



THE IBY AND ALADAR FLEISCHMAN FACULTY OF ENGINEERING  
The Zandman-Slaner Graduate School of Engineering

# Second-Order Multidimensional Independent Component Analysis: Theory and Methods

By

**Dana Lahat**

THESIS SUBMITTED TO THE SENATE OF TEL AVIV UNIVERSITY  
in partial fulfillment of the requirements for the degree of  
“DOCTOR OF PHILOSOPHY”

August 2012





THE IBY AND ALADAR FLEISCHMAN FACULTY OF ENGINEERING  
The Zandman-Slaner Graduate School of Engineering

# Second-Order Multidimensional Independent Component Analysis: Theory and Methods

By

**Dana Lahat**

THESIS SUBMITTED TO THE SENATE OF TEL AVIV UNIVERSITY  
in partial fulfillment of the requirements for the degree of  
“DOCTOR OF PHILOSOPHY”

Under the supervision of  
Dr. Jean-François Cardoso and Prof. Hagit Messer

August 2012



This work was carried out under the supervision of

**Dr. Jean-François Cardoso**  
**and**  
**Prof. Hagit Messer**



# Acknowledgements

I would like to express my deepest gratitude to my two supervisors, Dr. Jean-François Cardoso and Prof. Hagit Messer.

It has been a great honour and privilege to have the opportunity to work under the supervision of such prominent figures in the signal processing community. Thank you for the many hours of inspiring discussions. Thank you for guiding me in my first academic steps, and for sharing with me your sharp and uncompromising vision of scientific research.



I would like to thank the members of the ADAMIS group in the APC lab in Paris, France, which I have visited several times during my PhD. Especially I would like to thank Maude Le Jeune for helping me with the Planck Sky Model simulations, and Jacques Delabrouille for your very kind support and assistance in various matters.

I would like to thank the reviewers of my thesis for reading it and providing me with your important and helpful comments.

Prof. Lieven De Lathauwer from KU Leuven, Belgium, I would like to thank you for your interest in my work and for your pertinent remarks.

תודה לפרופ' אריה ירדאור מאוניברסיטת תל אביב על השיחות המועילות עמך.  
תודה לפרופ' ג'ורג' וויס מאוניברסיטת תל אביב, על אוזן קשבת, עידוד ועצה טובה.  
תודה לד"ר יעקב גולדברגר מאוניברסיטת בר אילן על תמיכתך ועצותיך.  
יעקב פיינגלרנט, תודה לך על שעזרת לי בהדרכת המעבדות, על דאגתך ועל תשומת לבך.  
תודה מקרב לב לד"ר שרון שדה על עזרתך הרבה בכל הקשור לעבודה עם Unix, להתקנת octave, פורטרן וhealpix, וכמובן על ידידותך הנעימה.  
תודה עד אין קץ ליונה אייל, על תשומת לבך אלי ועל נכונותך לעזור לי מייד, בכל רגע ושעה בכל הקשור בהתמודדותי עם מערכת המחשב, דואר, שרתים, התקנת תוכנות אוטוריות על לינוקס, ועוד.  
אלכס וייסמן, תודה לך על שעזרת לי שוב ושוב, תמיד מהר ככל האפשר ובנעימות, בשאלות מחשוב שונות.  
תודה לעירית נויילנדר ולזיוה ליפוביצקי על שהייתן נכונות לעזור לי בכל ענין שביכולתכן, ועל שהארתן לי פנים בכל עת.  
תודה לחברי, דוקטורים בהווה ולעתיד, על סבלנותכם הרבה לכל פְּרוֹנוֹיָתִי, על תמיכתכם הבלתי מתפשרת בכל רגע ומצב, על עצותיכם הטובות, עידודיכם ומעל לכל, על חברותכם ללא תנאי: ד"ר אמיר בן נתן, ד"ר אלונה ארז, ד"ר אהד ברק, ד"ר גדי מילר, ד"ר יובל קוכמן, יובל לומניץ, אלי חיים, ניר נוסנסון.  
תודה לשותפי במעבדה 313, שבזכותכם החודשים האחרונים של הדוקטורט שלי הפכו לנעימים במיוחד. בפרט, תודתי שלוחה לגיא יונה, לאנטון אנדרוסיר, ולטדי ריבר.

Sarah et Alain Ankierman, Eric, Sylvie et Raphaël, je vous remercie de m'avoir accueillie chez vous, d'avoir rendu mes longs séjours à Paris toujours agréables.

שרון גנות, תודה אינסופית על שלא חסכת מאמצים לעזור לי בכל דרך שרק אפשר.  
הוכי יהודית ומנשה, אחי איתן ודרור, תודה שתמכתם בי בכל שעה ובכל מצב.

אני מקדישה עבודה זאת לסבתותי ולסבִּי, יוכבד וחיים לאטי, אניה ודוד פלצ'ר, שנטעו בי את הרצון, התשוקה והצורך ללמוד, ולצערי לא זכיתי שיהיו עמי ברגע זה.

We acknowledge the use of the Planck Sky Model, developed by the Component Separation Working Group (WG2) of the Planck Collaboration.

This work has been partially supported by the Chateaubriand Fellowship of the French Government.

# Abstract

Independent component analysis (ICA) and blind source separation (BSS) deal with extracting a number of mutually independent elements from a set of observed mixtures thereof. Motivated by various applications, this work considers a more general and more flexible model: the sources can be partitioned into groups exhibiting dependence within a given group but independence between two different groups. We argue that this is tantamount to considering *multidimensional* components, as opposed to the standard ICA which is restricted to one-dimensional components.

In this work, we focus on second-order methods to separate statistically-independent multidimensional components from their linear instantaneous mixtures. The purpose of this work is to provide theoretical answers to questions which so far have been discussed mainly in the empirical domain. Namely, we provide a closed-form expression for the figure of merit, the mean square error (MSE), for multidimensional component separation, in two prominent scenarios: one is the optimal separation procedure. The other is a two-step procedure, which at the first step attributes the data a one-dimensional model, and at the second step clusters the data into multidimensional components. We prove that the latter is sub-optimal. Using the closed-form expressions, we can calculate the expected gain from using the correct multidimensional model over its one-dimensional counterpart. We derive two novel joint block diagonalization (JBD) algorithms which achieve the minimal mean square error (MMSE), when the model assumptions hold. We present new results on the identifiability of the model, including a rigorous proof. The identifiability results of the model are also the condition for the uniqueness of a solution to JBD of a set of mixtures of block-diagonal positive-definite symmetric matrices. As for the methods used in this work, it is shown that all the required derivations can be performed and presented in terms of well-defined quantities, which avoid the well-known scale ambiguity, prevalent in ICA and BSS representations. We demonstrate our methods and algorithms on an astrophysical application. Namely, extraction of the cosmic microwave background radiation (CMB) from its observations. We adapt our analysis to the spectral domain, in order to apply our methods on spectral density matrices. In this application, we extend our analysis to the overdetermined case, and deal with a situation where the component dimensions are not known a-priori. We support our theoretical results by numerical simulations.



# Contents

<b>Acknowledgements</b>	<b>i</b>
<b>Abstract</b>	<b>iii</b>
<b>Notations and Conventions</b>	<b>ix</b>
<b>Acronyms</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Preface . . . . .	1
1.2 Literature Survey and Our Related Contribution . . . . .	2
1.2.1 Optimal Performance of Second-Order Multidimensional ICA	2
1.2.2 Identifiability of Second-Order Multidimensional ICA . . . . .	5
1.2.3 Joint Block Diagonalization Algorithms . . . . .	6
1.2.4 Separation of Multidimensional Data Using a One-Dimensional Model . . . . .	7
1.2.5 Overdetermined Multidimensional ICA: Application to CMB Observations . . . . .	8
1.3 Thesis Outline . . . . .	11
<b>2 Mathematical Preliminaries</b>	<b>15</b>
2.1 Outline . . . . .	15
2.2 Dependent Sources vs. Multidimensional Components . . . . .	15
2.3 Model, Likelihood and Contrast Function . . . . .	17
2.3.1 Piecewise Stationary Model . . . . .	18
2.3.2 Likelihood . . . . .	18
2.3.3 Contrast Function for a Mixture of Dependent Sources . . . . .	20
2.3.4 Estimating Equations in Terms of the Mixing Matrix . . . . .	20
2.3.5 Estimating Equations in Terms of the Projectors . . . . .	22
2.4 Summary . . . . .	22
2.A Derivation of the Contrast Function . . . . .	23
<b>3 Optimal Performance of Second-Order Multidimensional ICA</b>	<b>25</b>
3.1 Outline . . . . .	25
3.2 Error Analysis . . . . .	26
3.2.1 Error Decomposition . . . . .	26
3.2.2 Influence Function . . . . .	27

3.2.3	Error Covariance for the Projectors . . . . .	29
3.2.4	Mean Square Error . . . . .	31
3.3	Numerical Examples . . . . .	32
3.3.1	Validating the Closed-Form Expression for the MSE . . . . .	33
3.3.2	Validating the Small-Errors Regime . . . . .	34
3.4	Discussion . . . . .	36
3.A	First-Order Expansion of the Estimating Equations . . . . .	37
3.B	Closed-Form Expression for $\text{Cov}([\mathbf{g}_{ij}^\dagger \ \mathbf{g}_{ji}^\dagger]^\dagger)$ . . . . .	39
3.C	Proof of $\mathcal{H}^{-1}\mathcal{H}_\Pi\mathcal{H}^{-\dagger} = \mathcal{H}_\Pi^\sharp$ . . . . .	41
3.D	Explicit Form for $\text{Cov}(\text{vec}\{\mathcal{E}_{ij}\})$ and $\text{Cov}(\text{vec}\{\mathcal{E}_{ij}\}, \text{vec}\{\mathcal{E}_{ji}\})$ . . . . .	41
<b>4</b>	<b>Identifiability of Second-Order Multidimensional ICA</b>	<b>45</b>
4.1	Outline . . . . .	45
4.2	Degrees of Freedom . . . . .	46
4.3	Derivation of the Identifiability Theorem . . . . .	46
4.4	Invertibility of $\mathcal{H}$ . . . . .	48
4.5	Example: Non-Identifiability . . . . .	50
4.6	Discussion . . . . .	51
4.A	Proof of the Identifiability Theorem . . . . .	52
<b>5</b>	<b>Joint Block Diagonalization Algorithms</b>	<b>55</b>
5.1	Outline . . . . .	55
5.2	Derivation of the Relative Variations . . . . .	55
5.3	Algorithms . . . . .	57
5.4	Numerical Examples . . . . .	58
5.5	Summary . . . . .	60
5.A	First-Order Expansion of the Estimating Equations . . . . .	61
<b>6</b>	<b>Separation of Multidimensional Data Using a One-Dimensional Model Followed by Clustering: Performance Analysis</b>	<b>65</b>
6.1	Outline . . . . .	65
6.2	Problem Formulation . . . . .	67
6.2.1	Component Estimation under Mismodeling . . . . .	68
6.2.2	Characterizing the Solutions to the Mismodeled Contrast Function . . . . .	70
6.3	Error Analysis . . . . .	73
6.3.1	Defining the Figure of Merit . . . . .	73
6.3.2	Error Decomposition . . . . .	74
6.3.3	Influence Function . . . . .	77
6.3.4	Error Covariance and Mean Square Error . . . . .	80
6.4	Numerical Examples . . . . .	81
6.4.1	JBD by JD Followed by Clustering . . . . .	82
6.4.2	Validation of the Performance Analysis . . . . .	82
6.4.3	Validating the Small-Errors Regime . . . . .	85
6.4.4	$\text{MSE}_i^{\text{JD}} / \text{MSE}_i^{\text{JBD}}$ as a Function of the Model Parameters . . . . .	86
6.5	Discussion . . . . .	87
6.A	First-Order Expansion of the Estimating Equations . . . . .	90

6.B	Closed-Form Expression for $\text{Cov}(\mathbf{g}_{ij})$ . . . . .	93
6.C	Explicit Form of $\mathcal{H}$ and $\text{Cov}(\mathbf{g}_{ij})$ . . . . .	97
<b>7</b>	<b>Overdetermined Multidimensional ICA: Application to CMB Observations</b>	<b>99</b>
7.1	Outline . . . . .	99
7.2	Preliminaries . . . . .	100
7.2.1	Astrophysical Background . . . . .	100
7.2.2	Spherical Harmonics . . . . .	103
7.2.3	Cosmic Variance . . . . .	104
7.3	Motivation . . . . .	104
7.4	Application to Sky Observation . . . . .	105
7.4.1	Overdetermined Data Model . . . . .	105
7.4.2	Derivation of the Spectral Density Matrices . . . . .	106
7.4.3	Cosmological Processes as Multidimensional Components . . . . .	108
7.4.4	Dimension Reduction . . . . .	108
7.4.5	Mean Square Error . . . . .	109
7.5	Numerical Experiments . . . . .	111
7.5.1	Model Order Selection for the Galactic Emission . . . . .	111
7.5.2	JBD vs. JD Followed by Clustering . . . . .	114
7.5.3	Empirical vs. Predicted MSE . . . . .	115
7.6	Discussion . . . . .	117
<b>8</b>	<b>Conclusion</b>	<b>121</b>
8.1	Summary . . . . .	121
8.2	Discussion . . . . .	123
8.3	Topics for Further Research . . . . .	124
<b>A</b>	<b>Some Algebraic Properties</b>	<b>127</b>



# Notations and Conventions

$a, b, \kappa, T$	scalar or constant
$\mathbf{a}, \mathbf{b}$	column vector
$\mathbf{A}, \mathbf{M}, \mathbf{Y}$	matrix
$\mathbf{A}^{-1}$	matrix inverse
$\mathbf{R}^{(q)}$	matrix $\mathbf{R}$ indexed by $q$
$\mathbf{R}^{-(q)}$	$(\mathbf{R}^{(q)})^{-1}$ , inverse of $\mathbf{R}^{(q)}$
$D(\cdot), C(\cdot)$	function or operator
$(\cdot)^\dagger$	transpose
$(\cdot)^H$	Hermitian
$a^*$	complex conjugate of scalar $a$
$(\cdot)^\#$	Moore-Penrose pseudoinverse
$\text{tr}\{\cdot\}$	trace
$ \mathbf{a} ^2$	$\mathbf{a}^\dagger \mathbf{a}$ for any vector $\mathbf{a}$
$\text{vec}\{\cdot\}$	the operator which stacks the columns of a $p \times q$ matrix into a $pq \times 1$ vector
$E\{\cdot\}$	expectation
$\text{Cov}(\mathbf{a})$	$E\{\mathbf{a}\mathbf{a}^\dagger\}$ for any stochastic vectors $\mathbf{a}$ with $E\{\mathbf{a}\} = \mathbf{0}$
$\text{Cov}(\mathbf{a}, \mathbf{b})$	$E\{\mathbf{a}\mathbf{b}^\dagger\}$ for any stochastic vectors $\mathbf{a}, \mathbf{b}$ with $E\{\mathbf{a}\} = \mathbf{0}$
$\ \cdot\ $	Frobenius norm
$\delta_{ij}$	Kronecker delta
$\otimes$	Kronecker product
$\langle \mathbf{M} \rangle$	$\frac{1}{T} \sum_{q=1}^Q n_q \mathbf{M}^{(q)}$ , a weighted average of any sequence indexed by $q$ with weights $n_q$ , $\sum_{q=1}^Q n_q = T$
$\Omega(f)$	zero mean stochastic terms whose standard deviation is proportional to $f$ , or to higher powers thereof
$O(f)$	deterministic terms which are bounded above, up to a constant factor, by $f$
$\mathcal{U}[a, b]$	Uniform distribution with parameters $a$ and $b$
$\mathcal{N}(\mu, \sigma^2)$	Normal (Gaussian) distribution with parameters $\mu$ and $\sigma^2$
$\mathbf{0}_{m \times n}$	matrix of zeros with $m$ rows and $n$ columns



# Acronyms

<b>AR</b>	auto-regressive
<b>ARMA</b>	auto-regressive moving-average
<b>BSS</b>	blind source separation
<b>CMB</b>	cosmic microwave background radiation
<b>CRLB</b>	Cramér-Rao lower bound
<b>ECG</b>	electrocardiography
<b>EEG</b>	electroencephalography
<b>FIM</b>	Fisher information matrix
<b>i.i.d.</b>	independent and identically distributed
<b>ICA</b>	independent component analysis
<b>ILC</b>	internal linear combination
<b>ISA</b>	independent subspace analysis
<b>ISR</b>	interference to source ratio
<b>IVA</b>	independent vector analysis
<b>JBD</b>	joint block diagonalization
<b>JBSS</b>	joint blind source separation
<b>JD</b>	joint diagonalization
<b>KLD</b>	Kullback-Leibler divergence
<b>LHS</b>	left-hand side
<b>LS</b>	least squares
<b>MA</b>	moving-average
<b>MEG</b>	magnetoencephalography

<b>MICA</b>	multidimensional ICA
<b>ML</b>	maximum likelihood
<b>MMSE</b>	minimal mean square error
<b>MSE</b>	mean square error
<b>PCA</b>	principal component analysis
<b>PSM</b>	Planck Sky Model
<b>QN</b>	quasi-Newton
<b>RHS</b>	right-hand side
<b>RG</b>	relative gradient
<b>SHT</b>	spherical harmonic transform
<b>TCA</b>	tree-dependent component analysis
<b>WMAP</b>	Wilkinson microwave anisotropy probe

# List of Figures

3.1	Validating the theoretical expression for $\text{Cov}(\delta\mathbf{P}_i)$ , as well as our choice of $n_q$ for the small-errors regime. . . . .	44
5.1	Convergence rate of the relative gradient and quasi-Newton algorithms	59
5.2	Numerical example to validate that the quasi-Newton algorithm has indeed converged to the correct minimum. . . . .	60
6.1	JBD by JD followed by clustering . . . . .	83
6.2	Validating the theoretical expression for $\text{Cov}(\delta\mathbf{P}_i^{\text{JD}})$ , as well as our choice of $n_q$ for the small-errors regime. . . . .	87
6.3	The ratio $\text{MSE}_i^{\text{JD}} / \text{MSE}_i^{\text{JBD}}$ for different values of $Q$ , $\mathbf{A}$ , $\mathbf{m}$ and $\mathbf{R}_S^{(q)}$ . . . . .	88
7.1	Temperature fluctuations over the full sky at 23GHz and 94GHz, and a processed map showing only the CMB . . . . .	102
7.2	Spherical coordinates . . . . .	103
7.3	CMB angular power spectrum . . . . .	105
7.4	Simulated component maps at the 217GHz detector channel. . . . .	112
7.5	Comparison of empirical $\text{MSE}_{\text{CMB}}$ with and without finite-data errors for $N_{\text{Gal}} = 3, 4, 5$ . . . . .	113
7.6	Kullback-Leibler-induced divergence between spectral density matrices and their block-diagonal model, with and without finite-data errors, for $N_{\text{Gal}} = 2, 3, 4, 5$ . . . . .	115
7.7	Kullback-Leibler-induced divergence between spectral density matrices and their diagonal and block-diagonal models, for $N_{\text{Gal}} = 2, 3, 4, 5$	116
7.8	Empirical vs. predicted $\text{MSE}_{\text{CMB}}$ . . . . .	120



# List of Tables

3.1	Performance of second-order multidimensional ICA: Analytical vs. empirical normalized MSE . . . . .	35
6.1	Performance of ICA of multidimensional data, when the separation is based on a one-dimensional model followed by clustering. . . . .	85



# Chapter 1

## Introduction

### 1.1 Preface

Independent component analysis (ICA) [1, 2, 3] and blind source separation (BSS) [4, 5] deal with extracting a number of mutually statistically independent elements from a set of observed mixtures thereof. Motivated by various applications, this work considers a more general and more flexible model: the mixed elements can be partitioned into groups exhibiting statistical dependence within a given group and statistical independence between two different groups. We argue that this is tantamount to considering *multidimensional* components, as opposed to the standard ICA which is restricted to one-dimensional components.

Multidimensional data may occur due to various complex relations within the dependent elements. The dimension of a dependent group may not always reflect the actual number of its underlying elements. As a result, in multidimensional models, there is not always a physically meaningful interpretation to separating the multidimensional components back into single-dimensional elements. Hence, a one-dimensional model for real data is often just an approximation. Various phenomena can yield data which is multidimensional in nature. For example, in the analysis of magnetoencephalography (MEG) stimuli [6], energies of evoked signals may become statistically dependent due to similar activation and termination times. In addition, a single isolated neural source is of little diagnostic value and consequently the number of detected sources may be reduced through a clustering procedure, based on spatial topography [7]. In natural image analysis, subspaces originate from features which represent positions, orientations, spatial frequencies and phases [8]. In fetal electrocardiography (ECG) detection [9, 10, 11], multidimensionality is explained by a spatially rotating electric dipole model of the heart, and may change during pregnancy. In [12], a multidimensional model, based on clustering of metabolites, is

suggested in order to recover independent metabolite profiles, each of which stands for a separate “direction” or cellular process in metabolic space. Finally, in the separation of astrophysical emissions, multidimensionality may reflect not only dependence between different emissions due to astrophysical processes [13, 14, 15], but also spatial nonstationarity, when observations are taken over a large enough region of the sky [16]. Additional examples of multidimensional data can be found, for example, in [17, Sec. 1].

This work deals with blind separation of statistically independent multidimensional components using second-order methods. We restrict our analysis to a simple model which still captures the essence of the problem. Namely, instantaneous linear mixtures, free from additive noise. The purpose of this work is to provide theoretical answers to questions which so far have been discussed mainly in the empirical domain. Namely, analytical expressions for the expected performance, and algorithms which achieve this expected performance.

## 1.2 Literature Survey and Our Related Contribution

### 1.2.1 Optimal Performance of Second-Order Multidimensional ICA

The idea of solving the dependent sources/multidimensional components problem in terms of blind source subspace separation through independent component analysis (ICA) was first demonstrated in [9], on fetal ECG recordings. The perspective of *multidimensional ICA*, of vector-valued components whose representation is based on unambiguous projections on the sources’ respective subspaces, was presented in [10]. An elaborate geometric framework for multidimensional ICA (MICA) was suggested in [10]. The notion of independent feature subspaces and independent subspace analysis (ISA) was introduced in [18, 8]. In the ISA framework, independent and identically distributed (i.i.d.) non-Gaussian observations are separated into feature-invariant subspaces by a criterion which maximizes the independence between norms of projections on linear subspaces. ISA algorithms usually search for a de-mixing matrix which is constrained to be orthogonal; examples are given in [8, 19, 20, 21]. This constraint is prevalent in source separation algorithms which are based on pre-whitening, as explained in [22, Sec. 3.1]. Source separation based on pre-whitening, in the context of [22, Sec. 3.1], simplifies the optimization by restricting the de-mixing matrix to be orthogonal; this results in a smaller num-

ber of variables to adjust. However, by assuming that a matrix which is based on finite-data observations can whiten the latent sources, whitening ignores finite-data effects. Whitening errors which are introduced in this preliminary stage cannot be compensated for by the following rotation stage [4, Sec. 4]. Hence, such methods are sub-optimal with respect to methods which are not based on pre-whitening [22, 23]. In the following, when we mention methods which use (pre-) whitening, we always refer to this entire procedure, including the orthogonality constraint. An algorithm which solves ISA through joint block diagonalization (JBD) of cumulant matrices is described in [24]. In this algorithm JBD is performed by joint diagonalization (JD) and permutation-recovery, and has a pre-whitening stage. An ISA algorithm without the whitening constraint is given, for example, in [25]. A non-parametric dependence measure, in the sense of no parametric assumption on the functional form of the joint distribution, is found in [26]. The minimization of this measure is based on one-dimensional ICA as a step in separating multidimensional channels for non-Gaussian i.i.d. pre-whitened data. The above-mentioned algorithms require as an input the correct subspace dimensions. An algorithm which extends [19] by automatically detecting the subspace dimensions and clustering the data is proposed in [27]. Conditions under which the two-step procedure of ICA and then clustering is sufficient for separation are discussed in [28]; this analysis is valid for i.i.d. pre-whitened data and specific distribution types. Another approach to weaken the assumption of independence in ICA has been proposed by [29]. Their approach, denoted as tree-dependent component analysis (TCA), looks for a transform which makes the data components well fit by a tree-structured graphical model. In particular, TCA allows the underlying graph to have multiple connected components, and thus the method is able to find “clusters” of components such that components are dependent within a cluster and independent between clusters. Bach and Jordan [29] also suggest an algorithm to learn the number and size of the clusters, even if they are not given a-priori. A flexible model which can incorporate multidimensional components, components with various parametric constraints, as well as additive noise, has been presented in [16]. The method in [16] is based on minimizing the mismatch between the empirical and model covariance matrices.

In the above-mentioned works, the performance of the algorithms is examined numerically [10, 9, 18, 8, 26, 27, 29], or their convergence to the correct separation point is discussed theoretically [19, 21, 20, 24, 25, 28]. However, to the best of our knowledge, no theoretical performance analysis, in the sense of closed-form expressions for an expected figure of merit or bound, has been conducted to any of these scenarios.

On the other hand, Cramér-Rao lower bounds on the estimation error of the mixing matrix and of the source parameters in ICA and BSS have already been discussed in the literature. Optimal performance and lower bounds were analyzed, for example, for the following scenarios: system identification for non-Gaussian sources [30]; noise-free static-mixture ICA for non-Gaussian sources [31, 32, 33]; noise-free stationary auto-regressive (AR) Gaussian sources [34]. The Cramér-Rao lower bound (CRLB), as well as a lower bound on the interference to source ratio (ISR), for noise-free stationary AR, moving-average (MA) and auto-regressive moving-average (ARMA) Gaussian sources, is given by Doron et al. [35]. Tichavský et al. [36] provide a lower bound on the ISR, based on the result of [35], for noise-free and non-stationary Gaussian AR processes. Cardoso and Pham [37] provide performance bounds for non-stationary Gaussian and non-Gaussian sources. Yeredor [38] provides the induced CRLB on the ISR for noise-free Gaussian non-stationary sources in two scenarios, one in which the temporal covariance matrices of the sources are known (“semi blind”) and one in which they are unknown (“fully blind”). For joint blind source separation (JBSS), which is a closely-related problem to BSS, an explicit expression for the induced CRLB [38] on the ISR has been derived by Anderson et al. [39]. Lower bounds on source separation due to whitening have been derived in [22, 23]. Closed-form error expressions for specific algorithms, without claim for optimality or lower bounds, can be found, for example, in the following works. Pham and Garat [40] discuss methods for separating noise-free mixtures of independent sources without any precise knowledge of their probability distribution. They obtain closed-form expression for the asymptotic performance of their proposed algorithms, and use these theoretical results to choose filters which yield the best performance. Belouchrani et al. [41] obtain a closed-form expression for the asymptotic performance of a second-order method based on JD of a set of lagged covariance matrices, for static-mixture ICA with additive noise.

**Our Contribution** In the first part of this work, we provide a rigorous small-error performance analysis. This analysis yields closed-form expressions for the Fisher information matrix (FIM), CRLB and mean square error (MSE), in terms of the model parameters. To the best of our knowledge, this is the first time that closed-form expressions for FIM, CRLB and MSE are derived for a multidimensional component separation procedure. Our analysis considers a second-order based method which extends the maximum likelihood (ML) treatment of [42] and the performance analysis of [37, 43] to the case of piecewise stationary Gaussian dependent sources/multidimensional components. In particular, we derive a contrast function,

whose minimization is equivalent to the ML solution for Gaussian piecewise stationary data. As in [42], there is no whitening constraint. Hence, the derived expressions represent the optimal performance under asymptotic conditions.

These results appear in Lahat et al. [44, 45].

### 1.2.2 Identifiability of Second-Order Multidimensional ICA

Once we have established the performance of component separation for our data model, we turn to the question of the identifiability of the model. That is, under which conditions the model itself is not identifiable and the proposed analysis is *not* valid. In this work we show that the optimal solution is obtained by minimizing a contrast function which is equivalent to JBD of a set of sample covariance matrices. Hence, our identifiability results are tightly related to results about the uniqueness of JBD of simultaneous mixtures of block-diagonal symmetric positive-definite matrices. An extensive and rigorous discussion and definition of the finest unique JBD of a set of symmetric positive-definite matrices is given by [46], in terms of matrix  $*$ -algebra. De Lathauwer [47, Theorem 6.1] has proved that uniqueness of JBD, up to trivial indeterminacies and for  $Q > 2$  matrices, is guaranteed with probability one when the entries of the input matrices are drawn from a continuous probability density function. It has been shown by [48] that blind identification of mixed one-dimensional sources is impossible when the variances of the different sources, as a function of the domain index, are proportional. Gutch et al. [49] have recently presented analogous conditions of (non-) identifiability for the multidimensional case. In [49], the discussion is based on whitened second-order wide-sense stationary sources and the proofs of the theorems in [49, Sec. 1.3] are omitted.

Recently, Anderson et al. [39] and Vía et al. [50] have presented results on the stability and identifiability of the JBSS problem using the independent vector analysis (IVA) framework and with Gaussian multivariate priors on the sources. The JBSS problem can be formulated (see, e.g., [39, Sec. II]) in a way which is in some sense reminiscent of the MICA model, although it deals with a different problem, of several datasets, each with its own *different* mixing matrix, and of statistically independent *one*-dimensional sources. Despite the essential difference in the model and in the final results, the identifiability analysis in the above works [39, 50] bears some interesting analogy to ours. This analogy is further discussed in Sec. 4.6.

**Our Contribution** In this work, we take a route different than that of [49] and [47]. As in the recently published work of [39], we, too, relate the identifiability

bility of the model to the positive-definiteness of the FIM. We provide a complete and rigorous proof for the identifiability of the multidimensional model. Our proof is based on the analysis of the FIM, which has been obtained from the previously-mentioned optimal performance analysis. We obtain conditions which guarantee uniqueness and identifiability of the model. These are also the necessary and sufficient conditions for the uniqueness (up to a certain block-diagonal scaling) solution to JBD of a set of mixtures of real, positive-definite block-diagonal JBD-irreducible symmetric matrices. By “JBD-irreducible” we mean that the set of block-diagonal matrices is constructed such that it cannot be further jointly block-diagonalized into smaller blocks. JBD-irreducibility is discussed in Sec. 2.3.1.

The main results have been published in Lahat et al. [51].

### 1.2.3 Joint Block Diagonalization Algorithms

In order to validate our theoretical results, the next step is to develop a JBD algorithm which minimizes the contrast function which we have derived. One partition of JBD algorithms is into unitary and non-unitary ones. Since our ML separation criterion is non-unitary, and our data is real, we focus on non-orthogonal algorithms.

There are two common criteria which define approximate<sup>1</sup> JBD of a set of matrices. One is the subspace-fitting least squares (LS) criterion, which seeks the best block-diagonal model which fits the original data. The second is the quadratic criterion, which quantifies the residual off block-diagonal terms. Non-orthogonal JBD by the subspace-fitting LS criterion is discussed, for example, in [52, 53]. The quadratic criterion is minimized using a non-orthogonal algorithm, for example, by [53, 54]. For the completeness of the picture, we mention also the JBD algorithm of [55]; it is based on an error-controlling parameter and on a certain randomized sampling scheme, which eventually produces a unitary de-mixing matrix.

However, all these error measures for JBD are different than our Kullback-Leibler divergence (KLD)-based criterion. A KLD-based criterion for JBD, in the context of source separation, has first been suggested by Bousbia-Salah et al. [56], in order to separate one-dimensional sources from their convolutive mixture. However, [56] do not specify the algorithm used to minimize their criterion. The fast algorithm for JD via KLD minimization, suggested by Pham [57], is extended for JBD of cyclostationary sources in [58]. However, to the best of our knowledge, an algorithm which minimizes our KLD-based criterion, which is equivalent to the (approximate) JBD

---

<sup>1</sup>In general, all JBD criteria (JD included) are designed such that they achieve exact JBD, up to unavoidable inherent indeterminacies, when the data can be exactly jointly block-diagonalized. They differ in the case that exact JBD cannot be achieved. It is only for simplicity of presentation that we omit the adjective “approximate”, in the sequel

of a set of mixtures of positive-definite symmetric JBD-irreducible block-diagonal matrices, could not be found in the literature.

**Our Contribution** For the first time, algorithms which minimize directly the KLD-based criterion for the block-diagonal case are derived. One algorithm is based on the relative gradient (also known as the *natural gradient* [59]) and the other one on a quasi-Newton method.

These algorithms have been published in Lahat et al. [60].

### 1.2.4 Separation of Multidimensional Data Using a One-Dimensional Model

Another aspect of our work is blind separation of multidimensional data, when the separation procedure consists of two steps. In the first step, the data is assigned a one-dimensional model. In the second step, called a clustering step, the one-dimensional output is assigned into groups, representing the multidimensional components. In this component separation scheme, the number of one-dimensional components in the first step is equal to the total number of signal-space dimensions.

The idea of resolving mixed multidimensional data in terms of blind source subspace separation was first demonstrated in [9], on fetal ECG recordings. Cardoso [10] has been the first to raise the question *how* ICA algorithms, designed to extract one-dimensional components, behave when processing a mixture of multidimensional independent components. That is, to which extent they are able to perform the first clustering step. Thenceforth, various works have been dedicated to the use of one-dimensional models and algorithms for the separation of multidimensional data, as well as to the study of their performance.

The relationship between blind separation methods and JBD of a set of matrices has been established, for example, in [56, 61, 24, 58, 54, 49, 12, 45]. An algorithm for ISA through JBD of cumulant matrices, where JBD is performed by JD and permutation-recovery, is described in [24]. In [62], an acoustic blind deconvolution problem is solved using JD of a set of covariance matrices, followed by clustering the filtered sources. For the case in which the data can be exactly jointly block-diagonalized, it is conjectured in [63], and supported by numerical simulations therein, that exact JBD can be achieved by JD followed by permutation resolving. This conjecture is partially proved by [24]. Conditions under which the two-step procedure of ICA and then clustering is sufficient for separation are discussed in [17, 28]. The analysis in [17, 28] is valid for i.i.d. pre-whitened data with at most one Gaussian variable, and specific distribution types and models; this

data model is different than the one used in this thesis. Tichavský et al. [64] prove that when the input consists of several symmetric block-diagonal matrices which have been mixed by the same invertible matrix, the “uniformly weighted exhaustive diagonalization with Gauss iterations” algorithm [65] can achieve exact JBD.

In some cases, one-dimensional methods now have multidimensional versions. Examples are the ISA method [8] and the FastISA [19] algorithm, which extends FastICA [66]; Nion’s non-unitary JBD algorithm [52], which minimizes a subspace-fitting LS criterion, and can be regarded as an extension of, e.g., [67, 68]; Ghennioui et al. [54], who suggest a non-unitary JBD algorithm to solve a quadratic criterion, and can be regarded as an extension of, e.g., [69, 70]; and Zhang et al. [71], who augment the stability of the quadratic criterion of [54] using a penalty term.

Tichavský et al. [64] compare, by numerical simulations, several JD and JBD algorithms on sets of mixed block-diagonal symmetric matrices. Koldovský et al. [72] present an extensive comparison of the performance of various JD and JBD, as well as ISA and ICA, algorithms for an acoustic convolutive BSS problem, using numerical simulations. Koldovský et al. [72] also report comparing multidimensional algorithms with their one-dimensional counterparts for the acoustic convolutive problem. These two works, [64] and [72], do not reveal a clear advantage of one approach over the other. Apart from the examples in [72], the gain expected (if at all) due to using the multidimensional version over its 1D counterpart on the same data has not yet been quantified, numerically or analytically.

**Our Contribution** For the first time, closed-form expressions for the expected MSE of component separation under the one-dimensional scheme are derived. Together with our closed-form expressions for the optimal case, this is the first time that one can calculate numerically, using analytical expressions only, the expected gain due to using the correct multidimensional model. We conjecture that the expected performance gain is always in favour of the method which uses the correct multidimensional model.

Results for the case of only two components, one of them one-dimensional and the other multidimensional, have been presented in Lahat et al. [73]. The results in this thesis are for any number of components and of any dimension.

### 1.2.5 Overdetermined Multidimensional ICA: Application to CMB Observations

Now that we have both a theoretical understanding of the optimal performance of second-order MICA, as well as an algorithm which achieves the theoretical results

under ideal conditions, it is interesting to test these methods in more realistic conditions. In this work, we demonstrate our results on astrophysical data. Namely, observations of the cosmic microwave background radiation (CMB). A short review of the CMB and its observations, as well as a more detailed explanation about this specific data model, is deferred to Sec. 7.2.

In this example, which is based on realistic sky simulations, the observations are composed of an instantaneous mixture of two components: one which follows the theoretical model (the CMB), and one which does not (the Galactic emission). Since our algorithms and performance analysis are based on a-priori knowledge of the component dimensions, then the first step in our analysis of this data set is to determine an effective dimension for the Galactic emission. Effective, since it turns out that the number of different cosmological processes which constitute what we call the Galactic emission component does not necessarily reflect the dimension required for appropriate component separation. We thus search for an integer number of dimensions which contains most of the energy of this component, while keeping this number small enough to allow for mathematical tractability and robustness within the framework of our model. In this application, there is no single number which is the “correct” answer, as opposed to the synthetic examples of the previous chapters. In this setup, the *total* number of statistically independent components is assumed to be known and set to two.

Various methods have been proposed for dealing with multidimensional data of unknown dimension. For example, the JBD algorithm of [55] should be able to find the finest block-diagonal structure, given an appropriate error-controlling parameter. Two methods to recover the actual block dimensions found by the algorithm of [55] are proposed by [12]. One is by reading a thresholded version of a matrix, whose entries correspond to the level of statistical dependence between any two pairs of components, as the adjacency matrix of a graph. The other method is based on viewing the sources as nodes of a hypergraph and then removing all hyperedges with weights lower than a certain threshold, where the weights are based on 4th order cumulants. Theis [24] deals with finding the correct block dimensions in the context of JBD algorithms, where in the first step JD is performed. In the second step, the optimal permutation should be found by iteratively permuting columns and rows in order to guarantee that all non-zeros of a matrix, which is constructed from the output of the JD algorithm, are clustered along the main diagonal as closely as possible. The algorithm proposed by [24] uses a threshold, whose choice is a matter of empirical trials. This scheme eventually recovers not only the desired permutation but also the block sizes. Bach and Jordan [29] propose, within the

framework of graphical models, an approach denoted as TCA. In this approach, learning the best possible non-spanning tree that fits a given distribution, provides a way to learn the number and size of the clusters for that particular distribution. A different dimension detection scheme is proposed by [27], based on the idea that using a wrong dimension in the algorithm leads to a larger variance of the output, when tested on repeated trials with different initializations.

In this work, we propose a different dimension detection scheme. Our scheme is based on singling-out the contribution of the model error (i.e., wrong choice of dimension) from the contribution of other type of errors (e.g., finite-data errors) to the figure of merit. When the correct model dimension is detected, the contribution of the model error becomes negligible with respect to the contribution of the other errors.

Besides the opportunity to demonstrate a method for determining the dimension of a component, this application allows us to extend our analysis to the overdetermined case, of more detectors than signal-space dimensions. This mission is achieved via principal component analysis (PCA). PCA has already been used to tread overdetermined MICA as early as [9].

Another variation of our theoretical analysis which is illustrated using the CMB application is the option to perform the entire analysis in the frequency domain. As demonstrated by [40, 74], in full analogy to the time-domain procedure, second-order BSS can also be performed in the frequency-domain on non-white stationary processes, due to the asymptotic properties of the Fourier coefficients. In this application, the CMB is a stationary field over the sky. That is, a two-dimensional spherical surface. It is through the spherical harmonic transform (SHT) (see Sec. 7.4) that the observations obtain their (approximately) piecewise-stationary form, required for our analysis.

The mapping between the temporal (or spatial) and spectral domains is linear and one-to-one. Therefore, our method, although formulated in terms of non-stationarity, can be readily applied also to stationary data. Applying our method to stationary data can be done either in time (or space, in the case of astrophysical observations), using the lagged covariance matrices (as done, e.g., in [41]), or in the spectral domain, in which the samples are non-stationary and decorrelated. The equivalence between these two domains has been shown, for example, by Pham and Garat [40]. The spectral domain may be, for example, the Fourier (as in [40]) or the spherical harmonics space. The latter is the way which was chosen for the astrophysical application in Chapter 7. Hence, although, for simplicity of presentation, our analysis is formulated in terms of non-stationarity, the sample index  $t$  is

not restricted to “time”. Similarly, the “covariance matrices”  $\mathbf{R}_X^{(t)}$  may be readily regarded as “lagged covariance matrices”, “cross-spectrum matrices” or any other appropriate statistics.

Part of these results can be found in Lahat et al. [75].

**Our Contribution** From the theoretical aspect, we extend our performance analysis to the overdetermined case. We recast our analysis into the spherical harmonics domain and show that our analysis is valid also in the spectral domain. We show that our methods are robust to small discrepancies between model and data. We demonstrate that it is possible to determine the (effective) number of dimensions of a component, even if this dimension is not known a-priori, and even if this dimension is not well-defined from the physical point of view, using only the methods presented in this thesis.

From the astrophysical aspect, we provide a preliminary test of feasibility. We show that our algorithms are capable of separating cosmological components in adequate scenarios. We demonstrate, for the first time, that a theoretical closed-form expression for the error can indeed predict, to a reasonable precision, the MSE in component separation of astrophysical emissions via MICA.

Most important, this realistic example demonstrates that our theoretical analysis and algorithms are likely to be useful in further applications and provide significant contribution to processing multidimensional data.

## 1.3 Thesis Outline

The outline of this thesis is as follows.

In Chapter 2 we describe the data model and the basic assumptions used throughout this work. We present the concept of multidimensional components as an alternative to the more widely used dependent sources, where the former are free from certain indeterminacies which the latter suffer from. We define the piecewise-stationary data model and give a concrete meaning to its multidimensionality. We derive a likelihood function, which is the basis for the statistical analysis in this work. Based on the likelihood, we derive the associated estimating equations for the Gaussian case. These equations are the basis for the small-error analysis in this work.

Chapter 3 is devoted to the error analysis of blind separation of multidimensional components based on second-order statistics, in a piecewise-stationary model. We obtain closed-form expressions for the FIM and the CRLB of the de-mixing param-

eters, as well as for the MSE of the component estimates. The derived MSE is valid also for non-Gaussian data. Our analysis is supported by numerical experiments and its performance is compared to classical ICA/BSS in various dependence scenarios.

In Chapter 4, we discuss the identifiability of second-order MICA. By analyzing the FIM, which was derived in Chapter 3, we derive necessary and sufficient conditions on the uniqueness and the identifiability the model. These are also the sufficient and necessary conditions for JBD of a set of mixtures of real positive-definite symmetric block-diagonal JBD-irreducible matrices to be unique (up to a certain block-diagonal scaling matrix).

In Chapter 5, two non-orthogonal algorithms which minimize the Kullback-Leibler-induced divergence between a set of real positive-definite symmetric matrices and a block-diagonal transformation thereof are suggested. One algorithm is based on the relative gradient (RG), and the other is based on a quasi-Newton (QN) method. Simulations demonstrate the convergence properties of the suggested algorithms.

Chapter 6 deals with the separation of multidimensional data, when the separation procedure consists of two steps. In the first step, the data is assigned a model designed originally for one-dimensional data. In the second phase, called a clustering step, the one-dimensional output is assigned into groups, representing the multidimensional components. We prove that for piecewise stationary data, and when only second-order statistics are used, this form of separation is suboptimal. In particular, we obtain a closed-form expression for the MSE for such separation, an expression which is based only on the model parameters. By comparing this expression with the optimal MSE, which was derived in Chapter 3, one can obtain the expected gain directly from the model parameters, without resorting to numerical simulations or Monte-Carlo trials. The derived MSE is valid also for non-Gaussian data. Our analysis is supported by numerical experiments. In addition, we demonstrate the theoretical gain in the accuracy of component recovery in the presence of multidimensional components for several dependence scenarios.

Chapter 7 deals with overdetermined MICA and its performance analysis, applied to multichannel observations of the CMB. We provide the mathematical preliminaries required for processing astrophysical observations taken over the celestial sphere, in the frequency domain. We then discuss the adaptations of our theoretical model to the special nature of this application. Namely, determining the components dimension, calculating spectral density matrices and extending the analysis to the overdetermined case by a dimension reduction scheme. We rewrite our expression for the figure of merit in a way which can reflect the component estimate's MSE in

each domain and observation channel. We discuss some of the special properties of CMB observations, such as the fact that there exists only one realization of the astrophysical processes which underlie the CMB observations, the non-stationarity of the Galactic emission component, and how these properties affect our analysis. Finally, we support our theoretical analysis and methods by simulated observations of the CMB. We show that despite various discrepancies between the theoretical model and the data, the basic assumption of statistical decorrelation of the multidimensional components is sufficiently robust to give a good fit between our predictions and the empirical results.



# Chapter 2

## Mathematical Preliminaries

### 2.1 Outline

In this chapter, we present the foundations on which the theoretical analysis in this thesis is based. The outline of this chapter is as follows. In Sec. 2.2 we present the two points of view, of dependent sources vs. multidimensional components. In Sec. 2.3 we describe our statistical model. We derive a contrast function, whose minimization is equivalent to the maximum likelihood (ML) solution for Gaussian piecewise stationary data. We show that this minimization can be obtained by joint block diagonalization (JBD) of a set of sample covariance matrices. We then derive the estimating equations, whose component-wise form forms the basis for the error analysis in this work.

### 2.2 Dependent Sources vs. Multidimensional Components

In their most basic setting, independent component analysis (ICA) and blind source separation (BSS) aim at extracting  $m$  mutually independent elements from  $m$  observed mixtures. The model is of  $T$  observations of an  $m \times 1$  vector  $\mathbf{x}(t)$ , modeled as

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \quad 1 \leq t \leq T \quad (2.1)$$

where  $\mathbf{A}$  is an  $m \times m$  full-rank matrix and  $\mathbf{s}(t)$  is a vector of independent *sources*. A natural extension of practical interest is to assume that the  $m$  sources can be partitioned into  $n \leq m$  groups with the sources of different groups being statistically independent while the sources in the same group are not independent and cannot be made independent by any linear transform on  $\mathbf{s}(t)$ . In the following, we use the

term *dependent sources* to indicate such a source model. As we shall see shortly, the multiplicative model (2.1) of dependent sources, which suffers from inherent indeterminacies, is equivalent to the additive model [10]

$$\mathbf{x}(t) = \sum_{i=1}^n \mathbf{x}_i(t) \quad (2.2)$$

of *multidimensional components*, in which these indeterminacies are avoided.

The separation problem of interest can be stated by partitioning the source vector  $\mathbf{s}(t)$  and matrix  $\mathbf{A}$  as

$$\mathbf{A} = [\mathbf{A}_1, \dots, \mathbf{A}_n], \quad \mathbf{s}(t) = [\mathbf{s}_1^\dagger(t), \dots, \mathbf{s}_n^\dagger(t)]^\dagger, \quad (2.3)$$

where  $\mathbf{A}_i$ , the  $i$ th column block of  $\mathbf{A}$ , has dimension  $m \times m_i$ , vector  $\mathbf{s}_i(t)$  has dimension  $m_i \times 1$  and  $\sum_{i=1}^n m_i = m$ . Given the block-pattern  $\mathbf{m} \triangleq [m_1, \dots, m_n]^\dagger$  and the observations  $\mathbf{x}(t)$ , the problem of blind separation of dependent sources is that of finding matrices  $\mathbf{A}_i$  such that  $\mathbf{A}$  is full-rank and such that the corresponding source vectors  $\mathbf{s}_1(t), \dots, \mathbf{s}_n(t)$  are *as independent as possible*. This notion is given a definite meaning in Sec. 2.3, where we set up a simple statistical model which, via its likelihood function, yields a quantitative measure of independence. However, before we write down a likelihood function, it is necessary to discuss the indeterminacies inherent to the blind separation of dependent sources and to explain how these indeterminacies lead to the alternate point of view of multidimensional ICA (MICA).

With the partition (2.3), the multiplicative, source-mixing model (2.1) can also be written as an additive model (2.2), where we define the  $i$ th component  $\mathbf{x}_i(t)$  as the  $m \times 1$  vector

$$\mathbf{x}_i(t) = \mathbf{A}_i \mathbf{s}_i(t). \quad (2.4)$$

In a blind context, the component vector  $\mathbf{x}_i(t)$  is better defined than the source vector  $\mathbf{s}_i(t)$ . Indeed, for any  $m_i \times m_i$  invertible matrix  $\mathbf{Z}_i$ , the pair  $(\mathbf{A}_i, \mathbf{s}_i(t))$  and the pair  $(\mathbf{A}_i \mathbf{Z}_i^{-1}, \mathbf{Z}_i \mathbf{s}_i(t))$  contribute the same quantity  $\mathbf{x}_i(t) = \mathbf{A}_i \mathbf{Z}_i^{-1} \mathbf{Z}_i \mathbf{s}_i(t) = \mathbf{A}_i \mathbf{s}_i(t)$  to the observations. It is thus impossible to discriminate between the representation of a component  $\mathbf{x}_i(t)$  by the pair  $(\mathbf{A}_i, \mathbf{s}_i(t))$  and by the pair  $(\mathbf{A}_i \mathbf{Z}_i^{-1}, \mathbf{Z}_i \mathbf{s}_i(t))$ . Therefore, matrix  $\mathbf{A}_i$  can, at best, be blindly identified only up to right multiplication by an invertible  $m_i \times m_i$  matrix. This is the familiar scale indeterminacy of standard ICA, carried over to dimension  $m_i$ , with an  $m_i \times m_i$  matrix factor  $\mathbf{Z}_i$  instead of a simple scalar factor.

Since matrix  $\mathbf{A}_i$  is determined only up to a right factor  $\mathbf{Z}_i$ , only its *column space*,  $\text{Span}(\mathbf{A}_i)$ , can be blindly identified. It is thus useful to introduce the separating

*projectors*: these are the  $m \times m$  oblique projection matrices  $\mathbf{P}_i$  onto  $\text{Span}(\mathbf{A}_i)$  along  $\text{Span}(\mathbf{A}_j)$  for all  $j \neq i$ . By definition, they satisfy  $\mathbf{P}_i \mathbf{A}_j = \delta_{ij} \mathbf{A}_i$ , are unaffected if  $\mathbf{A}_i$  is changed into  $\mathbf{A}_i \mathbf{Z}_i^{-1}$ , and allow one to write

$$\mathbf{x}_i(t) = \mathbf{P}_i \mathbf{x}(t). \quad (2.5)$$

The notation (2.5) is the geometric counterpart of (2.4). The set of  $n$  unambiguous oblique projections  $\mathbf{P}_i$  is the matrix-free counterpart of the inverse matrix  $\mathbf{B} = \mathbf{A}^{-1}$ .

For later reference we mention that if  $\mathbf{B}$  is partitioned into  $n$  horizontal blocks, where the  $i$ th block  $\mathbf{B}_i$  has dimension  $m_i \times m$ , then the  $i$ th oblique projector is given by

$$\mathbf{P}_i = \mathbf{A}_i \mathbf{B}_i. \quad (2.6)$$

We also define the orthogonal projection matrix  $\mathbf{\Pi}_i$  onto  $\text{Span}(\mathbf{A}_i)$ . That is,

$$\mathbf{\Pi}_i = \mathbf{A}_i (\mathbf{A}_i^\dagger \mathbf{A}_i)^{-1} \mathbf{A}_i^\dagger = \mathbf{A}_i \mathbf{A}_i^\sharp, \quad (2.7)$$

denoting by  $\mathbf{A}_i^\sharp = (\mathbf{A}_i^\dagger \mathbf{A}_i)^{-1} \mathbf{A}_i^\dagger$  the Moore-Penrose pseudoinverse ((A.2), Appendix A) of  $\mathbf{A}_i$ . Obviously, projector  $\mathbf{\Pi}_i$  is unaffected if  $\mathbf{A}_i$  is changed into  $\mathbf{A}_i \mathbf{Z}_i^{-1}$ .

In summary, the source separation model based on a mixing matrix is recast as a component separation model (2.2) where the  $i$ th component  $\mathbf{x}_i(t)$  is restricted to an  $m_i$ -dimensional subspace, represented by the uniquely-defined orthogonal projector  $\mathbf{\Pi}_i$ . The  $i$ th component is recovered via (2.5) using the oblique projector  $\mathbf{P}_i$ .

As a final note in this section, we would like to emphasize that the component perspective goes beyond the mere avoidance of scale indeterminacy. The various examples of multidimensional data in Chapter 1 demonstrate that in a multidimensional setup, a “mixing matrix” and “dependent sources” may not always have a physical interpretation of their own.

## 2.3 Model, Likelihood and Contrast Function

We derive a likelihood function for the separation of dependent sources by generalizing the Gaussian piecewise stationary model of Pham and Cardoso [42]. We show, following the same guidelines as [42], how this likelihood yields a contrast function for separating dependent sources, which is a JBD criterion. We then establish the estimating equations for the mixing matrix  $\mathbf{A}$ , that is, the equations satisfied by its ML value. Finally, we recast these equations in terms of the component model parameters, the oblique projections  $\mathbf{P}_i$ .

### 2.3.1 Piecewise Stationary Model

Let us consider a piecewise stationary model as follows. The observation interval  $[1, T]$  is partitioned into  $Q$  domains  $\mathcal{D}_q$ ,  $q = 1, \dots, Q$ , where domain  $q$  contains  $n_q$  samples, so that  $\sum_{q=1}^Q n_q = T$ . We assume that  $\mathbf{s}(t)$  is independent of  $\mathbf{s}(t')$  if  $t \neq t'$  and that, for any  $t \in \mathcal{D}_q$ ,  $\mathbf{s}(t)$  has zero mean and covariance matrix  $\mathbf{R}_S^{(q)}$ . The linear model (2.1) implies that

$$\mathbf{R}_X^{(q)} = \mathbf{A} \mathbf{R}_S^{(q)} \mathbf{A}^\dagger, \quad (2.8)$$

where  $\mathbf{R}_X^{(q)}$  is the covariance matrix of  $\mathbf{x}(t)$  for  $t \in \mathcal{D}_q$ . The empirical counterpart and natural estimate of  $\mathbf{R}_X^{(q)}$  is

$$\overline{\mathbf{R}}_X^{(q)} \triangleq \frac{1}{n_q} \sum_{t \in \mathcal{D}_q} \mathbf{x}(t) \mathbf{x}^\dagger(t). \quad (2.9)$$

The model of dependent sources, discussed in Sec. 2.2, corresponds to the block-diagonal structure

$$\mathbf{R}_S^{(q)} \triangleq \begin{bmatrix} \mathbf{R}_{S,11}^{(q)} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_{S,nn}^{(q)} \end{bmatrix}, \quad (2.10)$$

where  $\mathbf{R}_{S,ii}^{(q)}$  is the covariance matrix of  $\mathbf{s}_i(t)$  for  $t \in \mathcal{D}_q$ . The set  $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$  is such that it cannot be further jointly block-diagonalized into smaller blocks. Hence,  $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$  is *JBD-irreducible* (this notion of irreducibility is analogous to that proposed by [24, 76, 46, 77]).

In the following, we use  $\text{bdiag}\{\cdot, \dots, \cdot\}$  to denote a block-diagonal matrix constructed from the set of matrices in brackets. Therefore, (2.10) can be rewritten as

$$\mathbf{R}_S^{(q)} = \text{bdiag}\{\mathbf{R}_{S,11}^{(q)}, \dots, \mathbf{R}_{S,nn}^{(q)}\}.$$

We shall also use the related notation  $\text{bdiag}_m\{\mathbf{M}\}$  which, given an  $m \times m$  matrix  $\mathbf{M}$ , returns the block-diagonal matrix with block-pattern  $\mathbf{m}$  which has the same diagonal blocks as  $\mathbf{M}$  and has zeros in the off-diagonal blocks.

### 2.3.2 Likelihood

If  $\mathbf{s}(t)$  is normally distributed, then the log-likelihood for the model just described is

$$\log p(\{\mathbf{x}(t)\}_{t=1}^T) = \sum_{t=1}^T \log p(\mathbf{x}(t))$$

$$\begin{aligned}
&= - \sum_{t=1}^T \frac{1}{2} (\log \det 2\pi \mathbf{R}_X^{(t)} + \mathbf{x}^\dagger(t) \mathbf{R}_X^{- (t)} \mathbf{x}(t)) \\
&= - \sum_{q=1}^Q n_q \frac{1}{2} (\log \det 2\pi \mathbf{R}_X^{(q)} + \text{tr} \{ \overline{\mathbf{R}}_X^{(q)} \mathbf{R}_X^{- (q)} \}), \quad (2.11)
\end{aligned}$$

as we now explain. The first equality comes from the assumption of independence for  $t' \neq t$ . The second equality follows from the Gaussian assumption  $\mathbf{x}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_X^{(t)})$  and uses the notation  $(\mathbf{R}_X^{(t)})^{-1} = \mathbf{R}_X^{- (t)}$ . The third equality results from the piecewise stationary model that  $\mathbf{R}^{(t)} = \mathbf{R}^{(q)}$  for  $t \in \mathcal{D}_q$  and uses the property  $\mathbf{a}^\dagger \mathbf{R} \mathbf{a} = \text{tr} \{ \mathbf{R} \mathbf{a} \mathbf{a}^\dagger \}$  for any vector  $\mathbf{a}$  and matrix  $\mathbf{R}$  of appropriate dimensions.

