^{-1}W^*\right)^{\nu} G^*\delta\Lambda_1 G\right] + \mathrm{tr}(\delta\Lambda_1). \qquad (2.46)
\end{aligned}$$

The linear form $\nabla J_{\nu,\Lambda}(\cdot) := \delta J_\nu(\Lambda;\cdot)$ is the *gradient* of J_ν at Λ. In order to prove that $J_\nu(\Lambda) \in \mathcal{C}^1(\mathcal{L}_\nu^\Gamma)$ we have to show that $\delta(J_\nu(\Lambda);\delta\Lambda_1)$, for any fixed $\delta\Lambda_1$, is continuous in Λ. To this aim, consider a sequence $M_n \in \mathrm{Range}\,\Gamma$ such that $M_n \to 0$ and define $Q_N(z) = W(z)(I + G_1(z)^*NG_1(z))^{-1}W(z)^*$ with $N \in \mathcal{Q}_n$. By Lemma 5.2 in [56] and since W is bounded on $\mathbb{T}$, $Q_{\Lambda+M_n}$ converges uniformly to Q_Λ. Thus, applying elementwise the bounded convergence theorem, we obtain

$$\lim_{n\to\infty}\int GQ_{\Lambda+M_n}^\nu G^* = \int GQ_\Lambda^\nu G^*. \qquad (2.47)$$

Accordingly, $\delta(J_\nu(\Lambda);\delta\Lambda)$ is continuous, i.e. J_ν belongs to $\mathcal{C}^1(\mathcal{L}_\nu^\Gamma)$. In order to compute the second variation, consider the operator $\mathcal{I} : A \mapsto A^{-\nu}$. By applying the chain rule, we get

$$\delta(\mathcal{I}(A);\delta A) = -\sum_{l=1}^{\nu} A^{-l}\delta A A^{l-\nu-1}. \qquad (2.48)$$

Thus, for $\delta\Lambda_1, \delta\Lambda_2 \in \mathcal{Q}_n$ we have

$$\begin{aligned}
&\delta^2 J_\nu(\Lambda;\delta\Lambda_1,\delta\Lambda_2)\\
&\quad= \frac{1}{\nu}\sum_{l=1}^{\nu}\int \mathrm{tr}[Q_\Lambda^l G^*\delta\Lambda_2 G Q_\Lambda^{\nu+1-l}G^*\delta\Lambda_1 G]. \qquad (2.49)
\end{aligned}$$

The bilinear form $\mathcal{H}_{\nu,\Lambda}(\cdot,\cdot) = \delta^2 J_\nu(\Lambda;\cdot,\cdot)$ is the *Hessian* of J_ν at Λ. The continuity of $\delta^2 J_\nu$ can be established by using the previous argumentation. In similar way, we can show that J_ν has continuous directional derivatives of any order, i.e. $J_\nu \in \mathcal{C}^k(\mathcal{L}_\nu^\Gamma)$ for any k. Finally, it remains to be shown that J_ν is strictly convex on the open set $\mathcal{L}_\nu^\Gamma$. Since $J_\nu \in \mathcal{C}^\infty(\mathcal{L}_\nu^\Gamma)$, it is sufficient to show that $\mathcal{H}_\Lambda(\delta\Lambda,\delta\Lambda) \geq 0$ for each $\delta\Lambda \in \mathrm{Range}\,\Gamma$ and equality holds if and only if $\delta\Lambda = 0$. Since $\nu > 0$ and the integrands in (2.49) are positive semidefinite when $\delta\Lambda_1 = \delta\Lambda_2$, we have $\mathcal{H}_\Lambda(\delta\Lambda,\delta\Lambda) \geq 0$. If $\mathcal{H}_\Lambda(\delta\Lambda,\delta\Lambda) = 0$, then $G^*\delta\Lambda G \equiv 0$ namely $\delta\Lambda \in [\mathrm{Range}\,\Gamma]^\perp$ (see [56, Section III]). Since $\delta\Lambda \in \mathrm{Range}\,\Gamma$, it follows that $\delta\Lambda = 0$. In conclusion, the Hessian is positive

definite and the dual functional is strictly convex on $\mathcal{L}_\nu^\Gamma$. ■

In view of Theorem 2.9, the dual problem $\min_\Lambda \left\{ J_\nu(\Lambda) \mid \Lambda \in \mathcal{L}_\nu^\Gamma \right\}$ admits at most one solution Λ°. Since $\mathcal{L}_\nu^\Gamma$ is an open set, such a Λ° (if it does exist) annihilates the first directional derivative (2.46) for each $\delta\Lambda \in \mathcal{Q}_n$

$$\left\langle I - \int [G(\Psi^{-\frac{1}{\nu}} + \frac{1}{\nu} G^* \Lambda^\circ G)^{-\nu} G^*, \delta\Lambda \right\rangle = 0 \ \ \forall \delta\Lambda \in \mathcal{Q}_n \tag{2.50}$$

or, equivalently,

$$I = \int G(\Psi^{-\frac{1}{\nu}} + \frac{1}{\nu} G^* \Lambda^\circ G)^{-\nu} G^* = \int G \Phi_\nu(\Lambda^\circ) G^*. \tag{2.51}$$

This means that $\Phi_\nu(\Lambda^\circ) \in \mathbb{S}_+^m(\mathbb{T})$ satisfies the constraint in (2.3) and $\Phi_\nu(\Lambda^\circ)$ is therefore the unique solution to Problem 2.1.

The next step concerns the existence issue for the dual problem. Although the existence question is quite delicate, since set $\mathcal{L}_\nu^\Gamma$ is open and unbounded, we will show that a Λ° minimizing J_ν over $\mathcal{L}_\nu^\Gamma$ does exist.

Theorem 2.10. *Let $\nu \in \mathbb{N}_+ \setminus \{1\}$, then the dual functional J_ν has a unique minimum point in $\mathcal{L}_\nu^\Gamma$.*

Proof. Since the solution of the dual problem (if it does exist) is unique, we only need to show that J_ν takes a minimum value on $\mathcal{L}_\nu^\Gamma$. First of all, note that J_ν is continuous on $\mathcal{L}_\nu^\Gamma$, see Theorem 2.9. Secondly, we show that $\mathrm{tr}[\Lambda]$ is bounded from below on $\mathcal{L}_\nu^\Gamma$. Since Problem 2.1 is feasible, there exists $\Phi_I \in \mathbb{S}_+^m(\mathbb{T})$ such that $\int G \Phi_I G^* = I$. Thus,

$$\mathrm{tr}[\Lambda] = \mathrm{tr}\left[\int G \Phi_I G^* \Lambda \right] = \mathrm{tr}\left[\int G^* \Lambda G \Phi_I \right]. \tag{2.52}$$

Defining $\alpha = -\nu \mathrm{tr} \int \Psi^{-\frac{1}{\nu}} \Phi_I$, we obtain

$$\mathrm{tr}[\Lambda] = \nu \mathrm{tr}\left[\int (\Psi^{-\frac{1}{\nu}} + \frac{1}{\nu} G^* \Lambda G) \Phi_I \right] + \alpha. \tag{2.53}$$

Since $\Psi^{-\frac{1}{\nu}} + \frac{1}{\nu} G^* \Lambda G$ is positive definite on $\mathbb{T}$ for $\Lambda \in \mathcal{L}_\nu^\Gamma$, there exists a right spectral factor Δ such that $\Psi^{-\frac{1}{\nu}} + \frac{1}{\nu} G^* \Lambda G = \Delta^* \Delta$. Moreover, Φ_I is a coercive spectrum, namely there exists a constant $\mu > 0$ such that

$\Phi_I(\mathrm{e}^{j\vartheta}) \geq \mu I$, $\forall\ \mathrm{e}^{j\vartheta} \in \mathbb{T}$. Starting from the fact that the trace and the integral are monotonic functions, we get

$$\begin{aligned} \mathrm{tr}[\Lambda] &= \nu \mathrm{tr}\left[\int \Delta\Phi_I\Delta^*\right] + \alpha \geq \nu\mu \mathrm{tr}\left[\int \Delta\Delta^*\right] + \alpha \\ &= \nu\mu \mathrm{tr}\left[\int \Psi^{-\frac{1}{\nu}} + \frac{1}{\nu} G^*\Lambda G\right] + \alpha > \alpha \end{aligned} \tag{2.54}$$

where we have used the fact that $\mathrm{tr}\int \Psi^{\frac{1}{\nu}} + \frac{1}{\nu}G^*\Lambda G > 0$ when $\Lambda \in \mathcal{L}_\nu^\Gamma$. Finally, notice that $J_\nu(0) = -\frac{\nu}{1-\nu}\int \mathrm{tr}(\Psi^{\frac{\nu-1}{\nu}})$. Accordingly, we can restrict the search of a minimum point to the set $\left\{\Lambda \in \mathcal{L}_\nu^\Gamma \mid J_\nu(\Lambda) \leq J_\nu(0)\right\}$. We now show that this set is compact. Accordingly, the existence of the solution to the dual problem follows from the Weierstrass' Theorem. To prove the *compactness* of the set, it is sufficient to show that:

1. $\lim\limits_{\Lambda \to \partial\mathcal{L}_\nu^\Gamma} J_\nu(\Lambda) = +\infty$;

2. $\lim\limits_{\|\Lambda\| \to \infty} J_\nu(\Lambda) = +\infty$.

1) Firstly, $R_\Lambda(z) := \Psi^{-\frac{1}{\nu}}(z) + \frac{1}{\nu}G(z)^*\Lambda G(z)$ is a rational matrix function, thus $R_\Lambda(z)^{1-\nu}$ is rational as well. Observe that $\partial\mathcal{L}_\nu^\Gamma$ is the set of $\Lambda \in \mathrm{Range}\ \Gamma$ such that $R_\Lambda(\mathrm{e}^{j\vartheta}) \geq 0$ on $\mathbb{T}$ and there exists ϑ such that $R_\Lambda(\mathrm{e}^{j\vartheta})$ is singular. Thus, for $\Lambda \to \partial\mathcal{L}_\nu^\Gamma$ all the eigenvalues of $R_\Lambda(z)^{-1}$ are positive on $\mathbb{T}$ and at least one of them has a pole tending to the unit circle. Since $1-\nu < -1$, then also $R_\Lambda(z)^{1-\nu}$ has at least one eigenvalue with a pole tending to $\mathbb{T}$. Accordingly, $\mathrm{tr}[\int R^{1-\nu}] \to \infty$ as $\Lambda \to \partial\mathcal{L}_\nu^\Gamma$. In view of (2.54), we conclude that $J_\nu(\Lambda) = -\frac{\nu}{1-\nu}\mathrm{tr}\left[\int R^{1-\nu}\right] + \mathrm{tr}[\Lambda] \to \infty$ as $\Lambda \to \partial\mathcal{L}_\nu^\Gamma$.
2) Consider a sequence $\{\Lambda_k\}_{k\in\mathbb{N}} \in \mathcal{L}_\nu^\Gamma$, such that

$$\lim_{k\to\infty} \|\Lambda_k\| = \infty. \tag{2.55}$$

Let $\Lambda_k^0 = \frac{\Lambda_k}{\|\Lambda_k\|}$. Since $\mathcal{L}_\nu^\Gamma$ is convex and $0 \in \mathcal{L}_\nu^\Gamma$, if $\Lambda \in \mathcal{L}_\nu^\Gamma$ then $\xi\Lambda \in \mathcal{L}_\nu^\Gamma$ $\forall\ \xi \in [0,1]$. Therefore $\Lambda_k^0 \in \mathcal{L}_\nu^\Gamma$ for k sufficiently large. Let $\eta := \liminf \mathrm{tr}[\Lambda_k^0]$. In view of (2.54),

$$\mathrm{tr}[\Lambda_k^0] = \frac{1}{\|\Lambda_k\|}\mathrm{tr}[\Lambda_k] > \frac{1}{\|\Lambda_k\|}\alpha \to 0, \tag{2.56}$$

for $k \to \infty$, so $\eta \geq 0$. Thus, there exists a subsequence of $\{\Lambda_k^0\}$ such that the limit of its trace is equal to η. Moreover, this subsequence remains on the surface of the unit ball $\partial\mathcal{B} = \left\{\Lambda = \Lambda^T \mid \|\Lambda\| = 1\right\}$ which is compact.

Accordingly, it has a subsequence $\{\Lambda^0_{k_i}\}$ converging in $\partial\mathcal{B}$. Let $\Lambda^\infty \in \partial\mathcal{B}$ be its limit, thus $\lim_{i\to\infty} \mathrm{tr}[\Lambda^0_{k_i}] = \mathrm{tr}[\Lambda^\infty] = \eta$. We now prove that $\Lambda^\infty \in \mathcal{L}^\Gamma_\nu$. First of all, note that Λ^∞ is the limit of a sequence in the finite dimensional linear space Range Γ, hence $\Lambda^\infty \in$ Range Γ. It remains to be shown that $\Psi^{-\frac{1}{\nu}} + \frac{1}{\nu} G^* \Lambda^\infty G$ is positive definite on $\mathbb{T}$. Consider the unnormalized sequence $\{\Lambda_{k_i}\} \in \mathcal{L}^\Gamma_\nu$: We have that $\Psi^{-\frac{1}{\nu}} + \frac{1}{\nu} G^* \Lambda_{k_i} G > 0$ on $\mathbb{T}$ so that $\frac{1}{\|\Lambda_{k_i}\|}\Psi^{-\frac{1}{\nu}} + \frac{1}{\nu} G^* \Lambda^0_{k_i} G$ is also positive definite on $\mathbb{T}$ for each i. Taking the limit for $i \to \infty$, we get that $G^*\Lambda^\infty G$ is positive semidefinite on $\mathbb{T}$ so that $\Psi^{-\frac{1}{\nu}} + \frac{1}{\nu} G^* \Lambda^\infty G > 0$ on $\mathbb{T}$. Hence, $\Lambda^\infty \in \mathcal{L}^\Gamma_\nu$. Since Problem 2.1 is feasible, there exists $\Phi_I \in \mathbb{S}^m_+(\mathbb{T})$ such that $I = \int G\Phi_I G^*$, accordingly

$$\eta = \mathrm{tr}[\Lambda^\infty] = \mathrm{tr} \int G\Phi_I G^* \Lambda^\infty = \mathrm{tr} \int \Phi_I^{\frac{1}{2}} G^* \Lambda^\infty G \Phi_I^{\frac{1}{2}}. \tag{2.57}$$

Moreover, $G^*\Lambda^\infty G$ is not identically equal to zero. In fact, if $G^*\Lambda^\infty G \equiv 0$, then $\Lambda^\infty \in [\text{Range } \Gamma]^\perp$ and $\Lambda^\infty \neq 0$ since it belongs to the surface of the unit ball. This is a contradiction because $\Lambda^\infty \in$ Range Γ. Thus, $G^*\Lambda^\infty G$ is not identically zero and $\eta > 0$. Finally, we have

$$\begin{aligned} &\lim_{k\to\infty} J_\nu(\Lambda_k) \\ &\quad = \lim_{k\to\infty} -\frac{\nu}{1-\nu} \mathrm{tr}\left[\int (\Psi^{-\frac{1}{\nu}} + \frac{1}{\nu} G^* \Lambda_k G)^{1-\nu}\right] + \mathrm{tr}[\Lambda_k] \\ &\quad \geq \lim_{k\to\infty} \|\Lambda_k\| \mathrm{tr}[\Lambda^0_k] = \eta \lim_{k\to\infty} \|\Lambda_k\| = \infty. \end{aligned} \tag{2.58}$$

■

It remains to deal with the case $\nu \in \mathbb{Z}$ such that $\nu < 0$. In this situation, the dual problem may not have solution: The minimum point for $J_\nu(\Lambda)$ may lie on $\partial\mathcal{L}^\Gamma_\nu$, since J_ν takes finite values on the boundary of $\mathcal{L}^\Gamma_\nu$.

2.6 Computation of Λ°

We showed that the dual problem always admits a unique solution Λ° on $\mathcal{L}^\Gamma_\nu$ for $\nu \in \mathbb{N}_+$. In order to find Λ°, we exploit the following matricial *Newton* algorithm with backtracking stage proposed in [56]:

1. Set $\Lambda_0 = I \in \mathcal{L}^\Gamma_\nu$;
2. At each iteration, compute the *Newton* step Δ_{Λ_i} by solving the linear equation $\mathcal{H}_{\nu,\Lambda_i}(\Delta_{\Lambda_i}, \cdot) = -\nabla J_{\nu,\Lambda_i}(\cdot)$ where, once fixed Λ_i, $\nabla J_{\nu,\Lambda_i}(\cdot)$ and $\mathcal{H}_{\nu,\Lambda_i}(\cdot,\cdot)$ must be understood as a linear and bilinear form of (2.46) and (2.49), respectively;

3. Set $t_i^0 = 1$ and let $t_i^{k+1} = t_i^k/2$ until both of the following conditions hold:

$$\Lambda_i + t_i^k \Delta_{\Lambda_i} \in \mathcal{L}_\nu^\Gamma \tag{2.59}$$
$$J_\nu(\Lambda_i + t_i^k \Delta_{\Lambda_i}) < J_\nu(\Lambda_i) + \alpha t_i^k \langle \nabla J_{\nu,\Lambda_i}, \Delta_{\Lambda_i} \rangle \tag{2.60}$$

   with $0 < \alpha < 1/2$;

4. Set $\Lambda_{i+1} = \Lambda_i + t_i^k \Delta_{\Lambda_i} \in \mathcal{L}_\nu^\Gamma$;

5. Repeat steps 2, 3 and 4 until $\|\nabla J_{\nu,\Lambda_i}(\cdot)\| < \varepsilon$ where ε is a tolerance threshold. Then set $\Lambda^\circ = \Lambda_i$.

The computation of the search direction Δ_{Λ_i} is the most delicate part of the procedure. The corresponding linear equation reduces to

$$\frac{1}{\nu} \sum_{l=1}^{\nu} \int G Q_{\Lambda_i}^l G^* \Delta_{\Lambda_i} G Q_{\Lambda_i}^{\nu+1-l} G^* = \int G Q_{\Lambda_i}^\nu G^* - I \tag{2.61}$$

where $Q_\Lambda = W(I + G_1^* \Lambda G_1)^{-1} W^*$. By similar argumentations used in [27, Proposition 8.1], it is possible to prove that there exists a unique solution $\Delta_{\Lambda_i} \in \text{Range } \Gamma$ to (2.61). Accordingly, we can easily compute Δ_{Λ_i} in this way:

1. Compute
$$Y = \int G Q_{\Lambda_i}^\nu G^* - I; \tag{2.62}$$

2. Let $\{\Sigma_1 \dots \Sigma_M\}$ a basis for Range Γ (to see how to compute it, refer to equation (3.7) in Section 3.2) and for each Σ_k, $k = 1 \dots M$, compute
$$Y_k = \frac{1}{\nu} \sum_{l=1}^{\nu} \int G Q_{\Lambda_i}^l G^* \Sigma_k G Q_{\Lambda_i}^{\nu+1-l} G^*; \tag{2.63}$$

3. Find $\{\alpha_k\}$ such that $Y = \sum_k \alpha_k Y_k$. Then set $\Delta_{\Lambda_i} = \sum_k \alpha_k \Sigma_k$.

Concerning the evaluation of the integrals in (2.60), (2.62) and (2.63), a sensible and efficient method based on spectral factorization techniques may be employed. For further details, including the checking of condition (2.59), we refer to Section VI in [56].

Finally, it is possible to prove that:

1. $J_\nu(\cdot) \in \mathcal{C}^\infty(\mathcal{L}_\nu^\Gamma)$ is strongly convex on the sublevel set $\mathcal{K} = \{\Lambda \in \mathcal{L}_\nu^\Gamma \mid J_\nu(\Lambda) \leq J_\nu(\Lambda_0)\}$;

2. The *Hessian* is *Lipschitz* continuous in $\mathcal{K}$.

The proof follows the ones in [56, Section VII] and [26, Section VI-C] faithfully. These properties allow us to conclude that the proposed *Newton* algorithm globally converges, see Proposition 3.13. In particular the rate of convergence is quadratic during the last stage. In this way, the solution to Problem 2.1 may be efficiently computed.

2.7 Simulations results

In order to test the features of the family of solutions Φ_ν with $\nu \in \mathbb{N}_+$, we take into account the following comparison procedure:

1. Choose a zero mean wide sense stationary process $y = \{y_k; k \in \mathbb{Z}\}$ with spectral density $\Omega \in \mathbb{S}_+^m(\mathbb{T})$;
2. Design a filters bank $G(z)$ as in (1.3);
3. Set $\Psi = \int \Omega$ (Ψ is constant and equal to the zeroth moment of Ω)
4. Compute $\Sigma = \int G\Omega G^*$;
5. Solve Problem 2.1 (with $\mathcal{S}_\nu$) by means of the proposed algorithm with the chosen Ψ and $\Sigma^{-\frac{1}{2}}G(z)$ as filters bank.

In the above comparison procedure we assume to know Σ and $\int \Omega$. In this way, we avoid the approximation errors introduced by the estimation of Σ and $\int \Omega$ from the finite-length data $y(1) \dots y(N)$. Concerning the design of the filter, its role consists in providing the interpolation conditions for the solution to the spectrum approximation problem. More specifically, a higher resolution can be attained by selecting poles in the proximity of the unit circle with arguments in the range of frequency of interest, [14].

2.7.1 Scalar case

We start by taking into account Example described in [56, Section VIII-B] (the unique difference is that we assume to know Σ and $\int \Omega$). Consider the following ARMA process:

$$\begin{aligned} y(t) &= 0.5y(t-1) - 0.42y(t-2) + 0.602y(t-3) \\ &- 0.0425y(t-4) + 0.1192y(t-5) \\ &+ e(t) + 1.1e(t-1) + 0.08e(t-2) - 0.15e(t-3) \end{aligned} \tag{2.64}$$

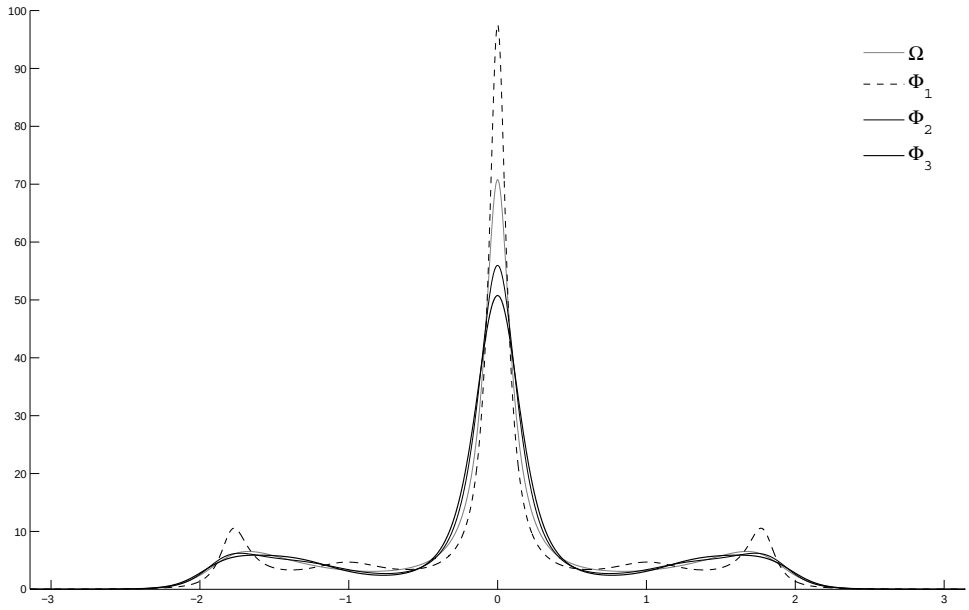


Figure 2.1: Approximation of an ARMA $(6, 4)$ spectral density.

where e is a zero-mean *Gaussian* white noise with unit variance. In Figure 2.1, the spectral density $\Omega \in \mathbb{S}^1_+(\mathbb{T})$ of the ARMA process is depicted (gray line). $G(z)$ is structured according to the covariance extension setting (1.7) with 6 covariance lags (i.e. $n = 6$). In Figure 2.1 the different solutions obtained by fixing $\nu = 1$, dashed line, $\nu = 2$, solid line, and $\nu = 3$, thick line, are shown. The solution obtained by minimizing the multivariate *Itakura-Saito* distance ($\nu = 1$) is characterized by peaks which are taller than these in Ω. In fact, this solution seems the most adequate for detecting spectral lines, see example of Section VII-A in [26]. On the contrary, the peaks are reduced by increasing ν. Note that, the solutions with $\nu = 2$ and $\nu = 3$ are closer to Ω than the one with $\nu = 1$.

As second example we consider the scalar *bandpass* random process with spectral density Ω depicted in Figure 2.2 (gray curve). The *cutoff* frequencies are $\vartheta_1 = 0.89$ and $\vartheta_2 = 2.46$. Moreover, $\Omega(\mathrm{e}^{j\vartheta}) \geq 2 \cdot 10^{-3}$ in the *stopband*, accordingly $\Omega \in \mathbb{S}^1_+(\mathbb{T})$. Matrix B is a column of ones. Matrix A is chosen as a block-diagonal matrix with one eigenvalue equal to zero and eight eigenvalues equispaced on the circle of radius 0.8

$$\pm 0.8,\ 0.8\mathrm{e}^{\pm j\frac{\pi}{4}},\ 0.8\mathrm{e}^{\pm j\frac{\pi}{2}},\ 0.8\mathrm{e}^{\pm j\frac{3}{4}\pi}. \tag{2.65}$$

Here, $\Psi = \int \Omega \simeq 1.5284$. Figure 2.2 also shows the obtained solutions. The one with $\nu = 1$ turns out inadequate. The solutions with $\nu = 2$ and $\nu = 3$ are, instead, similar and closer to Ω.

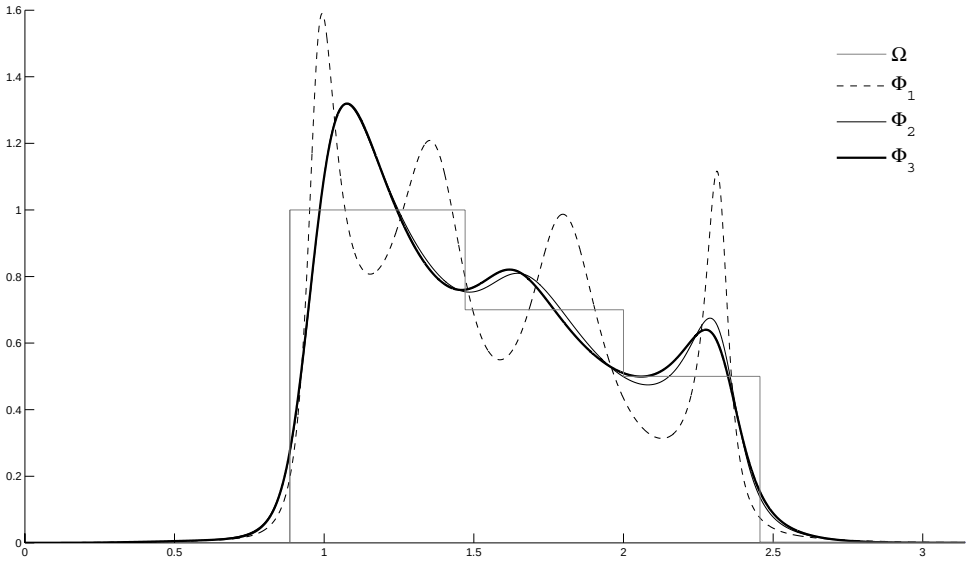


Figure 2.2: Approximation of the spectral density of a scalar *bandpass* random process.

2.7.2 Multivariate case

We consider a bivariate *bandpass* random process with spectral density Ω plotted in Figure 2.3 (gray curve). Here, the *cutoff* frequencies are $\vartheta_1 = 0.42$ and $\vartheta_2 = 1.94$, and $\Omega(\mathrm{e}^{j\vartheta}) \geq 2 \cdot 10^{-3} I$ in the whole range of frequencies. The constant *prior* is

$$\Psi = \int \Omega \simeq \begin{pmatrix} 0.9313 & 0.3314 \\ 0.3314 & 0.5128 \end{pmatrix}. \tag{2.66}$$

The matrix A of the filters bank has one eigenvalue equal to zero, two eigenvalues in ± 0.8 and three pairs of complex eigenvalues closer to the passband $0.8\mathrm{e}^{\pm j0.4}$, $0.8\mathrm{e}^{\pm j1.2}$, $0.8\mathrm{e}^{\pm j2}$. The solutions for $\nu = 1$ (dashed line) $\nu = 2$ (solid line) and $\nu = 3$ (thick line) are shown in Figure 2.3. It is apparent that the solution for $\nu = 2$ and $\nu = 3$ are the most appropriate.

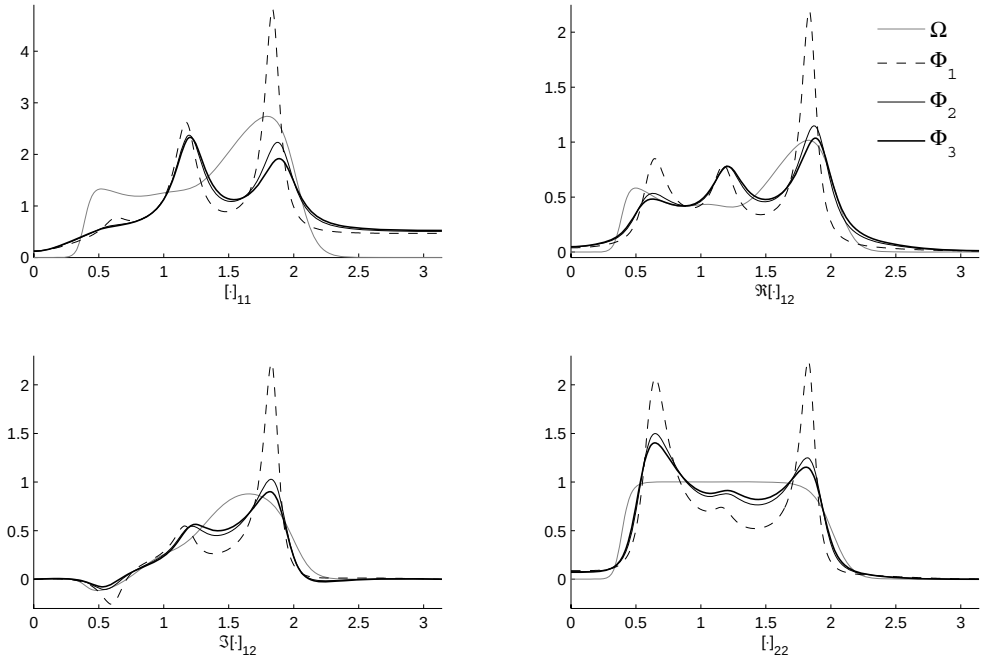


Figure 2.3: Approximation of the spectral density of a bivariate *bandpass* random process.

