APPROVAL SHEET

Title of Thesis: INDEPENDENT VECTOR ANALYSIS: THEORY, ALGORITHMS, AND APPLICATIONS

Name of Candidate: Matthew Anderson
Doctor of Philosophy, 2013

Thesis and Abstract Approved: ______________________
Dr. Tülay Adalı
Professor
Department of Computer Science and
Electrical Engineering

Date Approved: ____________________________

Curriculum Vitae

Name: Matthew Anderson
Degree and date to be conferred: Doctor of Philosophy, May 2013.
Collegiate institutions attended:
Johns Hopkins University, M.S. Electrical Engineering, 2006.
Cornell University, B.S. Electrical Engineering, 2000.
Ithaca College, B.A. Physics, 2000.

Professional publications:
Journal publications:

1. M. Anderson, G.-S. Fu, R. Phlypo, and T. Adalı. Independent vector analysis: Identification conditions and performance bounds. *IEEE Trans. Signal Process.* Submitted
2. M. Anderson, X.-L. Li, and T. Adalı. Complex-valued independent vector analysis: Application to multivariate Gaussian model. *Signal Process.*, (8):1821–1831, 2012. Latent Variable Analysis and Signal Separation
3. M. Anderson, T. Adalı, and X.-L. Li. Joint blind source separation of multivariate Gaussian sources: Algorithms and performance analysis. *IEEE Trans. Signal Process.*, 60(4):1672–1683, Apr. 2012
4. H. Li, M. Anderson, and T. Adalı. Kernel least mean pth power adaptive algorithm for nonlinear adaptive filtering. *IEEE Trans. Neural Netw. Learn. Syst.* Submitted
5. P. Rodriguez, M. Anderson, X.-L. Li, and T. Adalı. General non-orthogonal constrained ICA. *IEEE Trans. Signal Process.* Submitted
6. H. Li, X.-L. Li, M. Anderson, and T. Adalı. A class of adaptive algorithms based on entropy estimation achieving CRLB for linear non-Gaussian filtering. *IEEE Trans. Signal Process.*, 60(4):2049–2055, Apr. 2012
7. X.-L. Li, T. Adalı, and M. Anderson. Joint blind source separation by generalized joint diagonalization of cumulant matrices. *Signal Process.*, 91(10):2314–2322, Oct. 2011
8. X.-L. Li, M. Anderson, and T. Adalı. Noncircular principal component analysis and its application to model selection. *IEEE Trans. Signal Process.*, 59(10):4516–4528, Oct. 2011

Conference papers:

1. M. Anderson, G.-S. Fu, R. Phlypo, and T. Adalı. Independent vector analysis, the Kotz distribution, and performance bounds. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2013. Accepted
2. M. Anderson, X.-L. Li, P. Rodriguez, and T. Adalı. An effective decoupling method for matrix optimization and its application to the ICA problem. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pages 1885–1888, Mar. 2012
3. M. Anderson, X.-L. Li, and T. Adalı. Nonorthogonal independent vector analysis using multivariate Gaussian model. In *Latent Variable Analysis and Signal Separation*, volume 6365 of *Lecture Notes in Computer Science*, pages 354–361. Springer Berlin / Heidelberg, 2010
4. M. Anderson and T. Adalı. A general approach for robustification of ICA algorithms. In *Latent Variable Analysis and Signal Separation*, volume 6365 of *Lecture Notes in Computer Science*, pages 295–302. Springer Berlin / Heidelberg, 2010
5. G.-S. Fu, M. Anderson, R. Phlypo, and T. Adalı. Algorithms for Markovian source separation by entropy rate minimization. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2013. Accepted
6. T. Adalı, M. Anderson, and G. Fu. IVA and ICA: Use of diversity in independent decompositions. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 61–65, Aug. 2012
7. G. Fu, H. Li, M. Anderson, and T. Adalı. Order detection for dependent samples using entropy rate. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pages 2161–2164, 2012
8. J. T. Dea, M. Anderson, E. Allen, V. D. Calhoun, and T. Adalı. IVA for multi-subject FMRI analysis: A comparative study using a new simulation toolbox. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sept. 2011
9. 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 Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2011
10. 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)*, Beijing, China, Sept. 2011

11. X.-L. Li, M. Anderson, and T. Adalı. Second and higher-order correlation analysis of multiset multidimensional variables by joint diagonalization. In *Latent Variable Analysis and Signal Separation*, volume 6365 of *Lecture Notes in Computer Science*, pages 197–204. Springer Berlin / Heidelberg, 2010

12. X.-L. Li, , T. Adalı, and M. Anderson. Detection of circular and noncircular signals in the presence of circular white Gaussian noise. In *Proc. of 44th Asilomar Conference on Signals, Systems, and Computers*, 2010

13. C. M. Teixeira, M. Anderson, and C. H. Jaffe. Training data non-stationarity mitigation for STAP. In *MSS Tri-Service Radar Symposium*, June 2006

Professional positions held:

Northrop Grumman (August 2000 – Present)

ABSTRACT

Title of dissertation:	INDEPENDENT VECTOR ANALYSIS: THEORY, ALGORITHMS, AND APPLICATIONS
	Matthew Anderson Doctor of Philosophy, 2013
Dissertation directed by:	Professor Tülay Adalı Department of Computer Science and Electrical Engineering

The field of blind source separation (BSS) is a well studied discipline within the signal processing community due to its applicability to a variety of problems when the data observation model is poorly known or difficult to model. For example, in the study of the human brain with functional magnetic resonance imaging (fMRI), a neuroimaging sensor, BSS algorithms are able to provide medical researchers and practitioners with a decomposition of a three-dimensional 'movie' of the brain that is amenable to analysis. BSS algorithms achieve this decomposition with only a few justifiable assumptions; this is contrary to methods based on the general linear model, which require prespecified models of the expected or desired response to achieve analysis of fMRI data.

Most BSS algorithms consider just a single dataset, but it also desirable to have methods that can analyze multiple subjects or data collections in fMRI jointly, so as to provide insights beyond that achieved with individual analysis of single

datasets. Several frameworks for using BSS on multiple datasets jointly have been proposed. The subject of this dissertation is the study of one of these frameworks, which has been termed independent vector analysis (IVA). IVA is a recent extension of the classical independent component analysis (ICA) model to BSS of multiple datasets and it has been the subject of significant research interest. In this dissertation, we provide a formulation of IVA that accounts for sources which possess properties such as *a*) following Gaussian or non-Gaussian distributions; *b*) samples are independently and identically distributed (iid) or are dependent; and *c*) having either linear or nonlinear dependence of sources between datasets. The proposed IVA formulation utilizes the likelihood to define the objective function. This formulation admits to theoretical analysis. In particular, we provide the identification conditions, i.e., we determine when the sources can be 'blindly' recovered by IVA, and give a lower bound on the source separation performance.

Several algorithms exist for achieving IVA. We provide several new approaches to developing IVA algorithms and apply these approaches using a Gaussian distribution source model and a more general Kotz distribution model. The former, in addition to leading to efficient IVA algorithms, serves as the distribution model that directly connects canonical correlation analysis (CCA) and ICA.

INDEPENDENT VECTOR ANALYSIS: THEORY, ALGORITHMS, AND APPLICATIONS

by

Matthew Anderson

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Professor Tülay Adalı, Chair/Advisor
Professor Joel M. Morris
Dr. Mike Novey
Dr. Ronald Phlypo
Professor Anindya Roy

© Copyright by
Matthew Anderson
2013

To my wife Anna, my sons Kariso and Taloosh, my parents, and my loving family.

Acknowledgments

It is not possible for me to express my gratitude to all those who have made this dissertation a reality—but I will try.

My advisor, Professor Tülay Adalı, has provided a wonderful environment for me to ask questions, pursue answers, and learn about more ideas and concepts in this pursuit than I thought was possible. Her guidance, generosity, and constant encouragement have been, and will continue to be, deeply appreciated.

I would like to thank the postdoctoral associates of Dr. Adali's lab for the many hours of discussions and white-board sessions. I thank Dr. Xi-Lin Li for his patience with me while beginning my research, Dr. Hualiang Li for listening to my ideas no matter how poorly I explained them, and to Dr. Ronald Phlypo for always making research a collaborative experience.

My fellow researchers at the lab have made this an adventure and taught me to be open to different perspectives than my own—for this I am truly appreciative.

I would like to express gratitude to my committee members Dr. Mike Novey, Professor Joel Morris, and Professor Anindya Roy.

I owe my deepest thanks to my family—my mother and father who gave me everything, my forgiving and loving wife, and my sweet sons.