Using the notation

$$D(\mathbf{R}_1, \mathbf{R}_2) = \frac{1}{2} (\text{tr} \{ \mathbf{R}_1 \mathbf{R}_2^{-1} \} - \log \det(\mathbf{R}_1 \mathbf{R}_2^{-1}) - m) \quad (2.12)$$

for any two  $m \times m$  positive-definite matrices  $\mathbf{R}_1$  and  $\mathbf{R}_2$ , the log-likelihood (2.11) can be rewritten as

$$\begin{aligned}
\log p(\{\mathbf{x}(t)\}_{t=1}^T) &= - \sum_{q=1}^Q n_q D(\overline{\mathbf{R}}_X^{(q)}, \mathbf{R}_X^{(q)}) + \kappa \\
&= - \sum_{q=1}^Q n_q D(\mathbf{A}^{-1} \overline{\mathbf{R}}_X^{(q)} \mathbf{A}^{-\dagger}, \mathbf{R}_S^{(q)}) + \kappa \\
&= - T \langle D(\mathbf{A}^{-1} \overline{\mathbf{R}}_X \mathbf{A}^{-\dagger}, \mathbf{R}_S) \rangle + \kappa, \quad (2.13)
\end{aligned}$$

where  $\kappa = -\frac{1}{2}(mT + \sum_{q=1}^Q n_q \log \det 2\pi \overline{\mathbf{R}}_X^{(q)})$  denotes the term which is irrelevant to the maximization of the likelihood with respect to its parameters, since it depends only on the data, not on the model. The second equality uses (2.8) and then the invariance of (2.12) under any invertible transform: for any two positive-definite matrices  $\mathbf{R}_1, \mathbf{R}_2$  and for any invertible  $\mathbf{A}$ ,

$$D(\mathbf{R}_1, \mathbf{A} \mathbf{R}_2 \mathbf{A}^\dagger) = D(\mathbf{A}^{-1} \mathbf{R}_1 \mathbf{A}^{-\dagger}, \mathbf{R}_2). \quad (2.14)$$

The last step in (2.13) introduces the notation  $\langle \cdot \rangle$  to denote a weighted average of any sequence indexed by  $q$  with weights  $n_q$ :

$$\langle \mathbf{M} \rangle \triangleq \frac{1}{T} \sum_{q=1}^Q n_q \mathbf{M}^{(q)}. \quad (2.15)$$

### 2.3.3 Contrast Function for a Mixture of Dependent Sources

We can now focus on the case of interest: a mixture of dependent sources. The contrast function is obtained by maximizing the likelihood with respect to the nuisance parameters  $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$  for fixed  $\mathbf{A}$ . Following the derivation in Appendix 2.A, the ML estimate of  $\mathbf{R}_S^{(q)}$  is  $\text{bdiag}_m\{\mathbf{A}^{-1}\overline{\mathbf{R}}_X^{(q)}\mathbf{A}^{-\dagger}\}$  and

$$\max_{\mathbf{R}_S^{(q)}} \log p(\{\mathbf{x}(t)\}_{t=1}^T; \mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q) = -T C(\mathbf{A}) + \kappa, \quad (2.16)$$

where we define the *contrast function* [1]

$$C(\mathbf{A}) \triangleq \langle D(\mathbf{A}^{-1}\overline{\mathbf{R}}_X\mathbf{A}^{-\dagger}, \text{bdiag}_m\{\mathbf{A}^{-1}\overline{\mathbf{R}}_X\mathbf{A}^{-\dagger}\}) \rangle \quad (2.17)$$

and where, for brevity, the dependence of  $C(\mathbf{A})$  on the data via  $\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$  is not denoted explicitly. Note that (2.17) is the multidimensional analogue of its one-dimensional counterpart in [42].

The scalar  $D(\mathbf{R}_1, \mathbf{R}_2)$ , defined in (2.12), is equal to the Kullback-Leibler divergence (KLD) between the distributions  $\mathcal{N}(\mathbf{0}, \mathbf{R}_1)$  and  $\mathcal{N}(\mathbf{0}, \mathbf{R}_2)$  [78] and thus is a measure of mismatch between two positive-definite matrices  $\mathbf{R}_1$  and  $\mathbf{R}_2$ . Henceforth, we refer to  $D(\cdot, \cdot)$  as the Kullback-Leibler-induced divergence between two positive-definite symmetric matrices. Therefore, in our piecewise stationary model, maximizing (2.13) is equivalent to minimizing the average mismatch  $\langle D(\overline{\mathbf{R}}_X, \mathbf{R}_X) \rangle$  between the sample covariance matrices and their expected counterparts. Since  $D(\mathbf{R}, \text{bdiag}_m\{\mathbf{R}\}) \geq 0$  with equality if and only if  $\mathbf{R}$  is block-diagonal with block-pattern  $\mathbf{m}$ , then, for any positive-definite matrix  $\mathbf{R}$ , the divergence  $D(\mathbf{R}, \text{bdiag}_m\{\mathbf{R}\})$  is a measure of the block-diagonality of  $\mathbf{R}$ . Therefore, minimizing  $C(\mathbf{A})$  can be understood as *joint block diagonalization* of the set of covariance matrices  $\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$  by matrix  $\mathbf{A}^{-1}$  (see also [56, 58] and footnote 2 on page 70).

### 2.3.4 Estimating Equations in Terms of the Mixing Matrix

The next step is to solve for the mixing matrix  $\mathbf{A}$ . This is obtained by maximizing the log-likelihood (2.13) and thus also minimizing the contrast function (2.17). For this purpose, we calculate the derivative of the likelihood function  $\phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q) \triangleq \log p(\{\mathbf{x}(t)\}_{t=1}^T)$  with respect to  $\mathbf{A}$ , for fixed  $\mathbf{R}_S^{(q)}$  (we omit the dependence of  $\phi$  on the data, for brevity). The first-order variation of  $\phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q)$  when  $\mathbf{A}$  is replaced by  $\mathbf{A}(\mathbf{I} + \mathbf{E})$  (where  $\mathbf{I}$  denotes the identity matrix) can always be expressed by the

Taylor expansion

$$\begin{aligned} \phi(\mathbf{A}(\mathbf{I} + \mathbf{E}), \{\mathbf{R}_S^{(q)}\}_{q=1}^Q) &= \phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q) \\ &+ \text{tr} \left\{ (\nabla \phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q))^\dagger \mathbf{E} \right\} + \text{higher-order terms in } \mathbf{E}, \end{aligned} \quad (2.18)$$

for some  $m \times m$  matrix  $\nabla \phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q)$ , called the relative gradient (RG) of  $\phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q)$  with respect to  $\mathbf{A}$ . Similarly to the derivation for the one-dimensional case in [42], one obtains that

$$\nabla \phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q) = \langle \mathbf{R}_S^{-1} \mathbf{A}^{-1} \overline{\mathbf{R}}_X \mathbf{A}^{-\dagger} \rangle - \mathbf{I}. \quad (2.19)$$

In order to obtain the ML estimate of  $\mathbf{A}$ , we equate (2.19) to zero. Since (2.19) depends on the nuisance parameters, we can now replace  $\mathbf{R}_S^{(q)}$  with its ML estimate, derived in Sec. 2.3.3. This procedure yields the *estimating equations*

$$\langle \text{bdiag}_m^{-1} \{ \mathbf{A}^{-1} \overline{\mathbf{R}}_X \mathbf{A}^{-\dagger} \} \mathbf{A}^{-1} \overline{\mathbf{R}}_X \mathbf{A}^{-\dagger} \rangle = \mathbf{I}, \quad (2.20)$$

where  $\text{bdiag}_m^{-1} \{ \cdot \}$  implies  $(\text{bdiag}_m \{ \cdot \})^{-1}$ . It can be shown that  $\nabla C(\mathbf{A})$ , the first-order variation of  $C(\mathbf{A})$ , derived similarly to (2.18), obeys

$$\nabla \phi(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q) \Big|_{\mathbf{R}_S^{(q)} = \text{bdiag}_m \{ \mathbf{A}^{-1} \overline{\mathbf{R}}_X^{(q)} \mathbf{A}^{-\dagger} \}} = -\nabla C(\mathbf{A}). \quad (2.21)$$

Therefore, an ML estimate of  $\mathbf{A}$  is not only a minimizer of the contrast function (2.17) but also a solution of the estimating equations (2.20). The estimating equations (2.20) read block-wise

$$\langle ([\mathbf{A}^{-1} \overline{\mathbf{R}}_X \mathbf{A}^{-\dagger}]_{ii})^{-1} [\mathbf{A}^{-1} \overline{\mathbf{R}}_X \mathbf{A}^{-\dagger}]_{ij} \rangle = \mathbf{0}_{m_i \times m_j} \quad j \neq i, \quad (2.22)$$

where  $i, j$  are understood as *block* indices. Using the horizontal blocks  $\mathbf{B}_i$  of matrix  $\mathbf{B} = \mathbf{A}^{-1}$ , (2.22) is rewritten as

$$\langle (\mathbf{B}_i \overline{\mathbf{R}}_X \mathbf{B}_i^\dagger)^{-1} (\mathbf{B}_i \overline{\mathbf{R}}_X \mathbf{B}_j^\dagger) \rangle = \mathbf{0}_{m_i \times m_j} \quad j \neq i. \quad (2.23)$$

Note that the  $(i, i)$ th block of (2.20), that is,  $i = j$  of (2.22), degenerates into the identity matrix. This implies that the diagonal blocks  $i = j$  do not yield any constraints, reflecting the indeterminacy discussed in Sec. 2.2.

### 2.3.5 Estimating Equations in Terms of the Projectors

The estimating equations (2.23) can also be expressed as conditions on the oblique projectors  $\mathbf{P}_i$ . To do so, we multiply (2.23) on the left by  $\mathbf{A}_i^{\sharp\dagger}$  and on the right by  $\mathbf{A}_j^\dagger$ . In the middle we insert  $\mathbf{A}_i^\sharp \mathbf{A}_i = \mathbf{I}$ , to obtain

$$\langle \mathbf{A}_i^{\sharp\dagger} (\mathbf{B}_i \overline{\mathbf{R}}_X \mathbf{B}_i^\dagger)^{-1} \mathbf{A}_i^\sharp \mathbf{A}_i (\mathbf{B}_i \overline{\mathbf{R}}_X \mathbf{B}_j^\dagger) \mathbf{A}_j^\dagger \rangle = \mathbf{0}_{m \times m}. \quad (2.24)$$

We split the expression inside the angular brackets into two factors which can be re-expressed as follows. The leftmost factor is

$$\mathbf{A}_i^{\sharp\dagger} (\mathbf{B}_i \overline{\mathbf{R}}_X^{(q)} \mathbf{B}_i^\dagger)^{-1} \mathbf{A}_i^\sharp = (\mathbf{A}_i \mathbf{B}_i \overline{\mathbf{R}}_X^{(q)} \mathbf{B}_i^\dagger \mathbf{A}_i^\dagger)^\sharp = (\mathbf{P}_i \overline{\mathbf{R}}_X^{(q)} \mathbf{P}_i^\dagger)^\sharp, \quad (2.25)$$

where the first equality uses the property that for any invertible  $n \times n$  matrix  $\mathbf{M}$  and for any  $m \times n$ ,  $m \geq n$  rank- $n$  matrix  $\mathbf{V}$ ,

$$\mathbf{V}^{\sharp\dagger} \mathbf{M}^{-1} \mathbf{V}^\sharp = (\mathbf{V} \mathbf{M} \mathbf{V}^\dagger)^\sharp \quad (2.26)$$

(it is immediate to verify that (2.26) fulfills all four criteria of the Moore-Penrose pseudoinverse (A.2) in Appendix A) and the second equality uses definition (2.6) of the projectors. The rightmost factor is

$$\mathbf{A}_i (\mathbf{B}_i \overline{\mathbf{R}}_X^{(q)} \mathbf{B}_j^\dagger) \mathbf{A}_j^\dagger = \mathbf{P}_i \overline{\mathbf{R}}_X^{(q)} \mathbf{P}_j^\dagger. \quad (2.27)$$

Substituting (2.25) and (2.27) in (2.24), the estimating equations (2.23) also read

$$\langle (\mathbf{P}_i \overline{\mathbf{R}}_X \mathbf{P}_i^\dagger)^\sharp (\mathbf{P}_i \overline{\mathbf{R}}_X \mathbf{P}_j^\dagger) \rangle = \mathbf{0}_{m \times m} \quad j \neq i, \quad (2.28)$$

which is the desired form: the values of the oblique projectors for which (2.28) holds are the stationary points of the contrast function. Since the estimating equations (2.22) yield the ML estimates of  $\mathbf{A}$ , (2.28) yield the ML estimates of  $\mathbf{P}_i$ . The estimating equations (2.28) will allow us to lead the error analysis from the components' point of view, as we shall see in Sec. 3.2.

## 2.4 Summary

In this chapter, we presented the concept of BSS of multidimensional components as a new perspective on the dependent sources model. Based on a piecewise stationary model, we derived an ML-based criterion (2.17) which singles out the multi-

dimensional, unambiguous components from their sum. We have shown that the ML solution, for Gaussian data and non-orthogonal mixing matrix (that is, avoiding any whitening constraints) can be obtained by minimizing a contrast function, which is a JBD criterion. Hence, the ML solution for the de-mixing parameters can be obtained via JBD of a set of covariance matrices. By differentiating the log-likelihood, and thus also of the contrast function, with respect to their parameters, we derived the estimating equations. These are non-linear equations, which yield the ML estimate of the de-mixing parameters. The estimating equations form the basis for the error analysis in the upcoming chapters.

## 2.A Derivation of the Contrast Function

This appendix details the steps to obtain the contrast function (2.17), by maximizing the log-likelihood (2.13) with respect to  $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$ . This maximization can be obtained using Property 2.1 below, which implies that

$$\min_{\mathbf{N} \in \text{bdiag}_m} D(\mathbf{M}, \mathbf{N}) = D(\mathbf{M}, \text{bdiag}_m\{\mathbf{M}\}). \quad (2.29)$$

Therefore, for any domain  $q$ ,

$$\min_{\mathbf{R}_S^{(q)}} D(\mathbf{A}^{-1} \overline{\mathbf{R}}_X^{(q)} \mathbf{A}^{-\dagger}, \mathbf{R}_S^{(q)}) = D(\mathbf{A}^{-1} \overline{\mathbf{R}}_X^{(q)} \mathbf{A}^{-\dagger}, \text{bdiag}_m\{\mathbf{A}^{-1} \overline{\mathbf{R}}_X^{(q)} \mathbf{A}^{-\dagger}\})$$

since  $\mathbf{R}_S^{(q)}$  is block-diagonal by definition. Since the samples at different domains are statistically independent, the log-likelihood (2.13) can be maximized by optimizing each domain separately, which concludes the derivation.

**Property 2.1** (Closest Block-Diagonal Matrix). Given block-pattern  $\mathbf{m}$ , the best, in the Kullback-Leibler-induced divergence sense, block-diagonal approximation to a positive-definite matrix  $\mathbf{M}$ , is  $\text{bdiag}_m\{\mathbf{M}\}$ :

$$\arg \min_{\mathbf{N} \in \text{bdiag}_m} D(\mathbf{M}, \mathbf{N}) = \text{bdiag}_m\{\mathbf{M}\}$$

from which (2.29) follows.

*Proof.* For a given block-pattern  $\mathbf{m}$ , let  $\mathbf{M}$  and  $\mathbf{N} \triangleq \text{bdiag}_m\{\mathbf{N}\}$  be any two positive-definite matrices. Then, the Kullback-Leibler-induced divergence between

$\mathbf{M}$  and  $\text{bdiag}_m\{\mathbf{N}\}$  can be factorized into

$$D(\mathbf{M}, \text{bdiag}_m\{\mathbf{N}\}) = D(\mathbf{M}, \text{bdiag}_m\{\mathbf{M}\}) + D(\text{bdiag}_m\{\mathbf{M}\}, \text{bdiag}_m\{\mathbf{N}\}). \quad (2.30)$$

The proof of (2.30) is straightforward using the definition (2.12) of the Kullback-Leibler-induced divergence. In (2.30),  $D(\mathbf{M}, \text{bdiag}_m\{\mathbf{M}\})$  is independent of  $\mathbf{N}$ . Setting  $\mathbf{N} = \text{bdiag}_m\{\mathbf{M}\}$  zeroes (and minimizes, since  $D(\cdot, \cdot) \geq 0$ ) the rightmost term of (2.30).  $\square$

# Chapter 3

## Optimal Performance of Second-Order Multidimensional ICA

### 3.1 Outline

This chapter deals with the asymptotic error analysis of second-order multidimensional ICA (MICA) in a piecewise stationary model.

The outline of this chapter is as follows. Sec. 3.2 deals with the small-error analysis of the estimates, obtained from the minimization of the contrast function, derived in Chapter 2. The small-error analysis in this chapter is done exclusively in terms of well-defined quantities, i.e. components and projections. From the Taylor expansion of the estimating equations we obtain a closed-form expression for the error covariance of the maximum likelihood (ML) estimates of the model parameters, the oblique projections. This covariance is propagated to yield a closed-form expression for the total mean square error (MSE) in component estimation. We present the asymptotically-achievable Cramér-Rao lower bound (CRLB) and the Fisher information matrix (FIM) for this data model. Numerical examples which support our theoretical results are presented in Sec. 3.3. Numerical simulations demonstrate that our CRLB is achievable when the model assumptions hold, and that our MSE expression for component reconstruction holds also for non-Gaussian data, as predicted from our theoretical analysis. Furthermore, the simulations show a significant gain of the multidimensional approach over the cruder approach based on one-dimensional separation following a clustering of the dependent sources. The latter approach will be discussed in detail in Chapter 6.

The analysis in this chapter has been published in Lahat et al. [45].

## 3.2 Error Analysis

We turn to the error analysis of the estimates, obtained by minimizing the contrast function (2.17). Our purpose is to derive a closed-form expression for the MSE in component estimation. In Sec. 3.2.1 we define the error in component estimation. We express this error as a function of the error in the oblique projections and of the observations. The error in the oblique projections, which is due to all  $n$  components, is factorized into pairwise error terms. Following a first-order expansion of the estimating equations (2.28), we obtain in Sec. 3.2.2 an expression for the pairwise error terms, which depends on the model parameters and on the observations. In Sec. 3.2.3 we derive the covariance of the pairwise error terms and thus also of the error in the oblique projections. This derivation provides us with the FIM and CRLB for these estimated parameters, when the Gaussian model holds. Finally, in Sec. 3.2.4, based on the former results, we obtain a closed-form expression for the (normalized) MSE for component separation.

We consider an asymptotic analysis in the regime of small errors, in which the results are obtained from a first-order expansion of the estimating equations. In the following, we define asymptotic conditions as  $T \rightarrow \infty$  with  $\frac{nq}{T}$  fixed  $\forall q$ . The analysis is conducted under the assumption that the model of Sec 2.3.1 holds.

In order to avoid ambiguities, for certain parameters, a “ $\star$ ” is used to denote the “true” model quantities, and a hat denotes a value which minimizes the contrast function. For instance,  $\mathbf{A}^\star$  denotes the true mixing matrix, while  $\hat{\mathbf{A}}$  denotes a minimizer of (2.17), which is also an ML estimate if the Gaussian model holds.

### 3.2.1 Error Decomposition

A difficulty in error analysis for the multidimensional problem stems from the inability to characterize the estimation error of the mixing matrix, due to the severe indeterminacies it suffers from, as discussed in Sec. 2.2. We thus begin by defining convenient error terms. In order to focus on well-defined quantities, we consider the errors  $\delta\mathbf{P}_i$  in  $\hat{\mathbf{P}}_i$ , the ML estimates of the oblique projectors  $\mathbf{P}_i$ :

$$\delta\mathbf{P}_i = \hat{\mathbf{P}}_i - \mathbf{P}_i^\star. \quad (3.1)$$

The estimated  $i$ th component is thus

$$\hat{\mathbf{x}}_i(t) = \hat{\mathbf{P}}_i \mathbf{x}(t) = (\mathbf{P}_i^\star + \delta\mathbf{P}_i) \mathbf{x}(t) = \mathbf{x}_i(t) + \delta\mathbf{P}_i \mathbf{x}(t) \quad (3.2)$$

so that, using  $\mathbf{x}(t) = \sum_{j=1}^n \mathbf{x}_j(t)$ , the error in the  $i$ th component is factorized as

$$\widehat{\mathbf{x}}_i(t) - \mathbf{x}_i(t) = \delta \mathbf{P}_i \mathbf{x}(t) = \delta \mathbf{P}_i \sum_{j=1}^n \mathbf{x}_j(t) = \sum_{j=1}^n \delta \mathbf{P}_i \mathbf{\Pi}_j^* \mathbf{x}_j(t) = \sum_{j=1}^n \mathbf{\mathcal{E}}_{ij} \mathbf{x}_j(t), \quad (3.3)$$

where we have defined  $m \times m$  error matrices

$$\mathbf{\mathcal{E}}_{ij} \triangleq \delta \mathbf{P}_i \mathbf{\Pi}_j^* = (\widehat{\mathbf{P}}_i - \mathbf{P}_i^*) \mathbf{\Pi}_j^*. \quad (3.4)$$

The term  $\mathbf{\Pi}_j^*$  in (3.3) is inserted since it arises naturally in the derivation of the first-order expansion of the estimating equations (Appendix 3.A, (3.32)); recall that  $\mathbf{\Pi}_j^* \mathbf{x}_j(t) = \mathbf{x}_j(t) \forall j$ . The double-indexed term  $\mathbf{\mathcal{E}}_{ij}$  gives the linearized estimating equations their pairwise form, as will be seen shortly.

For  $i \neq j$ , the term  $\mathbf{\mathcal{E}}_{ij} \mathbf{x}_j(t)$  in (3.3) is called the  $(i, j)$ th-contamination error. That is, the contamination due to the  $j$ th component in the reconstruction of the  $i$ th component. The term  $\mathbf{\mathcal{E}}_{ii} \mathbf{x}_i(t)$  is called the  $i$ th-reconstruction error, since this term represents a distortion of  $\mathbf{x}_i(t)$  but not any contamination by the other components.

### 3.2.2 Influence Function

In order to evaluate the covariance of the estimation error, we first establish the first-order expansion of  $\mathbf{\mathcal{E}}_{ij}$  in terms of the finite-sample  $m \times m$  covariance matrices

$$\overline{\mathbf{R}}_{X_i X_j}^{(q)} \triangleq \frac{1}{n_q} \sum_{t \in \mathcal{D}_q} \mathbf{x}_i(t) \mathbf{x}_j^\dagger(t). \quad (3.5)$$

The key assumption for blind separation is block-decorrelation:

$$\mathbf{R}_{X_i X_j}^{(q)} \triangleq E \left\{ \overline{\mathbf{R}}_{X_i X_j}^{(q)} \right\} = \mathbf{0}_{m \times m} \quad \text{for } j \neq i. \quad (3.6)$$

However, because of finite sample size, this does not hold for its empirical counterpart, i.e.,  $\overline{\mathbf{R}}_{X_i X_j}^{(q)} \neq \mathbf{0}_{m \times m}$ . In this section, we develop the performance analysis in the regime of small errors. That is, we analyze the error terms  $\widehat{\mathbf{P}}_i - \mathbf{P}_i^*$  at first-order in  $\overline{\mathbf{R}}_{X_i X_j}^{(q)}$  when asymptotic conditions hold. From (3.4),  $\mathbf{\mathcal{E}}_{ij}$  decreases with  $T$  at the same rate as  $\delta \mathbf{P}_i$ . Assuming that asymptotic conditions hold, then  $\widehat{\mathbf{P}}_i \cong \mathbf{P}_i^*$  (see Appendix 3.A).

The first-order expansion of the estimating equations (2.28) yields (see Ap-

pendix 3.A) a set of *pairs* of equations:

$$\begin{aligned} -\langle \mathbf{R}_{X_i X_i}^\# \overline{\mathbf{R}}_{X_i X_j} \rangle &= \boldsymbol{\mathcal{E}}_{ji}^\dagger + \langle \mathbf{R}_{X_i X_i}^\# \boldsymbol{\mathcal{E}}_{ij} \mathbf{R}_{X_j X_j} \rangle + \Omega\left(\frac{1}{T}\right) \\ -\langle \mathbf{R}_{X_j X_j}^\# \overline{\mathbf{R}}_{X_j X_i} \rangle &= \boldsymbol{\mathcal{E}}_{ij}^\dagger + \langle \mathbf{R}_{X_j X_j}^\# \boldsymbol{\mathcal{E}}_{ji} \mathbf{R}_{X_i X_i} \rangle + \Omega\left(\frac{1}{T}\right) \end{aligned} \quad (3.7)$$

with one such pair of equations for each pair  $i \neq j$  of components. Equation (3.7) shows that asymptotically, for each pair of components, the projector error terms  $(\boldsymbol{\mathcal{E}}_{ij}, \boldsymbol{\mathcal{E}}_{ji})$  are related to the corresponding set of  $2Q$  matrices  $\{\overline{\mathbf{R}}_{X_i X_j}^{(q)}, \overline{\mathbf{R}}_{X_j X_i}^{(q)}\}_{q=1}^Q$ , which represents the block-decorrelation error. Such a pairwise decoupling is customary in the asymptotic analysis of blind source separation (BSS) algorithms, e.g. [40], [57].

In order to proceed, it is convenient to vectorize the matrices using the  $\text{vec}\{\cdot\}$  operator which stacks the columns of a  $p \times q$  matrix into a  $pq \times 1$  vector. The pair of equations (3.7) can thus be rewritten in matrix form as

$$\begin{bmatrix} \mathbf{g}_{ij} \\ \mathbf{g}_{ji} \end{bmatrix} = -\boldsymbol{\mathcal{H}} \cdot \begin{bmatrix} \text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\} \\ \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\} \end{bmatrix} + \Omega\left(\frac{1}{T}\right), \quad (3.8)$$

where

$$\mathbf{g}_{ij} = \langle \text{vec}\{\mathbf{R}_{X_i X_i}^\# \overline{\mathbf{R}}_{X_i X_j}\} \rangle \quad (3.9)$$

and

$$\boldsymbol{\mathcal{H}} = \begin{bmatrix} \mathbf{H}_{ij} & \boldsymbol{\mathcal{T}}_{m,m} \\ \boldsymbol{\mathcal{T}}_{m,m} & \mathbf{H}_{ji} \end{bmatrix} \quad (3.10)$$

is a symmetric matrix with

$$\mathbf{H}_{ij} = \langle \mathbf{R}_{X_j X_j} \otimes \mathbf{R}_{X_i X_i}^\# \rangle. \quad (3.11)$$

In (3.10) we have introduced the  $mn \times mn$  commutation matrix  $\boldsymbol{\mathcal{T}}_{m,n}$  [79], defined in (A.3) in Appendix A. Assuming that  $\boldsymbol{\mathcal{H}}$  is invertible (necessary and sufficient conditions are given in Sec. 4.4), then

$$\begin{bmatrix} \text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\} \\ \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\} \end{bmatrix} = -\boldsymbol{\mathcal{H}}^{-1} \begin{bmatrix} \mathbf{g}_{ij} \\ \mathbf{g}_{ji} \end{bmatrix} + \Omega\left(\frac{1}{T}\right). \quad (3.12)$$

Equation (3.12) shows how the empirical correlation between components, that is, the fact that  $\overline{\mathbf{R}}_{X_i X_j}^{(q)}$  is non-zero in finite sample size, results in non-zero errors  $\boldsymbol{\mathcal{E}}_{ij}$ . Note the similarity between (3.12) and its one-dimensional, source-wise counterpart in [42, 57]. Equation (3.12) is the desired closed-form, first-order expression for the

error terms in (3.3).

### 3.2.3 Error Covariance for the Projectors

We are almost ready now to calculate the covariance matrix of  $\delta\mathbf{P}_i$ , defined in (3.1). First, we notice that  $\delta\mathbf{P}_i = \delta\mathbf{P}_i \sum_{j=1}^n \mathbf{P}_j^* = \sum_{j=1}^n \delta\mathbf{P}_i \mathbf{\Pi}_j^* \mathbf{P}_j^* = \sum_{j=1}^n \mathcal{E}_{ij} \mathbf{P}_j^*$ , where  $\mathcal{E}_{ij}$  are given explicitly by (3.12) for  $i \neq j$ . For  $\mathcal{E}_{ii}$  one more step is required, because  $\mathcal{E}_{ii}$  is not given by (3.12). Since  $\mathbf{x}(t)$  is given (observed), we exploit it by constraining  $\sum_{j=1}^n \widehat{\mathbf{x}}_j(t) = \mathbf{x}(t)$ . Therefore,  $\sum_{j=1}^n \widehat{\mathbf{P}}_j = \mathbf{I}$ , which implies  $\sum_{j=1}^n \delta\mathbf{P}_j = \mathbf{0}_{m \times m}$ , hence  $\sum_{j=1}^n \mathcal{E}_{ji} = \mathbf{0}_{m \times m}$ . We can thus rewrite  $\delta\mathbf{P}_i = \sum_{j \neq i} (\mathcal{E}_{ij} \mathbf{P}_j^* - \mathcal{E}_{ji} \mathbf{P}_i^*)$ . Vectorizing  $\delta\mathbf{P}_i$ , its covariance matrix is given by

$$\begin{aligned} \text{Cov}(\text{vec}\{\delta\mathbf{P}_i\}) &= \text{Cov}\left(\sum_{j \neq i} ((\mathbf{P}_j^{*\dagger} \otimes \mathbf{I}) \text{vec}\{\mathcal{E}_{ij}\} - (\mathbf{P}_i^{*\dagger} \otimes \mathbf{I}) \text{vec}\{\mathcal{E}_{ji}\})\right) \\ &= \sum_{j \neq i} ((\mathbf{P}_j^{*\dagger} \otimes \mathbf{I}) \text{Cov}(\text{vec}\{\mathcal{E}_{ij}\}) (\mathbf{P}_j^* \otimes \mathbf{I}) \\ &\quad - (\mathbf{P}_j^{*\dagger} \otimes \mathbf{I}) \text{Cov}(\text{vec}\{\mathcal{E}_{ij}\}, \text{vec}\{\mathcal{E}_{ji}\}) (\mathbf{P}_i^* \otimes \mathbf{I}) \\ &\quad - (\mathbf{P}_i^{*\dagger} \otimes \mathbf{I}) \text{Cov}(\text{vec}\{\mathcal{E}_{ji}\}, \text{vec}\{\mathcal{E}_{ij}\}) (\mathbf{P}_j^* \otimes \mathbf{I}) \\ &\quad + (\mathbf{P}_i^{*\dagger} \otimes \mathbf{I}) \text{Cov}(\text{vec}\{\mathcal{E}_{ji}\}) (\mathbf{P}_i^* \otimes \mathbf{I}) + O\left(\frac{1}{T^2}\right), \end{aligned} \quad (3.13)$$

where the first equality is due to (A.1c) in Appendix A and the last equality is due to the lack of correlation between the components.

It remains now to calculate the four covariance matrices in (3.13). These will be taken from the covariance of (3.12). For this aim, we shall calculate the covariance matrix of the stochastic vector  $[\mathbf{g}_{ij}^\dagger, \mathbf{g}_{ji}^\dagger]^\dagger$ . We show in Appendix 3.B that

$$\text{Cov}\left(\begin{bmatrix} \mathbf{g}_{ij} \\ \mathbf{g}_{ji} \end{bmatrix}\right) = \frac{1}{T} \mathcal{H}_\Pi, \quad (3.14)$$

where

$$\mathcal{H}_\Pi = \begin{bmatrix} \mathbf{H}_{ij} & \mathcal{T}_{m,m}^{\Pi_{ij}} \\ \mathcal{T}_{m,m}^{\Pi_{ji}} & \mathbf{H}_{ji} \end{bmatrix} \quad (3.15)$$

and

$$\mathcal{T}_{m,m}^{\Pi_{ij}} \triangleq (\mathbf{\Pi}_j^* \otimes \mathbf{\Pi}_i^*) \mathcal{T}_{m,m} (\mathbf{\Pi}_i^* \otimes \mathbf{\Pi}_j^*). \quad (3.16)$$

Using (3.14), the covariance matrix of (3.12), for any  $i \neq j$ , is given by

$$\text{Cov} \left( \begin{bmatrix} \text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\} \\ \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\} \end{bmatrix} \right) = \frac{1}{T} \boldsymbol{\mathcal{H}}^{-1} \boldsymbol{\mathcal{H}}_{\Pi} \boldsymbol{\mathcal{H}}^{-\dagger} + O\left(\frac{1}{T^2}\right) = \frac{1}{T} \boldsymbol{\mathcal{H}}_{\Pi}^{\#} + O\left(\frac{1}{T^2}\right), \quad (3.17)$$

where the last step is proved in Appendix 3.C. Further manipulations on (3.17) (Appendix 3.D) yield, for  $i \neq j$ ,

$$\text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}) = T^{-1} (\langle \mathbf{R}_{X_j X_j} \otimes \mathbf{R}_{X_i X_i}^{\#} \rangle - \langle \mathbf{R}_{X_j X_j}^{\#} \otimes \mathbf{R}_{X_i X_i} \rangle^{\#})^{\#} + O\left(\frac{1}{T^2}\right) \quad (3.18)$$

and

$$\text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}, \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\}) = \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}) \langle \mathbf{R}_{X_j X_j}^{\#} \otimes \mathbf{R}_{X_i X_i} \rangle^{\#} \boldsymbol{\mathcal{T}}_{m,m} + O\left(\frac{1}{T^2}\right) \quad (3.19)$$

as the cross-covariance matrix between  $\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}$  and  $\text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\}$ . All the covariance matrices on the right-hand side (RHS) of (3.13) are now given in explicit form by (3.18) and (3.19), which concludes the closed-form derivation of  $\text{Cov}(\text{vec}\{\delta \mathbf{P}_i\})$ .

Matrix  $\boldsymbol{\mathcal{H}}_{\Pi}^{\#}$  and thus also  $\boldsymbol{\mathcal{H}}_{\Pi}$  have rank  $2m_i m_j$ . This rank is due to the fact that  $\mathbf{g}_{ij}$  reflects a correlation between an  $m_i$ -dimensional component and an  $m_j$ -dimensional component, as can be seen from (3.9). Hence,  $\text{Cov}(\mathbf{g}_{ij})$  is of rank  $m_i m_j$ . Due to (3.12),  $m_i m_j$  is also the rank of  $\text{Cov}(\boldsymbol{\mathcal{E}}_{ij})$ . Since (3.17) involves two such matrices, which are linearly independent, the rank of  $\boldsymbol{\mathcal{H}}_{\Pi}^{\#}$  (3.17) is twice this value. The rank of matrix  $\text{Cov}(\text{vec}\{\delta \mathbf{P}_i\})$  is, for each  $j$ , the sum of the rank of all the matrices (3.17) involved in (3.13) (if we ignore the higher-order terms, of course). Hence, its rank is  $2m_i \sum_{j \neq i} m_j$ . Both  $\text{Cov}(\text{vec}\{\delta \mathbf{P}_i\})$  and  $\boldsymbol{\mathcal{H}}_{\Pi}$  are rank-deficient.

Under the Gaussian assumption, the results in this section have the following interpretation.  $\mathbf{g}_{ij}$  is the first-order expansion of the relative gradient (RG) of the contrast function  $C(\mathbf{A})$ . From (2.16),  $T\mathbf{g}_{ij}$  is the first-order expansion of the RG (2.28) of the log-likelihood (2.11). Therefore, due to (3.14),  $T^2 \frac{1}{T} \boldsymbol{\mathcal{H}}_{\Pi} = T \boldsymbol{\mathcal{H}}_{\Pi}$  is the FIM for the pair  $(\text{vec}\{\mathbf{P}_i \boldsymbol{\Pi}_j^{\star}\}, \text{vec}\{\mathbf{P}_j \boldsymbol{\Pi}_i^{\star}\})$ . This interpretation can be verified from (3.8), since the pair  $(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}, \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\})$  denotes the estimation error of the pair  $(\text{vec}\{\mathbf{P}_i \boldsymbol{\Pi}_j^{\star}\}, \text{vec}\{\mathbf{P}_j \boldsymbol{\Pi}_i^{\star}\})$  (recall (3.4)). In (3.17) we obtain that the covariance matrix of the estimation errors is (approximately) equal to the pseudoinverse of the FIM. Therefore, (3.17) is the asymptotically achievable CRLB on the estimation of the pair  $(\text{vec}\{\mathbf{P}_i \boldsymbol{\Pi}_j^{\star}\}, \text{vec}\{\mathbf{P}_j \boldsymbol{\Pi}_i^{\star}\})$ , and (3.18) is the CRLB for  $\text{vec}\{\mathbf{P}_i \boldsymbol{\Pi}_j^{\star}\}$  alone. Since we have shown that (3.18) is the CRLB for  $\mathbf{P}_i \boldsymbol{\Pi}_j^{\star}$ , and  $\mathbf{P}_i$  is a linear function thereof (up to  $O(\frac{1}{T^2})$  terms), it is straightforward to show that (3.13) is the CRLB for  $\text{vec}\{\mathbf{P}_i\}$ . Naturally, the pseudoinverse of this CRLB is the FIM for the entries of  $\mathbf{P}_i$ .

### 3.2.4 Mean Square Error

The final step in our analysis is to propagate expression (3.13) for the covariance matrix of the oblique projection matrices  $\widehat{\mathbf{P}}_i$  into an expression for the component estimation error. Let us define the estimation error of a given component  $i$  by a normalized MSE:

$$\widehat{\text{MSE}}_i \triangleq \frac{1}{\sigma_i^2} \frac{1}{T} \sum_{t=1}^T |\widehat{\mathbf{x}}_i(t) - \mathbf{x}_i(t)|^2, \quad (3.20)$$

where the normalization is by the average power  $\sigma_i^2$  of the  $i$ th component,

$$\sigma_i^2 \triangleq \frac{1}{T} \sum_{t=1}^T E \{ |\mathbf{x}_i(t)|^2 \} = \text{tr} \{ \langle \mathbf{R}_{X_i X_i} \rangle \}, \quad (3.21)$$

and where the last step uses  $|\mathbf{x}_i(t)|^2 = \text{tr} \{ \mathbf{x}_i(t) \mathbf{x}_i^\dagger(t) \}$  and (2.9). Using (3.3), (3.20) can now be rewritten as

$$\begin{aligned} \widehat{\text{MSE}}_i &= \frac{1}{\sigma_i^2} \frac{1}{T} \sum_{t=1}^T |\delta \mathbf{P}_i \mathbf{x}(t)|^2 = \frac{1}{\sigma_i^2} \frac{1}{T} \sum_{t=1}^T \text{tr} \{ \delta \mathbf{P}_i \mathbf{x}(t) \mathbf{x}^\dagger(t) (\delta \mathbf{P}_i)^\dagger \} \\ &= \frac{1}{\sigma_i^2} \text{tr} \{ (\langle \overline{\mathbf{R}}_X \rangle \otimes \mathbf{I}) \text{vec} \{ \delta \mathbf{P}_i \} \text{vec}^\dagger \{ \delta \mathbf{P}_i \} \}, \end{aligned} \quad (3.22)$$

where the second equality is analogous to that in (3.21), and the last equality uses (2.9) and then Property A.1 in Appendix A.

Now, in order to obtain the expectation of (3.22), we employ the approximation that  $\widehat{\mathbf{P}}_i$ , and thus  $\delta \mathbf{P}_i$ , are statistically independent of the *total* power of the observations,  $\langle \overline{\mathbf{R}}_X \rangle$ . This approximation becomes more accurate with larger  $Q$ . Hence, we can write

$$\text{MSE}_i \triangleq E \{ \widehat{\text{MSE}}_i \} = \frac{1}{\sigma_i^2} \text{tr} \{ (\langle \overline{\mathbf{R}}_X \rangle \otimes \mathbf{I}) \text{Cov}(\text{vec} \{ \delta \mathbf{P}_i \}) \}. \quad (3.23)$$

Substituting (3.13) in (3.23),

$$\begin{aligned} \text{MSE}_i &= \frac{1}{\sigma_i^2} \sum_{j \neq i} \left( \text{tr} \left\{ (\mathbf{P}_j^* \otimes \mathbf{I}) (\langle \overline{\mathbf{R}}_X \rangle \otimes \mathbf{I}) (\mathbf{P}_j^{\star\dagger} \otimes \mathbf{I}) \text{Cov}(\text{vec} \{ \boldsymbol{\mathcal{E}}_{ij} \}) \right\} \right. \\ &\quad - \text{tr} \left\{ (\mathbf{P}_i^* \otimes \mathbf{I}) (\langle \overline{\mathbf{R}}_X \rangle \otimes \mathbf{I}) (\mathbf{P}_j^{\star\dagger} \otimes \mathbf{I}) \text{Cov}(\text{vec} \{ \boldsymbol{\mathcal{E}}_{ij} \}, \text{vec} \{ \boldsymbol{\mathcal{E}}_{ji} \}) \right\} \\ &\quad - \text{tr} \left\{ (\mathbf{P}_j^* \otimes \mathbf{I}) (\langle \overline{\mathbf{R}}_X \rangle \otimes \mathbf{I}) (\mathbf{P}_i^{\star\dagger} \otimes \mathbf{I}) \text{Cov}(\text{vec} \{ \boldsymbol{\mathcal{E}}_{ji} \}, \text{vec} \{ \boldsymbol{\mathcal{E}}_{ij} \}) \right\} \\ &\quad \left. + \text{tr} \left\{ (\mathbf{P}_i^* \otimes \mathbf{I}) (\langle \overline{\mathbf{R}}_X \rangle \otimes \mathbf{I}) (\mathbf{P}_i^{\star\dagger} \otimes \mathbf{I}) \text{Cov}(\text{vec} \{ \boldsymbol{\mathcal{E}}_{ji} \}) \right\} \right) + O\left(\frac{1}{T^2}\right) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\sigma_i^2} \sum_{j \neq i} \left( \text{tr} \left\{ \left( \langle \mathbf{R}_{X_j X_j} \rangle \otimes \mathbf{I} \right) \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}) \right\} \right. \\
&\quad - \text{tr} \left\{ \left( \langle \mathbf{R}_{X_i X_j} \rangle \otimes \mathbf{I} \right) \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}, \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\}) \right\} \\
&\quad - \text{tr} \left\{ \left( \langle \mathbf{R}_{X_j X_i} \rangle \otimes \mathbf{I} \right) \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\}, \text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}) \right\} \\
&\quad \left. + \text{tr} \left\{ \left( \langle \mathbf{R}_{X_i X_i} \rangle \otimes \mathbf{I} \right) \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\}) \right\} \right) + O\left(\frac{1}{T^2}\right) \\
&= \frac{1}{\sigma_i^2} \sum_{j \neq i} \left( \text{tr} \left\{ \left( \langle \mathbf{R}_{X_j X_j} \rangle \otimes \mathbf{I} \right) \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}) \right\} \right. \\
&\quad \left. + \text{tr} \left\{ \left( \langle \mathbf{R}_{X_i X_i} \rangle \otimes \mathbf{I} \right) \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\}) \right\} \right) + O\left(\frac{1}{T^2}\right). \tag{3.24}
\end{aligned}$$

The first equality of (3.24) results from the invariance of the trace operator under cyclic permutations (property (A.1d) in Appendix A). The second equality follows from (2.5), which implies  $\mathbf{P}_i^* \mathbf{R}_X^{(q)} \mathbf{P}_j^* = \mathbf{R}_{X_i X_j}^{(q)} \forall i, j$ . The third equality is due to the fact that  $\mathbf{R}_{X_i X_j}^{(q)} = \mathbf{0}_{m \times m} \forall i \neq j$ . Substituting the explicit expressions for  $\text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\})$  and  $\text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\})$ , given by (3.18), in (3.24),

$$\begin{aligned}
\text{MSE}_i &= \frac{1}{\sigma_i^2} \frac{1}{T} \sum_{j \neq i} \left( \text{tr} \left\{ \left( \langle \mathbf{R}_{X_j X_j} \rangle \otimes \mathbf{I} \right) \left( \langle \mathbf{R}_{X_j X_j} \otimes \mathbf{R}_{X_i X_i}^\# \rangle - \langle \mathbf{R}_{X_j X_j}^\# \otimes \mathbf{R}_{X_i X_i} \rangle^\# \right)^\# \right\} \right. \\
&\quad \left. + \text{tr} \left\{ \left( \langle \mathbf{R}_{X_i X_i} \rangle \otimes \mathbf{I} \right) \left( \langle \mathbf{R}_{X_i X_i} \otimes \mathbf{R}_{X_j X_j}^\# \rangle - \langle \mathbf{R}_{X_i X_i}^\# \otimes \mathbf{R}_{X_j X_j} \rangle^\# \right)^\# \right\} \right) \\
&\quad + O\left(\frac{1}{T^2}\right). \tag{3.25}
\end{aligned}$$

We have thus obtained a closed-form expression for the MSE, which can be fully expressed by the model parameters  $\{\mathbf{R}_{X_i X_i}^{(q)}\}_{q=1}^Q$  and the weights  $\{n_q\}_{q=1}^Q$ , up to  $O(\frac{1}{T^2})$  terms. In the Gaussian case, this is also the minimal mean square error (MMSE).

It should be noted that all the derivations in this section 3.2 and in the related appendices do not rely neither on the Gaussian distribution nor on statistics of order larger than 2. Therefore, (3.18), (3.19), (3.13) and (3.25) hold also for non-Gaussian observations. That is, they still reflect the error covariance and MSE if we apply (2.17) for their separation; however, the CRLB, FIM and MMSE interpretation of the derived expressions no longer applies.

### 3.3 Numerical Examples

In this section, we support the performance analysis of Sec. 3.2 by numerical examples. Two algorithms which minimize (2.17), and thus solve (3.8), are suggested in Chapter 5 and Lahat et al. [60]. Both converge to the same separation point. For the following simulations we preferred the quasi-Newton (QN) realization, due to

its faster convergence rate, in terms of number of required iterations given a certain threshold.

### 3.3.1 Validating the Closed-Form Expression for the MSE

In the following simulation, we construct the data so that the analysis requirements of Sec. 3.2 hold, including the small-errors regime. Therefore, the theoretical expression for the MSE is expected to be an accurate prediction of the measured error. We set  $Q = 5$  adjacent domains with  $n_q = 500$  samples for a total of  $T = 2500$  samples. The choice of  $n_q$  will be further discussed in Sec. 3.3.2. In each scenario, the diagonal blocks of the matrices  $\mathbf{R}_S^{(q)}$  are drawn as  $\mathbf{R}_{S,ii}^{(q)} = \mathbf{U}^\dagger \mathbf{U}$  where  $\mathbf{U}$  is an  $m_i \times m_i$  upper triangular matrix with independent entries uniformly distributed on  $[-\frac{1}{2}, \frac{1}{2}]$ . The underlying sources are created, at each Monte-Carlo trial, by left-multiplying the Cholesky factorization of  $\mathbf{R}_S^{(q)}$  with  $m \times n_q$  statistically independent, zero mean, unit variance samples. These samples are drawn from various distributions, in order to support our claim that the second-order analysis holds not only for Gaussian sources.

As explained in Sec. 2.2, there is no scale indeterminacy to resolve. Since the joint block diagonalization (JBD) algorithms in Chapter 5 do not guarantee global convergence, the following steps were taken in order to avoid permutation errors. We chose to initialize the QN algorithm with  $\mathbf{I}$ . In this case, permutation errors are avoided by choosing mixing matrices  $\mathbf{A}$  which are strictly diagonally-dominant. In the following simulations, the  $(i, j)$ th entry of matrix  $\mathbf{A}$  is given by  $\mathbf{A}_{ij} = 3\delta_{ij} + \frac{1}{10}r_{ij}$ ,  $r_{ij} \sim \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$  and  $r_{ij}$  are independent and identically distributed (i.i.d.). Such values allow for sufficient variability of the mixing matrix to test our small-error analysis, while maintaining global convergence. Cases in which an algorithm does converge to an undesired local minimum are due to permutation errors. These are easily detected, since they result in a significantly larger MSE. Therefore, as a final safety measure, we verified that no such large errors appeared in our results.

Table 3.1 on page 35 compares the empirical MSE with the analytical MSE for several scenarios with varying component dimensions and distributions. The second column states the arbitrary index given to each component within each scenario. The third column denotes the dimension of the  $i$ th component in the scenario. In each scenario, different  $\mathbf{A}$  and  $\mathbf{R}_S^{(q)}$  are drawn.  $\mathbf{A}$  and  $\mathbf{R}_S^{(q)}$  are fixed for the same scenario. The fourth column gives the analytical MSE for each component (3.25), which is calculated using the correct model parameters. Each scenario is evaluated using 5000 Monte-Carlo trials. For each scenario, two data types were tested. In columns 5–8, Gaussian, zero mean, unit-variance samples are used to create the un-

derlying sources. In columns 10–13, either uniform, Laplacian or Gaussian mixture (peaks centered at  $\pm 4$ ) zero mean, unit-variance samples, denoted U, L and GM, respectively, are used to create the underlying sources. Note that left-multiplication of the non-Gaussian samples with the Cholesky factorization of  $\mathbf{R}_S^{(q)}$  changes their distribution; however, it is still non-Gaussian. The non-Gaussian distribution, used to generate the data for each scenario, is given in column 9.

The fifth and tenth columns give the averaged empirical MSE for each component (3.20). Columns 6 and 11 give the ratio of MSE for component separation: simulated vs. analytical.

Columns 7–8 and 12–13 compare the averaged empirical MSE using our JBD criterion with the averaged empirical MSE obtained from one-dimensional modeling (joint diagonalization (JD)) and then grouping the separated elements into the multidimensional components, according to the known block-pattern  $\mathbf{m}$ . These values are denoted in Table 3.1 as  $\widehat{\text{MSE}}^{\text{JD}}$ . We point out that the conditions derived in [28] for the global optimum to be achieved by properly grouping the independent component analysis (ICA) elements refer to a *different* separation criterion and are thus inapplicable here. An elaborate discussion of the use of JD instead of JBD for multidimensional data is given in Chapter 6.

The last row of Table 3.1 summarizes the results of each column. First, note that all the (normalized) MSE values are much smaller than 1, illustrating the quality of the component separation. Second, note that all values in column 6 and 11 are close to 1, showing that our analysis predicts correctly the achievable separation accuracy. The good match between predicted and empirical MSE demonstrates that indeed only second-order statistics are required for our theoretical analysis, and that for Gaussian data, the CRLB is indeed achievable. As predicted from the derivations of Chapter 6, there is a significant gain (columns 8 and 13) due to using the correct model, as proposed in Chapter 3. An important result is that in scenarios #2 and #3, which include one-dimensional components along with higher-dimensional ones, the gain for the one-dimensional components is  $> 1$ , too. Obviously, when the sources are independent, as in the last scenario (#5), there is no difference between JD and JBD, hence we put ‘1’ in this cell.

### 3.3.2 Validating the Small-Errors Regime

In this section, we justify our choice of  $n_q$  for the simulations in Sec. 3.3.1. That is, we show that small-error conditions indeed apply when  $n_q = 500$ . For this aim, we compare the empirical covariance of the entries of  $\delta \mathbf{P}_i$  with their predicted covariance, given in closed-form by (3.13), with (3.18) and (3.19). Fig. 3.1 depicts

Table 3.1: Performance of second-order multidimensional ICA: Analytical vs. empirical normalized MSE, averaged over 5000 Monte-Carlo trials. Empirical results: JBD algorithm with correct  $\mathbf{m}$  ( $\widehat{\text{MSE}}_i$ ) vs.  $\mathbf{m} = [1, \dots, 1]^\dagger$  ( $\widehat{\text{MSE}}_i^{\text{JD}}$ ). Each scenario, that is, different  $\mathbf{m}$ , is tested once with Gaussian and once with non-Gaussian data. The underlying sources are drawn at each Monte-Carlo trial. The parameters  $\mathbf{A}$  and  $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$  are fixed for each scenario. The last row of the table summarizes the columns.

Scenario	$i$	$m_i$	MSE $_i$	Gaussian data				Non-Gaussian data			
				$\widehat{\text{MSE}}_i$	$\frac{\widehat{\text{MSE}}_i}{\text{MSE}_i}$	$\widehat{\text{MSE}}_i^{\text{JD}}$	$\frac{\widehat{\text{MSE}}_i^{\text{JD}}}{\text{MSE}_i}$	$\widehat{\text{MSE}}_i$	$\frac{\widehat{\text{MSE}}_i}{\text{MSE}_i}$	$\widehat{\text{MSE}}_i^{\text{JD}}$	$\frac{\widehat{\text{MSE}}_i^{\text{JD}}}{\text{MSE}_i}$
				Gaussian data				Non-Gaussian data			
#1	1	2	$1.40 \cdot 10^{-3}$	$1.40 \cdot 10^{-3}$	0.999	$2.11 \cdot 10^{-3}$	1.507	$1.40 \cdot 10^{-3}$	0.999	$2.08 \cdot 10^{-3}$	1.484
	2	2	$8.31 \cdot 10^{-4}$	$8.40 \cdot 10^{-4}$	1.010	$1.18 \cdot 10^{-3}$	1.410	$8.22 \cdot 10^{-4}$	0.989	$1.16 \cdot 10^{-3}$	1.415
	3	2	$1.57 \cdot 10^{-3}$	$1.56 \cdot 10^{-3}$	0.993	$2.46 \cdot 10^{-3}$	1.578	$1.58 \cdot 10^{-3}$	1.007	$2.43 \cdot 10^{-3}$	1.539
#2	1	6	$1.58 \cdot 10^{-4}$	$1.56 \cdot 10^{-4}$	0.991	$4.84 \cdot 10^{-4}$	3.099	$1.60 \cdot 10^{-4}$	1.018	$5.01 \cdot 10^{-4}$	3.125
	2	1	$4.52 \cdot 10^{-3}$	$4.60 \cdot 10^{-3}$	1.018	$1.70 \cdot 10^{-2}$	3.687	$4.62 \cdot 10^{-3}$	1.022	$1.78 \cdot 10^{-2}$	3.852
	3	1	$2.54 \cdot 10^{-3}$	$2.50 \cdot 10^{-3}$	0.985	$5.91 \cdot 10^{-3}$	2.364	$2.58 \cdot 10^{-3}$	1.019	$6.06 \cdot 10^{-3}$	2.343
#3	1	3	$8.22 \cdot 10^{-4}$	$8.31 \cdot 10^{-4}$	1.010	$1.44 \cdot 10^{-3}$	1.737	$8.20 \cdot 10^{-4}$	0.998	$1.42 \cdot 10^{-3}$	1.733
	2	2	$8.87 \cdot 10^{-4}$	$9.07 \cdot 10^{-4}$	1.022	$1.36 \cdot 10^{-3}$	1.500	$8.83 \cdot 10^{-4}$	0.996	$1.35 \cdot 10^{-3}$	1.524
	3	1	$2.60 \cdot 10^{-3}$	$2.59 \cdot 10^{-3}$	0.996	$5.09 \cdot 10^{-3}$	1.961	$2.63 \cdot 10^{-3}$	1.009	$4.98 \cdot 10^{-3}$	1.895
#4	1	5	$4.40 \cdot 10^{-4}$	$4.44 \cdot 10^{-4}$	1.009	$2.49 \cdot 10^{-3}$	5.595	$4.40 \cdot 10^{-4}$	0.998	$2.46 \cdot 10^{-3}$	5.585
	2	4	$5.85 \cdot 10^{-4}$	$5.90 \cdot 10^{-4}$	1.008	$3.30 \cdot 10^{-3}$	5.598	$5.85 \cdot 10^{-4}$	0.999	$3.26 \cdot 10^{-3}$	5.581
#5	1	1	$9.67 \cdot 10^{-4}$	$9.72 \cdot 10^{-4}$	1.005	$9.72 \cdot 10^{-4}$	1	$9.58 \cdot 10^{-4}$	0.991	$9.58 \cdot 10^{-4}$	1
	2	1	$2.30 \cdot 10^{-3}$	$2.31 \cdot 10^{-3}$	1.004	$2.31 \cdot 10^{-3}$	1	$2.27 \cdot 10^{-3}$	0.990	$2.27 \cdot 10^{-3}$	1
	3	1	$1.25 \cdot 10^{-4}$	$1.25 \cdot 10^{-4}$	0.998	$1.25 \cdot 10^{-4}$	1	$1.26 \cdot 10^{-4}$	1.007	$1.26 \cdot 10^{-4}$	1
			$\ll 1$	$\ll 1$	$\cong 1$	$\ll 1$	$\geq 1$	$\ll 1$	$\cong 1$	$\ll 1$	$\geq 1$

results for a scenario with block-pattern  $\mathbf{m} = [6, 2, 1]^\dagger$ . The parameters are drawn as for Table 3.1. The underlying sources are drawn from the uniform distribution.

The entries of the three oblique projections  $\mathbf{P}_i$  are illustrated in Fig. 3.1(c)–3.1(e). These entries are scaled to their  $\log_{10}(|\cdot|)$ , in order to enhance the difference between the smaller and larger values.

Fig. 3.1(a) and Fig. 3.1(b) refer to data corresponding to the first row of  $\delta\mathbf{P}_1$  and the sixth row of  $\delta\mathbf{P}_2$ , respectively. In each row there are  $m = 9$  entries. Fig. 3.1(a) and Fig. 3.1(b) compare the empirical standard deviation of each entry in the row with the predicted one. The empirical values are averaged over 500 Monte-Carlo trials and denoted by ‘ $\times$ ’ markers. The three largest standard deviation values in Fig. 3.1(a) correspond to values in the last three columns of the first row of  $\mathbf{P}_1$ . The two largest standard deviation values in Fig. 3.1(b) correspond to values in columns 7,8 of the sixth row of  $\mathbf{P}_2$ .

In the small-errors regime, the standard deviation of the entries of  $\delta\mathbf{P}_i$  is proportional to  $1/\sqrt{T}$ , as follows from (3.13), with (3.18) and (3.19). These predicted values are depicted by the continuous straight lines in Fig. 3.1(a) and Fig. 3.1(b). It is visible that the empirical covariances converge to the  $1/\sqrt{T}$  slope. At  $n_q = 500$ , the difference between the predicted and empirical values is already sufficiently small. Similar trends have been found for the other entries of the oblique projection estimates in this scenario, as well as for the entries of the oblique projection estimates in the other scenarios of Table 3.1. Using too low  $n_q$  results in bias and discrepancy between experiment and theory. These results also validate our theoretical expression for  $\text{Cov}(\text{vec}\{\delta\mathbf{P}_i\})$ , (3.13).