Chapter 3

Structured covariance estimation problem

3.1 Introduction

In THREE-like estimation procedures, a known filters bank $G(z) = (zI - A)^{-1}B$ is driven by an unknown stochastic process $y = \{y_k; \quad k \in \mathbb{Z}\}$. Let $\Sigma = E[x_k x_k^T]$ be the covariance matrix of the output process $x = \{x_k; \quad k \in \mathbb{Z}\}$. Here, we firstly have to compute an estimate $\hat{\Sigma}$ of the covariance Σ from a finite-length collection of sample data $y(1) \dots y(N)$ of the process y. This estimate must be positive definite and such that Problem 1.2 is feasible, i.e. there exists $\Phi \in \mathbb{S}_+^m(\mathbb{T})$ satisfying the constraint in (2.1). In this way, by replacing G with $\overline{G} = \hat{\Sigma}^{-\frac{1}{2}} G$ and (A, B) with $(\overline{A} = \hat{\Sigma}^{-\frac{1}{2}} A \hat{\Sigma}^{\frac{1}{2}}, \overline{B} = \hat{\Sigma}^{-\frac{1}{2}} B)$, we obtain Problem 2.1 (previously analyzed) which provides an estimate of the spectral density of y. This Chapter is devoted to the computation of such a $\hat{\Sigma}$.

To analyze the features of this task, we take into account the covariance extension problem introduced in Section 1.2.1. In this case, the covariance matrix $\Sigma = E[x_k x_k^T]$ of the output x has the form of a symmetric *Toeplitz* matrix having the first n covariance lags of y on the first row:

$$\Sigma := \begin{bmatrix} r_0 & r_1 & \dots & r_{n-1} \\ r_1 & r_0 & \ddots & r_{n-2} \\ \vdots & \ddots & \ddots & \ddots \\ r_{n-1} & \ddots & r_1 & r_0 \end{bmatrix}, \qquad r_h := E[y_{k+h} y_k]. \tag{3.1}$$

It is natural to impose that the estimate $\hat{\Sigma}$ be positive definite and have

Toeplitz structure. On the one hand, one can consider the estimate

$$\hat{\Sigma} = \begin{bmatrix} \hat{r}_0 & \hat{r}_1 & \dots & \hat{r}_{n-1} \\ \hat{r}_1 & \hat{r}_0 & \ddots & \hat{r}_{n-2} \\ \vdots & \ddots & \ddots & \ddots \\ \hat{r}_{n-1} & \ddots & \hat{r}_1 & \hat{r}_0 \end{bmatrix}, \quad \hat{r}_h = \frac{1}{N-h} \sum_{k=1}^{N-h} y(k+h)y(k) \tag{3.2}$$

which is a Toeplitz matrix. This estimate, however, is *not* guaranteed to be positive definite. On the other hand, one can compute the sample covariance $\hat{\Sigma}_C := \frac{1}{N} \sum_{k=1}^{N} x(k)x(k)^T$ of the output process x where $x(1) \dots x(N)$ is obtained by filtering $y(1) \dots y(N)$ through $G(z)$. The latter is typically, by construction, positive semi-definite but is *not* guaranteed to be Toeplitz. Notice, in passing, that the orthogonal projection of this estimate onto the linear space of Toeplitz matrices is no longer guaranteed to be positive definite. This problem, yet important, is very special due to the FIR structure of $G(z)$ in (1.7). In this case, it is well-known that the problem can be solved by computing, from $y(1) \dots y(N)$, the estimates $\hat{r}_h$ of the r_h in (3.1), with the *biased correlogram spectral estimator* [61].
The estimation of positive definite Toeplitz matrices is just an instance of the structured covariance estimation problem which will be introduced in the following section. Here, we consider the general case wherein the process y is $\mathbb{C}^m$-valued, $A \in \mathbb{C}^{n \times n}$, $B \in \mathbb{C}^{n \times m}$, and Σ is a positive definite Hermitian matrix. In Section 3.3 we introduce an optimization approach for estimating Σ based on the *infomation divergence index*. In Section 3.4 we extend the previous approach to the Beta matrix divergence index. Finally, in Section 3.5 we present a different method for computing $\hat{\Sigma}$ which can be viewed as a generalization of the *Blackman-Tukey* approach, [7].

3.2 Structured covariance estimation problem

Consider a transfer function

$$G(z) = (zI - A)^{-1}B, \quad A \in \mathbb{C}^{n \times n}, \quad B \in \mathbb{C}^{n \times m}, \quad n > m, \tag{3.3}$$

where A has all its eigenvalues in the open unit disk, B has full column rank, and (A, B) is a reachable pair. Suppose $G(z)$ models a bank of filters fed by a zero mean, wide sense stationary, purely nondeterministic, $\mathbb{C}^m$-valued process y with spectral density Ω which is coercive. Let x be the n-dimensional stationary output process

$$x_{k+1} = Ax_k + By_k, \quad k \in \mathbb{Z}. \tag{3.4}$$

We denote by Σ the covariance of x_k. Notice that $\Sigma > 0$ since A is a stable matrix, (A, B) is reachable and Φ is coercive. We denote by $\mathcal{V}(\mathbb{S}_+^m(\mathbb{T}))$ the linear space generated by $\mathbb{S}_+^m(\mathbb{T})$[1]. Consider now the linear operator

$$\begin{aligned} \Gamma : \mathcal{V}(\mathbb{S}_+^m(\mathbb{T})) &\to \mathcal{H}_n \\ \Phi &\mapsto \int G\Phi G^*, \end{aligned} \tag{3.5}$$

where integration takes place on $\mathbb{T}$ with respect to normalized Lebesgue measure $\mathrm{d}\vartheta/2\pi$ as in Chapter 2. Note that Σ belongs to the linear space

$$\begin{aligned} \text{Range } \Gamma : = \{P \in \mathcal{H}_n |\ & \exists \Phi \in \mathcal{V}(\mathbb{S}_+^m(\mathbb{T})) \\ & \text{such that } \int G\Phi G^* = P\}. \end{aligned} \tag{3.6}$$

Assume that a collection of sample data $y(1) \dots y(N)$ of the stochastic process y is available and let $\hat{\Sigma}$ be an estimate of Σ computed starting from $y(1) \dots y(N)$. It turns out that Problem 1.2 is feasible if and only if $\hat{\Sigma} \in$ Range $\Gamma \cap \mathcal{H}_{n,+}$ [36],[27]. Hence, we have to face the following structured covariance estimation problem.

Problem 3.1. *Compute an estimate* $\hat{\Sigma}$ *of* Σ *from* $y(1) \dots y(N)$ *such that* $\hat{\Sigma} \in$ Range $\Gamma \cap \mathcal{H}_{n,+}$.

First of all, we introduce the characterizing properties of the above vector space. In [33], [36] (see also [57]), it was shown that $P \in \mathcal{H}_n$ belongs to Range Γ if and only if there exists $H \in \mathbb{C}^{m \times n}$ such that

$$P - APA^* = BH + H^*B^* \tag{3.7}$$

or equivalently if and only if the following rank condition holds

$$\text{rank} \begin{bmatrix} P - APA^* & B \\ B^* & 0 \end{bmatrix} = 2m. \tag{3.8}$$

The dimension of Range Γ may now be established along the lines of [38, Lemma 4], which deals with the scalar case, and [34, Page 137], which treats the multivariate *real* case.

Proposition 3.2. *The linear space* Range Γ *has* real *dimension* $m(2n - m)$.

[1]Here, $\mathbb{S}_+^m(\mathbb{T})$ denotes the family of bounded and coercive $\mathbb{C}^{m \times m}$-valued spectral density functions on $\mathbb{T}$.

Proof. The dimension of the linear space Range Γ is invariant under a change of basis in the state space of G. Since B is assumed to be full column-rank, we can then assume that $B := \begin{bmatrix} I_m \\ 0 \end{bmatrix}$. From (3.7), we get that dim Range Γ equals the real dimension of the linear space of matrices that can be written in the form $BH + H^*B^*$, or, equivalently (given the structure of B), in the form $\begin{bmatrix} Q & H_2 \\ H_2^* & 0 \end{bmatrix}$, with $Q \in \mathcal{H}_m$ and $H_2 \in \mathbb{C}^{m\times(n-m)}$. Such a dimension is $m(2n-m)$. ∎

In [27, Proposition 2.1], it was shown that, after normalizing $P > 0$ to the identity matrix, condition (3.8) could be replaced by a geometric condition. We show next that the latter condition is equivalent to (3.8) for any Hermitian P.

Proposition 3.3. *Given $P \in \mathcal{H}_n$, a necessary and sufficient condition for $P \in$ Range Γ is that the following condition holds*

$$(I - \Pi_B)(P - APA^*)(I - \Pi_B) = 0, \tag{3.9}$$

where we denote by $\Pi_B := B(B^*B)^{-1}B^*$ the orthogonal projection onto Range B.

Proof. Necessity: We know that there exists $H \in \mathbb{C}^{m\times n}$ such that

$$P - APA^* = BH + H^*B^*. \tag{3.10}$$

Pre and post-multiplying this relation by $I - \Pi_B$, we obtain

$$\begin{aligned}(I - \Pi_B)(P - APA^*)(I - \Pi_B) \\ &= (I - \Pi_B)(BH + H^*B^*)(I - \Pi_B) \\ &= (I - \Pi_B)BH(I - \Pi_B) \\ &\quad +[(I - \Pi_B)BH(I - \Pi_B)]^* = 0.\end{aligned}$$

Sufficiency: We exploit condition (3.8). Let us first consider the matrix

$$T := \begin{bmatrix} C & B \end{bmatrix} \in \mathbb{C}^{n\times n}, \tag{3.11}$$

where $C \in \mathbb{C}^{n\times(n-m)}$ has full column rank and is such that (Range C)$\perp$(Range B), so that T is invertible. Moreover, C can be expressed as $C = (I - \Pi_B)V$, where $V \in \mathbb{C}^{n\times(n-m)}$ has full column rank. In view of (3.9), we have

$$C^*(P - APA^*)C = 0. \tag{3.12}$$

We now consider the matrices

$$\begin{bmatrix} P - APA^* & B \\ B^* & 0 \end{bmatrix} \tag{3.13}$$

and

$$\begin{aligned} \Delta : &= \begin{bmatrix} T^* & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} P - APA^* & B \\ B^* & 0 \end{bmatrix} \begin{bmatrix} T & 0 \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} T^*(P - APA^*)T & T^*B \\ B^*T & 0 \end{bmatrix}. \end{aligned} \tag{3.14}$$

By (3.11) and (3.12), we get

$$\begin{aligned} \Delta &= \left[\begin{array}{cc|c} C^*(P - APA^*)C & C^*(P - APA^*)B & C^*B \\ B^*(P - APA^*)C & B^*(P - APA^*)B & B^*B \\ \hline B^*C & B^*B & 0 \end{array}\right] \\ &= \left[\begin{array}{cc|c} 0 & \star & 0 \\ \star & \star & B^*B \\ \hline 0 & B^*B & 0 \end{array}\right] \end{aligned} \tag{3.15}$$

where B^*B is an invertible matrix, since B has full column rank. Recalling that the rank of a matrix is invariant under multiplication by an invertible matrix, we conclude that

$$\text{rank} \begin{bmatrix} P - APA^* & B \\ B^* & 0 \end{bmatrix} = \text{rank}\Delta = 2m \tag{3.16}$$

namely, by (3.8), $P \in \text{Range } \Gamma$. ■

Our way to attack Problem 3.1 consists in considering the following situation:

- The filter $G(z)$ is fed by the m-dimensional data $y(1) \dots y(N)$ and we collect the n-dimensional output data $x(1) \dots x(N)$.
- We compute the sample covariance $\hat{\Sigma}_C$ of Σ in the usual way

$$\hat{\Sigma}_C := \frac{1}{N} \sum_{k=1}^{N} x(k)x(k)^*. \tag{3.17}$$

Notice that $\hat{\Sigma}_C \in \mathcal{H}_n$ and $\hat{\Sigma}_C \geq 0$. Moreover, for $N \geq n$, $\hat{\Sigma}_C$ is positive definite with probability 1. In general, $\hat{\Sigma}_C$ does not belong to Range Γ. Indeed, $m < n$ and Range Γ has only dimension $m(2n-m) < n^2$ (Proposition 3.2). Moreover $\hat{\Sigma}_C \to \Sigma$ as $N \to \infty$. Thus, the estimate $\hat{\Sigma} \in \text{Range } \Gamma \cap \mathcal{H}_{n,+}$ should be, in a suitable sense, as close as possible to $\hat{\Sigma}_C$.

Projection method

In [27, Section 8] a simple-minded approach has been presented. It consists in projecting $\hat{\Sigma}_C$ given by (3.17) onto Range Γ thereby obtaining a new Hermitian matrix $\hat{\Sigma}_\Gamma$. For a large number N of samples, we expect $\hat{\Sigma}_\Gamma$ to be close to $\hat{\Sigma}_C$ since the true state covariance Σ does belong to Range Γ. The projection $\hat{\Sigma}_\Gamma$, however, might turn out to be indefinite and this is particularly likely when N is not large. In this case, $\hat{\Sigma}_\Gamma$ may be further adjusted by adding to it a matrix of the form $\varepsilon\Sigma_+$ with $\Sigma_+ \in$ Range Γ, $\Sigma_+ > 0$ and $\varepsilon > 0$ so large that

$$\hat{\Sigma}_{PJ} := \hat{\Sigma}_\Gamma + \varepsilon\Sigma_+ > 0. \tag{3.18}$$

In this way, a positive definite matrix belonging to Range Γ is obtained. Notice that a positive definite matrix $\Sigma_+ \in$ Range Γ indeed exists and can be easily computed as follows. Set $H_+ := \frac{1}{2}B^*$ and consider the equation

$$\Sigma_+ - A\Sigma_+A^* = BH_+ + H_+^*B^* = BB^*. \tag{3.19}$$

Since (A, B) is reachable and A is a stable matrix, we have that (3.19) admits a unique solution Σ_+ and such a solution is indeed positive definite. In view of (3.7), Σ_+ also belongs to Range Γ.

3.3 An optimization approach to estimating $\sum$

In this section, we present a new systematic procedure to find $\hat{\Sigma} \in$ Range $\Gamma \cap \mathcal{H}_{n,+}$ as close as possible to $\hat{\Sigma}_C$, [29]. Recall that a most fundamental (pseudo)-distance in mathematical statistics is the *information divergence* (Kullback-Leibler index, relative entropy), [23]. For two Gaussian distributions $p_P, p_{\hat{\Sigma}}$ on $\mathbb{R}^n$ with zero mean and covariance matrices $P > 0$ and $\hat{\Sigma} > 0$, respectively, it is given by

$$\mathcal{D}_I(p_P \| p_{\hat{\Sigma}}) := \frac{1}{2}\left[\log\det(P^{-1}\hat{\Sigma}) + \mathrm{tr}(\hat{\Sigma}^{-1}P) - n\right]. \tag{3.20}$$

Notice that the right-hand side of (3.20) provides a natural pseudo-distance, denoted henceforth by $\mathcal{D}_I(P\|\hat{\Sigma})$, on the space $\mathcal{H}_{n,+}$. This fact leads us to consider the following problem.

Problem 3.4. *Given $\hat{\Sigma}_C \in \mathcal{H}_{n,+}$ and $G(z)$ with the previous properties, solve*

$$\text{minimize } \mathcal{D}_I(P\|\hat{\Sigma}_C) \text{ over } P \in (\mathcal{H}_{n,+} \cap \text{Range } \Gamma). \tag{3.21}$$

The solution to Problem 3.4 provides the required estimate of Σ.

Remark 3.5. In [34], the Umegaki-von Neumann relative entropy [53] was proposed instead, restricting the search to covariances having the same trace as the sample covariance $\hat{\Sigma}_C$. In alternative, it was there suggested that one could use as distance the one induced by a matrix norm. Our choice is supported by the following considerations. First, as observed in [12, p.963], $\mathcal{D}_I(\cdot\|\cdot)$ "really comes from maximum-likelihood considerations and thus should, in some sense, give us a reasonable answer, even if the process is not Gaussian and the vector samples are not independent". Second, with this distance, the solution turns out to have a simple form and the variational analysis can be carried through to the very end, see below. Finally, simulation shows that THREE-like procedures initialised with the found estimate work extremely well.

In what follows, we assume that $\hat{\Sigma}_C > 0$ and use the compact notation $\Pi_B^\perp := I - \Pi_B$. In view of Proposition 3.3, Problem 3.4 finds $P \in \mathcal{H}_{n,+}$ minimizing $\mathcal{D}_{BG}(P\|\hat{\Sigma}_C) := 2\mathcal{D}_I(P\|\hat{\Sigma}_C)$ subject to the *linear* constraint

$$\Pi_B^\perp(P - APA^*)\Pi_B^\perp = 0. \tag{3.22}$$

Here, $\mathcal{D}_{BG}$ is the Burg matrix divergence, [25]. Thus, our problem resembles a most standard maximum entropy (or, equivalently, minimum relative entropy) problem [43], [23]. As a first step, we introduce the Lagrangian function

$$\begin{aligned} L_{BG}(P,\Delta) &= \mathcal{D}_{BG}(P\|\hat{\Sigma}_C) - \log\det\hat{\Sigma}_C + n + \mathrm{tr}\left[\Delta\Pi_B^\perp(P - APA^*)\Pi_B^\perp\right] \\ &= -\log\det P + \mathrm{tr}(\hat{\Sigma}_C^{-1}P) + \mathrm{tr}\left[\Delta\Pi_B^\perp(P - APA^*)\Pi_B^\perp\right] \end{aligned} \tag{3.23}$$

where we exploited the fact that the terms $\log\det\hat{\Sigma}_C$ and n play no role in the optimization problem. We consider the unconstrained minimization problem

$$\min_P\{L_{BG}(P,\Delta) \mid P \in \mathcal{H}_{n,+}\}. \tag{3.24}$$

Since $L_{BG}(\cdot,\Delta)$ is strictly convex over $\mathcal{H}_{n,+}$, its unique minimum point is given annihilating its first variation in each direction $\delta P \in \mathcal{H}_n$:

$$\begin{aligned} \delta L_{BG}(P,\Delta;\delta P) &= -\mathrm{tr}\left[P^{-1}\delta P\right] + \mathrm{tr}\left[\hat{\Sigma}_C^{-1}\delta P\right] \\ &\quad +\mathrm{tr}\left[\Delta\Pi_B^\perp(\delta P - A\delta PA^*)\Pi_B^\perp\right] \\ &= \mathrm{tr}[(-P^{-1} + \hat{\Sigma}_C^{-1} + \Pi_B^\perp\Delta\Pi_B^\perp \\ &\quad -A^*\Pi_B^\perp\Delta\Pi_B^\perp A)\delta P]. \end{aligned} \tag{3.25}$$

Thus (3.25) is zero for each $\delta P \in \mathcal{H}_n$ if and only if

$$P^{-1} = \hat{\Sigma}_C^{-1} + \Pi_B^{\perp}\Delta\Pi_B^{\perp} - A^*\Pi_B^{\perp}\Delta\Pi_B^{\perp}A. \tag{3.26}$$

It is then natural to restrict our attention to multiplier matrices belonging to the following set

$$\mathcal{L}_{BG} = \{\Delta \in \mathcal{H}_n \mid \hat{\Sigma}_C^{-1} + V_\Delta > 0\}, \tag{3.27}$$

where

$$V_\Delta := \Pi_B^{\perp}\Delta\Pi_B^{\perp} - A^*\Pi_B^{\perp}\Delta\Pi_B^{\perp}A. \tag{3.28}$$

Thus, given $\Delta \in \mathcal{L}_{BG}$, we get that the unique minimum point of the Lagrange functional has the form

$$P_{BG}(\Delta) := \left(\hat{\Sigma}_C^{-1} + V_\Delta\right)^{-1}. \tag{3.29}$$

It is quite interesting to notice that V_Δ gives another characterization of Range Γ as stated by the following proposition.

Proposition 3.6. *Let $\Delta \in \mathcal{H}_n$ and V_Δ be defined by (3.28). Then $V_\Delta \in$ Range $\Gamma^{\perp}$.*

Proof. Let be $P \in$ Range Γ, $\Delta \in \mathcal{H}_n$ and consider

$$\begin{aligned} \langle V_\Delta, P\rangle &= \operatorname{tr}(V_\Delta P) = \operatorname{tr}\left[(\Pi_B^{\perp}\Delta\Pi_B^{\perp} - A^*\Pi_B^{\perp}\Delta\Pi_B^{\perp}A)P\right] \\ &= \operatorname{tr}\left[(\Pi_B^{\perp}P\Pi_B^{\perp} - \Pi_B^{\perp}APA^*\Pi_B^{\perp})\Delta\right] \\ &= \operatorname{tr}\left[\Pi_B^{\perp}(P - APA^*)\Pi_B^{\perp}\Delta\right] = 0 \end{aligned} \tag{3.30}$$

where we employed condition (3.9). ∎

To sum up, we showed that $P_{BG}(\Delta)$ is the unique minimum point of $L_{BG}(\cdot, \Delta)$, namely

$$L_{BG}(P_{BG}(\Delta), \Delta) \leq L_{BG}(P, \Delta),\ \forall\ P \in \mathcal{H}_{n,+}. \tag{3.31}$$

Hence, if we produce $\Delta^\circ \in \mathcal{L}_{BG}$ such that $P_{BG}(\Delta^\circ)$ satisfies constraint (3.22), then inequality (3.31) implies

$$\mathcal{D}_{BG}(P_{BG}(\Delta^\circ)\|\hat{\Sigma}_C) \leq \mathcal{D}_{BG}(P\|\hat{\Sigma}_C),\ \forall\ P \in \text{Range}\ \Gamma \cap \mathcal{H}_{n,+}. \tag{3.32}$$

Thus, such a $\hat{\Sigma}_{ME} := P_{BG}(\Delta^\circ)$ is the unique solution to Problem 3.4. Here, we denote $P_{BL}(\Lambda^\circ)$ as $\hat{\Sigma}_{ME}$ (instead of $\hat{\Sigma}_{BG}$) in order to maintain the same notation employed in [29]. The abbreviation ME is employed to remark that Problem 3.4 resembles a most standard maximum entropy problem.

It remains to show the existence of such a Δ°. This is accomplished via duality theory. The dual problem consists in maximizing the following functional over $\mathcal{L}_{BG}$

$$\begin{aligned}
\inf_{P\in\mathcal{H}_{n,+}} L_{BG}(P,\Delta) &= L_{BG}(P_{BG}(\Delta),\Delta) = \log\det P_{BG}(\Delta)^{-1} + \mathrm{tr}(\hat{\Sigma}_C^{-1}P_{BG}(\Delta)) \\
&\quad +\mathrm{tr}\left[\Delta\Pi_B^\perp(P_{BG}(\Delta) - AP_{BG}(\Delta)A^*)\Pi_B^\perp\right] \\
&= \mathrm{tr}\Big[\log P_{BG}(\Delta)^{-1} + \hat{\Sigma}_C^{-1}P_{BG}(\Delta) \\
&\quad +\Delta\Pi_B^\perp(P_{BG}(\Delta) - AP_{BG}(\Delta)A^*)\Pi_B^\perp\Big] \\
&= \mathrm{tr}\Big[\log P_{BG}(\Delta)^{-1} + \Big(\hat{\Sigma}_C^{-1} + \Pi_B^\perp\Delta\Pi_B^\perp \\
&\quad -A^*\Pi_B^\perp\Delta\Pi_B^\perp A\Big)P_{BG}(\Delta)\Big] \\
&= \mathrm{tr}\left[\log P_{BG}(\Delta)^{-1} + P_{BG}(\Delta)^{-1}P_{BG}(\Delta)\right] \\
&= \mathrm{tr}\left[\log P_{BG}(\Delta)^{-1} + I\right] \\
&= \mathrm{tr}\left[\log\left(\hat{\Sigma}_C^{-1} + V_\Delta\right) + I\right]. \qquad (3.33)
\end{aligned}$$

Thus, it is equivalent to minimize the following function, hereafter referred to as dual functional:

$$J_{BG}(\Delta) := -\mathrm{tr}\log\left(\hat{\Sigma}_C^{-1} + V_\Delta\right). \qquad (3.34)$$

To perform this minimization it is convenient to restrict our attention to a subset of $\mathcal{L}_{BG}$ defined as follows. Consider the map

$$\begin{aligned}
\varphi: \quad & \mathcal{H}_n \to \mathcal{H}_n \\
& \Delta \mapsto \Pi_B^\perp\Delta\Pi_B^\perp. \qquad (3.35)
\end{aligned}$$

Such a map is self-adjoint because

$$\langle\varphi(\Delta),\Delta\rangle = \mathrm{tr}\left(\Pi_B^\perp\Delta\Pi_B^\perp\Delta\right) = \mathrm{tr}\left(\Delta\Pi_B^\perp\Delta\Pi_B^\perp\right) = \langle\Delta,\varphi(\Delta)\rangle.$$

Thus, $\ker\varphi = [\mathrm{Range}\,\varphi]^\perp$. Suppose now that J_{BG} takes the minimum value in $\Delta^\circ \in \mathcal{L}_{BG}$ and let $M \in [\mathrm{Range}\,\varphi]^\perp$. It is easy to see that

$$J_{BG}(\Delta^\circ + M) = J_{BG}(\Delta^\circ) \qquad (3.36)$$

so that the search for the solution of the dual problem can be restricted to the set

$$\mathcal{L}_{BG}^\varphi := \mathcal{L}_{BG} \cap \mathrm{Range}\,\varphi. \qquad (3.37)$$

Lemma 3.7. *Consider $J_{BG} : \mathcal{L}_{BG} \to \mathbb{R}$. Then:*

1. *J_{BG} is strictly convex on $\mathcal{L}_{BG}^{\varphi}$.*

2. *$J_{BG} \in \mathcal{C}^{\infty}(\mathcal{L}_{BG})$.*

Proof. 1. First of all, observe that $\mathcal{L}_{BG}$ is an open, convex subset of $\mathcal{H}_n$. Moreover, since J_{BG} is the negative of $\inf_P L_{BG}(P,\Delta)$, it is convex. For $\delta\Delta_1 \in \mathcal{H}_n$, we compute its directional derivative

$$\begin{aligned} \delta J_{BG}(\Delta;\delta\Delta_1) &= -\mathrm{tr}\left[P_{BG}(\Delta)(\Pi_B^{\perp}\delta\Delta_1\Pi_B^{\perp} - A^*\Pi_B^{\perp}\delta\Delta_1\Pi_B^{\perp}A)\right] \\ &= -\mathrm{tr}\left[P_{BG}(\Delta)V_{\delta\Delta_1}\right]. \end{aligned} \tag{3.38}$$

The second variation, in directions $\delta\Delta_1, \delta\Delta_2 \in \mathcal{H}_n$, is given by

$$\begin{aligned} \delta^2 J_{BG}(\Delta;\delta\Delta_1,\delta\Delta_2) &= \mathrm{tr}\left[P_{BG}(\Delta)(\Pi_B^{\perp}\delta\Delta_2\Pi_B^{\perp} - A^*\Pi_B^{\perp}\delta\Delta_2\Pi_B^{\perp}A)\right. \\ &\quad \left.\times P_{BG}(\Delta)(\Pi_B^{\perp}\delta\Delta_1\Pi_B^{\perp} - A^*\Pi_B^{\perp}\delta\Delta_1\Pi_B^{\perp}A)\right] \\ &= \mathrm{tr}\left[P_{BG}(\Delta)V_{\delta\Delta_2}P_{BG}(\Delta)V_{\delta\Delta_1}\right]. \end{aligned} \tag{3.39}$$

Consider now

$$\begin{aligned} \delta^2 J_{BG}(\Delta;\delta\Delta,\delta\Delta) &= \mathrm{tr}\left[P_{BG}(\Delta)V_{\delta\Delta}P_{BG}(\Delta)V_{\delta\Delta}\right] \\ &= \mathrm{tr}\left[P_{BG}(\Delta)^{\frac{1}{2}}V_{\delta\Delta}P_{BG}(\Delta)V_{\delta\Delta}P_{BG}(\Delta)^{\frac{1}{2}}\right] \end{aligned} \tag{3.40}$$

which, as expected, is a nonnegative quantity since $P_{BG}(\Delta)$ is positive definite. Suppose now that $\delta\Delta \in \mathrm{Range}\,\varphi$. The equation

$$\Pi_B^{\perp}\delta\Delta\Pi_B^{\perp} = A^*\Pi_B^{\perp}\delta\Delta\Pi_B^{\perp}A + V_{\delta\Delta} \tag{3.41}$$

is the Lyapunov equation associated to the (A^*, V_{δ_Δ}) pair where we regard as the unknown $\Pi_B^{\perp}\delta\Delta\Pi_B^{\perp}$. It follows that $\Pi_B^{\perp}\delta\Delta\Pi_B^{\perp}$ can be expressed as

$$\Pi_B^{\perp}\delta\Delta\Pi_B^{\perp} = \sum_{t=0}^{\infty}(A^*)^t V_{\delta\Delta}A^t. \tag{3.42}$$

Since $\delta\Delta \in \mathrm{Range}\,\varphi = [\ker\varphi]^{\perp}$, from $\Pi_B^{\perp}\delta\Delta\Pi_B^{\perp} = 0$ it follows that $\delta\Delta = 0$. Thus, taking (3.42) into account, we have that $V_{\delta\Delta} = 0$ implies $\delta\Delta = 0$. Accordingly, we have that $\delta^2 J_{BG}(\Delta;\delta\Delta,\delta\Delta)$ is strictly positive for any non zero $\delta\Delta \in \mathrm{Range}\,\varphi$, and consequently, J_{BG} is strictly convex on $\mathcal{L}_{BG}^{\varphi}$.
2. Notice that the first and the second variation of J_{BG} exist and are continuous on $\mathcal{L}_{BG}$. The same applies to the third variation in directions $\delta\Delta_1, \delta\Delta_2, \delta\Delta_3 \in \mathcal{H}_n$:

$$\begin{aligned} \delta^3 J_{BG}(\Delta;\delta\Delta_1,\delta\Delta_2,\delta\Delta_3) &= \mathrm{tr}[\delta P_{BG}(\Delta;\delta\Delta_3)V_{\delta\Delta_2}P_{BG}(\Delta)V_{\delta\Delta_1} \\ &\quad + P_{BG}(\Delta)V_{\delta\Delta_2}\delta P_{BG}(\Delta);\delta\Delta_3)V_{\delta\Delta_1}] \\ &= -\mathrm{tr}[P_{BG}(\Delta)V_{\delta\Delta_3}P_{BG}(\Delta)V_{\delta\Delta_2}P_{BG}(\Delta)V_{\delta\Delta_1} \\ &\quad + P_{BG}(\Delta)V_{\delta\Delta_2}P_{BG}(\Delta)V_{\delta\Delta_3}P_{BG}(\Delta)V_{\delta\Delta_1}]. \end{aligned}$$

Similarly, since $\delta P_{BG}(\Delta;\delta\Delta) = -P_{BG}(\Delta)V_{\delta\Delta}P_{BG}(\Delta)$, it can be shown that J_{BG} has continuous directional derivatives of any order in $\mathcal{L}_{BG}$. Thus $J_{BG} \in \mathcal{C}^k(\mathcal{L}_{BG})$ for any $k \geq 0$. ■

Corollary 3.8. *The dual problem*

$$\text{Find } \Delta \in \mathcal{L}_{BG}^{\varphi} \text{ minimizing } J(\Delta) \tag{3.43}$$

is a convex optimization problem which admits at most one solution.