Table of Contents

List of Figures vi

List of Abbreviations viii

1 Introductions 1
  1.1 Motivation 1
  1.2 Contributions 6
  1.3 Overview of Dissertation 7
  1.4 Mathematical Preliminaries 10

2 IVA Theory 13
  2.1 Blind Source Separation Framework 14
  2.2 Problem Formulation 17
  2.3 Objective Functions 18
    2.3.1 Dependent Samples 19
    2.3.2 iid Samples 23
    2.3.3 Gaussian Sources and Canonical Correlation Analysis 24
  2.4 Fisher Information Matrix 25
  2.5 Identification Conditions 28
    2.5.1 Special Cases 34
  2.6 Cramér-Rao Lower Bound 35
    2.6.1 CRLB with iid Samples 37
      2.6.1.1 Elliptical Distribution 39
      2.6.1.2 Kotz Distribution 44
      2.6.1.3 Multivariate Power Exponential Distribution 46
      2.6.1.4 Gaussian Distribution 49
    2.6.2 Example of Performance Bounds 50
    2.6.3 CRLB for ICA 52
  2.7 Performance Metrics 53
    2.7.1 Interference-to-Signal Power Ratio 53
    2.7.2 Intersymbol Interference 56
    2.7.3 Separation Success or Failure 61

3 IVA Algorithms 62
  3.1 Review of Existing Algorithms 63
    3.1.1 Linear Dependence 63
    3.1.2 Nonlinear Dependence 64
    3.1.3 Sample-to-Sample Dependence 65
  3.2 Optimization of IVA Objective Function 66
    3.2.1 Matrix (Natural) Gradient Descent 67
    3.2.2 Nonorthogonal Decoupling Trick 69
    3.2.3 Vector Gradient Descent 73
    3.2.4 Newton Update 75
    3.2.5 Block-Newton Update 76
  3.3 Optimization for Assumed Source Distribution Models 77

3.3.1 Multivariate Gaussian Distribution – IVA-Gauss 78
3.3.1.1 Matrix Gradient – IVA-G-M 78
3.3.1.2 Vector Gradient – IVA-G-V 79
3.3.1.3 Newton Update – IVA-G-N 81
3.3.1.4 Block Newton Update – IVA-G-B 83
3.3.2 IVA-Kotz 84
3.4 Simulations 85
3.4.1 Multivariate Gaussian Sources 85
3.4.2 Multivariate Non-Gaussian Sources 89
3.4.3 Orthogonal Generalized Joint Diagonalization with Second-Order Lags 98

4 Complex-Valued IVA 100
4.1 Problem Formulation 102
4.2 Likelihood 103
4.3 Review of Existing Algorithms 104
4.4 Nonunitary Decoupling 106
4.4.1 Vector Gradient Descent 108
4.4.2 Newton Update 110
4.4.3 Noncircular Gaussian Distribution 112
4.4.4 Circular Multivariate Gaussian Distribution 118
4.5 Simulations 118
4.5.1 Multivariate Gaussian Sources 119
4.5.2 Multivariate Non-Gaussian Sources 124

5 Conclusions and Future Work 127
5.1 Conclusions 127
5.2 Future Work 129

A IVA Fisher Information Matrix 132

B Decoupled ICA 138
B.1 Decoupled Maximum Likelihood ICA 139
B.1.1 Algorithm Details 142
B.1.2 Algorithm Performance 143
B.2 Conclusions 144

C Canonical Correlation Analysis 145
C.1 CCA 145
C.2 MCCA 147

D Online IVA 149
D.1 Algorithm 149
D.2 Simulation Results 151

E Kotz Sample Generation 156

Bibliography 162

List of Figures

2.1 Example IVA Fisher information matrix, for $N = 3$ sources. 27
2.2 Measure of non-Gaussianity for example Kotz distributions. 47
2.3 IVA performance bounds using three types of source diversity. 51

3.1 Comparison of convergence for various IVA-Gauss algorithms. 86
3.2 Comparison of different second-order algorithms for Gaussian sources. 87
3.3 Comparison of computational speed of different second-order algorithms. 88
3.4 Comparison of Newton IVA-G algorithm with performance bound. . . 89
3.5 Comparison of separation performance for various IVA algorithms with linearly combined GGD sources. 90
3.6 Comparison of separation performance for various IVA algorithms with second-order correlated Laplacian sources. 92
3.7 Comparison of performance of IVA-G and IVA-L with Laplacian sources versus a measure of the amount of second-order correlation. . 94
3.8 Comparison of success rate of IVA-MPE as the number of samples is varied. 96
3.9 Comparison of IVA-MPE source separation performance with exact knowledge of the shape parameter as compared the performance bound. 96
3.10 Comparison of IVA-MPE source separation performance without exact knowledge of the shape parameter as compared the performance bound. 97
3.11 Comparison of IVA-MPE source separation performance with the performance bound using the exact shape parameter. 98
3.12 Comparison of source separation performance of generalized joint diagonalization with the performance bound that accounts for sample dependence in IVA. 99

4.1 Comparison of convergence for various complex-valued IVA-Gauss algorithms. 120
4.2 Comparison of different second-order algorithms for complex-valued Gaussian sources. 121
4.3 Comparison of the source separation performance versus K, with $K{\cdot}T$ fixed. 122
4.4 Example showing that when a subset of the sources are nonidentifiable, the remaining sources can still be identified. 123
4.5 Comparison of source separation performance for versus the degree of noncircularity. 124
4.6 Comparison of source separation performance for various IVA algorithms with linearly combined GGD sources. 125

B.1 Comparison of source separation performance of D-ICA with two other ICA algorithms. 143

D.1 Comparison of source separation performance on the latest update for online IVA-Gauss for various number of datasets. 152
D.2 Comparison of number of iterations required on the latest update for online IVA-Gauss for various number of datasets. 153
D.3 Comparison of source separation performance for online IVA-Gauss for various memory factors. 154
D.4 Comparison of number of iterations required on last update for online IVA-Gauss for various memory factors. 154
D.5 Comparison of source separation performance for various memory factors in online IVA-Gauss with time varying mixing matrices. . . . 155
D.6 Comparison of number of iterations required on the latest update for various memory factors in online IVA-Gauss with time varying mixing matrices. 155

E.1 Empirical histograms and associated pdfs with varying shape parameters. 160
E.2 Empirical histograms and associated pdfs with varying hole parameters. 161

List of Abbreviations & Acronyms

BOLD blood oxygenation level dependent

BSS blind source separation

CCA canonical correlation analysis, see [49]

CRLB Cramér-Rao lower bound

D-ICA decoupled independent component analysis algorithm, see [13]

EBM entropy bound minimization, see [71, 74]

ECG electrocardiogram

EEG electroencephalography

EFICA efficient fast independent component analysis algorithm, see [60]

FIM Fisher information matrix

FastICA fastICA algorithm, see [50]

GGD generalized Gaussian distribution, also known as generalized normal distribution

ICA independent component analysis

ISI intersymbol-interference

ISR interference to source ratio

IVA-G-B IVA-G cost function optimize using block quasi-Newton algorithm

IVA-G-N IVA-G cost function optimize using Newton algorithm

IVA-G-V IVA-G cost function optimize using vector gradient descent approach

IVA-GL IVA algorithm using IVA-Gauss solution to initialize IVA-Lap

IVA-Gauss IVA with multivariate Gaussian distribution model

IVA-Lap IVA with second-order uncorrelated multivariate Laplace distribution model

IVA-MPE IVA with multivariate power exponential distribution model

IVA-NC-G IVA with non-circular multivariate Gaussian distribution model

IVA independent vector analysis

JBSS joint blind source separation

JDIAG-SOS joint diagonalization via second-order statistics, see [76, 75]

MCCA multiset canonical correlation analysis, see [53, 79]

MICA multidimensional independent component analysis, see [23]

MLE maximum likelihood estimate

MPE multivariate power exponential distribution, also known as MGGD

NDT nonorthogonal decoupling trick

SCM source component matrix, i.e., $\mathbf{S}_n = \left[\mathbf{s}_n^{[1]}, \ldots, \mathbf{s}_n^{[K]}\right]^{\mathsf{T}}$

SCV source component vector, i.e., $\mathbf{s}_n = \left[s_n^{[1]}, \ldots, s_n^{[K]}\right]^{\mathsf{T}}$

SNR signal to noise ratio

fMRI functional magnetic resonance imaging

iCRLB induced Cramér-Rao lower bound

sMRI structural magnetic resonance imaging

iff if and only if

iid independently and identically distributed

pdf probability density function

wrt with respect to

Chapter 1

Introductions

“You cannot open a book without learning something.” – Confucius

1.1 Motivation

We begin by stating our general problem: identify the distinct ‘original signals’ that have been sampled and then observed after some unknown transformation is applied. The observations may also be corrupted by noise and the transformation could change from sample to sample, i.e., might be a time varying transformation.

The above problem statement is well known to be the domain of blind source separation (BSS), a field of machine learning and signal processing research, which has now been studied and applied to real-world problems for several decades [51, 28]. Henceforth, we will refer to this as the BSS problem and the ‘original signals’ are more properly termed latent variables. In truth, although less well known or acknowledged, the BSS problem actually has roots dating all the way back to the seminal work of Hotelling, [49]. Hotelling proposed the canonical correlation analysis (CCA) framework, which we will show later to be a solution for a special case of the BSS problem.

The applications that can be solved using BSS techniques are vast, e.g., analysis of brain activities using either (or both) functional magnetic resonance imaging (fMRI) [84] and electroencephalography (EEG) [82] data, separation of audio [105], feature extraction from images [52], and in (antenna) array processing [25]. We have two principal applications which motivate our examination of BSS techniques.

The first, involves analyzing fMRI datasets collected on multiple subjects. The acquisition of brain activity maps from fMRI has been the subject of much research in the past fifteen years [101]. FMRI observes brain function spatially within a volume at the millimeter scale and temporally at nominally once per second. The ability to observe brain function and acquire activity maps from a sequence of volumetric images of the brain taken at a specific acquisition frequency derives from the sensitivity of the MRI scanner to the magnetic properties of blood as it becomes oxygenated (diamagnetic) and deoxygenated (paramagnetic). This sensitivity results in blood oxygenation level dependent (BOLD) contrast in the voxels collected over time. Insight about brain function derived from these BOLD contrast signals are complicated by many factors, such as the presence of subject motion, constant biological processes (heartbeat and breathing), low signal to noise ratio (SNR), random noise, and inter-subject anatomical variabilities to name a few [101]. This combination of challenges in addition to the difficulty in visualizing volumetric sequences is at the root of why so many processing techniques have been proposed, developed, and implemented to aid the medical practitioner in the analysis of fMRI. A variety of BSS techniques have been shown to be effective data-driven approaches for identifying meaningful brain activity maps from fMRI datasets, [21, 65, 79, 100]. In contrast to the model-driven approaches typically based on the general linear model which specifically model the expected BOLD response based on the experiment paradigm, the BSS approaches for analyzing fMRI data imposes nearly no assumptions about the nature of the time courses. For independent component analysis (ICA) in particular, the observed fMRI data is assumed to be a linear (invertible) mixture of *statistically independent* spatial maps. A main advantage of the data-driven approach is that in experiments with complicated time courses, e.g., simulated driving, and experiments with no predictable time courses, e.g., resting conditions, meaningful spatial brain activity maps can still be estimated. One general technique for fMRI analysis is to consider datasets collected from multiple subjects jointly or simultaneously. A natural argument for this technique is that the fundamental or common properties shared by the brains of most subjects is of high interest due to their stronger reliability. By considering multiple datasets together there is an inherent ability to combat the low SNR of the BOLD contrast/signal, intersubject variability, and to a lesser extent residual artifacts due to subject motion and the effects of background processes which remain after preprocessing steps.