## 3.4 Discussion

In this chapter, we obtained closed-form expressions for the FIM, CRLB and MMSE for the blind separation of statistically independent multidimensional Gaussian components, in a piecewise-stationary model. Our derivations and results were performed and presented in terms of the unambiguous and well-defined components, as well as the oblique projections on their subspaces. The oblique projections are the unambiguous counterparts of the mixing matrix in the component representation.

Our small-error analysis in this chapter was based on the first-order expansion of the contrast function and its related estimating equations, presented in Chapter 2. Error analysis of the contrast function, via its related estimating equations, provided us with a closed-form expression of the covariance of the oblique projections. We then propagated the covariance of the oblique projection estimates into the MSE. We

have thus obtained a closed-form expression for the MSE of the component estimates, in terms of the covariance matrices of the components. This MSE expression is valid, though no longer optimal, also for non-Gaussian data, when the other model assumptions hold.

Our derivations were validated by numerical simulations, in Sec. 3.3. We have shown that asymptotically, the empirical covariance of the oblique projections converges to the predicted value. We have shown good match between the predicted expected MSE, using our closed-form expression, and the empirical MSE, for various scenarios of multidimensional components. Since our contrast function is a JBD criterion, the performance of separation using JBD was compared with that of classical one-dimensional JD followed by clustering. Our treatment of multidimensional components was shown to yield a significant gain in their separation.

### 3.A First-Order Expansion of the Estimating Equations

In this appendix we show that the first-order expansion of the estimating equations (2.28) leads to the linear relation (3.7) between the projection error terms  $\boldsymbol{\varepsilon}_{ij} = \delta \mathbf{P}_i \boldsymbol{\Pi}_j^*$  and the sample error terms  $\overline{\mathbf{R}}_{X_i X_j}^{(q)}$ .

Let us begin by restating that the estimates  $\widehat{\mathbf{P}}_i$  are solutions of the estimating equations (2.28). That is, in the notation of Sec. 3.2,

$$\langle (\widehat{\mathbf{P}}_i \overline{\mathbf{R}}_X \widehat{\mathbf{P}}_i^\dagger)^\# (\widehat{\mathbf{P}}_i \overline{\mathbf{R}}_X \widehat{\mathbf{P}}_j^\dagger) \rangle = \mathbf{0}_{m \times m}. \quad (3.26)$$

In the following, we linearize these equations with respect to the error terms due to finite sample size,

$$\delta \mathbf{R}_X^{(q)} = \overline{\mathbf{R}}_X^{(q)} - \mathbf{R}_X^{(q)}. \quad (3.27)$$

Under asymptotic conditions, which are defined formally in Sec. 3.2,  $\overline{\mathbf{R}}_X^{(q)}$  converges, in the mean square, to  $\mathbf{R}_X^{(q)}$ , and the ML estimator  $\widehat{\mathbf{P}}_i$  converges, in probability, to  $\mathbf{P}_i^*$  (asymptotically, for non-Gaussian components, both converge in probability). As for the rate of convergence, the entries of both  $\delta \mathbf{R}_X^{(q)}$  and  $\delta \mathbf{P}_i$  are zero mean random variables with a standard deviation proportional to  $1/\sqrt{T}$ . Hence, asymptotically, terms which are proportional to  $\delta \mathbf{R}_X^{(q)}$  or  $\delta \mathbf{P}_i$  are  $\Omega(\frac{1}{\sqrt{T}})$  (this notation is defined in Chapter “Notations and Conventions” on page ix) and are considered as having the same order of magnitude.

Expanding the right-hand term within the  $\langle \cdot \rangle$  on the left-hand side (LHS)

of (3.26),

$$\begin{aligned}
\widehat{\mathbf{P}}_i \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{P}}_j^\dagger &= (\mathbf{P}_i^* + \delta \mathbf{P}_i)(\mathbf{R}_X^{(q)} + \delta \mathbf{R}_X^{(q)})(\mathbf{P}_j^{\dagger} + \delta \mathbf{P}_j^\dagger) \\
&= \overline{\mathbf{R}}_{X_i X_j}^{(q)} + \mathbf{P}_i^* \mathbf{R}_X^{(q)} \delta \mathbf{P}_j^\dagger + \delta \mathbf{P}_i \mathbf{R}_X^{(q)} \mathbf{P}_j^{\dagger} + \Omega(\frac{1}{T}) \\
&= \overline{\mathbf{R}}_{X_i X_j}^{(q)} + \mathbf{R}_{X_i X_i}^{(q)} \delta \mathbf{P}_j^\dagger + \delta \mathbf{P}_i \mathbf{R}_{X_j X_j}^{(q)} + \Omega(\frac{1}{T})
\end{aligned} \tag{3.28}$$

where in the last transition we used  $\mathbf{P}_i^* \mathbf{R}_X^{(q)} = \mathbf{R}_{X_i X_i}^{(q)}$ . Note that all the terms on the RHS of (3.28) are  $\Omega(\frac{1}{\sqrt{T}})$ , that is, first-order terms, since  $\mathbf{P}_i^* \mathbf{R}_X^{(q)} \mathbf{P}_j^{\dagger} = \mathbf{R}_{X_i X_j}^{(q)} = \mathbf{0}_{m \times m}$ .

As for the left-hand term within the  $\langle \cdot \rangle$  on the LHS of (3.26),

$$\begin{aligned}
(\widehat{\mathbf{P}}_i \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{P}}_i^\dagger)^\# &= (\widehat{\mathbf{A}}_i \widehat{\mathbf{B}}_i \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{B}}_i^\dagger \widehat{\mathbf{A}}_i^\dagger)^\# = \widehat{\mathbf{A}}_i^{\#\dagger} (\widehat{\mathbf{B}}_i \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{B}}_i^\dagger)^{-1} \widehat{\mathbf{A}}_i^\# \\
&= \mathbf{A}_i^{\#\dagger} (\mathbf{B}_i \mathbf{R}_X^{(q)} \mathbf{B}_i^\dagger)^{-1} \mathbf{A}_i^\# + \Omega(\frac{1}{\sqrt{T}}) = (\mathbf{P}_i^* \mathbf{R}_X^{(q)} \mathbf{P}_i^{\dagger})^\# + \Omega(\frac{1}{\sqrt{T}}).
\end{aligned} \tag{3.29}$$

The first equality in (3.29) is due to  $\widehat{\mathbf{P}}_i \triangleq \widehat{\mathbf{A}}_i \widehat{\mathbf{B}}_i$ , which follows from (2.6). The second equality follows from (2.26). The third transition is due to the fact that  $\overline{\mathbf{R}}_X^{(q)}$  and the ML estimates of  $\mathbf{A}$  and  $\mathbf{B}$  converge to their mean with a standard deviation proportional to  $1/\sqrt{T}$ . The last step follows again from (2.26) and then (2.6). Multiplying (3.29) with (3.28),

$$\begin{aligned}
(\widehat{\mathbf{P}}_i \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{P}}_i^\dagger)^\# (\widehat{\mathbf{P}}_i \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{P}}_j^\dagger) &= \mathbf{R}_{X_i X_i}^{\#\dagger(q)} (\overline{\mathbf{R}}_{X_i X_j}^{(q)} + \mathbf{R}_{X_i X_i}^{(q)} \delta \mathbf{P}_j^\dagger + \delta \mathbf{P}_i \mathbf{R}_{X_j X_j}^{(q)}) + \Omega(\frac{1}{T}) \\
&= \mathbf{R}_{X_i X_i}^{\#\dagger(q)} \overline{\mathbf{R}}_{X_i X_j}^{(q)} + \mathbf{\Pi}_i^* \delta \mathbf{P}_j^\dagger + \mathbf{R}_{X_i X_i}^{\#\dagger(q)} \delta \mathbf{P}_i \mathbf{R}_{X_j X_j}^{(q)} + \Omega(\frac{1}{T})
\end{aligned} \tag{3.30}$$

where in the first transition we used  $\mathbf{R}_{X_i X_i}^{\#\dagger(q)} = (\mathbf{P}_i^* \mathbf{R}_X^{(q)} \mathbf{P}_i^{\dagger})^\#$  and in the second one,

$$\mathbf{R}_{X_i X_i}^{\#\dagger(q)} \mathbf{R}_{X_i X_i}^{(q)} = \mathbf{\Pi}_i^*. \tag{3.31}$$

Averaging (3.30) over all domains, (3.26) is linearized as

$$\langle \mathbf{R}_{X_i X_i}^{\#\dagger} \overline{\mathbf{R}}_{X_i X_j} \rangle + \mathbf{\Pi}_i^* \delta \mathbf{P}_j^\dagger + \langle \mathbf{R}_{X_i X_i}^{\#\dagger} \delta \mathbf{P}_i \mathbf{R}_{X_j X_j} \rangle = \mathbf{0}_{m \times m} + \Omega(\frac{1}{T}). \tag{3.32}$$

Using the property  $\mathbf{R}_{X_i X_i}^{(q)} = \mathbf{\Pi}_i^* \mathbf{R}_{X_i X_i}^{(q)}$  and the notation  $\boldsymbol{\mathcal{E}}_{ij}$  of (3.4), (3.32) is rewritten as

$$\langle \mathbf{R}_{X_i X_i}^{\#\dagger} \overline{\mathbf{R}}_{X_i X_j} \rangle + \boldsymbol{\mathcal{E}}_{ji}^\dagger + \langle \mathbf{R}_{X_i X_i}^{\#\dagger} \boldsymbol{\mathcal{E}}_{ij} \mathbf{R}_{X_j X_j} \rangle = \mathbf{0}_{m \times m} + \Omega(\frac{1}{T}). \tag{3.33}$$

We have thus proved the first equation in the pair (3.7). The second equation is obtained by interchanging  $i$  and  $j$ .

### 3.B Closed-Form Expression for $\text{Cov}([\mathbf{g}_{ij}^\dagger \mathbf{g}_{ji}^\dagger]^\dagger)$

In this appendix we derive the expression  $\text{Cov}([\mathbf{g}_{ij}^\dagger \mathbf{g}_{ji}^\dagger]^\dagger) = \frac{1}{T} \mathbf{H}_\Pi$  (3.14) for the covariance matrix of the gradients  $\mathbf{g}_{ij}$  defined by (3.9). Since, by the assumptions in Sec. 2.3, these gradients have zero mean,

$$\text{Cov} \left( \begin{bmatrix} \mathbf{g}_{ij} \\ \mathbf{g}_{ji} \end{bmatrix} \right) = \begin{bmatrix} E \left\{ \mathbf{g}_{ij} \mathbf{g}_{ij}^\dagger \right\} & E \left\{ \mathbf{g}_{ij} \mathbf{g}_{ji}^\dagger \right\} \\ E \left\{ \mathbf{g}_{ji} \mathbf{g}_{ij}^\dagger \right\} & E \left\{ \mathbf{g}_{ji} \mathbf{g}_{ji}^\dagger \right\} \end{bmatrix}.$$

We shall now prove that

$$E \left\{ \mathbf{g}_{ij} \mathbf{g}_{ij}^\dagger \right\} = \frac{1}{T} \mathbf{H}_{ij} \quad (3.34)$$

$$E \left\{ \mathbf{g}_{ij} \mathbf{g}_{ji}^\dagger \right\} = \frac{1}{T} \mathcal{T}_{m,m}^{\Pi_{ij}} \quad (3.35)$$

for any  $i \neq j$ , which provide the desired result.

In a first step, we relate the covariance of the gradients to the covariance of the empirical matrices. For any  $i, j, k, l$ ,

$$E \left\{ \mathbf{g}_{ij} \mathbf{g}_{kl}^\dagger \right\} = \frac{1}{T^2} \sum_{\substack{q=1 \\ p=1}}^Q n_q n_p (\mathbf{I} \otimes \mathbf{R}_{X_i X_i}^{(q)\#}) E \left\{ \text{vec} \left\{ \overline{\mathbf{R}}_{X_i X_j}^{(q)} \right\} \text{vec}^\dagger \left\{ \overline{\mathbf{R}}_{X_k X_l}^{(p)} \right\} \right\} (\mathbf{I} \otimes \mathbf{R}_{X_k X_k}^{(p)\#}),$$

where we have used the alternate form of (3.9),

$$\mathbf{g}_{ij} = \langle \text{vec} \{ \mathbf{R}_{X_i X_i}^\# \overline{\mathbf{R}}_{X_i X_j} \mathbf{I} \} \rangle = \langle (\mathbf{I} \otimes \mathbf{R}_{X_i X_i}^{(q)\#}) \text{vec} \{ \overline{\mathbf{R}}_{X_i X_j} \} \rangle$$

and the last equality is due to (A.1c) in Appendix A. Since  $\mathbf{x}(t)$  is independent of  $\mathbf{x}(t')$  if  $t \neq t'$  (Sec. 2.3.1), the double sum merges into a single index  $q$ , leaving only

$$E \left\{ \mathbf{g}_{ij} \mathbf{g}_{kl}^\dagger \right\} = \frac{1}{T^2} \sum_{q=1}^Q n_q^2 (\mathbf{I} \otimes \mathbf{R}_{X_i X_i}^{(q)\#}) E \left\{ \text{vec} \{ \overline{\mathbf{R}}_{X_i X_j}^{(q)} \} \text{vec}^\dagger \{ \overline{\mathbf{R}}_{X_k X_l}^{(q)} \} \right\} (\mathbf{I} \otimes \mathbf{R}_{X_k X_k}^{(q)\#}). \quad (3.36)$$

In a second step, we work out the covariance matrix of the vectorized empirical matrices within (3.36) as follows. For  $i \neq j$ ,

$$\begin{aligned} E \left\{ \text{vec} \{ \overline{\mathbf{R}}_{X_i X_j}^{(q)} \} \text{vec}^\dagger \{ \overline{\mathbf{R}}_{X_i X_j}^{(q)} \} \right\} &= \frac{1}{n_q^2} \sum_{t \in \mathcal{D}_q} \sum_{r \in \mathcal{D}_q} E \left\{ \text{vec} \{ \mathbf{x}_i(t) \mathbf{x}_j^\dagger(t) \} \text{vec}^\dagger \{ \mathbf{x}_i(r) \mathbf{x}_j^\dagger(r) \} \right\} \\ &= \frac{1}{n_q^2} \sum_{t \in \mathcal{D}_q} \sum_{r \in \mathcal{D}_q} E \left\{ \mathbf{x}_j(t) \mathbf{x}_j^\dagger(r) \otimes \mathbf{x}_i(t) \mathbf{x}_i^\dagger(r) \right\} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{n_q^2} \sum_{t \in \mathcal{D}_q} \sum_{r \in \mathcal{D}_q} E \left\{ \mathbf{x}_j(t) \mathbf{x}_j^\dagger(r) \right\} \otimes E \left\{ \mathbf{x}_i(t) \mathbf{x}_i^\dagger(r) \right\} \\
&= \frac{1}{n_q^2} \sum_{t \in \mathcal{D}_q} \sum_{r \in \mathcal{D}_q} (\mathbf{R}_{X_j X_j}^{(q)} \otimes \mathbf{R}_{X_i X_i}^{(q)}) \delta_{rt} \\
&= \frac{1}{n_q} \mathbf{R}_{X_j X_j}^{(q)} \otimes \mathbf{R}_{X_i X_i}^{(q)}. \tag{3.37}
\end{aligned}$$

The first equality is an expansion of the definition of  $\overline{\mathbf{R}}_{X_i X_j}^{(q)}$ . The second equality uses Property A.2 in Appendix A. The third equality is by independence of components  $i$  and  $j$ . The fourth equality is based on the assumption of Sec. 2.3.1 that  $\forall t, r \in \mathcal{D}_q$ ,  $E \left\{ \mathbf{x}_i(t) \mathbf{x}_i^\dagger(r) \right\} = \mathbf{R}_{X_i X_i}^{(q)} \delta_{rt}$ .

Substituting (3.37) into (3.36) with  $k = i$ ,  $l = j$ ,

$$\begin{aligned}
E \left\{ \mathbf{g}_{ij} \mathbf{g}_{ij}^\dagger \right\} &= \frac{1}{T^2} \sum_{q=1}^Q \frac{n_q^2}{n_q} (\mathbf{I} \otimes \mathbf{R}_{X_i X_i}^{(q)\#}) (\mathbf{R}_{X_j X_j}^{(q)} \otimes \mathbf{R}_{X_i X_i}^{(q)}) (\mathbf{I} \otimes \mathbf{R}_{X_i X_i}^{(q)\#}) \\
&= \frac{1}{T^2} \sum_{q=1}^Q n_q (\mathbf{R}_{X_j X_j}^{(q)} \otimes \mathbf{R}_{X_i X_i}^{(q)\#}) = \frac{1}{T} \mathbf{H}_{ij}, \tag{3.38}
\end{aligned}$$

which establishes (3.34). In (3.38), the second equality is due to property (A.1a) and then property (A.2b) in Appendix A; the last step follows from (3.11). In order to establish (3.35), we right-multiply (3.37) by the commutation matrix  $\mathcal{T}_{m,m}$ , defined in Sec. 3.2.2, and use the equality  $\text{vec}^\dagger \{ \overline{\mathbf{R}}_{X_i X_j}^{(q)} \} \mathcal{T}_{m,m} = \text{vec}^\dagger \{ \overline{\mathbf{R}}_{X_j X_i}^{(q)} \}$ , which follows from (A.3). This turns (3.37) into

$$E \left\{ \text{vec} \{ \overline{\mathbf{R}}_{X_i X_j}^{(q)} \} \text{vec}^\dagger \{ \overline{\mathbf{R}}_{X_j X_i}^{(q)} \} \right\} = \frac{1}{n_q} (\mathbf{R}_{X_j X_j}^{(q)} \otimes \mathbf{R}_{X_i X_i}^{(q)}) \mathcal{T}_{m,m}. \tag{3.39}$$

Substituting (3.39) in (3.36) with  $k = j$ ,  $l = i$ ,

$$\begin{aligned}
E \left\{ \mathbf{g}_{ij} \mathbf{g}_{ji}^\dagger \right\} &= \frac{1}{T^2} \sum_{q=1}^Q \frac{n_q^2}{n_q} (\mathbf{I} \otimes \mathbf{R}_{X_i X_i}^{(q)\#}) (\mathbf{R}_{X_j X_j}^{(q)} \otimes \mathbf{R}_{X_i X_i}^{(q)}) \mathcal{T}_{m,m} (\mathbf{I} \otimes \mathbf{R}_{X_j X_j}^{(q)\#}) \\
&= \frac{1}{T} \langle \mathbf{R}_{X_j X_j} \mathbf{R}_{X_j X_j}^\# \otimes \mathbf{R}_{X_i X_i}^\# \mathbf{R}_{X_i X_i} \rangle \mathcal{T}_{m,m} \\
&= \frac{1}{T} \langle \mathbf{\Pi}_j^* \otimes \mathbf{\Pi}_i^* \rangle \mathcal{T}_{m,m} = \frac{1}{T} (\mathbf{\Pi}_j^* \otimes \mathbf{\Pi}_i^*) \mathcal{T}_{m,m}, \tag{3.40}
\end{aligned}$$

where in the second equality we applied property (A.4b), followed by property (A.1a)

in Appendix A and (2.15). In the third equality of (3.40) we applied (3.31). Finally,

$$\begin{aligned} (\mathbf{\Pi}_j^* \otimes \mathbf{\Pi}_i^*) \mathcal{T}_{m,m} &= (\mathbf{\Pi}_j^* \otimes \mathbf{\Pi}_i^*)^2 \mathcal{T}_{m,m} \\ &= (\mathbf{\Pi}_j^* \otimes \mathbf{\Pi}_i^*) \mathcal{T}_{m,m} (\mathbf{\Pi}_i^* \otimes \mathbf{\Pi}_j^*) = \mathcal{T}_{m,m}^{\Pi_{ij}}, \end{aligned} \quad (3.41)$$

where the second equality is due to (A.4b) in Appendix A, and the last step is by (3.16). Substituting (3.41) in (3.40) concludes the proof of (3.35).

### 3.C Proof of $\mathcal{H}^{-1}\mathcal{H}_{\Pi}\mathcal{H}^{-\dagger} = \mathcal{H}_{\Pi}^{\#}$

In this appendix we prove the second equality in (3.17). The desired identity is

$$\mathcal{H}^{-1}\mathcal{H}_{\Pi}\mathcal{H}^{-\dagger} = \mathcal{H}_{\Pi}^{\#}. \quad (3.42)$$

*Proof.* Denote  $\mathbf{M} = \mathcal{H}_{\Pi}$ ,  $\mathbf{N} = \mathcal{H} - \mathcal{H}_{\Pi}$  and  $\mathbf{\Pi} = \begin{bmatrix} \mathbf{\Pi}_j \otimes \mathbf{\Pi}_i & \mathbf{0}_{m^2 \times m^2} \\ \mathbf{0}_{m^2 \times m^2} & \mathbf{\Pi}_i \otimes \mathbf{\Pi}_j \end{bmatrix}$ . With these notations and the invertibility of  $\mathcal{H}$  (necessary and sufficient conditions are given in Sec. 4.4), the orthogonality conditions  $\mathbf{\Pi M \Pi} = \mathbf{M}$  and  $(\mathbf{I} - \mathbf{\Pi})\mathbf{N}(\mathbf{I} - \mathbf{\Pi}) = \mathbf{N}$ , which are required to fulfill Property 3.1 below, hold, as can be readily verified. The equality in (3.42) thus results from Property 3.1.  $\square$

**Property 3.1.** Let  $\mathbf{M}$  and  $\mathbf{N}$  be two symmetric matrices and  $\mathbf{M} + \mathbf{N}$  invertible. Assume that  $\mathbf{M}$  and  $\mathbf{N}$  have orthogonal range spaces. Then

$$(\mathbf{M} + \mathbf{N})^{-1} \mathbf{M} (\mathbf{M} + \mathbf{N})^{-\dagger} = \mathbf{M}^{\#}. \quad (3.43)$$

*Proof.* The outline of the proof is as follows. Multiply both sides of (3.43) with  $(\mathbf{M} + \mathbf{N})$ . Applying the four criteria of the Moore-Penrose pseudoinverse ((A.2) in Appendix A) to the resulting four summands on the RHS, three zero out using the orthogonal range spaces of  $\mathbf{M}$  and  $\mathbf{N}$  and the fact that  $\mathbf{M M}^{\#}$  and  $\mathbf{M}^{\#} \mathbf{M}$  are orthogonal projection operators. The only nonzero term is due to (A.2a) in Appendix A.  $\square$

### 3.D Explicit Form for $\text{Cov}(\text{vec}\{\mathcal{E}_{ij}\})$ and $\text{Cov}(\text{vec}\{\mathcal{E}_{ij}\}, \text{vec}\{\mathcal{E}_{ji}\})$

In this appendix we obtain an explicit expression for the blocks of matrix  $\mathcal{H}_{\Pi}^{\#}$ , where  $\mathcal{H}_{\Pi}$  is defined in (3.15). This calculation is the final step in the derivation of

$\text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\})$  and  $\text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}, \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\})$ , given in (3.18) and (3.19).

In order to factorize  $\boldsymbol{\mathcal{H}}_{\Pi}$  and  $\boldsymbol{\mathcal{H}}_{\Pi}^{\sharp}$  into more basic terms, we introduce the decomposition  $\boldsymbol{R}_{X_i X_i}^{(q)} = \boldsymbol{U}_i \boldsymbol{\Upsilon}_i^{(q)} \boldsymbol{U}_i^{\dagger}$  where, for convenience, the  $m_i \times m_i$  invertible matrices  $\boldsymbol{\Upsilon}_i^{(q)}$  are chosen so that the  $m \times m_i$  matrices  $\boldsymbol{U}_i$  have orthonormal columns.  $\boldsymbol{U}_i$  are thus the orthonormalized  $\boldsymbol{A}_i$  of (2.4). It then holds that  $\boldsymbol{U}_i^{\dagger} \boldsymbol{U}_i = \boldsymbol{I}$ ,

$$\boldsymbol{U}_i^{\sharp} = \boldsymbol{U}_i^{\dagger}, \quad (3.44)$$

and  $\boldsymbol{R}_{X_i X_i}^{\sharp(q)} = \boldsymbol{U}_i \boldsymbol{\Upsilon}_i^{-\sharp(q)} \boldsymbol{U}_i^{\dagger}$ , due to (2.26). With the notations  $\boldsymbol{\mathcal{I}} \triangleq \begin{bmatrix} \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\mathcal{T}}_{m_i, m_j} \end{bmatrix}$ ,

$$\boldsymbol{u} \triangleq \begin{bmatrix} \boldsymbol{U}_j \otimes \boldsymbol{U}_i & \mathbf{0} \\ \mathbf{0} & \boldsymbol{U}_i \otimes \boldsymbol{U}_j \end{bmatrix} \quad \text{and} \quad \boldsymbol{\mathcal{H}}_{\Upsilon} \triangleq \begin{bmatrix} \langle \boldsymbol{\Upsilon}_j \otimes \boldsymbol{\Upsilon}_i^{-1} \rangle & \boldsymbol{I} \\ \boldsymbol{I} & \langle \boldsymbol{\Upsilon}_j^{-1} \otimes \boldsymbol{\Upsilon}_i \rangle \end{bmatrix}$$

we can rewrite

$$\boldsymbol{\mathcal{H}}_{\Pi} = \boldsymbol{u} \boldsymbol{\mathcal{I}} \boldsymbol{\mathcal{H}}_{\Upsilon} \boldsymbol{\mathcal{I}}^{\dagger} \boldsymbol{u}^{\dagger}. \quad (3.45)$$

Then, if  $\boldsymbol{\mathcal{H}}_{\Upsilon}$  is indeed invertible (necessary and sufficient conditions are given in Theorem 4.1),

$$\boldsymbol{\mathcal{H}}_{\Pi}^{\sharp} = \boldsymbol{u} \boldsymbol{\mathcal{I}}^{-1} \boldsymbol{\mathcal{H}}_{\Upsilon}^{-1} \boldsymbol{\mathcal{I}}^{-\dagger} \boldsymbol{u}^{\dagger}, \quad (3.46)$$

which follows from (2.26), (3.44), and the fact that  $\boldsymbol{\mathcal{I}}$  is always a full-rank matrix. Since  $\boldsymbol{u}$  and  $\boldsymbol{\mathcal{I}}$  are block-diagonal matrices, then in order to calculate the blocks of  $\boldsymbol{\mathcal{H}}_{\Pi}^{\sharp}$ , all that remains is to invert  $\boldsymbol{\mathcal{H}}_{\Upsilon}$ .

For the upper-left  $m_i m_j \times m_i m_j$  block of  $\boldsymbol{\mathcal{H}}_{\Upsilon}^{-1}$ , a block-matrix inversion formula [80] yields

$$[\boldsymbol{\mathcal{H}}_{\Upsilon}^{-1}]^{\text{UL}} = (\langle \boldsymbol{\Upsilon}_j \otimes \boldsymbol{\Upsilon}_i^{-1} \rangle - \langle \boldsymbol{\Upsilon}_j^{-1} \otimes \boldsymbol{\Upsilon}_i \rangle^{-1})^{-1}. \quad (3.47)$$

Multiplying both sides by the upper-left blocks of  $\boldsymbol{u}$  and  $\boldsymbol{\mathcal{I}}^{-1}$ ,

$$\begin{aligned} [\boldsymbol{\mathcal{H}}_{\Pi}^{\sharp}]^{\text{UL}} &= (\boldsymbol{U}_j \otimes \boldsymbol{U}_i) (\langle \boldsymbol{\Upsilon}_j \otimes \boldsymbol{\Upsilon}_i^{-1} \rangle - \langle \boldsymbol{\Upsilon}_j^{-1} \otimes \boldsymbol{\Upsilon}_i \rangle^{-1})^{-1} (\boldsymbol{U}_j \otimes \boldsymbol{U}_i)^{\dagger} \\ &= \left( \langle \boldsymbol{R}_{X_j X_j} \otimes \boldsymbol{R}_{X_i X_i}^{\sharp} \rangle - \langle \boldsymbol{R}_{X_j X_j}^{\sharp} \otimes \boldsymbol{R}_{X_i X_i} \rangle^{\sharp} \right)^{\sharp} \end{aligned} \quad (3.48)$$

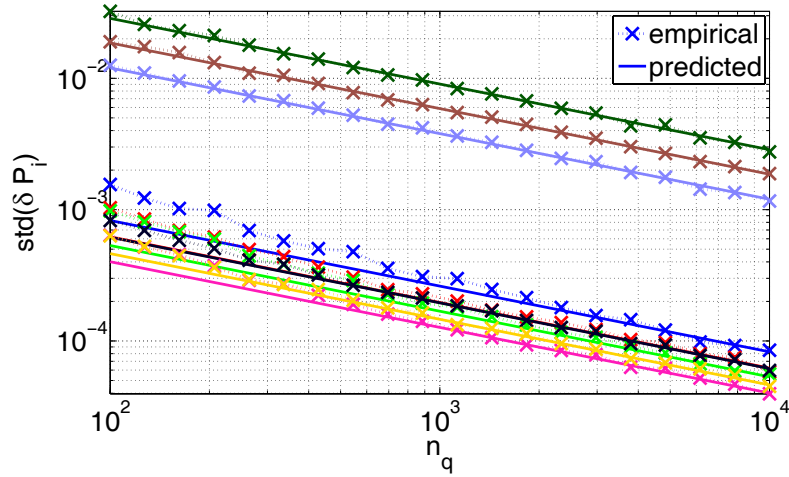
which follows from (2.26), (3.44), and (3.46). Similar arguments yield the expression

$$[\mathcal{H}_{\Pi}^{\#}]^{\text{UR}} = - (\mathbf{U}_j \otimes \mathbf{U}_i) \left( \langle \boldsymbol{\Upsilon}_j \otimes \boldsymbol{\Upsilon}_i^{-1} \rangle - \langle \boldsymbol{\Upsilon}_j^{-1} \otimes \boldsymbol{\Upsilon}_i \rangle^{-1} \right)^{-1} \cdot \langle \boldsymbol{\Upsilon}_j^{-1} \otimes \boldsymbol{\Upsilon}_i \rangle^{-1} \mathcal{T}_{m_i, m_j} (\mathbf{U}_i \otimes \mathbf{U}_j)^{\dagger} \quad (3.49)$$

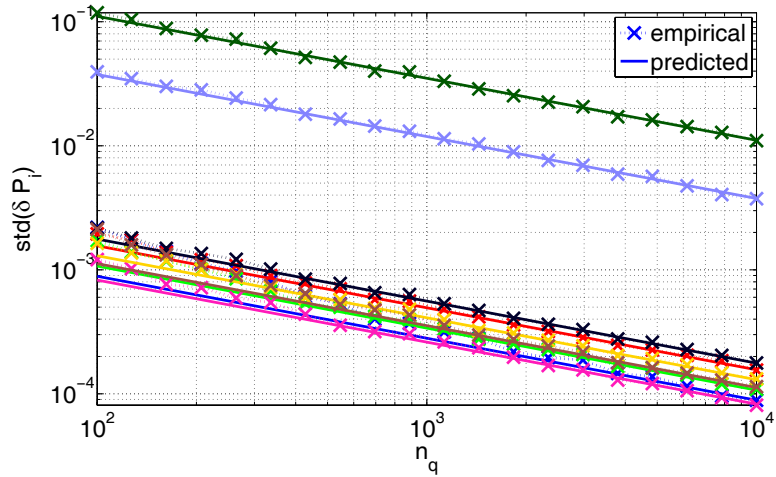
for the upper-right  $m_i m_j \times m_i m_j$  block of  $\mathcal{H}_{\Pi}^{\#}$ . Inserting the neutral term  $(\mathbf{U}_j^{\dagger} \otimes \mathbf{U}_i^{\dagger})(\mathbf{U}_j \otimes \mathbf{U}_i) = \mathbf{I}$  at the ‘ $\cdot$ ’ symbol in (3.49) and then applying algebraic operations similarly to (3.48),

$$[\mathcal{H}_{\Pi}^{\#}]^{\text{UR}} = - \left( \langle \mathbf{R}_{X_j X_j} \otimes \mathbf{R}_{X_i X_i}^{\#} \rangle - \langle \mathbf{R}_{X_j X_j}^{\#} \otimes \mathbf{R}_{X_i X_i} \rangle^{\#} \right)^{\#} \langle \mathbf{R}_{X_j X_j}^{\#} \otimes \mathbf{R}_{X_i X_i} \rangle^{\#} \mathcal{T}_{m, m}. \quad (3.50)$$

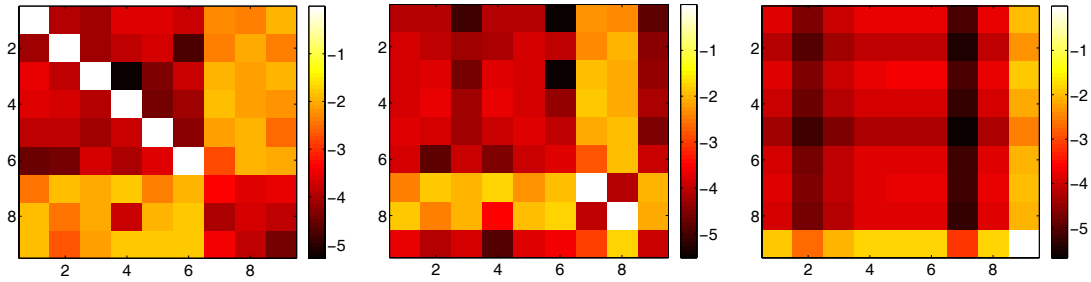
By exchanging  $i$  with  $j$  in (3.48) and (3.50), one obtains the two lower blocks of  $\mathcal{H}_{\Pi}^{\#}$ . Substituting these results in (3.17) yields (3.18) and (3.19).



(a) Standard deviation of the  $m = 9$  entries on the first row of  $\delta P_1$ : empirical and predicted, vs.  $n_q$ .



(b) Standard deviation of the  $m = 9$  entries on the sixth row of  $\delta P_2$ : empirical and predicted, vs.  $n_q$ .



(c)  $\log_{10}$  of the entries of  $P_1$ . (d)  $\log_{10}$  of the entries of  $P_2$ . (e)  $\log_{10}$  of the entries of  $P_3$ .

Figure 3.1: Validating the theoretical expression for  $\text{Cov}(\delta P_i)$ , as well as our choice of  $n_q$  for the small-errors regime. In Fig. 3.1(a) and 3.1(b), the ‘ $\times$ ’ symbol denotes the empirical values, averaged over 500 Monte-Carlo trials. The straight lines denote the predicted values, under the small-error assumption. Each colour refers to one entry of one row of  $P_1$  (Fig. 3.1(a)) and  $P_2$  (Fig. 3.1(b)). The data is drawn from the uniform distribution. Block-pattern  $m=[6,2,1]$ . The three oblique projections  $P_i$  are depicted in Fig. 3.1(c)–3.1(e).

# Chapter 4

## Identifiability of Second-Order Multidimensional ICA

### 4.1 Outline

This chapter deals with the identifiability of the data model which was presented in Chapter 2 and analyzed in Chapter 3. The analysis in Chapter 3 was carried out under the assumption that the model was identifiable. We now discuss the conditions under which this assumption holds.

As a first step in our analysis, a necessary condition on the number  $Q$  of matrices is derived in Sec. 4.2. In Sec. 4.3, we analyze the Fisher information matrix (FIM). We show that the identifiability of the model parameters relies on the positive-definiteness of a certain matrix. We derive a theorem, which states sufficient and necessary conditions for this matrix to be positive-definite. This theorem leads us to sufficient and necessary conditions for the identifiability of the model, in terms of the latent unnormalized, non-whitened, source covariance matrices. In Sec. 4.4, the relationship between the FIM and the invertibility of the matrix  $\mathcal{H}$ , defined in Sec. 3.2.2, is determined. The invertibility of this matrix  $\mathcal{H}$  is the prerequisite for the existence of the FIM. We prove that the conditions for the invertibility of  $\mathcal{H}$  are the same as those for the identifiability of the model. An example which illustrates the meaning of non-identifiability is given in Sec. 4.5. This example leads us to sufficient and necessary conditions on the uniqueness of a solution to joint block diagonalization (JBD) of a set of mixtures of positive-definite JBD-irreducible symmetric block-diagonal matrices. A short discussion of our results follows, in Sec. 4.6.

The core of the analysis in this chapter has been published in Lahat et al. [51].

## 4.2 Degrees of Freedom

Let us compare the number of degrees of freedom in the model with the number of constraints in the data. Since we focus on second-order methods, the data is represented only by their sample covariance matrices. These are  $Q$  symmetric  $m \times m$  matrices so that our model should try to fit

$$N_{\text{data}} = Q \frac{m(m+1)}{2}$$

scalar numbers. The model is adjusted by varying the mixing matrix and the source covariance matrices. However, there is some redundancy between these matrices, because of the factorizations discussed in Sec. 2.2: each submatrix  $\mathbf{A}_i$  has  $m \times m_i$  degrees of freedom, where  $m_i^2$  of them (which correspond to the entries of the arbitrary invertible scaling matrix  $\mathbf{Z}_i$ ) can be factored into the corresponding source covariance matrices. This leaves  $m^2 - \sum_{i=1}^n m_i^2 \geq 0$  effective degrees of freedom in the mixing matrix  $\mathbf{A}$ , and  $m_i(m_i + 1)/2$  degrees of freedom in each  $\mathbf{R}_{S,ii}^{(q)}$ . Hence, the model has

$$N_{\text{model}} = m^2 - \sum_{i=1}^n m_i^2 + Q \sum_{i=1}^n \frac{m_i(m_i + 1)}{2}$$

effective free scalar parameters. It turns out that

$$N_{\text{data}} - N_{\text{model}} = \left(\frac{Q}{2} - 1\right) \left(m^2 - \sum_{i=1}^n m_i^2\right).$$

Hence, as soon as  $Q \geq 2$ , we have  $N_{\text{data}} \geq N_{\text{model}}$ , that is, there are more (or, at least, as many) scalar statistics as free parameters in the model.

However, since *any* two positive-definite matrices can be *exactly* jointly diagonalized [81, Theorem 6], the JBD-irreducibility requirement of Sec. 2.3.1 will be violated if we let  $Q = 2$  for multidimensional data. The latter assumption implies that  $Q > 2$  is a necessary condition for identifiability in the presence of multidimensional components. Otherwise,  $Q = 2$  will suffice.

## 4.3 Derivation of the Identifiability Theorem

In Sec. 3.2.3 we derived the FIM (3.14) for the pair  $(\text{vec}\{\mathbf{P}_i \mathbf{\Pi}_j^*\}, \text{vec}\{\mathbf{P}_j \mathbf{\Pi}_i^*\})$ :

$$T^2 \text{Cov} \left( \begin{bmatrix} \mathbf{g}_{ij} \\ \mathbf{g}_{ji} \end{bmatrix} \right) = T^2 \frac{1}{T} \mathbf{H}_{\Pi} = T \mathbf{H}_{\Pi},$$

where the term  $T^2$  is due to multiplication by the number of samples  $T$  in (2.16). The FIM  $T\mathcal{H}_\Pi$  must reflect the number of non-degenerate parameters. This number is equal to the number of its non-zero eigenvalues, which is equal to its rank. If a solution to (3.8):

$$\begin{bmatrix} \mathbf{g}_{ij} \\ \mathbf{g}_{ji} \end{bmatrix} = -\mathcal{H} \cdot \begin{bmatrix} \text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\} \\ \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\} \end{bmatrix} + \Omega(\frac{1}{T})$$

exists, that is, the model is identifiable, then  $\mathcal{H}_\Pi$  is a covariance matrix of rank  $2m_i m_j$ . This rank is due to the fact that  $\mathbf{g}_{ij}$  and  $\boldsymbol{\mathcal{E}}_{ij}$  each reflect a relation between an  $m_i$ -dimensional element and an  $m_j$ -dimensional one. The term “2” is due to the concatenation of two such vectors,  $\mathbf{g}_{ij}$  and  $\mathbf{g}_{ji}$ . Therefore, we are interested in finding the condition for  $\mathcal{H}_\Pi$  to be positive-semidefinite symmetric with rank  $2m_i m_j$ .

In order to further simplify our task, we now break down  $\mathcal{H}_\Pi$  into more basic components. Using the algebraic properties (A.1), (A.2) and (A.4) in Appendix A,  $\mathcal{H}_\Pi$  can be factorized as

$$\mathcal{H}_\Pi = \mathcal{I} \mathcal{A} \mathcal{R} \mathcal{A}^\dagger \mathcal{I}, \quad (4.1)$$

where  $\mathcal{I} = \begin{bmatrix} \mathbf{I}_{m^2 \times m^2} & \mathbf{0}_{m^2 \times m^2} \\ \mathbf{0}_{m^2 \times m^2} & \mathcal{T}_{m,m} \end{bmatrix}$  is always invertible,  $\mathcal{A} = \begin{bmatrix} \mathbf{A}_j \otimes \mathbf{A}_i^\# & \mathbf{0}_{m^2 \times m_i m_j} \\ \mathbf{0}_{m^2 \times m_i m_j} & \mathbf{A}_j^\# \otimes \mathbf{A}_i \end{bmatrix}$  has rank  $2m_i m_j$  (we assume  $\mathbf{A}$  invertible), and

$$\mathcal{R} = \begin{bmatrix} \langle \mathbf{R}_{S,jj} \otimes \mathbf{R}_{S,ii}^{-1} \rangle & \mathbf{I}_{m_i m_j \times m_i m_j} \\ \mathbf{I}_{m_i m_j \times m_i m_j} & \langle \mathbf{R}_{S,jj}^{-1} \otimes \mathbf{R}_{S,ii} \rangle \end{bmatrix}$$

is a  $2m_i m_j \times 2m_i m_j$  symmetric matrix. If  $\mathcal{R}$  is invertible, that is,  $\mathcal{R}$  has rank  $2m_i m_j$ , then

$$\mathcal{H}_\Pi^\# = \mathcal{I} \mathcal{A}^\# \mathcal{R}^{-1} \mathcal{A}^\# \mathcal{I}^\dagger. \quad (4.2)$$

Notation (4.2) can be verified by checking the four properties of the pseudoinverse (A.2) in Appendix A. Recall from Sec. 3.2.3 and (3.17) that  $\mathcal{H}_\Pi^\#$  reflects the Cramér-Rao lower bound (CRLB) which corresponds to this FIM. Therefore, identifiability of the corresponding  $2m_i m_j$  parameters corresponds to  $\frac{1}{T} \mathcal{H}_\Pi^\#$  being well-defined, i.e. that  $\mathcal{R}^{-1}$  be finite. We have thus reduced the conditions on  $\mathcal{H}_\Pi$  or  $\mathcal{H}_\Pi^\#$  to the requirement that  $\mathcal{R}$  be positive-definite and symmetric. The conditions for  $\mathcal{R}$  to be strictly positive-definite and symmetric are given by the following theorem,

whose proof is given in Appendix 4.A.

**Theorem 4.1.** *Let  $\{\mathbf{R}_{S,ii}^{(q)}\}_{q=1}^Q$  and  $\{\mathbf{R}_{S,jj}^{(q)}\}_{q=1}^Q$ ,  $Q > 2$  be two sequences of JBD-irreducible positive-definite symmetric matrices of size  $m_i \times m_i$  and  $m_j \times m_j$ , respectively. Then  $\begin{bmatrix} \langle \mathbf{R}_{S,jj} \otimes \mathbf{R}_{S,ii}^{-1} \rangle & \mathbf{I} \\ \mathbf{I} & \langle \mathbf{R}_{S,jj}^{-1} \otimes \mathbf{R}_{S,ii} \rangle \end{bmatrix}$  is positive-definite if*

1.  $m_i \neq m_j$ , or
2.  $m_i = m_j$  and there does not exist an  $m_i \times m_i$  invertible matrix  $\mathbf{M}$  such that  $\forall q$ ,

$$\mathbf{R}_{S,jj}^{(q)} = \mathbf{M} \mathbf{R}_{S,ii}^{(q)} \mathbf{M}^\dagger. \quad (4.3)$$

Theorem 4.1 is applicable to our context for any normalization of the “true” latent source covariance matrices, regardless of the arbitrary block-diagonal scale ambiguity between  $\mathbf{A}$  and  $\mathbf{s}(t)$ . This is due to the fact that the invertible matrix  $\mathbf{M}$  in (4.3) is otherwise arbitrary. We have thus obtained all the necessary and sufficient conditions for  $\mathcal{R}$  to be positive-definite and symmetric, and consequently for the FIM  $T\mathcal{H}_\Pi$  to reflect identifiability for all its related parameters. As explained at the end of Sec. 3.2.3, the identifiability of the oblique projections  $\mathbf{P}_i$  follows from the identifiability of the pairs  $(\text{vec}\{\mathbf{P}_i \mathbf{\Pi}_j^*\}, \text{vec}\{\mathbf{P}_j \mathbf{\Pi}_i^*\})$ . Theorem 4.1 is defined for one pair  $(i, j)$ , and it should be applied pairwise to all components. Consequently, Theorem 4.1 provides the sufficient and necessary conditions for the identifiability of all the model parameters.

## 4.4 Invertibility of $\mathcal{H}$

In Sec. 3.2.2 we have shown that the pair of linearized estimating equations (3.7) could be rewritten in matrix form as (3.8):  $[\mathbf{g}_{ij}^\dagger \ \mathbf{g}_{ji}^\dagger]^\dagger = -\mathcal{H}[\text{vec}^\dagger\{\mathcal{E}_{ij}\} \ \text{vec}^\dagger\{\mathcal{E}_{ji}\}] + \Omega(\frac{1}{T})$ . Based on the assumption that  $\mathcal{H}$  was invertible, (3.8) was rewritten as (3.12):  $[\text{vec}^\dagger\{\mathcal{E}_{ij}\} \ \text{vec}^\dagger\{\mathcal{E}_{ji}\}] = -\mathcal{H}^{-1}[\mathbf{g}_{ij}^\dagger \ \mathbf{g}_{ji}^\dagger] + \Omega(\frac{1}{T})$ . Recall from (3.9) that  $\mathbf{g}_{ij}$  reflects the empirical correlation between components. Therefore, (3.12) shows how the empirical correlation between components, that is, the fact that  $\overline{\mathbf{R}}_{X_i X_j}^{(q)}$  is non-zero in finite sample size, results in non-zero errors  $\mathcal{E}_{ij}$ . Later in Sec. 3.2.3, we obtained the covariance matrix (3.17) of the vectorized error terms  $\mathcal{E}_{ij}$ :

$$\text{Cov} \left( \begin{bmatrix} \text{vec}\{\mathcal{E}_{ij}\} \\ \text{vec}\{\mathcal{E}_{ji}\} \end{bmatrix} \right) = \frac{1}{T} \mathcal{H}^{-1} \mathcal{H}_\Pi \mathcal{H}^{-\dagger} + O(\frac{1}{T^2}) = \frac{1}{T} \mathcal{H}_\Pi^\# + O(\frac{1}{T^2}).$$

That is, if  $\mathcal{H}$  is invertible, then  $\mathcal{H}_{\Pi}^{\sharp}$  is a covariance matrix of rank  $2m_i m_j$ . We now discuss the conditions for the invertibility of  $\mathcal{H}$ . We show that they are tightly related to the identifiability conditions of Sec. 4.3 and to Theorem 4.1.

The core of our proof is the claim that the inverse of  $\mathcal{H}$ , when exists, is given by

$$\mathcal{H}^{-1} = \mathcal{H}_{\Pi}^{\sharp} + \mathcal{H} - \mathcal{H}_{\Pi}. \quad (4.4)$$

Therefore, the problem of finding the condition for the invertibility of  $\mathcal{H}$  can be rephrased as finding the condition for  $\mathcal{H}_{\Pi}$  and  $\mathcal{H}_{\Pi}^{\sharp}$  to be positive-semidefinite symmetric with rank  $2m_i m_j$ . These conditions were given in Sec. 4.3 by Theorem 4.1.

It remains now to prove (4.4). The proof is based on showing that the term which we denote “ $\mathcal{H}^{-1}$ ” indeed obeys  $\mathcal{H}^{-1}\mathcal{H} = \mathbf{I} = \mathcal{H}\mathcal{H}^{-1}$ . Given (4.4),

$$\mathcal{H}^{-1}\mathcal{H} = (\mathcal{H}_{\Pi}^{\sharp} + \mathcal{J})(\mathcal{H}_{\Pi} + \mathcal{J}) = \mathcal{H}_{\Pi}^{\sharp}\mathcal{H}_{\Pi} + \mathcal{H}_{\Pi}^{\sharp}\mathcal{J} + \mathcal{J}\mathcal{H}_{\Pi} + \mathcal{J}\mathcal{J} \quad (4.5)$$

$$\mathcal{H}\mathcal{H}^{-1} = (\mathcal{H}_{\Pi} + \mathcal{J})(\mathcal{H}_{\Pi}^{\sharp} + \mathcal{J}) = \mathcal{H}_{\Pi}\mathcal{H}_{\Pi}^{\sharp} + \mathcal{J}\mathcal{H}_{\Pi}^{\sharp} + \mathcal{H}_{\Pi}\mathcal{J} + \mathcal{J}\mathcal{J}, \quad (4.6)$$

where  $\mathcal{J} \triangleq \mathcal{H} - \mathcal{H}_{\Pi}$ . From (3.10), (3.15) and (3.16),

$$\mathcal{J} = \begin{bmatrix} \mathbf{0}_{m^2 \times m^2} & \mathcal{T}_{m,m}(\mathbf{I} - \Pi_i \otimes \Pi_j) \\ \mathcal{T}_{m,m}(\mathbf{I} - \Pi_j \otimes \Pi_i) & \mathbf{0}_{m^2 \times m^2} \end{bmatrix}.$$

For the fourth term in (4.5) and (4.6),

$$\mathcal{J}\mathcal{J} = \begin{bmatrix} \mathbf{I} - \Pi_j \otimes \Pi_i & \mathbf{0}_{m^2 \times m^2} \\ \mathbf{0}_{m^2 \times m^2} & \mathbf{I} - \Pi_i \otimes \Pi_j \end{bmatrix} \quad (4.7)$$

due to (A.4a) in Appendix A. For the first term on the right-hand side (RHS) of (4.5) and (4.6), one can use the explicit forms (4.1) and (4.2) to obtain

$$\mathcal{H}_{\Pi}^{\sharp}\mathcal{H}_{\Pi} = \begin{bmatrix} \Pi_j \otimes \Pi_i & \mathbf{0}_{m^2 \times m^2} \\ \mathbf{0}_{m^2 \times m^2} & \Pi_i \otimes \Pi_j \end{bmatrix}, \quad (4.8)$$

and  $\mathcal{H}_{\Pi}^{\sharp}\mathcal{H}_{\Pi} = \mathcal{H}_{\Pi}\mathcal{H}_{\Pi}^{\sharp}$  due to the symmetry of  $\mathcal{H}_{\Pi}$ . For the second and third terms on the RHS of (4.5) and (4.6), combining (4.1) or (4.2) with the symmetry of  $\mathcal{J}$  yields

$$\mathcal{J}\mathcal{H}_{\Pi} = \mathcal{H}_{\Pi}\mathcal{J} = \mathcal{J}\mathcal{H}_{\Pi}^{\sharp} = \mathcal{H}_{\Pi}^{\sharp}\mathcal{J} = \mathbf{0}_{2m^2 \times 2m^2}. \quad (4.9)$$

Substituting (4.8), (4.9) and (4.7) in (4.5) and (4.6) yields  $\mathcal{H}\mathcal{H}^{-1} = \mathbf{I} = \mathcal{H}^{-1}\mathcal{H}$ . This confirms the desired identity (4.4) and thus the invertibility of  $\mathcal{H}$ . Note that

the derivation of (4.8) and (4.9) has relied on the fact that the decomposition (4.2) exists; that is, on the assumption that  $\mathcal{R}$  is invertible. Hence, the invertibility of  $\mathcal{H}$  follows from Theorem 4.1.

## 4.5 Example: Non-Identifiability

When the conditions in Theorem 4.1 hold, that is,  $\mathcal{H}$  is invertible and the FIM has the desired rank, the model is identifiable. In this case, one can obtain component separation, as demonstrated in the numerical examples of Sec. 3.3, Sec. 6.4, Sec. 5.4 and Sec. 7.5.

As a counter example, let us consider the case that Theorem 4.1 does not hold. Namely, that  $\mathbf{R}_{S,jj}^{(q)} = \mathbf{M}\mathbf{R}_{S,ii}^{(q)}\mathbf{M}^\dagger \forall q$  with  $\mathbf{M}$  an arbitrary invertible  $m_i \times m_i$  matrix. Assume that we have  $\mathbf{s}(t) = \begin{bmatrix} \mathbf{s}_1(t) \\ \mathbf{s}_2(t) \end{bmatrix}$  and  $\mathbf{A} = \mathbf{I}$ , such that  $\mathbf{x}(t) = \mathbf{s}(t)$ . As explained in Sec. 2.2, the components are correctly separated if  $\mathbf{A}$  is right-multiplied by an arbitrary invertible  $\mathbf{Z} \in \text{bdiag}_m$ . Since

$$\text{bdiag}_m\{\mathbf{Z}\mathbf{M}\mathbf{Z}^\dagger\} = \mathbf{Z} \text{bdiag}_m\{\mathbf{M}\}\mathbf{Z}^\dagger \quad \forall \mathbf{M} \in \mathbb{R}^{m \times m}, \quad (4.10)$$

then, from (2.17),

$$C(\mathbf{A}) = C(\mathbf{A}\mathbf{Z}). \quad (4.11)$$

That is, the output of the minimization of the contrast function (2.17), which is a JBD criterion, can be determined only up to block-diagonal scale ambiguity. Consider now the following invertible matrix,

$$\mathbf{B} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I} & -\mathbf{M}^{-1} \\ \mathbf{I} & \mathbf{M}^{-1} \end{bmatrix}.$$

Applying this matrix to the observation covariance matrices  $\mathbf{R}_X^{(q)} = \mathbf{R}_S^{(q)}$  yields

$$\mathbf{B}\mathbf{R}_X^{(q)}\mathbf{B}^\dagger = \begin{bmatrix} \mathbf{R}_{S,11}^{(q)} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{S,11}^{(q)} \end{bmatrix} \quad \forall q.$$

That is, matrix  $\mathbf{B}$  jointly block-diagonalizes  $\{\mathbf{R}_X^{(q)}\}_{q=1}^Q$ . Hence,  $\mathbf{B}$  is a legitimate output of a JBD algorithm applied to  $\{\mathbf{R}_X^{(q)}\}_{q=1}^Q$ . However, if we look at its application to the observations, we obtain

$$\mathbf{B}\mathbf{x}(t) = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I} & -\mathbf{M}^{-1} \\ \mathbf{I} & \mathbf{M}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{s}_1(t) \\ \mathbf{s}_2(t) \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{s}_1(t) - \mathbf{M}\mathbf{s}_2(t) \\ \mathbf{s}_2(t) + \mathbf{M}\mathbf{s}_2(t) \end{bmatrix} \neq \begin{bmatrix} \mathbf{Z}_1\mathbf{s}_1(t) \\ \mathbf{Z}_2\mathbf{s}_2(t) \end{bmatrix}. \quad (4.12)$$

Obviously,  $\mathbf{B}$  does not separate the sources. In other words, one can obtain exact JBD of the set of covariance matrices without actually statistically separating the underlying data. In this case, the errors terms  $\boldsymbol{\varepsilon}_{ij}$  and their covariances may be arbitrarily large, regardless of how close  $\overline{\mathbf{R}}_{X_i X_j}^{(q)}$  are to their (zero) mean. It is also clear that  $\mathbf{B}$  is not a block-diagonal matrix. Hence, it is not of the equivalence class of the original  $\mathbf{A}$ .

From the point of view of signal separation, this example demonstrates the inability to separate components when Theorem 4.1 is violated. From the point of view of JBD, since the contrast function (2.17) is also a JBD criterion, then Theorem 4.1 is also the sufficient and necessary condition for the uniqueness of JBD of a set of mixtures of JBD-irreducible block-diagonal positive-definite symmetric matrices. In this case, *uniqueness* is in the sense of left-multiplication of the block-diagonalizing matrix by an arbitrary invertible block-diagonal matrix with block-pattern  $\mathbf{m}$ .

## 4.6 Discussion

In this chapter, we obtained necessary and sufficient conditions for the identifiability of the multidimensional ICA (MICA) model of Sec. 2.3. These are also the sufficient and necessary conditions for JBD of a set of mixtures of real positive-definite JBD-irreducible block-diagonal symmetric matrices to be unique (up to a certain block-diagonal scaling matrix). Our analysis was based on the observation that the rank of the FIM should reflect identifiability of all the non-degenerate parameters. This amounts to the requirement that the consistency of errors between the model parameters and their corresponding estimates (i.e., solutions of (2.17)) depend on the invertibility of the 2-by-2 symmetric block matrix  $\mathcal{H}$ . Strictly speaking, since these conditions are based on small-error analysis, they imply *local* identifiability.

The proof of Lemma 4.1 is based on JBD-irreducibility. This lemma is the source for the requirement for JBD-irreducibility throughout this work.

The derived Theorem 4.1 is similar to that given in [49] and to the claim in [12, Sec. 2.2], without the whitening and orthonormality constraints. Theorem 4.1 is analogous to the notion of “simplicity”, as proposed by [49]. As argued in [49] (and recently also in [77, Sec. 4.2]), all the statistically independent sources for which (4.3) holds with equality should be gathered into one ‘simple’ factor, and the decomposition of  $\mathbf{s}$  into such ‘simple’ factors is unique, up to an arbitrary invertible mixture within each such factor. Unlike [49], we obtain Theorem 4.1 from the FIM. That is, from conditions on the identifiability of the model, and without any orthogonality prerequisite on the mixing or de-mixing matrix. We also provide a

full proof of the Theorem.

It should be noted that Theorem 4.1 reduces, in the one-dimensional case, to the condition that the variances of the different sources, as a function of the domain index, should not be proportional [41].