We now tackle the existence issue for the dual problem. To this aim, we prove that $\mathcal{L}_{BG}^{\varphi}$ is bounded. In doing that, we need to show a preliminary technical result.

Lemma 3.9. *Let V_Δ be given by (3.28) and let φ be given by (3.35). Given a sequence $\{\Delta_k\}_{k\geq 0}$ with $\Delta_k \in \operatorname{Range}\varphi$, if $\|\Delta_k\| \to +\infty$ then $\|V_{\Delta_k}\| \to +\infty$.*

Proof. The proof is divided into two steps.
Step 1: Consider a sequence $\{\Delta_k\}_{k\geq 0}$, $\Delta_k \in \operatorname{Range}\varphi$ such that $\|\Delta_k\| \to +\infty$ as $k \to +\infty$. Since $\operatorname{Range}\varphi = [\ker\varphi]^\perp$, the minimum singular value ρ of the map φ restricted to $\operatorname{Range}\varphi$ is strictly positive. Accordingly,

$$\|\Pi_B^\perp \Delta_k \Pi_B^\perp\| \geq \rho\|\Delta_k\| \to +\infty. \tag{3.44}$$

Step 2: It now remains to show that, if $\|\Pi_B^\perp \Delta_k \Pi_B^\perp\| \to +\infty$ as k tends to infinity, then also $\|V_{\Delta_k}\| \to +\infty$. We prove the following equivalent statement. Given a sequence $\{V_{\Delta_k}\}_{k\geq 0}$, if there exists α such that $\|V_{\Delta_k}\| \leq \alpha\ \forall k \geq 0$ then, there exists β such that $\|\Pi_B^\perp \Delta_k \Pi_B^\perp\| \leq \beta\ \forall k \geq 0$. Notice that equation

$$V_{\Delta_k} = \Pi_B^\perp \Delta_k \Pi_B^\perp - A^* \Pi_B^\perp \Delta_k \Pi_B^\perp A \tag{3.45}$$

is a discrete-time Lyapunov equation corresponding to the pair (A^*, V_{Δ_k}). Since A is a stable matrix, we have

$$\Pi_B^\perp \Delta_k \Pi_B^\perp = \sum_{t=0}^{\infty} (A^*)^t V_{\Delta_k} A^t. \tag{3.46}$$

Hence, $\forall\ k \geq 0$ we have

$$\begin{aligned}
\|\Pi_B^\perp \Delta_k \Pi_B^\perp\| &= \|\sum_{t=0}^{\infty} (A^*)^t V_{\Delta_k} A^t\| \leq \sum_{t=0}^{\infty} \|(A^*)^t V_{\Delta_k} A^t\| \\
&\leq \sum_{t=0}^{\infty} \|(A^*)^t\| \|V_{\Delta_k}\| \|A^t\| \\
&= \left(\sum_{t=0}^{\infty} \|A^t\|^2\right) \|V_{\Delta_k}\| \leq \gamma\alpha < +\infty
\end{aligned} \tag{3.47}$$

where

$$\gamma := \sum_{t=0}^{\infty} \|A^t\|^2. \tag{3.48}$$

The latter is a finite quantity since A is a stable matrix. ■

Proposition 3.10. *$\mathcal{L}_{BG}^{\varphi}$ is an open and bounded set.*

Proof. By (3.27) and (3.37), it follows that $\mathcal{L}_{BG}^{\varphi}$ is an open set. We know that each $\Delta \in \mathcal{L}_{BG}$ must satisfy the following inequality

$$V_\Delta > -\hat{\Sigma}_C^{-1}. \tag{3.49}$$

Therefore

$$\min \sigma\,(V_\Delta) > -\varepsilon^2 \tag{3.50}$$

where $\sigma(X)$ denotes the spectrum of the matrix X and

$$\varepsilon^2 := \max \sigma(\hat{\Sigma}_C^{-1}). \tag{3.51}$$

We now show that a sequence $\{\Delta_k\}_{k\geq 0}$, with $\Delta_k \in \operatorname{Range} \varphi$, and $\|\Delta_k\| \to +\infty$, cannot belong to $\mathcal{L}_{BG}^{\varphi}$, i.e., $\mathcal{L}_{BG}^{\varphi}$ is bounded. To this end, it suffices to show that the minimum eigenvalue of V_{Δ_k} tends to $-\infty$ so that, for k large enough, Δ_k does not satisfy (3.50). By Lemma 3.9, $\|\Delta_k\| \to +\infty$ implies that, as k approaches infinity, $\|V_{\Delta_k}\| \to +\infty$. Hence,

$$\|P^{\frac{1}{2}} V_{\Delta_k} P^{\frac{1}{2}}\| \to +\infty, \tag{3.52}$$

for any given positive definite matrix $P > 0$. In particular, if we choose $P \in \operatorname{Range} \Gamma$, since $V_{\Delta_k} \in \operatorname{Range} \Gamma^{\perp}$ (see Proposition 3.6), we have

$$\langle V_{\Delta_k}, P\rangle = \operatorname{tr}(V_{\Delta_k} P) = \operatorname{tr}(P^{\frac{1}{2}} V_{\Delta_k} P^{\frac{1}{2}}) = 0. \tag{3.53}$$

Since $P^{\frac{1}{2}} V_{\Delta_k} P^{\frac{1}{2}}$ is Hermitian, from (3.52) and (3.53) it follows that $P^{\frac{1}{2}} V_{\Delta_k} P^{\frac{1}{2}}$, and hence V_{Δ_k}, have at least one eigenvalue tending to $-\infty$ as k approaches infinity. In conclusion, there exist an integer $\bar{k} > 0$ and an eigenvalue λ_k of V_{Δ_k} such that

$$\lambda_k \leq -\varepsilon^2 \qquad \forall\ k > \bar{k}. \tag{3.54}$$

Hence, in view of (3.50), $\Delta_k \notin \mathcal{L}_{BG}^{\varphi}$, $\forall\ k > \bar{k}$, and we may conclude that $\mathcal{L}_{BG}^{\varphi}$ is a bounded set. ■

We are now ready to prove existence of the minimum point.

Theorem 3.11. *The dual functional (3.34) has a unique minimum point in* $\mathcal{L}_{BG}^{\varphi}$.

Proof. In view of Corollary 3.8, we only need to show that J takes a minimum value on $\mathcal{L}_{BG}^{\varphi}$. First we observe that J_{BG} is continuous on its domain. We now demonstrate that J_{BG} is inf-compact, i.e., the image of $(-\infty, r]$ under the map J_{BG}^{-1} is a compact set. It is then sufficient to apply Weierstrass' theorem which states that a continuous function defined on a compact set admits a minimum. Indeed, observing that $J_{BG}(0) = \log\det\hat{\Sigma}_C$, we can restrict the search for a minimum point to the image of $(-\infty, \log\det\hat{\Sigma}_C]$ under J_{BG}^{-1}. Since, as stated in Proposition 3.10, $\mathcal{L}_{BG}^{\varphi}$ is a bounded set, to prove inf-compactness of J_{BG} it is sufficient to show that

$$\lim_{\Delta\to\partial\mathcal{L}_{BG}^{\varphi}} J_{BG}(\Delta) = +\infty. \tag{3.55}$$

Notice that $\partial\mathcal{L}_{BG}^{\varphi}$ is the set of $\Delta \in \operatorname{Range}\varphi$ for which

$$\hat{\Sigma}_C^{-1} + V_\Delta \tag{3.56}$$

is a singular positive semidefinite matrix. Thus, for $\Delta \to \partial\mathcal{L}_{BG}^{\varphi}$ all the eigenvalues of (3.56) remain bounded and at least one of them tends to 0^+. We denote with $\lambda_1, \ldots, \lambda_n > 0$ the eigenvalues of $\hat{\Sigma}_C + V_\Delta$ and, without loss generality, we suppose that, for $\Delta \to \partial\mathcal{L}_{BG}^{\varphi}$, $\lambda_1 \to 0^+$. Hence

$$\begin{aligned}\lim_{\Delta\to\partial\mathcal{L}_{BG}^{\varphi}} J_{BG}(\Delta) &= \lim_{\lambda_1\to 0^+} -\log\prod_{i=1}^{n}\lambda_i \\ &= \lim_{\lambda_1\to 0^+}\sum_{i=1}^{n}\log\frac{1}{\lambda_i} = +\infty.\end{aligned} \tag{3.57}$$

■

Corollary 3.12. *The set*

$$\tilde{S} := \{\Delta \in \mathcal{L}_{BG}^{\varphi} \mid J_{BG}(\Delta) \le J_{BG}(0) = \log\det\hat{\Sigma}_C\} \tag{3.58}$$

is compact.

3.3.1 A matricial Newton algorithm

In this section, we present a matricial Newton algorithm with backtracking stage for finding the minimum point of J_{BG} over $\mathcal{L}_{BG}^{\varphi}$. To this end we introduce the linear functional

$$\nabla J_\Delta = \nabla J_\Delta(\cdot) := \delta J_{BG}(\Delta, \cdot) \tag{3.59}$$

which may be interpreted as the *gradient* of J_{BG} at Δ. Here, $\delta J_{BG}(\Delta; \delta\Delta)$ is the first variation of J_{BG} at Δ in direction $\delta\Delta$. The bilinear form

$$\mathcal{H}_\Delta = \mathcal{H}_\Delta(\cdot, \cdot) := \delta^2 J_{BG}(\Delta; \cdot, \cdot) \tag{3.60}$$

may be interpreted as the *Hessian* of J_{BG} at Δ. Here, $\delta^2 J_{BG}(\Delta; \delta\Delta_1, \delta\Delta_2)$ is the second variation of J_{BG} at Δ in directions $\delta\Delta_1, \delta\Delta_2$. The algorithm steps are the following:

1. Set the initial condition $\Delta_0 = 0 \in \mathcal{L}_{BG}^{\varphi}$.

2. At each iteration, compute the Newton step $\delta\Delta_i$ over $\operatorname{Range}\varphi$ by solving the following equation

$$\mathcal{H}_{\Delta_i}(\delta\Delta_i, \cdot) = -\nabla J_{\Delta_i}(\cdot), \tag{3.61}$$

   where the gradient ∇J_{Δ_i} and the Hessian $\mathcal{H}_{\Delta_i}$ are defined by (3.59) and (3.60), respectively. Taking into account (3.38) and (3.39), the latter equation may be written explicitly as

$$\begin{aligned} \Pi_B^{\perp} \left[P_{BG}(\Delta_i) V_{\delta\Delta_i} P_{BG}(\Delta_i) - A P_{BG}(\Delta_i) V_{\delta\Delta_i} P_{BG}(\Delta_i) A^* \right] \Pi_B^{\perp} = \\ = -\Pi_B^{\perp} (A P_{BG}(\Delta_i) A^* - P_{BG}(\Delta_i)) \Pi_B^{\perp}. \end{aligned} \tag{3.62}$$

3. Set $t_i^0 = 1$, and let $t_i^{k+1} = t_i^k/2$ until both of the following conditions hold:

$$\hat{\Sigma}_C^{-1} + V_{\Delta_i + t_i^k \delta\Delta_i} > 0 \tag{3.63}$$

$$J(\Delta_i + t_i^k \delta\Delta_i) < J(\Delta_i) + \alpha t_i^k \operatorname{tr}\left[\nabla J_{\Delta_i} \delta\Delta_i\right] \tag{3.64}$$

   where $\alpha \in (0, \frac{1}{2})$ is a real constant. Notice that, since $\Delta_i + t_i^k \delta\Delta_i \in \operatorname{Range}\varphi$, condition (3.63) implies that

$$\Delta_i + t_i^k \delta\Delta_i \in \mathcal{L}_{BG}^{\varphi}. \tag{3.65}$$

4. Set $\Delta_{i+1} = \Delta_i + t_i^k \delta\Delta_i$.

5. Repeat steps 2, 3 and 4 until the condition $\|\nabla J_{\Delta_i}\|_2 < \varepsilon$ is satisfied, where ε is a (small) tolerance threshold, then set $\Delta^\circ = \Delta_i$.

We suggest the following procedure to solve Equation (3.61) by taking into account the explicit form (3.62):

1. Take a basis $\{L_1, \ldots, L_l\}$ of $\operatorname{Range}\varphi$.

2. Compute
$$Y = \Pi_B^{\perp}(P_{BG}(\Delta_i) - AP_{BG}(\Delta_i)A^*)\Pi_B^{\perp} \tag{3.66}$$

3. For each L_k, compute:
$$Y_k = \Pi_B^{\perp}\left[P_{BG}(\Delta_i)V_{L_k}P_{BG}(\Delta_i) - AP_{BG}(\Delta_i)V_{L_k}P_{BG}(\Delta_i)A^*\right]\Pi_B^{\perp} \tag{3.67}$$

4. Solve, by means of linear algebraic methods (the Moore-Penrose pseudoinverse), the equation
$$\sum_k \alpha_k Y_k = Y \tag{3.68}$$

5. By linearity, the solution to (3.61) is:
$$\delta\Delta_i = \sum_k \alpha_k L_k \in \operatorname{Range}\varphi. \tag{3.69}$$

Since the minimum of J_{BG} exists and is unique, we investigate the global convergence of our Newton algorithm. To prove the convergence we need of the following result.

Proposition 3.13. *Consider a function $f : D \subset \mathbb{R}^k \to \mathbb{R}$ twice differentiable on D with H_x the Hessian of f at x. Suppose moreover that f is strongly convex on a set $S \subset D$, i.e. there exists a constant $m > 0$ such that $H_x \geq mI$ for $x \in S$, and H_x is Lipschitz continuous on S. Let $\{x_i\} \in S$ be the sequence generated by the Newton algorithm. Under these assumptions, Newton's algorithm with backtracking converges globally. More specifically, $\{x_i\}$ decreases in linear way for a finite number of steps, and converges in a quadratic way to the minimum point after the linear stage.*

Proof. See [11, 9.5.3, p. 488]. ■

To prove the convergence of our algorithm, we proceed in the following manner: Identify a compact set $\tilde{S}$ such that $\Delta_i \in \tilde{S}$ and prove that the second variation is coercive and Lipschitz continuous on $\tilde{S}$. We then apply Proposition 3.13 in order to prove convergence.
Since $\Delta_0 = 0$, we consider the set
$$\tilde{S} := \{\Delta \in \mathcal{L}_{BG}^{\varphi} \mid J_{BG}(\Delta) \leq J_{BG}(\Delta_0) = \log\det\hat{\Sigma}_C\} \tag{3.70}$$
which is compact (see Corollary 3.12). The presence of the backtracking stage in the algorithm guarantees that the sequence $J_{BG}(\Delta_0), J_{BG}(\Delta_1), \ldots$ is decreasing. Thus $\Delta_i \in \tilde{S}$, $\forall i \geq 0$.

Proposition 3.14. *Consider the Hessian $\mathcal{H}_\Delta$ defined in (3.60) and the associated quadratic form. The following facts hold:*

1. *As a quadratic form, $\mathcal{H}_\Delta$ is coercive and bounded on $\tilde{S}$, namely there exist $m, M > 0$ such that*

$$m\|\delta\Delta\|^2 \leq \mathcal{H}_\Delta(\delta\Delta, \delta\Delta) \leq M\|\delta\Delta\|^2, \ \forall\, \Delta \in \tilde{S}. \tag{3.71}$$

2. *$\mathcal{H}_\Delta$ is Lipschitz continuous on $\tilde{S}$.*

Proof. 1. First, observe that $\tilde{S} \subset \mathcal{L}_{BG}$. Since $\tilde{S}$ is a compact set, there exists $\varepsilon > 0$ such that

$$P_{BG}(\Delta) = \left(\hat{\Sigma}_C + V_\Delta\right)^{-1} \geq \varepsilon I. \tag{3.72}$$

Accordingly, for $\delta\Delta \neq 0$,

$$\begin{aligned}
\mathcal{H}_\Delta(\delta\Delta, \delta\Delta) &= \mathrm{tr}[P_{BG}(\Delta)^{\frac{1}{2}} V_{\delta\Delta} P_{BG}(\Delta) V_{\delta\Delta} P_{BG}(\Delta)^{\frac{1}{2}}] \\
&\geq \varepsilon \mathrm{tr}[P_{BG}(\Delta)^{\frac{1}{2}} V_{\delta\Delta} V_{\delta\Delta} P_{BG}(\Delta)^{\frac{1}{2}}] = \varepsilon \mathrm{tr}[V_{\delta\Delta} P_{BG}(\Delta) V_{\delta\Delta}] \\
&\geq \varepsilon^2 \mathrm{tr}[V_{\delta\Delta} V_{\delta\Delta}] = \varepsilon^2 \gamma > 0,
\end{aligned}$$

where $\gamma := \mathrm{tr}[V_{\delta\Delta} V_{\delta\Delta}] > 0$, since $V_{\delta\Delta}$ is not the zero matrix when $\delta\Delta \neq 0$, as observed in the proof of Lemma 3.7. Since $J_{BG} \in \mathcal{C}^\infty(\mathcal{L}_{BG})$, it follows that $\mathcal{H}_\Delta$ is continuous on the compact $\tilde{S}$ where it is also strictly positive definite. Hence, $\mathcal{H}_\Delta$ is coercive and bounded on $\tilde{S}$.
2. $\mathcal{H}_\Delta \in \mathcal{C}^1(\tilde{S})$ and $\|\mathcal{H}_\Delta\| \leq M \ \forall\, \Delta \in \tilde{S}$, therefore $\mathcal{H}_\Delta$ is Lipschitz continuous on $\tilde{S}$. ■

Proposition 3.15. *The sequence $\{\Delta_i\}_{i\geq 0}$ generated by the proposed Newton algorithm converges to the unique minimum point of J_{BG} in $\mathcal{L}^{\varphi}_{BG}$.*

Proof. Proposition 3.13 applied to functions with domain contained in $\mathbb{R}^k$, $k \in \mathbb{N}$. The functional J_{BG} is defined over a subset of the linear space $\mathrm{Range}\,\varphi$ which has finite dimension d on $\mathbb{R}$. We define $x := vect(X)$ as the column vector (with n^2 entries) obtained by stacking the columns of X one over the other, and we consider the following change of representation

$$\Delta \mapsto \lambda = vect(\Delta). \tag{3.73}$$

Let be D, S, f, H_λ the corresponding representation of $\mathcal{L}_{BG}, \tilde{S}, J_{BG}, \mathcal{H}_\Delta$. It follows that:

- By Lemma 3.7, $f : D \subset \mathbb{R}^d \to \mathbb{R}$ is twice differentiable on $S \subset D$

- By Proposition 3.14 and since $\mathcal{H}_\Delta(\cdot,\cdot)$ is a bilinear form, follows that f is strongly convex on S and H_λ is Lipschitz continuous on S.

Therefore all the hypothesis of Proposition 3.13 are satisfied and the conclusion follows. ■

3.3.2 Performance comparison

In this section, we use the following notation:

- *PJ method* to denote the projection method outlined in the last part of Section 3.2;

- *ME method* to denote the maximum entropy method based on the minimization of the information divergence (or equivalently the Burg matrix divergence).

A performance comparison procedure

Suppose that we have a finite sequence $y(1)\dots y(N)$ extracted from a sample path of a zero-mean, weakly stationary discrete-time process y. We want to compare the estimates $\hat{\Sigma}_{PJ}$ and $\hat{\Sigma}_{ME}$ obtained by the PJ and ME methods, respectively. In order to make the comparison reasonably independent of the specific data set, we average over $M = 500$ experiments performed with sequences extracted from different sample paths. We are now ready to describe the comparison procedure:

- Fix the transfer function $G(z)$.

- At the k-th experiment $G(z)$ is fed by the data $y^k(1)\dots y^k(N)$ and we collect the output data $x^k(1)\dots x^k(N)$.

- Compute the consistent estimate $\hat{\Sigma}_C(k)$ of the covariance matrix of the output from $x^k(\tilde{N})\dots x^k(N)$, with $\tilde{N} < N$, as in (3.17). Note that the first $\tilde{N}-1$ output samples $x^k(1)\dots x^k(\tilde{N}-1)$ are discarded so that the filter can be considered to operate in steady state.

- From $\hat{\Sigma}_C(k)$, estimate $\hat{\Sigma}_{PJ}(k)$ and $\hat{\Sigma}_{ME}(k)$ using PJ and ME method respectively.

- Compute the relative error norm[2] between Σ and its estimates $\hat{\Sigma}_{PJ}(k)$ and $\hat{\Sigma}_{ME}(k)$

$$e_{PJ}(k) = \frac{\|\hat{\Sigma}_{PJ}(k) - \Sigma\|}{\|\Sigma\|},\ e_{ME}(k) = \frac{\|\hat{\Sigma}_{ME}(k) - \Sigma\|}{\|\Sigma\|}$$

- When the experiments are completed, compute the mean and the variance of the relative error norm

$$\begin{aligned}
\mu_{PJ} &= \frac{1}{M}\sum_{k=1}^{M} e_{PJ}(k),\ \mu_{ME} = \frac{1}{M}\sum_{k=1}^{M} e_{ME}(k), \\
\sigma^2_{PJ} &= \frac{1}{M}\sum_{k=1}^{M}(e_{PJ}(k) - \mu_{PJ})^2, \\
\sigma^2_{ME} &= \frac{1}{M}\sum_{k=1}^{M}(e_{ME}(k) - \mu_{ME})^2.
\end{aligned} \tag{3.74}$$

- Count the times that the PJ method adjusts the projected estimation $\hat{\Sigma}_{\Gamma}(k)$ by adding to it the quantity $\varepsilon\Sigma_+$. This number is denoted as $\sharp F$.

The output of this procedure are the parameters μ_j, σ_j^2 and $\sharp F$. Clearly, the smaller these parameters, the better estimation is expected.

Simulation results for the real scalar case

We choose a real scalar process y with a high-order spectral density Ω (represented by the solid line in Figure 3.1). The bank of filters has the following structure.

$$A = \begin{bmatrix} a_1 & 1 & 0 & 0 & 0 & 0 \\ 0 & a_2 & 1 & 0 & 0 & 0 \\ 0 & 0 & a_3 & 1 & 0 & 0 \\ 0 & 0 & 0 & a_4 & 1 & 0 \\ 0 & 0 & 0 & 0 & a_5 & 1 \\ 0 & 0 & 0 & 0 & 0 & a_6 \end{bmatrix},\ B = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}. \tag{3.75}$$

First we choose

$$a_1 = a_2 = a_3 = 0.4, \quad a_4 = a_5 = a_6 = 0.5. \tag{3.76}$$

[2]Here the norm $\|\cdot\|$ is the *spectral norm.* i.e. the matrix norm induced by the Euclidean norm in $\mathbb{C}^n$.

In this case, the true covariance Σ of the process x has the following eigenvalues $\lambda_1 = 0.2408, \lambda_2 = 0.4775, \lambda_3 = 1.9235, \lambda_4 = 2.8125, \lambda_5 = 21.1455, \lambda_6 = 285.2539$. Thus Σ has a condition number of the order of 10^3. In Table 3.1, we present the results obtained for different lengths N of the observed sequences $y(1) \dots y(N)$. In this case, the ME method appears produce only a

N	μ_{PJ}	μ_{ME}	σ^2_{PJ}	σ^2_{ME}	$\sharp F$
300	0.19827	0.19045	0.026206	0.018779	11
500	0.14329	0.14169	0.013137	0.011047	1
700	0.12701	0.1269	0.0092071	0.0091365	0
1000	0.10679	0.10677	0.0064781	0.0064709	0

Table 3.1: Parameters μ_{PJ}, μ_{ME}, σ^2_{PJ}, σ^2_{ME}, $\sharp F$ for A, B given by (3.75)-(3.76).

very marginal improvement with respect to the projection method. Moreover, as N increases, μ and σ^2 decrease for both methods: In fact, $\hat{\Sigma}_C \to \Sigma$ with probability one as $N \to +\infty$. Therefore, as N increases, the performances of the two methods are more and more similar. This picture, however, changes dramatically if the time-constants of the dynamics of the filter $G(z)$ are significatively different. Consider, for example, a filters bank with the same structure (3.75) but with

$$a_1 = 0.3,\ a_2 = 0.4,\ a_3 = 0.5,\ a_4 = 0.6,\ a_5 = 0.7,\ a_6 = 0.8. \tag{3.77}$$

In this case, the eigenvalues of Σ are $\lambda_1 = 0.3, \lambda_2 = 0.7, \lambda_3 = 2.1, \lambda_4 = 8.5, \lambda_5 = 123.8, \lambda_6 = 3862.3$. Thus, the condition number of Σ is of the order of 10^4. In Table 3.2 we present the results obtained for different lengths N of the observed sequences $y(1) \dots y(N)$.

N	μ_{PJ}	μ_{ME}	σ^2_{PJ}	σ^2_{ME}	$\sharp F$
300	1.6509	0.24068	8.1929	0.030232	197
500	0.99964	0.17711	3.1593	0.018377	141
700	0.60446	0.15266	1.2813	0.01464	99
1000	0.49333	0.13648	0.95235	0.011395	78

Table 3.2: Parameters μ_{PJ}, μ_{ME}, σ^2_{PJ}, σ^2_{ME}, $\sharp F$ for A, B given by (3.75)-(3.77).

In this situation, the condition number of Σ is larger than in the previous case. Thus, the projection of $\hat{\Sigma}_C$ (that is a perturbed version of Σ) onto Range Γ yields a matrix $\hat{\Sigma}_\Gamma$ that, in many cases, fails to be positive definite

(or even positive semidefinite). This explains why the number of failures $\sharp F$ is significant. Recall that, when the projection fails to be positive definite, the PJ method adjusts $\hat{\Sigma}_\Gamma$ by adding a positive definite matrix Σ_+ belonging to Range Γ. For each experiment, Σ_+ is the same. Hence, the adjustment cannot provide a good estimate of Σ. This is the heuristic reason why, in this case, the estimates provided by our method largely outperform those obtained by the projection method. Indeed, even increasing N to 1000 (so that the observed sequences are pretty long), the differences in the performances remain remarkable.

Remark 3.16. We hasten to anticipate that even in the case of the filters bank (3.75)-(3.76), with $N = 500$ or larger, when the estimation errors of the PJ and ME methods have practically the same mean and variance, the THREE-like spectral estimator performs much better when initialized with $\hat{\Sigma}_{ME}$ than when initialized with $\hat{\Sigma}_{PJ}$ (see next section).

Simulation results for the real multivariable case

We consider a bivariate real process y with a high-order spectral density Ω. As for the scalar case, we consider two filters banks with the same structure:

$$A = \begin{bmatrix} a_1 & 1 & 0 & 0 & 0 & 0 \\ 0 & a_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & a_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & a_2 & 1 & 0 \\ 0 & 0 & 0 & 0 & a_2 & 1 \\ 0 & 0 & 0 & 0 & 0 & a_2 \end{bmatrix}, \; B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3.78}$$

In the first case

$$a_1 = 0.55, \; a_2 = 0.65. \tag{3.79}$$

In this case, the true Σ has the following eigenvalues: $\lambda_1 = 0.0031, \lambda_2 = 0.0120, \lambda_3 = 0.9863, \lambda_4 = 2.3744, \lambda_5 = 8.1872, \lambda_6 = 84.0289$. The corresponding error means and variances for the two estimation methods PJ and ME are reported in Table 3.3 for different values of the length N of the observed data sequences $y(1) \dots y(N)$.