Our second application is data fusion. Data fusion is a diverse field that means many things to many people. For this work, we consider the fusion of biomedical data collected from multiple patients using different and distinct sensors. The purpose of doing so is to exploit the complementary nature of the collecting sensor properties in a manner that improves data or medical analysis. Specifically, we have the data fusion as formulated by Correa [29], where features extracted from each subject in analysis of fMRI datasets are combined with those from EEG and structural magnetic resonance imaging (sMRI) datasets. The key set of complementary properties being combined by such data fusion are the poor spatial resolution and high temporal resolution of EEG combined with high spatial resolution and low temporal resolution of fMRI and sMRI. We aim with this work to provide some tools which will be useful in achieving this difficult task of combining such disparate data sources.

Both of the motivating applications above have both been solved using CCA. Multidataset extensions of CCA are summarized in [53] and has been termed multiset canonical correlation analysis (MCCA). MCCA utilizes a number of cost functions based on second-order statistics. Both CCA and MCCA have been used in other applications, e.g., hyperspectral data analysis [90], blind equalization [115].

We have found that both applications can be studied using a BSS technique called independent vector analysis (IVA) [58]. The formulation of IVA is an extension of classical ICA to multiple datasets which assumes a source within one dataset is dependent on at most one source in another dataset while sources within a dataset are mutually independent (as in ICA). Using the IVA framework, one can exploit the dependencies of sources across datasets potentially resulting in performance beyond what is achievable with single-set ICA algorithms applied to each dataset individually. Additionally, IVA potentially provides an automatic method for 'aligning' dependent sources across the datasets. Using ICA algorithms individually does not provide such an alignment. Rather, an additional algorithm is used to minimize the amount of local or internal permutation ambiguity, e.g., see [31, 80].

The formulation of IVA can also be applied to single dataset problems in multiple ways. For example, CCA has been successfully used to achieve BSS with the kernel ICA approach of [15], which considers a representation of the sources in the reproducing kernel Hilbert space where CCA is then applied. A different method for using CCA to achieve BSS is to exploit the sample-to-sample dependence present in a single subject fMRI dataset as proposed in [39]. Another recent application of IVA using an extension of diagonalization methods to multiple datasets is presented in [75] for the separation of the weak fetal electrocardiogram (ECG) heartbeat signal from the stronger but slower ECG heartbeat of the mother.

IVA as originally posed and implemented is limited by several assumptions, namely the source dependency across datasets are assumed to be second-order uncorrelated, to follow a multivariate Laplacian distribution (implying super-Gaussian sources), and the cost function assumes independently and identically distributed (iid) samples. Several questions naturally arise from the IVA formulation above:

- What if sources possess dependence across datasets that are linear, i.e., the cross-correlation is not zero?
- What if the sources are Gaussian or even sub-Gaussian?
- What if the samples are not iid?
- Under what conditions can the sources be identified?
- What are the performance bounds on the errors in the estimation of these sources?
- Under what conditions can the sources that possess dependence across datasets be aligned?

Seeking answers to these questions motivated our desire to expand the IVA algorithms for more general sources and to provide the domain of IVA with insightful and useful theoretical analysis.

1.2 Contributions

The contributions of the thesis are primarily theoretical and algorithmic.

Our theoretical contributions first extend the definition of IVA as originally proposed for iid samples to include sample-to-sample dependence. For both the iid IVA and the more generalized IVA model we give the identification conditions, i.e., what are the prerequisites to recovering the 'original' sources. We have also posed IVA as a maximum likelihood problem, which enables the performance bounds for IVA to be determined via Cramér-Rao lower bound (CRLB) analysis. From the theoretical analysis, the connections of the IVA results with prior work for ICA are discussed and the analogous roles of sample dependency for ICA and dataset dependency for IVA are illuminated.

Algorithms are given for achieving IVA when the samples are iid. We have utilized a novel decoupled optimization procedure to implement vector gradient and Newton-based algorithms. The algorithms are applied assuming the sources ('original' signals) are Gaussian. More flexible models for the sources are permitted by introducing the use of the Kotz distribution for IVA.

Additional algorithms are also given for complex-valued data generated by sources that are possibly noncircular. In fact, our noncircular IVA solution provides the first complex-valued linear (as opposed to widely-linear, [102]) CCA solution. More precisely, the noncircular Gaussian source model is able to achieve CCA for sources which heretofore was not possible.

In the process of developing algorithms for achieving IVA, we found a new quasi-Newton update procedure for ICA. The algorithm uses a nonorthogonal decoupling trick.

To promote understanding, utilization, and future research we have published MATLAB code [83] that implements the IVA algorithms, the decoupling procedure, and the decoupled ICA algorithm for the public at `http://mlsp.umbc.edu/resources.html`.

1.3 Overview of Dissertation

In Chapter 2, we provide two frameworks which equivalently define IVA. The first, casts IVA as a subset of the broader BSS family of problems described previously; this perspective is new. The more common perspective of IVA as a generalization of the BSS problem from one dataset to multiple datasets is also discussed. We then proceed to develop the objective (cost) function for IVA when the observation samples are iid and for the more general case when the samples are not iid. The equivalence between the objective functions given by likelihood maximization, minimization of entropy (rate), and minimization of mutual information (rate) are established for IVA. We also explain how IVA is the connection which bridges the gap between ICA and CCA—an overview of CCA and its multidataset extension, MCCA, are provided in Appendix C. The form of the Fisher information matrix (FIM) associated with the IVA likelihood function is given—derivation details are reserved for Appendix A. The FIM is used to prove that IVA is identifiable, to determine the identification conditions, and to give the performance bounds via a CRLB analysis. The identification conditions and the CRLBs are shown to be the multivariate extensions of the ICA identification conditions and the CRLBs. Furthermore, we can simplify the performance bounds expressions considerably when the samples are iid and the sources follow (multivariate) elliptical distributions. We conclude the chapter with a discussion on performance metrics for IVA.

In Chapter 3, we discuss IVA algorithms. We begin by reviewing existing algorithms and placing the algorithms into one of three categories. The categories differentiate based on the type of dataset dependencies that are assumed or exploited by the algorithms. We then focus on optimization of the IVA objective function when the samples are iid. We proceed by motivating and deriving a decoupling procedure which allows optimization of vectors rather than matrices and permits gradient descent, Newton, and quasi-Newton optimization algorithms. In fact, this decoupling procedure is used to derive a new quasi-Newton algorithm for ICA, which we describe in Appendix B. The decoupling algorithms are derived without assuming any particular source distribution model. By assuming that the source model follows a multivariate Gaussian distribution we then develop computationally efficient methods for achieving IVA of linearly dependent sources. The use of the Kotz distribution as a source model is introduced by us. This source model greatly increases the flexibility of the sources which can be accurately separated by

IVA algorithms under the iid assumption. We conclude the chapter by presenting simulations which demonstrate the benefits of the nonorthogonal decoupling procedure, the computational efficiency of the Newton and quasi-Newton algorithms, and compare the achieved source separation performance with the theoretical bounds. All of the algorithms in Chapter 3 assume batch processing. In Appendix D, we give some preliminary online processing algorithms for IVA.

In Chapter 4, we consider the case when the observations are complex-valued. In the previous chapters, the analysis and algorithms assume the observations are real-valued. Complex-valued data is frequently encountered in many applications. It is desirable for the processing of the data to be performed, "...in such a way that the complete information—represented by the interrelationship of the real and imaginary parts . . . of the signal—can be fully exploited" [2]. We achieve such processing by using Wirtinger calculus. Wirtinger calculus enables working completely in the complex domain for the derivation and analysis of IVA—and avoids unnecessary complications related to transformations from the complex to the real domain, i.e., $\mathbb{C}^N \mapsto \mathbb{R}^{2N}$. By using Wirtinger calculus we observe that the algorithm derivations for complex-valued IVA closely parallel the real-valued case. The complex-valued IVA problem is formulated, the objective function is defined, and algorithms using the complex version of the nonorthogonal decoupling procedure are described and analyzed via simulation. By assuming the source model follows the complex noncircular multivariate Gaussian distribution, we can develop computationally efficient methods for achieving IVA of linearly dependent complex-valued sources. The connection between CCA and IVA implies that IVA with the noncircular Gaussian source model extends the domain of *linear* CCA to include noncircular sources.

Chapter 5 provides a summary of conclusions to the dissertation and suggestions for future work.

1.4 Mathematical Preliminaries

A review of the mathematical notations and conventions are provided in the following before proceeding to the next chapter.

For this document, the number sets (domains) are restricted to the sets of nonnegative natural ($\mathbb{N}$), real ($\mathbb{R}$), and complex ($\mathbb{C}$) numbers. Matrices and vectors from each domain are indicated by $\mathbb{D}^{N \times N}$ and $\mathbb{D}^{N}$, respectively, where $\mathbb{D}$ denotes any number set. When no particular number set is assumed then we use $\mathbb{D}$. Scalar, (column) vector, and matrix quantities are denoted as lower-case light face, lower-case bold face, and upper-case bold face, respectively. The mth element of a vector $\mathbf{v}$, $[\mathbf{v}]_m$, and an element in the mth row and nth column of a matrix $\mathbf{A}$, $[\mathbf{A}]_{m,n}$, are often denoted v_m and $a_{m,n}$, respectively.