It is interesting to compare the analysis and the results of this chapter to those obtained by Vía et al. [50] and Anderson et al. [39]. Although their model is different, and consists of several datasets of one-dimensional sources, each mixed by a different mixing matrix, their separation scheme is also based on second-order statistics. Therefore, their identifiability conditions are based on constraints on the covariance matrices of the sources, as is the case in this thesis. In particular, [50, Definition 1] and [39, Lemma 1] have a similar form as our Theorem 4.1. However, in Theorem 4.1 the invertible matrices  $\mathbf{M}$  are arbitrary, whereas in [50] and [39] they must be diagonal. The route taken in [50] is based on analyzing eigenvalues, and is closely related to the proof of Schur's first lemma, which is used in [46, Lemma A.4]. The route taken in [39] is based on characterizing the non-singularity of the FIM, which is the same as the basic idea of our analysis. Of course, the FIM, as well as the Hessian function used in [39], are different than those of our MICA model.

## 4.A Proof of the Identifiability Theorem

In this appendix we prove Theorem 4.1.

Let  $\Xi_i^{(q)}$  and  $\Xi_j^{(q)}$ ,  $q = 1, \dots, Q$  denote positive-definite JBD-irreducible symmetric matrices of dimensions  $m_i \times m_i$  and  $m_j \times m_j$ , respectively, and

$$\mathcal{H}_{\Xi} = \begin{bmatrix} \langle \Xi_j \otimes \Xi_i^{-1} \rangle & \mathbf{I} \\ \mathbf{I} & \langle \Xi_j^{-1} \otimes \Xi_i \rangle \end{bmatrix} \quad (4.13)$$

a matrix with non-negative eigenvalues. We look for sufficient and necessary conditions on the sequences  $\{\Xi_i^{(q)}\}_{q=1}^Q$  and  $\{\Xi_j^{(q)}\}_{q=1}^Q$  for the strict positivity:  $\mathcal{H}_{\Xi} > 0$ .

$\mathcal{H}_{\Xi}$  is not positive-definite if and only if  $\mathbf{x}^\dagger \mathcal{H}_{\Xi} \mathbf{x} = 0$  for all non-zero vectors  $\mathbf{x}$  with real entries. Since  $n_q > 0$ , the condition  $\mathbf{x}^\dagger \mathcal{H}_{\Xi} \mathbf{x} = 0$  is equivalent to  $\mathbf{x}^\dagger \mathcal{H}_{\Xi}^{(q)} \mathbf{x} = 0 \forall q$ , where each

$$\mathcal{H}_{\Xi}^{(q)} \triangleq \begin{bmatrix} \Xi_j^{(q)} \otimes \Xi_i^{-1} & \mathbf{I} \\ \mathbf{I} & \Xi_j^{-1} \otimes \Xi_i^{(q)} \end{bmatrix}$$

is *always* singular. Without loss of generality, we can look for  $\mathbf{x}$  in the form

$$\mathbf{x} = \begin{bmatrix} \text{vec}\{\mathbf{M}\} \\ -\text{vec}\{\mathbf{N}\} \end{bmatrix},$$

where  $\mathbf{M}$  and  $\mathbf{N}$  are  $m_i \times m_j$  matrices. With the factorization

$$\mathcal{H}_{\Xi}^{(q)} = \begin{bmatrix} \Xi_j^{\frac{1}{2}(q)} \otimes \Xi_i^{-\frac{1}{2}(q)} \\ \Xi_j^{-\frac{1}{2}(q)} \otimes \Xi_i^{\frac{1}{2}(q)} \end{bmatrix} \begin{bmatrix} \Xi_j^{\frac{1}{2}(q)} \otimes \Xi_i^{-\frac{1}{2}(q)} \\ \Xi_j^{-\frac{1}{2}(q)} \otimes \Xi_i^{\frac{1}{2}(q)} \end{bmatrix}^\dagger$$

we have

$$\mathbf{x}^\dagger \mathcal{H}_{\Xi}^{(q)} \mathbf{x} = \left| \begin{bmatrix} \Xi_j^{\frac{1}{2}(q)} \otimes \Xi_i^{-\frac{1}{2}(q)} \\ \Xi_j^{-\frac{1}{2}(q)} \otimes \Xi_i^{\frac{1}{2}(q)} \end{bmatrix}^\dagger \begin{bmatrix} \text{vec}\{\mathbf{M}\} \\ -\text{vec}\{\mathbf{N}\} \end{bmatrix} \right|^2$$

and therefore, the condition for positive-definiteness becomes

$$\begin{bmatrix} \Xi_j^{\frac{1}{2}(q)} \otimes \Xi_i^{-\frac{1}{2}(q)} & \Xi_j^{-\frac{1}{2}(q)} \otimes \Xi_i^{\frac{1}{2}(q)} \end{bmatrix} \begin{bmatrix} \text{vec}\{\mathbf{M}\} \\ -\text{vec}\{\mathbf{N}\} \end{bmatrix} = 0 \quad \forall q.$$

Using equation (A.1c) in Appendix A, the latter can be rewritten as

$$\Xi_i^{-\frac{1}{2}(q)} \mathbf{M} \Xi_j^{\frac{1}{2}(q)} = \Xi_i^{\frac{1}{2}(q)} \mathbf{N} \Xi_j^{-\frac{1}{2}(q)} \quad \forall q, \quad (4.14)$$

which is equivalent to

$$\mathbf{M} \Xi_j^{(q)} = \Xi_i^{(q)} \mathbf{N} \quad \forall q. \quad (4.15)$$

Equation (4.15) can be simplified into  $\mathbf{M}' \mathbf{T}_j^{(q)} = \mathbf{T}_i^{(q)} \mathbf{M}' \forall q$ , where

$$\mathbf{T}_i^{(q)} = \langle \Xi_i \rangle^{-\frac{1}{2}} \Xi_i^{(q)} \langle \Xi_i \rangle^{-\frac{1}{2}}, \quad (4.16)$$

and  $\mathbf{M}' = \langle \Xi_i \rangle^{-\frac{1}{2}} \mathbf{M} \langle \Xi_j \rangle^{\frac{1}{2}}$  (analogously for index  $j$ ) so that  $\langle \mathbf{T}_i \rangle = \mathbf{I}$ .

We now introduce a Lemma. This lemma and its proof can be found in Murota et al. [46, Lemma A.4].

**Lemma 4.1.** *Let  $\{\mathbf{R}^{(q)}\}_{q=1}^Q$  and  $\{\mathbf{P}^{(q)}\}_{q=1}^Q$  be two sequences of  $Q > 2$  JBD-irreducible positive-definite symmetric matrices of size  $p \times p$  and  $r \times r$ , respectively. Then, the equation*

$$\mathbf{M} \mathbf{R}^{(q)} = \mathbf{P}^{(q)} \mathbf{M} \quad \forall q \quad (4.17)$$

for a  $p \times r$  matrix  $\mathbf{M}$  has a non-zero solution only if  $r = p$  and that solution must be proportional (i.e., up to a scalar factor) to a  $p \times p$  orthonormal matrix.

By identifying  $\mathbf{P}^{(q)}$  with  $\mathbf{T}_i^{(q)}$  and  $\mathbf{R}^{(q)}$  with  $\mathbf{T}_j^{(q)}$ , and applying Lemma 4.1, it

turns out that for  $m_i \neq m_j$ ,  $\mathcal{H}_\Xi$  is always positive-definite and symmetric. For  $m_i = m_j$ ,  $\mathcal{H}_\Xi$  is *not* positive-definite if and only if there exists an orthonormal matrix  $\mathbf{O}$  such that

$$\mathbf{O}\mathbf{T}_j^{(q)} = \mathbf{T}_i^{(q)}\mathbf{O} \quad \forall q. \quad (4.18)$$

Equation (4.18) defines an orthogonal equivalence relation between  $\mathbf{T}_i^{(q)}$  and  $\mathbf{T}_j^{(q)}$ . By rewriting (4.18) explicitly we obtain

$$\mathbf{O}\langle\Xi_j\rangle^{-\frac{1}{2}}\Xi_j^{(q)}\langle\Xi_j\rangle^{-\frac{1}{2}} = \langle\Xi_i\rangle^{-\frac{1}{2}}\Xi_i^{(q)}\langle\Xi_i\rangle^{-\frac{1}{2}}\mathbf{O} \quad \forall q,$$

where  $\langle\Xi_j\rangle^{-\frac{1}{2}}\Xi_j^{(q)}\langle\Xi_j\rangle^{-\frac{1}{2}}$  is the whitened version of  $\Xi_j^{(q)}$ , and similarly for  $\Xi_i^{(q)}$ . By changing sides, the latter equation can be rewritten as

$$\Xi_j^{(q)} = \langle\Xi_j\rangle^{\frac{1}{2}}\mathbf{O}^\dagger\langle\Xi_i\rangle^{-\frac{1}{2}}\Xi_i^{(q)}\langle\Xi_i\rangle^{-\frac{1}{2}}\mathbf{O}\langle\Xi_j\rangle^{\frac{1}{2}} \quad \forall q.$$

This is equivalent to normalizing  $\Xi_i^{(q)}$  by  $\langle\Xi_j\rangle^{\frac{1}{2}}\mathbf{O}^\dagger\langle\Xi_i\rangle^{-\frac{1}{2}}$ . However, since  $\langle\Xi_j\rangle$  and  $\langle\Xi_i\rangle$  are in fact arbitrary symmetric positive-definite matrices, then  $\langle\Xi_j\rangle^{\frac{1}{2}}\mathbf{O}^\dagger\langle\Xi_i\rangle^{-\frac{1}{2}}$  can be any invertible matrix. We summarize the results of this section in Theorem 4.1, which also replaces  $\Xi_i^{(q)}$  and  $\Xi_j^{(q)}$  with  $\mathbf{R}_{S,ii}^{(q)}$  and  $\mathbf{R}_{S,jj}^{(q)}$ , respectively.

# Chapter 5

## Joint Block Diagonalization Algorithms

### 5.1 Outline

In this chapter, we present two algorithms which (approximately) jointly block-diagonalize a set of weighted real positive-definite matrices. As demonstrated by [82], relative-variation algorithms enjoy equivariant performance and are thus preferred for our problem over their non-relative counterparts.

In Sec. 5.2, we derive the relative gradient (RG) and its first-order variation for the update step of our algorithms. A detailed description of the proposed RG and quasi-Newton (QN) algorithms is given in Sec. 5.3. Numerical simulations which demonstrate the convergence properties of the algorithms can be found in Sec. 5.4.

The algorithms in this chapter have been presented in Lahat et al. [60].

### 5.2 Derivation of the Relative Variations

As explained in Sec. 2.3.3, the likelihood of the observations can be maximized by joint block diagonalization (JBD) of a set of covariance matrices of the observations. This maximization amounts to minimization of the contrast function  $C(\mathbf{A})$  (2.17). The first-order variation of  $C(\mathbf{A})$  when  $\mathbf{A}$  is replaced by  $\mathbf{A}(\mathbf{I} + \mathbf{E})$  (where  $\mathbf{I}$  denotes the identity matrix and  $\mathbf{E}$  is an arbitrary matrix of the same size) can always be expressed by the Taylor expansion

$$C(\mathbf{A}(\mathbf{I} + \mathbf{E})) = C(\mathbf{A}) + \text{tr} \{ (\nabla C(\mathbf{A}))^\dagger \mathbf{E} \} + \text{higher-order terms in } \mathbf{E}, \quad (5.1)$$

for some  $m \times m$  matrix  $\nabla C(\mathbf{A})$ , defined as the RG of  $C(\mathbf{A})$  with respect to  $\mathbf{A}$ . Similarly to the derivation for the one-dimensional case in [42], one obtains the RG

$$\nabla C(\mathbf{A}) = -\langle \text{bdiag}_m^{-1} \{ \mathbf{A}^{-1} \overline{\mathbf{R}}_X^{(q)} \mathbf{A}^{-\dagger} \} \mathbf{A}^{-1} \overline{\mathbf{R}}_X^{(q)} \mathbf{A}^{-\dagger} \rangle + \mathbf{I} \quad (5.2)$$

on which the RG algorithm, explained in Sec. 5.3, is based. The equation in (5.2) was previously mentioned in an implicit form in (2.21). As explained in Sec. 2.3.4, solving (5.2) yields the stationary points of the contrast function (2.17).

Another algorithm can be derived based on the Newton method. In order to realize a QN method in the sense of [83], we obtain a first-order approximation of the gradient, using the following steps. First, we note that the maximum likelihood (ML) estimate of  $\mathbf{A}$  can be obtained by setting the RG (5.2) to zero. This yields the estimating equations (2.20) (this is true due to the relation (2.21)), or in their block-wise form (2.22).

In the second step, the first-order expansion of the estimating equations (2.22) under asymptotic conditions ( $T \rightarrow \infty$  for  $\frac{n_q}{T}$  fixed  $\forall q$ ) can be expressed (see Appendix 5.A), as

$$\begin{bmatrix} \text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\} \\ \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\} \end{bmatrix} = \boldsymbol{\mathcal{H}}^{-1} \begin{bmatrix} \mathbf{g}_{ij} \\ \mathbf{g}_{ji} \end{bmatrix} + \Omega(\frac{1}{T}) \quad i \neq j, \quad (5.3)$$

where

$$\boldsymbol{\mathcal{H}} = \begin{bmatrix} \mathbf{H}_{ij} & \boldsymbol{\mathcal{T}}_{m_j, m_i} \\ \boldsymbol{\mathcal{T}}_{m_i, m_j} & \mathbf{H}_{ji} \end{bmatrix}, \quad (5.4a)$$

$$\mathbf{H}_{ij}^{(q)} = \mathbf{R}_{S,jj}^{(q)} \otimes \mathbf{R}_{S,ii}^{-(q)}, \quad \mathbf{H}_{ij} = \langle \mathbf{R}_{S,jj} \otimes \mathbf{R}_{S,ii}^{-1} \rangle \quad (5.4b)$$

$$\mathbf{g}_{ij}^{(q)} = -\mathbf{R}_{S,ii}^{-(q)} \overline{\mathbf{R}}_{S,ij}^{(q)}, \quad \mathbf{g}_{ij} = -\langle \mathbf{R}_{S,ii}^{-1} \overline{\mathbf{R}}_{S,ij} \rangle, \quad (5.4c)$$

$$\overline{\mathbf{R}}_{S,ij}^{(q)} = \frac{1}{n_q} \sum_{t \in \mathcal{D}_q} \mathbf{s}_i(t) \mathbf{s}_j^\dagger(t), \quad (5.5)$$

$\mathbf{R}_{S,ii}^{-(q)} \triangleq (\mathbf{R}_{S,ii}^{(q)})^{-1}$  and  $\boldsymbol{\mathcal{E}}_{ij}$ , the  $m_i \times m_j$  blocks of an  $m \times m$  matrix  $\boldsymbol{\mathcal{E}}$ , defined formally in (5.7), reflect the relative change in  $\mathbf{A}$  due to the difference between  $\mathbf{R}_S^{(q)}$  and  $\overline{\mathbf{R}}_S^{(q)}$ .  $\boldsymbol{\mathcal{T}}_{m_i, m_j}$  is the  $m_i m_j \times m_i m_j$  commutation matrix, defined in (A.3) in Appendix A. It should be emphasized that since  $\text{bdiag}_m\{\nabla C(\mathbf{A})\}$  is invariant to changes in  $\overline{\mathbf{R}}_X^{(q)}$  (this is a direct result of (2.20)), then  $\boldsymbol{\mathcal{E}}_{ii} \equiv \mathbf{0}_{m_i \times m_i}$ . It is assumed that  $\boldsymbol{\mathcal{H}}$  is invertible; sufficient and necessary conditions are given by Theorem 4.1. The set of matrices  $\boldsymbol{\mathcal{E}}_{ij}$ ,  $\forall i \neq j$ , obtained from (5.3), constitute the Newton step.

This leads to our QN algorithm, explained in Sec. 5.3.

## 5.3 Algorithms

The pseudocode of the iterative algorithms is given in Algorithm 1. The part pertaining to each of the RG and QN algorithms is given in Algorithm 2 and 3, respectively. The RG algorithm works as follows: according to (5.1), if  $\mathbf{A}$  is changed into  $\mathbf{A}(\mathbf{I} + \mathbf{E})$ , then  $C(\mathbf{A})$  changes by the amount  $\text{tr} \{ (\nabla C(\mathbf{A}))^\dagger \mathbf{E} \} +$  higher-order terms in  $\mathbf{E}$ . Given  $\mathbf{E} = -\lambda \nabla C(\mathbf{A})$  and  $\lambda > 0$  a real scalar, the updating rule (line 4 in Algorithm 2) changes  $C(\mathbf{A})$  into  $C(\mathbf{A}) - \lambda \|\nabla C(\mathbf{A})\|^2 +$  higher-order terms in  $\nabla C(\mathbf{A})$ . Hence, the decrease of the contrast function  $C(\mathbf{A})$  is guaranteed for small enough  $\lambda$ . The updating rule is iterated until  $\|\nabla C(\mathbf{A})\| \leq \text{threshold}$ . In the QN algorithm, the relative change in  $\mathbf{A}$  is determined directly by  $\mathcal{E}$ , as explained in Sec. 5.2. The transformation matrix  $\mathbf{T}$  in the algorithms' pseudocode reflects the relative change in  $\mathbf{A}$  at each iteration.

The choice of the step-size in a RG algorithm determines its convergence rate, in terms of the number of required iterations; see [54], for example. For the simulations in Sec. 5.4 we chose to set  $\lambda$  by backtracking line search. Since only  $\overline{\mathbf{R}}_X^{(q)}$  is available to the algorithm, then within the iterations,  $\mathbf{A}^{-1} \mathbf{R}^{(q)} \mathbf{A}^\dagger$  is used to approximate both  $\overline{\mathbf{R}}_S^{(q)}$  and  $\mathbf{R}_S^{(q)}$  of (5.4c) and (5.4b). Therefore, within Algorithm 3,  $\mathbf{g}_{ij}$  is equal to the evaluated  $(i, j)$ th sub-block of  $\nabla C(\mathbf{A})$ .

---

### Algorithm 1 An Iterative JBD Algorithm

---

```

1: function JBD( $\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q, \{n_q\}_{q=1}^Q, \mathbf{m}, \text{threshold}$ )
2:    $\mathbf{A} \leftarrow \mathbf{I}$  ▷ Init
3:    $\mathbf{R}^{(q)} \leftarrow \overline{\mathbf{R}}_X^{(q)} \forall q$  ▷ Init
4:   while  $\|\nabla C(\mathbf{A})\| > \text{threshold}$  do
5:      $\nabla C(\mathbf{A}) \leftarrow \mathbf{I} - \langle \text{bdiag}_m^{-1} \{ \mathbf{R}^{(q)} \} \mathbf{R}^{(q)} \rangle$  ▷ (5.2)
6:     Evaluate  $\mathbf{T}$  ▷ Algorithm 2 for RG, Algorithm 3 for QN
7:      $\mathbf{R}^{(q)} \leftarrow \mathbf{T}^{-1} \mathbf{R}^{(q)} \mathbf{T}^{-\dagger}, q = 1, \dots, Q$ 
8:      $\mathbf{A} \leftarrow \mathbf{A} \mathbf{T}$  ▷ For output only
9:   end while
10:  return  $\mathbf{A}$ 
11: end function

```

---

It is interesting to compare our algorithm to the Gaussian ML independent vector analysis (IVA) algorithm of [84], which has a QN-type structure similar to Algorithm 3. Vía et al. [84], too, minimize the Kullback-Leibler divergence (KLD). However, the purpose of their algorithm is to block-diagonalize a single matrix.

---

**Algorithm 2** Update Step for the Relative Gradient Algorithm

---

- 1:  $\lambda \leftarrow 1$   $\triangleright$  Choose  $\lambda$ , e.g. by backtracking line search
  - 2: **while**  $C(\mathbf{A}(\mathbf{I} - \lambda \nabla C(\mathbf{A}))) > C(\mathbf{A}) - \alpha \lambda \text{tr} \{ \|\nabla C(\mathbf{A})\|^2 \}$  **do**  $\lambda \leftarrow \beta \lambda$
  - 3: **end while**
  - 4:  $\mathbf{T} \leftarrow \mathbf{I} - \lambda \nabla C(\mathbf{A})$
- 

---

**Algorithm 3** Update Step for the Quasi-Newton algorithm

---

- 1: **for**  $i=1:n, j=1:i-1$  **do**
  - 2:    $\mathbf{g}_{ij} \leftarrow [\nabla C(\mathbf{A})]_{ij}$   $\triangleright$  (5.4c)
  - 3:    $\mathbf{H}_{ij}^{(q)} \leftarrow \mathbf{R}_{jj}^{(q)} \otimes \mathbf{R}_{ii}^{-(q)}, q = 1, \dots, Q$   $\triangleright$  (5.4b)
  - 4:   Evaluate  $\mathcal{E}_{ij}, \mathcal{E}_{ji}$   $\triangleright$  (5.3)
  - 5: **end for**
  - 6: Reconstruct  $\mathcal{E}$  from  $\{\mathcal{E}_{ij}\}_{i \neq j}$   $\triangleright \mathcal{E}_{ii} \equiv \mathbf{0}_{m_i \times m_i}$ , see Sec. 5.2
  - 7:  $\mathbf{T} \leftarrow \mathbf{I} - \mathcal{E}$
- 

## 5.4 Numerical Examples

In this section, we study the convergence properties of the proposed algorithms by numerical simulations.

The data for the simulations is as follows. The input to the algorithm is a set of  $Q$   $m \times m$  matrices  $\overline{\mathbf{R}}_X^{(q)}$ , defined, following (2.9) and (2.4), as

$$\overline{\mathbf{R}}_X^{(q)} = \mathbf{A} \overline{\mathbf{R}}_S^{(q)} \mathbf{A}^\dagger. \quad (5.6)$$

Matrices reflecting the latent  $\overline{\mathbf{R}}_S^{(q)}$  are drawn from the Wishart distribution with  $n_q$  degrees of freedom and matrix parameter  $\mathbf{R}_S^{(q)}$ , mimicking  $n_q$  observations at each domain  $\mathcal{D}_q$  (recall the definitions in Chapter 2). The real positive-definite matrices  $\mathbf{R}_S^{(q)}$ , with block-pattern  $\mathbf{m}$ , are drawn as  $\mathbf{R}_{S,ii}^{(q)} = \mathbf{U}^\dagger \mathbf{U}$ , where  $\mathbf{U}$  is an  $m_i \times m_i$  upper triangular matrix whose independent and identically distributed (i.i.d.) entries  $\sim \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$ . The condition number of each  $\mathbf{R}_{S,ii}^{(q)}$  is limited by 500, to assure proper invertibility.  $\mathbf{A}$  is realized as  $\mathbf{A} = \mathbf{I} + \mathbf{Y}$ , where the entries of  $\mathbf{Y}$  are i.i.d. and  $\sim \mathcal{U}[-\frac{1}{4}, \frac{1}{4}]$ . Since the contrast function (2.17) is invariant to block-diagonal scale ambiguity (see (4.10) and (4.11)), we are concerned only about permutation ambiguity. The said choice of  $\mathbf{A}$ , together with initializing  $\mathbf{A}$  with  $\mathbf{I}$  (line 2 in Algorithm 1) allows for sufficient variability in our simulations, while usually assuring convergence to the desired minimum.

First, we observe and compare the convergence rate of the two proposed algorithms, in terms of the required number of iterations. The stopping threshold is set to  $10^{-4}$ . In the RG algorithm we set  $\lambda$  at each iteration using backtracking line search (lines 1–3 in Algorithm 2) with  $\alpha = 0.3, \beta = 0.2$ . The convergence rate of the

two algorithms is illustrated in Fig. 5.1. For each algorithm, there are 20 realizations of  $\mathbf{A}$ , with fixed latent  $\overline{\mathbf{R}}_S^{(q)}$  and  $\mathbf{R}_S^{(q)}$ . The fast convergence of the QN algorithm is very distinct from that of the RG algorithm. Both algorithms converge, eventually, in all trials, to the same value of the contrast function.

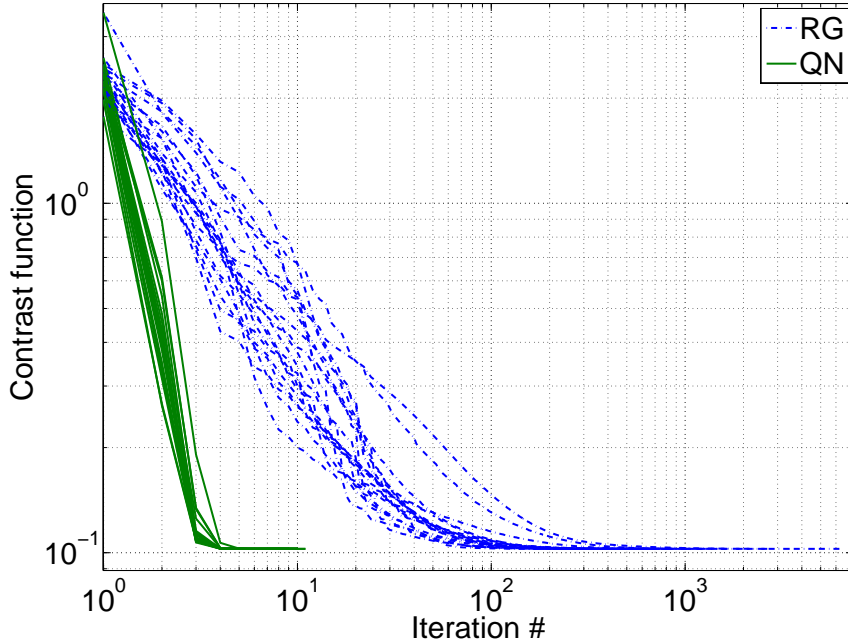


Figure 5.1: Convergence rate of the relative gradient and quasi-Newton algorithms, 20 trials each. Only  $\mathbf{A}$  varies at each trial.  $n_q = 100 \forall q$ .  $Q = 8$ . Block-pattern  $\mathbf{m} = [4, 3, 2]^\dagger$ . Threshold= $10^{-4}$ .

We now turn to verifying that the algorithms indeed converge to the minimum of the contrast function. That is, they are not trapped in a local minimum. As the former experiment has validated that both algorithms converge to the same value, we proceed with the QN realization, which is much faster. Since there is no closed-form solution for the minimization of (2.17), we perform a randomized test. This is achieved by perturbing the de-mixing matrix found by the QN algorithm,  $\widehat{\mathbf{B}}^{\text{JBD}} \triangleq (\widehat{\mathbf{A}}^{\text{JBD}})^{-1}$ , and recalculating the contrast function with the new matrix. The perturbation is obtained as follows. Define an  $m \times m$  matrix  $\mathbf{P}$ , whose entries are drawn i.i.d. from  $r \cdot u/50$ ,  $r \sim \mathcal{N}(0, 1)$ ,  $u \sim U[-\frac{1}{2}, \frac{1}{2}]$ . Then,  $\widehat{\mathbf{B}}^{\text{JBD}}$  is left-multiplied by  $\mathbf{I} + \mathbf{P}$ . The outcome is a slightly perturbed version of  $\widehat{\mathbf{B}}^{\text{JBD}}$ . For each such randomized  $\mathbf{P}$ , we calculate the contrast function  $C(\widehat{\mathbf{A}}^{\text{JBD}}(\mathbf{I} + \mathbf{P})^{-1})$ . If this value is smaller than  $C(\widehat{\mathbf{A}}^{\text{JBD}})$ , it means that the algorithm has failed in finding the minimum. Fig. 5.2 depicts graphically the results of such a test, for a certain choice of parameters and for a large number ( $10^4$ ) of trials. In each trial,  $\mathbf{m}$ ,  $n_q$ ,  $Q$  and the threshold are fixed.  $\mathbf{A}$  and  $\mathbf{R}_S^{(q)}$  are randomized as explained earlier in this section. In order to visualize the diversity of the perturbation matrices, the X-axis in Fig. 5.2

is the Frobenius norm of the perturbation matrix, minus the Frobenius norm of the  $m \times m$  identity matrix. The Y-axis depicts the value of the contrast function  $C(\mathbf{A})$  with  $\mathbf{A} = ((\mathbf{I} + \mathbf{P})\widehat{\mathbf{B}}^{\text{JBD}})^{-1}$ , minus  $C(\widehat{\mathbf{A}}^{\text{JBD}})$ , at different randomization of  $\mathbf{P}$ . A negative value on the Y-axis would imply that there exists another de-mixing matrix which yields a smaller  $C(\mathbf{A})$ . We have run such tests on various parameters, and all have shown that the QN algorithm indeed converges well to the desired minimum.

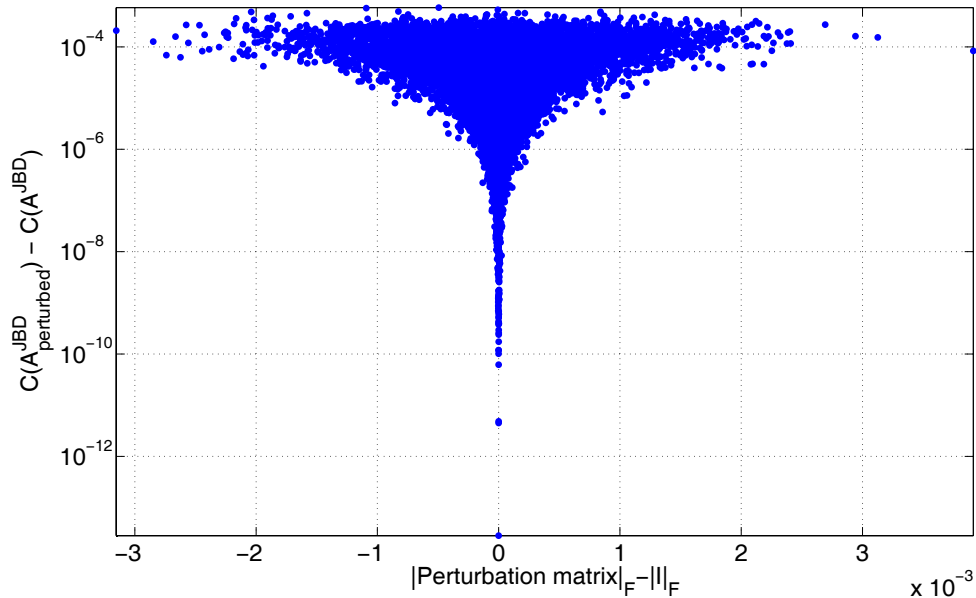


Figure 5.2: Numerical example to validate that the quasi-Newton algorithm has indeed converged to the correct minimum. Each dot stands for one of the  $10^4$  trials, i.e. a new randomized perturbation matrix  $\mathbf{P}$ . The X-axis quantifies the divergence of the perturbed matrix from the de-mixing matrix found by the algorithm. Negative values on the Y-axis would have meant that the algorithm did not converge to the (local) minimum. Fixed parameters are  $\mathbf{m} = [2, 2, 2]^\dagger$ ,  $Q = 12$ ,  $n_q = 100 \forall q$ , threshold= $10^{-9}$ .

## 5.5 Summary

In this chapter, we presented two non-orthogonal JBD algorithms: a quasi-Newton and a relative gradient one. These algorithms minimize the Kullback-Leibler-induced divergence between a set of symmetric positive-definite covariance matrices and a block-diagonal function thereof. Simulations demonstrate the correct convergence of these algorithms, given an appropriate initialization, to the desired minimum. As predicted, the QN algorithm converges much faster, in terms of the number of required iterations, than the RG one, by several orders of magnitude. Since the equations on which these algorithms are based are derived from a first-order ex-

pansion (of the estimating equations), then convergence is guaranteed only locally. That is, if the initialization is not too far from the global minimum.

## 5.A First-Order Expansion of the Estimating Equations

The purpose of this section is to derive the first-order expansion of the estimating equations (2.22) in terms of the latent *sources*. The need for this new representation stems from the fact that the block-diagonal structure, which is the core of the JBD algorithms, corresponds only to the covariance matrices of the sources, and not of the components. This derivation has some analogy to the derivation of the first-order expansion of the estimating equations in terms of components (2.28) in Appendix 3.A. However, in the current case, additional care has to be taken since the sources can be identified only up to the block-diagonal scale-ambiguity matrix  $\mathbf{Z}$ , discussed in Sec. 2.2.

Let us define an  $m \times m$  error matrix  $\boldsymbol{\mathcal{E}}$ , such that

$$\widehat{\mathbf{A}}^{-1} \mathbf{A} = \mathbf{Z}(\mathbf{I} + \boldsymbol{\mathcal{E}}), \quad (5.7)$$

where  $\widehat{\mathbf{A}}$  is a matrix which solves (2.22), now rewritten as

$$\langle ([\widehat{\mathbf{A}}^{-1} \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{A}}^{-\dagger}]_{ii})^{-1} [\widehat{\mathbf{A}}^{-1} \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{A}}^{-\dagger}]_{ij} \rangle = \mathbf{0}_{m_i \times m_j} \quad j \neq i. \quad (5.8)$$

In the notation of (5.7), matrix  $\boldsymbol{\mathcal{E}}$  reflects the relative change in the product of the estimated de-mixing matrix  $\widehat{\mathbf{A}}^{-1}$  with the “true” mixing matrix  $\mathbf{A}$ , up to scale ambiguity. When the estimation is precise, the product is block-diagonal. Using the notation (5.7), the terms within square brackets in (5.8) can be rewritten as

$$\begin{aligned} \widehat{\mathbf{A}}^{-1} \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{A}}^{-\dagger} &= \mathbf{Z}(\mathbf{I} + \boldsymbol{\mathcal{E}}) \mathbf{A}^{-1} \cdot \mathbf{A} \overline{\mathbf{R}}_S^{(q)} \mathbf{A}^\dagger \cdot \mathbf{A}^{-\dagger} (\mathbf{I} + \boldsymbol{\mathcal{E}})^\dagger \mathbf{Z}^\dagger \\ &= \mathbf{Z}(\mathbf{I} + \boldsymbol{\mathcal{E}}) \overline{\mathbf{R}}_S^{(q)} (\mathbf{I} + \boldsymbol{\mathcal{E}})^\dagger \mathbf{Z}^\dagger, \end{aligned} \quad (5.9)$$

where in the first equality we used  $\overline{\mathbf{R}}_X^{(q)} = \mathbf{A} \overline{\mathbf{R}}_S^{(q)} \mathbf{A}^\dagger$ , the empirical counterpart of (2.8). Since for any matrix  $\mathbf{M}$

$$[\mathbf{Z} \mathbf{M} \mathbf{Z}^\dagger]_{ij} = \mathbf{Z}_{ii} \mathbf{M}_{ij} \mathbf{Z}_{jj}^\dagger,$$

where  $i, j$  are the block indices, then if  $\mathbf{M}$  and  $\mathbf{Z}$  are invertible,

$$([\mathbf{Z}\mathbf{M}\mathbf{Z}^\dagger]_{ii})^{-1}[\mathbf{Z}\mathbf{M}\mathbf{Z}^\dagger]_{ij} = \mathbf{Z}_{ii}^{-\dagger} \mathbf{M}_{ii}^{-1} \mathbf{M}_{ij} \mathbf{Z}_{jj}^\dagger.$$

Now, by identifying  $\mathbf{M}$  with  $(\mathbf{I} + \boldsymbol{\mathcal{E}})\overline{\mathbf{R}}_S^{(q)}(\mathbf{I} + \boldsymbol{\mathcal{E}})^\dagger$ , (5.8) can be rewritten as

$$\mathbf{Z}_{ii}^{-\dagger} \langle ([(\mathbf{I} + \boldsymbol{\mathcal{E}})\overline{\mathbf{R}}_S(\mathbf{I} + \boldsymbol{\mathcal{E}})^\dagger]_{ii})^{-1} [(\mathbf{I} + \boldsymbol{\mathcal{E}})\overline{\mathbf{R}}_S(\mathbf{I} + \boldsymbol{\mathcal{E}})^\dagger]_{ij} \rangle \mathbf{Z}_{jj}^\dagger = \mathbf{0}_{m_i \times m_j}. \quad (5.10)$$

Since  $\mathbf{Z}$  is defined as a full-rank matrix, it can be removed from both sides of (5.10) without changing the solution with respect to  $\boldsymbol{\mathcal{E}}$ . Therefore, we now set out to linearize the estimating equations

$$\langle ([(\mathbf{I} + \boldsymbol{\mathcal{E}})\overline{\mathbf{R}}_S(\mathbf{I} + \boldsymbol{\mathcal{E}})^\dagger]_{ii})^{-1} [(\mathbf{I} + \boldsymbol{\mathcal{E}})\overline{\mathbf{R}}_S(\mathbf{I} + \boldsymbol{\mathcal{E}})^\dagger]_{ij} \rangle = \mathbf{0}_{m_i \times m_j} \quad j \neq i, \quad (5.11)$$

with respect to the error terms  $\boldsymbol{\mathcal{E}}$ .

From this point and on, the derivation follows lines similar to those in Appendix 3.A. Similarly to the arguments in Appendix 3.A, under asymptotic conditions,  $\overline{\mathbf{R}}_S^{(q)}$  converges, in the mean square, to  $\mathbf{Z}\mathbf{R}_S^{(q)}\mathbf{Z}^\dagger$ , and the ML estimator  $\widehat{\mathbf{A}}$  converges, in probability, to  $\mathbf{A}\mathbf{Z}^{-1}$  (asymptotically, for non-Gaussian components, both converge in probability). As for the rate of convergence, the entries of both  $\delta\mathbf{R}_S^{(q)}$  and  $\delta\boldsymbol{\mathcal{E}}$  are zero mean random variables with a standard deviation proportional to  $1/\sqrt{T}$ . Hence, asymptotically, terms which are proportional to  $\delta\mathbf{R}_S^{(q)}$  or  $\boldsymbol{\mathcal{E}}$  are  $\Omega(\frac{1}{\sqrt{T}})$  (this notation is defined in Chapter “Notations and Conventions”) and are considered as having the same order of magnitude.

By decomposing

$$\overline{\mathbf{R}}_S^{(q)} = \mathbf{R}_S^{(q)} + \delta\mathbf{R}_S^{(q)},$$

The term on the right-hand side (RHS) of (5.11) within the angle brackets can be expanded,  $\forall i, j$ , as

$$\begin{aligned} [(\mathbf{I} + \boldsymbol{\mathcal{E}})\overline{\mathbf{R}}_S(\mathbf{I} + \boldsymbol{\mathcal{E}})^\dagger]_{ij} &= \left[ \overline{\mathbf{R}}_S^{(q)} + \boldsymbol{\mathcal{E}}(\mathbf{R}_S^{(q)} + \delta\mathbf{R}_S^{(q)}) + (\mathbf{R}_S^{(q)} + \delta\mathbf{R}_S^{(q)})\boldsymbol{\mathcal{E}}^\dagger \right]_{ij} + \Omega(\frac{1}{T}) \\ &= \left[ \overline{\mathbf{R}}_S^{(q)} \right]_{ij} + \left[ \boldsymbol{\mathcal{E}}\mathbf{R}_S^{(q)} \right]_{ij} + \left[ \mathbf{R}_S^{(q)}\boldsymbol{\mathcal{E}}^\dagger \right]_{ij} + \Omega(\frac{1}{T}) \\ &= \overline{\mathbf{R}}_{S,ij}^{(q)} + \boldsymbol{\mathcal{E}}_{ij}\mathbf{R}_{S,jj}^{(q)} + \mathbf{R}_{S,ii}^{(q)}(\boldsymbol{\mathcal{E}}^\dagger)_{ij} + \Omega(\frac{1}{T}), \end{aligned} \quad (5.12)$$

where in the third equality we use the block-diagonal structure of  $\mathbf{R}_S^{(q)}$  and the notation (5.5).

As for the left-hand term within the angle brackets of (5.11),

$$\begin{aligned} \left( [(I + \boldsymbol{\mathcal{E}})\overline{\mathbf{R}}_S(I + \boldsymbol{\mathcal{E}})^\dagger]_{ii} \right)^{-1} &= \left( \mathbf{R}_{S,ii}^{(q)} + \delta \mathbf{R}_{S,ii}^{(q)} + \boldsymbol{\mathcal{E}}_{ii} \mathbf{R}_{S,ii}^{(q)} + \mathbf{R}_{S,ii}^{(q)} \boldsymbol{\mathcal{E}}_{ii}^\dagger + \Omega\left(\frac{1}{T}\right) \right)^{-1} \\ &= \mathbf{R}_{S,ii}^{-{(q)}} + \Omega\left(\frac{1}{\sqrt{T}}\right), \end{aligned} \quad (5.13)$$

where the first equality is due to (5.12). Multiplying (5.13) with (5.12),

$$\begin{aligned} \left( [(I + \boldsymbol{\mathcal{E}})\overline{\mathbf{R}}_S(I + \boldsymbol{\mathcal{E}})^\dagger]_{ii} \right)^{-1} [(I + \boldsymbol{\mathcal{E}})\overline{\mathbf{R}}_S(I + \boldsymbol{\mathcal{E}})^\dagger]_{ij} \\ &= \mathbf{R}_{S,ii}^{-{(q)}} \left( \overline{\mathbf{R}}_{S,ij}^{(q)} + \boldsymbol{\mathcal{E}}_{ij} \mathbf{R}_{S,jj}^{(q)} + \mathbf{R}_{S,ii}^{(q)} \boldsymbol{\mathcal{E}}_{ji}^\dagger \right) + \Omega\left(\frac{1}{T}\right) \\ &= \mathbf{R}_{S,ii}^{-{(q)}} \overline{\mathbf{R}}_{S,ij}^{(q)} + \mathbf{R}_{S,ii}^{-{(q)}} \boldsymbol{\mathcal{E}}_{ij} \mathbf{R}_{S,jj}^{(q)} + \boldsymbol{\mathcal{E}}_{ji}^\dagger + \Omega\left(\frac{1}{T}\right). \end{aligned} \quad (5.14)$$

Averaging (5.14) over all domains yields the linearized form of (5.11),

$$\langle \mathbf{R}_{S,ii}^{-{(q)}} \overline{\mathbf{R}}_{S,ij}^{(q)} \rangle + \langle \mathbf{R}_{S,ii}^{-{(q)}} \boldsymbol{\mathcal{E}}_{ij} \mathbf{R}_{S,jj}^{(q)} \rangle + \boldsymbol{\mathcal{E}}_{ji}^\dagger = \mathbf{0}_{m \times m} + \Omega\left(\frac{1}{T}\right) \quad , i \neq j. \quad (5.15)$$

Note the analogy between (5.15) and (3.33). Vectorizing (5.15) using (A.1c) in Appendix A yields

$$-\langle \text{vec}\{\mathbf{R}_{S,ii}^{-{(q)}} \overline{\mathbf{R}}_{S,ij}^{(q)}\} \rangle = \langle \mathbf{R}_{S,jj}^{(q)} \otimes \mathbf{R}_{S,ii}^{-{(q)}} \rangle \text{vec}\{\boldsymbol{\mathcal{E}}_{ij}\} + \boldsymbol{\mathcal{T}}_{m_j m_i} \text{vec}\{\boldsymbol{\mathcal{E}}_{ji}\} = \mathbf{0}_{m \times m} + \Omega\left(\frac{1}{\sqrt{T}}\right), \quad (5.16)$$

which leads to the desired form (5.3).



# Chapter 6

## Separation of Multidimensional Data Using a One-Dimensional Model Followed by Clustering: Performance Analysis

### 6.1 Outline

In this chapter, we compare the performance of two non-unitary criteria for the blind separation of multidimensional components. One has been shown in Sec. 2.3 to be optimal for piecewise-stationary multidimensional data, when second-order statistics are used. The other is the one-dimensional counterpart thereof [42], followed by grouping the one-dimensional components into their respective multidimensional subspaces. The latter procedure is denoted as clustering. In other words, we discuss a procedure for dependent component analysis which consists of two steps: the first step is the separation of the observations into terms which correspond to one-dimensional subspaces, and the second step consists of clustering these one-dimensional terms into multidimensional components. In Sec. 3.3, the performance, in terms of mean square error (MSE), of the one-dimensional criterion applied to multidimensional data, has been evaluated via numerical simulations. In this chapter, we obtain a closed-form expression for the expected MSE in the one-dimensional procedure. The proposed analysis allows us to predict the expected gain, in terms of the ratio of the MSE in both approaches, theoretically, without resorting to numerical simulations.

The outline of Chapter 6 is as follows. In Sec. 6.2 we formulate mathematically the problem which we denote by “mismodeling”, i.e. the use of a one-dimensional

model in order to separate multidimensional data. We describe the contrast function, the estimating equations and the component estimates for the mismodeled scenario. We then characterize the solutions to the contrast function in the mismodeled scenario in terms of invariance to certain transformations, that is, their equivalence class, and uniqueness. We prove that for matrices which can be exactly jointly block-diagonalized, exact joint block diagonalization (JBD) can be obtained by joint diagonalization (JD). Sec. 6.3 is dedicated to the error analysis. In this section, we obtain a closed-form expression for the MSE of multidimensional component separation under the mismodeling assumption. We begin in Sec. 6.3.1 by defining error terms and the desired figure of merit, the MSE, in terms of the well-defined quantities, i.e. components and projections. We then turn in Sec. 6.3.2 to representing the error terms using their one-dimensional factors. In particular, we obtain an expression for the MSE as a function of rank-1 error matrices. These rank-1 error matrices are approximated in Sec. 6.3.3 using the first-order Taylor expansion of the estimating equations. Finally, in Sec. 6.3.4, the covariance of these rank-1 error matrices is derived, and propagated into the closed-form terms of the MSE. In Sec. 6.4, we support our theoretical analysis by some numerical examples. First, in Sec. 6.4.1, we demonstrate that when the data can be exactly jointly block-diagonalized, a JD algorithm which minimizes our Kullback-Leibler divergence (KLD)-based criterion can indeed perform perfect JBD, up to clustering and permutation. Then, in Sec. 6.4.2, we validate the theoretical analysis of Sec. 6.3 by comparing the predicted MSE with the empirical one. We show for several multidimensional scenarios that the closed-form expression for the MSE is achievable, when small-error conditions hold. These numerical simulations demonstrate that our theoretical MSE expressions hold also for non-Gaussian data, as predicted from our analysis. The expected gain due to using the correct multidimensional model over the one-dimensional one is now given using the analytical expressions, whereas in Table 3.1 it was based on empirical experiments. In Sec. 6.4.3 we verify that the number of samples per domain used for the scenarios in Table 6.1 indeed guarantees asymptotic conditions. This is done by comparing the empirical variance of the oblique projection estimates with their predicted variance, as a function of the number of samples per domain. When asymptotic conditions hold, these values converge. In Sec. 6.4.4 we exploit the closed-form expressions for the MSE, for the mismodeled and optimal case, in order to observe the predicted ratio in MSE in various multidimensional scenarios. Chapter 6 is concluded with a discussion in Sec. 6.5.

Results for the specific case of only two components, one one-dimensional and

the other multidimensional, have been presented in Lahat et al. [73].

## 6.2 Problem Formulation

In this chapter, we focus on the case in which the data is multidimensional with block-pattern  $\mathbf{m}$ , but the analysis is performed under the (possibly false) assumption that the data consists of  $m$  one-dimensional components. In this case, instead of the contrast function (2.17), another target function, which we denote  $C^{\text{JD}}$ , is minimized:

$$C^{\text{JD}}(\mathbf{A}, \{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q) \triangleq \langle D(\mathbf{A}^{-1}\overline{\mathbf{R}}_X\mathbf{A}^{-\dagger}, \text{diag}\{\mathbf{A}^{-1}\overline{\mathbf{R}}_X\mathbf{A}^{-\dagger}\}) \rangle. \quad (6.1)$$

Naturally,  $C^{\text{JD}}(\mathbf{A}) \equiv C(\mathbf{A})$  when the *data* model follows  $\mathbf{m} = [1, \dots, 1]^\dagger$ . The minimization of (6.1) can be achieved by solving the one-dimensional version of the estimating equations (2.20),

$$\langle \text{diag}^{-1}\{\mathbf{A}^{-1}\overline{\mathbf{R}}_X\mathbf{A}^{-\dagger}\} \mathbf{A}^{-1}\overline{\mathbf{R}}_X\mathbf{A}^{-\dagger} \rangle = \mathbf{I}, \quad (6.2)$$

where  $\text{diag}^{-1}\{\cdot\}$  implies  $(\text{diag}\{\cdot\})^{-1}$ . Now, recall that when the model suits the data, there exists a matrix  $\mathbf{A}^{\text{JBD}}$  which solves

$$\langle \text{bdiag}_m^{-1}\{\mathbf{A}^{-1}\mathbf{R}_X\mathbf{A}^{-\dagger}\} \mathbf{A}^{-1}\mathbf{R}_X\mathbf{A}^{-\dagger} \rangle = \mathbf{I}$$

(this is (2.20) with  $\overline{\mathbf{R}}_X^{(q)}$  replaced with their expectation) and also satisfies  $C(\mathbf{A}^{\text{JBD}}, \{\mathbf{R}_X^{(q)}\}_{q=1}^Q) = 0$ . In the mismodeled case, the estimating equations

$$\langle \text{diag}^{-1}\{\mathbf{A}^{-1}\mathbf{R}_X\mathbf{A}^{-\dagger}\} \mathbf{A}^{-1}\mathbf{R}_X\mathbf{A}^{-\dagger} \rangle = \mathbf{I} \quad (6.3)$$

correspond to maximizing the likelihood of a one-dimensional model which does not suit the multidimensional data. They also correspond to minimizing  $C^{\text{JD}}(\mathbf{A}, \{\mathbf{R}_X^{(q)}\}_{q=1}^Q)$ . However, the latter amounts to jointly *diagonalizing*  $\{\mathbf{R}_X^{(q)}\}_{q=1}^Q$  — and by construction (recall Sec. 2.3.1), they are JBD-irreducible with block-pattern  $\mathbf{m}$ . Therefore, although a solution to (6.3) *can* be found<sup>1</sup>, since (6.3) reflects the stationary points of  $C^{\text{JD}}(\mathbf{A}, \{\mathbf{R}_X^{(q)}\}_{q=1}^Q)$ ,  $C^{\text{JD}}(\mathbf{A}, \{\mathbf{R}_X^{(q)}\}_{q=1}^Q) = 0$  *cannot* be satisfied.

In the sequel, in order to avoid ambiguities, for certain parameters a “ $\star$ ” is used to denote the “true” model quantities, and a hat denotes the value which minimizes the contrast function. For instance,  $\mathbf{A}^\star$  denotes the true mixing matrix,

<sup>1</sup>The existence of such a solution is discussed in Sec. 6.2.2.1.

while  $\widehat{\mathbf{A}}^{\text{JBD}}$  denotes a minimizer of the contrast function (2.17), which is also a maximum likelihood (ML) estimate if the Gaussian model holds. A matrix which minimizes (6.1) and thus also solves (6.2) is denoted  $\widehat{\mathbf{A}}^{\text{JD}}$ . Matrices which minimize  $C^{\text{JD}}(\mathbf{A}, \{\mathbf{R}_X^{(q)}\}_{q=1}^Q)$  and  $C(\mathbf{A}, \{\mathbf{R}_X^{(q)}\}_{q=1}^Q)$  are denoted  $\mathbf{A}^{\text{JD}}$  and  $\mathbf{A}^{\text{JBD}}$ , respectively.

## 6.2.1 Component Estimation under Mismatching

In the correct model setup, one obtains  $n$  oblique projections  $\widehat{\mathbf{P}}_i^{\text{JBD}}$  and their related component estimates  $\widehat{\mathbf{x}}_i^{\text{JBD}}(t)$  (defined in Sec. 3.2.1, where they were denoted simply as  $\widehat{\mathbf{P}}_i$  and  $\widehat{\mathbf{x}}_i(t)$ , respectively). In the mismatching setup, in the first stage, each column of the de-mixing matrix is associated with a rank-1 oblique projection matrix. By clustering, or grouping (correctly!) the  $m$  oblique projection matrices, each  $m_i$ -dimensional component can be regarded as a sum of  $m_i$  one-dimensional components:

$$\widehat{\mathbf{x}}_i^{\text{JD}}(t) \triangleq \widehat{\mathbf{P}}_i^{\text{JD}} \mathbf{x}(t) = \sum_{r=1}^{m_i} \widehat{\mathbf{P}}_{i_r}^{\text{JD}} \mathbf{x}(t) = \sum_{r=1}^{m_i} \widehat{\mathbf{x}}_{i_r}^{\text{JD}}(t), \quad (6.4)$$

where we have indexed the  $m$  rank-1 oblique projection matrices as  $\widehat{\mathbf{P}}_{i_1}^{\text{JD}}, \dots, \widehat{\mathbf{P}}_{i_{m_i}}^{\text{JD}}, \widehat{\mathbf{P}}_{j_1}^{\text{JD}}, \dots, \widehat{\mathbf{P}}_{j_{m_j}}^{\text{JD}}$ , etc. As a convention, the notation  $\mathbf{P}_{i_r}^{\text{JD}}$  implies an oblique projection constructed from the  $r$ th column of  $\mathbf{A}_i^{\text{JD}}$ . In the sequel, this double indexing system is used *only* for mismatched parameters. Hence, we often omit the superscript “JD” from double-indexed symbols. Throughout the text, indices  $i, j$  pertain to components. The first step of (6.4) follows from (2.5) (see also (6.8b)). The second step of (6.4) is due to

$$\widehat{\mathbf{P}}_i^{\text{JD}} \triangleq \sum_{r=1}^{m_i} \widehat{\mathbf{P}}_{i_r}^{\text{JD}}, \quad i = 1, \dots, n, \quad (6.5)$$

which is the mathematical formulation of “clustering”, as used in this thesis. The last step is again due to (2.5). At the last step of (6.4), the  $m$  one-dimensional components were indexed as  $\widehat{\mathbf{x}}_{i_1}^{\text{JD}}(t), \dots, \widehat{\mathbf{x}}_{i_{m_i}}^{\text{JD}}(t), \widehat{\mathbf{x}}_{j_1}^{\text{JD}}(t), \dots, \widehat{\mathbf{x}}_{j_{m_j}}^{\text{JD}}(t)$ , etc. For ease of reference, the various oblique projections used in Chapter 6 and the components which they generate are summarized in (6.6) and (6.8), respectively.

$$\widehat{\mathbf{P}}_{i_r}^{\text{JD}} \leftarrow \text{JD}(\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q) \quad (6.6a)$$

$$\widehat{\mathbf{P}}_i^{\text{JD}} = \sum_{r=1}^{m_i} \widehat{\mathbf{P}}_{i_r}^{\text{JD}} \leftarrow \text{JD}(\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q) + \text{clustering} \quad (6.6b)$$

$$\widehat{\mathbf{P}}_i^{\text{JBD}} \leftarrow \text{JBD}(\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q) \quad (6.6c)$$

$$\mathbf{P}_{i_r}^{\text{JD}} \leftarrow \text{JD}(\{\mathbf{R}_X^{(q)}\}_{q=1}^Q) \quad (6.6d)$$

$$\mathbf{P}_i^{\text{JD}} = \sum_{r=1}^{m_i} \mathbf{P}_{i_r}^{\text{JD}} \leftarrow \text{JD}(\{\mathbf{R}_X^{(q)}\}_{q=1}^Q) + \text{clustering} \quad (6.6e)$$

$$\mathbf{P}_i^* = \mathbf{A}_i \mathbf{B}_i \quad (6.6f)$$

$$\mathbf{P}_i^{\text{JBD}} \leftarrow \text{JBD}(\{\mathbf{R}_X^{(q)}\}_{q=1}^Q). \quad (6.6g)$$

In (6.6), the compound operation of minimizing  $C(\cdot, \cdot)$  (respectively,  $C^{\text{JD}}(\cdot, \cdot)$ ) and constructing the  $n$  (respectively,  $m$ ) oblique projections from the resulting de-mixing matrix is denoted as  $\text{JBD}(\cdot)$  (respectively,  $\text{JD}(\cdot)$ ). In this new notation we omit the mixing matrix, as it is only a transitional variable. Note that

$$\mathbf{P}_i^* = \mathbf{P}_i^{\text{JD}} = \mathbf{P}_i^{\text{JBD}}, \quad (6.7)$$

which follows from Sec. 6.2.2.1.

$$\widehat{\mathbf{x}}_{i_r}^{\text{JD}}(t) = \widehat{\mathbf{P}}_{i_r}^{\text{JD}} \mathbf{x}(t) \quad (6.8a)$$

$$\widehat{\mathbf{x}}_i^{\text{JD}}(t) = \widehat{\mathbf{P}}_i^{\text{JD}} \mathbf{x}(t) = \sum_{r=1}^{m_i} \widehat{\mathbf{P}}_{i_r}^{\text{JD}} \mathbf{x}(t) \quad (6.8b)$$

$$\widehat{\mathbf{x}}_i^{\text{JBD}}(t) = \widehat{\mathbf{P}}_i^{\text{JBD}} \mathbf{x}(t) \quad (6.8c)$$

$$\mathbf{x}_{i_r}^{\text{JD}}(t) = \mathbf{P}_{i_r}^{\text{JD}} \mathbf{x}(t) \quad (6.8d)$$

$$\mathbf{x}_i(t) = \mathbf{P}_i^* \mathbf{x}(t) = \mathbf{P}_i^{\text{JD}} \mathbf{x}(t) = \mathbf{P}_i^{\text{JBD}} \mathbf{x}(t) \quad (6.8e)$$

$$= \sum_{r=1}^{m_i} \mathbf{P}_{i_r}^{\text{JD}} \mathbf{x}(t) \quad (6.8f)$$

$$= \sum_{r=1}^{m_i} \mathbf{x}_{i_r}^{\text{JD}}(t). \quad (6.8g)$$

(6.8c) is identical to (3.2). (6.8e) follows from (6.7). (6.8g) follows from (6.8f) and (6.8d).

## 6.2.2 Characterizing the Solutions to the Mismodeled Contrast Function

In this section, we discuss the properties of the possible solutions to the minimization of the contrast function  $C^{\text{JD}}(\mathbf{A}, \{\bar{\mathbf{R}}_X^{(q)}\}_{q=1}^Q)$ , defined in (6.1).