The second filters bank has the same structure (3.78), but the eigenvalues of A are closer to the unit circle:

$$a_1 = 0.6, \; a_2 = 0.7. \tag{3.80}$$

In this case, the true Σ has the following eigenvalues: $\lambda_1 = 0.0034, \lambda_2 = 0.0169, \lambda_3 = 1.4706, \lambda_4 = 2.9195, \lambda_5 = 11.8157, \lambda_6 = 159.1730$. The corresponding error means and variances are reported in the Table 3.4. As it

N	μ_{PJ}	μ_{ME}	σ^2_{PJ}	σ^2_{ME}	$\sharp F$
300	0.3372	0.17694	0.46351	0.017164	33
500	0.16107	0.1431	0.043056	0.011703	6
700	0.12044	0.11778	0.010054	0.006674	1
1000	0.09712	0.09696	0.005345	0.005333	0

Table 3.3: Parameters μ_{PJ}, μ_{ME}, σ^2_{PJ}, σ^2_{ME}, $\sharp F$ for A, B given by (3.78)-(3.79).

N	μ_{PJ}	μ_{ME}	σ^2_{PJ}	σ^2_{ME}	$\sharp F$
300	1.0234	0.20392	3.5055	0.022113	93
500	0.45285	0.14658	1.4239	0.013541	35
700	0.25175	0.12041	0.55549	0.0082425	16
1000	0.16576	0.11102	0.22098	0.006274	7

Table 3.4: Parameters μ_{PJ}, μ_{ME}, σ^2_{PJ}, σ^2_{ME}, $\sharp F$ for A, B given by (3.78)-(3.80).

can be observed from the tables, the scenario is the same as in the scalar case: The ME method performs remarkably better than the PJ method, particularly for the second filters bank.

Simulation results for the complex case

So far we have considered only real examples because this situation is more common in control engineering applications. Since the theory has, however, been developed for the more general complex case, we also include the following complex example where the process y is a high order (the McMillan degree of the corresponding spectral density Ω is 80) complex-valued scalar process. Let A and B be defined by:

$$A = \begin{bmatrix} a_1 & 1 & 0 & 0 & 0 & 0 \\ 0 & a_2 & 1 & 0 & 0 & 0 \\ 0 & 0 & a_3 & 1 & 0 & 0 \\ 0 & 0 & 0 & a_4 & 1 & 0 \\ 0 & 0 & 0 & 0 & a_5 & 1 \\ 0 & 0 & 0 & 0 & 0 & a_6 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \tag{3.81}$$

where $a_i = 0.7\mathrm{e}^{\mathrm{j}\omega_i}$ and $\omega_1 := 0.8148$, $\omega_2 := 0.9058$, $\omega_3 := 0.1270$, $\omega_4 := 0.9133$, $\omega_5 := 0.6324$, $\omega_6 := 0.0976$. The eigenvalues of the matrix Σ are $\lambda_1 = 0.5$, $\lambda_2 = 3.06$, $\lambda_3 = 17.65$, $\lambda_4 = 132.79$, $\lambda_5 = 1.55 \cdot 10^3$, $\lambda_6 = 2.67 \cdot 10^4$. Table 3.5, where the performances of our method are compared to those of the

projection method, shows that also in this case our approach is particularly convenient:

N	μ_{PJ}	μ_{ME}	σ^2_{PJ}	σ^2_{ME}	$\sharp F$
300	0.6861	0.1481	0.79881	0.011669	148
500	0.32225	0.11871	0.3374	0.006533	59
700	0.20526	0.10044	0.18662	0.004495	29
1000	0.13541	0.08525	0.0919	0.003015	14

Table 3.5: Parameters μ_{PJ}, μ_{ME}, σ^2_{PJ}, σ^2_{ME}, $\sharp F$ for A, B given by (3.81).

3.3.3 Application to spectral estimation

Next, we compare the estimated spectral densities, obtained by one of the THREE-like spectral estimation procedures, when initialized with the true variance Σ and with the two estimates $\hat{\Sigma}_{PJ}$ and $\hat{\Sigma}_{ME}$. We stress that, while the results of Section 3.3.2 compare the estimated covariance $\hat{\Sigma}_{ME}$ or $\hat{\Sigma}_{PJ}$ to the true Σ, the following comparison evaluates the different performances directly in terms of the main applications of the methods, i.e. spectral estimation.

Simulation results for the scalar case using the Prior-THREE algorithm

From the procedure presented in Subsection 3.3.2, we get the state covariance estimates $\hat{\Sigma}_{PJ}(k)$ and $\hat{\Sigma}_{ME}(k)$ for $k = 1 \dots, M$. Thus, we exploit this set of estimates (with $M = 500$ experiments and $N = 500$) as input state covariances for the spectrum approximation problem with $\mathcal{S}_{\mathrm{KL0}}(\Psi \| \Phi)$ in (2.6):

- We consider a prior spectral density $\Psi_k(\mathrm{e}^{j\vartheta})$ that may depend on the data $y^k(1) \dots y^k(N)$ and hence is indexed on k.

- For each experiment k, we compute the spectrum estimate $\hat{\Phi}_{T,k}(\mathrm{e}^{j\vartheta})$ solving the spectrum approximation problem associated to $\mathcal{S}_{\mathrm{KL0}}(\Psi \| \Phi)$ with inputs $\Psi_k(\mathrm{e}^{j\vartheta})$ and the true variance Σ.

- For each experiment k, we compute the spectrum estimates $\hat{\Phi}_{PJ,k}(\mathrm{e}^{j\vartheta})$, and $\hat{\Phi}_{ME,k}(\mathrm{e}^{j\vartheta})$ which are solution to the spectrum approximation problem associated to $\mathcal{S}_{\mathrm{KL0}}(\Psi \| \Phi)$ using the same "a priori" spectral density $\Psi_k(\mathrm{e}^{j\vartheta})$ and taking $\hat{\Sigma}_{PJ}(k)$ and $\hat{\Sigma}_{ME}(k)$, respectively, as state variance.

- When the spectral estimates are completed, we compute the mean estimates

$$\Phi_T(\mathrm{e}^{j\vartheta}) := \frac{1}{M}\sum_{k=1}^{M}\hat{\Phi}_{T,k}(\mathrm{e}^{j\vartheta})$$
$$\Phi_{PJ}(\mathrm{e}^{j\vartheta}) := \frac{1}{M}\sum_{k=1}^{M}\hat{\Phi}_{PJ,k}(\mathrm{e}^{j\vartheta})$$
$$\Phi_{ME}(\mathrm{e}^{j\vartheta}) := \frac{1}{M}\sum_{k=1}^{M}\hat{\Phi}_{ME,k}(\mathrm{e}^{j\vartheta}) \tag{3.82}$$

and the mean of the error norm for each method with respect to $\hat{\Phi}_{T,k}(\mathrm{e}^{j\vartheta})$

$$E_{PJ}(\vartheta) : = \frac{1}{M}\sum_{k=1}^{M}|\hat{\Phi}_{PJ,k}(\mathrm{e}^{j\vartheta}) - \hat{\Phi}_{T,k}(\mathrm{e}^{j\vartheta})|$$
$$E_{ME}(\vartheta) : = \frac{1}{M}\sum_{k=1}^{M}|\hat{\Phi}_{ME,k}(\mathrm{e}^{j\vartheta}) - \hat{\Phi}_{T,k}(\mathrm{e}^{j\vartheta})|. \tag{3.83}$$

Remark 3.17. Notice that the very same procedure may be employed to deal with the solution (2.5) which is just the special case of the (2.7) corresponding to the choice $\Psi_k(\mathrm{e}^{j\vartheta}) \equiv 1$ for the prior spectral density.
Notice also that, in the above procedure, an essential degree of freedom is the filter bank $G(z)$. Indeed, the choice of $G(z)$ has profound implications (see [14],[38],[28] and [27]). As noticed before, it turns out that the spectrum estimate has better resolution in those sectors of the unit circle where more eigenvalues are located close to the unit circle.

To perform the comparison, we have chosen the two filters (3.75)-(3.76), (3.75)-(3.77) and we have set the prior spectral density to be $\Psi_k(z) := W_k(z)W_k^*(z)$ where $W_k(z) = \left[\hat{\sigma}_e\frac{c(z)}{a(z)}\right]_k$ is a three-order AR model estimated from the sequence $y^k(1)\dots y^k(N)$ extracted from the k-th sample path of the process y.

In Figure 3.1 the mean spectra corresponding to the filters bank (3.75)-(3.76) are depicted. In Figure 3.2 the corresponding mean error norms are represented.

It is apparent that our method produces an estimate $\hat{\Sigma}_{ME}$ for which the corresponding spectral density Φ_{ME} approximates the true Φ almost as well as the estimation produced starting from the true Σ, while the estimation

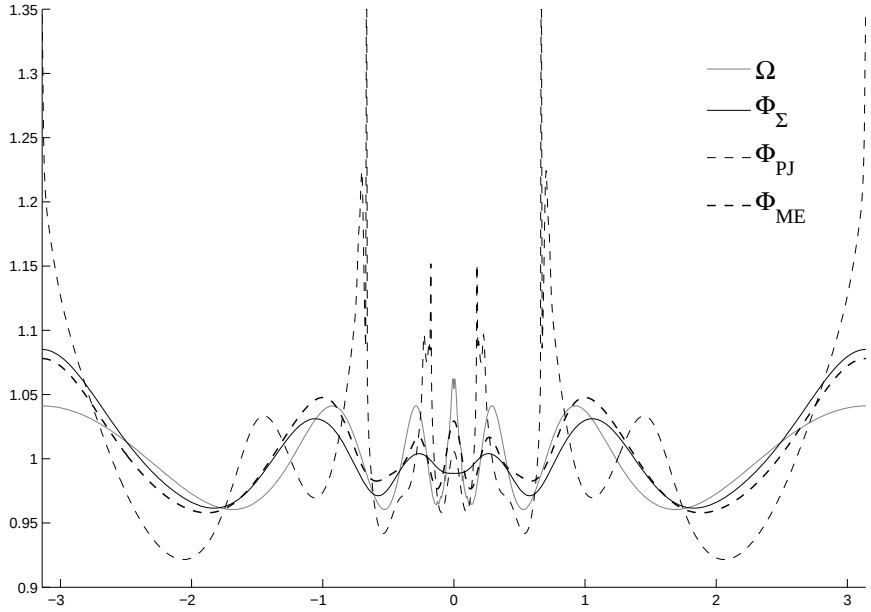


Figure 3.1: Mean spectra comparison using the solution (2.7), with the bank of filters (3.75)-(3.76).

corresponding to $\hat{\Sigma}_{PJ}$ is highly unsatisfactory. Notice also that, although in this case $\hat{\Sigma}_{PJ}$ and $\hat{\Sigma}_{ME}$ appear quite similar (see the table in the previous section), the estimated spectra are very different and the ME method provides a considerable improvement, cf. Remark 3.16.

Figures 3.3 and 3.4 show the mean spectra and the mean error norm, respectively, when the filters bank (3.75)-(3.77) is employed. As expected, in this case the inferior performance of the PJ method when compared to the ME method is more salient while the ME method practically performs as well as the estimation Φ_Σ produced by employing the true Σ. Similar results are obtained when we consider solution in (2.5), i.e. $\Psi_k(\mathrm{e}^{j\vartheta}) \equiv 1$.

Simulation results for the multivariable case

We have carried out this comparison along the very same lines of the previous simulation employing the same $\Omega(z)$ and the same two filters $G(z)$ used in Subsection 3.3.2. The only differences with respect to the above simulation for the scalar case are the following:

1. For the spectral estimation, we have employed the maximum entropy solution (2.5) in which we have plugged the true variance Σ and the two estimates $\hat{\Sigma}_{PJ}$ and $\hat{\Sigma}_{ME}$.

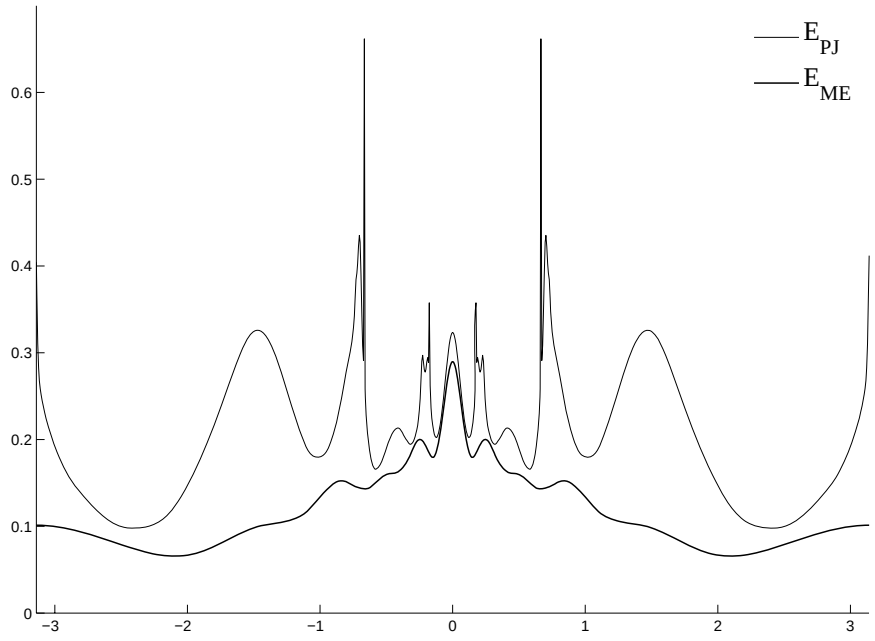


Figure 3.2: Mean error norm comparison using the solution (2.7), with the bank of filters (3.75)-(3.76).

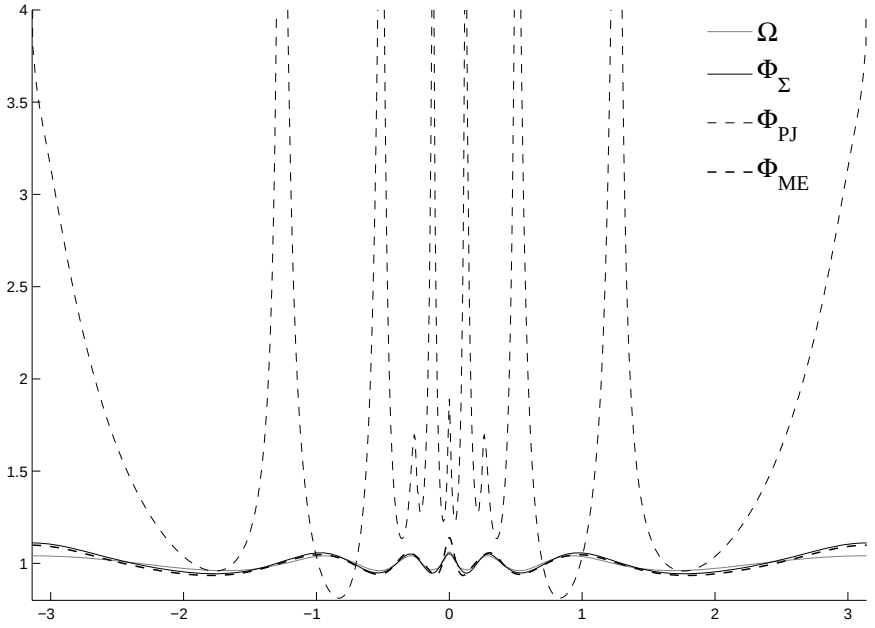


Figure 3.3: Mean spectra comparison using the solution (2.7), with the bank of filters (3.75)-(3.77).

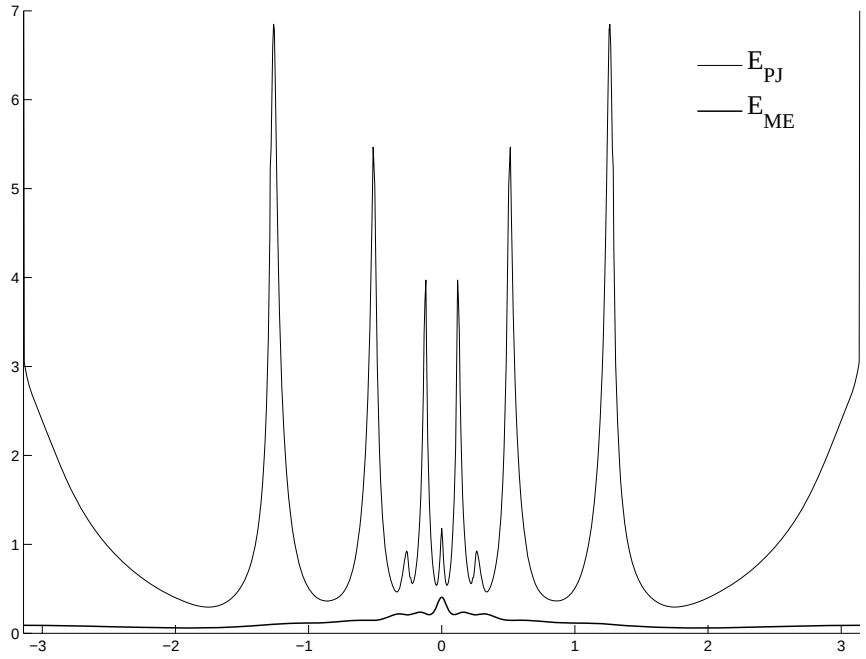


Figure 3.4: Mean error norm comparison using the solution (2.7), with the bank of filters (3.75)-(3.77).

2. We have modified (3.83) by using the matrix induced norms in place of the absolute values.

3. We have illustrated only the mean of the errors norm since comparing the 2×2 spectral densities would require four pictures for each case.

Figure 3.5 shows the mean of the error norm in the case of filters bank

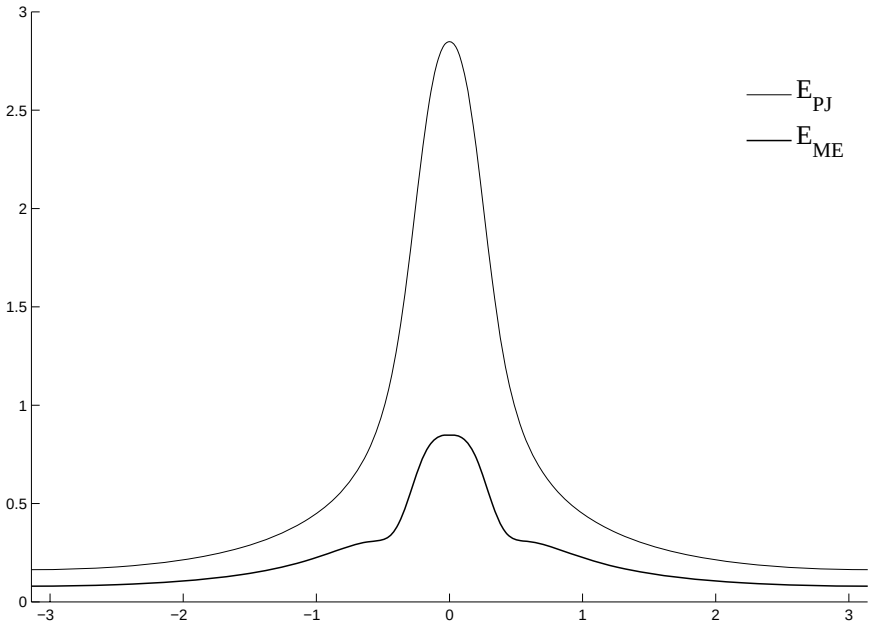


Figure 3.5: Mean of the error norm comparison using the maximum entropy solution (2.5), with the bank of filters (3.78)-(3.79).

(3.78)-(3.79). Although $\hat{\Sigma}_{PJ}$ and $\hat{\Sigma}_{ME}$ are quite similar in this case (see the table in the previous section), the difference among the mean error norms is more evident and the ME method provides an estimate closer to the estimate obtained using the "true" Σ. Finally, in Figure 3.6, the mean of the error norm is depicted for the case of the filters bank (3.78)-(3.80). The spectral estimate obtained using $\hat{\Sigma}_{PJ}$ is clearly unsatisfactory with respect to the one obtained using $\hat{\Sigma}_{ME}$. In conclusion, the significant improvement in spectral estimation brought about by our method occurs also in the multivariable setting.

We conclude that, in several critical cases, the projection method of Section 3.2 provides a poor estimate of the covariance matrix Σ, compromising the quality of the spectral estimator. Moreover, simulation shows that, even when the projection-based estimate $\hat{\Sigma}_{PJ}$ looks close to our estimate $\hat{\Sigma}_{ME}$, the

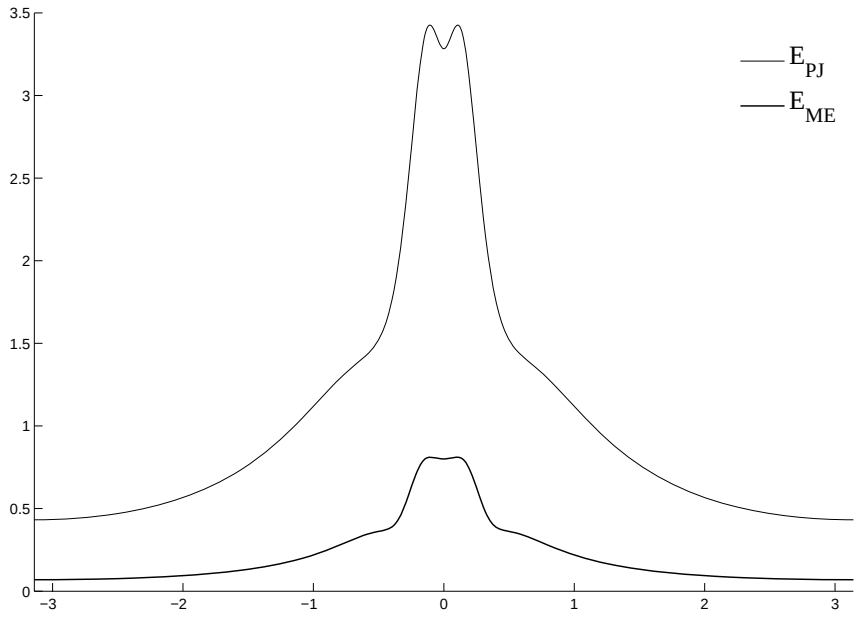


Figure 3.6: Mean of the error norm comparison using the maximum entropy solution (2.5), with the bank of filters (3.78)-(3.80).

spectral estimator initialized with $\hat{\Sigma}_{ME}$ significantly outperforms the other one. Indeed, it often performs nearly as well as the estimator initialized with the true state covariance Σ.

3.4 Estimates with the Beta matrix divergence

Problem 3.4 may be extended by considering a generic index divergence $\mathcal{D}$ among (positive definite) covariance matrices.

Problem 3.18. *Given $\hat{\Sigma}_C \in \mathcal{H}_{n,+}$ and $G(z)$ with the previous properties, solve*

$$\text{minimize } \mathcal{D}(P\|\hat{\Sigma}_C) \text{ over } P \in (\mathcal{H}_{n,+} \cap \text{Range } \Gamma)\,. \tag{3.84}$$

One possible matrix divergence index is the Beta matrix divergence (family) which is defined as follows:

$$\mathcal{D}_\beta(P\|\hat{\Sigma}) := \text{tr}[\frac{1}{\beta-1}(P^\beta - P\hat{\Sigma}^{\beta-1}) - \frac{1}{\beta}(P^\beta - \hat{\Sigma}^\beta)] \tag{3.85}$$

where $P, \hat{\Sigma} \in \mathcal{H}_{n,+}$ and $\beta \in \mathbb{R} \setminus \{0,1\}$. In fact, $\mathcal{D}_\beta(P\|\hat{\Sigma})$ is the Beta divergence $\mathcal{S}_\beta(\Phi\|\Psi)$, introduced in Section 2.3, among the two constant spectral densities $\Phi(\mathrm{e}^{j\vartheta}) \equiv P$ and $\Psi(\mathrm{e}^{j\vartheta}) \equiv \hat{\Sigma}$ [3]. Since $\mathcal{D}_\beta$ is a special case of $\mathcal{S}_\beta$, it is strictly convex with respect to the first argument. Moreover, it is a continuous function of real variable $\beta \in \mathbb{R}$ with

$$\begin{aligned} \lim_{\beta\to 0} \mathcal{D}_\beta(P\|\hat{\Sigma}) &= \mathcal{D}_{\mathrm{BG}}(P\|\hat{\Sigma}) \\ \lim_{\beta\to 1} \mathcal{D}_\beta(P\|\hat{\Sigma}) &= \mathcal{D}_{\mathrm{KL}}(P\|\hat{\Sigma}) \end{aligned} \tag{3.86}$$

where

$$\mathcal{D}_{\mathrm{KL}}(P\|\hat{\Sigma}) := \mathrm{tr}[P(\log(P) - \log(\hat{\Sigma})) - P + \hat{\Sigma}] \tag{3.87}$$

is the extension of the *Umegaki-von Neumann*'s relative entropy, [53], to non equal-trace matrices.

In what follows, we consider the parametrization $\beta = -\frac{1}{\nu}+1$ with $\nu \in \mathbb{N}_+$, and we define $\mathcal{D}_\nu(P\|\hat{\Sigma}_C) := \mathcal{D}_\beta(P\|\hat{\Sigma}_C)$ with $\beta = -\frac{1}{\nu} + 1$. In Section 3.3, we already showed that there exists a unique solution, say $\hat{\Sigma}_{ME}$, to Problem 3.18 with $\mathcal{D}_\nu$ and $\nu = 1$. Consider now Problem 3.18 with $\nu \in \mathbb{N}_+ \setminus \{1\}$. The corresponding *Lagrange* functional is

$$\begin{aligned} L_\nu(P,\Delta) &:= \mathcal{D}_\nu(P\|\hat{\Sigma}_C) + \frac{\nu}{1-\nu}\mathrm{tr}(\hat{\Sigma}_C^{\frac{\nu-1}{\nu}}) + \mathrm{tr}[\Delta\Pi_B^\perp(P - A^*PA)\Pi_B^\perp] \\ &= \mathcal{D}_\nu(P\|\hat{\Sigma}_C) + \frac{\nu}{1-\nu}\mathrm{tr}(\hat{\Sigma}_C^{\frac{\nu-1}{\nu}}) + \mathrm{tr}(PV_\Delta). \end{aligned} \tag{3.88}$$

Since $L_\nu(P,\Delta+\bar{\Delta}) = L_\nu(P,\Delta)\ \forall\bar{\Delta} \in \ker\varphi = [\mathrm{Range}\,\varphi]^\perp$, we can assume that $\Delta \in \mathrm{Range}\,\varphi$. Moreover, $L_\nu(\cdot,\Delta)$ is strictly convex over $\mathcal{H}_{n,+}$. Thus, the unique minimum point of $L_\nu(\cdot,\Delta)$ is given by annihilating the first directional derivative of $L_\nu(\cdot,\Delta)$ in each direction $\delta P \in \mathcal{H}_n$

$$\delta L_\nu(P,\Delta;\delta\Delta) = \mathrm{tr}[(-\nu P^{-\frac{1}{\nu}} + \nu\hat{\Sigma}_C^{-\frac{1}{\nu}} + V_\Delta)\delta P]. \tag{3.89}$$

Thus, (3.89) is zero for each $\delta P \in \mathcal{H}_n$ if and only if

$$P^{-\frac{1}{\nu}} = \hat{\Sigma}_C^{-\frac{1}{\nu}} + \frac{1}{\nu}V_\Delta. \tag{3.90}$$

Since $P^{-\frac{1}{\nu}} \in \mathcal{H}_{n,+}$, the set of the admissible *Lagrange* multipliers is

$$\mathcal{L}_\nu^\varphi := \left\{\Delta \in \mathcal{H}_n \mid \hat{\Sigma}_C^{-\frac{1}{\nu}} + \frac{1}{\nu}V_\Delta > 0\right\} \cap \mathrm{Range}\,\varphi. \tag{3.91}$$

[3] In this case the spectral densities conrresponds to $\mathbb{C}^m$-valued processes and the definition of $\mathcal{S}_\beta$ in (2.23) is still well-defined.

We conclude that the unique minimum point for $L_\nu(\cdot, \Delta)$ is

$$P_\nu(\Delta) := [\hat{\Sigma}_C^{-\frac{1}{\nu}} + \frac{1}{\nu} V_\Delta]^{-\nu}. \tag{3.92}$$

Similarly to the case $\nu = 1$, if we produce $\Delta^\circ \in \mathcal{L}_\nu^\varphi$ such that $P_\nu(\Delta^\circ)$ satisfies constraint (3.22), then $\hat{\Sigma}_\nu := P_\nu(\Delta^\circ)$ is the unique solution to Problem 3.18 with $\mathcal{D}_\nu$ and $\nu \in \mathbb{N}_+ \setminus \{1\}$. Such a Λ° is given by minimizing the dual functional

(3.93)

$$J_\nu(\Delta) := -L_\nu(P_\nu(\Delta), \Delta) = \frac{\nu}{\nu - 1} \mathrm{tr}[\hat{\Sigma}_C^{-\frac{1}{\nu}} + \frac{1}{\nu} V_\Delta]^{1-\nu}. \tag{3.94}$$

Theorem 3.19. *The dual problem*

$$\text{Find } \Delta \in \mathcal{L}_\nu^\varphi \text{ minimizing } J_\nu(\Delta) \tag{3.95}$$

admits a unique solution.