The Kronecker delta, $\delta_{m,n}$, is one when $m = n$ and zero otherwise. The mth entry of the standard basis vector, $\mathbf{e}_n$, is $[\mathbf{e}_n]_m = \delta_{m,n}$. The $\mathbf{0}$ and $\mathbf{1}$ denote matrices (or vectors) with all entries of zeros and ones, respectively, where the dimensions of the matrices are either known from the context or indicated by an additional subscript.

The superscripts $*$, T, and $\dagger$ denote the complex conjugate, matrix transpose, and the complex conjugate matrix transpose operations, respectively. The element-wise (Hadamard) product, element-wise division, and Kronecker products are denoted by $\mathbf{A} \circ \mathbf{B}$, $\mathbf{A} \oslash \mathbf{B}$, and $\mathbf{A} \otimes \mathbf{B}$, respectively.

We use $\operatorname{vec}(\mathbf{A}) \in \mathbb{R}^{MN} = \sum_{n=1}^{N} \mathbf{e}_n \otimes (\mathbf{A}\mathbf{e}_n)$, where $\mathbf{e}_n \in \mathbb{R}^N$, to compactly denote the stacking of the columns of $\mathbf{A} \in \mathbb{R}^{M \times N}$. Additionally, let a subset of the rows in $\mathbf{A}$ be listed by the vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_d]^{\mathsf{T}} \in \mathbb{N}^d$, where $0 \leq d \leq M$. Then we use a corresponding indexing matrix $\mathbf{E}_{[\boldsymbol{\alpha}]} = [\mathbf{e}_{\alpha_1}, \ldots, \mathbf{e}_{\alpha_d}]^{\mathsf{T}} \in \mathbb{R}^{d \times K}$ to select the subset of rows in $\mathbf{A}$ indicated by $\boldsymbol{\alpha}$ via $\mathbf{E}_{[\boldsymbol{\alpha}]}\mathbf{A}$. For compactness we use $\operatorname{vec}_{\boldsymbol{\alpha}}(\mathbf{A}) \triangleq \operatorname{vec}\left(\mathbf{E}_{[\boldsymbol{\alpha}]}\mathbf{A}\right)$. The complementing subset of $\boldsymbol{\alpha}$ is indicated by $\boldsymbol{\alpha}^c \in \mathbb{N}^{M-d}$.

A diagonal matrix with entries given by $\mathbf{d}$ is denoted by $\operatorname{Diag}(\mathbf{d}) = \sum_{n=1}^{N} \mathbf{e}_n \mathbf{e}_n^{\mathsf{T}} \mathbf{d} \mathbf{e}_n^{\mathsf{T}}$, and thus the identity matrix, $\mathbf{I}_N = \operatorname{Diag}(\mathbf{1}_N)$. The square matrix, $\mathbf{A}$, has diagonal entries, $\operatorname{diag}(\mathbf{A}) = \sum_{n=1}^{N} \mathbf{e}_n \mathbf{e}_n^{\mathsf{T}} \mathbf{A} \mathbf{e}_n$, a trace, $\operatorname{tr}(\mathbf{A}) = \sum_{n=1}^{N} [\operatorname{diag}(\mathbf{A})]_n$, and a determinant, $\det(\mathbf{A})$. We indicate $\mathbf{A} - \mathbf{B}$ is positive definite using $\mathbf{A} \succ \mathbf{B}$ and positive semidefinite with $\mathbf{A} \succeq \mathbf{B}$. The $|\cdot|$ denotes the magnitude.

For a matrix $\mathbf{A}$ with block structure, the matrix $\mathbf{A}_{m,n}$ is the mth row and nth column in the block representation of the matrix $\mathbf{A}$ using M row partitions and N column partitions. The special block diagonal matrix is necessarily a square matrix (implying $M = N$) that has off-diagonal partitions being zero, i.e., $\mathbf{A}_{m,n} = \mathbf{0}$ for $1 \leq m \neq n \leq M$, and is denoted with the *direct sum* notation, $\mathbf{A} = \mathbf{A}_{1,1} \oplus \mathbf{A}_{2,2} \oplus \ldots \oplus \mathbf{A}_{M,M} = \oplus \sum_{m=1}^{M} \mathbf{A}_{m,m}$.

The common functions of random variables such as the expectation operator, entropy[1], and mutual information are denoted using $E\{\cdot\}$, $\mathcal{H}\{\cdot\}$, and $\mathcal{I}\{\cdot\}$, respec-

---

[1] Since discrete-valued variables are not considered in this work, we refer to differential entropy as simply entropy. Furthermore, $\mathcal{H}\{\mathbf{x}\} = -E\{\log p(\mathbf{x})\}$ is the entropy of $\mathbf{x}$, where $p(\mathbf{x})$ is the probability density function (pdf) for $\mathbf{x}$.

tively. For mutual information if there is one nonscalar argument then we mean the mutual information of all components/entries of the argument, e.g., $\mathcal{I}\left\{\mathbf{x}\right\} = \sum_i \mathcal{H}\left\{x_i\right\} - \mathcal{H}\left\{\mathbf{x}\right\}$. Otherwise, there are multiple arguments and the mutual information is given by $\mathcal{I}\left\{\mathcal{X};\mathcal{Y};\ldots;\mathcal{Z}\right\} = \mathcal{H}\left\{\mathcal{X}\right\}+\mathcal{H}\left\{\mathcal{Y}\right\}+\ldots+\mathcal{H}\left\{\mathcal{Z}\right\}-\mathcal{H}\left\{\mathcal{X},\mathcal{Y},\ldots,\mathcal{Z}\right\}$, where $\mathcal{X}$, $\mathcal{Y}$, $\ldots\mathcal{Z}$ are any random variables, vectors, matrices, or processes. The Kullback-Leibler divergence metric for two continuous distributions, p and q, is denote by $\mathcal{D}_{KL}\left\{p||q\right\}$. We use $\mathcal{X} \perp\!\!\!\perp \mathcal{Y}$ to denote that $\mathcal{X}$ is independent of $\mathcal{Y}$. A random vector $\mathbf{x}$ following the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$ is denoted $\mathbf{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{\Sigma}\right)$.

We use standard elementary functions such as $\log\left(\cdot\right)$, $\exp\left(\cdot\right)$, $\Gamma\left(\cdot\right)$ for the natural logarithm, the anti-logarithm, and the complete gamma function. We reserve e for $\exp\left(1\right)$.

We use $\mathcal{O}$ to denote the order of (computational) operations.

Chapter 2

IVA Theory

> "Life is really simple, but we insist on making it complicated."
>
> – Confucius

The recent interest in IVA is motivated by various application domains such as when analyzing multisubject datasets in biomedical studies using fMRI or EEG data and when solving the convolutive ICA problem in the frequency domain using multiple frequency bins. We begin this chapter by casting IVA as special case of the very general BSS problem introduced in Chapter 1. This explanation makes the connections between IVA and CCA with other classes of BSS problems clear. We then proceed with the more intuitive definition of IVA as an extension of ICA to multiple datasets. The IVA objective function to be optimized is given by the likelihood function. The likelihood function is related to the mutual information (rate) and entropy (rate) as the number of observation samples approaches infinity. Of particular interest for us is to determine the conditions required for identification of sources and the bounds on performance. For a real-valued problem it more recently shown that independent sources can be blindly identified up to a permutation and scaling ambiguity when no two sources are Gaussian with proportional sample-to-sample correlation matrices [28, Chapter 4]. We show that IVA possesses an additional type of diversity—chiefly, dependence of sources between datasets—which can be exploited for identifying sources that cannot be identified using ICA.

2.1 Blind Source Separation Framework

Clearly, the BSS problem formulation given in the introduction is very general. So general in fact that it does not permit much analysis without the imposition of additional assumptions for a solution to become realizable. Appropriate assumptions about the nature of the sources, the noise, and the transformation can lead to a tractable subclass of BSS problems, the most famous of which is ICA.

A more recent development, termed joint blind source separation (JBSS), has been noted to be a generalization of the BSS problem for multiple datasets [79, 113]. Although this is one valid perspective, here we offer an alternative view in which we consider the IVA formulation of JBSS as a special case of the BSS problem stated in the introduction.

We start by formalizing the broad BSS problem mathematically and then enumerate the assumptions that systematically lead us to the IVA formulation. As we enumerate the assumptions used, we can readily observe the limits of IVA. Formulated mathematically we have,

$$\mathbf{x}\left(t\right) = \mathcal{A}\left(\mathbf{s}\left(t\right), \mathbf{n}\left(t\right), t\right), \tag{2.1}$$

where $\mathbf{s}\left(t\right) = \left[s_1\left(t\right), \ldots, s_M\left(t\right)\right]^{\mathsf{T}}$ is tth sample of the M latent variables (which we call sources for brevity), $\mathcal{A}\left(\cdot, \cdot, t\right) : \mathbb{D}^M \mapsto \mathbb{D}^{M'}$ is the unknown mapping of the noise and sources which could change over the samples, and $\mathbf{n}\left(t\right) \in \mathbb{D}^M$ is the tth sample of corrupting noise.

The most common BSS assumptions are:

Assumption 1: Transformation is independent of the sample index t, i.e., $\mathcal{A}(\cdot,\cdot,t) = \mathcal{A}(\cdot,\cdot)$

Assumption 2: No corrupting noise, i.e., $\mathcal{A}(\cdot,\cdot) = \mathcal{A}(\cdot)$

Assumption 3: Transformation is linear in $\mathbf{s}(t)$, i.e., $\mathcal{A}(\mathbf{s}(t)) = \mathbf{A}\mathbf{s}(t)$

Assumption 4: Rank of $\mathbf{A}$, M'', is at least equal to M

At this point the BSS problem has been reduced to a fixed noiseless linear unknown mapping of M sources to an M'' dimensional space which is at least as large as M. Without the presence of noise it is a simple matter to determine the M dimensional subspace of M'' which the sources reside in, we can without loss of generality add the assumption that $M'' = M$, i.e., $\mathbf{A}$ is a square invertible matrix. The motivations for these assumptions are for tractability and are justified for many real-world applications. The weakest assumption is the noise-free assumption. If additive noise is considered the identification of the sources—with noise removal or reduction—requires additional assumptions.