### 6.2.2.1 JBD by JD

In the first step, we show that JD, that is, minimization of (6.1) or solution of (6.2), can yield *exact* separation of multidimensional data when the input matrices can be exactly jointly block-diagonalized. That is, that there exists a matrix  $\mathbf{A}^{\text{JD}}$  such that

$$\mathbf{A}^{\text{JD}} \triangleq \arg \min_{\mathbf{A}} C^{\text{JD}}(\mathbf{A}, \{\mathbf{R}_X^{(q)}\}_{q=1}^Q) \quad (6.9)$$

and

$$C(\mathbf{A}^{\text{JD}}, \{\mathbf{R}_X^{(q)}\}_{q=1}^Q) = 0. \quad (6.10)$$

The *existence* of a solution to (6.9) and (6.10) can be proved by construction, in a way similar to that proposed by [64]. First, we find an  $m_i \times m_i$  matrix  $\mathbf{G}_{ii}^{\text{JD}}$  which satisfies<sup>2</sup>

$$\begin{aligned} \mathbf{G}_{ii}^{\text{JD}} &= \arg \min_{\mathbf{G}_{ii}} C^{\text{JD}}(\mathbf{G}_{ii}^{-1}, \{\mathbf{R}_{S,ii}^{(q)}\}_{q=1}^Q) \\ &= \arg \min_{\mathbf{G}_{ii}} \langle D(\mathbf{G}_{ii} \mathbf{R}_{S,ii} \mathbf{G}_{ii}^\dagger, \text{diag}\{\mathbf{G}_{ii} \mathbf{R}_{S,ii} \mathbf{G}_{ii}^\dagger\}) \rangle. \end{aligned} \quad (6.11)$$

By definition,  $\mathbf{G}_{ii}^{\text{JD}}$  also satisfies (6.2), such that

$$\langle \text{diag}^{-1}\{\mathbf{G}_{ii}^{\text{JD}} \mathbf{R}_{S,ii} \mathbf{G}_{ii}^{\text{JD}\dagger}\} \mathbf{G}_{ii}^{\text{JD}} \mathbf{R}_{S,ii} \mathbf{G}_{ii}^{\text{JD}\dagger} \rangle = \mathbf{I}. \quad (6.12)$$

Equation (6.11) amounts to trying to jointly diagonalize the  $Q$  JBD-irreducible matrices  $\{\mathbf{R}_{S,ii}^{(q)}\}_{q=1}^Q$ . Next, we define a matrix  $\mathbf{W} \triangleq \mathbf{G}^{\text{JD}} \mathbf{A}^{\star-1}$ , where  $\mathbf{G}^{\text{JD}} \triangleq$

---

<sup>2</sup>At this point, some caution should be exercised. Theoretically, when the input matrices cannot be exactly jointly (block-) diagonalized, there may be cases in which only an infimum, but not a minimum, of  $C^{\text{JD}}(\cdot, \cdot)$  or  $C(\cdot, \cdot)$ , exists (see for example [85, Sec. 5], [86, Sec. 3.3]). In our numerical experiments, we have always found a stable solution. Characterizing the conditions for the existence of the minimum is beyond the scope of this work. In the sequel, we assume that an optimum exists.

$\text{bdiag}\{\mathbf{G}_{11}^{\text{JD}}, \dots, \mathbf{G}_{11}^{\text{JD}}\}$ . Replacing  $\mathbf{A}^{-1}$  with  $\mathbf{W}$  in the estimating equations (6.3) yields

$$\langle \text{diag}^{-1}\{\mathbf{W}\mathbf{R}_X\mathbf{W}^\dagger\} \mathbf{W}\mathbf{R}_X\mathbf{W}^\dagger \rangle = \langle \text{diag}^{-1}\{\mathbf{G}^{\text{JD}}\mathbf{R}_S\mathbf{G}^{\text{JD}\dagger}\} \mathbf{G}^{\text{JD}}\mathbf{R}_S\mathbf{G}^{\text{JD}\dagger} \rangle = \mathbf{I}, \quad (6.13)$$

where the first equality is due to (2.8). The last equality is due to the fact that both  $\mathbf{G}^{\text{JD}}$  and  $\mathbf{R}_S^{(q)}$  are block-diagonal with block-pattern  $\mathbf{m}$ , and (6.12). Equation (6.13) implies that there exists a matrix  $\mathbf{W}$  such that

$$\mathbf{W} = \arg \min_{\mathbf{W}'} C^{\text{JD}}(\mathbf{W}'^{-1}, \{\mathbf{R}_X^{(q)}\}_{q=1}^Q),$$

which settles (6.9) with  $\mathbf{A}^{\text{JD}} = \mathbf{W}^{-1}$ . On the other hand,  $\mathbf{W}\mathbf{R}_X^{(q)}\mathbf{W} \in \text{bdiag}_{\mathbf{m}} \forall q$ . Hence,  $\mathbf{W}$  is also the joint block-diagonalizer of  $\{\mathbf{R}_X^{(q)}\}_{q=1}^Q$  such that (6.10) is fulfilled. This concludes the proof for the existence of a solution to (6.9) which also solves (6.10). A graphical example which illustrates JBD by JD is given in Sec. 6.4.1.

Note: we did *not* prove that *any* solution of (6.9) is a solution of (6.10), although this is a plausible conjecture. This would require to prove that any solution to  $\min_{\mathbf{A}} C^{\text{JD}}(\mathbf{A}, \{\mathbf{R}_S^{(q)}\}_{q=1}^Q)$  is block-diagonal. This is beyond the scope of this work.

### 6.2.2.2 Equivalence Class

Given the existence of a minimum to (6.1) and a solution to (6.2) (see Sec. 6.2.2.1), we now discuss its equivalence class. That is, the subspace of solutions it generates which cannot be further reduced into smaller subspaces without additional constraints. The  $\text{diag}$  operator commutes with a diagonal matrix  $\mathbf{\Lambda}$  and a permutation matrix  $\boldsymbol{\pi}$ , both  $m \times m$ , in the following manner:

$$\begin{aligned} \text{diag}\{\mathbf{\Lambda}\mathbf{M}\mathbf{\Lambda}\} &= \mathbf{\Lambda} \text{diag}\{\mathbf{M}\}\mathbf{\Lambda} \\ \text{diag}\{\boldsymbol{\pi}\mathbf{M}\boldsymbol{\pi}^\dagger\} &= \boldsymbol{\pi} \text{diag}\{\mathbf{M}\}\boldsymbol{\pi}^\dagger \end{aligned} \quad \forall \mathbf{M} \in \mathbb{R}^{m \times m}. \quad (6.14)$$

Due to (6.14) and to (2.12), the contrast function  $C^{\text{JD}}(\mathbf{A}, \{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q)$  (6.1) and its related estimating equations (6.2) are invariant to an arbitrary invertible diagonal matrix and a permutation matrix. Mathematically,

$$C^{\text{JD}}(\mathbf{A}) = C^{\text{JD}}(\mathbf{A}\mathbf{\Lambda}\boldsymbol{\pi}), \quad (6.15)$$

with the additional constraint that  $\mathbf{\Lambda}$  be invertible. This invariance implies that a matrix which minimizes  $C^{\text{JD}}(\mathbf{A}, \{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q)$  (6.1) or solves its related estimating

equations (6.2) can be determined up to right-multiplication by any permuted arbitrary  $m \times m$  invertible diagonal matrix.

Let us now consider the case in which the permutation matrix is block-diagonal with block-pattern  $\mathbf{m}$ . This amounts to permutation within the scalar entries of each source vector  $\mathbf{s}_i(t)$ , such that for each component  $i$ ,  $\widehat{\mathbf{P}}_{i_r}^{\text{permuted}} = \widehat{\mathbf{P}}_{i_s}$ ,  $(r, s) \in i$  (this is a shorter notation for “ $(r, s) \in [1, \dots, m_i]$ ”). In this case, component reconstruction by clustering (6.5) is not affected:

$$\widehat{\mathbf{P}}_i^{\text{permuted}} = \sum_{s \in i} \widehat{\mathbf{P}}_{i_s} = \widehat{\mathbf{P}}_i^{\text{JD}}. \quad (6.16)$$

Recall from (4.10) and (4.11) that the contrast function (2.17) commutes with any invertible *block*-diagonal matrix with block-pattern  $\mathbf{m}$ . Hence, any solution to  $\min_{\mathbf{A}} C^{\text{JD}}(\mathbf{A}, \{\mathbf{R}_X^{(q)}\}_{q=1}^Q)$  is also a solution to  $C(\mathbf{A}, \{\mathbf{R}_X^{(q)}\}_{q=1}^Q) = 0$ , as demonstrated in Sec. 6.2.2.1. The converse is in general false. This property of the mismodeling solution, that it commutes only with diagonal matrices, and not with a block-diagonal matrix, is the key point in the error analysis of the mismodeling scenario, as detailed in Sec. 6.3. This issue is further stressed in Sec. 6.2.2.3.

In the sequel, we assume that permutations are limited to within the block-pattern  $\mathbf{m}$ . Global permutation, that is, between different components, is not considered in this work.

### 6.2.2.3 Uniqueness

We now turn to the question of *uniqueness*: can there exist two different solutions to (6.9) which are not related by an invertible diagonal permuted matrix (6.14)? This would imply that  $C^{\text{JD}}(\mathbf{A}, \{\mathbf{R}_X^{(q)}\}_{q=1}^Q)$  has more than one global minimum, as conjectured in [24, Sec. 2.2]. For the correct model case, this question has been discussed in Sec. 4.5 and answered by Theorem 4.1. In the sequel, as a working assumption, we postulate that at least in the small-errors regime, which will be defined in Sec. 6.3, such different global minima, if exist, are sufficiently far apart such that, given that the permutation ambiguity (6.16) has been resolved,

$$\widehat{\mathbf{P}}_{i_r}^{\text{JD}} \xrightarrow{\text{asymptotically}} \mathbf{P}_{i_r}^{\text{JD}} \quad \forall i_r. \quad (6.17)$$

Violation of this working assumption will cause our theoretical error expression to fail. However, as we see in the numerical examples (Sec. 6.4), when the small-error assumptions hold, our analysis indeed follows the empirical results. To conclude, in the sequel we assume that the solution of (6.1) is unique up to an  $\mathbf{m}$ -permuted

diagonal matrix.

## 6.3 Error Analysis

The purpose of our analysis is to obtain a closed-form expression, which does not depend on empirical finite-length data, for the MSE in component separation, given the mismodeling assumption of Sec. 6.2. The MSE is defined in Sec. 6.3.1. The MSE is essentially expressed as a product of the components' energy and the covariance of the error in the oblique projections. In Sec. 6.3.2 we express this covariance in terms of error matrices, which can be obtained in closed-form from the linearized estimating equations. These are derived in Sec. 6.3.3. Sec. 6.3.4 is dedicated to obtaining the closed-form expression for this covariance. That is, only in terms of the model parameters, without resorting to finite-data terms.

We consider an asymptotic analysis in the regime of small errors, in which the results are obtained from a first-order expansion of the estimating equations (6.2). In the following, we define asymptotic conditions as  $T \rightarrow \infty$  with  $\frac{n_q}{T}$  fixed  $\forall q$ . The analysis is conducted under the assumption that the data model of Sec. 2.3.1 holds.

For readability, whenever there is no risk of ambiguity, we have omitted the superscript ‘‘JD’’. Whenever we have a partition of a subspace into one-dimensional subspaces, which is denoted by double indexing, we always refer only to the partition defined by (6.6d). This shorter notation is used in the sequel.

### 6.3.1 Defining the Figure of Merit

A difficulty in error analysis for the mismodeling problem stems from the inability to characterize the estimation error of the mixing matrix, due to the severe indeterminacies it suffers from, as discussed in Sec. 2.2. As in Sec. 3.2.1, we begin by defining convenient error terms. In order to focus on well-defined quantities, we consider the errors  $\delta \mathbf{P}_i^{\text{JD}}$  in  $\widehat{\mathbf{P}}_i^{\text{JD}}$ , the estimates (6.6b) of the oblique projectors  $\mathbf{P}_i$ . The estimated  $i$ th component is thus

$$\widehat{\mathbf{x}}_i^{\text{JD}}(t) = \widehat{\mathbf{P}}_i^{\text{JD}} \mathbf{x}(t) = (\mathbf{P}_i^* + \delta \mathbf{P}_i^{\text{JD}}) \mathbf{x}(t) = \mathbf{x}_i(t) + \delta \mathbf{P}_i^{\text{JD}} \mathbf{x}(t),$$

where the first equality follows from (6.8b) and the last one from (6.8e). Hence, the component estimation error is defined as

$$\Delta \mathbf{x}_i^{\text{JD}}(t) \triangleq \widehat{\mathbf{x}}_i^{\text{JD}}(t) - \mathbf{x}_i(t) = \delta \mathbf{P}_i^{\text{JD}} \mathbf{x}(t). \quad (6.18)$$

Let us define the normalized MSE of the estimation error of a given component  $i$  as

$$\widehat{\text{MSE}}_i^{\text{JD}} \triangleq \frac{1}{\sigma_i^2} \frac{1}{T} \sum_{t=1}^T |\widehat{\mathbf{x}}_i^{\text{JD}}(t) - \mathbf{x}_i(t)|^2, \quad (6.19)$$

where the normalization is by the average power  $\sigma_i^2$  of the  $i$ th component (3.21). Substituting (6.18) in (6.19),

$$\begin{aligned} \widehat{\text{MSE}}_i^{\text{JD}} &= \frac{1}{\sigma_i^2} \frac{1}{T} \sum_{t=1}^T |\delta \mathbf{P}_i^{\text{JD}} \mathbf{x}(t)|^2 \\ &= \frac{1}{\sigma_i^2} \frac{1}{T} \sum_{t=1}^T \text{tr} \{ \delta \mathbf{P}_i^{\text{JD}} \mathbf{x}(t) \mathbf{x}^\dagger(t) (\delta \mathbf{P}_i^{\text{JD}})^\dagger \} \\ &= \frac{1}{\sigma_i^2} \text{tr} \{ (\langle \overline{\mathbf{R}}_X \rangle \otimes \mathbf{I}) \text{vec} \{ \delta \mathbf{P}_i^{\text{JD}} \} \text{vec}^\dagger \{ \delta \mathbf{P}_i^{\text{JD}} \} \}, \end{aligned} \quad (6.20)$$

where the second equality is analogous to that in (3.21), and the last equality uses (2.9) and then Property A.1 in Appendix A.

In order to obtain the expectation of (6.20), we employ the approximation that  $\widehat{\mathbf{P}}_i^{\text{JD}}$ , and thus  $\delta \mathbf{P}_i^{\text{JD}}$ , are statistically independent of the *total* power of the observations,  $\langle \overline{\mathbf{R}}_X \rangle$ . This approximation becomes more accurate with larger  $Q$ . A similar approximation has been used in Sec. 3.2.4 for  $\widehat{\mathbf{P}}_i^{\text{JBD}}$ . Hence, we can write

$$\begin{aligned} \text{MSE}_i^{\text{JD}} &\triangleq E \left\{ \widehat{\text{MSE}}_i^{\text{JD}} \right\} \\ &= \frac{1}{\sigma_i^2} \text{tr} \{ (\langle \mathbf{R}_X \rangle \otimes \mathbf{I}) \text{Cov}(\text{vec} \{ \delta \mathbf{P}_i^{\text{JD}} \}) \}. \end{aligned} \quad (6.21)$$

It now remains to evaluate the matrices  $\text{Cov}(\text{vec} \{ \delta \mathbf{P}_i^{\text{JD}} \})$ . Their derivation is given in the following subsections.

### 6.3.2 Error Decomposition

Analogously to the error analysis in the correct model case in Sec. 3.2, the first step of the current error analysis is to factorize the estimation error (6.18) into uniquely-defined terms. Equation (6.15) implies that  $\text{span}(\mathbf{A}_{i,r})$  is uniquely defined, where  $\mathbf{A}_{i,r}$  denotes the  $r$ th column of  $\mathbf{A}_i^{\text{JD}}$ ,  $r = 1, \dots, m_i$ . Consequently, also the rank-1 oblique projections  $\mathbf{P}_{i,r}^{\text{JD}}$  are uniquely defined. This is the key point to the error analysis in this chapter, which will henceforth be based on these projections. In addition, we

use the notion of global uniqueness as in Sec. 6.2.2.3. Summoning (6.8b) and (6.8f),

$$\Delta \mathbf{x}_i^{\text{JD}}(t) = \sum_{r \in i} (\widehat{\mathbf{P}}_{i_r} \mathbf{x}(t) - \mathbf{P}_{i_r} \mathbf{x}(t)) = \sum_{r \in i} \delta \mathbf{P}_{i_r} \mathbf{x}(t). \quad (6.22)$$

In the last step we have introduced the notation

$$\delta \mathbf{P}_{i_r} \triangleq \widehat{\mathbf{P}}_{i_r} - \mathbf{P}_{i_r} \quad (6.23)$$

as well as the implied factorization/reconstruction

$$\delta \mathbf{P}_i = \sum_{r \in i} \delta \mathbf{P}_{i_r}. \quad (6.24)$$

For the error notation (6.23) to be consistent, we assume that (6.17) holds. Obviously, in practice, this assumption cannot be guaranteed, as the permutation matrix generated by the algorithm is arbitrary by nature (recall Sec. 6.2.2.2); however, for the sake of the theoretical analysis, one can assume that the same arbitrary permutation has been applied both to  $\widehat{\mathbf{P}}_{i_r}$  and  $\mathbf{P}_{i_r}$ . This assumption is sufficient for our proposed small-error framework to be valid. Furthermore, since we are eventually interested in reconstructing components, not in their rank-1 decompositions, then (6.16) guarantees that our small-error analysis for components be valid regardless of the arbitrary permutation.

Proceeding with the factorization into rank-1 terms, (6.22) can be rewritten as

$$\begin{aligned} \Delta \mathbf{x}_i^{\text{JD}}(t) &= \sum_{r \in i} \delta \mathbf{P}_{i_r} \sum_{j=1}^n \sum_{s \in j} \mathbf{x}_{j_s}(t) = \sum_{j=1}^n \sum_{r \in i} \sum_{s \in j} \delta \mathbf{P}_{i_r} \mathbf{\Pi}_{j_s} \mathbf{x}_{j_s}(t) \\ &= \sum_{j=1}^n \sum_{r \in i} \sum_{s \in j} \mathcal{E}_{i_r j_s} \mathbf{x}_{j_s}(t). \end{aligned} \quad (6.25)$$

The first equality stems from summing (6.8g) over all  $n$  components. The second equality consists of a change of order of the summations, plus the insertion of the orthogonal projection on the subspace spanned by the  $j_s$ th column of  $\mathbf{A}^{\text{JD}}$ . The term  $\mathbf{\Pi}_{j_s}$  in (6.25) is inserted since it arises naturally in the first-order expansion of the estimating equations (Appendix 6.A, (6.47)). Recall that  $\mathbf{\Pi}_{j_s} \mathbf{x}_{j_s}(t) = \mathbf{x}_{j_s}(t) \forall j$  and  $s \in j$ . For the last equality in (6.25), we have defined  $m \times m$  error matrices

$$\mathcal{E}_{i_r j_s} \triangleq \delta \mathbf{P}_{i_r} \mathbf{\Pi}_{j_s}. \quad (6.26)$$

Similarly to the error matrices (3.4), also here, the double-indexed term  $\mathcal{E}_{i_r j_s}$  gives

the linearized estimating equations their pairwise form, as will be seen in Sec. 6.3.3.

In analogy to the error terms in Sec. 3.2.1, for  $i \neq j$ , the term  $\boldsymbol{\mathcal{E}}_{i_r j_s} \mathbf{x}_j(t)$  is called the  $(i_r, j_s)$ -*contamination error*, that is, the contamination due to the  $j_s$ th component in the reconstruction of the  $i_r$ th component. The term  $\boldsymbol{\mathcal{E}}_{i_r i_r} \mathbf{x}_i(t)$  is called the  $i_r i_r$ -*reconstruction error*, since this term represents a distortion of  $\mathbf{x}_i(t)$  but not any contamination by the other components.

Given the notation (6.26), we now turn to writing  $\delta \mathbf{P}_i^{\text{JD}}$  as a function of these terms. Let us begin by factorizing  $\delta \mathbf{P}_i^{\text{JD}}$  as

$$\begin{aligned} \delta \mathbf{P}_i^{\text{JD}} &= \delta \mathbf{P}_i^{\text{JD}} \sum_{j=1}^n \mathbf{P}_j^{\text{JD}} = \sum_{r \in i} \delta \mathbf{P}_{i_r} \sum_{j=1}^n \sum_{s \in j} \mathbf{P}_{j_s} = \sum_{r \in i} \sum_{j=1}^n \sum_{s \in j} \delta \mathbf{P}_{i_r} \boldsymbol{\Pi}_{j_s} \mathbf{P}_{j_s} \\ &= \sum_{j=1}^n \sum_{r \in i} \sum_{s \in j} \boldsymbol{\mathcal{E}}_{i_r j_s} \mathbf{P}_{j_s} = \sum_{r \in i} \sum_{s \in i} \boldsymbol{\mathcal{E}}_{i_r i_s} \mathbf{P}_{i_s} + \sum_{j \neq i} \sum_{r \in i} \sum_{s \in j} \boldsymbol{\mathcal{E}}_{i_r j_s} \mathbf{P}_{j_s}. \end{aligned} \quad (6.27)$$

The first equality is due to  $\sum_{j=1}^n \mathbf{P}_j^{\text{JD}} = \mathbf{I}$ , which always holds for any complete set of oblique projections onto a subspace. The second equality follows from (6.6e) and (6.24). The third equality is due to  $\boldsymbol{\Pi}_{j_s} \mathbf{P}_{j_s} = \mathbf{P}_{j_s}$ , which always holds, by definition of the orthogonal and oblique projections onto the same subspaces. In the fourth equality, we have used (6.26). In the last step, we separated the terms with index  $j = i$  from  $j \neq i$ . The terms on the second summand at the last step of (6.27) can be obtained from the linearized estimating equations, which shall be presented in Sec. 6.3.3. However, the terms on the first summand:  $\sum_{r \in i} \sum_{s \in i} \boldsymbol{\mathcal{E}}_{i_r i_s} \mathbf{P}_{i_s} = \delta \mathbf{P}_i \mathbf{P}_i^{\text{JD}}$ , cannot. We therefore resort to the following factorization. Since  $\mathbf{x}(t)$  is given (observed), we exploit it by constraining  $\sum_{j=1}^n \widehat{\mathbf{x}}_j^{\text{JD}}(t) = \mathbf{x}(t)$ . Therefore,  $\sum_{j=1}^n \widehat{\mathbf{P}}_j^{\text{JD}} = \mathbf{I}$ , which implies  $\sum_{j=1}^n \delta \mathbf{P}_j^{\text{JD}} = \mathbf{0}_{m \times m}$ . Multiplying  $\delta \mathbf{P}_i^{\text{JD}} = -\sum_{j \neq i} \delta \mathbf{P}_j^{\text{JD}}$  on the right with  $\mathbf{P}_i^{\text{JD}}$ , we obtain

$$\begin{aligned} \delta \mathbf{P}_i^{\text{JD}} \mathbf{P}_i^{\text{JD}} &= -\sum_{j \neq i} \delta \mathbf{P}_j^{\text{JD}} \mathbf{P}_i^{\text{JD}} = -\sum_{j \neq i} \sum_{s \in j} \delta \mathbf{P}_{j_s} \sum_{r \in i} \mathbf{P}_{i_r} \\ &= -\sum_{j \neq i} \sum_{s \in j} \sum_{r \in i} \delta \mathbf{P}_{j_s} \boldsymbol{\Pi}_{i_r} \mathbf{P}_{i_r} = -\sum_{j \neq i} \sum_{s \in j} \sum_{r \in i} \boldsymbol{\mathcal{E}}_{j_s i_r} \mathbf{P}_{i_r}. \end{aligned} \quad (6.28)$$

The second transition follows from (6.24) and (6.6e). In the third step, we have changed the order of summation and introduced  $\boldsymbol{\Pi}_{i_r} \mathbf{P}_{i_r} = \mathbf{P}_{i_r}$ , which always holds. In the last step, we used (6.26). Substituting the final result of (6.28) in the last equality of (6.27) yields

$$\delta \mathbf{P}_i^{\text{JD}} = \sum_{j \neq i} \left( \sum_{r \in i} \sum_{s \in j} \boldsymbol{\mathcal{E}}_{i_r j_s} \mathbf{P}_{j_s} - \sum_{r \in i} \sum_{s \in j} \boldsymbol{\mathcal{E}}_{j_s i_r} \mathbf{P}_{i_r} \right),$$

which is the desired form: an expression for  $\delta \mathbf{P}_i^{\text{JD}}$  which consists only of terms which can be calculated from the linearized estimating equations of Sec. 6.3.3 and from the model parameters. Vectorizing  $\delta \mathbf{P}_i^{\text{JD}}$  using (A.1c) in Appendix A, its covariance matrix is given by

$$\begin{aligned} \text{Cov}(\text{vec}\{\delta \mathbf{P}_i^{\text{JD}}\}) = & \sum_{j \neq i} \left( \sum_{\substack{l \in j \\ s \in j}} (\mathbf{P}_{jl}^\dagger \otimes \mathbf{I}) \sum_{\substack{m \in i \\ r \in i}} \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{imjl}\}, \text{vec}\{\boldsymbol{\mathcal{E}}_{irjs}\}) (\mathbf{P}_{js} \otimes \mathbf{I}) \right. \\ & - \sum_{\substack{l \in j \\ r \in i}} (\mathbf{P}_{jl}^\dagger \otimes \mathbf{I}) \sum_{\substack{m \in i \\ s \in j}} \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{imjl}\}, \text{vec}\{\boldsymbol{\mathcal{E}}_{jsir}\}) (\mathbf{P}_{ir} \otimes \mathbf{I}) \\ & - \sum_{\substack{m \in i \\ s \in j}} (\mathbf{P}_{im}^\dagger \otimes \mathbf{I}) \sum_{\substack{l \in j \\ r \in i}} \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{jlim}\}, \text{vec}\{\boldsymbol{\mathcal{E}}_{irjs}\}) (\mathbf{P}_{js} \otimes \mathbf{I}) \\ & \left. + \sum_{\substack{m \in i \\ r \in i}} (\mathbf{P}_{im}^\dagger \otimes \mathbf{I}) \sum_{\substack{l \in j \\ s \in j}} \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{jlim}\}, \text{vec}\{\boldsymbol{\mathcal{E}}_{jsir}\}) (\mathbf{P}_{ir} \otimes \mathbf{I}) \right). \end{aligned} \quad (6.29)$$

Substituting (6.29) in (6.21), followed by using (A.1d) in Appendix A and the property that any two different components are statistically independent, it turns out that the two middle terms in (6.29) do not contribute to the MSE. Hence, (6.21) can be rewritten as

$$\begin{aligned} \text{MSE}_i^{\text{JD}} = \sigma_i^{-2} \text{tr} \left\{ (\langle \mathbf{R}_X \rangle \otimes \mathbf{I}) \sum_{j \neq i} \left( \right. \right. & \quad (6.30) \\ & \sum_{\substack{l \in j \\ s \in j}} (\mathbf{P}_{jl}^\dagger \otimes \mathbf{I}) \sum_{\substack{m \in i \\ r \in i}} \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{imjl}\}, \text{vec}\{\boldsymbol{\mathcal{E}}_{irjs}\}) (\mathbf{P}_{js} \otimes \mathbf{I}) \\ & \left. \left. + \sum_{\substack{m \in i \\ r \in i}} (\mathbf{P}_{im}^\dagger \otimes \mathbf{I}) \sum_{\substack{l \in j \\ s \in j}} \text{Cov}(\text{vec}\{\boldsymbol{\mathcal{E}}_{jlim}\}, \text{vec}\{\boldsymbol{\mathcal{E}}_{jsir}\}) (\mathbf{P}_{ir} \otimes \mathbf{I}) \right) \right\}. \end{aligned}$$

It now remains to evaluate the covariance matrices in (6.30). Their derivation is given in the following subsections.

### 6.3.3 Influence Function

In order to evaluate the covariance of the estimation error, required for the MSE (6.30), we first establish the first-order expansion of  $\boldsymbol{\mathcal{E}}_{irjs}$  in terms of the finite-sample  $m \times m$  covariance matrices

$$\overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)} \triangleq \frac{1}{n_q} \sum_{t \in \mathcal{D}_q} \mathbf{x}_{i_r}(t) \mathbf{x}_{j_s}^\dagger(t). \quad (6.31)$$

The key assumption for blind separation is block-decorrelation:

$$\mathbf{R}_{X_{i_r} X_{j_s}}^{(q)} \triangleq E \left\{ \overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)} \right\} = \mathbf{P}_{i_r} \mathbf{R}_X^{(q)} \mathbf{P}_{j_s}^\dagger = \mathbf{0}_{m \times m} \quad (6.32)$$

for  $j \neq i$ . However, because of finite sample size, this does not hold for its empirical counterpart, i.e.,  $\overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)} \neq \mathbf{0}_{m \times m}$ . In this section, we develop the performance analysis in the regime of small errors. That is, we analyze the error terms  $\widehat{\mathbf{P}}_{i_r} - \mathbf{P}_{i_r}$  at first-order in  $\overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)}$  when asymptotic conditions hold. From (6.26),  $\boldsymbol{\mathcal{E}}_{i_r j_s}$  decreases with  $T$  at the same rate as  $\delta \mathbf{P}_{i_r}$ . Assuming that asymptotic conditions hold, then  $\widehat{\mathbf{P}}_{i_r} \cong \mathbf{P}_{i_r}$  (see Appendix 6.A).

The estimating equations (6.2) can also be expressed in terms of the oblique projections. When the model was multidimensional, the estimating equations (2.20) could be rewritten as a set of  $n(n-1)$  equations (2.28). Analogously, when the model is one-dimensional, the estimating equations (6.2) can be rewritten as a set of  $(m^2 - \sum_{i=1}^n m_i^2)$  equations

$$\langle (\mathbf{P}_{i_r} \overline{\mathbf{R}}_X \mathbf{P}_{i_r}^\dagger)^\sharp (\mathbf{P}_{i_r} \overline{\mathbf{R}}_X \mathbf{P}_{j_s}^\dagger) \rangle = 0, \quad i \neq j. \quad (6.33)$$

Note that  $n(n-1) = m^2 - \sum_{i=1}^n m_i^2$  for the simplest case of  $n = 2$ ,  $m_1 = m_2 = 1$ . In this case, in which these two models coincide, there are only two equations: one for  $(i, j)$  and one for  $(j, i)$ . The estimating equations (6.33) allow us to lead the error analysis from the components' point of view. The linearized form of (6.33), which will be introduced shortly, yields the set of pairwise error matrices  $(\boldsymbol{\mathcal{E}}_{i_r j_s}, \boldsymbol{\mathcal{E}}_{j_s i_r})$ ,  $i \neq j$ , whose covariance shall be used in (6.30). Indeed, there are additional equations, with  $j = i$ , which reflect the main block-diagonal of (6.2), and which obey

$$\langle (\mathbf{P}_{i_r} \overline{\mathbf{R}}_X \mathbf{P}_{i_r}^\dagger)^\sharp (\mathbf{P}_{i_r} \overline{\mathbf{R}}_X \mathbf{P}_{i_m}^\dagger) \rangle = 0, \quad i_r \neq i_m.$$

However, these are not required for the error analysis since the error matrices which they provide:  $\boldsymbol{\mathcal{E}}_{i_r i_m}$ , do not take part in the expression (6.30) for the MSE. Note that the  $(i_r, i_r)$ th entry of (6.2), that is,  $i_r = j_s$  of (6.33), degenerates into the identity matrix: the diagonal entries  $i_r = j_s$  do not yield any constraints, reflecting the indeterminacy discussed in Sec. 2.2.

The first-order expansion of the estimating equations (6.33) yields (see Appendix 6.A), for each pair of  $i \neq j$ , the following set of equations:

$$-\text{vec} \{ \langle \mathbf{R}_{i_r i_r}^\sharp \overline{\mathbf{R}}_{i_r j_s} \rangle \} = \sum_{l=1}^{m_j} \langle \mathbf{R}_{j_s j_l} \otimes \mathbf{R}_{i_r i_r}^\sharp \rangle \text{vec} \{ \boldsymbol{\mathcal{E}}_{i_r j_l} \}$$

$$\begin{aligned}
& + \sum_{k \neq r}^{m_i} (I \otimes \langle \mathbf{R}_{i_r i_r}^\# \mathbf{R}_{i_r i_k} \rangle) \mathcal{T} \text{vec}\{\mathcal{E}_{j_s i_k}\} + \mathcal{T} \text{vec}\{\mathcal{E}_{j_s i_r}\} + \Omega\left(\frac{1}{T}\right) \\
-\text{vec}\{\langle \mathbf{R}_{j_s j_s}^\# \overline{\mathbf{R}}_{j_s i_r} \rangle\} & = \mathcal{T} \text{vec}\{\mathcal{E}_{i_r j_s}\} + \sum_{l \neq s}^{m_j} (I \otimes \langle \mathbf{R}_{j_s j_s}^\# \mathbf{R}_{j_s j_l} \rangle) \mathcal{T} \text{vec}\{\mathcal{E}_{i_r j_l}\} \\
& + \sum_{k=1}^{m_i} \langle \mathbf{R}_{i_r i_k} \otimes \mathbf{R}_{j_s j_s}^\# \rangle \text{vec}\{\mathcal{E}_{j_s i_k}\} + \Omega\left(\frac{1}{T}\right), \tag{6.34}
\end{aligned}$$

where  $r = 1, \dots, m_i$ ,  $s = 1, \dots, m_j$ ,  $i \neq j$ .  $\mathcal{T}$  denotes the commutation matrix [79],  $\text{vec}\{\mathbf{M}^\dagger\} = \mathcal{T} \text{vec}\{\mathbf{M}\}$  for any  $m \times m$  matrix  $\mathbf{M}$ , defined in (A.3).

Equation (6.34) shows that asymptotically, for each pair of components, the  $2m_i m_j$  projector error terms  $\{\mathcal{E}_{i_r j_s}, \mathcal{E}_{j_s i_r}\}_{r=1, s=1}^{m_i, m_j}$  are related to the corresponding set of  $2Qm_i m_j$  matrices  $\{\overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)}, \overline{\mathbf{R}}_{X_{j_s} X_{i_r}}^{(q)}\}_{q=1}^Q$ , which represents the block-decorrelation error. We have thus obtained the same type of pairwise decoupling as in Sec. 3.2.2.

For each pair  $i \neq j$ , the  $2m_i m_j$  equations (6.34) can be rewritten in matrix form as

$$\mathbf{g}_{ij} = -\mathcal{H} \mathbf{e}_{ij} + \Omega\left(\frac{1}{T}\right), \tag{6.35}$$

as we now explain.  $\mathbf{g}_{ij}$  and  $\mathbf{e}_{ij}$  are  $2m_i m_j m^2 \times 1$  vectors,

$$\mathbf{g}_{ij} \triangleq \begin{bmatrix} \mathbf{g}_{i_1 j_1} \\ \mathbf{g}_{j_1 i_1} \\ \vdots \\ \mathbf{g}_{i_1 j_{m_j}} \\ \mathbf{g}_{j_{m_j} i_1} \\ \vdots \\ \mathbf{g}_{i_{m_i} j_1} \\ \mathbf{g}_{j_1 i_{m_i}} \\ \vdots \\ \mathbf{g}_{i_{m_i} j_{m_j}} \\ \mathbf{g}_{j_{m_j} i_{m_i}} \end{bmatrix} \quad \text{and} \quad \mathbf{e}_{ij} \triangleq \begin{bmatrix} \text{vec}\{\mathcal{E}_{i_1 j_1}\} \\ \text{vec}\{\mathcal{E}_{j_1 i_1}\} \\ \vdots \\ \text{vec}\{\mathcal{E}_{i_1 j_{m_j}}\} \\ \text{vec}\{\mathcal{E}_{j_{m_j} i_1}\} \\ \vdots \\ \text{vec}\{\mathcal{E}_{i_{m_i} j_1}\} \\ \text{vec}\{\mathcal{E}_{j_1 i_{m_i}}\} \\ \vdots \\ \text{vec}\{\mathcal{E}_{i_{m_i} j_{m_j}}\} \\ \text{vec}\{\mathcal{E}_{j_{m_j} i_{m_i}}\} \end{bmatrix}, \tag{6.36}$$

where the matrices on the left-hand side (LHS) of (6.34) are collected into  $m^2 \times 1$  vectors denoted as

$$\mathbf{g}_{i_r j_s} = \text{vec}\{\langle \mathbf{R}_{X_{i_r} X_{i_r}}^\# \overline{\mathbf{R}}_{X_{i_r} X_{j_s}} \rangle\}. \tag{6.37}$$

The  $2m_i m_j m^2 \times 2m_i m_j m^2$  matrix  $\mathcal{H}$  collects the terms which precede the vectorized error matrices in (6.34). The indexing order suggested in (6.36) allows to write  $\mathcal{H}$  as a symmetric matrix, which is not always the case with an arbitrary indexing choice. In any case, the symmetry of  $\mathcal{H}$  is only for the aesthetic value and has no practical implications in the sequel. An example of the explicit form of  $\mathcal{H}$ , for the simplest and non-trivial case of  $m_i = 1$ ,  $m_j = 2$ , is given in (6.63) in Appendix 6.C.

Assuming that  $\mathcal{H}$  is invertible (it is conjectured that the conditions in Theorem 4.1 are necessary and sufficient also for the mismodeled case, since asymptotically, both JBD and JD attain the same results; further analysis of this issue is beyond the scope of this work), then

$$\mathbf{e}_{ij} = -\mathcal{H}^{-1} \mathbf{g}_{ij} + \Omega\left(\frac{1}{T}\right). \quad (6.38)$$

Equation (6.38) shows how the empirical correlation between components, that is, the fact that  $\overline{\mathbf{R}}_{X_{ir} X_{js}}^{(q)}$  is non-zero in finite sample size, results in non-zero errors  $\mathcal{E}_{irjs}$ . Note the analogy between (6.38) and its multidimensional counterpart (3.12). Both equations, (6.38) and (3.12), merge when  $\mathbf{R}_S^{(q)} = \text{diag}\{\mathbf{R}_S^{(q)}\} \forall q$ . Equation (6.38) is the desired closed-form, first-order expression for the error terms in (6.25).

### 6.3.4 Error Covariance and Mean Square Error

We shall now use the explicit expression (6.38) for the error terms, in order to derive an explicit expression for the covariances used in Sec. 6.3.2, in (6.29) and in (6.30). The covariance matrices

$$\begin{aligned} & \text{Cov}(\text{vec}\{\mathcal{E}_{imjl}\}, \text{vec}\{\mathcal{E}_{irjs}\}) \quad \text{and} \\ & \text{Cov}(\text{vec}\{\mathcal{E}_{imjl}\}, \text{vec}\{\mathcal{E}_{jsir}\}), \quad \forall i, j, r \in i, s \in j, \end{aligned}$$

required for (6.29) and (6.30) are taken from the blocks of the covariance of  $\mathbf{e}_{ij}$  in (6.38), where this covariance is given by

$$\text{Cov}(\mathbf{e}_{ij}) = \mathcal{H}^{-1} \text{Cov}(\mathbf{g}_{ij}) \mathcal{H}^{-\dagger} + O\left(\frac{1}{T^2}\right). \quad (6.39)$$

In order to obtain  $\text{Cov}(\mathbf{e}_{ij})$ , we shall calculate the covariance matrix of the stochastic vector  $\mathbf{g}_{ij}$ , defined in (6.36). An explicit expression for the entries of  $\text{Cov}(\mathbf{g}_{ij})$  is derived in Appendix 6.B. We hereby quote only the final results, given in (6.54). Based on the zero-mean assumption, we can write  $\text{Cov}(\mathbf{g}_{ij}) = E\left\{\mathbf{g}_{ij} \mathbf{g}_{ij}^\dagger\right\}$ . Then,

the building blocks of  $\text{Cov}(\mathbf{g}_{ij})$ , based on the ordering in (6.36), are

$$\begin{aligned} E \left\{ \mathbf{g}_{irj_s} \mathbf{g}_{imj_l}^\dagger \right\} &= \frac{1}{T} \langle \mathbf{R}_{X_{j_s} X_{j_l}} \otimes (\mathbf{R}_{X_{ir} X_{ir}}^\# \mathbf{R}_{X_{ir} X_{im}} \mathbf{R}_{X_{im} X_{im}}^\#) \rangle, \\ E \left\{ \mathbf{g}_{j_s i_r} \mathbf{g}_{j_l i_m}^\dagger \right\} &= \frac{1}{T} \langle \mathbf{R}_{X_{ir} X_{im}} \otimes (\mathbf{R}_{X_{j_s} X_{j_s}}^\# \mathbf{R}_{X_{j_s} X_{j_l}} \mathbf{R}_{X_{j_l} X_{j_l}}^\#) \rangle, \\ E \left\{ \mathbf{g}_{irj_s} \mathbf{g}_{j_l i_m}^\dagger \right\} &= \frac{1}{T} \langle \mathbf{R}_{X_{j_s} X_{j_l}} \mathbf{R}_{X_{j_l} X_{j_l}}^\# \otimes \mathbf{R}_{X_{ir} X_{ir}} \mathbf{R}_{X_{ir} X_{im}} \rangle \mathcal{T}, \\ E \left\{ \mathbf{g}_{j_s i_r} \mathbf{g}_{imj_l}^\dagger \right\} &= \frac{1}{T} \langle \mathbf{R}_{X_{ir} X_{im}} \mathbf{R}_{X_{im} X_{im}}^\# \otimes \mathbf{R}_{X_{j_s} X_{j_s}}^\# \mathbf{R}_{X_{j_s} X_{j_l}} \rangle \mathcal{T}, \end{aligned}$$

where we use the convention that  $r, m \in i, s, l \in j, i \neq j$ . Unlike the correct model, in which the covariance matrix has been simplified (3.17) into a single matrix, in the mismodeled case, the structure of  $\text{Cov}(\mathbf{g}_{ij})$  and  $\mathcal{H}$  is more complex. We thus leave the possible derivation of a simplified expression for (6.39) for future work. As an illustrative example, the explicit expression of  $\text{Cov}(\mathbf{g}_{ij})$  for the simplest case of  $m_i = 1, m_j = 2$  is given in (6.64) in Appendix 6.C. We therefore content ourself with the numerical calculation of  $\text{Cov}(\mathbf{e}_{ij})$  using their explicit *analytical* building blocks  $\mathcal{H}$  and  $\text{Cov}(\mathbf{g}_{ij})$ .

Given the explicit expression for (6.39), we have thus obtained an expression for the MSE (6.30), which can be calculated numerically by the model parameters  $\{\mathbf{R}_{X_{ir} X_{j_s}}^{(q)}\}_{q=1}^Q$  and the weights  $\{n_q\}_{q=1}^Q$ , up to  $O(\frac{1}{T^2})$  terms.

It should be noted that all the derivations in this section 6.3 and in the related appendices do not rely neither on Gaussian distribution nor on statistics of order larger than 2. Therefore, the derived numerical expression for (6.29) and (6.30) hold also for non-Gaussian observations.

It is immediate to verify that in the presence of only one-dimensional components, that is,  $m_i = 1 \forall i$ , the results in this chapter identify with their counterparts in Chapter 3.

## 6.4 Numerical Examples

In this section, we validate the theoretical results which were derived in previous sections. The ability of a JD algorithm which minimizes our contrast function to perform JBD, up to clustering and permutation, is demonstrated in Sec. 6.4.1. The theoretical prediction of the expected MSE is compared with empirical values in Sec. 6.4.2. The asymptotic conditions required for the validity of the small-errors regime of the theoretical analysis are justified in Sec. 6.4.3. Finally, in Sec. 6.4.4, we use the closed-form expressions for the MSE, in the correct and mismodeled cases, to examine numerically the possible gain due to using the correct model

(alternatively, the loss due to using the one-dimensional model). We illustrate how this gain depends on some of the model parameters.

In all the following numerical experiments, the symmetric positive-definite block-diagonal matrices  $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$  are drawn with  $\mathbf{R}_{S,ii}^{(q)} = \mathbf{U}^\dagger \mathbf{U}$  where  $\mathbf{U}$  is an  $m_i \times m_i$  upper triangular matrix with independent entries uniformly distributed on  $[-\frac{1}{2}, \frac{1}{2}]$ . JBD is performed using the quasi-Newton (QN) algorithm of Chapter 5 and Lahat et al. [60]. JD is performed using Pham's algorithm [57]. The latter achieves exactly the same solution as our JBD algorithm with  $\mathbf{m} = [1, \dots, 1]^\dagger$ . However, it is sometimes faster.

### 6.4.1 JBD by JD Followed by Clustering

First, we illustrate the theoretical result of Sec. 6.2.2.1, that JD followed by clustering can achieve exact JBD when the original data consists of JBD-irreducible positive-definite block-diagonal matrices, mixed by the same invertible matrix. In the following example, the entries of the mixing matrix are independent and identically distributed (i.i.d.), and taken from the standard normal distribution  $\mathcal{N}(0, 1)$ . The  $Q = 3$  block-diagonal matrices are created as explained at the beginning of Sec. 6.4, with  $\mathbf{m} = [5, 5, 5, 5]^\dagger$ . Fig. 6.1(a) illustrates the original matrices  $\mathbf{R}_S^{(q)}$ . Fig. 6.1(b) illustrates the output of the JD algorithm. Fig. 6.1(c) depicts Fig. 6.1(b) after clustering using the algorithm suggested by [64, Sec. 3.1]. The data in Fig. 6.1 is given after scaling the matrix entries to their  $\log_{10}(|\cdot|)$ . This scaling makes visible the small numerical errors on the off block-diagonal, after clustering. These negligible numerical errors are inevitable, since the algorithm stops when a certain non-zero threshold is achieved. It should be noted that these results do not contradict a similar experiment conducted in [64, Sec. 4], since in the latter, the original block-diagonal matrices were symmetric but not positive-definite, as in our case.

### 6.4.2 Validation of the Performance Analysis

In this section, we validate experimentally the performance analysis of Sec. 6.3. In the following simulation, we construct the data such that the analysis requirements of Sec. 6.3 hold, including the small-errors regime. Therefore, the theoretical expression for the MSE is expected to be an accurate prediction of the measured error. The underlying sources are created by left-multiplying the Cholesky factorization of  $\mathbf{R}_S^{(q)}$  with  $m \times n_q$  statistically independent, zero mean, unit variance samples. These samples are drawn from various distributions, in order to support our claim that the second-order analysis holds not only for Gaussian sources.

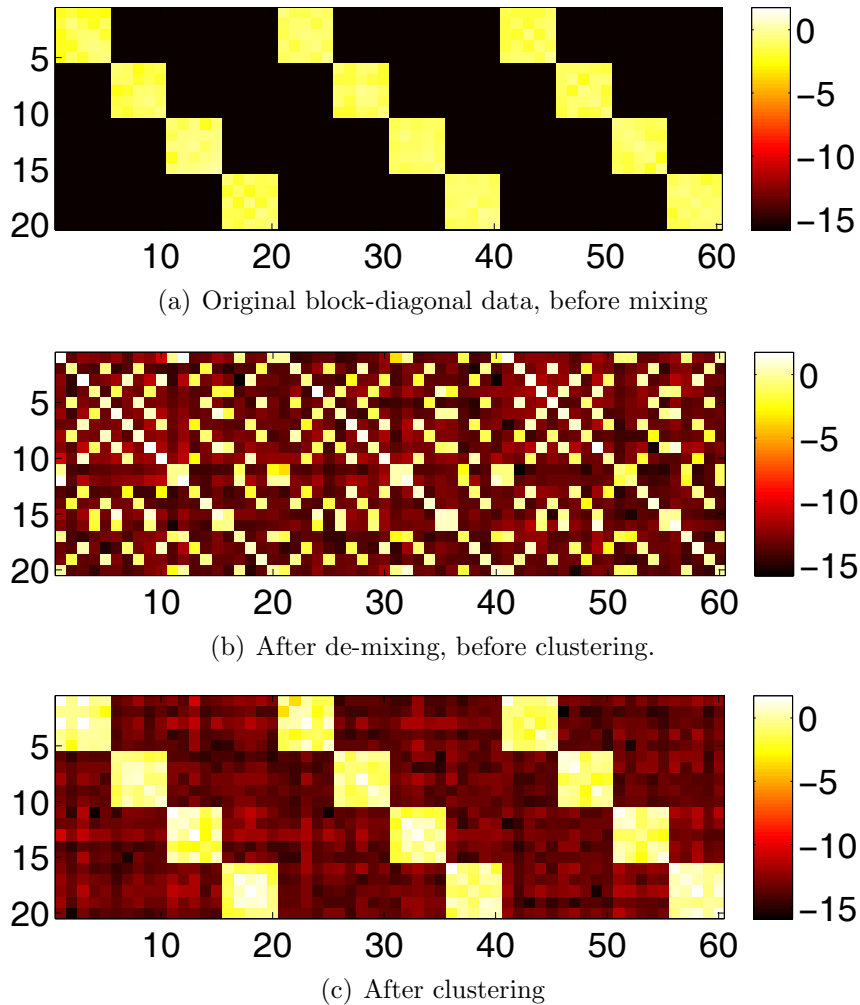


Figure 6.1: JBD by JD followed by clustering.  $Q = 3$ ,  $\mathbf{m} = [5, 5, 5, 5]^\dagger$ . Entries of  $\mathbf{A}$  taken from  $\mathcal{N}(0, 1)$ . Demonstration on one draw of  $\mathbf{A}$  and  $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$ . The displayed matrix entries are scaled to  $\log_{10}(|\cdot|)$  of their original value.

As explained in Sec. 6.2.2.2, there is no scale indeterminacy to resolve. Given that our analysis is correct, the only origin of discrepancy between prediction and experiment should be due to permutation or clustering which mix different components. In order to avoid the need to check for such errors at each Monte-Carlo trial, the following precautions were taken. We set  $Q = 5$  adjacent domains with  $n_q = 5000$  samples for a total of  $T = 25000$  samples. According to the results in Sec. 6.4.3, this  $n_q$  guarantees that the small-error assumption is valid. In addition, we chose to initialize the JD algorithm with  $\mathbf{I}$ . In this case, permutation errors and the need to cluster the data are avoided by choosing mixing matrices  $\mathbf{A}$  which are strictly diagonally-dominant. In the following simulations, the  $(i, j)$ th entry of matrix  $\mathbf{A}$  is given by  $\mathbf{A}_{ij} = 3\delta_{ij} + \frac{1}{10}r_{ij}$ ,  $r_{ij} \sim \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$  and  $r_{ij}$  are i.i.d. Such values allow for sufficient variability of the mixing matrix to test our small-error analysis,

while avoiding the need to resolve the permutation and cluster the data into components at each trial. Permutation and clustering errors are easily detected, since they result in a significantly larger MSE. Therefore, as a final safety measure, we verified that no such large errors appeared in our results.

Table 6.1 compares the empirical with the analytical MSE for several scenarios with varying component dimensions and distributions. The second column states the arbitrary index given to each component. The third column denotes the dimension of the  $i$ th component in the scenario. In each scenario, different  $\mathbf{A}$  and  $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$  are drawn.  $\mathbf{A}$  and  $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$  are fixed at each scenario. The fourth column gives the analytical MSE for each component (6.30), which is calculated using the correct model parameters. The fifth column gives the ratio between the analytical MSE given a one-dimensional model (6.30) and the MSE given the correct block-pattern  $\mathbf{m}$ . The MSE given the correct block-pattern is given in (3.25). This ratio represents the theoretical gain of performance due to using the correct model vs. the one-dimensional one. The theoretical MSE, as well as the ratio, are a function of both  $\mathbf{A}$  and  $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$ . Each scenario is evaluated using 5000 Monte-Carlo trials. For each scenario, two data types were tested. In columns 6–7, Gaussian, zero mean, unit-variance samples are used to create the underlying sources. In columns 9–10, either uniform or Gaussian mixture (peaks centered at  $\pm 4$ ) zero mean, unit-variance samples, denoted U and GM, respectively, are used to create the underlying sources. Note that left-multiplication of the non-Gaussian samples with the Cholesky factorization of  $\mathbf{R}_S^{(q)}$  changes their distribution; however, it is still non-Gaussian. The non-Gaussian distribution, used to generate the data for each scenario, is given in column 8. The sixth and ninth columns give the averaged empirical MSE for each component (6.19). Columns 7 and 10 give the ratio of MSE for component separation under the one-dimensional assumption: simulated vs. analytical.

The last row of Table 6.1 summarizes the results of each column. First, note that all the (normalized) MSE values are much smaller than 1, illustrating the quality of the component separation. Second, note that all values in column 7 and 10 are close to 1, showing that our analysis predicts correctly the achievable separation accuracy. The good match between predicted and empirical MSE also in column 10 is due to the fact that only second-order statistics are required for our theoretical analysis. The analytical results in column 5 validate the empirical results of Table 3.1, that there is indeed a significant gain due to using the correct model, rather than JD followed by clustering. An important result is that in scenarios #2 and #3, which include one-dimensional components along with higher-dimensional ones, the gain for the one-dimensional components is  $> 1$ , too. Obviously, when the sources are

independent, as in the last scenario (#5), there is no difference between JD and JBD, hence we put ‘1’ in this cell.

Table 6.1: Performance of ICA of multidimensional data, when the separation is based on a one-dimensional model and second-order statistics. The figure of merit is the normalized MSE. Each scenario, that is, different  $\mathbf{m}$ , is tested once with Gaussian and once with non-Gaussian data. In each scenario,  $\mathbf{A}$  and  $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$  are fixed. The empirical MSE is averaged over 5000 Monte-Carlo trials. In each Monte-Carlo trial, different underlying sources are drawn. The analytical MSE is compared with that of using the correct  $\mathbf{m}$  ( $\text{MSE}_i^{\text{JBD}}$ ). The last row of the table summarizes the columns.

Scenario	$i$	$m_i$	$\text{MSE}_i^{\text{JD}}$	$\frac{\text{MSE}_i^{\text{JD}}}{\text{MSE}_i^{\text{JBD}}}$	$\widehat{\text{MSE}}_i^{\text{JD}}$	$\frac{\widehat{\text{MSE}}_i^{\text{JD}}}{\text{MSE}_i^{\text{JD}}}$	Non-Gaussian Data Type	$\widehat{\text{MSE}}_i^{\text{JD}}$	$\frac{\widehat{\text{MSE}}_i^{\text{JD}}}{\text{MSE}_i^{\text{JD}}}$
Model			Gaussian data			Non-Gaussian data			
#1	1	2	$6.75 \cdot 10^{-4}$	2.167	$6.78 \cdot 10^{-4}$	1.006	GM	$6.68 \cdot 10^{-4}$	0.991
	2	2	$4.48 \cdot 10^{-4}$	1.471	$4.46 \cdot 10^{-4}$	0.995		$4.45 \cdot 10^{-4}$	0.993
	3	2	$6.00 \cdot 10^{-4}$	2.216	$6.01 \cdot 10^{-4}$	1.002		$6.13 \cdot 10^{-4}$	1.021
	4	2	$8.47 \cdot 10^{-4}$	2.741	$8.53 \cdot 10^{-4}$	1.007		$8.55 \cdot 10^{-4}$	1.009
#2	1	6	$1.58 \cdot 10^{-4}$	2.812	$1.59 \cdot 10^{-4}$	1.008	U	$1.57 \cdot 10^{-4}$	0.994
	2	2	$5.84 \cdot 10^{-4}$	2.021	$5.94 \cdot 10^{-4}$	1.018		$5.83 \cdot 10^{-4}$	1.000
	3	1	$1.41 \cdot 10^{-3}$	3.603	$1.41 \cdot 10^{-3}$	0.998		$1.39 \cdot 10^{-3}$	0.990
#3	1	1	$7.60 \cdot 10^{-4}$	3.674	$7.56 \cdot 10^{-4}$	0.995	U	$7.57 \cdot 10^{-4}$	0.996
	2	2	$3.58 \cdot 10^{-4}$	1.713	$3.57 \cdot 10^{-4}$	0.998		$3.59 \cdot 10^{-4}$	1.004
	3	5	$1.28 \cdot 10^{-4}$	2.486	$1.28 \cdot 10^{-4}$	0.996		$1.28 \cdot 10^{-4}$	1.000
#4	1	5	$2.41 \cdot 10^{-4}$	5.485	$2.41 \cdot 10^{-4}$	0.997	GM	$2.39 \cdot 10^{-4}$	0.990
	2	4	$3.21 \cdot 10^{-4}$	5.485	$3.20 \cdot 10^{-4}$	0.997		$3.18 \cdot 10^{-4}$	0.990
#5	1	1	$5.26 \cdot 10^{-5}$	1	$5.31 \cdot 10^{-5}$	0.992	GM	$5.21 \cdot 10^{-5}$	0.992
	2	1	$2.61 \cdot 10^{-5}$	1	$2.63 \cdot 10^{-5}$	0.998		$2.61 \cdot 10^{-5}$	0.998
	3	1	$4.85 \cdot 10^{-5}$	1	$4.82 \cdot 10^{-5}$	1.003		$4.87 \cdot 10^{-5}$	1.003
			$\ll 1$	$\geq 1$	$\ll 1$	$\cong 1$		$\ll 1$	$\cong 1$

### 6.4.3 Validating the Small-Errors Regime

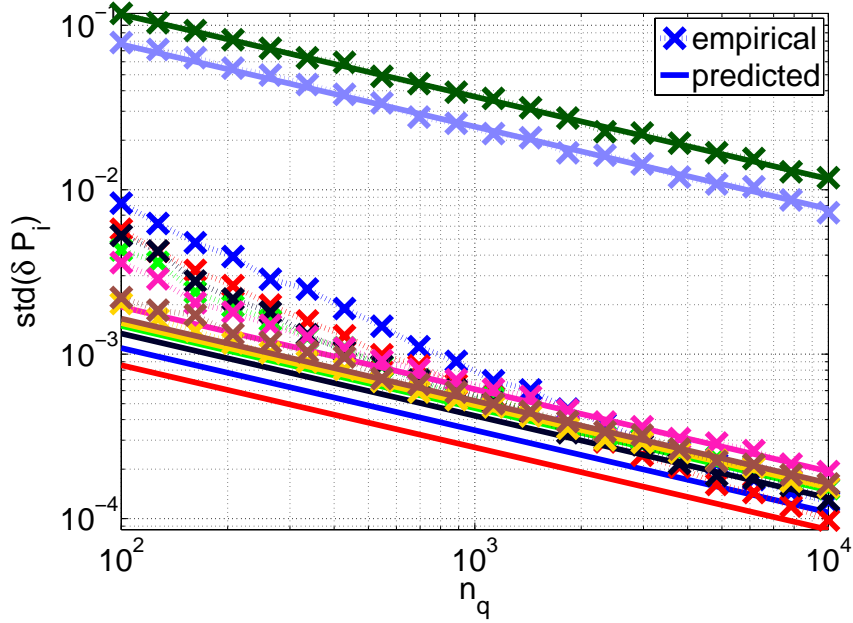
In this section, we justify our choice of  $n_q$  for the simulations in Sec. 6.4.2. That is, we show that small-error conditions indeed apply when  $n_q = 5000$ . For this aim, we compare the empirical covariance of the entries of  $\delta \mathbf{P}_i^{\text{JD}}$  with their predicted value, given by (6.29). Fig. 6.2 depicts results for scenario #2 in Table 6.1, when the underlying sources are drawn from the uniform distribution. The entries of  $\mathbf{P}_i^{\text{JD}}$ ,

that is, the oblique projections onto the subspace of the  $i$ th component in scenario #2, are illustrated in Fig. 6.2(c). These entries are scaled to their  $\log_{10}(|\cdot|)$ , in order to enhance the difference between the smaller and larger values. Fig. 6.2(a) refers to data taken from the 6th row of  $\mathbf{P}_2^{\text{JD}}$ . For each of the  $m = 9$  entries in this row, Fig. 6.2(a) compares their empirical standard deviation with the predicted one. The empirical values are averaged over 500 Monte-Carlo trials and denoted by ‘ $\times$ ’ markers. In the small-errors regime, the standard deviation of the entries of  $\delta\mathbf{P}_i^{\text{JD}}$  is proportional to  $1/\sqrt{T}$ , as follows from (6.54), (6.39) and (6.29). These predicted values are depicted by the continuous straight lines in Fig. 6.2(a). The two largest standard deviation values in Fig. 6.2(a) correspond to values in columns 7,8 of the sixth row of  $\mathbf{P}_2$ . It is visible that the empirical covariances converge to the  $1/\sqrt{T}$  slope at around  $n_q = 3000$ . At  $n_q = 5000$ , the difference between the predicted and empirical values is already sufficiently small. Hence, in Table 6.1 we get a good match between the empirical and theoretical values. Similar trends have been found for the other entries of the oblique projection estimates in all the scenarios of Table 6.1. Using too low  $n_q$  results in bias and discrepancy between experiment and theory. These results also validate our theoretical expression (6.29).