Proof. Firstly, note that $\mathcal{L}_\nu^\varphi$ is open and bounded. The proof follows the one of Proposition 3.10 faithfully.
Secondly, $J_\nu \in \mathcal{C}^2(\mathcal{L}_\nu^\varphi)$. In fact, its first and second variation, respectively,

$$\begin{aligned} \delta J_\nu(\Delta; \delta\Delta) &= -\mathrm{tr}[(\hat{\Sigma}_C^{-\frac{1}{\nu}} + \frac{1}{\nu} V_\Delta)^{-\nu} V_{\delta\Delta}] \\ \delta^2 J_\nu(\Delta; \delta\Delta, \delta\Delta) &= \frac{1}{\nu} \sum_{l=1}^{\nu} \mathrm{tr}[(\hat{\Sigma}_C^{-\frac{1}{\nu}} + \frac{1}{\nu} V_\Delta)^{-l} V_{\delta\Delta} (\hat{\Sigma}_C^{-\frac{1}{\nu}} + \frac{1}{\nu} V_\Delta)^{l-\nu-1} V_{\delta\Delta}] \end{aligned}$$

are continuous over $\mathcal{L}_\nu^\varphi$. Note that $\delta J_\nu(\Delta; \delta\Delta, \delta\Delta) \geq 0$. Since $V_{\delta\Delta} \neq 0$ for each $\delta\Delta \neq 0$ with $\delta\Delta \in \mathrm{Range}\, \varphi$ (as observed in the proof of Lemma 3.7) and $\hat{\Sigma}_C^{-\frac{1}{\nu}} + \frac{1}{\nu} V_\Delta > 0$, it follows that $\delta^2 J_\nu(\Delta; \delta, \delta\Delta)$ is strictly positive over $\mathcal{L}_\nu^\varphi$. Thus, J_ν is strictly convex over $\mathcal{L}_\nu^\varphi$, and the dual problem admits at most one solution.
Finally, it remains to prove the existence of such a solution. Note that $J_\nu(0) = \frac{\nu}{\nu-1} \mathrm{tr}(\hat{\Sigma}_C^{\frac{\nu-1}{\nu}})$ and we can restrict therefore the search of a minimum point to the set $\mathcal{L}^\star := \{\Delta \in \mathcal{L}_\nu^\varphi \mid J_\nu(\Delta) \leq J_\nu(0)\} \subset \mathcal{L}_\nu^\varphi$ which is bounded. Following the same lines in the proof of Theorem 3.11 it is possible to prove that $\lim_{\Delta \to \partial\mathcal{L}_\nu^\varphi} J_\nu(\Delta) = +\infty$ (the limit diverges because the exponent in (3.94) is negative). Thus, $\mathcal{L}^\star$ is a compact set (i.e. closed and bounded) and J_ν, which is continuous over $\mathcal{L}_\nu^\varphi$, admits a minimum point Δ° over $\mathcal{L}^\star$ by the Weierstrass' Theorem. ∎

Also in this case, a globally convergent matricial *Newton* algorithm for finding Δ°, similar to the one of Section 3.3.1, may be employed. Finally, the same analysis may be extended to $\mathcal{D}_{\mathrm{KL}}$. In this case, $P_{KL}(\Delta) = \mathrm{e}^{\log(\hat{\Sigma}_C) - V_\Delta}$.

3.4.1 Application to spectral estimation

In Section 2.7, we tested the features of the family of solutions Φ_ν with $\nu \in \mathbb{N}_+$ to Problem 2.1 by exploiting the knowledge of the covariance matrix Σ. Here, we consider the bivariate *bandpass* random process y of Section 2.7.2 and we consider the corresponding THREE-like spectral estimation procedure:

1. We start from a finite sequence $y(1) \dots y(N)$ extracted from a realization of the process y;
2. Fix $G(z)$ as in Section 2.7.2;
3. Set $\Psi = \frac{1}{N} \sum_{k=\tilde{N}}^{N} y(k)y(k)^T$ with $\tilde{N} < N$;
4. Feed the filters bank with the data sequence $y(1) \dots y(N)$, collect the output data $x(1) \dots x(N)$ and compute $\hat{\Sigma}_C = \frac{1}{N} \sum_{k=\tilde{N}}^{N} x(k)x(k)^T$;
5. Compute $\hat{\Sigma}_\nu \in \text{Range } \Gamma \cap \mathcal{Q}_{n,+}$ with $\nu \in \mathbb{N}_+$
6. Compute Φ_ν by solving Problem 2.1 (with $\mathcal{S}_\nu$) with the chosen Ψ and $\hat{\Sigma}_\nu^{-\frac{1}{2}} G(z)$ as filters bank.

Note that, in point 5 of the above procedure we assume that $\hat{\Sigma}_\nu = \hat{\Sigma}_{ME}$ when $\nu = 1$.

In Figure 3.7, the obtained estimates with $N = 50$ (i.e. we have considered a short-length data) are depicted. Also in this case, the peaks of the estimates are reduced by increasing ν. For the extracted sequence, the estimators for $\nu = 2$ and $\nu = 3$ appear to perform better than the one for $\nu = 1$. Finally, the same procedure can be applied to the other processes considered in Section 2.7. We conclude that the presented family of estimators Φ_ν provides a relevant tool in multivariate spectral estimation.

3.5 Generalization of the Blackman-Tukey method

The structured covariance estimation problem introduced in Section 3.2 is just an instance of a class of problems in digital signal processing where the covariance matrix of the output process of a general linear filter has to be estimated with the knowledge of the input sample data.

For the special case of linear filters $G(z)$ whose output is the state of the filter, the problem of characterizing the output covariance Σ has been addressed by Georgiou in [33] and [36]. In Section 3.3 and Section 3.4 we

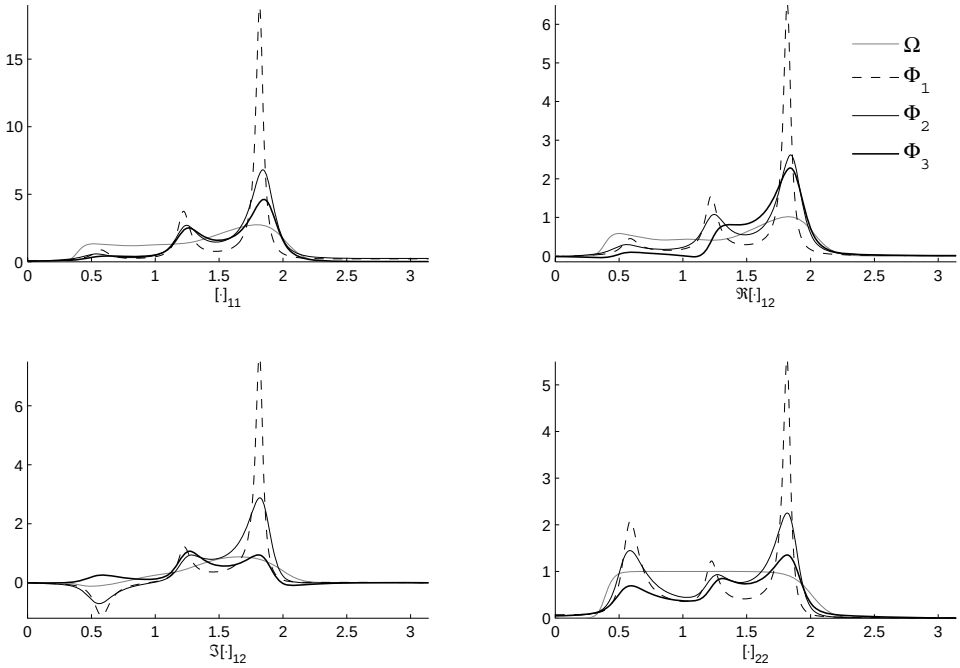


Figure 3.7: Estimation of the spectral density of a bivariate *bandpass* random process.

employed this characterization in order to estimate the state covariance by solving an optimization problem. Notice that these techniques require that the state covariance Σ and the sample covariance $\hat{\Sigma}_C$ are strictly positive definite and that the filter's output and state coincide. On the other hand, these techniques do not exploit the knowledge of $y(1) \dots y(N)$ that, in the THREE-like methods, are the problem data.
In this section we introduce a different approach, [67] (see also [44] and [30]), based on the knowledge of the input sample data $y(1) \dots y(N)$, to compute a positive semi-definite estimate $\hat{\Sigma}$ whose structure is consistent with an arbitrary, finite dimensional, stable, linear filter $G(z)$. This method, which is an extension of the one for estimating the *Toeplitz* covariance matrix of order M of the process y based on the *Blackman-Tukey estimator* [7], hinges on the characterization of Σ in terms of the filter $G(z)$ and the covariance lags sequence of the input process y. Thus, given an estimate of the covariance lags sequence of the input process, we can compute an estimate $\hat{\Sigma}$ consistent with the structure imposed by the filter. It will be shown that if we consider the sample covariance lags used in the biased correlogram spectral estimator we can guarantee that $\hat{\Sigma} \geq 0$.

In what follows, we present a more precise formulation of the problem

with $G(z)$ arbitrary, finite dimensional, stable, linear filter. We successively introduce the approach based on the covariance lags.

3.5.1 Generalized structured covariance estimation problem

Consider a linear filter

$$\begin{aligned} x_{k+1} &= Ax_k + By_k \\ w_k &= Cx_k + Dy_k, \quad k \in \mathbb{Z}, \end{aligned} \tag{3.96}$$

where $A \in \mathbb{C}^{n\times n}, B \in \mathbb{C}^{n\times m}, C \in \mathbb{C}^{p\times n}, D \in \mathbb{C}^{p\times m}$ and A has all its eigenvalues in the open unit disk. The input process y is zero mean, $\mathbb{C}^m$-valued, wide sense stationary and purely nondeterministic. $\Sigma = \Sigma^* = E[w_k w_k^*] \geq 0$ denotes the covariance matrix of the (stationary) output process w and we denote by

$$G(z) = C(zI - A)^{-1}B + D \tag{3.97}$$

the filter transfer function. We denote by $\mathcal{C}(\mathbb{T}, \mathcal{H}_m)$ the family $\mathcal{H}_m$-valued, continuous functions on the unit circle $\mathbb{T}$. Consider now the linear operator

$$\begin{aligned} \Gamma \;:\; & \mathcal{C}(\mathbb{T}, \mathcal{H}_m) \to \mathcal{H}_p \\ & \Phi \mapsto \int G\Phi G^*. \end{aligned} \tag{3.98}$$

It follows that Σ must belong to the linear space

$$\begin{aligned} \text{Range } \Gamma := \{M \in \mathcal{H}_p |\; & \exists \Phi \in \mathcal{C}(\mathbb{T}, \mathcal{H}_m) \\ & \text{such that } \int G\Phi G^* = M\}. \end{aligned} \tag{3.99}$$

Note that the above definition of Γ hinges on the filter $G(z) = C(zI - A)^{-1}B + D$. Thus, (3.98) generalizes the definition in (3.5).

Suppose now that A, B, C, D are known and a sample data $y(1) \dots y(N)$ is given. We therefore consider the following problem.

Problem 3.20. *Compute an estimate $\hat{\Sigma}$ of Σ from $y(1) \dots y(N)$ such that*

$$\hat{\Sigma} \in \text{Range } \Gamma \cap \overline{\mathcal{H}}_{p,+}. \tag{3.100}$$

If we feed $G(z)$ with the data $y(1) \dots y(N)$ and we collect the output data $w(1) \dots w(N)$, an estimate of Σ is given by the sample covariance $\hat{\Sigma}_C := \frac{1}{N}\sum_{k=1}^{N} w(k)w(k)^* \geq 0$. This estimate, as it happened in the example discussed in Section 3.1, normally fails to belong to Range Γ. In fact,

Range Γ is a linear vector subspace usually strictly contained in $\mathcal{H}_p$. One could project $\hat{\Sigma}_C$ onto Range Γ obtaining a new Hermitian matrix $\hat{\Sigma}_\Gamma$. This matrix $\hat{\Sigma}_\Gamma$, however, may be indefinite and this is particularly likely when N is not large. In addition, when the linear filter $G(z)$ does not satisfies particular properties, the computation of a basis for Range Γ is not trivial.

3.5.2 Characterization of Range Γ.

We start by considering a particular, yet very relevant, situation. We will later deal with the general case.

State covariance matrices

Next we restrict attention to the case when $C = I_n$ and $D = 0_{n\times m}$, with $m < n$, so that Σ is a state covariance matrix. Under the additional assumptions that (A, B) is a reachable pair and B has full column rank, we know that a matrix $M \in \mathcal{H}_n$ belongs to Range Γ if and only if condition (3.7) holds for some $H \in \mathbb{C}^{m\times n}$. Moreover, Range Γ has *real dimension* equal to $m(2n-m)$.

We now want to relax the reachability assumption. To this end, we derive a preliminary result. Consider an (A, B) pair and the operator Γ corresponding to $G(z) = (zI-A)^{-1}B$. We perform a state space transformation induced by an invertible matrix $T \in \mathbb{C}^{n\times n}$,

$$\tilde{A} := T^{-1}AT, \quad \tilde{B} := T^{-1}B. \tag{3.101}$$

We define the corresponding operator

$$\begin{aligned} \tilde{\Gamma} \;\; : \;\; & \mathcal{C}(\mathbb{T}, \mathcal{H}_m) \to \mathcal{H}_n \\ & \Phi \mapsto \int \tilde{G}\Phi\tilde{G}^* \end{aligned} \tag{3.102}$$

with $\tilde{G}(z) = (zI - \tilde{A})^{-1}\tilde{B} = T^{-1}G(z)$. Note that

$$\int G\Phi G^* = \int T\tilde{G}\Phi\tilde{G}^*T^*, \quad \forall \Phi \in \mathcal{C}(\mathbb{T}, \mathcal{H}_m). \tag{3.103}$$

Thus, Range Γ and Range $\tilde{\Gamma}$ are isomorphic vector spaces and

$$\tilde{M} \in \text{Range}\,\tilde{\Gamma} \;\Leftrightarrow\; T\tilde{M}T^* \in \text{Range}\,\Gamma. \tag{3.104}$$

Theorem 3.21. *Consider an (A, B) pair with B full column rank. Let $T \in \mathbb{C}^{n\times n}$ be a state space transformation such that the pair $(T^{-1}AT, T^{-1}B)$ is in standard reachability form. Let l be the dimension of the reachable subspace.*

Assume $l > m$. Then, Range Γ *has real dimension equal to* $m(2l-m)$ *and* $M \in$ Range Γ *if and only if there exists* $H_1 \in \mathbb{C}^{m\times l}$ *such that*

$$M - AMA^* = B \begin{bmatrix} H_1 & 0 \end{bmatrix} T^* + T \begin{bmatrix} H_1^* \\ 0 \end{bmatrix} B^*. \tag{3.105}$$

Proof. The proof is divided in three steps.
Step 1) By assumption, we have

$$\tilde{A} := T^{-1}AT = \begin{bmatrix} A_1 & A_{12} \\ 0 & A_2 \end{bmatrix}, \quad \tilde{B} := T^{-1}B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix} \tag{3.106}$$

where $A_1 \in \mathbb{C}^{l\times l}$, $A_{12} \in \mathbb{C}^{l\times(n-l)}$, $A_2 \in \mathbb{C}^{(n-l)\times(n-l)}$, $B_1 \in \mathbb{C}^{l\times m}$, (A_1, B_1) reachable and B_1 full column rank. Then, it is easy to see that

$$\tilde{G}(z) = (zI - \tilde{A})^{-1}\tilde{B} = \begin{bmatrix} G_1(z) \\ 0 \end{bmatrix} \tag{3.107}$$

with $G_1(z) = (zI - A_1)^{-1}B_1$. Moreover, for each $\Phi \in \mathcal{C}(\mathbb{T}, \mathcal{H}_m)$ we have

$$\int \tilde{G}\Phi\tilde{G}^* = \begin{bmatrix} \int G_1\Phi G_1^* & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.108}$$

Accordingly,

$$\text{Range}\,\tilde{\Gamma} = \left\{ \begin{bmatrix} M_1 & 0 \\ 0 & 0 \end{bmatrix} \text{ s.t. } M_1 \in \text{Range}\,\Gamma_1 \right\} \tag{3.109}$$

where

$$\begin{aligned} \Gamma_1 \;:\; & \mathcal{C}(\mathbb{T}, \mathcal{H}_m) \to \mathcal{H}_l \\ & \Phi \mapsto \int G_1\Phi G_1^*. \end{aligned} \tag{3.110}$$

It follows that Range Γ has the same dimension of Range Γ_1 and, since (A_1, B_1) is reachable and B_1 full column rank, as recalled before, such dimension is equal to $m(2l-m)$.
Step 2) Since (A_1, B_1) is reachable and B_1 full column rank, exploiting condition (3.7), we have that $\tilde{M} \in$ Range $\tilde{\Gamma}$ if and only if

$$\tilde{M} = \begin{bmatrix} M_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad M_1 \in \mathbb{C}^{l\times l} \tag{3.111}$$

and there exists $H_1 \in \mathbb{C}^{m\times l}$ such that

$$M_1 - A_1M_1A_1^* = B_1H_1 + H_1^*B_1^*. \tag{3.112}$$

The above condition is equivalent to the existence of $H_1 \in \mathbb{C}^{m\times l}$ such that

$$\begin{aligned}\tilde{M} - \tilde{A}\tilde{M}\tilde{A}^* &= \tilde{B}\begin{bmatrix} H_1 & 0 \end{bmatrix} + \begin{bmatrix} H_1^* \\ 0 \end{bmatrix}\tilde{B}^* \\ &= \begin{bmatrix} B_1H_1 + H_1^*B_1^* & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned} \tag{3.113}$$

Here, we have exploited the fact that the (unique) solution $\tilde{M}$ of the Lyapunov equation (3.113) has the block-diagonal structure (3.111) with M_1 being the solution of (3.112).
Step 3) Pre and post multiplying (3.113) by T and T^*, respectively, we see that $\tilde{M} \in \operatorname{Range}\tilde{\Gamma}$ if and only if $\exists H_1 \in \mathbb{C}^{m\times l}$ such that $M := T\tilde{M}T^*$ satisfies (3.105). Exploiting (3.104) we obtain the statement. ■

The previous theorem enables us to easily compute a basis for Range Γ also when the pair (A, B) is not reachable.

Characterization of Range Γ in the general case

We now consider a general linear filter $G(z) = C(zI - A)^{-1}B + D$ and the corresponding linear operator Γ defined in (3.98). Moreover, we define the linear operator

$$\begin{aligned}\Theta \;:\; & \mathcal{C}(\mathbb{T}, \mathcal{H}_m) \to \mathcal{H}_{n+p} \\ & \Phi \mapsto \int L\Phi L^* \end{aligned} \tag{3.114}$$

where

$$L(z) := \left(zI - \begin{bmatrix} A & 0 \\ C & 0 \end{bmatrix}\right)^{-1}\begin{bmatrix} B \\ D \end{bmatrix} = \begin{bmatrix} G_S(z) \\ z^{-1}G(z) \end{bmatrix} \tag{3.115}$$

and $G_S(z) = (zI - A)^{-1}B$.

Theorem 3.22. *$M \in$ Range Γ if and only if there exist $P \in \mathbb{C}^{n\times n}$ and $Q \in \mathbb{C}^{n\times p}$ such that*

$$X := \begin{bmatrix} P & Q \\ Q^* & M \end{bmatrix} \in \operatorname{Range}\Theta. \tag{3.116}$$

Proof. Assume that $M \in$ Range Γ, then there exists $\Phi \in \mathcal{C}(\mathbb{T}, \mathcal{H}_m)$ such that $M = \int G\Phi G^*$. Define

$$P := \int G_S\Phi G_S^*, \;\; Q := \int e^{i\vartheta}G_S\Phi G^*. \tag{3.117}$$

It follows that

$$X := \begin{bmatrix} P & Q \\ Q^* & M \end{bmatrix} = \begin{bmatrix} \int G_S \Phi G_S^* & \int e^{i\vartheta} G_S \Phi G^* \\ \int e^{-i\vartheta} G \Phi G_S^* & \int G \Phi G^* \end{bmatrix}$$
$$= \int \begin{bmatrix} G_S \\ e^{-i\vartheta} G \end{bmatrix} \Phi \begin{bmatrix} G_S^* & e^{i\vartheta} G^* \end{bmatrix} = \int L \Phi L^*. \tag{3.118}$$

Accordingly $X \in \operatorname{Range} \Theta$.
Conversely, assume that there exist P and Q such that (3.116) holds. Then there exists $\Phi \in \mathcal{C}(\mathbb{T}, \mathcal{H}_m)$ such that

$$X = \begin{bmatrix} P & Q \\ Q^* & M \end{bmatrix} = \int L \Phi L^*$$
$$= \begin{bmatrix} \int G_S \Phi G_S^* & \int e^{i\vartheta} G_S \Phi G^* \\ \int e^{-i\vartheta} G \Phi G_S^* & \int G \Phi G^* \end{bmatrix}. \tag{3.119}$$

Accordingly, $M = \int G \Phi G^*$, namely $M \in \operatorname{Range} \Gamma$. ■
Note that, $L(z)$ satisfies the hypothesis of Theorem 3.21. Accordingly, we can compute a basis for $\operatorname{Range} \Theta$.

3.5.3 Projection method in the general case

We now show how to exploit the results of Section 3.5.2 to extend the projection method considered in Section 3.2 to the general setting.
Let us first consider the situation where (A, B) may be non reachable (so that $\Sigma \geq 0$ may be singular) but still $C = I$ and $D = 0$. In view of Theorem 3.21, we can easily compute a basis for $\operatorname{Range} \Gamma$. Accordingly, we are able to compute the corresponding projected matrix $\hat{\Sigma}_\Gamma$ of $\hat{\Sigma}$. Here $\Sigma_+ \geq 0$ may be singular because we have removed the reachability condition. However, when $\hat{\Sigma}_\Gamma$ is indefinite, there always exists $\varepsilon > 0$ such that $\hat{\Sigma}_{PJ} := \hat{\Sigma}_\Gamma + \varepsilon \Sigma_+ \geq 0$ because the null space of Σ_+ coincides with the orthogonal complement of the reachable subspace of the pair (A, B).
In view of Theorem 3.22, we can now extend the projection method to the general case. Consider the linear filter $L(z)$ as in (3.115). Let v be the output process when $L(z)$ is fed by y

$$v_{k+1} = \begin{bmatrix} A & 0 \\ C & 0 \end{bmatrix} v_k + \begin{bmatrix} B \\ D \end{bmatrix} y_k, \ k \in \mathbb{Z}. \tag{3.120}$$

Define then $X := E[v_k v_k^*]$ as the corresponding output covariance matrix. We are now ready to outline the generalization of the projection method. Let $v(1) \dots v(N)$ be the output data when $L(z)$ is fed with the sample data

$y(1)\dots y(N)$. Compute then the sample matrix $\hat{X}_C := \frac{1}{N}\sum_{k=1}^{N} v(k)v(k)^*$. Notice that X is a state covariance matrix. Applying the projection method presented in Section 3.2, we obtain an estimate $\hat{X}_{PJ} \geq 0$ belonging to Range Θ. Finally, exploiting Theorem 3.22, we have

$$\hat{\Sigma}_{PJ} := \begin{bmatrix} 0 & I_p \end{bmatrix} \hat{X}_{PJ} \begin{bmatrix} 0 \\ I_p \end{bmatrix}. \tag{3.121}$$

3.5.4 Constrained covariance estimation method

In this section we first characterize the output covariance Σ in terms of the filter parameters and the covariance lags $\{R_j\}_{j=0}^{\infty}$ of y (Theorem 3.23). This enables us to define an estimate $\hat{\Sigma}_{CL}$ of the output covariance depending on an estimate $\hat{R}_j$ of the input covariance lags and to characterize the key feature $\hat{\Sigma}_{CL} \in$ Range Γ in terms of a property of the $\hat{R}_j$'s (Corollary 3.24). Finally, we present a method to compute the $\hat{R}_j$'s guaranteeing $\hat{\Sigma}_{CL} \in$ Range $\Gamma \cap \overline{\mathcal{H}}_{n,+}$.

Let $R_j := E[y_{k+j}y_k^*]$, $j \in \mathbb{Z}$, be the j-th covariance lag of y. Notice that $R_j = R_{-j}^*$.

Theorem 3.23. *Let y and w be the input and output processes of the linear filter $G(z)$ as defined in (3.97). Then, the covariance matrix of w_k is given by*

$$\Sigma = CPC^* + CQD^* + DQ^*C^* + DR_0D^* \tag{3.122}$$

where

$$Q := \sum_{j=1}^{\infty} A^{j-1}BR_j^* \tag{3.123}$$

and P is the (unique) solution of the Lyapunov equation

$$P - APA^* = AQB^* + BQ^*A^* + BR_0B^*. \tag{3.124}$$

Proof. From (3.96) we have

$$w_kw_k^* = Cx_kx_k^*C^* + Cx_ky_k^*D^* + Dy_kx_k^*C^* + Dy_ky_k^*D^*.$$

Taking expectations on both sides, we get (3.122), where $P := E[x_kx_k^*]$. Equation (3.124) follows from [36, Theorem 1]. ∎

We now define the block-Toeplitz matrix

$$T_M(R) := \begin{bmatrix} R_0 & R_{-1} & & R_{-M} \\ R_1 & \ddots & \ddots & \\ & \ddots & \ddots & R_{-1} \\ R_M & & R_1 & R_0 \end{bmatrix} \tag{3.125}$$

as the covariance matrix of order M of the process y. Notice that, $T_M(R) \geq 0$ for each $M \in \mathbb{N}$.

Corollary 3.24. *Let $\{\hat{R}_j\}_{j=0}^{\infty}$ be a sequence of $m \times m$ matrices such that $T_M(\hat{R}) \geq 0$ for each $M \in \mathbb{N}$. Define*

$$\hat{\Sigma}_{CL} := C\hat{P}C^* + C\hat{Q}D^* + D\hat{Q}^*C^* + D\hat{R}_0D^* \tag{3.126}$$

where $\hat{Q} := \sum_{j=1}^{\infty} A^{j-1}B\hat{R}_j^*$ and $\hat{P}$ is the (unique) solution to the Lyapunov equation

$$\hat{P} - A\hat{P}A^* = A\hat{Q}B^* + B\hat{Q}^*A^* + B\hat{R}_0B^*. \tag{3.127}$$

Then, $\hat{\Sigma}_{CL} \in \text{Range } \Gamma \cap \overline{\mathcal{H}}_{n,+}$.

Proof. Since $T_M(\hat{R}) \geq 0$ for each $M \in \mathbb{N}$, there exists a wide sense stationary $\mathbb{C}^m$-valued process $\hat{y}$ with covariance lags sequence $\{\hat{R}_j\}_{j=0}^{\infty}$. If we feed the filter $G(z)$ with $\hat{y}$, we get a stationary output process $\hat{w}$. In view of Theorem 3.23, it follows that the covariance matrix of $\hat{w}$ is $\hat{\Sigma}_{CL} \in \text{Range } \Gamma \cap \overline{\mathcal{H}}_{n,+}$. ■

Thus, once we have an estimate $\{\hat{R}_j\}_{j=0}^{\infty}$ of the covariance lags sequence of y satisfying $T_M(\hat{R}) \geq 0$ for each $M \in \mathbb{N}$, a positive semi-definite estimate $\hat{\Sigma}_{CL} \in \text{Range } \Gamma$ of the true covariance Σ is given by (3.126). It remains to choose a method to estimate $\{\hat{R}_j\}_{j=0}^{\infty}$ from the sample data $y(1) \dots y(N)$ in such a way that $T_M(\hat{R}) \geq 0$ for each $M \in \mathbb{N}$. We consider the correlogram spectral estimator, [61], $\hat{\Phi} = \sum_{j=-\infty}^{\infty} \hat{R}_j e^{-i\vartheta j}$ where

$$\hat{R}_j = \begin{cases} \frac{1}{N}\sum_{k=1}^{N-j} y(k+j)y^*(k), & 0 \leq j < N \\ 0_{m\times m}, & j \geq N. \end{cases} \tag{3.128}$$

This method suffers from the drawback that the reliability of the estimate $\hat{R}_j$ decreases considerably as j grows, especially for relatively short time series, [46]. The corresponding estimated joint correlation $\hat{Q}$ is, however, a finite sum. Moreover, it is easy to see that $T_M(\hat{R}) = Y_M Y_M^* \geq 0$ where $Y_M = \frac{1}{\sqrt{N}}\mathcal{C}$ with $\mathcal{C} \in \mathbb{C}^{mM\times(M-1+N)}$ being the left block-circulant (block Hankel) matrix, with m block rows, having $\begin{bmatrix} 0_{m\times 1} & \dots & \dots & 0_{m\times 1} & y(1) & \dots & y(N) \end{bmatrix}$ as the first block row. Notice that, in view of (3.122), (3.123) and (3.124), the term A^{j-1} in $\hat{\Sigma}_{CL}$ acts as "reliability index" for the estimate $\hat{R}_j$: Due to the presence of the term A^{j-1}, the influence of $\hat{R}_j$ on $\hat{\Sigma}_{CL}$ decreases as j increases. Accordingly, we can truncate the covariance lags sequence in (3.128) to L

$$\hat{R}_j = \begin{cases} \frac{1}{N}\sum_{k=1}^{N-j} y(k+j)y^*(k), & 0 \leq j < L \\ 0_{m\times m}, & j \geq L. \end{cases} \tag{3.129}$$

L is chosen in such way that $\|A^{L-1}\| < \varepsilon$, where ε is a threshold constant. Notice that (3.129) is the covariance lags sequence obtained by the *Blackman-Tukey method* [7] using a rectangular lag window of width equal to L. Thus, the corresponding block-Toeplitz matrix $T_M(\hat{R})$ is positive semi-definite for each M, see [61]. Hence, (3.129) is a natural choice for computing $\hat{\Sigma}_{CL}$.

The previous results suggest the following simple procedure, which we shall refer to as *input covariance lags method*, to compute $\hat{\Sigma}_{CL}$ given the sample data $y(1) \dots y(N)$:

1. Choose L such that $\|A^{L-1}\| < \varepsilon$
2. Compute

$$\hat{R}_0 = \frac{1}{N}\sum_{k=1}^{N} y(k)y^*(k), \quad \hat{Q} = \frac{1}{N}\sum_{j=1}^{L-1}\sum_{k=1}^{N-j} A^{j-1}By(k)y^*(k+j)$$

3. Solve in $\hat{P}$ the Lyapunov equation (3.127)
4. Compute the estimate $\hat{\Sigma}_{CL}$ of the true covariance Σ using (3.126).