We have yet to make any assumptions about the sources. If we make just one assumption, we arrive at ICA. Namely, assume that the sources are statistically independent, i.e., $p(\mathbf{s}(t)) = \prod_{m=1}^{M} p_m(s_m(t)) \ \forall t$.

For IVA, dependency within $\mathbf{s}(t)$ is permitted, at least in a structured way by using the following assumptions about the sources:

Assumption 5: $\exists N \in \mathbb{N}, K \in \mathbb{N} : M = NK$

Assumption 6: A matrix can be formed from the entries of $\mathbf{s}(t)$ such that the rows are statistically independent, i.e., $\exists \mathbf{S}(t) = [\mathbf{s}_1(t), \ldots, \mathbf{s}_N(t)]^\mathsf{T} \in \mathbb{D}^{N \times K} : p(\mathbf{s}(t)) = \prod_{n=1}^{N} p_n(\mathbf{s}_n(t)) \;\; \forall t$

Using the above assumptions we have specified a generalization of ICA known as multidimensional independent component analysis (MICA)[1][23] . Upon first reading, the nature of the assumptions made about the sources above seems contrived. The justification for these assumptions becomes clearer after applying two more assumptions to the mixture process, namely:

Assumption 7: $\mathbf{A} = \oplus \sum_{k=1}^{M} \mathbf{A}^{[k]}$, where $\mathbf{A}^{[k]} \in \mathbb{D}^{N \times N}$

Assumption 8: $\mathbf{x}(t) = \left(\oplus \sum_{k=1}^{M} \mathbf{A}^{[k]}\right) \mathrm{vec}(\mathbf{S}(t))$

The above description of the problem at hand is admittedly cumbersome. We use it to assert that IVA can be viewed as a subclass of BSS and to connect IVA to other BSS frameworks more directly. First, the IVA formulation is clearly a special case of MICA. Second, if $K = 1$, which implies $M = N$, then we also have that IVA is equivalent to ICA. Furthermore, if we have that $K = 2$ then the assumptions are equivalent to the CCA formulation, i.e., CCA is a special case of BSS.

For reasons which become clearer when we provide the original description of the IVA problem as an extension of BSS for multiple datasets given in Section 2.2, we call K the number of datasets. Considering IVA as either a special case of

---

[1] Technically, we have enumerated a special case of MICA. For additional information about identifiability of MICA see [62].

BSS or as extension of BSS to multiple datasets can be useful depending on one's perspective.

2.2 Problem Formulation

A more natural construction of the IVA problem is now utilized. IVA is an extension of BSS to multiple datasets. The original formulation of IVA in [57] assumes the samples are iid. We extend the formulation here to permit sample dependency, i.e., samples are not iid.

There are K datasets, each containing T samples, formed from the linear mixture of N independent sources,

$$\mathbf{X}^{[k]} = \mathbf{A}^{[k]}\mathbf{S}^{[k]} \in \mathbb{R}^{N\times T}, 1 \leq k \leq K.$$

The entry in the nth row and tth column of $\mathbf{S}^{[k]}$ is $s_n^{[k]}(t)$, the nth row of $\mathbf{S}^{[k]}$ is denoted with the column vector $\mathbf{s}_n^{[k]} = \left[s_n^{[k]}(1), \ldots, s_n^{[k]}(T)\right]^\mathsf{T} \in \mathbb{R}^T$, and the tth column of $\mathbf{S}^{[k]}$ is denoted by the column vector $\mathbf{s}^{[k]}(t) = \left[s_1^{[k]}(t), \ldots, s_N^{[k]}(t)\right]^\mathsf{T} \in \mathbb{R}^N$. The invertible mixing matrices, $\mathbf{A}^{[k]} \in \mathbb{R}^{N\times N}$, and the sources $\mathbf{S}$ are unknown real-valued quantities to be estimated. The source matrices in each dataset can be concatenated to form $\mathbf{S} = \left[\left(\mathbf{S}^{[1]}\right)^\mathsf{T}, \ldots, \left(\mathbf{S}^{[K]}\right)^\mathsf{T}\right]^\mathsf{T} \in \mathbb{R}^{NK\times T}$. Using this notation, we can denote the IVA data model with a single equation, namely $\mathbf{X} = \mathbf{A}\mathbf{S}$, where $\mathbf{A} = \oplus \sum_{k=1}^{K} \mathbf{A}^{[k]}$. The nth source component matrix (SCM), $\mathbf{S}_n = \left[\mathbf{s}_n^{[1]}, \ldots, \mathbf{s}_n^{[K]}\right]^\mathsf{T} \in \mathbb{R}^{K\times T}$, is independent of all other SCMs. Then the pdf of the concatenated source vector, $\mathbf{S}$, can be written as $p(\mathbf{S}) = \prod_{n=1}^{N} p_n(\mathbf{S}_n)$.

The IVA solution finds K demixing matrices and the corresponding source estimates for each dataset, with the kth ones denoted as $\mathbf{W}^{[k]}$ and $\mathbf{Y}^{[k]} \triangleq \mathbf{W}^{[k]}\mathbf{X}^{[k]} = \mathbf{W}^{[k]}\mathbf{A}^{[k]}\mathbf{S}^{[k]}$, respectively. The estimate of the nth component from the tth sample of the kth dataset is given by $y_n^{[k]}(t) = \left(\mathbf{w}_n^{[k]}\right)^{\mathsf{T}} \mathbf{x}^{[k]}(t) = \sum_{l=1}^{N} w_{n,l}^{[k]} x_l^{[k]}(t)$, where $\left(\mathbf{w}_n^{[k]}\right)^{\mathsf{T}}$ is the nth row of $\mathbf{W}^{[k]}$. Furthermore, it is assumed that the mixing matrices possess no known relationship.

2.3 Objective Functions

In this section, we give several related objective functions, which at their optimum all yield a valid solution for IVA. Additional choices for objective functions exist and they serve as the basis for other IVA achieving algorithms which we will review briefly in Chapter 3. We will begin by considering the general case when samples are not iid.

Before doing so we clarify that the parameters to be estimated are the mixing matrices $\mathbf{A}^{[k]}$ of each dataset. For convenience, an equivalent set of parameters are estimated, namely the inverse of the mixing matrices, $\mathbf{W}^{[k]} = \left(\hat{\mathbf{A}}^{[k]}\right)^{-1}$. It is clear that by the IVA formulation, $\mathbf{X} = \mathbf{A}\mathbf{S} = \left(\oplus \sum_{k=1}^{K} \mathbf{A}^{[k]}\right)\mathbf{S}$. We then have that $\mathbf{W} = \oplus \sum_{k=1}^{K} \mathbf{W}^{[k]}$. It is a bit misleading to say that the optimization parameter is $\mathbf{W}$, since most of the terms in this matrix are zero. Thus we choose in the sequel to use $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{N\times N\times K}$, i.e., a three-dimensional array, to denote the set of parameters to be estimated.

2.3.1 Dependent Samples

Since IVA is a parameter estimation problem, a natural objective function is given by the likelihood function, or equivalently the natural logarithm of the likelihood. In addition to the demixing parameters $\boldsymbol{\mathcal{W}}$, we use $\boldsymbol{\Theta}$ to represent all additional parameters required to describe the source pdf in the likelihood function,

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\mathcal{W}},\boldsymbol{\Theta}\right) &\triangleq \log\left(p_{\mathbf{X}}\left(\mathbf{X};\boldsymbol{\mathcal{W}},\boldsymbol{\Theta}\right)\right)\\ &= \log\left(\prod_{n=1}^{N} p_n\left(\mathbf{Y}_n;\boldsymbol{\theta}_n\right)\left|\det\mathbf{W}\right|^{T}\right)\\ &= \sum_{n=1}^{N}\log\left(p_n\left(\mathbf{Y}_n;\boldsymbol{\theta}_n\right)\right) + T\sum_{k=1}^{K}\log\left|\det\mathbf{W}^{[k]}\right|, \qquad (2.2)\end{aligned}$$

where $p_n\left(\cdot;\boldsymbol{\theta}_n\right)$ is the model for the distribution characterizing the multivariate source $\mathbf{S}_n$ with associated distribution parameters $\boldsymbol{\theta}_n$ and we have used $\det\mathbf{W} = \prod_{k=1}^{K}\det\mathbf{W}^{[k]}$. Note that if $\mathbf{X} = \mathbf{A}\mathbf{S}$, then $\mathrm{vec}\left(\mathbf{S}\right) = \left(\mathbf{I}_T\otimes\mathbf{A}^{-1}\right)\mathrm{vec}\left(\mathbf{X}\right)$, which implies $p_{\mathbf{X}}\left(\mathbf{X};\mathbf{A}\right) = \left|\det\left(\mathbf{I}_T\otimes\mathbf{A}^{-1}\right)\right| p_{\mathbf{S}}\left(\left(\mathbf{I}_T\otimes\mathbf{A}^{-1}\right)\mathrm{vec}\left(\mathbf{X}\right)\right) = \left|\det\mathbf{A}^{-1}\right|^{T} p_{\mathbf{S}}\left(\mathbf{S}\right)$. A useful reference for properties related to Kronecker products is provided by Seber [104].

The regularization term, $\log\left|\det\left(\mathbf{W}^{[k]}\right)\right|$, penalizes demixing matrices with s mall determinants, thus preventing the minimization of the first term by simply making the volume described by $\mathbf{W}^{[k]}$ arbitrarily small. The regularization term also prevents the estimated demixing matrices from becoming singular. The likelihood has no restrictions on $\mathbf{W}^{[k]}$ beyond being invertible, thus the solution space is the general linear group, which of course includes nonorthogonal matrices. When the smaller parameter space of orthogonal demixing matrices is considered, then some penalty in performance will be incurred. This will be discussed in more detail in Section 3.2.2.