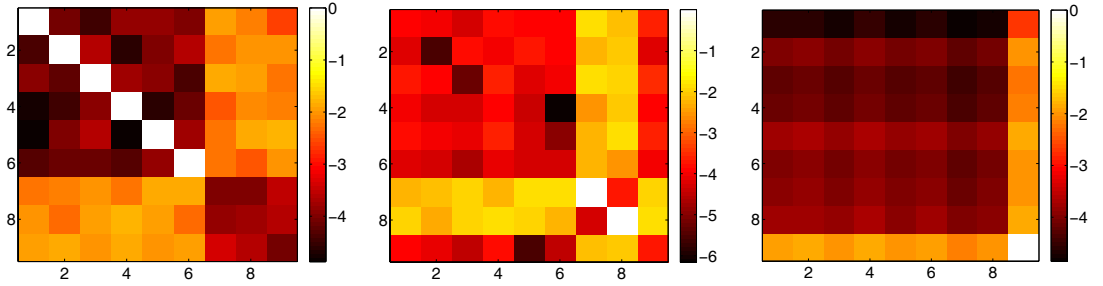
#### 6.4.4 $\text{MSE}_i^{\text{JD}} / \text{MSE}_i^{\text{JBD}}$ as a Function of the Model Parameters

Now that we have a closed-form expression for the expected MSE in the optimal and mismodeled scenarios, we can perform various numerical inquiries. In the following, we examine numerically the range of gain values:  $\text{MSE}_i^{\text{JD}} / \text{MSE}_i^{\text{JBD}}$ , as a function of various parameters. At this point, we do not perform an analytical examination of the gain, since the closed-form expression for (6.30) is quite prohibitive.

In each subfigure in Fig. 6.3, a new  $\mathbf{A}$  is drawn as in Sec. 6.4.2. The block-pattern  $\mathbf{m}$  is fixed, and  $Q = [5, 10, 20, 50, 100]$ .  $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$  are drawn as mentioned at the beginning of Sec. 6.4. Each marker in each plot refers to a new draw of  $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$ . That is, for fixed  $Q$  and component index (in each subfigure), only  $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$  changes. From the subfigures in Fig. 6.3 we observe that several parameters influence the gain. First, and most obvious, is that the closer the block-pattern to the one-dimensional model, the smaller the average gain, and vice versa. Second, we note again, as in Table 3.1 and Table 6.1, that the gain is considerable not only for the components with  $m_i > 1$ , but also for those with  $m_i = 1$ . Third, there is a trend that the gain in using the correct model increases with the number of domains  $Q$ . This may be explained by the conjecture that the more matrices to jointly (block-) diagonalize, the more accurate the output JBD is, and thus the difference between a diagonal and block-diagonal model has more impact.



(a) Standard deviation of the  $m = 9$  entries on the second row of  $\delta \mathbf{P}_2^{\text{JD}}$ : empirical and predicted, vs.  $n_q$ .



(b)  $\log_{10}$  of the entries of  $\mathbf{P}_1^{\text{JD}}$ . (c)  $\log_{10}$  of the entries of  $\mathbf{P}_2^{\text{JD}}$ . (d)  $\log_{10}$  of the entries of  $\mathbf{P}_3^{\text{JD}}$ .

Figure 6.2: Validating the theoretical expression for  $\text{Cov}(\delta \mathbf{P}_i^{\text{JD}})$ , as well as our choice of  $n_q$  for the small-errors regime. In Fig. 6.2(a), the ‘ $\times$ ’ symbol denotes the empirical values, averaged over 500 Monte-Carlo trials. The straight lines denote the predicted values, under the small-error assumption. Each colour refers to one entry of the 6th row of  $\mathbf{P}_2^{\text{JD}}$ . The data is drawn from the uniform distribution, using the same model parameters as in scenario #2 in Table 6.1. Block-pattern  $\mathbf{m}=[6,2,1]$ . The three oblique projections  $\mathbf{P}_i^{\text{JD}}$  are depicted in Fig. 6.2(b)–6.2(d).

## 6.5 Discussion

In this chapter, we derived a closed-form expression for the expected MSE in the separation of multidimensional components, when the separation criterion first assigns a one-dimensional model to the components, and then assigns the one-dimensional outputs to their multidimensional groups. In Chapter 3 it was shown that for Gaussian data, the minimal mean square error (MMSE) is obtained with a separation criterion which uses the correct component dimensions. That is, JBD with the cor-

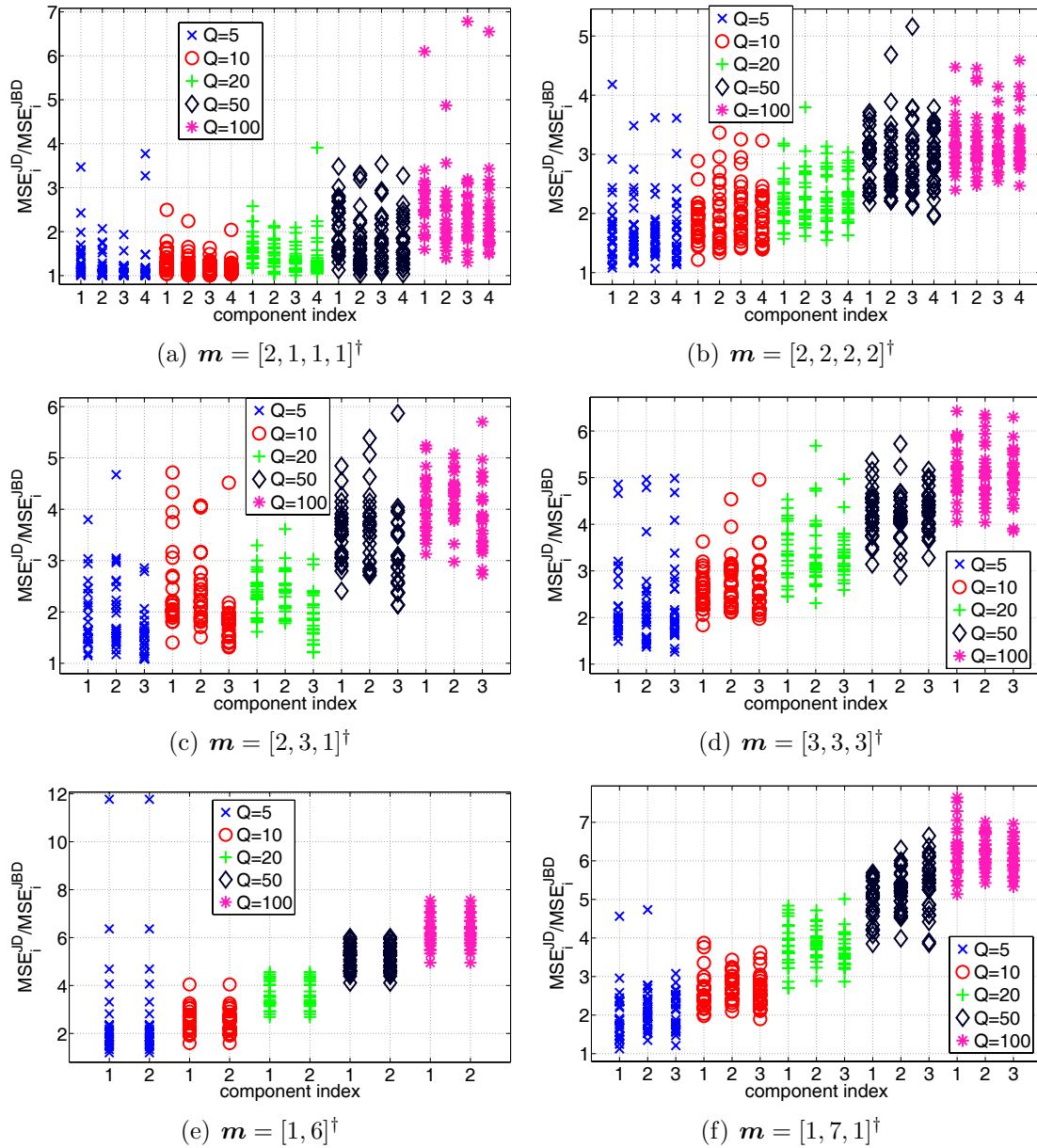


Figure 6.3: The ratio  $\text{MSE}_i^{\text{JD}} / \text{MSE}_i^{\text{JBD}}$  for different values of  $Q$ ,  $\mathbf{A}$ ,  $\mathbf{m}$  and  $\mathbf{R}_S^{(q)}$ .  $\text{MSE}_i^{\text{JBD}}$  and  $\text{MSE}_i^{\text{JD}}$  are taken from the analytical closed-form expressions (3.25) and (6.30), respectively. In each subfigure,  $\mathbf{A}$  and  $\mathbf{m}$  are fixed. A different colour and shape of a symbol represent a different number of domains  $Q$ . For each  $Q$ , 25 different values of  $\{\mathbf{R}_S^{(q)}\}_{q=1}^Q$  were drawn.  $n_q = 500$  in all subfigures.

rect block-pattern. Since the MSE which we derive in Sec. 6.3 is different than the MMSE, we conjecture that for our data model, the one-dimensional assumption followed by clustering is sub-optimal, when using second-order statistics. The numerical examples in Fig. 6.3 and in Table 6.1 support this conjecture.

In this chapter, the data was modeled as piecewise-stationary, and the separation was achieved by optimizing a non-unitary criterion, which is equivalent to JD

of a set of covariance matrices. The theoretical analysis was performed in terms of components and oblique projections, which are the unambiguous counterparts of the latent sources and the mixing matrix. The starting point of our asymptotic error analysis was the contrast function, designed for one-dimensional piecewise-stationary Gaussian data. First, we proved that asymptotically, minimizing this contrast function yields perfect separation also for multidimensional data. This property is the fundamental justification for using the one-dimensional model for separating multidimensional components. Next, we have shown that it is possible to partition the estimated component subspace into uniquely-defined rank-1 subspaces, up to internal permutation, or ordering, of these rank-1 subspaces. This property is fundamentally different from the case of the correct model, in which any partition of the component subspace into smaller subspaces is completely arbitrary. This property is the key to the error analysis, which is based on decomposing the subspaces into rank-1 factors. We then performed an error analysis, which is valid in the small-errors regime. This analysis is based only on second-order statistics. Hence, all the derived expressions are valid also for non-Gaussian data. Our theoretical results were supported by numerical experiments, in Table 6.1 and Fig. 6.2.

The closed-form MSE, derived in this chapter, allows us to calculate numerically the gain due to using the correct model against the one-dimensional model. That is, of separating multidimensional components using JBD instead of JD followed by clustering. Until now, the expected gain in performance due to using the correct multidimensional model, as opposed to the one-dimensional separation followed by clustering, could be derived only empirically, using numerical experiments, as in Table 3.1. Using both analytical closed-form expressions for the MSE, in the optimal and mismodeled procedures, we can now calculate numerically the expected gain in the MSE due to using the correct multidimensional model. By comparing these theoretical gains for several multidimensional scenarios, in Fig. 6.3, we obtain the following observations.

- The gain is larger as the block-pattern of the components diverges from one.
- The gain is equally significant for all the components, even for those of dimension one. Hence, there is benefit in using the correct multidimensional model even if a particular component of interest is one-dimensional, when the mixture contains multidimensional components.
- The gain is larger as the number of domains increases. This may be explained by the conjecture that the algorithm can better disclose the underlying block-diagonal structure with more data, i.e. more matrices to jointly

block-diagonalize.

All these observations favour the use of JBD over JD in the presence of multi-dimensional components. By comparing the empirical behaviour of parameter and component estimates under both procedures, we observe that more samples per domain are required to achieve small-error conditions for the mismodeled case than for the correct-model case. This result can be interpreted as the observation that the JD algorithm is less suitable for performing JBD, and thus for the same MSE, it requires less perturbed input. Hence, another point in favour of using JBD vs. JD. Finally, one should keep in mind that the superior performance of JBD over JD is only on average. Given the same model parameters, a specific randomization of the underlying sources may well yield the opposite result.

## 6.A First-Order Expansion of the Estimating Equations

In this appendix, we show that the first-order expansion of the estimating equations (6.33) leads to the linear relation (6.34) between the projection error terms  $\mathcal{E}_{i_r j_s}$ , defined in (6.26), and the sample error terms  $\bar{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)}$ , defined in (6.31).

Let us begin by restating that the estimates  $\{\hat{\mathbf{P}}_{i_r}\}_{i=1, r \in i}^n$  are solutions of the estimating equations (6.33). That is,

$$\langle (\hat{\mathbf{P}}_{i_r} \bar{\mathbf{R}}_X \hat{\mathbf{P}}_{i_r}^\dagger)^\# (\hat{\mathbf{P}}_{i_r} \bar{\mathbf{R}}_X \hat{\mathbf{P}}_{j_s}^\dagger) \rangle = 0, \quad i \neq j. \quad (6.40)$$

In the following, we linearize these equations with respect to the error terms due to finite sample size (3.27). Under asymptotic conditions, which are defined formally in Sec. 6.3,  $\bar{\mathbf{R}}_X^{(q)}$  converges, in the mean square, to  $\mathbf{R}_X^{(q)}$ ; for non-Gaussian components, it converges in probability. Although  $\hat{\mathbf{P}}_{i_r}$  is not an ML estimator (as opposed to  $\hat{\mathbf{P}}_i^{\text{JBD}}$ , recall Sec. 3.2.1), the estimating equations (6.33) do achieve the ML solution asymptotically (recall Sec. 6.2.2.2 and 6.2.2.3). Hence,  $\hat{\mathbf{P}}_{i_r}$  converges in probability. As for the rate of convergence, the entries of  $\delta \mathbf{R}_X^{(q)}$  are zero mean random variables with a standard deviation proportional to  $1/\sqrt{T}$ . Asymptotically, this is true also for  $\delta \mathbf{P}_{i_r}$ , since the estimate of  $\mathbf{P}_{i_r}$  is a function of a mean of  $T$  random variables. Hence, asymptotically, terms which are proportional to  $\delta \mathbf{R}_X^{(q)}$  or  $\delta \mathbf{P}_{i_r}$  are  $\Omega(\frac{1}{\sqrt{T}})$  (this notation is defined in Chapter “Notations and Conventions”) and are considered as having the same order of magnitude. The asymptotic behaviour of  $\delta \mathbf{P}_{i_r}$  is illustrated with numerical simulations in Sec. 6.4.3.

Expanding the right-hand term within the angle brackets in (6.40),

$$\begin{aligned}\widehat{\mathbf{P}}_{i_r} \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{P}}_{j_s}^\dagger &= (\mathbf{P}_{i_r} + \delta \mathbf{P}_{i_r})(\mathbf{R}_X^{(q)} + \delta \mathbf{R}_X^{(q)})(\mathbf{P}_{j_s}^\dagger + \delta \mathbf{P}_{j_s}^\dagger) \\ &= \mathbf{P}_{i_r} \overline{\mathbf{R}}_X^{(q)} \mathbf{P}_{j_s}^\dagger + \mathbf{P}_{i_r} \mathbf{R}_X^{(q)} \delta \mathbf{P}_{j_s}^\dagger + \delta \mathbf{P}_{i_r} \mathbf{R}_X^{(q)} \mathbf{P}_{j_s}^\dagger + \Omega\left(\frac{1}{T}\right).\end{aligned}\quad (6.41)$$

The first step follows from (3.27) and (6.23), the latter being the key point in the entire analysis. As for the left-hand term within the angle brackets on the LHS of (6.40),

$$\begin{aligned}(\widehat{\mathbf{P}}_{i_r} \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{P}}_{i_r}^\dagger)^\# &= (\widehat{\mathbf{A}}_{i_r} \widehat{\mathbf{B}}_{i_r} \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{B}}_{i_r}^\dagger \widehat{\mathbf{A}}_{i_r}^\dagger)^\# \\ &= \widehat{\mathbf{A}}_{i_r}^\# (\widehat{\mathbf{B}}_{i_r} \overline{\mathbf{R}}_X^{(q)} \widehat{\mathbf{B}}_{i_r}^\dagger)^{-1} \widehat{\mathbf{A}}_{i_r}^\# \\ &= \mathbf{A}_{i_r}^\# (\mathbf{B}_{i_r} \mathbf{R}_X^{(q)} \mathbf{B}_{i_r}^\dagger)^{-1} \mathbf{A}_{i_r}^\# + \Omega\left(\frac{1}{\sqrt{T}}\right) \\ &= (\mathbf{P}_{i_r} \mathbf{R}_X^{(q)} \mathbf{P}_{i_r}^\dagger)^\# + \Omega\left(\frac{1}{\sqrt{T}}\right).\end{aligned}\quad (6.42)$$

The first equality in (6.42) is due to  $\widehat{\mathbf{P}}_{i_r} = \widehat{\mathbf{A}}_{i_r} \widehat{\mathbf{B}}_{i_r}$ , which follows from (2.6). The second equality follows from (2.26). The third transition is due to the fact that  $\overline{\mathbf{R}}_X^{(q)}$  and the estimates of  $\mathbf{A}$  and  $\mathbf{B}$  converge to their mean with a standard deviation proportional to  $1/\sqrt{T}$  (the same explanation as for  $\widehat{\mathbf{P}}_i$  applies). The last step follows again from (2.26) and then (2.6).

Multiplying (6.42) with (6.41) and substituting them in (6.40),

$$\begin{aligned}\langle (\mathbf{P}_{i_r} \mathbf{R}_X \mathbf{P}_{i_r}^\dagger)^\# (\mathbf{P}_{i_r} \overline{\mathbf{R}}_X \mathbf{P}_{j_s}^\dagger + \mathbf{P}_{i_r} \mathbf{R}_X \delta \mathbf{P}_{j_s}^\dagger + \delta \mathbf{P}_{i_r} \mathbf{R}_X \mathbf{P}_{j_s}^\dagger) \rangle + \Omega\left(\frac{1}{T}\right) \\ = \langle \mathbf{R}_{X_{i_r} X_{i_r}}^\# \overline{\mathbf{R}}_{X_{i_r} X_{j_s}} \rangle + \langle \mathbf{R}_{X_{i_r} X_{i_r}}^\# \mathbf{P}_{i_r} \mathbf{R}_X \delta \mathbf{P}_{j_s}^\dagger \rangle + \langle \mathbf{R}_{X_{i_r} X_{i_r}}^\# \delta \mathbf{P}_{i_r} \mathbf{R}_X \mathbf{P}_{j_s}^\dagger \rangle \\ = \mathbf{0}_{m \times m} + \Omega\left(\frac{1}{T}\right),\end{aligned}\quad (6.43)$$

where in the first equality we used  $\mathbf{R}_{X_{i_r} X_{i_r}}^{(q)\#} = (\mathbf{P}_{i_r} \mathbf{R}_X^{(q)} \mathbf{P}_{i_r}^\dagger)^\#$ . We now start simplifying (6.43), with the aim of rewriting it using the error terms (6.26). Let us introduce the following equality,

$$\mathbf{P}_{i_r} \mathbf{R}_X^{(q)} = \sum_{k=1}^{m_i} \mathbf{R}_{X_{i_r} X_{i_k}}^{(q)} \quad \forall i \text{ and } r \in i, \quad (6.44)$$

which follows from (6.8d). Substituting (6.44) in (6.43),

$$\langle \mathbf{R}_{i_r i_r}^\# \overline{\mathbf{R}}_{i_r j_s} \rangle + \langle \mathbf{R}_{i_r i_r}^\# \left( \sum_{k=1}^{m_i} \mathbf{R}_{i_r i_k} \right) \delta \mathbf{P}_{j_s}^\dagger \rangle + \langle \mathbf{R}_{i_r i_r}^\# \delta \mathbf{P}_{i_r} \sum_{l=1}^{m_j} \mathbf{R}_{j_l j_s} \rangle = \mathbf{0}_{m \times m} + \Omega\left(\frac{1}{T}\right). \quad (6.45)$$

Note that in (6.45), for brevity,  $\mathbf{R}_X^{(q)}$  was replaced with  $\mathbf{R}^{(q)}$ . Taking the sum out of the brackets in (6.45),

$$\langle \mathbf{R}_{i_r i_r}^\# \bar{\mathbf{R}}_{i_r j_s} \rangle + \sum_{k=1}^{m_i} \langle \mathbf{R}_{i_r i_r}^\# \mathbf{R}_{i_r i_k} \delta \mathbf{P}_{j_s}^\dagger \rangle + \sum_{l=1}^{m_j} \langle \mathbf{R}_{i_r i_r}^\# \delta \mathbf{P}_{i_r} \mathbf{R}_{j_l j_s} \rangle = \mathbf{0}_{m \times m} + \Omega\left(\frac{1}{T}\right). \quad (6.46)$$

Using the fact that  $\mathbf{R}_{X_{i_r} X_{j_s}}^{(q)} = \mathbf{\Pi}_{i_r} \mathbf{R}_{X_{i_r} X_{j_s}}^{(q)} \quad \forall (i, j)$ , (6.46) can be rewritten as

$$\langle \mathbf{R}_{i_r i_r}^\# \bar{\mathbf{R}}_{i_r j_s} \rangle + \sum_{k=1}^{m_i} \langle \mathbf{R}_{i_r i_r}^\# \mathbf{R}_{i_r i_k} \mathbf{\Pi}_{i_k} \delta \mathbf{P}_{j_s}^\dagger \rangle + \sum_{l=1}^{m_j} \langle \mathbf{R}_{i_r i_r}^\# \delta \mathbf{P}_{i_r} \mathbf{\Pi}_{j_l} \mathbf{R}_{j_l j_s} \rangle = \mathbf{0}_{m \times m} + \Omega\left(\frac{1}{T}\right). \quad (6.47)$$

Since

$$\mathbf{R}_{X_{i_r} X_{i_r}}^{\#(q)} \mathbf{R}_{X_{i_r} X_{i_r}}^{(q)} = \mathbf{\Pi}_{i_r},$$

(6.47) can be rewritten as

$$\langle \mathbf{R}_{i_r i_r}^\# \bar{\mathbf{R}}_{i_r j_s} \rangle + \mathbf{\Pi}_{i_r} \delta \mathbf{P}_{j_s}^\dagger + \sum_{k \neq r}^{m_i} \langle \mathbf{R}_{i_r i_r}^\# \mathbf{R}_{i_r i_k} \mathbf{\Pi}_{i_k} \delta \mathbf{P}_{j_s}^\dagger \rangle + \sum_{l=1}^{m_j} \langle \mathbf{R}_{i_r i_r}^\# \delta \mathbf{P}_{i_r} \mathbf{\Pi}_{j_l} \mathbf{R}_{j_l j_s} \rangle = \mathbf{0}_{m \times m} + \Omega\left(\frac{1}{T}\right). \quad (6.48)$$

The error terms in (6.48) can be emphasized by rewriting them as

$$\langle \mathbf{R}_{i_r i_r}^\# \bar{\mathbf{R}}_{i_r j_s} \rangle + (\delta \mathbf{P}_{j_s} \mathbf{\Pi}_{i_r})^\dagger + \sum_{k \neq r}^{m_i} \langle \mathbf{R}_{i_r i_r}^\# \mathbf{R}_{i_r i_k} (\delta \mathbf{P}_{j_s} \mathbf{\Pi}_{i_k})^\dagger \rangle + \sum_{l=1}^{m_j} \langle \mathbf{R}_{i_r i_r}^\# (\delta \mathbf{P}_{i_r} \mathbf{\Pi}_{j_l}) \mathbf{R}_{j_l j_s} \rangle = \mathbf{0}_{m \times m} + \Omega\left(\frac{1}{T}\right). \quad (6.49)$$

Equation (6.49), when repeated for all  $r \in m_i$ ,  $s \in m_j$ , as well as changing roles between  $i$  and  $j$ , yields the  $2m_j m_i$  linear *estimating equations* for  $\delta \mathbf{P}_{i_k} \mathbf{\Pi}_{j_l}$ . Using the error notation (6.26), the estimating equations (6.49) can be rewritten as

$$\langle \mathbf{R}_{i_r i_r}^\# \bar{\mathbf{R}}_{i_r j_s} \rangle + \boldsymbol{\mathcal{E}}_{j_s, i_r}^\dagger + \sum_{k \neq r}^{m_i} \langle \mathbf{R}_{i_r i_r}^\# \mathbf{R}_{i_r i_k} \boldsymbol{\mathcal{E}}_{j_s, i_k}^\dagger \rangle + \sum_{l=1}^{m_j} \langle \mathbf{R}_{i_r i_r}^\# \boldsymbol{\mathcal{E}}_{i_r, j_l} \mathbf{R}_{j_l j_s} \rangle = \mathbf{0}_{m \times m} + \Omega\left(\frac{1}{T}\right). \quad (6.50)$$

By enumerating over all  $r \in i$  and  $s \in j$ , one obtains  $m_i m_j$  estimating equations. The other half is obtained by exchanging  $i$  with  $j$  and keeping the index  $r$  with  $i$

and  $s$  with  $j$ ,

$$\begin{aligned} \langle \mathbf{R}_{j_s j_s}^\# \bar{\mathbf{R}}_{j_s i_r} \rangle + \boldsymbol{\mathcal{E}}_{i_r, j_s}^\dagger + \sum_{l \neq s}^{m_j} \langle \mathbf{R}_{j_s j_s}^\# \mathbf{R}_{j_s j_l} \boldsymbol{\mathcal{E}}_{i_r, j_l}^\dagger \rangle + \sum_{k=1}^{m_i} \langle \mathbf{R}_{j_s j_s}^\# \boldsymbol{\mathcal{E}}_{j_s, i_k} \mathbf{R}_{i_k i_r} \rangle \\ = \mathbf{0}_{m \times m} + \Omega\left(\frac{1}{T}\right). \end{aligned} \quad (6.51)$$

(6.50) and (6.51) can be reordered and presented together as

$$\begin{aligned} -\langle \mathbf{R}_{i_r i_r}^\# \bar{\mathbf{R}}_{i_r j_s} \rangle &= \boldsymbol{\mathcal{E}}_{j_s, i_r}^\dagger + \sum_{k \neq r}^{m_i} \langle \mathbf{R}_{i_r i_r}^\# \mathbf{R}_{i_r i_k} \rangle \boldsymbol{\mathcal{E}}_{j_s, i_k}^\dagger + \sum_{l=1}^{m_j} \langle \mathbf{R}_{i_r i_r}^\# \boldsymbol{\mathcal{E}}_{i_r, j_l} \mathbf{R}_{j_l j_s} \rangle + \Omega\left(\frac{1}{T}\right) \\ -\langle \mathbf{R}_{j_s j_s}^\# \bar{\mathbf{R}}_{j_s i_r} \rangle &= \boldsymbol{\mathcal{E}}_{i_r, j_s}^\dagger + \sum_{l \neq s}^{m_j} \langle \mathbf{R}_{j_s j_s}^\# \mathbf{R}_{j_s j_l} \rangle \boldsymbol{\mathcal{E}}_{i_r, j_l}^\dagger + \sum_{k=1}^{m_i} \langle \mathbf{R}_{j_s j_s}^\# \boldsymbol{\mathcal{E}}_{j_s, i_k} \mathbf{R}_{i_k i_r} \rangle + \Omega\left(\frac{1}{T}\right), \end{aligned} \quad (6.52)$$

where  $r = 1, \dots, m_i$ ,  $s = 1, \dots, m_j$ . In order to proceed, it is convenient to vectorize the matrices in (6.52) with the  $\text{vec}\{\cdot\}$  operator, defined in Chapter ‘‘Notations and Conventions’’. Using property (A.1c) in Appendix A, (6.52) can be rewritten as

$$\begin{aligned} -\text{vec}\{\langle \mathbf{R}_{i_r i_r}^\# \bar{\mathbf{R}}_{i_r j_s} \rangle\} &= \mathcal{T} \text{vec}\{\boldsymbol{\mathcal{E}}_{j_s, i_r}\} + \sum_{k \neq r}^{m_i} (\mathbf{I} \otimes \langle \mathbf{R}_{i_r i_r}^\# \mathbf{R}_{i_r i_k} \rangle) \mathcal{T} \text{vec}\{\boldsymbol{\mathcal{E}}_{j_s, i_k}\} \\ &\quad + \sum_{l=1}^{m_j} \langle \mathbf{R}_{j_s j_l} \otimes \mathbf{R}_{i_r i_r}^\# \rangle \text{vec}\{\boldsymbol{\mathcal{E}}_{i_r, j_l}\} + \Omega\left(\frac{1}{T}\right) \\ -\text{vec}\{\langle \mathbf{R}_{j_s j_s}^\# \bar{\mathbf{R}}_{j_s i_r} \rangle\} &= \mathcal{T} \text{vec}\{\boldsymbol{\mathcal{E}}_{i_r, j_s}\} + \sum_{l \neq s}^{m_j} (\mathbf{I} \otimes \langle \mathbf{R}_{j_s j_s}^\# \mathbf{R}_{j_s j_l} \rangle) \mathcal{T} \text{vec}\{\boldsymbol{\mathcal{E}}_{i_r, j_l}\} \\ &\quad + \sum_{k=1}^{m_i} \langle \mathbf{R}_{i_r i_k} \otimes \mathbf{R}_{j_s j_s}^\# \rangle \text{vec}\{\boldsymbol{\mathcal{E}}_{j_s, i_k}\} + \Omega\left(\frac{1}{T}\right) \end{aligned}$$

where  $r \in i$ ,  $s \in j$ ,  $i \neq j$  and  $\mathcal{T}$  denotes the commutation matrix [79], defined in (A.3). Changing order in the first equation, (6.34) results.

## 6.B Closed-Form Expression for $\text{Cov}(\mathfrak{g}_{ij})$

In this appendix, we obtain a closed-form expression for the covariance matrix of the stochastic vector  $\mathfrak{g}_{ij}$ , defined in (6.36). Since, by the assumptions in Sec. 2.3.1, it has zero mean, then we can write

$$\text{Cov}(\mathfrak{g}_{ij}) = E \left\{ \mathfrak{g}_{ij} \mathfrak{g}_{ij}^\dagger \right\}. \quad (6.53)$$

In the following, we show that

$$E \left\{ \mathbf{g}_{i_r j_s} \mathbf{g}_{i_m j_l}^\dagger \right\} = \frac{1}{T} \langle \mathbf{R}_{X_{j_s} X_{j_l}} \otimes (\mathbf{R}_{X_{i_r} X_{i_r}}^\# \mathbf{R}_{X_{i_r} X_{i_m}} \mathbf{R}_{X_{i_m} X_{i_m}}^\#) \rangle, \quad (6.54a)$$

$$E \left\{ \mathbf{g}_{j_s i_r} \mathbf{g}_{j_l i_m}^\dagger \right\} = \frac{1}{T} \langle \mathbf{R}_{X_{i_r} X_{i_m}} \otimes (\mathbf{R}_{X_{j_s} X_{j_s}}^\# \mathbf{R}_{X_{j_s} X_{j_l}} \mathbf{R}_{X_{j_l} X_{j_l}}^\#) \rangle, \quad (6.54b)$$

$$E \left\{ \mathbf{g}_{i_r j_s} \mathbf{g}_{j_l i_m}^\dagger \right\} = \frac{1}{T} \langle \mathbf{R}_{X_{j_s} X_{j_l}} \mathbf{R}_{X_{j_l} X_{j_l}}^\# \otimes \mathbf{R}_{X_{i_r} X_{i_r}} \mathbf{R}_{X_{i_r} X_{i_m}} \rangle \mathcal{T}, \quad (6.54c)$$

$$E \left\{ \mathbf{g}_{j_s i_r} \mathbf{g}_{i_m j_l}^\dagger \right\} = \frac{1}{T} \langle \mathbf{R}_{X_{i_r} X_{i_m}} \mathbf{R}_{X_{i_m} X_{i_m}}^\# \otimes \mathbf{R}_{X_{j_s} X_{j_s}}^\# \mathbf{R}_{X_{j_s} X_{j_l}} \rangle \mathcal{T}, \quad (6.54d)$$

where we use the convention that  $r, m \in i, s, l \in j, i \neq j$ .

In a first step, let  $r, s, t$  and  $u$  denote the indices within the four components  $i, j, k$  and  $l$ , respectively. Then,

$$\begin{aligned} & E \left\{ \mathbf{g}_{i_r j_s} \mathbf{g}_{k_t l_u}^\dagger \right\} \quad (6.55) \\ &= E \left\{ \frac{1}{T} \sum_{q=1}^Q n_q (\mathbf{I} \otimes \mathbf{R}_{X_{i_r} X_{i_r}}^{(q)\#}) \text{vec} \{ \overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)} \} \frac{1}{T} \sum_{p=1}^Q n_p \text{vec}^\dagger \{ \overline{\mathbf{R}}_{X_{k_t} X_{l_u}}^{(p)} \} (\mathbf{I} \otimes \mathbf{R}_{X_{k_t} X_{k_t}}^{(p)\#}) \right\} \\ &= \frac{1}{T^2} \sum_{q=1}^Q \sum_{p=1}^Q n_q n_p (\mathbf{I} \otimes \mathbf{R}_{X_{i_r} X_{i_r}}^{(q)\#}) E \left\{ \text{vec} \{ \overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)} \} \text{vec}^\dagger \{ \overline{\mathbf{R}}_{X_{k_t} X_{l_u}}^{(p)} \} \right\} (\mathbf{I} \otimes \mathbf{R}_{X_{k_t} X_{k_t}}^{(p)\#}), \end{aligned}$$

where in the first step we used (A.1c) in Appendix A to rewrite (6.37) as

$$\mathbf{g}_{i_r j_s} = \langle (\mathbf{I} \otimes \mathbf{R}_{X_{i_r} X_{i_r}}^\#) \text{vec} \{ \overline{\mathbf{R}}_{X_{i_r} X_{j_s}} \} \rangle,$$

and in the last step we have left within the expectation operator only the stochastic terms. Since  $\mathbf{x}(t)$  is independent of  $\mathbf{x}(t')$  if  $t \in \mathcal{D}_q, t' \in \mathcal{D}_{q'}, q \neq q'$ , then

$$E \left\{ \text{vec} \{ \overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)} \} \text{vec}^\dagger \{ \overline{\mathbf{R}}_{X_{k_t} X_{l_u}}^{(p)} \} \right\} = \delta_{pq} E \left\{ \text{vec} \{ \overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)} \} \text{vec}^\dagger \{ \overline{\mathbf{R}}_{X_{k_t} X_{l_u}}^{(p)} \} \right\}.$$

Therefore, (6.55) can be rewritten as

$$\begin{aligned} & E \left\{ \mathbf{g}_{i_r j_s} \mathbf{g}_{k_t l_u}^\dagger \right\} = \quad (6.56) \\ & \frac{1}{T^2} \sum_{q=1}^Q n_q^2 (\mathbf{I} \otimes \mathbf{R}_{X_{i_r} X_{i_r}}^{(q)\#}) E \left\{ \text{vec} \{ \overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)} \} \text{vec}^\dagger \{ \overline{\mathbf{R}}_{X_{k_t} X_{l_u}}^{(q)} \} \right\} (\mathbf{I} \otimes \mathbf{R}_{X_{k_t} X_{k_t}}^{(q)\#}). \end{aligned}$$

We now restrict the four component indices to those relevant to (6.53).

For  $(i = k) \neq (j = l)$ , substituting (6.61) of Lemma 6.1 (given below) in (6.56)

yields

$$\begin{aligned} E \left\{ \mathbf{g}_{i_r j_s} \mathbf{g}_{i_m j_l}^\dagger \right\} &= \frac{1}{T^2} \sum_{q=1}^Q n_q^2 (\mathbf{I} \otimes \mathbf{R}_{X_{i_r} X_{i_r}}^{(q)\#}) \frac{1}{n_q} (\mathbf{R}_{X_{j_s} X_{j_l}}^{(q)} \otimes \mathbf{R}_{X_{i_r} X_{i_m}}^{(q)}) (\mathbf{I} \otimes \mathbf{R}_{X_{i_m} X_{i_m}}^{(q)\#}) \\ &= \frac{1}{T^2} \sum_{q=1}^Q n_q^2 \frac{1}{n_q} (\mathbf{R}_{X_{j_s} X_{j_l}}^{(q)} \otimes (\mathbf{R}_{X_{i_r} X_{i_r}}^{(q)\#} \mathbf{R}_{X_{i_r} X_{i_m}}^{(q)} \mathbf{R}_{X_{i_m} X_{i_m}}^{(q)\#})), \end{aligned} \quad (6.57)$$

which leads to (6.54a). The last step in (6.57) is due to (A.1a). By exchanging  $i \leftrightarrow j$  and their respective sub-indices, (6.54a) rewrites as (6.54b).

For  $(i = l) \neq (j = k)$ , we need to simplify  $E \left\{ \text{vec}\{\overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)}\} \text{vec}^\dagger\{\overline{\mathbf{R}}_{X_{j_l} X_{i_m}}^{(q)}\} \right\}$ . We first use the commutation matrix  $\mathcal{T}$ , defined in (A.3), to write

$$\text{vec}^\dagger\{\overline{\mathbf{R}}_{X_{j_l} X_{i_m}}^{(q)}\} = (\mathcal{T} \text{vec}\{\overline{\mathbf{R}}_{X_{i_m} X_{j_l}}^{(q)}\})^\dagger = \text{vec}^\dagger\{\overline{\mathbf{R}}_{X_{i_m} X_{j_l}}^{(q)}\} \mathcal{T}. \quad (6.58)$$

Then,

$$\begin{aligned} E \left\{ \text{vec}\{\overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)}\} \text{vec}^\dagger\{\overline{\mathbf{R}}_{X_{j_l} X_{i_m}}^{(q)}\} \right\} &= E \left\{ \text{vec}\{\overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)}\} \text{vec}^\dagger\{\overline{\mathbf{R}}_{X_{i_m} X_{j_l}}^{(q)}\} \right\} \mathcal{T} \\ &= \frac{1}{n_q} (\mathbf{R}_{X_{j_s} X_{j_l}}^{(q)} \otimes \mathbf{R}_{X_{i_r} X_{i_m}}^{(q)}) \mathcal{T}, \end{aligned} \quad (6.59)$$

where the first equality in (6.59) is due to (6.58) and the second follows from Lemma 6.1. Back to (6.56) with  $(i = l) \neq (j = k)$ , we now have

$$\begin{aligned} E \left\{ \mathbf{g}_{i_r j_s} \mathbf{g}_{j_l i_m}^\dagger \right\} &= \frac{1}{T^2} \sum_{q=1}^Q n_q^2 (\mathbf{I} \otimes \mathbf{R}_{X_{i_r} X_{i_r}}^{(q)\#}) \frac{1}{n_q} (\mathbf{R}_{X_{j_s} X_{j_l}}^{(q)} \otimes \mathbf{R}_{X_{i_r} X_{i_m}}^{(q)}) \mathcal{T} (\mathbf{I} \otimes \mathbf{R}_{X_{j_l} X_{j_l}}^{(q)\#}) \\ &= \frac{1}{T^2} \sum_{q=1}^Q n_q^2 \frac{1}{n_q} (\mathbf{I} \otimes \mathbf{R}_{X_{i_r} X_{i_r}}^{(q)\#}) (\mathbf{R}_{X_{j_s} X_{j_l}}^{(q)} \otimes \mathbf{R}_{X_{i_r} X_{i_m}}^{(q)}) (\mathbf{R}_{X_{j_l} X_{j_l}}^{(q)\#} \otimes \mathbf{I}) \mathcal{T} \\ &= \frac{1}{T} \langle \mathbf{R}_{X_{j_s} X_{j_l}} \mathbf{R}_{X_{j_l} X_{j_l}}^\# \otimes \mathbf{R}_{X_{i_r} X_{i_r}}^\# \mathbf{R}_{X_{i_r} X_{i_m}} \rangle \mathcal{T}, \end{aligned} \quad (6.60)$$

where in the second equality of (6.60) we have used  $\mathcal{T}(\mathbf{I} \otimes \mathbf{R}_{X_{j_l} X_{j_l}}^{(q)\#}) = (\mathbf{R}_{X_{j_l} X_{j_l}}^{(q)\#} \otimes \mathbf{I}) \mathcal{T}$ , which follows from (A.4b). The last equality of (6.60) follows from (A.1a) in Appendix A and then (2.15). This yields (6.54c). By change of appropriate indices, (6.54d) results.

**Lemma 6.1.**

$$E \left\{ \text{vec}\{\overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)}\} \text{vec}^\dagger\{\overline{\mathbf{R}}_{X_{i_m} X_{j_l}}^{(q)}\} \right\} = \frac{1}{n_q} \mathbf{R}_{X_{j_s} X_{j_l}}^{(q)} \otimes \mathbf{R}_{X_{i_r} X_{i_m}}^{(q)}. \quad (6.61)$$

*Proof.* By (6.31),

$$\begin{aligned} E \left\{ \text{vec} \{ \overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)} \} \text{vec}^\dagger \{ \overline{\mathbf{R}}_{X_{i_m} X_{j_l}}^{(q)} \} \right\} \\ = \frac{1}{n_q^2} \sum_{\substack{t \in \mathcal{D}_q \\ r \in \mathcal{D}_q}} E \left\{ \text{vec} \{ \mathbf{x}_{i_r}(t) \mathbf{x}_{j_s}^\dagger(t) \} \text{vec}^\dagger \{ \mathbf{x}_{i_m}(r) \mathbf{x}_{j_l}^\dagger(r) \} \right\}. \end{aligned}$$

From Property A.2 in Appendix A,

$$\text{vec} \{ \mathbf{x}_{i_r}(t) \mathbf{x}_{j_s}^\dagger(t) \} \text{vec}^\dagger \{ \mathbf{x}_{i_m}(r) \mathbf{x}_{j_l}^\dagger(r) \} = \mathbf{x}_{j_s}(t) \mathbf{x}_{j_l}^\dagger(r) \otimes \mathbf{x}_{i_r}(t) \mathbf{x}_{i_m}^\dagger(r).$$

Therefore,

$$\begin{aligned} E \left\{ \text{vec} \{ \overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)} \} \text{vec}^\dagger \{ \overline{\mathbf{R}}_{X_{i_m} X_{j_l}}^{(q)} \} \right\} \\ = \frac{1}{n_q^2} \sum_{\substack{t \in \mathcal{D}_q \\ r \in \mathcal{D}_q}} E \left\{ \mathbf{x}_{j_s}(t) \mathbf{x}_{j_l}^\dagger(r) \otimes \mathbf{x}_{i_r}(t) \mathbf{x}_{i_m}^\dagger(r) \right\}. \quad (6.62) \end{aligned}$$

Since  $\mathbf{x}_{i_r}(t), \mathbf{x}_{j_s}(t)$  are uncorrelated for  $i \neq j$ ,

$$E \left\{ \mathbf{x}_{j_s}(t) \mathbf{x}_{j_l}^\dagger(r) \otimes \mathbf{x}_{i_r}(t) \mathbf{x}_{i_m}^\dagger(r) \right\} = E \left\{ \mathbf{x}_{j_s}(t) \mathbf{x}_{j_l}^\dagger(r) \right\} \otimes E \left\{ \mathbf{x}_{i_r}(t) \mathbf{x}_{i_m}^\dagger(r) \right\}.$$

If we assume that component samples are uncorrelated also within each domain, i.e. for all  $t \neq r \in \mathcal{D}_q$ , such that

$$\begin{aligned} E \left\{ \mathbf{x}_{j_s}(t) \mathbf{x}_{j_l}^\dagger(r) \right\} &= \mathbf{R}_{X_{j_s} X_{j_l}} \delta_{rt}, \\ E \left\{ \mathbf{x}_{i_r}(t) \mathbf{x}_{i_m}^\dagger(r) \right\} &= \mathbf{R}_{X_{i_r} X_{i_m}}^{(q)} \delta_{rt}, \end{aligned}$$

then

$$E \left\{ \mathbf{x}_{j_s}(t) \mathbf{x}_{j_l}^\dagger(r) \otimes \mathbf{x}_{i_r}(t) \mathbf{x}_{i_m}^\dagger(r) \right\} = (\mathbf{R}_{X_{j_s} X_{j_l}}^{(q)} \otimes \mathbf{R}_{X_{i_r} X_{i_m}}^{(q)}) \delta_{rt}.$$

Hence, (6.62) is simplified into

$$\begin{aligned} E \left\{ \text{vec} \{ \overline{\mathbf{R}}_{X_{i_r} X_{j_s}}^{(q)} \} \text{vec}^\dagger \{ \overline{\mathbf{R}}_{X_{i_m} X_{j_l}}^{(q)} \} \right\} &= \frac{1}{n_q^2} \sum_{\substack{t \in \mathcal{D}_q \\ r \in \mathcal{D}_q}} (\mathbf{R}_{X_{j_s} X_{j_l}}^{(q)} \otimes \mathbf{R}_{X_{i_r} X_{i_m}}^{(q)}) \delta_{rt} \\ &= \frac{1}{n_q} \mathbf{R}_{X_{j_s} X_{j_l}}^{(q)} \otimes \mathbf{R}_{X_{i_r} X_{i_m}}^{(q)}, \end{aligned}$$

which is the desired result (6.61). □

## 6.C Explicit Form of $\mathcal{H}$ and $\text{Cov}(\mathfrak{g}_{ij})$

The  $m \times m$  block matrices which compose  $\mathcal{H}$  are defined by (6.34). The  $m \times m$  block matrices which compose  $\text{Cov}(\mathfrak{g}_{ij})$  are defined in Sec. 6.B. Each of these matrices is composed of  $2m_i m_j \times 2m_i m_j$  such blocks. When  $m_i = m_j = 1$ , these matrices coincide with  $\mathcal{H}$  (3.10) and  $\mathcal{H}_{\Pi}$  (3.15). The simplest case where our analysis is not trivial is  $m_i = 1$ ,  $m_j = 2$ . For this case, these two matrices are given in explicit form in (6.63) and (6.64) on page 98. For larger dimensions, the number of blocks increases, but the blocks have essentially the same structure as those shown in these equations.

$$\mathcal{H} = \left[ \begin{array}{c|c|c} \langle \mathbf{R}_{j_1 j_1} \otimes \mathbf{R}_{ii}^\# \rangle & \mathcal{T} & \langle \mathbf{R}_{j_1 j_2} \otimes \mathbf{R}_{ii}^\# \rangle \\ \mathcal{T} & \langle \mathbf{R}_{ii} \otimes \mathbf{R}_{j_1 j_1}^\# \rangle & (\mathbf{I} \otimes \langle \mathbf{R}_{j_1 j_1}^\# \mathbf{R}_{j_1 j_2} \rangle) \mathcal{T} \\ \hline \langle \mathbf{R}_{j_2 j_1} \otimes \mathbf{R}_{ii}^\# \rangle & \mathbf{0}_{m^2 \times m^2} & \langle \mathbf{R}_{j_2 j_2} \otimes \mathbf{R}_{ii}^\# \rangle \\ (\mathbf{I} \otimes \langle \mathbf{R}_{j_2 j_2}^\# \mathbf{R}_{j_2 j_1} \rangle) \mathcal{T} & \mathbf{0}_{m^2 \times m^2} & \mathcal{T} \\ \hline & & \langle \mathbf{R}_{ii} \otimes \mathbf{R}_{j_2 j_2}^\# \rangle \end{array} \right] \quad (6.63)$$

$$\text{Cov}(\mathbf{g}_{ij}) = \left[ \begin{array}{c|c|c} \langle \mathbf{R}_{j_1 j_1} \otimes \mathbf{R}_{ii}^\# \rangle & (\mathbf{\Pi}_{j_1} \otimes \mathbf{\Pi}_i) \mathcal{T} & \langle \mathbf{R}_{j_1 j_2} \otimes \mathbf{R}_{ii}^\# \rangle \\ (\mathbf{\Pi}_i \otimes \mathbf{\Pi}_{j_1}) \mathcal{T} & \langle \mathbf{R}_{ii} \otimes \mathbf{R}_{j_1 j_1}^\# \rangle & (\mathbf{\Pi}_i \otimes \langle \mathbf{R}_{j_1 j_1}^\# \mathbf{R}_{j_1 j_2} \rangle) \mathcal{T} \\ \hline \langle \mathbf{R}_{j_2 j_1} \otimes \mathbf{R}_{ii}^\# \rangle & ((\mathbf{R}_{j_2 j_1} \mathbf{R}_{j_1 j_1}^\#) \otimes \mathbf{\Pi}_i) \mathcal{T} & \langle \mathbf{R}_{j_2 j_2} \otimes \mathbf{R}_{ii}^\# \rangle \\ (\mathbf{\Pi}_i \otimes \langle \mathbf{R}_{j_2 j_2}^\# \mathbf{R}_{j_2 j_1} \rangle) \mathcal{T} & \langle \mathbf{R}_{ii} \otimes \mathbf{R}_{j_2 j_2}^\# \mathbf{R}_{j_2 j_1} \rangle & (\mathbf{\Pi}_i \otimes \mathbf{\Pi}_{j_2}) \mathcal{T} \\ \hline & & \langle \mathbf{R}_{ii} \otimes \mathbf{R}_{j_2 j_2}^\# \rangle \end{array} \right] \quad (6.64)$$

# Chapter 7

## Overdetermined Multidimensional ICA: Application to CMB Observations

### 7.1 Outline

This chapter deals with applying our theoretical analysis and algorithms to realistic data. Namely, observations of the cosmic microwave background radiation (CMB). Sec. 7.2.1 provides the necessary background and mathematical tools specific to this application. It gives the astrophysical background of CMB research, and how it is related to multidimensional ICA (MICA). Sec. 7.2.2 describes spherical harmonics and the spherical harmonic transform (SHT). Sec. 7.2.3 presents the concept of “cosmic variance”: it is an inherent error in CMB estimation, due to the fact that only one image of the sky at CMB observation frequencies is available. Sec. 7.3 presents the motivation for our analysis from the perspective of cosmological research; the motivation from the more theoretical MICA perspective has been presented in Chapter 1. Sec. 7.4 discusses various aspects in applying our analysis to the astrophysical data. Sec. 7.4.1 discusses the required notational adaptations to the overdetermined data. In Sec. 7.4.2 we define the spectral density matrices, which are the spectral-domain counterparts of the covariance matrices in the piecewise-stationary model of Sec. 2.3.1. The cosmological processes which constitute our components are described in terms of our model in Sec. 7.4.3. In Sec. 7.4.4 we describe a dimension reduction scheme, to be applied on the overdetermined data. In CMB observations, the intensity and spectral density of the CMB vary significantly at each observation frequency and at each angular frequency, both absolutely and with respect to the other components. Therefore, in Sec. 7.4.5 we factorize the closed-form expression

for the mean square error (MSE) to express the MSE as a function of the observation frequency and angular frequency. Now that all the necessary adaptations are ready, we can proceed to applying our analysis to the astrophysical data, in Sec. 7.5. We begin by describing the specific data used in our simulations, which consists of simulated observations of CMB and Galactic emissions, created by the Planck Sky Model (PSM). In Sec. 7.5.1 we discuss methods to determine the effective dimension of the Galactic emission component. This dimension is not known in advance. In Sec. 7.5.2 we validate our claim that the Galactic emission component is indeed composed of several dependent components, and not of independent components. This is done by comparing the output of joint block diagonalization (JBD) with that of joint diagonalization (JD) followed by clustering. Finally, in Sec. 7.5.3, we apply our theoretical closed-form expression for the MSE to the simulated CMB observations. We show that for the two final candidate dimensions of the Galactic emission component, obtained in Sec. 7.5.1, we achieve good match between the theoretical closed-form expression and the empirical MSE. We discuss our results in Sec. 7.6.

Part of the analysis and results in this chapter has been presented in Lahat et al. [75].

## 7.2 Preliminaries

### 7.2.1 Astrophysical Background

According to the cosmological model known as the “Big Bang”, the early Universe began in a highly uniform, hot, dense, and compact state. Subsequently, it expanded and cooled down. This cooling and expansion continue to the present day [87]. According to cosmological theory [88, 89], the Universe today should be filled with radiation that is literally the remnant light left over from these early processes. This radiation was discovered and identified in 1965 [90, 91]. The electromagnetic spectrum of this radiation peaks in the microwave range. Hence, it is denoted as the *cosmic microwave background* radiation. One of the most prominent features of the CMB is its uniformity, or isotropy, across the sky. Only with very sensitive instruments, fluctuations in the CMB temperature could be detected [92]. These departures from isotropy are denoted as *anisotropies*. The statistical distribution of the CMB anisotropies depends on cosmological parameters [89, 93]. Hence, mapping the CMB correctly yields key quantitative information about the formation of our Universe and its structure [93].

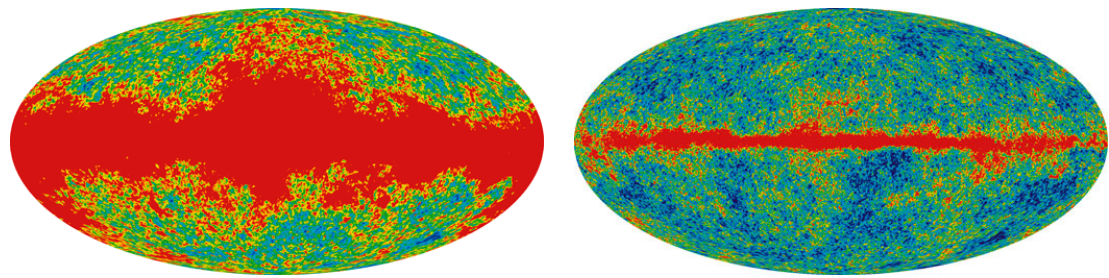
A multitude of physical processes contribute to today’s sky at the same range of frequencies in which the CMB is observed. Therefore, at each detector, the received

image is a superposition of different astrophysical components, in addition to the CMB. These processes were generated at various cosmological epochs, millions of years after the CMB has been formed. Since they are closer to us in time, as well as in space, they are denoted as *foregrounds*. The main foregrounds are [94, 95] (1) Galactic foregrounds, from our own galaxy, the Milky Way: synchrotron, free-free, and dust emission, (2) extra-galactic sources, such as the thermal and kinematic Sunyaev-Zeldovich (SZ) effect [96, 97] and point sources, e.g. distant radio and far-infrared sources, and (3) detector noise. A few hundreds of these point sources are strong enough such that they should be masked out and discarded prior to undertaking any CMB analysis [98, 95].

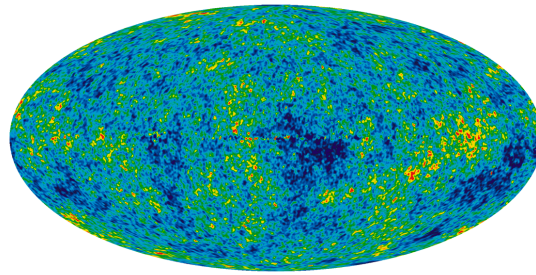
A fundamental property of the CMB and all foregrounds, except detector noise, is that they vary very slowly in time and space. For all practical experimental purposes, there exists only one image of the sky at each frequency. In order to reduce the effective noise variance, the same region of the sky is scanned several times by the observation instruments. However, there is no practical way to “average out” any of the other foregrounds, neither by extending the observation time nor by changing the position of the instruments in space. The fact that we have only one realization of the sky determines the variance of the estimate of the CMB angular power spectrum (this term shall be explained in Sec. 7.2.2). This variance is termed “cosmic variance”, and will be discussed in Sec. 7.2.3.