Note that, in the covariance extension setting (1.7) we have

$$\hat{\Sigma}_{CL} = \frac{1}{N}\begin{bmatrix} \dots & \dots & 0 & y(1) & \dots & y(L) \\ \dots & 0 & y(1) & \dots & y(L) & 0 \\ \iddots & \iddots & \iddots & \iddots & \iddots & \iddots \end{bmatrix} \begin{bmatrix} \vdots & \vdots & \iddots \\ \vdots & 0 & \iddots \\ 0 & y^*(1) & \iddots \\ y^*(1) & \vdots & \iddots \\ \vdots & y^*(L) & \iddots \\ y^*(L) & 0 & \iddots \end{bmatrix}$$

which coincides with the one given by the Blackman-Tukey method. Thus, the above method generalizes the Blackman-Tukey approach.

3.5.5 Performance comparison

In this section, we want to test the method presented in Section 3.5.4 with the other methods introduced in the Chapter. We use the following notation:
• *CL method* to denote the input covariance lags method.
• *PJ method* to denote the extended projection method presented in Section 3.5.3.

• *ME method* to denote the maximum entropy method (only employed to estimate state covariance matrices).
For a fair interpretation of the comparison results, we hasten to point out that while other methods exploits only a finite sample of the output of the linear filter, our method uses only the corresponding sample of the input process. Notice that in the case of estimate of state covariance matrices (and assuming the matrix B to be full column rank) the sample of the input process are easily obtainable from that of the output process and the converse is also true. Therefore, the available information is really the same for the three methods. On the contrary, for general filters the available information may be different. Notice also that in the applications related with THREE-like estimation methods the available data are a finite sample of the input process.

A performance comparison procedure

Suppose that we have a finite sequence $y(1)\dots y(N)$ extracted from a sample path of a zero-mean, weakly stationary discrete-time process y. We want to compare the estimates $\hat{\Sigma}_{CL}, \hat{\Sigma}_{PJ}, \hat{\Sigma}_{ME}$ obtained by employing CL, PJ and ME method, respectively. In order to make the comparison reasonably independent of the specific data set, we average over 500 experiments performed with sequences extracted from different sample paths. We are now ready to describe the comparison procedure:
• Fix the transfer function $G(z)$.
• At the j-th experiment $G(z)$ is fed by the data $y^j(1)\dots y^j(N)$. From $y^j(1)\dots y^j(N)$ estimate $\hat{\Sigma}_{CL}(j)$, $\hat{\Sigma}_{PJ}(j)$ and $\hat{\Sigma}_{ME}(j)$ using CL, PJ and ME method respectively.
• Compute the relative error norm[4] between Σ and the estimate $\hat{\Sigma}_{CL}(j)$

$$e_{CL}(j) = \frac{\|\hat{\Sigma}_{CL}(j) - \Sigma\|}{\|\Sigma\|}. \tag{3.130}$$

In similar way, compute the relative error norms $e_{PJ}(j)$ and $e_{ME}(j)$ between Σ and the estimates $\hat{\Sigma}_{PJ}(j)$ and $\hat{\Sigma}_{ME}(j)$ respectively.
• Once completed the experiments, compute the means $\mu_{CL}, \mu_{PJ}, \mu_{ME}$ and the variances $\sigma^2_{CL}, \sigma^2_{PJ}, \sigma^2_{ME}$ of the corresponding sequences $\{e_{CL}(j)\}_{j=1}^{500}$, $\{e_{PJ}(j)\}_{j=1}^{500}$, $\{e_{ME}(j)\}_{j=1}^{500}$. For example, for the CL method:

$$\mu_{CL} = \frac{1}{500}\sum_{j=1}^{500} e_{CL}(j),\ \sigma^2_{CL} = \frac{1}{500}\sum_{j=1}^{500}(e_{CL}(j) - \mu_{CL})^2.$$

[4]Here the norm $\|\cdot\|$ is the spectral norm i.e. the matrix norm induced by the Euclidean norm in $\mathbb{C}^p$

N	μ_{CL}	μ_{PJ}	σ^2_{CL}	σ^2_{PJ}	$\sharp F$
300	0.1360	0.4130	0.0089	2.0937	174
500	0.1016	0.2127	0.0045	0.5544	126
700	0.0865	0.1893	0.0031	0.8592	97

Table 3.6: Parameters μ_{CL}, μ_{PJ}, σ^2_{CL}, σ^2_{PJ}, $\sharp F$ for $G(z)$ considered in Subsection 3.5.5.

• Count the number $\sharp F$ of times that the PJ method adjusts the estimate $\hat{X}_\Gamma$ by adding the quantity εX_+. Notice that the ME method can be only used when Σ is a state covariance matrix (and not in the general case). For the sake of comparison, we consider the parameters μ_i, σ_i^2 and $\sharp F$. Clearly, the smaller these parameters, the better estimation is expected.

Simulation results: The general case

We have considered a bivariate real process y with a high-order spectral density $\Omega(z)$ and a filter $G(z)$ with a 3-dimensional output with 4 poles equispaced on the circle of radius 0.8. The true covariance matrix Σ is positive definite with eigenvalues: $\lambda_1 = 3.12 \cdot 10^4$, $\lambda_2 = 1.15 \cdot 10^2$, $\lambda_3 = 3.33 \cdot 10^2$. The corresponding error means and variances for PJ and CL method are reported in Table 3.6 for different values of the length N of the observed data sequences $y(1) \dots y(N)$. It is clear that the CL method largely outperform the PJ method. The heuristic reason follows. As noted in Section 3.3.2, the projection of $\hat{X}_C$ (that is a perturbed version of the state covariance X) onto Range Θ yields a matrix $\hat{X}_\Theta$ that, in many cases, in particular when N is small, fails to be positive definite (or even positive semi-definite). This, explains why the number of failures $\sharp F$ is significant. Moreover, when $\hat{X}_\Theta$ is indefinite the projection method add it the positive definite matrix $X_+ \in$ Range Θ. For each experiment, X_+ is the same. In view of (3.121), the adjustment cannot be expected to provide a good estimate of $\hat{\Sigma}_{PJ}$. Note that μ_{PJ}, σ^2_{PJ} decrease as N increases: In fact, $\hat{X}_C \to X$ with probability one as $N \to \infty$. Notice that also μ_{CL} and σ^2_{CL} decrease as N grows. Indeed, each $\hat{R}_j$ approaches the true covariance lag R_j as $N \to \infty$. Accordingly $\hat{\Sigma}_{CL} \to \Sigma$. Moreover, each estimate $\hat{\Sigma}_{CL}$ is positive definite. We conclude that the CL method is remarkably preferable to the PJ method.

N	μ_{CL}	μ_{PJ}	μ_{ME}	σ^2_{CL}	σ^2_{PJ}	σ^2_{ME}	$\sharp F$
300	0.18	0.81	0.18	0.018	2.65	0.02	73
500	0.16	0.47	0.15	0.013	1.37	0.013	37
700	0.13	0.29	0.13	0.001	0.74	0.009	18

Table 3.7: Parameters μ_{CL}, μ_{PJ}, μ_{ME}, σ^2_{CL}, σ^2_{PJ}, σ^2_{ME}, $\sharp F$ for $G(z)$ considered in Subsection 3.5.5.

Simulation results: State covariance estimation

Consider $G(z)$ corresponding to $C = I_6,\ D = 0_{6\times 2}$,

$$A = \begin{bmatrix} 0.6 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0.6 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.7 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0.7 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0.7 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

We choose the bivariate real process y with a coercive high-order spectral density $\Omega(z)$ considered in Section 3.3.2. The true covariance Σ is positive definite with eigenvalues: $\lambda_1 = 3.4 \cdot 10^{-3}$, $\lambda_2 = 1.69 \cdot 10^{-2}$, $\lambda_3 = 1.47$, $\lambda_4 = 2.92$, $\lambda_5 = 1.18 \cdot 10$, $\lambda_6 = 1.59 \cdot 10^2$. In Table 3.7, we present the results obtained for different lengths N of the observed sequences $y(1) \dots y(L)$. CL and ME methods provide quite similar performances. The PJ method provides bad estimates when N is small. In this situation, the PJ method must adjust the projection $\hat{\Sigma}_C$ onto Range Γ in many experiments. Accordingly, its performance gets remarkably poor with respect to the other methods when N is not large. Also in this case each estimate $\hat{\Sigma}_{CL}$ is positive definite.

Remark 3.25. As for the computational burden, the PJ method described in Section 3.5.3 normally compensates for the poor performances with a very high numerical efficiency. The ME and CL methods are very hardly comparable. In fact, the number of operations of the ME and CL methods is highly dependent on the problem's parameters. Moreover, the ME method is an optimization procedure whose computational burden is also dependent on the tolerance threshold fixed for the convergence of the algorithm. On the other hand the CL does not require any optimization procedure. For example, in the cases illustrated above the two methods perform very similarly also with respect to the computational burden (while the PJ method is much faster). On the other hand, extensive simulation shows that the ME method presents numerical problems and leads to extremely slow convergence, if we consider a case when the state covariance Σ is close to singularity. The CL

method, on the contrary, does not require any optimization procedure and does not present any of these problems.

To conclude, the CL approach hinges on an explicit representation of Σ in terms of the given filter and the covariance lags sequence of the input process. Not only the estimated matrix was shown to be positive semi-definite, but extensive simulation suggests also that the estimate is strictly positive definite with high probability when $\Sigma > 0$.

Chapter 4

Quantum process tomography

4.1 Introduction

In this Chapter, we consider an *identification problem arising in the reconstruction of quantum dynamical models from experimental data*. This is a key issue in many quantum information processing tasks [53],[10], [54],[50], [9]. For example, a precise knowledge of the behavior of a channel to be used for quantum computation or communications is needed in order to ensure that optimal encoding/decoding strategies are employed, and verify that the noise thresholds for hierarchical error-correction protocols, or for effectiveness of quantum key distribution protocols, are met [53],[10]. In many cases of interest, for example in free-space communication [63], channels are not stationary and to ensure good performances, repeated and fast estimation steps would be needed as a prerequisite for adaptive encodings. In addition to this, when the goal is to embed the system used for probing the channel in a moving vehicle or a satellite, one seeks the simplest implementation, or at least a compromise between estimation accuracy and the number of experimental resources needed.

To this aim, we here focus on: (i) characterizing the minimal experimental setting (in terms of available probe states and measured observables) needed for a consistent estimation of the channel; (ii) exploring how a minimal parametrization of the models can be exploited to reduce the complexity of the estimation algorithm; and (iii) simulating the *minimal experimental setting*, and comparing it to "richer" experimental resources. In doing this, we present a general framework for the estimation of physically-admissible trace preserving quantum channels by minimizing a suitable class of (convex) loss functions which contains, as special cases, commonly used maximum likelihood (ML) functionals. In the large body of literature regarding

channel estimation, or *quantum process tomography* (see e.g. [54],[50] and references therein), the experimental resources are usually assumed to be given. Mohseni et al. [50] compare different strategies, but focus on the role of having entangled states as an additional resource. However, it can be argued (see e.g. [65]) that the information acquired through an entanglement-assisted method by measuring a certain number K of independent observables can be equivalently gathered by the method we describe, by properly choosing L probe states and M observables such that $L \cdot M = K$. Accordingly, our result on the minimal resources for TPCP map estimation can be easily adapted to this approach. The problem we study is close in spirit to that taken in [62] while studying minimal *state* tomography.
Our analysis of the problem and the standard tomography methods (including "inversion" and ML methods) leads to a necessary and sufficient condition for *identifiability* of the channel and to the characterization of the minimal experimental resources (or *quorum*, in the language of [24]) for *Trace-Preserving (TP) channels estimation*, [68],[69]. While the existing ML approaches introduce the TP constraint through a Lagrange multiplier [54],[50], the method we propose constrains the set of channels of the optimization problem to TP maps from the beginning. In a d level quantum system, this allow for an immediate reduction from d^4 to $d^4 - d^2$ free parameters in the estimation problem. However, determining which conditions on probe states and output measurements must be satisfied to ensure identifiability has not been made explicit so far. Our analysis can also be considered as complementary to the one presented in [8], where the TP assumption is relaxed to include losses. We pursue a rigorous presentation of the results and we try, whenever possible, to make contact with ideas and methods of (classical) system identification. The same explicit parametrization for TP channels is also used to develop a Newton-type algorithm with barriers, which ensures convergence in the set of physically-admissible maps. Numerical simulation evidences that experimental settings richer than the minimal one do not lead to better performances, once the total number of available "trials" is fixed.

4.2 Preliminaries: Quantum channels and χ-representation

As explained in Section 1.2.2, a quantum channel (in Schrödinger's picture) is a map $\mathcal{E} : \mathfrak{D}(\mathbb{H}) \to \mathfrak{D}(\mathbb{H})$ with $\mathfrak{D}(\mathbb{H}) = \{\rho \in \overline{\mathcal{H}}_{n,+} | \mathrm{tr}(\rho) = 1\}$. Moreover, a physically admissible quantum channel must be linear and *Completely Posi-*

tive (CP), namely it must admit an Operator-Sum Representation (OSR)

$$\mathcal{E}(\rho) = \sum_{j=1}^{d^2} K_j \rho K_j^* \tag{4.1}$$

where $K_i \in \mathbb{C}^{d\times d}$ are called *Kraus operators*, [47]. In order to be *Trace Preserving* (TP), a necessary condition to map states to states, it must also hold that

$$\sum_{j=1}^{d^2} K_j^* K_j = I_d \tag{4.2}$$

where I_d is the $d\times d$ identity matrix. TPCP maps can be thought as the quantum equivalent of Markov transition matrices in the classical setting. An alternative way to describe a CPTP channel is offered by the χ-*representation*. Each Kraus operator $K_j \in \mathbb{C}^{d\times d}$ can be expressed as a linear combination (with complex coefficients) of $\{F_m\}_{m=1}^{d^2}$, F_m being the elementary matrix E_{jk}, (whose entries are all zero except the one in position jk which is 1) with $m = (j-1)d + k$. Accordingly, the OSR (4.1) can be rewritten as

$$\mathcal{E}(\rho) = \sum_{m,n=1}^{d^2} \chi_{m,n} F_m \rho F_n^*. \tag{4.3}$$

Let χ be the $d^2 \times d^2$ matrix with element $\chi_{m,n}$ in position (m,n). It is easy to see that it must satisfy

$$\chi = \chi^* \geq 0 \tag{4.4}$$

and (following from (4.2))

$$\sum_{m,n=1}^{d^2} \chi_{m,n} F_n^* F_m = I_d. \tag{4.5}$$

The map $\mathcal{E}$ is completely determined by the matrix χ. The χ matrix can be used directly to calculate the effect of the map on a given state, and the probability measurements outcomes, as well as observable expectations[1]. Before providing the explicit formulas in the next lemmas we need to recall the definition of *partial trace*. Consider two finite-dimensional vector spaces

[1]These results implicitly relate the χ matrix emerging from the basis of elementary matrices we chose to the Choi matrix $C_{\mathcal{E}} = \sum_{mn} E_{mn} \otimes \mathcal{E}(E_{mn})$[55]. In fact, either by direct computation or by confronting formula (4.8) with its equivalent for the Choi matrix $C_{\mathcal{E}}$ (see e.g. [54], chapter 2), it is easy to see that $C_{\mathcal{E}} = O\chi O^*$, where O is the unique unitary such that $O(X \otimes Y)O^* = Y \otimes X$ [6].

$\mathcal{V}_1\,\mathcal{V}_2$, with $\dim \mathcal{V}_1 = n_1$, $\dim \mathcal{V}_2 = n_2$. Let us denote by $\mathcal{M}_n$ the set of complex matrices of dimension $n \times n$. Let $\{M_j\}$ be a basis for $\mathcal{M}_{n_1}$, and $\{N_j\}$ be a basis for $\mathcal{M}_{n_2}$, representing linear maps on $\mathcal{V}_1$ and $\mathcal{V}_2$, respectively. Consider $\mathcal{M}_{n_1\times n_2} = \mathcal{M}_{n_1} \otimes \mathcal{M}_{n_2}$: It is easy to show that the $n_1^2 \times n_2^2$ linearly independent matrices $\{M_j \otimes N_k\}$ form a basis for $\mathcal{M}_{n_1\times n_2}$, where $\otimes$ denotes the Kronecker product. Thus, one can express any $X \in \mathcal{M}_{n_1\times n_2}$ as

$$X = \sum_{jk} c_{jk} M_j \otimes N_k.$$

The *partial trace* over $\mathcal{V}_2$ is the linear map

$$\begin{aligned} \mathrm{tr}_{\mathcal{V}_2} \quad : \quad & \mathcal{M}_{n_1\times n_2} \to M_{n_1} && (4.6)\\ & X \mapsto \mathrm{tr}_{\mathcal{V}_2}(X) := \sum_{j,k} (c_{jk}\mathrm{tr}(N_k)) M_j. && (4.7) \end{aligned}$$

An analogous definition can be given for the partial trace over $\mathcal{V}_1$. If the two vector spaces have the same dimension, $n_1 = n_2$, we will indicate with tr_1 and tr_2 the partial traces over $\mathcal{V}_1$ and $\mathcal{V}_2$, respectively. The partial trace can be also implicitly defined (without reference to a specific basis) as the only linear function such that for any pair $X \in \mathcal{M}_{n_1}$, $Y \in \mathcal{M}_{n_2}$:

$$\mathrm{tr}_{\mathcal{V}_2}(X \otimes Y) = \mathrm{tr}(Y)X.$$

By linearity, this clearly implies

$$\mathrm{tr}((A \otimes I)B) = \mathrm{tr}(A\,\mathrm{tr}_2(B)).$$

Lemma 4.1. *Let $\mathcal{E}_\chi$ be a CPTP map associated with a given χ. Then for any $\rho \in \mathfrak{D}(\mathbb{H})$*

$$\mathcal{E}_\chi(\rho) = \mathrm{tr}_2(\chi(I_d \otimes \rho^T)). \tag{4.8}$$

Proof. Let us rewrite each F_j as the corresponding elementary matrix E_{lm}, with $j = (l-1)d+m$, $k = (n-1)d+p$, and relabel χ_{jk} as $\hat{\chi}_{lmnp}$ accordingly. Hence we get

$$\chi = \sum_{l,m,n,p} \hat{\chi}_{lmnp} E_{ln} \otimes E_{mp}, \tag{4.9}$$

and

$$\mathcal{E}_\chi(\rho) = \sum_{l,m,n,p} \hat{\chi}_{lmnp} E_{lm} \rho E_{pn}.$$

We can also expand $\rho = \sum_{rs} \rho_{rs} E_{rs}$, and substitute it in the above expression. Taking into account that $E_{lm} E_{np} = \delta_{mn} E_{lp}$, and $\mathrm{tr}(\rho^T E_{rs}) = \rho_{rs}$, we get:

$$\begin{aligned}
\mathcal{E}_\chi(\rho) &= \sum_{l,m,n,p} \hat{\chi}_{lmnp} E_{lm} \rho E_{np}^* = \sum_{l,m,n,p,r,s} \rho_{rs} \hat{\chi}_{lmnp} E_{lm} E_{rs} E_{pn} \qquad (4.10) \\
&= \sum_{l,n,r,s} \rho_{rs} \hat{\chi}_{lrns} E_{ln} = \sum_{l,n,r,s} \hat{\chi}_{lrns} \mathrm{tr}(\rho^T E_{rs}) E_{ln} \\
&= \mathrm{tr}_2 \left(\sum_{l,n,r,s} \hat{\chi}_{lrns} E_{ln} \otimes E_{rs} \rho^T \right) = \mathrm{tr}_2 \left(\sum_{l,n,r,s} \hat{\chi}_{lrns} (E_{ln} \otimes E_{rs})(I \otimes \rho^T) \right) \\
&= \mathrm{tr}_2(\chi(I \otimes \rho^T)). \qquad (4.11)
\end{aligned}$$

■

This leads to a useful expression for the computation of the expectations.

Corollary 4.2. *Let us consider a state ρ, a projector Π and a quantum channel $\mathcal{E}$ with associated a χ-representation matrix χ. Then*

$$p_{\mathcal{E}(\rho)}(\Pi) = \mathrm{tr}(\mathcal{E}(\rho)\Pi) = \mathrm{tr}(\chi(\Pi \otimes \rho^T)).$$

Proof. It suffices to substitute (4.8) in $p_{\chi,\rho}(\Pi) = \mathrm{tr}(\mathcal{E}(\rho)\Pi)$, and use the identity $\mathrm{tr}((X \otimes I)Y) = \mathrm{tr}(X \mathrm{tr}_2(Y))$. ■

The TP condition (4.5) can also be re-expressed directly in terms of the χ matrix.

Corollary 4.3. *Let us consider a CP map $\mathcal{E}_\chi$ with associated χ-representation matrix χ. Then $\mathcal{E}_\chi$ is TP if and only if*

$$\mathrm{tr}_1(\chi) = I_d. \qquad (4.12)$$

Proof. Using the same notation we used in the proof of Lemma 4.1, we can re-espress the TP condition (4.5) as:

$$I_d = \sum_{l,m,n,p} \hat{\chi}_{lmnp} E_{pn} E_{lm} = \sum_{l,m,p} \hat{\chi}_{lmlp} E_{pm} = \mathrm{tr}_1(\chi).$$

■

4.3 Main Results: Identifiability Condition and Minimal Setting

4.3.1 The Channel Identification Problem

Consider the following setting: A quantum system prepared in a *known pure state* ρ [2] is fed to an unknown channel $\mathcal{E}$. The system in the *output state* $\mathcal{E}(\rho)$ is then subjected to a projective measurement of an *observable*. By noting that an observable (being represented by an Hermitian matrix in our setting) admits a decomposition in orthogonal projections representing mutually incompatible quantum events, we can without loss of generality restrict ourselves to consider measurements associated to orthogonal projections $\Pi = \Pi^* = \Pi^2$. For each one of these, the outcome x is in the set $\{0, 1\}$, and can be interpreted as a sample of the (classical) random variable X which has distribution

$$P_{\chi(x),\rho} = \begin{cases} p_{\chi,\rho}(\Pi), & \text{if } x = 1 \\ 1 - p_{\chi,\rho}(\Pi), & \text{if } x = 0 \end{cases} \tag{4.13}$$

where $p_{\chi,\rho}(\Pi) = \mathrm{tr}(\mathcal{E}_\chi(\rho)\Pi)$ is the probability that the measurement of Π returns outcome 1 when the state is $\mathcal{E}_\chi(\rho)$.

Assume that the experiment is repeated with a series of known input (pure) states $\{\rho_k\}_{k=1}^L$, and to each trial the same orthogonal projections $\{\Pi_j\}_{j=1}^M$ are measured N times, obtaining a series of outcomes $\{x_l^{jk}\}$. We consider the sampled frequencies to be our *data*, namely

$$f_{jk} := \frac{1}{N} \sum_{l=1}^{N} x_l^{jk}. \tag{4.14}$$

The channel identification problem we are concerned with consists in constructing a *Kraus map* $\mathcal{E}_{\hat{\chi}}$ that fits the experimental data (in some optimal way), in particular estimating a matrix $\hat{\chi}$ satisfying constraints (4.4), (4.5).

4.3.2 Necessary and sufficient conditions for identifiability

It is well known [58],[54] that by imposing linear constraints associated to the TP condition (4.5), or equivalently (4.12), one reduces the d^4 real degrees of freedom of χ to $d^4 - d^2$. This will be made explicit in the following, by

[2] A state is called *pure* if ρ is an orthogonal projection matrix on a one-dimensional subspace.

parameterizing χ in a "generalized" Pauli basis (also known as gell-mann matrices, Fano basis or coherence vector representation in the case of states [3], [5],[54]). Usually the trace preserving constraint is not directly included in the standard tomography method [50], since in principle it should emerge from the physical properties of the channel, or it is imposed through a (non-linear) Lagrange multiplier in the maximum likelihood approach [54]. Here, in order to investigate the minimum number of probe (input) states and measured projectors needed to uniquely determine χ, it is convenient to include this constraint from the very beginning. Doing so, we lose the possibility of exploiting a Cholesky factorization in order to impose positive semidefiniteness of χ: Noentheless, we show in Section 4.4 that semidefiniteness of the solution can be imposed algorithmically by using a barrier method [11].

Before proceeding to the main results, a number of definitions are in order. Consider an orthonormal basis for $d^2 \times d^2$ Hermitian matrices of the form $\{\sigma_j \otimes \sigma_k\}_{j,k=0,1,\ldots,d^2-1}$, where $\sigma_0 = 1/\sqrt{d} I_d$, while $\{\sigma_j\}_{j=1,\ldots,d^2-1}$ is a basis for the traceless subspaces. We can now write

$$\chi = \sum_{jk} s_{jk} \sigma_j \otimes \sigma_k.$$

If we now substitute it into (4.12), we get:

$$I_d = \mathrm{tr}_1(\chi) = \sum_{jk} s_{jk} \mathrm{tr}(\sigma_j) \sigma_k = \sum_k \sqrt{d}\, s_{0k} \sigma_k,$$

and hence, since the σ_j are linearly independent, we can conclude that $s_{00} = 1$, $s_{0j} = 0$ for $j = 1, \ldots, d^2 - 1$. Hence, the free parameters for a TP map (at this point not necessarily CP, since we have not imposed the positivity of χ yet) are $d^4 - d^2$, and we can write any TP χ as $\chi = d^{-1} I_{d^2} + \sum_{j=1,k=0}^{d^2-1,d^2-1} s_{jk} \sigma_j \otimes \sigma_k$, or, in a more compact notation,

$$\chi(\underline{\theta}) = d^{-1} I_{d^2} + \sum_{\ell=1}^{d^4-d^2} \theta_\ell Q_\ell, \tag{4.15}$$

by rearranging the double indexes j, k in a single index ℓ, and defining the corresponding $Q_\ell = \sigma_j \otimes \sigma_k$. Thus, there exists a one to one correspondence among χ and the $d^4 - d^2$-dimensional real vector $\underline{\vartheta} = \begin{bmatrix} \vartheta_1 & \ldots & \vartheta_{d^4-d^2} \end{bmatrix}^T$, and the χ matrices corresponding to TP maps form an affine space, its linear part being

$$\mathcal{S}_{TP} := \mathrm{span}\,\{Q_\ell\} = \mathrm{span}\,\{\sigma_j \otimes \sigma_k\}_{j=1,\ldots,d^2-1,k=0,\ldots,d^2-1}\,.$$

In order to find necessary and sufficient conditions for identifiability, it is convenient to define

$$B_{jk} = (\Pi_j - \frac{r_j}{d} I) \otimes \rho_k^T \tag{4.16}$$

where r_j is the rank of Π_j and $\mathrm{tr}(\Pi_j) = r_j$. Moreover, we define $\mathcal{B} = \mathrm{span}\{B_{jk}\}_{j=1,\ldots,M,k=1,\ldots,L}$. Intuitively, $\mathcal{B}$ represents the space of input/output combination that can be probed by the set of experimental resources $\{\rho_k\}, \{\Pi_j\}$ we choose. The definition of the B_{jk} is motivated by the fact that, since $Q_\ell = \sigma_{j\neq 0} \otimes \sigma_k$, it holds that

$$\mathrm{tr}(Q_\ell(\Pi_j \otimes \rho_k^T)) = \mathrm{tr}(Q_\ell B_{jk}). \tag{4.17}$$

By recalling that $\sigma_j, j = 1, \ldots d^2 - 1$ is a basis for the traceless subspace of Hermitian matrices it is immediate to show that $\mathcal{B} \subseteq \mathcal{S}_{TP}$. Finally, let us introduce the function g that maps the space of TP channels in the (theoretical) set of probabilities for the input states/measured projectors combinations:

$$\begin{aligned} g \;:\; & \mathbb{R}^{d^4-d^2} \to \mathbb{R}^{M\times L} \\ & \underline{\vartheta} \mapsto g(\underline{\vartheta}) \end{aligned} \tag{4.18}$$

where the component of $g(\underline{\vartheta})$ in position (j,k) is defined as

$$g_{jk}(\underline{\vartheta}) = p_{\chi(\underline{\vartheta}),\rho_k}(\Pi_j) = \mathrm{tr}(\chi(\underline{\vartheta})(\Pi_j \otimes \rho_k^T)). \tag{4.19}$$

The key result on identifiability is the following:

Proposition 4.4. *g is injective if and only if $\mathcal{S}_{TP} = \mathcal{B}$.*

Proof. Given (4.19), we have that

$$\begin{aligned} g_{jk}(\underline{\vartheta}_1) - g_{jk}(\underline{\vartheta}_2) &= \mathrm{tr}[(\chi(\underline{\vartheta}_1) - \chi(\underline{\vartheta}_2))(\Pi_j \otimes \rho_k^T)] && (4.20) \\ &= \mathrm{tr}[S(\underline{\vartheta}_1 - \underline{\vartheta}_2) B_{jk}] \\ &= \langle S(\underline{\vartheta}_1 - \underline{\vartheta}_2), B_{jk}\rangle && (4.21) \end{aligned}$$

where $S(\underline{\vartheta}_1 - \underline{\vartheta}_2) = \chi(\underline{\vartheta}_1) - \chi(\underline{\vartheta}_2) = \sum_{l=1}^{d^4-d^2} (\vartheta_{1,l} - \vartheta_{2,l}) Q_l \in \mathcal{S}_{TP}$. So, we have that

$$g(\underline{\vartheta}_1) = g(\underline{\vartheta}_2) \;\Leftrightarrow\; \langle S(\underline{\vartheta}_1 - \underline{\vartheta}_2), B_{jk}\rangle = 0 \;\; \forall\, j,k. \tag{4.22}$$

Assume $\mathcal{S}_{TP} = \mathcal{B}$: The only element of $\mathcal{S}_{TP}$ for which the r.h.s. of (4.22) could be true is zero. Since by definition $S(\underline{\vartheta}_1 - \underline{\vartheta}_2) = 0$ if and only if $\underline{\vartheta}_1 = \underline{\vartheta}_2$, g is injective. On the other hand, assume that $\mathcal{B} \subsetneq \mathcal{S}_{TP}$: Therefore there exists $T \neq 0 \in \mathcal{S}_{TP} \bigcap \mathcal{B}^\perp$ such that

$$T = \sum_\ell \gamma_\ell Q_\ell, \quad \langle T, B_{jk}\rangle = 0 \;\forall j,k.$$

But this would mean that $\underline{\vartheta}$ and $\underline{\vartheta} + \underline{\gamma}$ have the same image $g(\underline{\vartheta})$, and hence g is not injective. ■

We anticipate here that g being injective is a necessary and sufficient condition for *a priori* identifiability of χ, and thus for having a unique solution of the problem for both inversion (standard process tomography) and convex optimization-based (e.g. maximum likelihood) methods, up to some basic assumptions on the cost functional. The proof of these facts is given in full detail in Section 4.3.3 and 4.3.4.