Another principled approach to defining an IVA objective function is the minimization of mutual information rate among the estimated source component vectors (SCVs). Before specifying this objective function it is helpful to recall the definition of entropy rate [30, Eq 4.10]

$$\mathcal{H}_r\left\{\mathbf{s}_n\right\} \triangleq \lim_{T\to\infty} \frac{1}{T}\mathcal{H}\left\{\mathbf{s}_n\left(1\right),\ldots,\mathbf{s}_n\left(T\right)\right\} = \lim_{T\to\infty} \frac{1}{T}\mathcal{H}\left\{\mathbf{S}_n\right\}. \tag{2.3}$$

In order to utilize the entropy rate we have introduced the SCV, $\mathbf{s}_n \in \mathbb{R}^K$, which denotes a random vector process. Furthermore, the entropy (rate) of a random variable that is related to $\mathbf{x}$ through a linear invertible transformation, $\mathbf{x} = \mathbf{W}^{-1}\mathbf{y}$, is given by $\log\left|\det \mathbf{W}\right| + \mathcal{H}\left\{\mathbf{x}\right\}$.

From the above, it is clear that

$$\begin{aligned}\mathcal{I}_r\left\{\mathbf{y}_1;\ldots;\mathbf{y}_N\right\} &= \sum_{n=1}^{N}\mathcal{H}_r\left\{\mathbf{y}_n\right\} - \mathcal{H}_r\left\{\mathbf{W}^{[1]}\mathbf{x}^{[1]}\ldots\mathbf{W}^{[K]}\mathbf{x}^{[K]}\right\} \\ &= \sum_{n=1}^{N}\mathcal{H}_r\left\{\mathbf{y}_n\right\} - \sum_{k=1}^{K}\log\left|\det \mathbf{W}^{[k]}\right| - \mathcal{H}_r\left\{\mathbf{x}^{[1]}\ldots\mathbf{x}^{[K]}\right\}. \end{aligned} \tag{2.4}$$

Thus, to minimize the mutual information we must estimate and minimize the entropy rate of the estimated SCV $\mathbf{y}_n$. This is a difficult task and so we proceed by connecting the entropy rate measure to the likelihood of (2.2) by considering the following

$$\begin{aligned}
\mathcal{H}_r\left\{\mathbf{y}_n\right\} &= \lim_{T\to\infty}\frac{1}{T}\mathcal{H}\left\{\mathbf{y}_n\left(1\right),\ldots,\mathbf{y}_n\left(T\right)\right\} \\
&= -\lim_{T\to\infty}\frac{1}{T}\int_{\Omega_n} p_n\left(\mathbf{Y}_n\right)\log\left(p_n\left(\mathbf{Y}_n\right)\right)d\mathbf{Y}_n \\
&= \lim_{T\to\infty}\frac{1}{T}\left(-\int_{\Omega_n} p_n\left(\mathbf{Y}_n\right)\log\left(\frac{p_n\left(\mathbf{Y}_n\right)}{p_n\left(\mathbf{Y}_n;\boldsymbol{\theta}_n\right)}\right)d\mathbf{Y}_n\right. \\
&\qquad\qquad \left. -\int_{\Omega_n} p_n\left(\mathbf{Y}_n\right)\log\left(p_n\left(\mathbf{Y}_n;\boldsymbol{\theta}_n\right)\right)d\mathbf{Y}_n\right) \\
&= \lim_{T\to\infty}\frac{1}{T}\left(-\mathcal{D}_{KL}\left\{p_n\left(\mathbf{Y}_n\right)||p_n\left(\mathbf{Y}_n;\boldsymbol{\theta}_n\right)\right\} - E\left\{\log\left(p_n\left(\mathbf{Y}_n;\boldsymbol{\theta}_n\right)\right)\right\}\right).
\end{aligned}$$

By rearranging the expression above we have that

$$\lim_{T\to\infty}\frac{1}{T}E\left\{\log\left(p_n\left(\mathbf{Y}_n;\boldsymbol{\theta}_n\right)\right)\right\} = -\mathcal{H}_r\left\{\mathbf{y}_n\right\} - \lim_{T\to\infty}\frac{1}{T}\mathcal{D}_{KL}\left\{p_n\left(\mathbf{Y}_n\right)||p_n\left(\mathbf{Y}_n;\boldsymbol{\theta}_n\right)\right\},$$

which indicates that maximizing the likelihood will be minimizing the entropy rate of $\mathbf{y}_n$ *and* finding a distribution within the family $p_n\left(\mathbf{Y}_n;\boldsymbol{\theta}_n\right)$ that minimizes the 'distance' to the actual distribution $p_n\left(\mathbf{Y}_n\right)$. When the distribution parameters are such that the distribution model can exactly match the true source distribution then the 'error' represented by $\mathcal{D}_{KL}\left\{p_n\left(\mathbf{Y}_n\right)||p_n\left(\mathbf{Y}_n;\boldsymbol{\theta}_n\right)\right\}$ is zero, which implies that

$$\mathcal{H}_r\left\{\mathbf{y}_n\right\} = -\lim_{T\to\infty}\frac{1}{T}E\left\{\log\left(p_n\left(\mathbf{Y}_n;\boldsymbol{\theta}_n\right)\right)\right\} = -\lim_{T\to\infty}\frac{1}{T}E\left\{\log\left(p_n\left(\mathbf{Y}_n\right)\right)\right\}.$$

Because $\mathcal{H}_r\left\{\mathbf{x}^{[1]}\ldots\mathbf{x}^{[K]}\right\}$ is a constant with respect to (wrt) $\boldsymbol{\mathcal{W}}$, we have that minimization of mutual information rate in (2.4) is equivalent to maximizing the likelihood in (2.2) as the sample size goes to infinity provided that the model can exactly match the true source distribution. Under these conditions, it can also be stated that IVA minimizes the entropy (rate) of the estimated SCVs subject to a regularization term.

Also, we provide an additional insight into the cost function for IVA by noting that $\mathcal{I}_r\left\{\mathbf{y}_n\right\}=\sum_{k=1}^{K}\mathcal{H}_r\left\{y_n^{[k]}\right\}-\mathcal{H}_r\left\{\mathbf{y}_n\right\}$, which shows that

$$\begin{aligned}\mathcal{C}_{\mathrm{IVA}}\left(\boldsymbol{\mathcal{W}}\right) &\triangleq \sum_{n=1}^{N}\mathcal{H}_r\left\{\mathbf{y}_n\right\}-\sum_{k=1}^{K}\log\left|\det\mathbf{W}^{[k]}\right| &(2.5)\\ &=\sum_{n=1}^{N}\left(\sum_{k=1}^{K}\mathcal{H}_r\left\{y_n^{[k]}\right\}-\mathcal{I}_r\left\{\mathbf{y}_n\right\}\right)-\sum_{k=1}^{K}\log\left|\det\mathbf{W}^{[k]}\right|. &(2.6)\end{aligned}$$

This representation illustrates that minimizing the IVA cost function will equally weight the minimization of the source entropy rates and the maximization of the across dataset dependence measure provided by the mutual information rate of $\mathbf{y}_n$. It is also clear that the mutual information rate portion of the IVA cost function is responsible for resolving the permutation ambiguity across multiple datasets, since without the mutual information rate of the SCVs the cost function would be identical to using ICA on each of the K datasets. This representation will prove handy for our identifiability discussion in Section 2.5.

Throughout this work it is sufficient to consider the simplification when the source distribution parameters are known, which we indicate by suppressing of $\mathbf{\Theta}$ and $\boldsymbol{\theta}_n$. Accordingly, in the sequel, we will find it useful to express the multivariate score function $\mathbf{\Phi}_n\triangleq\mathbf{\Phi}_n\left(\mathbf{Y}_n\right)=-\partial\log\left(p_n\left(\mathbf{Y}_n\right)\right)/\partial\mathbf{Y}_n\in\mathbb{R}^{K\times T}$ and $\boldsymbol{\phi}_n^{[k]}=\mathbf{\Phi}_n^{\mathsf{T}}\mathbf{e}_k$.

2.3.2 iid Samples

The particular case when the samples are iid is of interest for us and will be the problem solved by the algorithms described in Chapter 3. So we start with (2.2) and use the independent sampling assumption first, followed by the identically distributed assumption, i.e.,

$$\begin{aligned}\mathcal{L}_{\text{iid}}\left(\boldsymbol{\mathcal{W}}\right) &= -\sum_{n=1}^{N}\log\left(p_n\left(\mathbf{Y}_n\right)\right) - T\sum_{k=1}^{K}\log\left|\det\mathbf{W}^{[k]}\right| \\ &= -\sum_{n=1}^{N}\log\left(\prod_{t=1}^{T}p_{n,t}\left(\mathbf{y}_n\left(t\right)\right)\right) - T\sum_{k=1}^{K}\log\left|\det\mathbf{W}^{[k]}\right| \\ &= -\sum_{n=1}^{N}\sum_{t=1}^{T}\log\left(p_n\left(\mathbf{y}_n\left(t\right)\right)\right) - T\sum_{k=1}^{K}\log\left|\det\mathbf{W}^{[k]}\right|, \end{aligned} \tag{2.7}$$

where $p_{n,t}\left(\cdot\right)$ denotes the pdf associated with the tth sample and nth source, which if iid reduces to $p_n\left(\cdot\right)$.