Figs. 7.1(a) and 7.1(b) demonstrate the sky measurements at two of the five microwave frequencies used for CMB observation by the Wilkinson microwave anisotropy probe (WMAP) mission [99]. A full-sky CMB-only map based on seven-year WMAP measurements [99] is given in Fig. 7.1(c). WMAP is a NASA Explorer mission launched June 2001, and whose data collection has recently ended. The contamination of the CMB by foregrounds, especially due to our Galaxy, is obvious. The minimum in Galactic foregrounds and the clearest window on CMB fluctuations occurs near 70GHz [100, Chapter 1].

A variety of novel methods and models have been developed to extract the CMB from its observations. The internal linear combination (ILC) method [98, Sec. 5.2],[102], widely used for WMAP data analysis [103], can obtain highly accurate component separation. However, ILC is known to be sensitive to calibration errors [104, 105]. Therefore, other methods which do not assume perfect knowledge of specific astrophysical quantities have been recently explored. In particular, blind or semi-blind source separation and independent component analysis (ICA). For example, [106, 107, 108]. These methods are motivated by the widely accepted assumption [93, Sec. 6], that the generating mechanism of the CMB is stochastic in



(a) Full-sky at 23GHz, showing significant foreground contamination at the Galactic axis. (b) Full-sky at 94GHz, showing foreground contamination at the Galactic axis.



(c) Processed map showing CMB only, after removal of foregrounds.

Figure 7.1: Temperature fluctuations over the full sky, based on seven years of WMAP data. 23GHz is the lowest detector frequency of the WMAP mission, 94GHz is the highest. The CMB map in Fig. 7.1(c) was created by processing data from all five WMAP channels. Temperature fluctuations are shown as color differences. Temperature range is  $\pm 200 \mu\text{Kelvin}$ . The sphere is projected onto the plane using the Mollweide projection in Galactic coordinates. Images source: WMAP Science Team [101]

nature. That is, the observed CMB is the single realization of a stationary Gaussian process on the sphere [93, Sec. 6]. In terms of the analysis presented in this thesis, the CMB can also be regarded as a one-dimensional component, which is statistically independent of all other foregrounds, at least to a high level of approximation.

The Galactic foregrounds, however, demonstrate significant statistical dependence between each other. This dependence is due to the fact that matter in our Galaxy emits radiation through various processes: the more matter, the stronger the emission of all processes. Therefore, the Galactic components tend to follow the same spatial Galactic pattern. In addition, a completely stochastic model of the Galactic foregrounds is currently unavailable [95]. Therefore, more recently, blind and semi-blind methods which are based on dependent component analysis (as opposed to classical ICA) were suggested by [109, 13, 110, 16, 14] and others.

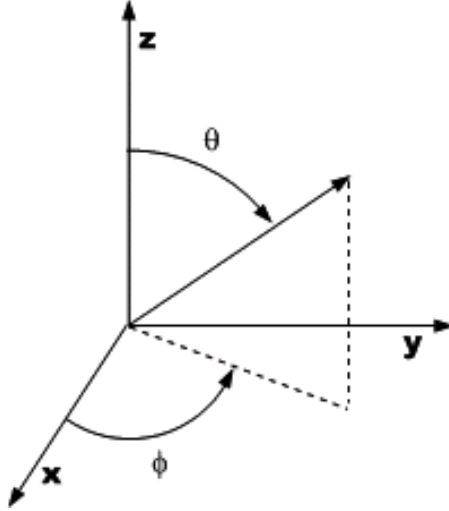


Figure 7.2: Spherical coordinates:  $\theta \in [0, \pi)$  is the polar angle and  $\phi \in [0, 2\pi)$  is the azimuthal angle. On the two-dimensional surface of the sphere, the radius is fixed.

## 7.2.2 Spherical Harmonics

In this chapter, the data is the full sky, observed at several spectral windows. As explained above, the observed CMB is regarded as a single realization of a stationary Gaussian process on the sphere. In the following, we shall conduct our analysis in the spectral domain. Therefore, we introduce the concept of spectral analysis on the sphere.

Consider a function  $f(\theta, \phi) \in L_2$  on the surface of the sphere, where the spherical coordinates  $\theta \in [0, \pi)$  and  $\phi \in [0, 2\pi)$  are illustrated in Fig. 7.2. Then,  $f(\theta, \phi)$  can be expanded as

$$f(\theta, \phi) = \sum_{\ell \geq 0} \sum_{m=-\ell}^{\ell} Y_{\ell m}(\theta, \phi) a_{\ell m}, \quad (7.1)$$

where  $a_{\ell m}$ ,  $l \geq 0$ ,  $m = \{-\ell, \dots, \ell\}$ , are the spherical harmonic coefficients and  $Y_{\ell m}(\theta, \phi)$  are the spherical harmonics.  $Y_{\ell m}(\theta, \phi)$ ,  $l \geq 0$ ,  $m = \{-\ell, \dots, \ell\}$  form a complete set of orthonormal basis functions. The inverse transform is

$$a_{\ell m} = \int_0^\pi \int_0^{2\pi} f(\theta, \phi) Y_{\ell m}^*(\theta, \phi) d\phi \sin \theta d\theta. \quad (7.2)$$

Let us now consider the case that  $f(\theta, \phi)$  is a zero-mean, real-valued stationary random field on the sphere. Due to the orthonormality of the spherical harmonics,

$$E \{a_{\ell m} a_{\ell' m'}^*\} = \delta_{\ell \ell'} \delta_{m m'} c_\ell, \quad (7.3)$$

where (7.3) defines the *angular power spectrum*  $c_\ell$  of the field. Index  $\ell$  is called

the *angular frequency* or multipole moment, and index  $m$  denotes the phase. The empirical angular power spectrum of the field is estimated by

$$\widehat{c}_\ell \triangleq \frac{1}{2\ell + 1} \sum_{m=-\ell}^{\ell} |a_{\ell m}|^2 \quad (7.4)$$

such that

$$c_\ell = E \{ \widehat{c}_\ell \} . \quad (7.5)$$

Since the SHT (7.2) is a linear operation, then if  $f(\theta, \phi)$  is Gaussian, the  $a_{\ell m}$  are complex Gaussian with variance  $E \{ |a_{\ell m}|^2 \} = c_\ell$ .

We have thus described a linear transform through which the real stationary Gaussian random field on the sphere, which represents the CMB, can be transformed into a sequence of statistically independent complex Gaussian coefficients  $a_{\ell m}$ , whose variance is the angular power spectrum  $c_\ell$  of the CMB. The angular power spectrum of the CMB, based on WMAP observations, is illustrated in Fig. 7.3.

### 7.2.3 Cosmic Variance

Equations (7.3) and (7.4) imply, for Gaussian stationary data, that every  $\widehat{c}_\ell$  is in fact an average of  $2\ell + 1$  statistically independent terms. Since there exists, for all practical purposes, only one realization of the whole sky, the variance of each  $\widehat{c}_\ell$  is  $\frac{2}{2\ell+1} c_\ell^2$ . This variance is due to the fact that  $\widehat{c}_\ell \sim \chi_{2\ell+1}^2$ . That is, an inherent finite sample-size error due to the unique realization of the sky on one hand, and the finite number of the SHT  $m$ -modes, on the other. This property is called “cosmic variance”. By its nature, it is more significant for lower  $\ell$ . It is one of the unique features of CMB analysis, compared with other applications. The effect of cosmic variance on the angular power spectrum of the CMB is illustrated in Fig. 7.3.

A more detailed discussion of the relationship between CMB, SHT and cosmic variance can be found, for example, in [112, 113, 114].

## 7.3 Motivation

The European space Agency’s (ESA) satellite Planck [100], launched in 2009, is designed to observe the CMB over the whole sky with unprecedented frequency coverage, angular resolution, and sensitivity [115]. It is expected that the limit of the performance of CMB extraction methods will eventually stem from cosmological restrictions, rather than by instrumental limitations [95]. Therefore, there is high

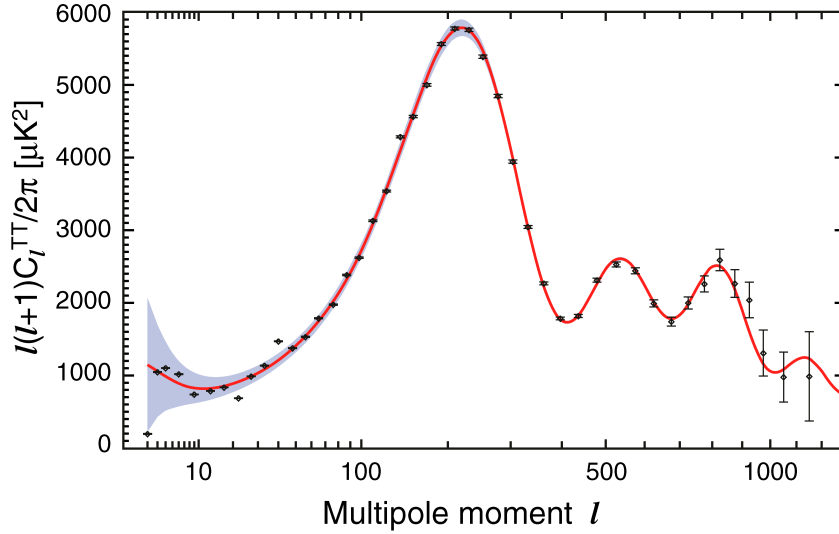


Figure 7.3: CMB angular power spectrum, based on seven years of WMAP measurements. The red curve is the  $\Lambda$ CDM cosmological model best fit to the 7-year WMAP data. The error bars denote instrument noise. The shaded region indicates the uncertainty in the model spectrum arising from cosmic variance. The Y-axis is rescaled by  $\frac{\ell(\ell+1)}{2\pi}$  for clarity of presentation. Image source: [111]

interest in being able to predict the best achievable performance that component extraction methods can provide.

In this chapter, we present a preliminary study which shows that the framework of second-order multidimensional ICA may also be appropriate for the analysis of CMB observations. This is obtained by demonstrating that our JBD algorithm is capable of separating CMB and Galactic emission components from their mixtures.

From the theoretical aspect, until now, the performance analysis of existing methods relied on processing simulated or real sky maps. In this work, we propose to use a closed-form analytical expression for the expected error, which takes into account the stochastic nature of the CMB component.

## 7.4 Application to Sky Observation

In this section, we discuss the required adaptations of the theoretical model and analysis of previous chapters to the astrophysical data.

### 7.4.1 Overdetermined Data Model

The input data for the analysis in this chapter is simulated sky observations at Planck's  $N_d = 9$  channel frequencies, in the range 30–857GHz. These maps are similar to those in Fig. 7.1(a) and 7.1(b). As a first simplifying step, we assume

that all maps are taken at the same resolution. That is, all channels are band-limited at  $0 \leq \ell \leq T$ . The data in each map is collected, after being transformed with the SHT, into the  $f$ th entry of the  $N_d \times 1$  vector  $\mathbf{x}_{\ell m}$ . That is,  $\mathbf{x}_{\ell m}$  are the complex-valued SHT coefficients of the sky maps on the sphere. We assume that these maps can be modeled as a linear instantaneous sum of  $N_c$  components, such that

$$\mathbf{x}_{\ell m} = \sum_{i=1}^{N_c} \mathbf{x}_{i,\ell m}.$$

The  $m_i$ -dimensional components  $\mathbf{x}_{i,\ell m}$  stand for CMB, Galactic emission, etc. For each component we consider a latent model

$$\mathbf{x}_{i,\ell m} = \mathbf{A}_i \mathbf{s}_{i,\ell m},$$

where  $\mathbf{s}_{i,\ell m}$  is an  $m_i \times 1$  vector, and

$$N_s = \sum_{i=1}^{N_c} m_i$$

denotes the dimension of the signal space.  $\mathbf{A}_i$  is a full column-rank  $N_d \times m_i$  matrix, such that

$$\mathbf{A} = [\mathbf{A}_1, \dots, \mathbf{A}_{N_c}]$$

is an  $N_d \times N_s$  full column-rank matrix. Accordingly,  $\mathbf{s}_{\ell m} = [\mathbf{s}_{1,\ell m}^\dagger, \dots, \mathbf{s}_{N_c,\ell m}^\dagger]^\dagger$  is an  $N_s \times 1$  vector, such that the observations can be rewritten as

$$\mathbf{x}_{\ell m} = \mathbf{A} \mathbf{s}_{\ell m}.$$

In the previous chapters, we had  $m = N_s = N_d$ . We now allow  $N_s \leq N_d$ . That is, an overdetermined model. This notation shall be used in Sec. 7.4.4.

## 7.4.2 Derivation of the Spectral Density Matrices

As explained in Sec. 7.2.2, we process our data in the spherical harmonics domain. Therefore, the matrices required for our analysis are constructed in a way which is slightly different than that in Sec. 2.3.1.

In practice, when processing CMB observations, it is common to mask out certain regions of the sky where the foregrounds are too bright for CMB analysis [103], for example at the Galactic plane. The fraction of the sky which is not masked out is denoted by  $0 \leq f_{\text{sky}} \leq 1$ . In this case, a good approximation to the increase in

variance is to divide it by  $f_{\text{sky}}$  [116].

The empirical spectral density matrix of the observations at multipole  $\ell$  is thus given by

$$\overline{\mathbf{R}}_X^{(\ell)} = \frac{1}{2\ell + 1} \sum_{m=-\ell}^{\ell} \mathbf{x}_{\ell m} \mathbf{x}_{\ell m}^H \cdot \frac{1}{f_{\text{sky}}}. \quad (7.6)$$

When the field is real, as is the case with the temperature measurements over the sky,

$$a_{\ell, -m} = (-1)^m a_{\ell m}^* \quad m \geq 0 \quad (7.7)$$

so that  $\overline{\mathbf{R}}_X^{(\ell)}$  is real. Similarly,

$$\overline{\mathbf{R}}_{X_i X_j}^{(\ell)} = \frac{1}{2\ell + 1} \sum_{m=-\ell}^{\ell} \mathbf{x}_{i, \ell m} \mathbf{x}_{j, \ell m}^H \cdot \frac{1}{f_{\text{sky}}}. \quad (7.8)$$

Let us now define a domain  $q$ , for which  $\ell_{\min}(q) \leq \ell \leq \ell_{\max}(q)$ . Then, in analogy to (2.9),

$$\begin{aligned} \overline{\mathbf{R}}_X^{(q)} &= \frac{1}{\sum_{\ell=\ell_{\min}(q)}^{\ell_{\max}(q)} (2\ell + 1)} \sum_{\ell=\ell_{\min}(q)}^{\ell_{\max}(q)} \frac{1}{2\ell + 1} \sum_{m=-\ell}^{\ell} \mathbf{x}_{\ell m} \mathbf{x}_{\ell m}^H \cdot \frac{1}{f_{\text{sky}}} \\ &= \frac{1}{n_q} \sum_{\ell \in \mathcal{D}_q} \overline{\mathbf{R}}_X^{(\ell)}, \end{aligned} \quad (7.9)$$

where now

$$n_q = f_{\text{sky}} \sum_{\ell=\ell_{\min}(q)}^{\ell_{\max}(q)} (2\ell + 1). \quad (7.10)$$

Note that in this case,  $n_q$  is not necessarily an integer. In analogy to (3.5),

$$\overline{\mathbf{R}}_{X_i X_j}^{(q)} = \frac{1}{n_q} \sum_{\ell \in \mathcal{D}_q} \overline{\mathbf{R}}_{X_i X_j}^{(\ell)}. \quad (7.11)$$

Following (7.5) we define the expected spectral density matrix as

$$\mathbf{R}_X^{(q)} = \frac{1}{n_q} \sum_{\ell \in \mathcal{D}_q} E \left\{ \overline{\mathbf{R}}_X^{(\ell)} \right\}, \quad (7.12)$$

and

$$\mathbf{R}_{X_i X_j}^{(q)} = \frac{1}{n_q} \sum_{\ell \in \mathcal{D}_q} E \left\{ \overline{\mathbf{R}}_{X_i X_j}^{(\ell)} \right\}. \quad (7.13)$$

The above definitions are used, for example, in [109, Sec. 2.1]. It should be remarked

that the angular power spectrum is well defined only for stationary or isotropic random fields on the sphere. We have thus obtained the real-valued (due to (7.7)) spectral density matrices required for our analysis.

### 7.4.3 Cosmological Processes as Multidimensional Components

Following cosmological theory [93, Sec. 6], we model the CMB as a Gaussian stationary random field on the sphere whose angular power spectrum  $c_\ell$  is based on the  $\Lambda$ CDM cosmological model. The  $\Lambda$ CDM model is in agreement with the latest WMAP results [93, 111], see Fig. 7.3. The emission law of the CMB anisotropies is well known, and is given, to first order, by the derivative (with respect to the temperature) of the blackbody law. Therefore, we model the CMB as an  $N_{\text{CMB}} = 1$  dimensional component. The CMB emission law at each of the  $N_d$  detector frequencies is collected into the  $N_d \times 1$  vector  $\mathbf{a}_{\text{CMB}}$ . Vector  $\mathbf{a}_{\text{CMB}}$  shall be used as the first column of the mixing matrix  $\mathbf{A}$  in the additive model representation (2.4). The spectral density matrix of the CMB component at domain  $q$  is fully determined by  $\mathbf{R}_{X_{\text{CMB}}}^{(q)} = c_q \mathbf{a}_{\text{CMB}} \mathbf{a}_{\text{CMB}}^\dagger$ .

In this thesis, we present a preliminary test of feasibility. Therefore, we content ourselves with the two most dominant components, CMB and Galactic emission. We model the Galactic emission as a combination of synchrotron, free-free and dust emissions. The Galactic emission is non-stationary and non-isotropic on the sky and therefore cannot be represented with a well-defined mixing model, nor with a well-defined power spectrum. However, we can still calculate the empirical  $\overline{\mathbf{R}}_{X_{\text{Gal}}}^{(q)}$  using (7.9). Therefore, in the following analysis, we treat the Galactic emission as a deterministic component by using only one realization thereof, whereas the CMB has a different realization in each Monte-Carlo trial. Since the CMB temperature fluctuations are zero-mean random variables uncorrelated with all other foregrounds, then (2.10) holds empirically with respect to the Galactic emission component.

### 7.4.4 Dimension Reduction

According to astrophysical observations, the Galactic emission should be modeled as a high-dimensional component ( $N_{\text{Gal}} \gg 1$ ). However, the minimization of the contrast function with a large  $N_{\text{Gal}}$  is numerically ill-conditioned, because the eigenvalues of  $\langle \overline{\mathbf{R}}_{X_{\text{Gal}}} \rangle$  vary within 10 orders of magnitude. Fortunately, it turns out that the effective dimension of the Galactic component can be safely taken as  $N_{\text{Gal}} < N_d - 1$ , for which there is no conditioning problem, as we shall show shortly. This implies

$N_s = N_{\text{CMB}} + N_{\text{Gal}} < N_d$ . Hence, an overdetermined problem.

When  $N_d = N_s$ , that is,  $\mathbf{A}$  invertible and  $\mathbf{R}_X^{(q)}$  full-rank for  $1 \leq q \leq Q$ , we can write the log-likelihood as in (2.13). Consequently, maximizing the likelihood as in Sec. 2.3.3 is equivalent to minimizing the contrast function (2.17). As explained in Sec. 2.3.3, the minimization of (2.17) can be understood as the JBD of the set of spectral density matrices  $\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$  by  $\mathbf{A}^{-1}$ . However, it is clear that the derivation of the contrast function (2.17) and thus also its minimization by the inverse of the mixing matrix cannot be obtained if  $N_s < N_d$ . We now show how the overdetermined case can still fit in into this framework.

In order to turn the overdetermined problem into a determined one, we project the  $N_d$ -dimensional observations onto the  $N_s$ -dimensional signal subspace,

$$\tilde{\mathbf{R}}_X^{(q)} = \mathbf{U}_s^\# \overline{\mathbf{R}}_X^{(q)} \mathbf{U}_s^{\#\dagger}, \quad 1 \leq q \leq Q, \quad (7.14)$$

where  $\mathbf{U}_s$  is the  $N_d \times N_s$  matrix of singular vectors which correspond to the largest  $N_s$  singular values of the SVD of  $\langle \overline{\mathbf{R}}_X \rangle$ . Therefore, the columns of  $\mathbf{U}_s$  span the subspace which holds most of the signal energy. This amounts to principal component analysis (PCA), which has already been used for multidimensional data in [9], for example. The dimension-reduced empirical spectral density matrices  $\{\tilde{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$  are now the input to the contrast function (2.17). If our choice of  $N_{\text{Gal}}$  is still too large, we shall again have a conditioning problem with the JBD algorithm. If our choice of  $N_{\text{Gal}}$  is sufficiently small to allow proper conditioning, but not too small to discard a significant part of the signal energy, then  $\{\tilde{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$  can be jointly block-diagonalized into matrices  $\{\tilde{\mathbf{R}}_S^{(q)}\}_{q=1}^Q$  with (approximate) block-pattern  $\mathbf{m} = [N_{\text{CMB}}, N_{\text{Gal}}]^\dagger$ ,

$$\tilde{\mathbf{R}}_X^{(q)} = \tilde{\mathbf{A}} \tilde{\mathbf{R}}_S^{(q)} \tilde{\mathbf{A}}^\dagger, \quad 1 \leq q \leq Q,$$

where  $\tilde{\mathbf{A}}$  is the new  $N_s \times N_s$  mixing matrix. If  $N_{\text{Gal}}$  is too small, then there will not be a conditioning problem, but obviously the output  $\tilde{\mathbf{R}}_S^{(q)}$  will have large off block-diagonal terms and thus low separation quality. With the inverse projection we expand  $\hat{\mathbf{A}} = \mathbf{U}_s \tilde{\mathbf{A}}$  back to  $N_d$  dimensions. Using this  $\hat{\mathbf{A}}$  one can reconstruct the CMB and Galactic components via  $\hat{\mathbf{P}}_i = \hat{\mathbf{A}}_i \hat{\mathbf{B}}_i$  (2.6) and then  $\hat{\mathbf{x}}_i(\theta, \phi) = \hat{\mathbf{P}}_i \mathbf{x}(\theta, \phi)$  (6.8c),  $i = \{\text{CMB}, \text{Gal}\}$ .

### 7.4.5 Mean Square Error

Until now, we have applied the MSE expression (3.25) only to data with equal weighting. In the CMB application, the weighting is different at each domain.

This is due to the term  $2\ell + 1$  in (7.4). We are also potentially interested in the contribution of each observation channel to the total MSE. Therefore, in this section, the MSE (3.25) is rewritten per each domain and per channel.

The empirical MSE in component  $i$  at detector  $f$  and domain  $q$  is defined as

$$\widehat{\text{MSE}}_i^{(f,q)} = \frac{1}{n_q} \sum_{t \in \mathcal{D}_q} |\mathbf{e}_f^\dagger (\widehat{\mathbf{x}}_i(t) - \mathbf{x}_i(t))|^2, \quad (7.15)$$

where  $\mathbf{e}_f$  is an  $N_d \times 1$  vector whose  $f$ th entry is 1 and the rest is zero. Using (3.3),

$$\begin{aligned} \widehat{\text{MSE}}_i^{(f,q)} &= \frac{1}{n_q} \sum_{t \in \mathcal{D}_q} |\mathbf{e}_f^\dagger \delta \mathbf{P}_i \mathbf{x}(t)|^2 = \mathbf{e}_f^\dagger \delta \mathbf{P}_i \overline{\mathbf{R}}_X^{(q)} (\delta \mathbf{P}_i)^\dagger \mathbf{e}_f \\ &= \text{tr} \left\{ (\overline{\mathbf{R}}_X^{(q)} \otimes \mathbf{e}_f \mathbf{e}_f^\dagger) \text{vec}\{\delta \mathbf{P}_i\} \text{vec}^\dagger\{\delta \mathbf{P}_i\} \right\}, \end{aligned} \quad (7.16)$$

where the second equality uses (2.9). The third expression follows from properties (A.1e), (A.1c) and (A.1d) in Appendix A. Note the analogy between (7.16) and (3.22). The fundamental difference is that now the identity matrix in (3.22) is replaced with the rank-1 term  $\mathbf{e}_f \mathbf{e}_f^\dagger$ . As in Sec. 3.2.4, in order to obtain the expectation of (7.16), we employ the approximation that  $\delta \mathbf{P}_i$  are statistically independent of the *total* power of the observations,  $\langle \overline{\mathbf{R}}_X \rangle$ . Hence, we can write

$$\begin{aligned} \overline{\text{MSE}}_i^{(f,q)} &\triangleq E \left\{ \widehat{\text{MSE}}_i^{(f,q)} \right\} \\ &= \text{tr} \left\{ (\mathbf{R}_X^{(q)} \otimes \mathbf{e}_f \mathbf{e}_f^\dagger) \text{Cov}(\text{vec}\{\delta \mathbf{P}_i\}) \right\}. \end{aligned} \quad (7.17)$$

Substituting  $\text{Cov}(\text{vec}\{\delta \mathbf{P}_i\})$  (3.13) in (7.17) and following the same steps as in (3.24), the total MSE in component  $i$  at detector  $f$  and domain  $q$  is

$$\begin{aligned} \text{MSE}_i^{(f,q)} &= \frac{1}{T} \sum_{j \neq i} \left( \right. & (7.18) \\ & \text{tr} \left\{ (\mathbf{R}_{X_j X_j}^{(q)} \otimes \mathbf{e}_f \mathbf{e}_f^\dagger) \left( \langle \mathbf{R}_{X_j X_j} \otimes \mathbf{R}_{X_i X_i}^\# \rangle - \langle \mathbf{R}_{X_j X_j}^\# \otimes \mathbf{R}_{X_i X_i} \rangle^\# \right) \right\} \\ & \left. + \text{tr} \left\{ (\mathbf{R}_{X_i X_i}^{(q)} \otimes \mathbf{e}_f \mathbf{e}_f^\dagger) \left( \langle \mathbf{R}_{X_i X_i} \otimes \mathbf{R}_{X_j X_j}^\# \rangle - \langle \mathbf{R}_{X_i X_i}^\# \otimes \mathbf{R}_{X_j X_j} \rangle^\# \right) \right\} \right) + O\left(\frac{1}{T^2}\right). \end{aligned}$$

Summing up (7.18) over all  $f$  and  $q$  leads back to (3.25), up to normalization. Expression (7.18) is the basis of our analysis in Sec. 7.5 .

## 7.5 Numerical Experiments

The sky simulations in this chapter are based on the PSM [16, 95, 117, 118], a software package developed by the Planck Collaboration Component Separation Working Group (WG2). This software package [118] can generate random realizations of the sky emission, constrained to match observational data within their uncertainties. The simulated observations are given at frequencies ranging from a few GHz to a few THz. This is the frequency range of typical CMB experiments, and in particular of the Planck sky mission [100].

Since our analysis is restricted to noise-free measurements, we completely omit instrumental noise from our simulated sky. In order not to deviate too much from realistic constraints, we limit our analysis to angular frequencies  $2 \leq \ell \leq 900$ , in which it is possible to ignore instrumental noise with respect to the other components. We set the angular resolution to 14 arcmin at all 9 channels. The frequency range is partitioned into consecutive domains, each contains 5  $\ell$ -modes, so that  $Q = 179$ . These domains are sufficiently small to approximate the piecewise stationarity assumption. The approximation that the samples within each domain are mutually decorrelated is due to (7.3). In more realistic setups, different domains may consist of a different number of  $\ell$ -modes [16, Sec. V.B]. The effective number of modes per domain,  $n_q$ , is as defined in (7.10). The simulation below uses all  $N_d = 9$  Planck channels in the range 30–857GHz.

In order to exclude observations where the Galactic emission is so strong that it completely obscures the CMB, we use a mask in our preprocessing stage. We chose a straight-band mask with smooth edges which lets through 58% of the input power. The original components and their sum, on which we perform component separation, are shown in Fig. 7.4 in one of the detector channels, after masking.

The contrast function is minimized via the JBD quasi-Newton (QN) algorithm of Sec. 5.3. The JBD algorithm is initialized with the identity matrix. The desired block-pattern  $\mathbf{m}$  is input into the algorithm. Since this algorithm may, in principle, converge to a local minimum, we verified that our results indeed reflected separation of the desired components.

### 7.5.1 Model Order Selection for the Galactic Emission

There are several considerations in setting  $N_{\text{Gal}}$ . On one hand,  $N_{\text{Gal}}$  should be small enough such that the minimization of the contrast function (2.17) does not suffer from conditioning problems. On the other hand, if too many dimensions of the Galactic emission are discarded, the model assumption (2.10) will *not* hold and our

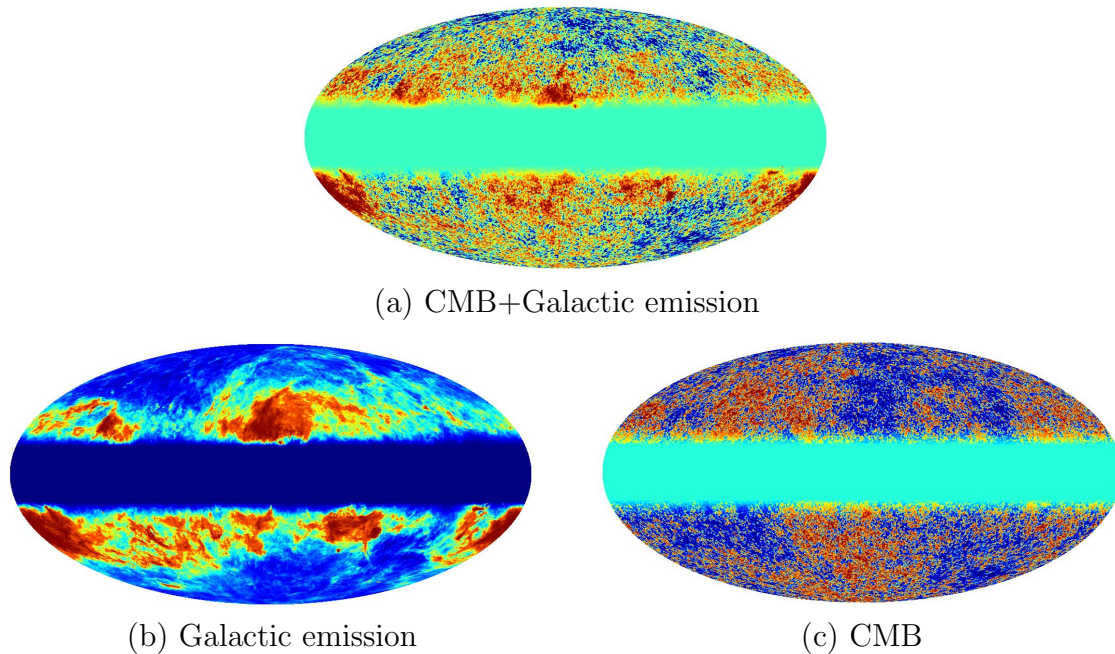


Figure 7.4: Simulated component maps at the 217GHz detector channel. Each map has a different and equalized colour scale, reflecting the brightness temperature; the value of the masked equatorial stripe is zero in all maps. The sphere is projected onto the plane using the Mollweide projection.

theoretical error prediction (Sec. 7.4.5) will fail. Since the (effective) dimension of the Galactic emission cannot be deduced from astrophysical theory, we conduct the following experiment to set it.

When the model assumptions hold, the MSE (as defined in Sec. 7.4.5) is only due to finite-data estimation (2.9) of  $\mathbf{R}_X^{(q)}$ . Therefore, if the spectral density matrices (2.9), which are the input to (2.17), do not have finite-data estimation errors, the contrast (2.17) should be zero and the components reconstructed perfectly, i.e., MSE zero. If the input spectral density matrices are free from finite-data errors, then a non-zero MSE can only stem from inadequate modeling.

Based on these arguments, we now suggest the following procedure. For each candidate  $N_s = 1 + N_{\text{Gal}}$  we calculate the empirical MSE (7.15) twice. Once, when our separation procedure uses  $\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$  followed by dimension reduction as the input to the contrast function (2.17): this is the ordinary way. The second calculation uses  $\{\mathbf{R}_{X_{\text{CMB}}}^{(q)} + \overline{\mathbf{R}}_{X_{\text{Gal}}}^{(q)}\}_{q=1}^Q$  followed by dimension reduction as the input to the contrast function (2.17). Recall from Sec. 7.4.3 that we take  $\{\overline{\mathbf{R}}_{X_{\text{Gal}}}^{(q)}\}_{q=1}^Q$  to be the model spectral density matrices of the Galactic emission component. Hence, the latter option neutralizes the finite-data errors. In both cases, the dimension of the localized spectral density matrices has to be reduced to  $N_s$ , before applying JBD,

as explained in (7.14). In each case, the output of the contrast function is used to construct the oblique projections which reconstruct the components, as explained at the end of Sec. 7.4.4. Once the components are reconstructed, we calculate the empirical MSE using (7.15).

The results for  $\text{MSE}_{\text{CMB}}$  with  $N_{\text{Gal}} = 3, 4, 5$  are given in Fig. 7.5. For this figure, this experiment was repeated for 40 Monte-Carlo trials. In each trial, a new CMB map was drawn, and new  $\{\bar{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$  were calculated. It is evident from Fig. 7.5 that the MSE for  $N_{\text{Gal}} = 3$  is much larger. More importantly, for  $N_{\text{Gal}} = 3$ , the MSE without finite-data errors is almost identical to the MSE when finite-data errors are present. This indicates that the model error is dominant. The “ordinary” MSE for  $N_{\text{Gal}} = 4$  is similar to that of  $N_{\text{Gal}} = 5$ . For  $N_{\text{Gal}} = 4$  and  $N_{\text{Gal}} = 5$ , the gap between the “ordinary” and the MSE without finite-data errors is quite large. Due to the logarithmic scale, the gap looks larger for  $N_{\text{Gal}} = 5$ . Qualitatively, both MSE values without finite-data errors, for  $N_{\text{Gal}} = 4$  and  $N_{\text{Gal}} = 5$ , are negligible with respect to their “ordinary” MSE. Similar trends were obtained with  $\text{MSE}_{\text{Gal}}$  and for other detector frequencies.  $N_{\text{Gal}} = 5$  was the largest dimension that we checked, as numerical conditioning issues prevented us from running the JBD algorithm (or the JD followed by clustering, as shall be discussed in Sec. 7.5.2) on  $N_{\text{Gal}} \geq 6$ . According to these results, we conclude that there is no significant gain in separation quality by adding the fifth dimension to the Galactic component, and thus  $N_{\text{Gal}} = 4$  seems a reasonable choice.

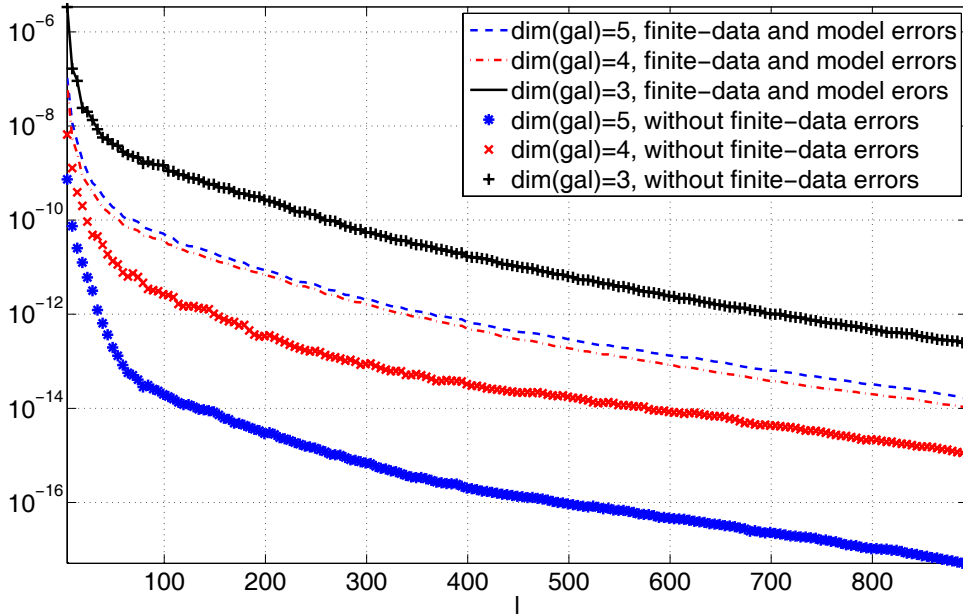


Figure 7.5: Comparison of empirical  $\text{MSE}_{\text{CMB}}$  with and without finite-data errors for  $N_{\text{Gal}} = 3, 4, 5$ . Each point of data “with finite-data errors” is the average of MC=40 randomizations of the CMB. Channel=143GHz.

In order to illustrate the meaning of the minimization of the contrast function, and to compare it to the MSE, we provide the following demonstration. In Fig. 7.6 we observe the value of the minimized contrast function (2.17) for one such experiment, i.e. with a single randomization of the CMB, for  $N_{\text{Gal}} = [2, 3, 4, 5]$ . The minimum of the contrast function is the Kullback-Leibler-induced divergence between the original data and its closest block-diagonal model, calculated on the dimension-reduced matrices:  $\min_{\mathbf{A}} C(\mathbf{A}, \{\tilde{\mathbf{R}}_X^{(q)}\}_{q=1}^Q)$ , with block-pattern  $\mathbf{m} = [1, N_{\text{Gal}}]^\dagger$ . Since the Kullback-Leibler-induced divergence in this figure originates from spectral density matrices of different dimensions, comparing results for different  $N_{\text{Gal}}$  on Fig. 7.6 is only indicative, not quantitative. As a result, for the case that the data consists of the original  $\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$ , we observe values of Kullback-Leibler-induced divergence which are more or less of the same order of magnitude, although they correspond to different values of  $N_{\text{Gal}}$ . This is different than Fig. 7.5, in which the MSE is calculated on the full  $N_d$ -dimensional components, and thus one can compare quantitatively between the MSE of different  $N_{\text{Gal}}$ . It should also be noted that the Kullback-Leibler-induced divergence measures the divergence, or fit, between two sets of matrices, namely the spectral density matrices, and their block-diagonal model. The MSE, on the other hand, measures the error between the reconstructed components and the true ones. Hence, if our purpose is component estimation, then the Kullback-Leibler-induced divergence is only an indirect and indicative figure of merit, whereas the MSE is our final target. When the finite-data errors are removed, we observe in Fig. 7.6 that the Kullback-Leibler-induced divergence decreases. For  $N_{\text{Gal}} = [2, 3]$ , the new Kullback-Leibler-induced divergence is similar. For  $N_{\text{Gal}} = 4$  we observe a larger decrease, and the best results are for  $N_{\text{Gal}} = 5$ . These results are in agreement with those of Fig. 7.5.

The advantage of using the Kullback-Leibler divergence (KLD) is that it is the natural measure of mismatch, in the sense that it arises from the maximum likelihood (ML) derivation (Sec. 2.3.3), and is the immediate outcome of the JBD algorithm. If we have at hand both the observations  $\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$  and their model  $\{\mathbf{R}_{X_{\text{CMB}}}^{(q)} + \overline{\mathbf{R}}_{X_{\text{Gal}}}^{(q)}\}_{q=1}^Q$ , then the suggested experiment yields similar conclusions about the effective  $N_{\text{Gal}}$  either by relying on the KLD or the MSE, as can be seen from Fig. 7.5 and Fig. 7.6. If, however, only the observations  $\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$  are available, then the KLD is not indicative to the effective  $N_{\text{Gal}}$ , but the empirical MSE is.

### 7.5.2 JBD vs. JD Followed by Clustering

In Fig. 7.7 we compare the minimal value of the contrast function  $C(\mathbf{A}, \tilde{\mathbf{R}}_X^{(q)})$  (2.17) vs.  $C^{\text{JD}}(\mathbf{A}, \tilde{\mathbf{R}}_X^{(q)})$  (6.1), for  $N_{\text{Gal}} = [2, 3, 4, 5]$ . The purpose is to evaluate the quality

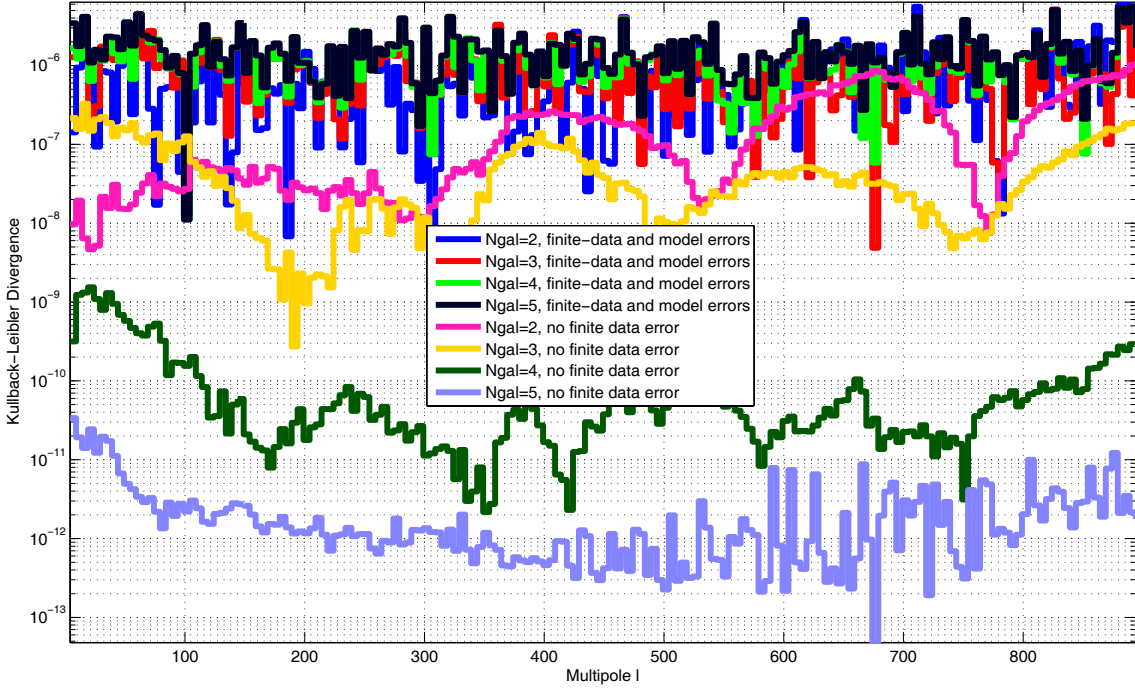


Figure 7.6: Kullback-Leibler-induced divergence between spectral density matrices and their block-diagonal model, for  $N_{\text{Gal}} = 2, 3, 4, 5$ . For each  $N_{\text{Gal}}$ , two graphs are plotted: one when the data is represented by  $\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$ , and one when the data is represented by  $\{\mathbf{R}_{X_{\text{CMB}}}^{(q)} + \overline{\mathbf{R}}_{X_{\text{Gal}}}^{(q)}\}_{q=1}^Q$ . That is, without finite-data errors. The graphs for  $\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$  are based on one realization of the CMB.

of separation of JD followed by clustering, as opposed to JBD with  $\mathbf{m} = [1, N_{\text{Gal}}]^\dagger$ , and thus to validate the multidimensional model of the Galactic emission component. The input to both algorithms is the set of matrices  $\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$ , followed by dimension reduction, as explained in Sec. 7.4.4. The data is the same as in Fig. 7.6. Therefore, the graphs for JBD are identical to those in Fig. 7.6. We observe that JD followed by clustering yields significantly poorer results, with respect to JBD, for all  $N_{\text{Gal}}$ . These results indicate that indeed the Galactic component has strong dependence within its sub-components, and thus the multidimensional approach is appropriate for this type of data.

### 7.5.3 Empirical vs. Predicted MSE

In Fig. 7.8 we compare the empirical MSE (7.15) with the predicted MSE (7.18). Until now, in the experiments in this Chapter, only the empirical MSE was calculated. We now turn to using also the closed-form expression (7.18).  $\mathbf{R}_{X_{\text{CMB}}}^{(q)}$ , set as in Sec. 7.4.3, is the spectral density matrix for CMB in (7.18). Since  $\overline{\mathbf{R}}_{X_{\text{Gal}}}^{(q)}$  is fixed for all MC trials, as explained in Sec. 7.4.3, it will serve as the spectral density

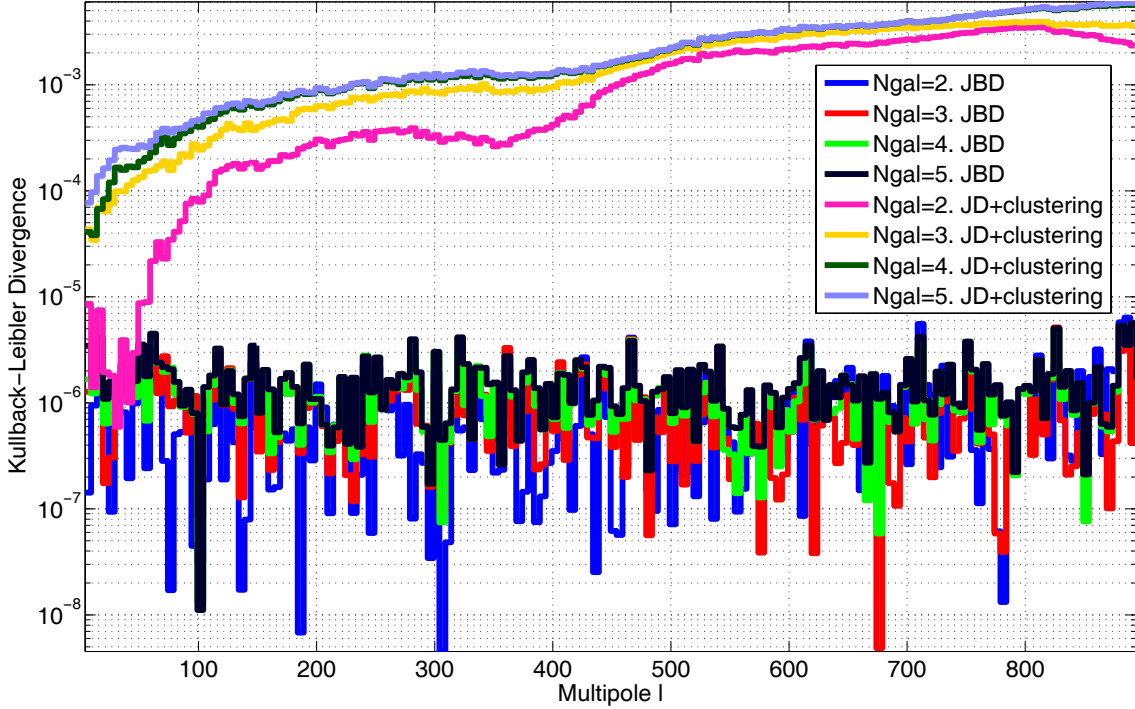


Figure 7.7: Kullback-Leibler-induced divergence between spectral density matrices and their diagonal and block-diagonal models, for  $N_{\text{Gal}} = 2, 3, 4, 5$ . For each  $N_{\text{Gal}}$ , two graphs are plotted: one when the separation is based on JBD with block-pattern  $[1, N_{\text{Gal}}]$ , and one with JD followed by clustering. The matrices are input to the algorithms after the dimension reduction scheme, explained in Sec. 7.4.4. The graphs for  $\{\overline{\mathbf{R}}_X^{(q)}\}_{q=1}^Q$  are based on one realization of the CMB.

matrix for Galactic emission in (7.18) once its dimension is reduced to the model dimension  $N_{\text{Gal}}$ . As explained in Sec. 7.4.4, SVD on  $\langle \overline{\mathbf{R}}_{X_{\text{Gal}}} \rangle$  yields an  $N_d \times N_{\text{Gal}}$  matrix  $\mathbf{U}_{\text{Gal}}$ . With  $\mathbf{P} \triangleq \mathbf{U}_{\text{Gal}} \mathbf{U}_{\text{Gal}}^\dagger$ ,  $\mathbf{P} \overline{\mathbf{R}}_{X_{\text{Gal}}}^{(q)} \mathbf{P}^\dagger$  is the desired  $N_d \times N_d$  rank- $N_{\text{Gal}}$  spectral density matrix.  $\mathbf{P} \overline{\mathbf{R}}_{X_{\text{Gal}}}^{(q)} \mathbf{P}^\dagger$  stands for the “model” spectral density matrix of the Galactic emission component, after dimension reduction, but expanded back to the observations subspace. Using these two spectral density matrices in (7.18), the predicted  $\text{MSE}_i$ ,  $i = \{\text{CMB}, \text{Gal}\}$  can be calculated.

The results are depicted in Fig. 7.8. Fig. 7.8(a) and Fig. 7.8(b) depict the theoretical  $\text{MSE}_{\text{CMB}}$  (7.18), normalized by  $c_q$ , for  $N_{\text{Gal}} = 4$  and  $N_{\text{Gal}} = 5$ , respectively. The mean and standard deviation  $\sigma$  of the empirical MSE (7.15) from 40 Monte-Carlo trials (randomizations of the CMB only) were calculated. The vertical lines indicate a range of  $\pm\sigma/2$  around the mean. The red line depicting the theoretical value is well within these  $\pm\sigma/2$  margins for most  $\ell$ . Similar trends were obtained with  $\text{MSE}_{\text{Gal}}$  and at other channel frequencies.

These results, similarly to those in Sec. 7.5.1, do not give a clear-cut preference to  $N_{\text{Gal}} = 4$  or  $N_{\text{Gal}} = 5$ . Since the difference is minor, it may be preferable to

choose the smaller dimension, as it requires less variables to adjust.

## 7.6 Discussion

In this chapter, we applied the MICA performance analysis of Chapter 3 and the QN JBD algorithm of Chapter 5 to realistic astrophysical data. In order to analyze the (simulated) sky observations in the spectral domain, we have recast our analysis into the spherical harmonics domain.

The chosen data does not follow precisely all the model assumptions of Sec. 2.3. More specifically, the CMB component is expected to behave as a perfect one-dimensional Gaussian random component. By choosing small enough domains  $\mathcal{D}_q$ , in which its angular power spectrum (depicted in Fig. 7.3) is almost constant, the CMB component is close to piecewise-stationary. The CMB component can be regarded as statistically independent from the Galactic emission component. As for the Galactic emission component, its most interesting and relevant property, for our purpose, is that it cannot be regarded as a superposition of one-dimensional statistically independent components. On the other hand, it cannot be assigned a priori with any predefined dimension. In addition, the Galactic emission component cannot be fully described using any known stochastic model. Nevertheless, due to the stochastic properties of the CMB, we can still obtain empirical decorrelation between the CMB and the Galactic emission component. This motivates the use of our MICA analysis for this data.

In a first step, we show that the problem can be regarded as overdetermined. We have thus extended our notations to overdetermined observations. Using PCA, we show that we can reduce the dimension of the data to a dimension in which the problem is well-conditioned. We then demonstrate how an effective dimension for the Galactic component can be selected, based on the observations. We show that this can be achieved either using the closed-form MSE expression, or the KLD. The advantage of using the KLD is that it is the natural measure of divergence between the data and its block-diagonal model. Natural, in the sense that it arises from minimizing our contrast function by the JBD algorithm, without additional mathematical effort. The disadvantage is that the KLD is not suited to compare quantitatively results on matrices of different dimensions. Hence, it can serve for qualitative comparison only. The MSE, on the other hand, is well-normalized, and in addition, it measures the actual estimation error of the reconstructed components. Once an effective dimension for the Galactic emission component is set, we can proceed to apply our error analysis. As predicted, the results of the dimension

determining procedure are not clear-cut: we can exclude the option of dimension 3 (or less) for the Galactic emission component. However, the choice between 4 and 5 is less obvious. It may well be that the “true” dimension is somewhere in between. For both options,  $N_{\text{Gal}} = 4$  or 5, we demonstrate good match between our theoretical closed-form MSE and the empirical results.

Another experiment, which compares the output of JBD with that of JD followed by clustering, validates our multidimensional model for the Galactic emission component, as opposed to a representation by  $N_{\text{Gal}}$  statistically independent components.

In the numerical simulations in this chapter, we validated our methods also for  $n_q \neq n_{q'} \forall q \neq q'$ . In the previous, more theoretical chapters, the numerical examples consisted only of data with equal-sized domains.

There are several sources of discrepancy between the theoretical statistical model and the data. Mainly,

- The data is not stationary within domains, and the partition into domains is ad-hoc, in some sense.
- The emission laws, which constitute the columns of the mixing matrix  $\mathbf{A}$ , can be expressed in different units. This, by itself, is of course not a problem, when the whole data is used. The output of the SVD operation in our dimension reduction scheme (Sec. 7.4.4), however, is not invariant to the choice of units. Hence, our method is not invariant to the choice of units, only indicative, and slightly different results may be obtained with different units.
- Our model assumes that each component can be exactly confined within a finite-dimensional signal space. However, as explained in Sec. 7.2.1, the dimension of the Galactic emission is not well-defined.

Consequently, in this chapter, we validate our theoretical closed-form expression for the MSE also for data for which the model assumptions hold only to some approximation.

To conclude, the simulation results indicate that good component separation has been obtained with our algorithm for this type of data. These results form the motivation for further research in this direction. It may well be that for a larger number of Monte-Carlo trials, other randomizations of the Galactic component, a different choice of domains, observation resolutions, masks etc., different results and insights may be achieved.

In a broader view of the astrophysical application, we note that the data used for this demonstration is realistic, but simplified. Obviously, when more realistic

conditions are considered, such as detector noise, additional foregrounds (other astrophysical emissions), wider observation beams (which imply poorer resolution), as well as working on more difficult regions of the sky (that is, where the CMB component is weak compared with the other foregrounds) – a more careful tuning of the separation procedure: e.g., choice of domains and masks, has to be taken into consideration.

One goal of applying our methods to CMB observations is to serve as an indicator for the best achievable performance given certain data. It can be used as a fast method to evaluate the possible gain due to fine-tuning various parameters in the model, rather than the existing situation, which is based on Monte-Carlo trials. Monte-Carlo trials consume much more resources, as they require repeated simulations of the components and of their observations (high-resolution sky maps), applying a component separation procedure to the maps, and calculating the figure of merit. The closed-form expression requires only the covariance or spectral density matrices. Hence, significant compression of the data, with a similar outcome.

In actual CMB research, the fully-blind component separation which we propose can serve as a first step in component separation, especially in cases where emission laws are not known [119, Sec. 7]. Although our method cannot separate dependent components into physically-meaningful elements, it can well serve as a preliminary step, by discarding the CMB component from the maps. In a second stage, other methods should be used to separate the dependent components. Methods which separate the dependent foregrounds usually take advantage of prior information such as additional maps from other experiments, known emission laws and other parameters, etc. See [103, 120], for example. In any case, in astrophysical study, it cannot be expected that JBD alone can do the whole separation work – at most, it can be used as one of the many steps in the very complex pipeline [121] of the data processing.

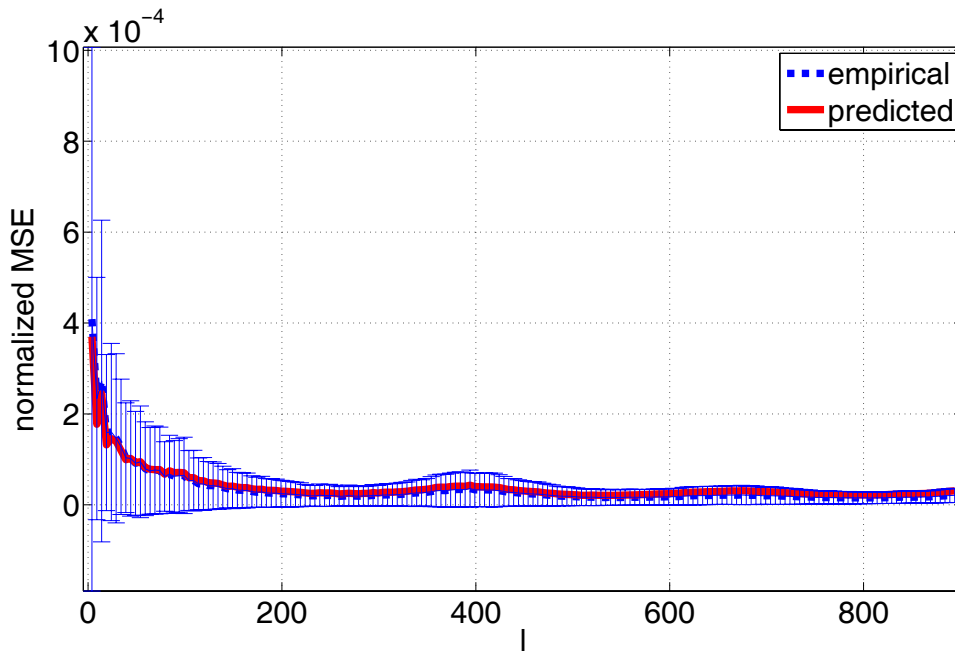
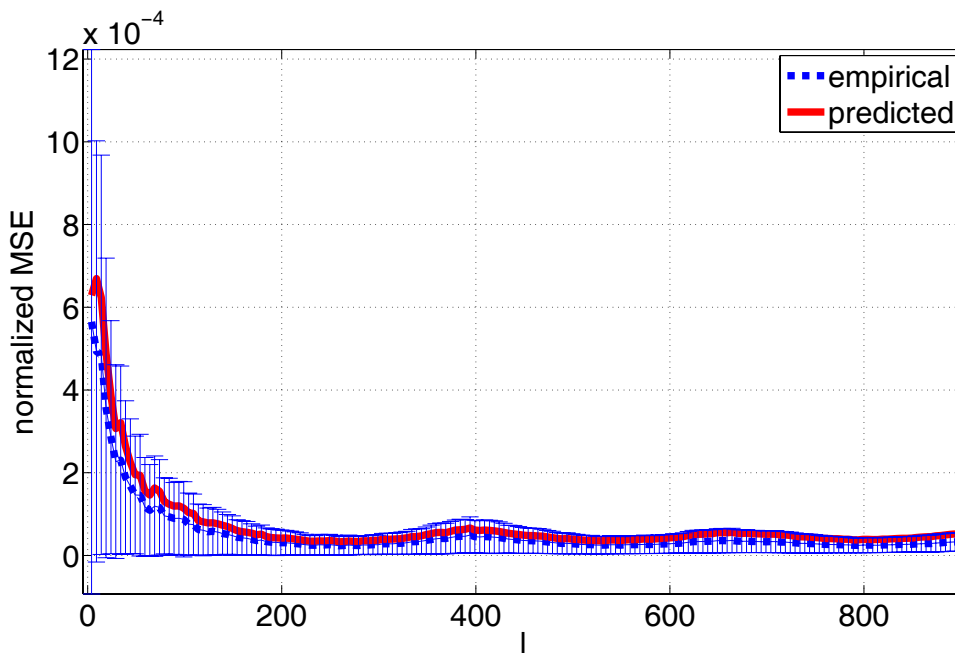
(a)  $N_{\text{Gal}} = 4$ .(b)  $N_{\text{Gal}} = 5$ .