As a consequence of these facts, we can determine the *minimal experimental resources*, in terms of input states and measured projectors, needed for faithfully reconstructing χ from noiseless data $\{f^{\circ}_{jk}\}$, where $f^{\circ}_{jk} = p_{\chi,\rho_k}(\Pi_j)$. In the light of Proposition 4.4, the minimal experimental setting is characterized by a choice of $\{\Pi_j, \rho_k\}$ such that $\mathcal{S}_{TP} = \mathcal{B}$. Recalling the definition of $\mathcal{B}$, through (4.16), it is immediate to see that $\mathcal{S}_{TP} = \mathcal{B}$ if and only if $\text{span}\left\{\Pi_j - \frac{r_j}{d} I_d\right\} = \text{span}\left\{\sigma_j, j = 1, \ldots, d^2 - 1\right\}$ and $\text{span}\left\{\rho_k\right\} = \mathcal{H}_d$. We can summarize this fact as a corollary of Proposition 4.4.

Corollary 4.5. *g is injective if and only if we have* at least *d^2 linearly independent input states $\{\rho_k\}$, and $d^2 - 1$ measured $\{\Pi_j\}$ such that*

$$\text{span}\left\{\Pi_j - \frac{r_j}{d} I_d\right\} = \text{span}\left\{\sigma_j, j = 1, \ldots, d^2 - 1\right\}.$$

We call such a set a *minimal experimental setting*. Notice that, using the terminology of [54],[24], the minimal *quorum* of observables consists of $d^2 - 1$ properly chosen elements. While in most of the literature at least d^2 observables are considered [31],[50], we showed it is in principle possible to spare a measurement channel at the output. A physically-inspired interpretation for this fact is that, since we *a priori* know, or assume, that the channel is TP, measuring the component of the observables along the identity does not provide useful information. This is clearly not true if one relaxes the TP condition, as it has been done in [8]: In that case, by the same line of reasoning, d^2 linearly independent observables are the necessary and sufficient for g to be injective.

As an example relevant to many experimental situation, consider the qubit case, i.e. $d = 2$. A minimal set of projector has to span the traceless subspace of $\mathcal{H}_2$: one can choose e.g.:

$$\Pi_j = \frac{1}{2}(I_2 + \sigma_j), \; j = x, y, z.$$

$$\rho_{x,y} = \frac{1}{2}(I_2 + \sigma_{x,y}), \quad \rho_{\pm} = \frac{1}{2} I_2 \pm \sigma_z. \tag{4.23}$$

It is clear that there is an asymmetry between the role of output and inputs: In fact, exchanging the number of $\{\Pi_j\}$ and $\{\rho_k\}$ can not lead to an injective g.

4.3.3 Process Tomography by inversion

Assume that $\mathcal{S}_{TP} = \mathcal{B}$, and that the data $\{f_{jk}\}$ have been collected. Since f_{jk} is an estimate of $p_{\chi(\underline{\vartheta}),\rho_k}(\Pi_j)$, consider the following least mean square problem

$$\min_{\underline{\vartheta}\in\mathbb{R}^{d^4-d^2}} \|\underline{g}(\underline{\vartheta}) - \underline{f}\| \tag{4.24}$$

where $\underline{g}(\underline{\vartheta})$ and $\underline{f}$ are the vectors obtained by stacking the $g_{jk}(\underline{\vartheta})$ and f_{jk} $j = 1 \dots M$, $k = 1 \dots L$, respectively. In view of (4.15) and (4.19) we have that $\underline{g}(\underline{\vartheta}) = T\underline{\vartheta} + d^{-1}\underline{r}$ where

$$T = \begin{bmatrix} \ddots & \vdots & \\ & \mathrm{tr}(B_{jk}Q_\ell) & \\ & \vdots & \ddots \end{bmatrix} \tag{4.25}$$

and

$$\underline{r} = \begin{bmatrix} r_1 & \dots & r_M \end{bmatrix}^T. \tag{4.26}$$

Notice that the ℓth column of T is formed with the inner products of Q_ℓ with each B_{jk}. Since $\mathcal{S}_{TP} = \mathcal{B}$, the Q_ℓ are linearly independent and the B_{jk} are the generators of $\mathcal{B}$, then T is *full column rank*, namely has rank $d^4 - d^2$. Hence, in principle, one can reconstruct $\hat{\vartheta}$ as

$$\underline{\hat{\vartheta}} = T^{\#}(\underline{f} - \frac{1}{d}\underline{r}), \tag{4.27}$$

$T^{\#}$ being the Moore-Penrose pseudo inverse of T [42]. If the experimental setting is minimal, the usual inverse suffices. However, as it is well known, when computing χ this way from real (noisy) data, the positivity character is typically lost [54],[1]. We better illustrate this fact in Section 4.5, through numerical simulations.

4.3.4 Convex methods: general framework

More robust approaches for the estimation of physically-acceptable χ (or equivalent parametrizations) have been developed, most notably by resorting to Maximum Likelihood methods [31],[58],[54],[64]. The optimal channel estimation problem can be stated, by using the parametrization for $\chi(\underline{\theta}) =$

$d^{-1}I_{d^2} + \sum_\ell \vartheta_\ell Q_\ell$ presented in the previous section, as it follows: Consider a set of data $\{f_{jk}\}$ as above, and a cost functional $J(\underline{\vartheta}) := h \circ g(\underline{\vartheta})$ where $h : \mathbb{R}^{M\times L} \to \mathbb{R}$ is a suitable function which depends on the data $\{f_{jk}\}$. We aim to find

$$\underline{\hat{\vartheta}} = \arg\min_{\underline{\vartheta}} J(\underline{\vartheta}) \tag{4.28}$$

subject to $\underline{\vartheta}$ belonging to some constrained set $\mathcal{C} \subset \mathbb{R}^{d^4-d^2}$. In our case

$$\mathcal{C} = \mathcal{A}_+ \quad \text{or} \quad \mathcal{C} = \mathcal{A}_+ \cap \mathcal{I},$$

with $\mathcal{A}_+ = \{\underline{\vartheta} \mid \chi(\theta) \geq 0\}$, while $\mathcal{I} = \{\underline{\vartheta} \mid 0 < \mathrm{tr}(\chi(\underline{\vartheta})(\Pi_j \otimes \rho_k^T)) < 1,\ \forall j,k\}$. The second constraint may be used when a cost functional which is not well-defined for extremal probabilities, or in order to ensure that the estimated channel exhibits some noise in each of the measured directions, as it is expected in realistic experimental settings. Since the analysis does not change significantly in the two settings, we will not distinguish between them where it is not strictly necessary. The following result will be instrumental to prove the existence of a unique solution.

Proposition 4.6. *$\mathcal{C}$ is a bounded set.*

Proof. First, we remark that $\mathcal{C}$ is neither closed nor open in general. Since $\mathcal{C} \subset \mathcal{A}_+$, it is sufficient to show that $\mathcal{A}_+$ is bounded or, equivalently, that a sequence $\{\underline{\vartheta}_j\}_{j\geq 0}$, with $\underline{\vartheta}_j \in \mathbb{R}^{d^4-d^2}$, and $\|\underline{\vartheta}_j\| \to +\infty$, cannot belong to $\mathcal{A}_+$. To this end, it is sufficient to show that, as $\|\underline{\vartheta}_j\| \to +\infty$, the minimum eigenvalue of $\chi(\underline{\vartheta}_j)$ tends to $-\infty$ so that, for j large enough, $\underline{\vartheta}_j$ does not satisfy condition $\chi(\underline{\vartheta}_j) \geq 0$. Notice that the map $\underline{\vartheta} \mapsto \chi(\underline{\vartheta})$ is affine. Moreover, since the Q_ℓ are linearly independent, this map is injective. Accordingly, $\|\chi(\underline{\vartheta}_j)\|$ approach infinity as $\|\underline{\vartheta}_j\| \to +\infty$. Since $\chi(\underline{\vartheta}_j)$ is a Hermitian matrix, $\chi(\underline{\vartheta}_j)$ has an eigenvalue λ_j such that $|\lambda_j| \to +\infty$ as $\|\chi(\underline{\vartheta}_j)\| \to +\infty$. Recall that $\chi(\underline{\vartheta}_j)$ satisfies (4.12) by construction which implies that $\mathrm{tr}(\chi(\underline{\vartheta}_j)) = d$ namely the sum of its eigenvalues is always equal to d. Thus, there exists an eigenvalue of $\chi(\underline{\vartheta}_j)$ which approaches $-\infty$ as $j \to +\infty$, which is in contrast with its positivity. So, $\mathcal{C}$ is bounded. ■

Here we focus on the following issue: Under which conditions on the experimental setting (or, mathematically, on the set $\mathcal{B}$ defined above) do the optimization approach have a unique solution? In either of the cases above, $\mathcal{C}$ is the intersection of convex nonempty sets: In fact, $\mathcal{S}_{TP}$ and $\chi \geq 0$ are convex and so must be the corresponding sets of $\underline{\vartheta}$, and it is immediate to verify that $\mathcal{I}$ is convex as well; all of these contain $\underline{\theta} = 0$, corresponding to $\frac{1}{d}I_{d^2}$, and hence they are non empty. In the light of this, it is possible to derive sufficient conditions on J for existence and uniqueness of the minimum in the presence of arbitrary constraint set $\mathcal{C}$. Define $\partial\mathcal{C}_0 := \partial\mathcal{C} \setminus (\partial\mathcal{C} \cap \mathcal{C})$.

Proposition 4.7. *Assume h is continuous and strictly convex on $g(\mathcal{C})$, and*

$$\lim_{\underline{\vartheta}\to\partial\mathcal{C}_0} J(\underline{\vartheta}) = \lim_{\underline{\vartheta}\to\partial\mathcal{C}_0} h\circ g(\underline{\vartheta}) = +\infty. \tag{4.29}$$

If $\mathcal{S}_{TP} = \mathcal{B}$, then the functional J has a unique minimum point in $\mathcal{C}$.

Proof. Since h is strictly convex on $g(\mathcal{C})$ and the linear function g, in view of Proposition 4.4, is injective on $\mathcal{C}$, J is strictly convex on $\mathcal{C}$. So, we only need to show that J takes a minimum value on $\mathcal{C}$. In order to do so, it is sufficient to show that J is inf-compact, i.e., the image of $(-\infty, k]$ under the map J^{-1} is a compact set. Existence of the minimum for J then follows from Weierstrass' theorem. Define $\underline{\vartheta}_0 := \begin{bmatrix} 0 & \dots & 0 \end{bmatrix}^T \in \mathbb{R}^{d^4-d^2}$. Observe that $\chi(\underline{\vartheta}_0) = d^{-1}I_{d^2} \geq 0$. Moreover,

$$\mathrm{tr}(\chi(\underline{\vartheta}_0)\Pi_j \otimes \rho_k^T) = \frac{r_j}{d} < 1 \ \ \forall j,k. \tag{4.30}$$

Therefore, $\underline{\vartheta}_0 \in \mathcal{C}$ and call $J(\underline{\vartheta}_0) = J_0 < \infty$. So, we can restrict the search for a minimum point to the image of $(-\infty, J_0]$ under J^{-1}. Since $\mathcal{C}$ is a bounded set by construction, to prove inf-compactness of J it is sufficient to guarantee that

$$\lim_{\underline{\vartheta}\to\partial\mathcal{C}_0} J(\underline{\vartheta}) = +\infty.$$

■

4.3.5 ML Binomial functional

Assume a certain set of data $\{f_{jk}\}$ have been obtained, by repeating N times the measurement of each pair (ρ_k, Π_j). For technical reasons (strict convexity of the ML functional on the optimization set) and experimental considerations (noise typically irreversibly affects any state), it is typically assumed that $0 < f_{jk} < 1$. The probability of obtaining a series of outcomes with $c_{jk} = f_{jk}N$ ones for the pair (j,k) is then

$$P_\chi(c_{jk}) = \binom{N}{c_{jk}} \mathrm{tr}(\chi\Pi_j \otimes \rho_k^T)^{c_{jk}}[1 - \mathrm{tr}(\chi\Pi_j \otimes \rho_k^T)]^{N-c_{jk}} \tag{4.31}$$

so that the overall probability of $\{c_{jk}\}$, may be expressed as: $P_\chi(\{c_{jk}\}) = \prod_{j=1}^M \prod_{k=1}^L P_\chi(c_{jk})$. By adopting the Maximum Likelihood (ML) criterion, once fixed the $\{c_{jk}\}$ describing the recorded data, the optimal estimate $\hat{\chi}$ of χ is given by maximizing $P_\chi(\{c_{jk}\})$ with respect to χ over a suitable set $\mathcal{C}$. Let us consider our parametrization of the TP $\chi(\underline{\vartheta})$ as in (4.15) . If we assume $0 < \mathrm{tr}(\chi(\underline{\vartheta})(\Pi_j \otimes \rho_k^T)) < 1$, since the logarithm function is monotone,

it is equivalent (up to a constant emerging from the binomial coefficients) to minimize over $\mathcal{C} = \mathcal{A}_+ \cap \mathcal{I}$ [3] the function

$$\begin{aligned} J(\underline{\vartheta}) &= -\frac{1}{N}\log P_{\chi(\underline{\vartheta})}(\{c_{jk}\}) + \frac{1}{N}\sum_{j,k}\log\left(\begin{array}{c} N \\ c_{jk} \end{array}\right) \\ &= -\sum_{j,k} f_{jk}\log[\mathrm{tr}(\chi(\underline{\vartheta})(\Pi_j \otimes \rho_k^T)] \\ &\quad +(1-f_{jk})\log[1-\mathrm{tr}(\chi(\underline{\vartheta})(\Pi_j \otimes \rho_k^T))]. \end{aligned} \tag{4.32}$$

Here, $h(X) = -\sum_{j,k} f_{jk}\log(x_{jk}) + (1-f_{jk})\log(1-x_{jk})$ with $x_{jk} = [X]_{jk}$ and $X \in \mathbb{R}^{M\times L}$ is strictly convex on $\mathbb{R}^{M\times L}$ because $0 < f_{jk} < 1$ by assumption. Notice that $\partial\mathcal{C}_0$ is the set of $\underline{\vartheta} \in \mathcal{A}_+$ for which there exists at least one pair $(\tilde{i},\tilde{k})$ such that $\mathrm{tr}(\chi(\underline{\vartheta})(\Pi_{\tilde{j}} \otimes \rho_{\tilde{k}}^T)) = 0, 1$. Suppose that $\mathrm{tr}(\chi(\underline{\vartheta})(\Pi_{\tilde{j}} \otimes \rho_{\tilde{k}}^T)) \to 0$ as $\underline{\vartheta} \to \partial\mathcal{C}_0$. Therefore, $\log[\mathrm{tr}(\chi(\underline{\vartheta})(\Pi_j \otimes \rho_k^T))] \to -\infty$. Since $c_{\tilde{j},\tilde{k}} > 0$ by assumption, we have that

$$\begin{aligned} \lim_{\underline{\vartheta}\to\partial\mathcal{C}_0} J(\underline{\vartheta}) &= -\lim_{\underline{\vartheta}\to\partial\mathcal{C}_0}\sum_{j,k} f_{jk}\log[\mathrm{tr}(\chi(\underline{\vartheta})(\Pi_j \otimes \rho_k^T))] && (4.33) \\ &\quad +(1-f_{jk})\log[1-\mathrm{tr}(\chi(\underline{\vartheta})(\Pi_j \otimes \rho_k^T))] \\ &= -f_{\tilde{j},\tilde{k}}\lim_{\underline{\vartheta}\to\partial\mathcal{C}_0}\log[\mathrm{tr}(\chi(\underline{\vartheta})(\Pi_{\tilde{j}} \otimes \rho_{\tilde{k}}^T))] && (4.34) \\ &= +\infty. && (4.35) \end{aligned}$$

In similar way, we obtain the same result from the other case, and the conditions for existence and uniqueness of the minimum of Proposition 4.7 are satisfied.

We now discuss *consistency* of this method. Let $\underline{\vartheta}^\circ$ be the "true" parameter and $\chi = \chi(\underline{\vartheta}^\circ)$ be the corresponding χ-matrix of the "true" channel. First observe that, once fixed the sample frequencies f_{jk} (or, equivalently, c_{jk}),

$$J(\underline{\vartheta}) \geq -\sum_{j,k} f_{jk}\log[f_{jk}] + (1-f_{jk})\log[1-f_{jk}],$$

so that if there exists $\hat{\underline{\vartheta}} \in \mathcal{C}$ such that $\mathrm{tr}[\chi(\hat{\underline{\vartheta}})(\Pi_{\tilde{j}} \otimes \rho_{\tilde{k}}^T)] = f_{jk}$, then such a $\hat{\underline{\vartheta}}$ is optimal. Hence, in particular, the (unique) optimal solution corresponding to the f_{jk} equal to the "true" probabilities $\mathrm{tr}[\chi(\Pi_j \otimes \rho_k^T)]$ is exactly $\underline{\vartheta}^\circ$. On the other hand, as the number of experiments N increases, the sample frequencies f_{jk} tend to the "true" probabilities $\mathrm{tr}[\chi(\Pi_j \otimes \rho_k^T)]$. Therefore, in

[3] If the optimization is constrained to $\mathcal{A}_+ \cap \mathcal{I}$, we are guaranteed that f_{jk} will tend to be positive for a sufficiently large numbers of trials.

view of convexity of J and of the continuity of J and its first two derivatives, the corresponding optimal solution tends to the "true" parameter $\underline{\vartheta}^\circ$. This proves consistency.

4.3.6 ML Gaussian functional

Assume a certain data $\{f_{jk}\}$ have been obtained. For each ρ_k consider the sample vector $\underline{f}_k = \left[\begin{array}{ccc} f_{1k} & \dots & f_{Mk} \end{array} \right]^T \in \mathbb{R}^M$, that can be thought as a sample of $\underline{p}^k_\chi = \left[\begin{array}{ccc} \mathrm{tr}(\chi(\Pi_1 \otimes \rho_k^T)) & \dots & \mathrm{tr}(\chi(\Pi_M \otimes \rho_k^T)) \end{array} \right]^T$. Accordingly, we can consider the probabilistic model $\underline{f}_k = \underline{p}^k_\chi + \underline{v}_k$ where $\underline{v}_k \sim \mathcal{N}(0, \Sigma), \Sigma > 0$ is gaussian noise. This noise model is a good representation of certain experimental settings in quantum optics, where the sampled frequencies are obtained with high number of counts c_j and the gaussian noise is due to the electronic of the measurement devices, typically photodiodes. In our model, we can think that to each measured Π_j is associated a different device with noise component v_j. Notice that, the noise components are in general correlated. Let $\underline{\mathcal{D}}_j$ denote the device associated to Π_j. Then, $\underline{\mathcal{D}}_j$ will measure the data $f_{j1}, \dots, f_{jL}$. Since $\underline{f}_k \sim \mathcal{N}(\underline{p}^k_\chi, \Sigma)$, the probability of obtaining the outcomes $\underline{f}_k$ is then

$$P^k_\chi(\underline{f}_k) = \frac{1}{\sqrt{(2\pi)^M \det \Sigma}} \exp\{-\frac{1}{2}(\underline{f}_k - \underline{p}^k_\chi)\Sigma^{-1}(\underline{f}_k - \underline{p}^k_\chi)^T\} \tag{4.36}$$

so that the overall probability of $\{f_{jk}\}$ is equal to $P_\chi(\{f_{jk}\}) = \prod_{k=1}^L P^k_\chi(\underline{f}_k)$. By adopting the ML criterion, given $\{f_{jk}\}$, the optimal estimate $\hat{\chi}$ of χ is given by maximizing $P_\chi(\{f_{jk}\})$ with respect to χ. Taking into account the parametrization $\chi(\underline{\vartheta})$ as in (4.15), it is equivalent to minimize over $\mathcal{C} = \mathcal{A}_+$ the function

$$\begin{aligned} J(\underline{\vartheta}) &= -2 \log \left(\sqrt{(2\pi)^M \det(\Sigma)}\, P_{\chi(\underline{\vartheta})}(\{f_{jk}\}) \right) \\ &= \sum_{k=1}^L (\underline{f}_k - \underline{p}^k_{\chi(\underline{\vartheta})})\Sigma^{-1}(\underline{f}_k - \underline{p}^k_{\chi(\underline{\vartheta})})^T. \end{aligned} \tag{4.37}$$

Then, it easy to see that the conditions of Proposition 4.7 are satisfied. Accordingly, the minimum $\hat{\underline{\vartheta}}$ of J is unique. Also in this case it is possible to show, along the same lines used for the previous functional, the consistency of the method.

4.4 A convergent Newton-type algorithm

In Sections 4.3.5 and 4.3.6 we have presented two ML functionals and showed the uniqueness of their solution. Now, we face the problem of (numerically) finding the solution $\hat{\underline{\vartheta}}$ minimizing J over the prescribed set. In the following we will refer to the binomial functional (4.32), but the results can be easily extended for the Gaussian case.

Consider J as in (4.32) and assume that $\mathcal{S}_{TP} = \mathcal{B}$. Problem (4.28), with $\mathcal{C} = \mathcal{A}_+ \cap \mathcal{I}$, is equivalent to minimize J over $\mathcal{I}$ with the linear inequality constraint $\chi(\underline{\vartheta}) \geq 0$. Rewrite the problem, making the inequality constraint implicit in the objective

$$\hat{\underline{\vartheta}} = \min_{\underline{\vartheta} \in \mathcal{I}} J(\underline{\vartheta}) + I_-(\underline{\vartheta}) \tag{4.38}$$

where $I_- : \mathbb{R}^{d^4-d^2} \to \mathbb{R}$ is the indicator function for the non positive semidefinite matrices $\chi(\underline{\vartheta})$

$$I_-(\underline{\vartheta}) := \begin{cases} 0, & \underline{\vartheta} \text{ s.t. } \chi(\underline{\vartheta}) \geq 0 \\ +\infty, & \text{elsewhere.} \end{cases} \tag{4.39}$$

The basic idea is to approximate the indicator function I_- by the convex function

$$\hat{I}_-(\underline{\vartheta}) := -\frac{1}{q} \log \det(\chi(\underline{\vartheta})) \tag{4.40}$$

where $q > 0$ is a parameter that sets the accuracy of the approximation (the approximation becomes more accurate as q increases). Then, we take into account the approximated problem

$$\hat{\underline{\vartheta}}^q = \min_{\underline{\vartheta} \in \operatorname{int}(\mathcal{C})} G_q(\underline{\vartheta}) \tag{4.41}$$

where $\operatorname{int}(\mathcal{C})$ denotes the interior of $\mathcal{C}$ and the convex function

$$G_q(\underline{\vartheta}) := qJ(\underline{\vartheta}) - \log \det(\chi(\underline{\vartheta})). \tag{4.42}$$

The solution $\hat{\underline{\vartheta}}^q$ can be computed employing the following Newton algorithm with backtracking stage:

1. Set the initial condition $\underline{\vartheta}_0 \in \operatorname{int}(\mathcal{C})$.

2. At each iteration, compute the Newton step

$$\Delta\underline{\vartheta}_l = -H^{-1}_{\underline{\vartheta}_l} \nabla G_{\underline{\vartheta}_l} \in \mathbb{R}^{d^4-d^2} \tag{4.43}$$

where

$$[\nabla G_{\underline{\vartheta}}]_s := \frac{\partial G_q(\underline{\vartheta})}{\partial \vartheta_s}$$

$$= q \sum_{j,k} \left\{ \frac{1 - f_{jk}}{1 - \mathrm{tr}[\chi(\underline{\vartheta}) B_{jk}]} - \frac{f_{jk}}{\mathrm{tr}[\chi(\underline{\vartheta}) B_{jk}]} \right\} \times \tag{4.44}$$

$$\times \mathrm{tr}(Q_s B_{jk}) - \mathrm{tr}[\chi(\underline{\vartheta})^{-1} Q_s] \tag{4.45}$$

$$[H_{\underline{\vartheta}}]_{r,s} := \frac{\partial G_q(\underline{\vartheta})}{\partial \vartheta_r \vartheta_s}$$

$$= q \sum_{j,k} \left\{ \frac{1 - f_{jk}}{[1 - \mathrm{tr}(\chi(\underline{\vartheta}) B_{jk})]^2} + \frac{f_{jk}}{[\mathrm{tr}(\chi(\underline{\vartheta}) B_{jk})]^2} \right\} \times \tag{4.46}$$

$$\times \mathrm{tr}(Q_r B_{jk}) \mathrm{tr}(Q_s B_{jk}) + \mathrm{tr}[\chi(\underline{\vartheta})^{-1} Q_r \chi(\underline{\vartheta})^{-1} Q_s] \tag{4.47}$$

are the element in position s of the gradient (understood as column vector) and the element in position (r, s) of the Hessian of G_q both computed at $\underline{\vartheta}$.

3. Set $t_l^0 = 1$, and let $t_l^{p+1} = t_l^p / 2$ until all the following conditions hold:

$$0 < \mathrm{tr}[\chi(\underline{\vartheta}_l + t_l^p \Delta \underline{\vartheta}_l) B_{jk}] < 1 \ \ \forall\, j, k \tag{4.48}$$

$$\chi(\underline{\vartheta}_l + t_l^p \Delta \underline{\vartheta}_l) \geq 0 \tag{4.49}$$

$$G_q(\underline{\vartheta}_l + t_l^p \Delta \underline{\vartheta}_l) < G_q(\underline{\vartheta}_l) + \gamma t_l^p \nabla G_{\underline{\vartheta}_l}^T \Delta \underline{\vartheta}_l \tag{4.50}$$

where γ is a real constant, $0 < \gamma < \frac{1}{2}$.

4. Set $\underline{\vartheta}_{l+1} = \underline{\vartheta}_l + t_l^p \Delta \underline{\vartheta}_l \in \mathrm{int}\,(\mathcal{C})$.

5. Repeat steps 2, 3 and 4 until the condition $\|\nabla G_{\underline{\vartheta}_l}\| < \epsilon$ is satisfied, where ϵ is a (small) tolerance threshold, then set $\hat{\underline{\vartheta}}^q = \underline{\vartheta}_l$.

To prove the convergence of our Newton algorithm we exploit Proposition 3.13. We proceed in the following way: Identify a compact set D such that $\underline{\vartheta}_l \in D$ and prove that the Hessian is coercive and Lipschitz continuous on D. We then apply Proposition 3.13 in order to prove the convergence.
Since $\underline{\vartheta}_0 \in \mathrm{int}\,(\mathcal{C})$ we consider the set

$$D := \{\underline{\vartheta} \in \mathbb{R}^{d^4 - d^2} \mid G_q(\underline{\vartheta}) \leq G_q(\underline{\vartheta}_0)\}. \tag{4.51}$$

The presence of the backtracking stage in the algorithm guarantees that the sequence $G_q(\underline{\vartheta}_0), G_q(\underline{\vartheta}_1), \ldots$ is decreasing. Thus $\underline{\vartheta}_l \in D$, $\forall l \geq 0$.

Proposition 4.8. *The following facts hold:*

1. *D is a compact set.*

2. *$H_{\underline{\vartheta}}$ is coercive and bounded on D, namely there exist $s, S > 0$ such that*

$$sI \leq H_{\underline{\vartheta}} \leq SI \qquad \forall\, \underline{\vartheta} \in D. \tag{4.52}$$

3. *$H_{\underline{\vartheta}}$ is Lipschitz continuous on D.*

Proof. 1) D is contained into the bounded set $\mathcal{C}$. Since D is a finite dimensional space, it is sufficient to show that

$$\lim_{\underline{\vartheta} \to \partial \mathcal{C}} G_q(\underline{\vartheta}) = +\infty. \tag{4.53}$$

Here, we have three kind of boundary: $\partial\mathcal{I} \cap \text{int}\,(A_+)$, $\text{int}\,(\mathcal{I}) \cap \partial A_+$ and $\partial\mathcal{I} \cap \partial A_+$. Notice that, $\log\det(\chi(\underline{\vartheta}))$ takes finite values on $\partial\mathcal{I} \cap \text{int}\,(A_+)$. Accordingly, taking (4.29) into account,

$$\lim_{\underline{\vartheta} \to \partial\mathcal{I} \cap \text{int}(A_+)} G_q(\underline{\vartheta}) = q \lim_{\underline{\vartheta} \to \partial\mathcal{I} \cap \text{int}(A_+)} J(\underline{\vartheta}) = +\infty. \tag{4.54}$$

Then, $\text{int}\,(\mathcal{I}) \cap \partial A_+$ is the set of $\underline{\vartheta}$ for which J is bounded and there exists at least one eigenvalue of $\chi(\underline{\vartheta})$ equal to zero. Thus,

$$\lim_{\underline{\vartheta} \to \text{int}(\mathcal{I}) \cap \partial A_+} G_q(\underline{\vartheta}) = -\lim_{\underline{\vartheta} \to \text{int}(\mathcal{I}) \cap \partial A_+} \log\det(\chi(\underline{\vartheta})) = +\infty. \tag{4.55}$$

Finally, from (4.54) and (4.55) it follows that $G_q(\underline{\vartheta})$ diverges as $\underline{\vartheta}$ approach $\partial\mathcal{I} \cap \partial A_+$.
2) First, observe that $D \subset \text{int}\,(\mathcal{C})$. Since D is a compact set, there exists $s > 0$ such that

$$\chi(\underline{\vartheta})^{-1} \geq sI \;\; \forall\, \underline{\vartheta} \in D. \tag{4.56}$$

Define

$$\begin{aligned} \delta_{jk} &:= \frac{1 - f_{jk}}{[1 - \text{tr}(\chi(\underline{\vartheta})B_{jk})]^2} + \frac{f_{jk}}{[\text{tr}(\chi(\underline{\vartheta})B_{jk})]^2} > 0 \\ [M_{jk}]_{r,s} &:= \text{tr}(Q_r B_{jk})\text{tr}(Q_s B_{jk}) \end{aligned} \tag{4.57}$$

where M_{jk} is a positive semidefinite matrix with rank equal to one. Accord-

ingly,

$$\begin{aligned}
[H_{\underline{\vartheta}}]_{r,s} &= q\sum_{j,k}\delta_{jk}[M_{jk}]_{r,s} + \\
&\quad +\mathrm{tr}[\chi(\underline{\vartheta})^{-\frac{1}{2}}Q_r\chi(\underline{\vartheta})^{-1}Q_s\chi(\underline{\vartheta})^{-\frac{1}{2}}] \\
&\geq q\sum_{j,k}\delta_{jk}[M_{jk}]_{r,s} + s\mathrm{tr}[Q_r\chi(\underline{\vartheta})^{-1}Q_s] \qquad (4.58)\\
&\geq q\sum_{j,k}\delta_{jk}[M_{jk}]_{r,s} + s^2\mathrm{tr}[Q_rQ_s] \\
&\geq q\sum_{j,k}\delta_{jk}[M_{jk}]_{r,s} + s^2\langle Q_r, Q_s\rangle. \qquad (4.59)
\end{aligned}$$

Since $\{Q_l\}_{l=1}^{12}$ are orthonormal matrices and $\delta_{jk}M_{jk} \geq 0$, we have that

$$H_{\underline{\vartheta}} \geq q\sum_{j,k}\delta_{jk}M_{jk} + s^2 I \geq s^2 I. \qquad (4.60)$$

Notice that, $H_{\underline{\vartheta}}$ is continuous on $\mathrm{int}\,(\mathcal{C})$. Since $D \subset \mathrm{int}\,(\mathcal{C})$, it follows that $H_{\underline{\vartheta}}$ is continuous on the compact D. Hence, there exists $S > 0$ such that $H_{\underline{\vartheta}} \leq SI \ \forall\ \underline{\vartheta} \in D$. We conclude that $H_{\underline{\vartheta}}$ is coercive and bounded on D.
3) $H_{\underline{\vartheta}}$ is continuous on D and $\|H_{\underline{\vartheta}}\| \leq S \ \forall\ \underline{\vartheta} \in D$, therefore $H_{\underline{\vartheta}}$ is Lipschitz continuous on D. ■

Since all the hypothesis of the Proposition 3.13 are satisfied, we have the following proposition.