The connection of (2.7) with mutual information follows from the discussion of (2.4). Thus, an objective function for IVA using mutual information with iid samples is as follows

$$\mathcal{C}_{\text{IVA}}\left(\boldsymbol{\mathcal{W}}\right) = \sum_{n=1}^{N}\mathcal{H}\left\{\mathbf{y}_n\right\} - \sum_{k=1}^{K}\log\left|\det\mathbf{W}^{[k]}\right|. \tag{2.8}$$

This new cost function is equivalent to minimizing the mutual information between the estimated SCVs when the source distribution model follows the actual (true) distribution.

2.3.3 Gaussian Sources and Canonical Correlation Analysis

The Gaussian assumption has found wide use in many signal processing and data analysis applications [54], and in [79] it has been shown that second-order statistics are successfully able to solve the IVA problem.[2] The original motivation for us to consider the Gaussian distribution source model for IVA are the previous works utilizing CCA [29] and MCCA [79], which utilize second-order statistics. Hence, we use the zero-mean and real-valued multivariate Gaussian distribution,

$$p\left(\mathbf{s}_n|\mathbf{\Sigma}_n\right) = \frac{1}{\left(2\pi\right)^{K/2}\det\left(\mathbf{\Sigma}_n\right)^{1/2}}\exp\left(-\tfrac{1}{2}\mathbf{s}_n^{\mathsf{T}}\mathbf{\Sigma}_n^{-1}\mathbf{s}_n\right),$$

as the assumed form of the K dimensional SCV distribution for the nth source with covariance matrix $\mathbf{\Sigma}_n = E\left\{\mathbf{s}_n\mathbf{s}_n^{\mathsf{T}}\right\}$. The covariance matrix is in general not known a priori and is thus estimated a part of the IVA optimization procedure.

The Gaussian assumption applied to IVA provides a connection to the earlier works of CCA and MCCA (for a review of CCA and MCCA see Appendix C). Substituting into (2.8) the entropy of a K dimensional multivariate Gaussian vector, $H\left(\mathbf{s}_n\right) = 1/2\log\left[\left(2\pi e\right)^K \prod_{k=1}^{K}\lambda_n^{[k]}\right]$, where $\lambda_n^{[k]}$ is the kth eigenvalue of the covariance matrix $\mathbf{\Sigma}_n$, gives the IVA with multivariate Gaussian distribution model (IVA-Gauss) cost function,

$$\mathcal{C}_{\text{IVA-G}}\left(\boldsymbol{\mathcal{W}}\right) = \frac{NK\log\left(2\pi e\right)}{2} + \frac{1}{2}\log\left(\prod_{n=1}^{N}\prod_{k=1}^{K}\lambda_n^{[k]}\right) - \sum_{k=1}^{K}\log\left|\det\left(\mathbf{W}^{[k]}\right)\right|.$$

---

[2] We would like to note that that the necessary identification conditions given in [79] are only sufficient conditions and that less stringent conditions have been identified in ths work.

The cost function indicates that the product of the SCV covariance eigenvalues should be minimized. For the purposes of this discussion, we can consider the constraint that the sum of the eigenvalues is fixed. Under this constraint, then the minimal cost is when each covariance matrix is as ill-conditioned as possible, i.e., the eigenvalues are maximally spread apart. This condition is equivalent to finding SCVs with maximally correlated components.

In MCCA, five cost functions are proposed to indirectly achieve this condition—see Appendix C for a brief revision of CCA and MCCA. The GENVAR cost function is one of these five ad hoc—yet reasonably motivated—cost functions. Since MCCA adopts a deflationary approach, the GENVAR cost function is the product of eigenvalues associated with a single SCV covariance matrix. If the GENVAR cost function is naturally extended to be the product of the covariance eigenvalues for all N estimated SCVs and the demixing matrices are orthogonal, then the MCCA and IVA with a multivariate Gaussian source model have identical cost functions. In this way, IVA bridges the gap between CCA and ICA.

2.4 Fisher Information Matrix

The FIM provides a measure of how much information each sample or observation of a random variable carries about the parameters being estimated. Thus in rough manner, one can say the 'larger' the FIM the more valuable, or informative, a given observation is to the estimator. Here we derive the FIM of (2.2) wrt $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{N \times N \times K}$. The KN^2 parameters result in $KN^2 \times KN^2$ dimension FIM with the entry associated with $w_{m_1,n_1}^{[k_1]}$ and $w_{m_2,n_2}^{[k_2]}$ denoted by and computed as:

$$[\mathbf{F}(\boldsymbol{\mathcal{W}})]_{k_2,m_2,n_2}^{k_1,m_1,n_1} \triangleq E\left\{\frac{\partial\mathcal{L}(\boldsymbol{\mathcal{W}})}{\partial w_{m_1,n_1}^{[k_1]}}\frac{\partial\mathcal{L}(\boldsymbol{\mathcal{W}})}{\partial w_{m_2,n_2}^{[k_2]}}\right\} = -E\left\{\frac{\partial^2\mathcal{L}(\mathbf{W})}{\partial w_{m_2,n_2}^{[k_2]}\partial w_{m_1,n_1}^{[k_1]}}\right\}. \tag{2.9}$$

The expression above holds when the regularity condition $E\left\{\frac{\partial\mathcal{L}(\boldsymbol{\mathcal{W}})}{\partial w_{m,n}^{[k]}}\right\} = 0$, i.e., an unbiased estimator, holds for all k, m, and n. Indeed, this regularity condition holds, see (A.15) in Appendix A.

For the purposes of determining identifiability and the performance bound, we need only consider the FIM locally around a solution, i.e., $\mathbf{W} = \mathbf{A}^{-1}$, where $\mathbf{A}^{-1}$ and $\mathbf{W}$ are “freely” chosen so as to alleviate all scale and permutation ambiguities. In general, this leads to a complex expression that depends on $\mathbf{A}$; fortunately this complexity is unnecessary. Due to the invariance of the induced Cramér-Rao lower bound (iCRLB) wrt $\mathbf{G} = \mathbf{WA}$, the global demixing-mixing matrix, we need only consider $\mathbf{A} = \mathbf{I}$, i.e., the CRLB of $\mathbf{G}$ depends only on the statistics of the sources, [24]. This is similar to that case in [119] since the mixing matrices $\mathbf{A}^{[k]}$ are unrelated. Thus the matrix of interest is

$$[\mathbf{F}]_{k_2,m_2,n_2}^{k_1,m_1,n_1} \triangleq [\mathbf{F}(\boldsymbol{\mathcal{W}})]_{k_2,m_2,n_2}^{k_1,m_1,n_1}\Big|_{\mathbf{A}=\mathbf{I},\mathbf{W}=\mathbf{I}}. \tag{2.10}$$

It will prove useful to define $\mathcal{K}_{m,n}^{[k_1,k_2]} \triangleq \frac{1}{T}E\left\{\left(\boldsymbol{\phi}_m^{[k_1]}\right)^{\mathsf{T}}\mathbf{s}_n^{[k_1]}\left(\mathbf{s}_n^{[k_2]}\right)^{\mathsf{T}}\boldsymbol{\phi}_m^{[k_2]}\right\}$, $1 \leq m,n \leq N$, $\mathbf{R}_n^{[k_1,k_2]} \triangleq E\left\{\mathbf{s}_n^{[k_1]}\left(\mathbf{s}_n^{[k_2]}\right)^{\mathsf{T}}\right\} \in \mathbb{R}^{T\times T}$, and $\boldsymbol{\Gamma}_n^{[k_1,k_2]} \triangleq E\left\{\boldsymbol{\phi}_n^{[k_1]}\left(\boldsymbol{\phi}_n^{[k_2]}\right)^{\mathsf{T}}\right\} \in \mathbb{R}^{T\times T}$, to describe the form of the block diagonal FIM compactly. In Appendix A, it is shown that the first N block entries are given by $\mathbf{F}_n \triangleq \operatorname{cov}\left\{\operatorname{diag}\left(\boldsymbol{\Phi}_n\mathbf{S}_n^{\mathsf{T}} - \mathbf{I}_T\right)\right\} =$

$$\mathbf{F} = V \begin{bmatrix} \frac{1}{V}\mathbf{F_1} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & {\color{blue}\boldsymbol{\mathcal{K}}_{1,2}} & \cdot & {\color{blue}\mathbf{I}} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & {\color{green}\boldsymbol{\mathcal{K}}_{1,3}} & \cdot & \cdot & \cdot & {\color{green}\mathbf{I}} & \cdot & \cdot \\ \cdot & {\color{blue}\mathbf{I}} & \cdot & {\color{blue}\boldsymbol{\mathcal{K}}_{2,1}} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \frac{1}{V}\mathbf{F_2} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & {\color{red}\boldsymbol{\mathcal{K}}_{2,3}} & \cdot & {\color{red}\mathbf{I}} & \cdot \\ \cdot & \cdot & {\color{green}\mathbf{I}} & \cdot & \cdot & \cdot & {\color{green}\boldsymbol{\mathcal{K}}_{3,1}} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & {\color{red}\mathbf{I}} & \cdot & {\color{red}\boldsymbol{\mathcal{K}}_{3,2}} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{1}{V}\mathbf{F_3} \end{bmatrix}$$

Figure 2.1: Form of FIM when $N = 3$ sources. All entries are of $K \times K$ matrices and we use $\cdot$ denote zero blocks. The entries of FIM associated with $\mathbf{F}_{1,2}$, $\mathbf{F}_{1,3}$, and $\mathbf{F}_{2,3}$ are indicated by blue, green, and red, respectively.