Figure 7.8: Empirical vs. predicted  $\text{MSE}_{\text{CMB}}$ , normalized by the CMB power at each domain. The red line denotes theoretical values. The blue line denotes the empirical values, averaged over  $\text{MC}=40$ . The vertical lines denote  $\pm\sigma/2$  deviation of the MC trials around the empirical mean at each domain. Channel=100GHz.

# Chapter 8

## Conclusion

### 8.1 Summary

In this thesis, we presented the concept of independent component analysis (ICA)/blind source separation (BSS) of multidimensional components as a new perspective on the dependent sources model.

We have shown that the maximum likelihood (ML) solution, for Gaussian data and non-orthogonal mixing matrix (that is, avoiding any whitening constraints) can be obtained by minimizing a contrast function, which is a joint block diagonalization (JBD) criterion. Hence, the ML solution for the de-mixing parameters can be obtained via JBD of a set of covariance matrices.

We presented two novel non-orthogonal JBD algorithms which achieve this ML solution. One algorithm is based on the relative gradient and the other on a quasi-Newton method. These algorithms minimize the Kullback-Leibler-induced divergence between a set of symmetric positive-definite matrices and a block-diagonal function thereof. We were the first to present algorithms which achieve this minimization.

We obtained closed-form expressions for the Fisher information matrix (FIM), Cramér-Rao lower bound (CRLB) and mean square error (MSE) for the blind separation of statistically independent multidimensional Gaussian components, in a piecewise-stationary model. These expressions depend only on the model parameters. The derived expressions represent the optimal performance under asymptotic conditions. The closed-form expression for the MSE is valid also for non-Gaussian data, when the other model assumptions hold. Our work is the first one in which the performance of multidimensional ICA is given in closed-form expressions.

A different and prevalent procedure for the separation of multidimensional components consists of a two-step procedure. In the first step, applying a method which

is designed for separating one-dimensional components. In the second step, assigning the one-dimensional outputs to their multidimensional groups. This procedure is denoted in this work as “mismodeling”. We derived a closed-form expression for the expected MSE for this procedure, as a function of the model parameters. This expression is different than the minimal mean square error (MMSE) expression, obtained using the correct model and Gaussian data. Therefore, we conjecture that for our data model, the one-dimensional assumption followed by clustering is sub-optimal, when using second-order statistics. Our work is the first one to present a theoretical performance analysis for this separation procedure.

Both derivations, for the optimal multidimensional ICA (MICA) and for the mismodeled procedure, were performed in terms of components and oblique projections, which are the unambiguous counterparts of the latent sources and the mixing matrix. The derived expressions for the MSE are valid also for non-Gaussian data.

In order to complete the picture, we provide necessary and sufficient conditions which guarantee uniqueness and identifiability of the MICA model in terms of the latent covariance matrices of the mixed sources. Strictly speaking, since these conditions are based on small-error analysis, they imply *local* identifiability. We provide a complete and rigorous proof for the identifiability of the second-order MICA piecewise-stationary model. Our proof is based on the analysis of the FIM. To the best of our knowledge, no other proof for this model can be found in the literature. The identifiability conditions for the model are also the necessary and sufficient conditions for the uniqueness (up to a certain block-diagonal scaling matrix) solution to JBD of a mixture of a set of real, positive-definite symmetric JBD-irreducible block-diagonal matrices.

In order to demonstrate the practical value of our theoretical methods, we applied the optimal MICA performance analysis and our JBD algorithm to realistic astrophysical data. In this application, the purpose is to separate observations of the cosmic microwave background radiation (CMB) into their underlying components. This application is challenging because the chosen data does not follow precisely all the model assumptions. Since the number of detectors is larger than the total dimension of the involved components, we have extended our analysis to the overdetermined case. Since the dimension of the Galactic emission component is not known a-priori, we have proposed methods to determine the (effective) dimension of a component

Throughout this thesis, our theoretical expressions were supported by numerical experiments. We have shown good match between theory and empirical results.

## 8.2 Discussion

In this section, we discuss the main contributions of our work and their implications, in a more global view.

The first contribution is providing closed-form error expressions for MICA. There are several advantages to having closed-form error expressions.

1. They give insight into the influence of the model parameters on the error.
2. A closed-form expression is based on a relatively small set of parameters and can thus be regarded as a way of compressing the data. This can be significant in experiments which involve large amounts of data, as in astrophysics, where simulations may take considerable time, and high-resolution sky images consume much storage volume.
3. They are the natural way to compare different methods.

The second contribution is that we have demonstrated analytically that the one-dimensional approximation is suboptimal in the presence of finite-data errors and multidimensional components. In particular, we suggest the following observations:

1. The gain is larger as the dimensions of the components diverge from one.
2. The gain is equally significant for all the components, even for those of dimension one.
3. The gain is larger as the number of domains increases.
4. The asymptotic conditions required to achieve the expected errors are obtained with less samples per domain in the correct model setup than in the mismodeled one.

All these observations favour the use of the correct multidimensional model in the presence of multidimensional components.

The third contribution is demonstrating that all the analytical results for (multi-dimensional) ICA can be represented only in terms of well-defined quantities: projections and components. That is, a representation which does not rely on the assumption of the existence of some “mixing matrix” and “underlying sources”. This is particularly advantageous for components whose multidimensional nature does not stem from underlying well-defined lower-dimensional components but from other processes. For example, spatial localization, as with the Galactic emission or neurological measurements.

The fourth contribution is from the applicative aspect. We have provided a preliminary test of feasibility for the application of our model to astrophysical data. We have shown that our algorithm is capable of separating cosmological components in an adequate scenario. We have demonstrated that our theoretical closed-form expression for the error could indeed predict, to a reasonable precision, the MSE in such component separation. Most important, this realistic example indicates that a theoretical analysis of multidimensional methods and algorithms which takes into consideration the multidimensional character of the signals is likely to be advantageous in further applications and provide a significant contribution to processing multidimensional data.

### 8.3 Topics for Further Research

In this section, we survey some issues which were not covered in this work and are natural extensions thereof.

1. In real-life applications, noise is usually unavoidable, and cannot be ignored. In this work, we focused on a noise-free model. Still, in the presence of additive noise, our closed-form error expression can be used as a lower bound.

A paper which presents a generalization and extension of the model which is used in this thesis can be found in [16]. The model in [16] is flexible in the sense that it can treat not only multidimensional components, but also components with various parameterizations or constraints, as well as additive noise. The component separation method in [16] is based on second-order statistics, by minimizing the mismatch between sets of model and empirical covariance matrices. An important extension of our work could thus be a theoretical performance and identifiability analysis also to this model.

2. In this work, our simulations and analysis were tested with components of relatively small dimensions. Therefore, the complexity of the algorithms and their running time were not our primary concern. When larger dimensions are involved, it would be interesting to compare the complexity and speed of the quasi-Newton (QN) JBD algorithm with the state of the art joint diagonalization (JD) and clustering algorithms. This examination and comparison may be extended to other JBD algorithms in particular, and to other multidimensional methods in general.
3. An important obstacle in all JD and JBD algorithms is that in the presence of perturbations and/or noise, they sometimes get stuck in local minima. Various

methods have been suggested in the literature to cope with this problem. However, until now, no satisfactory solution has been found. This is still an open problem. It would be helpful to find a solution for the global convergence issue, both for our specific JBD algorithms, and for this problem in general.

4. In this work, our theoretical analysis has focused on the case of MICA when the number of components  $n$  and their dimensions  $m_i$  are known a-priori. In Chapter 7 we have looked at the case where  $n$  is known but  $m_i$  is not, for one of the (two) components. In general, and in real-life applications, it may be more appropriate to assume that  $n$  and  $m_i$  are generally unknown. This calls, from the empirical aspect, to the development of methods to determine these parameters. A survey of existing methods has been presented in Sec. 1.2.5. However, optimality has not been claimed for any of these methods, and there may be room for proposing additional ones. From the theoretical aspect, to the best of our knowledge, there is not yet available a performance bound for MICA for the case that  $n$  or  $m_i$  have to be estimated. It would be beneficial to derive such bounds. These bounds may help determine the efficiency of the existing algorithms, as well as lead to the derivation of closer-to-optimal ones. In addition, such a theoretical analysis may lead to yet-unavailable (to the best of our knowledge) theoretical results regarding identifiability of the MICA model when these parameters need to be estimated.
5. In this work, it is assumed that the observations are real. Hence, the JBD algorithms in Chapter 5 were developed only for real matrices. It is natural to extend all derivations and the resulting algorithms to complex data.
6. In this work, we have shown numerically that the MSE gain due to using the correct model over the one-dimensional model is larger than one. We have conjectured that since the analytical expression for mismodeling is different than the optimal one, then it is suboptimal. A natural extension of our work is to prove analytically that the MSE in the mismodeled case is always larger than the optimal one.
7. Analogously to the mismodeling analysis in Chapter 6, in which we derived the MSE when the separation procedure ignored the data's multidimensional characteristics, it would be interesting to develop closed-form expression for the figure of merit in cases that other model assumptions are violated. For example, a certain non-stationarity within domains.
8. In Chapter 7 we examined empirically the effect of JD on the performance. It

would be interesting to apply the closed-form mismodeling analysis of Chapter 6 to CMB observations, to see if its prediction is also as good as that of the correct-model MSE.

9. Once we have shown that our JBD algorithms and performance analysis can yield good results in separating astrophysical components, it would be interesting to apply the theoretical analysis and the algorithm to other types of data of potential multidimensional nature; for example: electroencephalography (EEG), magnetoencephalography (MEG), electrocardiography (ECG).

As in the case of our astrophysical application, it is expected that applying the theoretical analysis and algorithms to these data types will not be straightforward, but will require some interesting adaptations and extensions.

A natural extension of this research is to derive an analogous *optimal* performance analysis for data with different distribution types. For example, neurological activity, which consists of a superposition of isolated spikes. This analysis could point out if existing methods to extract such components already achieve the optimal performance. If not, the analysis could be used to derive more suitable algorithms.

# Appendix A

## Some Algebraic Properties

For ease of reference, we list some useful algebraic properties. Properties which are not proved below can be found in [79, 80, 122].

For any matrices  $M, N, P, Q$  (with appropriate dimensions),

$$(N \otimes M)(P \otimes Q) = NP \otimes MQ \quad (\text{A.1a})$$

$$(N \otimes M)^\dagger = N^\dagger \otimes M^\dagger \quad (\text{A.1b})$$

$$\text{vec}\{MQN\} = (N^\dagger \otimes M)\text{vec}\{Q\} \quad (\text{A.1c})$$

$$\text{tr}\{PQ\} = \text{tr}\{QP\} \quad (\text{A.1d})$$

$$\text{tr}\{P^\dagger Q\} = \text{vec}^\dagger\{P\}\text{vec}\{Q\}. \quad (\text{A.1e})$$

The Moore-Penrose pseudoinverse of a matrix  $M$  is the matrix  $M^\#$  that obeys

$$MM^\#M = M \quad (\text{A.2a})$$

$$M^\#MM^\# = M^\# \quad (\text{A.2b})$$

$$MM^\# = (MM^\#)^\dagger \quad (\text{A.2c})$$

$$M^\#M = (M^\#M)^\dagger. \quad (\text{A.2d})$$

The  $mn \times mn$  commutation matrix  $\mathcal{T}_{m,n}$  [79] is such that

$$\text{vec}\{M^\dagger\} = \mathcal{T}_{m,n}\text{vec}\{M\} \quad (\text{A.3})$$

for any  $m \times n$  matrix  $M$ . The commutation matrix is isometric and satisfies  $\mathcal{T}_{n,m}\mathcal{T}_{m,n} = I$ . Hence,  $\mathcal{T}_{n,m} = \mathcal{T}_{m,n}^{-1}$ . One also has  $\mathcal{T}_{n,m} = \mathcal{T}_{m,n}^\dagger$ . For any two

matrices  $\mathbf{M}_{m \times n}$  and  $\mathbf{N}_{p \times q}$ ,

$$\mathbf{N} \otimes \mathbf{M} = \mathcal{T}_{p,m}(\mathbf{M} \otimes \mathbf{N})\mathcal{T}_{n,q} \quad (\text{A.4a})$$

$$\mathcal{T}_{m,p}(\mathbf{N} \otimes \mathbf{M}) = (\mathbf{M} \otimes \mathbf{N})\mathcal{T}_{n,q}. \quad (\text{A.4b})$$

**Property A.1.** Let  $\mathbf{V}, \mathbf{M}, \mathbf{N}$  be  $n \times n$  matrices. Then

$$\text{tr} \{ \mathbf{M} \mathbf{V} \mathbf{N}^\dagger \} = \text{tr} \{ (\mathbf{V}^\dagger \otimes \mathbf{I}) \text{vec} \{ \mathbf{M} \} \text{vec}^\dagger \{ \mathbf{N} \} \}.$$

*Proof.*

$$\begin{aligned} \text{tr} \{ \mathbf{M} \mathbf{V} \mathbf{N}^\dagger \} &= \text{vec}^\dagger \{ \mathbf{N} \} \text{vec} \{ \mathbf{M} \mathbf{V} \} = \text{vec}^\dagger \{ \mathbf{N} \} (\mathbf{V}^\dagger \otimes \mathbf{I}) \text{vec} \{ \mathbf{M} \} \\ &= \text{tr} \{ (\mathbf{V}^\dagger \otimes \mathbf{I}) \text{vec} \{ \mathbf{M} \} \text{vec}^\dagger \{ \mathbf{N} \} \} \end{aligned}$$

□

**Property A.2.** For any four vectors  $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ ,

$$\text{vec} \{ \mathbf{a} \mathbf{b}^\dagger \} \text{vec}^\dagger \{ \mathbf{c} \mathbf{d}^\dagger \} = \mathbf{d} \mathbf{b}^\dagger \otimes \mathbf{a} \mathbf{c}^\dagger. \quad (\text{A.5})$$

*Proof.* Equation (A.5) follows from

$$\begin{aligned} \text{vec} \{ \mathbf{a} \mathbf{b}^\dagger \} \text{vec}^\dagger \{ \mathbf{c} \mathbf{d}^\dagger \} &= (\mathbf{b} \otimes \mathbf{a}) (\mathbf{d} \otimes \mathbf{c})^\dagger \\ &= (\mathbf{b} \otimes \mathbf{a}) (\mathbf{d}^\dagger \otimes \mathbf{c}^\dagger) = \mathbf{b} \mathbf{d}^\dagger \otimes \mathbf{a} \mathbf{c}^\dagger. \end{aligned} \quad (\text{A.6})$$

The first equality of (A.6) is based on the property that for any two vectors  $\mathbf{a}$  and  $\mathbf{b}$ ,  $\text{vec} \{ \mathbf{a} \mathbf{b}^\dagger \} = \text{vec} \{ \mathbf{a} \cdot \mathbf{1} \cdot \mathbf{b}^\dagger \} = \mathbf{b} \otimes \mathbf{a}$ , where the second step is due to (A.1c) in Appendix A. The second equality of (A.6) is based on (A.1b) and the third on (A.1a) in Appendix A. □

# Bibliography

- [1] P. Comon, “Independent component analysis,” in *Proc. International Signal Processing Workshop on Higher Order Statistics*, Chamrousse, France, Jul. 1991, pp. 111–120, keynote address. Republished in *HOS*, J.-L. Lacoume ed., Elsevier, 1992, pp. 29–38.
- [2] —, “Independent component analysis, a new concept?” *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994.
- [3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, 2001.
- [4] J.-F. Cardoso, “Blind signal separation: Statistical principles,” *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.
- [5] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, 1st ed. Academic Press, Feb. 2010.
- [6] F. Kohl, G. Wübbeler, D. Kolossa, C. Elster, M. Bär, and R. Orglmeister, “Non-independent BSS: A model for evoked MEG signals with controllable dependencies,” in *Independent Component Analysis and Signal Separation*, T. Adalı, C. Jutten, J. M. T. Romano, and A. K. Barros, Eds. Heidelberg: Springer, 2009, pp. 443–450.
- [7] A. Ossadtchi, S. Baillet, J. C. Mosher, D. Thyerlei, W. Sutherling, and R. M. Leahy, “Automated interictal spike detection and source localization in magnetoencephalography using independent components analysis and spatio-temporal clustering.” *Clinical Neurophysiology*, vol. 115, no. 3, pp. 508–522, Mar. 2004.
- [8] A. Hyvärinen and P. O. Hoyer, “Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces,” *Neural Computation*, vol. 12, no. 7, pp. 1705–1720, Jul. 2000.
- [9] L. De Lathauwer, B. De Moor, and J. Vandewalle, “Fetal electrocardiogram extraction by source subspace separation,” in *Proc. IEEE Signal Processing / ATHOS Workshop on Higher-Order Statistics*, Girona, Spain, Jun. 1995, pp. 134–138.

- [10] J.-F. Cardoso, “Multidimensional independent component analysis,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Seattle, WA, May 1998, pp. 1941–1944.
- [11] L. De Lathauwer, B. De Moor, and J. Vandewalle, “Fetal electrocardiogram extraction by blind source subspace separation,” *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 5, pp. 567–572, May 2000.
- [12] H. W. Gutch, J. Krumsiek, and F. J. Theis, “An ISA algorithm with unknown group sizes identifies meaningful clusters in metabolomics data,” in *Proc. European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, Aug. 29 – Sep. 2 2011, pp. 1733–1737.
- [13] L. Bedini, D. Herranz, E. Salerno, C. Baccigalupi, E. E. Kuruoğlu, and A. Tonazzini, “Separation of correlated astrophysical sources using multiple-lag data covariance matrices,” *EURASIP JASP*, vol. 2005, no. 15, pp. 2400–2412, 2005.
- [14] E. E. Kuruoğlu, “Dependent component analysis for cosmology: A case study,” in *Latent Variable Analysis and Signal Separation*, ser. Lecture Notes in Computer Science, V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, and E. Vincent, Eds. Heidelberg: Springer, 2010, vol. 6365, pp. 538–545.
- [15] Planck Collaboration, “Planck Early Results XVIII: The power spectrum of cosmic infrared background anisotropies,” *Astronomy & Astrophysics*, vol. 536, no. A18, pp. 1–30, Dec. 2011.
- [16] J.-F. Cardoso, M. Le Jeune, J. Delabrouille, M. Betoule, and G. Patanchon, “Component separation with flexible models – application to multichannel astrophysical observations,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 5, pp. 735–746, Oct. 2008.
- [17] Z. Szabó, B. Póczos, and A. Lőrincz, “Separation theorem for independent subspace analysis and its consequences,” *Pattern Recognition*, vol. 45, no. 4, pp. 1782–1791, Apr. 2012.
- [18] A. Hyvärinen, “Beyond independent components,” in *Proc. 9th International Conference on Artificial Neural Networks (ICANN)*, vol. 2, Edinburgh, UK, Sep. 1999, pp. 809–814.
- [19] A. Hyvärinen and U. Köster, “FastISA: A fast fixed-point algorithm for independent subspace analysis,” in *Proc. European Symposium on Artificial Neural Networks (ESANN)*, Bruges (Belgium), Apr. 2006.

- [20] H. Shen, K. Hüper, and M. Kleinsteuber, “Block Jacobi-type methods for log-likelihood based linear independent subspace analysis,” in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2007)*, Thessaloniki, Greece, Aug. 2007, pp. 133–138.
- [21] H. Shen and K. Hüper, “Generalised FastICA for independent subspace analysis,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Honolulu, HI, Apr. 2007, pp. IV–1409–IV–1412.
- [22] J.-F. Cardoso, “On the performance of orthogonal source separation algorithms,” in *Proc. European Signal Processing Conference (EUSIPCO)*, Edinburgh, Sep. 1994, pp. 776–779.
- [23] L. De Lathauwer, B. De Moor, and J. Vandewalle, “A prewhitening-induced bound on the identification error in independent component analysis,” *IEEE Transactions on Circuits and Systems—Part I: Regular Papers*, vol. 52, no. 3, pp. 546–554, Mar. 2005.
- [24] F. J. Theis, “Towards a general independent subspace analysis,” in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 1361–1368.
- [25] H. Choi and S. Choi, “Relative gradient learning for independent subspace analysis,” in *Proc. International Joint Conference on Neural Networks (IJCNN)*, Vancouver, Canada, Jul. 2006, pp. 3919–3924.
- [26] S. Kirshner and B. Póczos, “ICA and ISA using Schweizer-Wolff measure of dependence,” in *Proc. 25th International Conference on Machine Learning (ICML)*. Helsinki, Finland: ACM Press, 2008, pp. 464–471.
- [27] P. Gruber, H. W. Gutch, and F. J. Theis, “Hierarchical extraction of independent subspaces of unknown dimensions,” in *Independent Component Analysis and Signal Separation*, T. Adalı, C. Jutten, J. M. T. Romano, and A. K. Barros, Eds. Heidelberg: Springer, 2009, pp. 259–266.
- [28] Z. Szabó, B. Póczos, and A. Lőrincz, “Undercomplete blind subspace deconvolution,” *Journal of Machine Learning Research*, vol. 8, pp. 1063–1095, May 2007.
- [29] F. R. Bach and M. I. Jordan, “Beyond independent components: Trees and clusters,” *Journal of Machine Learning Research*, vol. 4, pp. 1205–1233, Dec. 2003.
- [30] O. Shalvi and E. Weinstein, “Maximum likelihood and lower bounds in system identification with non-Gaussian inputs,” *IEEE Transactions on Information Theory*, vol. 40, no. 2, pp. 328–339, Mar. 1994.

- [31] V. Vigneron and C. Jutten, *Independent Component Analysis and Blind Signal Separation*, ser. Lecture Notes in Computer Science. Heidelberg: Springer, Sep. 2004, vol. 3195, ch. Fisher Information in Source Separation Problems, pp. 168–176.
- [32] P. Tichavský, Z. Koldovský, and E. Oja, “Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis,” *IEEE Transactions on Signal Processing*, vol. 54, no. 4, pp. 1189–1203, Apr. 2006.
- [33] E. Ollila, H.-J. Kim, and V. Koivunen, “Compact Cramér-Rao bound expression for independent component analysis,” *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1421–1428, Apr. 2008.
- [34] S. Dégerine and A. Zaïdi, “Separation of an instantaneous mixture of Gaussian autoregressive sources by the exact maximum likelihood approach,” *IEEE Transactions on Signal Processing*, vol. 52, no. 6, pp. 1499–1512, Jun. 2004.
- [35] E. Doron, A. Yeredor, and P. Tichavský, “Cramér-Rao-induced bound for blind separation of stationary parametric Gaussian sources,” *IEEE Signal Processing Letters*, vol. 14, no. 6, pp. 417–420, Jun. 2007.
- [36] P. Tichavský, A. Yeredor, and Z. Koldovský, “A fast asymptotically efficient algorithm for blind separation of a linear mixture of block-wise stationary autoregressive processes,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 3133–3136.
- [37] J.-F. Cardoso and D.-T. Pham, *Independent component analysis: principles and practice*. Cambridge, United Kingdom: Cambridge University Press, 2001, ch. Separation of non-stationary sources: algorithms and performance, pp. 158–180.
- [38] A. Yeredor, “Blind separation of Gaussian sources with general covariance structures: Bounds and optimal estimation,” *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5057–5068, Oct. 2010.
- [39] M. Anderson, T. Adalı, and X.-L. Li, “Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis,” *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1672–1683, Apr. 2012.
- [40] D.-T. Pham and P. Garat, “Blind separation of mixtures of independent sources through a quasi-maximum likelihood approach,” *IEEE Transactions on Signal Processing*, vol. 45, no. 7, pp. 1712–1725, Jul. 1997.
- [41] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, “A blind source separation technique based on second-order statistics,” *IEEE Transactions on Signal Processing*, vol. 45, no. 2, pp. 434–444, Feb. 1997.

- [42] D.-T. Pham and J.-F. Cardoso, “Blind separation of instantaneous mixtures of non stationary sources,” *IEEE Transactions on Signal Processing*, vol. 49, no. 9, pp. 1837–1848, Sep. 2001.
- [43] J.-F. Cardoso, “Separation of non stationary sources; achievable performance,” in *Proc. IEEE 10th Workshop Stat. Signal, Array Processing*, Pocono Manor, PA, Aug. 2000, pp. 359–363.
- [44] D. Lahat, J.-F. Cardoso, and H. Messer, “Optimal performance of second-order multidimensional ICA,” in *Independent Component Analysis and Signal Separation*, T. Adalı, C. Jutten, J. M. T. Romano, and A. K. Barros, Eds. Heidelberg: Springer, 2009, pp. 50–57.
- [45] —, “Second-order multidimensional ICA: Performance analysis,” *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 1–13, Sep. 2012.
- [46] K. Murota, Y. Kanno, M. Kojima, and S. Kojima, “A numerical algorithm for block-diagonal decomposition of matrix \*-algebras with application to semidefinite programming,” *Japan J. Indust. Appl. Math.*, vol. 27, no. 1, pp. 125–160, Jun. 2010.
- [47] L. De Lathauwer, “Decompositions of a higher-order tensor in block terms. Part II: Definitions and uniqueness,” *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 1033–1066, 2008.
- [48] L. Tong, R. Liu, V. Soon, and Y. F. Huang, “Indeterminacy and identifiability of blind identification,” *IEEE Transactions on Circuits and Systems*, vol. 38, no. 5, pp. 499–509, May 1991.
- [49] H. W. Gutch, T. Maehara, and F. J. Theis, “Second order subspace analysis and simple decompositions,” in *Latent Variable Analysis and Signal Separation*, ser. Lecture Notes in Computer Science, V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, and E. Vincent, Eds., vol. 6365. Heidelberg: Springer, 2010, pp. 370–377.
- [50] J. Vía, M. Anderson, X.-L. Li, and T. Adalı, “Joint blind source separation from second-order statistics: Necessary and sufficient identifiability conditions,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 2520–2523.
- [51] D. Lahat, J.-F. Cardoso, and H. Messer, “Identifiability of second-order multidimensional ICA,” in *Proc. European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, Aug. 2012.
- [52] D. Nion, “A tensor framework for nonunitary joint block diagonalization,” *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4585–4594, Oct. 2011.

- [53] P. Tichavský and Z. Koldovský, “Algorithms for nonorthogonal approximate joint block-diagonalization,” in *Proc. European Signal Processing Conference (EU-SIPCO)*, Bucharest, Romania, Aug. 2012.
- [54] H. Ghennioui, N. Thirion-Moreau, E. Moreau, and D. Aboutajdine, “Gradient-based joint block diagonalization algorithms: Application to blind separation of FIR convolutive mixtures,” *Signal Processing*, vol. 90, no. 6, pp. 1836–1849, 2010.
- [55] T. Maehara and K. Murota, “Algorithm for error-controlled simultaneous block-diagonalization of matrices,” *SIAM Journal on Matrix Analysis and Applications*, vol. 32, no. 2, pp. 605–620, 2011.
- [56] H. Bousbia-Salah, A. Belouchrani, and K. Abed-Meraim, “Blind separation of non stationary sources using joint block diagonalization,” in *Proc. 11th IEEE Signal Processing Workshop on Statistical Signal Processing*, Aug. 2001, pp. 448–451.
- [57] D.-T. Pham, “Joint approximate diagonalization of positive definite hermitian matrices,” *SIAM Journal on Matrix Analysis and Applications*, vol. 22, no. 4, pp. 1136–1152, 2001.
- [58] ———, “Blind separation of cyclostationary sources using joint block approximate diagonalization,” in *Independent Component Analysis and Signal Separation*, ser. Lecture Notes in Computer Science, vol. 4666. Heidelberg: Springer, 2007, pp. 244–251.
- [59] S. Amari, A. Cichocki, and H. H. Yang, “A new learning algorithm for blind signal separation,” in *Adv. Neural Inf. Process. Syst.* MIT Press, 1996, pp. 757–763.
- [60] D. Lahat, J.-F. Cardoso, and H. Messer, “Joint block diagonalization algorithms for optimal separation of multidimensional components,” in *Latent Variable Analysis and Signal Separation*, ser. Lecture Notes in Computer Science, F. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, Eds., vol. 7191. Heidelberg: Springer, 2012, pp. 155–162.
- [61] C. Févotte and C. Doncarli, “A unified presentation of blind separation methods for convolutive mixtures using block-diagonalization,” in *Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003)*, Nara, Japan, Apr. 2003, pp. 349–354.
- [62] Z. Koldovský and P. Tichavský, “Time-domain blind audio source separation using advanced component clustering and reconstruction,” in *Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Trento, Italy, 6–8 May 2008, pp. 216–219.

- [63] K. Abed-Meraim and A. Belouchrani, “Algorithms for joint block diagonalization,” in *Proc. European Signal Processing Conference (EUSIPCO)*, Vienna, Austria, 2004, pp. 209–212.
- [64] P. Tichavský, A. Yeredor, and Z. Koldovský, “On computation of approximate joint block-diagonalization using ordinary AJD,” in *Latent Variable Analysis and Signal Separation*, ser. Lecture Notes in Computer Science, F. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, Eds., vol. 7191. Springer, 2012, pp. 163–171.
- [65] P. Tichavský and A. Yeredor, “Fast approximate joint diagonalization incorporating weight matrices,” *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 878–891, Mar. 2009.
- [66] A. Hyvärinen, “Fast and robust fixed-point algorithms for independent component analysis,” *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, May 1999.
- [67] A.-J. van der Veen, “Joint diagonalization via subspace fitting techniques,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, Salt Lake City, UT, May 7–11 2001, pp. 2773–2776.
- [68] A. Yeredor, “Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation,” *IEEE Transactions on Signal Processing*, vol. 50, no. 7, pp. 1545–1553, Jul. 2002.
- [69] M. Joho and H. Mathis, “Joint diagonalization of correlation matrices by using gradient methods with application to blind signal separation,” in *Proc. Sensor Array and Multichannel Signal Processing Workshop*, Rosslyn, VA, 4–6 Aug. 2002, pp. 273–277.
- [70] M. Joho and K. Rahbar, “Joint diagonalization of correlation matrices by using newton methods with application to blind signal separation,” in *Proc. Sensor Array and Multichannel Signal Processing Workshop*, Rosslyn, VA, 4–6 Aug. 2002, pp. 403–407.
- [71] W.-T. Zhang, S.-T. Lou, and H.-M. Lu, “Fast nonunitary joint block diagonalization with degenerate solution elimination for convolutive blind source separation,” *Digital Signal Processing*, vol. 22, no. 5, pp. 808–819, Sep. 2012.
- [72] Z. Koldovský and P. Tichavský, “A comparison of independent component and independent subspace analysis algorithms,” in *Proc. European Signal Processing Conference (EUSIPCO)*, Glasgow, Scotland, 24–28 Aug. 2009.
- [73] D. Lahat, J.-F. Cardoso, and H. Messer, “ICA of correlated sources mismodeled as uncorrelated: performance analysis,” in *Proc. IEEE/SP 15th Workshop Stat. Signal Process.*, Aug. 31–Sep. 3 2009, pp. 489–492.

- [74] J.-F. Cardoso, “The three easy routes to independent component analysis; contrasts and geometry,” in *Proc. International Workshop on Independent Component Analysis and Blind Signal Separation*, San Diego, CA, Dec. 2001, pp. 1–6.
- [75] D. Lahat, J.-F. Cardoso, M. Le Jeune, and H. Messer, “Multidimensional ICA and its performance analysis applied to CMB observations,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 22–27 May 2011, pp. 3724–3727.
- [76] H. W. Gutch and F. J. Theis, “Independent subspace analysis is unique, given irreducibility,” in *Independent Component Analysis and Signal Separation*, ser. Lecture Notes in Computer Science, M. E. Davies, C. J. James, S. A. Abdallah, and M. D. Plumbley, Eds., vol. 4666. Heidelberg: Springer, 2007, pp. 49–56.
- [77] —, “Uniqueness of linear factorizations into independent subspaces,” *Journal of Multivariate Analysis*, vol. 112, pp. 48–62, Nov. 2012.
- [78] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951.
- [79] J. R. Magnus and H. Neudecker, “The commutation matrix: Some properties and applications,” *The Annals of Statistics*, vol. 7, no. 2, pp. 381–394, Mar. 1979.
- [80] K. B. Petersen and M. S. Pedersen, “The matrix cookbook,” Oct. 2008, version 20081110. [Online]. Available: <http://www2.imm.dtu.dk/pubdb/p.php?3274>
- [81] R. E. Bellman, *Introduction to Matrix Analysis*. McGraw-Hill, New York, 1960.
- [82] J.-F. Cardoso and B. Laheld, “Equivariant adaptive source separation,” *IEEE Transactions on Signal Processing*, vol. 44, no. 12, pp. 3017–3030, Dec. 1996.
- [83] D.-T. Pham, “Information approach to blind source separation and deconvolution,” in *Information Theory and Statistical Learning*, F. Emmert-Streib and M. Dehmer, Eds. Springer, 2009, ch. 7, pp. 153–182.
- [84] J. Vía, M. Anderson, X.-L. Li, and T. Adalı, “A maximum likelihood approach for independent vector analysis of Gaussian data sets,” in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2011)*, Beijing, China, Sep. 2011.
- [85] L. De Lathauwer and D. Nion, “Decompositions of a higher-order tensor in block terms. Part III: Alternating least squares algorithms,” *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 1067–1083, 2008.
- [86] T. G. Kolda and B. W. Bader, “Tensor decompositions and applications,” *SIAM Review*, vol. 51, no. 3, pp. 455–500, Sep. 2009.

- [87] N. Jarosik, “The wilkinson microwave anisotropy probe (WMAP): Using precision microwave measurements to study cosmology,” in *Proc. 13th International Symposium on Antenna Technology and Applied Electromagnetics and the Canadian Radio Science Meeting. ANTEM/URSI*, Banff, AB, Canada, 15–18 Feb. 2009, pp. 1–4.
- [88] G. Gamow, “Expanding universe and the origin of elements,” *Phys. Rev.*, vol. 70, pp. 572–573, Oct. 1946.
- [89] W. Hu and S. Dodelson, “Cosmic microwave background anisotropies,” *Annual Review of Astronomy and Astrophysics*, vol. 40, pp. 171–216, Sep. 2002.
- [90] A. A. Penzias and R. W. Wilson, “A measurement of excess antenna temperature at 4080 Mc/s,” *The Astrophysical Journal*, vol. 142, pp. 419–421, Jul. 1965.
- [91] R. H. Dicke, P. J. E. Peebles, P. G. Roll, and D. T. Wilkinson, “Cosmic black-body radiation,” *The Astrophysical Journal*, vol. 142, pp. 414–419, Jul. 1965.
- [92] G. F. Smoot, C. L. Bennett, A. Kogut, E. L. Wright, J. Aymon, N. W. Boggess, E. S. Cheng, G. de Amici, S. Gulkis, M. G. Hauser, G. Hinshaw, P. D. Jackson, M. Janssen, E. Kaita, T. Kelsall, P. Keegstra, C. Lineweaver, K. Loewenstein, P. Lubin, J. Mather, S. S. Meyer, S. H. Moseley, T. Murdock, L. Rokke, R. F. Silverberg, L. Tenorio, R. Weiss, and D. T. Wilkinson, “Structure in the COBE differential microwave radiometer first-year maps,” *Astrophysical Journal, Part 2 - Letters*, vol. 396, no. 1, pp. L1–L5, Sep. 1992.
- [93] E. Komatsu, K. M. Smith, J. Dunkley, C. L. Bennett, B. Gold, G. Hinshaw, N. Jarosik, D. Larson, M. R. Nolta, L. Page, D. N. Spergel, M. Halpern, R. S. Hill, A. Kogut, M. Limon, S. S. Meyer, N. Odegard, G. S. Tucker, J. L. Weiland, E. Wollack, and E. L. Wright, “Seven-year Wilkinson microwave anisotropy probe (WMAP) observations: Cosmological interpretation,” *The Astrophysical Journal Supplement Series*, vol. 192, no. 2, Feb. 2011.
- [94] M. Tegmark, D. J. Eisenstein, W. Hu, and A. de Oliveira-Costa, “Foregrounds and forecasts for the cosmic microwave background,” *The Astrophysical Journal*, vol. 530, no. 1, pp. 133–165, Feb. 2000, astro-ph/9905257.
- [95] J. Delabrouille, M. Betoule, J.-B. Melin, M.-A. Miville-Deschênes, J. Gonzalez-Nuevo, M. L. Jeune, G. Castex, G. De Zotti, S. Basak, M. Ashdown, J. Aumont, C. Baccigalupi, A. Banday, J.-P. Bernard, F. R. Bouchet, D. L. Clements, A. da Silva, C. Dickinson, F. Dodu, K. Dolag, F. Elsner, L. Fauvet, G. Faÿ, G. Giardino, S. Leach, J. Lesgourgues, M. Liguori, J. F. Macias-Perez, M. Massardi, S. Matarrese, P. Mazzotta, L. Montier, S. Mottet, R. Paladini, B. Partridge, R. Piffaretti, G. Prezeau, S. Prunet, S. Ricciardi, M. Roman, B. Schaefer, and L. Toffolatti, “The pre-launch Planck Sky Model: a model of sky emission

- at submillimetre to centimetre wavelengths,” 2012, submitted to *Astronomy & Astrophysics*. [Online]. Available: <http://arxiv.org/abs/1207.3675v1>
- [96] Y. Rephaeli, “Comptonization of the cosmic microwave background: The Sunyaev-Zeldovich effect,” *Annual Review of Astronomy and Astrophysics*, vol. 33, pp. 541–579, 1995.
- [97] Planck Collaboration, “Planck early results. VIII. The all-sky early Sunyaev-Zeldovich cluster sample,” *Astronomy & Astrophysics*, vol. 536, no. A8, pp. 1–28, Dec. 2011.
- [98] G. Hinshaw, M. R. Nolta, C. L. Bennett, R. Bean, O. Doré, M. R. Greason, M. Halpern, R. S. Hill, N. Jarosik, A. Kogut, E. Komatsu, M. Limon, N. Odegard, S. S. Meyer, L. Page, H. V. Peiris, D. N. Spergel, G. S. Tucker, L. Verde, J. L. Weiland, E. Wollack, and E. L. Wright, “Three-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Temperature analysis,” *The Astrophysical Journal Supplement Series*, vol. 170, pp. 288–334, Jun. 2007.
- [99] N. Jarosik, C. L. Bennett, J. Dunkley, B. Gold, M. R. Greason, M. Halpern, R. S. Hill, G. Hinshaw, A. Kogut, E. Komatsu, D. Larson, M. Limon, S. S. Meyer, M. R. Nolta, N. Odegard, L. Page, K. M. Smith, D. N. Spergel, G. S. Tucker, J. L. Weiland, E. Wollack, and E. L. Wright, “Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Sky maps, systematic errors, and basic results,” *The Astrophysical Journal Supplement Series*, vol. 192, no. 14, Feb. 2011.
- [100] Planck Collaboration, *Planck: The Scientific Programme*. ESA-SCI, 2005, vol. 1. [Online]. Available: [http://www.rssd.esa.int/SA/PLANCK/docs/Bluebook-ESA-SCI\(2005\)1\\_V2.pdf](http://www.rssd.esa.int/SA/PLANCK/docs/Bluebook-ESA-SCI(2005)1_V2.pdf)
- [101] NASA, “WMAP seven year data release images,” [http://map.gsfc.nasa.gov/resources/featured\\_images\\_7yr\\_release.html](http://map.gsfc.nasa.gov/resources/featured_images_7yr_release.html).
- [102] R. Vio and P. Andreani, “A statistical analysis of the “internal linear combination” method in problems of signal separation as in cosmic microwave background observations,” *Astronomy & Astrophysics*, vol. 487, no. 2, pp. 775–780, Aug. 2008.
- [103] B. Gold, N. Odegard, J. L. Weiland, R. S. Hill, A. Kogut, C. L. Bennett, G. Hinshaw, X. Chen, J. Dunkley, M. Halpern, N. Jarosik, E. Komatsu, D. Larson, M. Limon, S. S. Meyer, M. R. Nolta, L. Page, K. M. Smith, D. N. Spergel, G. S. Tucker, E. Wollack, and E. L. Wright, “Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Galactic foreground emission,” *The Astrophysical Journal Supplement Series*, vol. 192, no. 15, Feb. 2011.

- [104] J. Dick, M. Remazeilles, and J. Delabrouille, “Impact of calibration errors on CMB component separation using FastICA and ILC,” *Monthly Notices of the Royal Astronomical Society*, vol. 401, pp. 1602–1612, 2010.
- [105] R. Vio and P. Andreani, “Considerations on some neglected but important issues concerning the internal linear combination method in astronomy,” 2009, submitted to *Astronomy & Astrophysics*. [Online]. Available: <http://arxiv.org/abs/0910.4294v1>
- [106] C. Baccigalupi, L. Bedini, C. Burigana, G. De Zotti, A. Farusi, D. Maino, M. Maris, F. Perrotta, E. Salerno, L. Toffolatti, and A. Tonazzini, “Neural networks and the separation of cosmic microwave background and astrophysical signals in sky maps,” *Monthly Notices of the Royal Astronomical Society*, vol. 318, no. 3, pp. 769–780, Nov. 2000.
- [107] D. Maino, A. J. Banday, C. Baccigalupi, F. Perrotta, and K. M. Górski, “Astrophysical component separation of COBE-DMR 4-yr data with FastICA,” *Monthly Notices of the Royal Astronomical Society*, vol. 344, no. 2, pp. 544–552, Sep. 2003.
- [108] D. Maino, S. Donzelli, A. J. Banday, F. Stivoli, and C. Baccigalupi, “Cosmic microwave background signal in Wilkinson microwave anisotropy probe three-year data with FastICA,” *Monthly Notices of the Royal Astronomical Society*, vol. 374, no. 4, pp. 1207–1215, Feb. 2007.
- [109] J. Delabrouille, J.-F. Cardoso, and G. Patanchon, “Multidetector multicomponent spectral matching and applications for cosmic microwave background data analysis,” *Monthly Notices of the Royal Astronomical Society*, vol. 346, no. 4, pp. 1089–1102, Dec. 2003.
- [110] A. Bonaldi, L. Bedini, E. Salerno, C. Baccigalupi, and G. De Zotti, “Estimating the spectral indices of correlated astrophysical foregrounds by a second-order statistical approach,” *Monthly Notices of the Royal Astronomical Society*, vol. 373, no. 1, pp. 271–279, Nov. 2006.
- [111] D. Larson, J. Dunkley, G. Hinshaw, E. Komatsu, M. R. Nolta, C. L. Bennett, B. Gold, M. Halpern, R. S. Hill, N. Jarosik, A. Kogut, M. Limon, S. S. Meyer, N. Odegard, L. Page, K. M. Smith, D. N. Spergel, G. S. Tucker, J. L. Weiland, E. Wollack, and E. L. Wright, “Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Power spectra and WMAP-derived parameters,” *The Astrophysical Journal Supplement Series*, vol. 192, no. 2, 2011.
- [112] M. White, L. M. Krauss, and J. Silk, “Cosmic variance in CMB anisotropies: From  $1^\circ$  to COBE,” *The Astrophysical Journal*, vol. 418, pp. 535–543, Dec. 1993.

- [113] M. Kamionkowski and A. Loeb, “Getting around cosmic variance,” *Physical Review D*, vol. 56, no. 8, pp. 4511–4513, Oct. 1997.
- [114] J.-F. Cardoso, “Precision cosmology with the cosmic microwave background,” *IEEE Signal Processing Magazine*, vol. 27, no. 1, pp. 55–66, Jan. 2010.
- [115] Planck Collaboration, “Planck Early Results: The Planck mission,” *Astronomy & Astrophysics*, vol. 536, no. A1, pp. 1–17, Dec. 2011.
- [116] D. Scott, M. Srednicki, and M. White, ““Sample variance” in small-scale cosmic microwave background anisotropy experiments,” *The Astrophysical Journal*, vol. 421, no. 1, pp. L5–L7, Jan. 1994.
- [117] S. M. Leach, J.-F. Cardoso, C. Baccigalupi, R. B. Barreiro, M. Betoule, J. Bobin, A. Bonaldi, J. Delabrouille, G. de Zotti, C. Dickinson, H. K. Eriksen, J. González-Nuevo, F. K. Hansen, D. Herranz, M. Le Jeune, M. López-Caniego, E. Martínez-González, M. Massardi, J.-B. Melin, M.-A. Miville-Deschênes, G. Patanchon, S. Prunet, S. Ricciardi, E. Salerno, J. L. Sanz, J.-L. Starck, F. Stivoli, V. Stolyarov, R. Stompor, and P. Vielva, “Component separation methods for the PLANCK mission,” *Astronomy & Astrophysics*, vol. 491, no. 2, pp. 597–615, Nov. 2008.
- [118] J. Delabrouille, “The Planck Sky Model,” <http://www.apc.univ-paris7.fr/~delabrou/PSM/psm.html>, 2012.
- [119] J. Delabrouille and J.-F. Cardoso, “Diffuse source separation in CMB observations,” in *Data Analysis in Cosmology*, ser. Lecture Notes in Physics, V. J. Martinez, E. Saar, E. M. Gonzales, and M. J. Pons-Borderia, Eds. Heidelberg: Springer, 2009, vol. 665, pp. 159–205.
- [120] A. Q. Carretero and S. P. Wilson, “Dependent Gaussian mixture models for source separation,” in *Proc. European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, Aug. 29 – Sep. 2 2011, pp. 1723–1727.
- [121] Planck HFI Core Team, “Planck early results. VI. The High Frequency Instrument data processing,” *Astronomy & Astrophysics*, vol. 536, no. A6, pp. 1–47, Dec. 2011.
- [122] A. Graham, *Kronecker Products and Matrix Calculus with Applications*, ser. Mathematics and its Applications. Chichester, West Sussex, England: Ellis Horwood Limited, 1981.

55	5. לכסון במשותף בבלוקים
55	5.1 מבנה הפרק
55	5.2 פיתוח השינויים היחסיים
57	5.3 אלגוריתמים
57	5.4 דוגמאות נומריות
60	5.5 סיכום
61	5.A פיתוח מסדר ראשון של משואות השערוך
65	6. הפרדה של נתונים רב מימדיים בעזרת מודל חד מימדי ומיונם לקבוצות: ניתוח ביצועים
65	6.1 מבנה הפרק
67	6.2 הצגת הבעיה
68	6.2.1 שערוך רכיבים בהנחת מודל לא מתאים
70	6.2.2 אפיון פתרונותיה של פונקציית הניגוד עבור מודל לא מתאים
73	6.3 ניתוח השגיאה
73	6.3.1 הגדרת אומד השגיאה
74	6.3.2 פירוט השגיאה לרכיביה
77	6.3.3 קירוב מסדר ראשון לשגיאה
80	6.3.4 קווריאנס השגיאה והשגיאה הריבועית הממוצעת
81	6.4 דוגמאות נומריות
82	6.4.1 לכסון במשותף בבלוקים על ידי לכסון רגיל ומיון לקבוצות
82	6.4.2 אישוש ניתוח הביצועים
85	6.4.3 אישוש הנחת שגיאות קטנות
86	6.4.4 $MSE_i^{JD}/MSE_i^{JBD}$ כפונקציה של פרמטרי המודל
87	6.5 דיון
90	6.A פיתוח מסדר ראשון של משואות השערוך
93	6.B ביטוי סגור ל $Cov(\mathbf{g}_{ij})$
97	6.C ביטוי מפורש ל $H$ ול $Cov(\mathbf{g}_{ij})$
99	7. ICA רב מימדי עם יתרות: יישום לתצפיות בקרינת הרקע הקוסמית
99	7.1 מבנה הפרק
100	7.2 מונחים בסיסיים
100	7.2.1 רקע אסטרופיזיקלי
103	7.2.2 הרמוניות ספריות
104	7.2.3 שונות קוסמית
104	7.3 מוטיבציה
105	7.4 יישום לתצפיות בשמיים
105	7.4.1 מודל הנתונים בעלי היתרות
106	7.4.2 מטריצות צפיפות התדר
108	7.4.3 תהליכים קוסמולוגים כרכיבים רב מימדיים
108	7.4.4 הורדת מימד
109	7.4.5 השגיאה הריבועית הממוצעת
111	7.5 ניסויים נומריים
111	7.5.1 בחירת סדר המודל עבור קרינת הגלקסיה
114	7.5.2 לכסון במשותף בבלוקים לעומת לכסון רגיל ומיון
115	7.5.3 השגיאה הריבועית הממוצעת: חזויה לעומת נסיונית
117	7.6 דיון
121	8. סיכום
121	8.1 סיכום התוצאות
123	8.2 תרומתנו לתחום המחקר
124	8.3 כיוונים למחקר עתידי
127	A כמה תכונות אלגבריות

# תוכן העניינים

i	תודות
iii	תקציר
ix	רשימת סימונים
xi	רשימת קיצורים
1	1. מבוא
1	1.1 הקדמה
2	1.2 סקר ספרות ותרומתנו לתחום
2	1.2.1 ביצועים אופטימליים של ICA רב מימדי מסדר שני
5	1.2.2 יכולת ההבחנה של ICA רב מימדי מסדר שני
6	1.2.3 אלגוריתמים ללכסון במשותף בבלוקים
7	1.2.4 הפרדה של נתונים רב מימדיים על ידי מודל חד מימדי
8	1.2.5 ICA רב מימדי עם יתרות: יישום לתצפיות בקרינת הרקע הקוסמית
11	1.3 מבנה העבודה
15	2. בסיס מתמטי
15	2.1 מבנה הפרק
15	2.2 מקורות תלויים סטטיסטית לעומת רכיבים רב מימדיים
17	2.3 המודל, פונקציית הנראות ופונקציית הניגוד
18	2.3.1 המודל הסטציונרי למקוטעין
18	2.3.2 פונקציית הנראות
20	2.3.3 פונקציית ניגוד לתערובת של מקורות תלויים סטטיסטית
20	2.3.4 משוואות השערוך במונחי מטריצת הערבוב
22	2.3.5 משוואות השערוך במונחי מטריצות ההיטל
22	2.4 סיכום
23	2.A פיתוח פונקציית הניגוד
25	3. ביצועים אופטימליים של ICA מסדר שני
25	3.1 מבנה הפרק
26	3.2 ניתוח השגיאה
26	3.2.1 פירוק השגיאה לרכיביה
27	3.2.2 קירוב מסדר ראשון לשגיאה
29	3.2.3 קוריאנס השגיאה במונחי מטריצות ההיטל
31	3.2.4 השגיאה הריבועית הממוצעת
32	3.3 דוגמאות נומריות
33	3.3.1 אישוש הביטוי הסגור לשגיאה הריבועית הממוצעת
34	3.3.2 אישוש הנחת שגיאות קטנות
36	3.4 דיון
37	3.A פיתוח מסדר ראשון של משוואות השערוך
39	3.B ביטוי סגור ל $\text{Cov}([\mathbf{g}_{ij}^\dagger, \mathbf{g}_{ji}^\dagger]^\dagger)$
41	3.C הוכחה של $\mathbf{H}^{-1} \mathbf{H}_\Pi \mathbf{H}^{-\dagger} = \mathbf{H}_\Pi^\#$
41	3.D ביטוי מפורש ל $\text{Cov}(\text{vec}\{E_{ij}\})$ ול $\text{Cov}(\text{vec}\{E_{ji}\})$
45	4. יכולת ההבחנה של ICA רב מימדי מסדר שני
45	4.1 מבנה הפרק
46	4.2 דרגות החופש
46	4.3 פיתוח משפט יכולת ההבחנה
48	4.4 הפיכותה של H
50	4.5 דוגמה: אי-יכולת להבחין בין רכיבים
51	4.6 דיון
52	4.A הוכחה של משפט יכולת ההבחנה



## תקציר

התחום של ניתוח מקורות בלתי תלויים סטטיסטית (ICA, Independent Component Analysis), או, בשמו האחר, הפרדה עיוורת של מקורות (BSS, Blind Source Separation), עוסק בחילוץ מספר אלמנטים בלתי תלויים סטטיסטית מתוך מדידות של צרופים לינאריים שלהם. עבודה זו מתייחסת למודל כללי יותר וגמיש יותר: ניתן לחלק את המקורות למספר קבוצות, כך שמקורות באותה קבוצה הם בעלי תלות סטטיסטית, ואילו בין קבוצות שונות אין תלות סטטיסטית. ייצוג זה שקול לעיסוק ברכיבים (components) רב-מימדיים. זאת, בניגוד ל ICA הסטנדרטי, שבו הטיפול מוגבל לרכיבים חד-מימדיים. הצורך בהרחבת המודל נובע מכך שבמגוון יישומים הנחת אי-תלות סטטיסטית בין כל הרכיבים אינה ראלית, בעוד שהנחת אי תלות סטטיסטית בין רכיבים רב ממדיים בהחלט כן.

משפך, בחרתי להתמקד בשיטות מסדר שני להפרדה של רכיבים רב מימדים בלתי תלויים סטטיסטית מתוך קבוצה של תערובות לינאריות רגועות שלהם. מטרתה של עבודה זו היא להציג תשובות תאורטיות לשאלות אשר עד כה נידונו בעיקר אמפירית. בפרט, אציג ביטוי מפורש לאומד הביצועים (במובן של השגיאה הריבועית הממוצעת) עבור הפרדת רכיבים רב מימדיים, בשני תרחישים מרכזיים: האחד הוא שיטת ההפרדה האופטימלית (בהנחת רכיבים גאוסיים). השני הוא שיטה דו-שלבית, אשר בשלב ראשון מפעילה על המדידות אלגוריתם עבור מודל חד מימדי. בשלב השני השיטה ממיינת את תוצרי האלגוריתם לרכיבים הרב-מימדיים המתאימים. אוכיח, כי הגישה השניה היא תת-אופטימלית, כלומר – שגיאת השערוך בה אינה קטנה מזו של השיטה הראשונה. בעזרת הביטוי המפורש לשגיאה ניתן לחשב את השיפור הצפוי משימוש במודל הרב מימדי (הנכון), ביחס לשימוש בגרסה החד-מימדית של המודל.

את השיטה האופטימלית אממש ע"י שני אלגוריתמים חדישים ללכסון-בבלוקים במשותף של מטריצות. בעזרת אלגוריתמים אלה ניתן להשיג את השגיאה הריבועית המינימלית הממוצעת, כאשר הנחות המודל מתקיימות. אציג תוצאות תאורטיות חדשות לגבי יכולתו של המודל להבחין בין רכיבים שונים (identifiability). תוצאות אלו ביחס ליכולת ההבחנה של המודל בין רכיבים שונים הן גם התנאי ליחידות של פתרון לבעיית הלכסון-בבלוקים במשותף של קבוצת מטריצות בלוק-אלכסוניות סימטריות חיוביות-מוגדרות, שעורבבו. לגבי השיטות שנעשה בהן שימוש בעבודה זו, נראה שניתן לבצע ולהציג את כל הפיתוחים הנדרשים במונחים של גדלים מוגדרים היטב. כך אנו נמנעים מהבעיה הידועה של אי-הודאות בהגדרה, השגורה בייצוגים של ICA ו BSS.

את השיטות והאלגוריתמים שפיתחתי אפעיל על יישום אסטרופיזיקלי של הפרדה של קרינת הרקע הקוסמית (CMB) מתוך מדידות הכוללות רכיבים נוספים, ובעיקר הגלקסיה. לשם כך אתאים את הניתוח לתחום התדר, ואפעיל את השיטות שפיתחנו על מטריצות של צפיפות תדר. ביישום זה, ארחיב את הניתוח למקרה של יתרות במספר המקלטים, ואטפל במקרה שבו מימד של רכיב אינו ידוע מראש. ביישום זה נראה התאמה טובה של הנוסחאות התאורטיות לתוצאותיהן של סימולציות נומריות.



עבודה זו נעשתה בהדרכת

**פרופ' חגית מסר**

**וֹדֵר' Jean-François Cardoso**



אוניברסיטת תל-אביב  TEL AVIV UNIVERSITY  
הפקולטה להנדסה ע"ש איבי ואלדר פליישמן  
בית הספר לתארים מתקדמים ע"ש זנדמן סליינר

**אנליזה מסדר שני**  
**של רכיבים רב-מימדיים**  
**בלתי תלויים סטטיסטית:**  
**תאוריה ושיטות**

**דנה להט**

חיבור לשם קבלת התואר "דוקטור לפילוסופיה"

הוגש לסנאט של אוניברסיטת תל-אביב

עבודה זו נעשתה באוניברסיטת תל-אביב בפקולטה להנדסה

בהדרכת פרופ' חגית מסר ודר' Jean-François Cardoso

אלול תשע"ב



TEL AVIV UNIVERSITY  אוניברסיטת תל-אביב  
הפקולטה להנדסה ע"ש איבי ואלדר פליישמן  
בית הספר לתארים מתקדמים ע"ש זנדמן סליינר

**אנליזה מסדר שני  
של רכיבים רב-מימדיים  
בלתי תלויים סטטיסטית:  
תאוריה ושיטות**

חיבור לשם קבלת התואר "דוקטור לפילוסופיה"

**דנה להט**

הוגש לסנאט של אוניברסיטת תל-אביב

אלול תשע"ב