Proposition 4.9. *The sequence $\{\underline{\vartheta}_l\}_{l\geq 0}$ generated by the Newton algorithm converges to the unique minimum point $\hat{\underline{\vartheta}}^q \in \mathrm{int}\,(\mathcal{C})$ of G_q.*

Then, it is possible to show [11, p. 597] that

$$J(\hat{\underline{\vartheta}}) \leq J(\hat{\underline{\vartheta}}^q) \leq J(\hat{\underline{\vartheta}}) + \frac{d}{q}. \qquad (4.61)$$

Hence, d/q is the accuracy (with respect to $\hat{\underline{\vartheta}}$) of the solution $\hat{\underline{\vartheta}}^q$ found. This method, however, works well only setting a moderate accuracy.

An extension of the previous procedure is given by the Barrier method [11, p. 569] which solves (4.38) with a specified accuracy $\xi > 0$:

1. Set the initial conditions $q_0 > 0$ and $\underline{\vartheta}^{q_0} = \begin{bmatrix} 0 & \dots & 0 \end{bmatrix}^T \in \mathrm{int}\,(\mathcal{C})$.
2. Centering step: At the k-th iteration compute $\hat{\underline{\vartheta}}^{q_k} \in \mathrm{int}\,(\mathcal{C})$ by minimizing G_{q_k} with starting point $\hat{\underline{\vartheta}}^{q_{k-1}}$ using the Newton method previously presented.

3. Set $q_{k+1} = \mu q_k$.

4. Repeat steps 2 and 3 until the condition $\frac{d}{q_k} < \xi$ is satisfied, then set $\hat{\underline{\vartheta}} = \hat{\underline{\vartheta}}^{q_k}$.

So, at each iteration we compute $\hat{\underline{\vartheta}}^{q_k}$ starting from the previously computed point $\hat{\underline{\vartheta}}^{q_{k-1}}$, and then increase q_k by a factor $\mu > 1$. The choice of the value of the parameters q_0 and μ is discussed in [11, p. 574]. Since the Newton method used in the centering step globally converges, the sequence $\{\hat{\underline{\vartheta}}^{q_k}\}_{k\geq 0}$ converges to the unique minimum point $\hat{\underline{\vartheta}}$ of J with accuracy ξ. Moreover, the number of centering steps required to compute $\hat{\underline{\vartheta}}$ with accuracy ξ starting with q_0 is equal to $\left\lceil \frac{\log \frac{d}{\xi q_0}}{\log \mu} \right\rceil + 1$, [11, p. 601].

4.5 Simulation results

4.5.1 Performance comparison

We use the following notation:
• *IN method* to denote the process tomography by inversion of Section 4.3.3.
• *ML method* to denote the ML method, using the functional (4.32) of Section 4.3.4.
Here, we want to compare the performance of IN and ML method for the qubit case $d = 2$. Consider a set of CPTP map $\{\chi_l\}_{l=1}^{100}$ randomly generated and the minimal setting (4.23). Once the number of measurements N for each couple (ρ_k, Π_j) is fixed, we consider the following comparison procedure:
• At the l-th experiment, let $\{c_{jk}^l\}$ be the data corresponding to the map χ_l. Then, compute the corresponding frequencies $f_{jk}^l = c_{jk}^l/N$.
• From $\{f_{jk}^l\}$ compute the estimates $\hat{\chi}_l^{IN}$ and $\hat{\chi}_l^{ML}$ using IN and ML method respectively.
• Compute the relative errors

$$e_{IN}(l) = \frac{\|\hat{\chi}_l^{IN} - \chi_l\|}{\|\chi_l\|},\ e_{ML}(l) = \frac{\|\hat{\chi}_l^{ML} - \chi_l\|}{\|\chi_l\|}. \tag{4.62}$$

• When the experiments are completed, compute the mean of the relative error

$$\mu_{IN} = \frac{1}{100}\sum_{l=1}^{100} e_{IN}(l),\ \mu_{ML} = \frac{1}{100}\sum_{l=1}^{100} e_{ML}(l). \tag{4.63}$$

In Figure 4.1 the results obtained for different lengths N of measurements related to $\{c_{jk}^l\}$ are depicted. The mean error norm of ML method is smaller

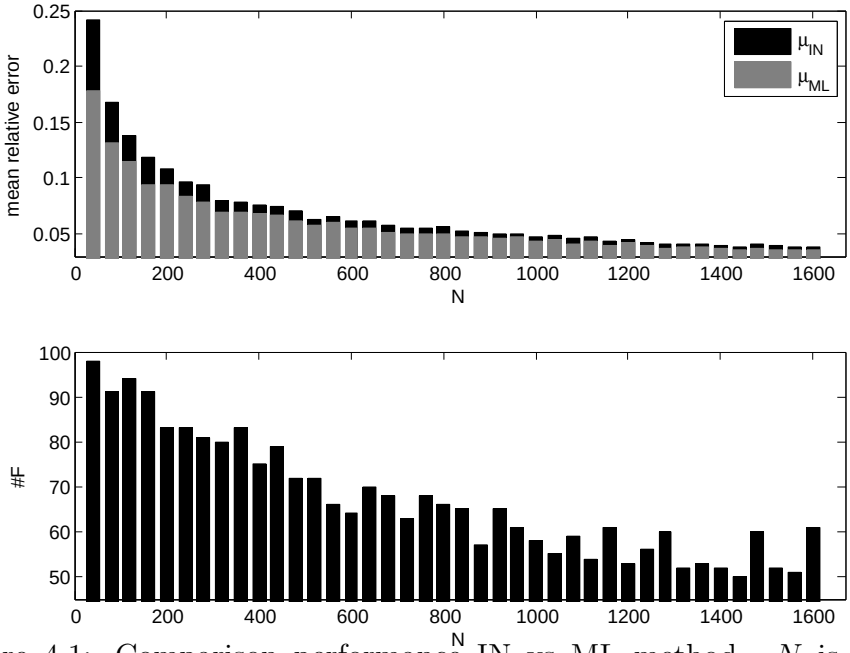


Figure 4.1: Comparison performance IN vs ML method. N is the total number of measurements for each (ρ_k, Π_j), μ is the mean relative error as introduced in (4.63), while $\sharp F$ denotes the number of failures of the IN method, i.e. the times in which the reconstructed χ is not positive.

than the one corresponding to the IN method, in particular when N is small (typical situation in the practice). In addition, more than half of the estimates obtained by the IN method are not positive semidefinite, i.e not physically acceptable, even when N is sufficient large. Finally, we observe that for both methods the mean error decrease as N grows. This fact confirms in the practice their consistency.

4.5.2 Minimal setting

Let $\mathcal{T}_{M,L}$ denote the set of the experimental settings with L input states and M observables satisfying Proposition 4.4. Accordingly the set of the minimal experimental settings is $\mathcal{T}_{d^2-1,d^2}$. Here, we consider the case $d = 2$. We want to compare the performance of the minimal settings in $\mathcal{T}_{3,4}$ with those settings that employ more input states and observables. We shall do so by picking a test channel, finding a minimal setting that performs well, and comparing its performance with a non minimal setting in $\mathcal{T}_{M,L}$, $M > 3, L \geq 4$ that performs well in this set while the total number N_T of trials is fixed.

Consider the Kraus map (4.1) representing a perturbed amplitude damping operation ($\gamma = 0.5$) with

$$K_1 = \sqrt{0.9}\begin{bmatrix} \sqrt{0.5} & 0 \\ 0 & 0 \end{bmatrix}, K_2 = \sqrt{0.9}\begin{bmatrix} 1 & 0 \\ 0 & \sqrt{0.5} \end{bmatrix},$$

$K_3 = \sqrt{0.1}/2I_2$, $K_j = \sqrt{0.1}/2\sigma_{l(j)}$, $j = 4, 5, 6$, $l(j) = x, y, z$ corresponding to the χ-representation

$$\chi = \begin{bmatrix} 0.95 & 0 & 0 & 0.6364 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.05 & 0 \\ 0.6364 & 0 & 0 & 0.5 \end{bmatrix}.$$

We set the total number of trials $N_T = 3600$. Fixed the set $\mathcal{T}_{M,L}$ $M \geq 3$ $L \geq 4$, we take into account the following procedure:
• Set $N = N_T \setminus (LM)$ and choose a randomly generated collection $\{\mathrm{T}_m\}_{m=1}^{100}$, $\mathrm{T}_m \in \mathcal{T}_{M,L}$.
• Perform 50 experiments for each T_m. At the l-th experiment we have a sample data $\{f_{jk}^m(l)\}$ corresponding to χ and T_m. From $\{f_{jk}^m(l)\}$ compute the estimate $\hat{\chi}_m(l)$ using the ML method and the corresponding error norm $e_m(l) = \|\hat{\chi}_m(l) - \chi\|/\|\chi\|$.
• When the experiments corresponding to T_m are completed, compute the mean error norm $\mu_m = \frac{1}{50}\sum_{l=1}^{50} e_m(l)$.

• When we have μ_m for $m = 1 \dots 100$, compute

$$\bar{\mu}_{L,M} = \min_{m \in \{1,\dots,100\}} \mu_m.$$

In Figure 4.2, $\bar{\mu}_{L,M}$ is depicted for different values of M and L. As we

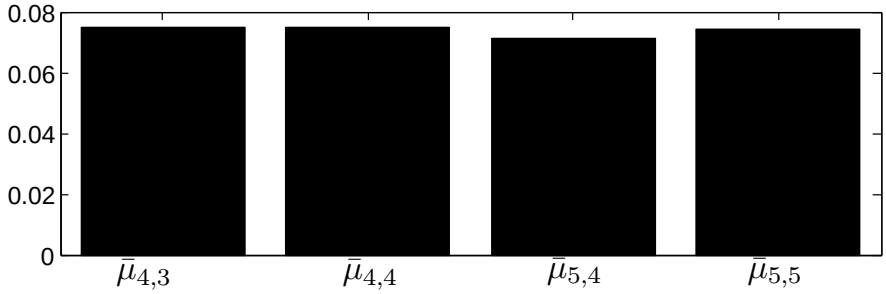


Figure 4.2: $\bar{\mu}_{L,M}$ for different values of L and M.

can see, incrementing the number of input states/observables does not lead to an improvement in the performance index. Analogous results have been observed with other choices of test maps and N_T. Finally, in Figure 4.3 is depicted the true χ and the averaged estimation $\bar{\chi}_{ML} = \frac{1}{50} \sum_{l=1}^{50} \chi_m(l)$ with $m = \arg\min_{m \in \{1,\dots,100\}} \mu_m$ for $M = 3$ and $L = 4$.

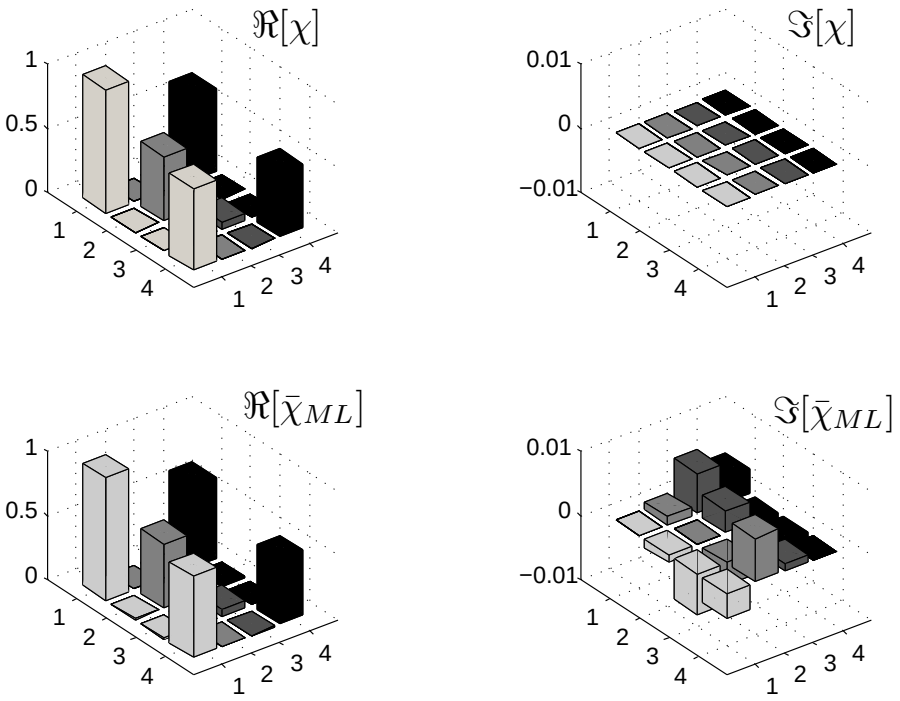


Figure 4.3: Real and imaginary part of χ (top) and the averaged estimation $\bar{\chi}_{ML}$ (bottom). In order to improve readability, the vertical scale of the imaginary part has been magnified in order to show the errors are below 0.01.

Appendix A

On the exponentiation of positive definite matrices

We collect some technical result concerning the exponentiation of positive definite matrices to an arbitrary real number. We start by introducing the differential of the matrix exponential and the matrix logarithm (see [37]).

Proposition A.1. *Given $Y \in \mathcal{Q}_n$, the differential of $Y \mapsto \mathrm{e}^Y$ in the direction $\Delta \in \mathcal{Q}_n$ is given by the linear map*

$$M_Y : \Delta \mapsto \int_0^1 \mathrm{e}^{(1-\tau)Y} \Delta \mathrm{e}^{\tau Y} \mathrm{d}\tau. \tag{A.1}$$

Proposition A.2. *Given $Y \in \mathcal{Q}_{n,+}$, the differential of $Y \mapsto \log(Y)$ in the direction $\Delta \in \mathcal{Q}_n$ is given by the linear map*

$$N_Y : \Delta \mapsto \int_0^\infty (Y + tI)^{-1} \Delta (Y + tI)^{-1} \mathrm{d}t. \tag{A.2}$$

Let us consider now a positive definite matrix $X \in \mathcal{Q}_{n,+}$ and a real number c. The exponentiation of X to c may be rewritten in the following way

$$X^c = e^{c \log X}. \tag{A.3}$$

Accordingly, by applying the chain rule, the differential of $X \mapsto X^c$ in the direction $\Delta \in \mathcal{Q}_n$ is given by

$$M_{c \log X}(cN_X(\Delta)) = c \int_0^1 X^{c(1-\tau)} \int_0^\infty (X + tI)^{-1} \Delta (X + tI)^{-1} \mathrm{d}t X^{c\tau} \mathrm{d}\tau.$$

We summarize this result below.

Proposition A.3. *The differential of $X \mapsto X^c$ in direction $\Delta \in \mathcal{Q}_n$ is given by the linear map*

$$O_{X,c} : \Delta \mapsto c \int_0^1 X^{c(1-\tau)} \int_0^\infty (X + tI)^{-1} \Delta (X + tI)^{-1} \mathrm{d}t X^{c\tau} \mathrm{d}\tau. \quad \text{(A.4)}$$

Corollary A.4. *The first variation of $X \mapsto \mathrm{tr}(X^c)$ in direction $\Delta \in \mathcal{Q}_n$ is*

$$\delta(\mathrm{tr}(X^c); \Delta) = c\mathrm{tr}(X^{c-1}\Delta). \quad \text{(A.5)}$$

Proof. Since $X^{c\tau}$ and $(X + tI)$ commute, we get

$$\begin{aligned} \delta(\mathrm{tr}(X^c); \Delta) &= \mathrm{tr}(O_{X,c}(\Delta)) \\ &= c\mathrm{tr}\left\{ X^c \int_0^\infty (X + tI)^{-2} \mathrm{d}t \Delta \right\} \\ &= c\mathrm{tr}\left\{ X^c X^{-1} \Delta \right\} = c\mathrm{tr}\left\{ X^{c-1} \Delta \right\}. \end{aligned}$$

■

Bibliography

[1] A. Aiello, G. Puentes, D. Voigt, and J. P. Woerdman. Maximum-likelihood estimation of mueller matrices. *Opt. Lett.*, 31(6):817–819, 2006.

[2] N.I. Akhiezer. *The classical moment problem: and some related questions in analysis*. Hafner Publishing, New York, 1965.

[3] R. Alicki and K. Lendi. *Quantum Dynamical Semigroups and Applications*. Springer-Verlag, Berlin, 1987.

[4] A. Basu, I.R. Harris, N.L. Hjort, and M.C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[5] G. Benenti and G. Strini. Simple representation of quantum process tomography. *Phys. Rev. A*, 80(2):022318, 2009.

[6] R. Bhatia. *Matrix Analysis*. Springer-Verlag, New York, 1997.

[7] R. B. Blackman and J. W. Tukey. *The measurement of Power Spectra from the Point of View of Computation Engineering*. Dover, New York, 1958.

[8] I. Bongioanni, L. Sansoni, F. Sciarrino, G. Vallone, and P. Mataloni. Experimental quantum process tomography of non-trace-preserving maps. *Phys. Rev. A*, 82(4):042307, 2010.

[9] N. Boulant, T. F. Havel, M. A. Pravia, and D. G. Cory. Robust method for estimating the lindblad operators of a dissipative quantum process from measurements of the density operator at multiple time points. *Phys. Rev. A*, 67(4):042322, 2003.

[10] D. Bouwmeester, A. Ekert, and A. Zeilinger, editors. *The Physics of Quantum Information: Quantum Cryptography, Quantum Teleportation, Quantum Computation*. Springer-Verlag, 2000.

[11] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, U.K., 2004.

[12] J. P. Burg, D. G. Luenberger, and D. L. Wenger. Estimation structured covariance matrices. *Proceedings of the IEEE*, 70:963–974, 1982.

[13] C.I Byrnes, Per Enqvist, and A. Lindquist. Identifiability and well-posedness of shaping-filter parameterizations: A global analysis approach. *SIAM J. Control and Optimization*, 41(1):23–59, Mar. 2002.

[14] C.I. Byrnes, T.T. Georgiou, and A. Lindquist. A new approach to spectral estimation: A tunable high-resolution spectral estimator. *IEEE Trans. Signal Processing*, 48(11):3189–3205, Nov. 2000.

[15] C.I. Byrnes, T.T. Georgiou, and A. Lindquist. A generalized entropy criterion for Nevanlinna-Pick interpolation with degree constraint. *IEEE Trans. Autom. Control*, 46(5):822–839, May 2001.

[16] C.I. Byrnes, T.T. Georgiou, A. Lindquist, and A. Megretski. Generalized interpolation in H-infinity with a complexity constraint. *Trans. American Math. Society*, 358(3):965–987, Dec. 2004.

[17] C.I Byrnes, S. Gusev, and A. Lindquist. A convex optimization approach to the rational covariance extension problem. *SIAM J. Control and Optimization*, 37(1):211–229, Oct. 1998.

[18] C.I. Byrnes and A. Lindquist. On the partial stochastic realization problem. *IEEE Trans. Autom. Control*, 42(8):1049–1070, Aug 1997.

[19] C.I. Byrnes, A. Lindquist, S.V. Gusev, and A.S. Matveev. A complete parameterization of all positive rational extensions of a covariance sequence. *IEEE Trans. Autom. Control*, 40(11):1841–1857, Nov 1995.

[20] A. Cichocki and S. Amari. Families of Alpha- Beta- and Gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, Jun. 2010.

[21] A. Cichocki, S. Cruces, and S. Amari. Generalized Alpha-Beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170, Jan. 2011.

[22] A. Cichocki, R. Zdunek, A.H. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations*. John Wiley & Sons, Chichester, U.K., 2009.

[23] T. M. Cover and J. A. Thomas. *Information Theory*. Wiley, New York, 1991.

[24] G. M. D'Ariano, L. Maccone, and M. G. A. Paris. Quorum of observables for universal quantum estimation. *Journal of Physics A: Mathematical and General*, 34(1):93, 2001.

[25] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *in NIPS 2006 Workshop on Learning to Compare Examples*, 2007.

[26] A. Ferrante, C. Masiero, and M. Pavon. Time and spectral domain relative entropy: A new approach to multivariate spectral estimation. *IEEE Trans. Autom. Control*, 57(10):2561–2575, Oct. 2012.

[27] A. Ferrante, M. Pavon, and F. Ramponi. Hellinger versus Kullback-Leibler multivariable spectrum approximation. *IEEE Trans. Autom. Control*, 53(4):954–967, May 2008.

[28] A. Ferrante, M. Pavon, and F. Ramponi. Constrained approximation in the Hellinger distance. In *Proceedings of the European Control Conference 2007 (ECC'07)*, pages 322–327, Kos, Greece, July 2007.

[29] A. Ferrante, M. Pavon, and M. Zorzi. A maximum entropy enhancement for a family of high-resolution spectral estimators. *IEEE Trans. Autom. Control*, 57(2):318–329, Feb. 2012.

[30] A. Ferrante, M. Pavon, and M. Zorzi. Structured covariance estimation in high resolution spectral analysis. In *Proc. of Int. Symp. Mathematical Theory of Network and Systems, MTNS 2012*. Melbourne, 2012.

[31] J. Fiurášek and Z. Hradil. Maximum-likelihood estimation of quantum processes. *Phys. Rev. A*, 63(2):020101, Jan 2001.

[32] T. T. Georgiou. Realization of power spectra from partial covariance sequences. *IEEE Trans. Acoust., Speech Signal Processing*, 35(4):438–449, Apr. 1987.

[33] T. T. Georgiou. Spectral estimation by selective harmonic amplification. *IEEE Trans. Aut. Control*, 46:29–42, Jan. 2001.

[34] T. T. Georgiou. Structured covariances and related approximation questions. In A. Rantzer and C. Byrnes, editors, *Directions in Mathematical*

Systems Theory and Optimization, volume 286 of *Lecture Notes in Control and Information Sciences*, pages 135–140. Springer Berlin / Heidelberg, 2003.

[35] T.T. Georgiou. Spectral analysis based on the state covariance: The maximum entropy spectrum and linear fractional parametrization. *IEEE Trans. Autom. Control*, 47(11):1811–1823, Nov. 2002.

[36] T.T. Georgiou. The structure of state covariances and its relation to the power spectrum of the input. *IEEE Trans. Autom. Control*, 47(7):1056–1066, Jul. 2002.

[37] T.T. Georgiou. Relative entropy and the multivariable multidimensional moment problem. *IEEE Trans. Inform. Theory*, 52(3):1052–1066, Mar. 2006.

[38] T.T. Georgiou and A. Lindquist. Kullback-Leibler approximation of spectral density functions. *IEEE Trans. Inform. Theory*, 49(11):2910–2917, Nov. 2003.

[39] T.T. Georgiou and A. Lindquist. Remarks on control design with degree constraint. *IEEE Trans. Autom. Control*, 51(7):1150–1156, Jul. 2006.

[40] T.T. Georgiou and A. Lindquist. A convex optimization approach to ARMA modeling. *IEEE Trans. Autom. Control*, 53(5):1108–1119, Jun. 2008.

[41] H. Hamburger. Über eine Erweiterung des Stieltjesschen Momentenproblems. *Math. Ann.*, 81(2-4):235–319, 1920.

[42] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, 1990.

[43] E.T. Jaynes. On the rationale of maximum-entropy methods. *Proc. of the IEEE*, 70:939–952, Sep. 1982.

[44] J. Karlsson and P. Enqvist. Input-to-state covariances for spectral analysis: The biased estimate. In *Proc. of Int. Symp. Mathematical Theory of Network and Systems, MTNS 2012*. Melbourne, 2012.

[45] J. Karlsson and T. T. Georgiou. Uncertainty Bounds for Spectral Estimation. *ArXiv e-prints*, 2012.

[46] M.G. Kendall, J.K. Stuart, and J.K. Ord. *Advanced theory of statistics*, volume 2. Macmillan, New York, 4th edition, 1983.

[47] K. Kraus. *States, Effects, and Operations: Fundamental Notions of Quantaum Theory*. Lecture notes in Physics. Springer-Verlag, Berlin, 1983.

[48] M.G. Krein and A.A. Nudelman. *The Markov Moment Problem and Extremal Problems*. American Mathematical Society, 1977.

[49] M. Minami and S. Eguchi. Robust blind source separation by Beta divergence. *Neural Computation*, 14(8):1859–1886, Aug. 2002.

[50] M. Mohseni, A. T. Rezakhani, and D. A. Lidar. Quantum-process tomography: Resource analysis of different strategies. *Phys. Rev. A*, 77(3):032322, 2008.

[51] M. Mollah, N. Sultana, M. Minami, and S. Eguchi. Robust extraction of local structures by the minimum of Beta-divergence method. *Neural Networks*, 23(2):226–238, 2010.

[52] A. Nasiri Amini, E.S. Ebbini, and T.T. Georgiou. Noninvasive estimation of tissue temperature via high-resolution spectral analysis techniques. *IEEE Trans. Biomed. Eng.*, 52(2):221–228, Feb. 2005.

[53] M.A. Nielsen and I.L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge, U.K., 2004.

[54] M. G. A. Paris and J. Řeháček, editors. *Quantum States Estimation*, volume 649 of *Lecture Notes Physics*. Springer, Berlin Heidelberg, 2004.

[55] D. Petz. *Quantum Information Theory and Quantum Statistics*. Springer Verlag, 2008.

[56] F. Ramponi, A. Ferrante, and M. Pavon. A globally convergent matricial algorithm for multivariate spectral estimation. *IEEE Trans. Autom. Control*, 54(10):2376–2388, Oct. 2009.

[57] F. Ramponi, A. Ferrante, and M. Pavon. On the well-posedness of multivariate spectrum approximation and convergence of high-resolution spectral estimators. *Systems & Control Letters*, 59:167–172, 2010.

[58] M. F. Sacchi. Maximum-likelihood reconstruction of completely positive maps. *Phys. Rev. A*, 63(5):054104, Apr 2001.

[59] J.A. Shohat and J.D. Tamarkin. *The Problem of Moments*. American Mathematical Society, New York, 1943.

[60] T.J. Stieltjes. *Recherches sur les fractions continues*. 1894.

[61] P. Stoica and R. Moses. *Introduction to Spectral Analysis*. Prentice Hall, New York, 1997.

[62] J. Řeháček, B.-G. Englert, and D. Kaszlikowski. Minimal qubit tomography. *Phys. Rev. A*, 70(5):052321, 2004.

[63] P. Villoresi, T. Jennewein, F. Tamburini, C. Bonato M. Aspelmeyer, R. Ursin, C. Pernechele, V. Luceri, G. Bianco, A. Zeilinger, and C. Barbieri. Experimental verification of the feasibility of a quantum channel between space and earth. *New Journal of Physics*, 10:033038, 2008.

[64] M. Ziman, M. Plesch, V. Bužek, and P. Štelmachovič. Process reconstruction: From unphysical to physical maps via maximum likelihood. *Phys. Rev. A*, 72(2):022106, 2005.

[65] Mário Ziman. Incomplete quantum process tomography and principle of maximal entropy. *Phys. Rev. A*, 78:032118, 2008.

[66] M. Zorzi. A new family of high-resolution multivariate spectral estimators. *ArXiv e-prints*, October 2012.

[67] M. Zorzi and A. Ferrante. On the estimation of structured covariance matrices. *Automatica*, 48(9):2145–2151, Sep. 2012.

[68] M. Zorzi, F. Ticozzi, and A. Ferrante. On Quantum Channel Estimation with Minimal Resources. *ArXiv e-prints*, June 2011.

[69] M. Zorzi, F. Ticozzi, and A. Ferrante. Estimation of quantum channels: Identifiability and ML methods. 51st IEEE Conference on Decision and Control.