$T\left(\boldsymbol{\mathcal{K}}_{n,n} - T\mathbf{1}_{K\times K}\right) \in \mathbb{R}^{K\times K}$ and the remaining block entries are defined for $n > m$ as

$$\mathbf{F}_{m,n} \triangleq \operatorname{cov}\left\{ \begin{bmatrix} \operatorname{diag}\left(\boldsymbol{\Phi}_m \mathbf{S}_n^{\mathsf{T}}\right) \\ \operatorname{diag}\left(\boldsymbol{\Phi}_n \mathbf{S}_m^{\mathsf{T}}\right) \end{bmatrix} \right\} = T \begin{bmatrix} \boldsymbol{\mathcal{K}}_{m,n} & \mathbf{I}_K \\ \mathbf{I}_K & \boldsymbol{\mathcal{K}}_{n,m} \end{bmatrix} \in \mathbb{R}^{2K\times 2K}, \qquad (2.11)$$

where the (k_1, k_2) entry of $\boldsymbol{\mathcal{K}}_{m,n} \in \mathbb{R}^{K\times K}$ is $T^{-1}\operatorname{tr}\left(\boldsymbol{\Gamma}_m^{[k_2,k_1]}\mathbf{R}_n^{[k_1,k_2]}\right)$ when $m \neq n$.

The form of the FIM is a multivariate extension of the single dataset forms given in [109, 93, 119, 28]. The FIM has a form that is a block matrix version of the single dataset result, e.g., see Fig. 2.1 and compare to the similar form given in [81] for complex-valued ICA. The 2×2 blocks with ones in the off-diagonal elements and pair-wise cross terms in the two diagonal elements of the ICA FIM are here replaced with 2×2 block matrices with identity matrices in the off-diagonal blocks and the cross terms in the two diagonal block matrices, i.e., $\mathbf{F}_{m,n}$.

2.5 Identification Conditions

The identification of sources in (real-valued) ICA is possible so long as no two sources are Gaussian with proportional covariance matrices [28, Chapter 4]. When sources are said to be identifiable for ICA, this means that the sources can be recovered up to a scale factor and arbitrary ordering, i.e., the true mixing matrix $\mathbf{A}_0$ can be identified upto $\mathbf{A}_0\mathbf{\Lambda}\mathbf{P}$, where $\mathbf{\Lambda}$ is any nonsingular diagonal matrix and $\mathbf{P}$ is any permutation matrix. The identifiability of the mixing matrix and consequently of the sources "is defined with respect to a certain model structure" [112]. Because the model structure of IVA is a generalization of the model structure for ICA, we expect a generalization of the identification conditions for ICA. Intuitively, the identification conditions for IVA are related to the dependence of the sources across the datasets. More specifically, when sources possess dependence across datasets we expect that these estimated sources can be 'aligned'—this is the original motivation of IVA [58, 46]. However, if there are sources for which no alignment exhibits dependence, then under the ICA identification conditions, sources can be separated but not necessarily aligned. That is, without dependence across datasets the estimated sources of IVA would be no different than using ICA on each dataset individually since there is no dependency to exploit. The identification conditions, which we present in this section, capture both cases, i.e., whether or not there is dependence between sources across datasets.

To discuss identifiability of IVA we need to provide a notation that allows us to indicate a particular subset of rows in an SCM. For this section, we let

$\boldsymbol{\alpha} = [\alpha_1 \ldots \alpha_{d_\alpha}]^\mathsf{T} \in \mathbb{N}^{K_\alpha}$, where $0 \leq K_\alpha \leq K$. The complementing subset of $\boldsymbol{\alpha}$ in $\{1, \ldots, K\}$ is indicated by $\boldsymbol{\alpha}^c \in \mathbb{N}^{K-K_\alpha}$. The IVA identification conditions use the following definition:

Definition 1 ($\boldsymbol{\alpha}$*-Gaussian*)**.** *A source,* $\mathbf{S} \in \mathbb{R}^{K \times V}$*, has an* $\boldsymbol{\alpha}$-Gaussian *component when* $\mathrm{vec}_{\boldsymbol{\alpha}}(\mathbf{S}) \perp\!\!\!\perp \mathrm{vec}_{\boldsymbol{\alpha}^c}(\mathbf{S})$*, and* $\mathrm{vec}_{\boldsymbol{\alpha}}(\mathbf{S}) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_\alpha)$*, where* $\mathbf{R}_\alpha = E\left\{\mathrm{vec}_{\boldsymbol{\alpha}}(\mathbf{S})\, \mathrm{vec}_{\boldsymbol{\alpha}}^\mathsf{T}(\mathbf{S})\right\} \in \mathbb{R}^{K_\alpha V \times K_\alpha V}$ *is nonsingular.*

The $\boldsymbol{\alpha}$-Gaussian definition is used to identify that there exist a subset of rows in an SCM that is independent of the other rows in the same SCM and that the given subset follows a multivariate Gaussian distribution. The theorem stating the IVA identification conditions and its proof follow.

Theorem 1 (IVA Nonidentifiability)**.** *The sources cannot be identified if and only if (iff)* $\exists\, \boldsymbol{\alpha} \neq \emptyset$ *and* $\exists\, m \neq n$ *such that* $\mathbf{S}_m$ *and* $\mathbf{S}_n$ *have* $\boldsymbol{\alpha}$-Gaussian *components for which* $\mathbf{R}_{m,\alpha} = (\mathbf{I}_V \otimes \mathbf{D})\, \mathbf{R}_{n,\alpha}\, (\mathbf{I}_V \otimes \mathbf{D}) \in \mathbb{R}^{K_\alpha V \times K_\alpha V}$*, where* $\mathbf{D} \in \mathbb{R}^{K_\alpha \times K_\alpha}$ *is any full rank diagonal matrix.*

Proof of IVA Nonidentifiability. Given the FIM (A.9), (A.10), (A.11), since $\mathbf{F}_{m,n}$ is a covariance matrix, it must be positive semidefinite and is singular iff $\exists\, (\mathbf{a}, \mathbf{b}) \neq (\mathbf{0}, \mathbf{0}) : \mathbf{a}^\mathsf{T} \mathrm{diag}\left(\boldsymbol{\Phi}_m \mathbf{S}_n^\mathsf{T}\right) - \mathbf{b}^\mathsf{T} \mathrm{diag}\left(\boldsymbol{\Phi}_n \mathbf{S}_m^\mathsf{T}\right) = 0$, $\forall\, \mathbf{S}_m \in \Omega_{\mathbf{S}_m}, \mathbf{S}_n \in \Omega_{\mathbf{S}_n}$, where $\Omega_{\mathbf{X}}$ denotes the sample space of the random matrix $\mathbf{X}$.

It is convenient to rewrite the following:

$$\mathrm{diag}\left(\boldsymbol{\Phi}_m \mathbf{S}_n^\mathsf{T}\right) = \mathrm{diag}\left(\sum_{t=1}^{T} \boldsymbol{\phi}_m(t)\, \mathbf{s}_n^\mathsf{T}(t)\right) = \sum_{t=1}^{T} \boldsymbol{\phi}_m(t) \circ \mathbf{s}_n(t), \tag{2.12}$$

where $\mathbf{s}_n(t)$ and $\boldsymbol{\phi}_m(t)$ denote the tth columns of $\mathbf{S}_n$ and $\boldsymbol{\Phi}_m$, respectively.

Hence, the following statements are all equivalent conditional on $\exists(\mathbf{a},\mathbf{b}) \neq (\mathbf{0},\mathbf{0})$:

$$\begin{aligned}
& && \mathbf{F}_{m,n} \text{ is singular} && (2.13)\\
&\Leftrightarrow && 0 = \mathbf{a}^{\mathsf{T}}\mathrm{diag}\left(\boldsymbol{\Phi}_m\mathbf{S}_n^{\mathsf{T}}\right) - \mathbf{b}^{\mathsf{T}}\mathrm{diag}\left(\boldsymbol{\Phi}_n\mathbf{S}_m^{\mathsf{T}}\right) && (2.14)\\
&\Leftrightarrow && 0 = \mathbf{a}^{\mathsf{T}}\sum_{v=1}^{V}\boldsymbol{\phi}_m(v)\circ\mathbf{s}_n(v) - \mathbf{b}^{\mathsf{T}}\sum_{q=1}^{V}\boldsymbol{\phi}_n(q)\circ\mathbf{s}_m(q) && (2.15)\\
&\Leftrightarrow && 0 = (\mathbf{1}_V\otimes\mathbf{a})^{\mathsf{T}}\left(\mathrm{vec}(\boldsymbol{\Phi}_m)\circ\mathrm{vec}(\mathbf{S}_n)\right) - (\mathbf{1}_V\otimes\mathbf{b})^{\mathsf{T}}\left(\mathrm{vec}(\boldsymbol{\Phi}_n)\circ\mathrm{vec}(\mathbf{S}_m)\right) && (2.16)\\
&\Leftrightarrow && 0 = \mathrm{vec}_{\boldsymbol{\alpha}}^{\mathsf{T}}(\boldsymbol{\Phi}_m)\left(\mathbf{I}_V\otimes\mathbf{D}_{\mathbf{a}[\boldsymbol{\alpha}]}\right)\mathrm{vec}_{\boldsymbol{\alpha}}(\mathbf{S}_n) - \mathrm{vec}_{\boldsymbol{\beta}}^{\mathsf{T}}(\boldsymbol{\Phi}_n)\left(\mathbf{I}_V\otimes\mathbf{D}_{\mathbf{b}[\boldsymbol{\beta}]}\right)\mathrm{vec}_{\boldsymbol{\beta}}(\mathbf{S}_m) && (2.17)\\
&\Leftrightarrow && 0 = \mathrm{vec}_{\boldsymbol{\alpha}}^{\mathsf{T}}(\boldsymbol{\Phi}_m)\left(\mathbf{I}_V\otimes\mathbf{D}_{\mathbf{a}[\boldsymbol{\alpha}]}\right)\mathrm{vec}_{\boldsymbol{\alpha}}(\mathbf{S}_n) - \mathrm{vec}_{\boldsymbol{\alpha}}^{\mathsf{T}}(\boldsymbol{\Phi}_n)\left(\mathbf{I}_V\otimes\mathbf{D}_{\mathbf{b}[\boldsymbol{\alpha}]}\right)\mathrm{vec}_{\boldsymbol{\alpha}}(\mathbf{S}_m) && (2.18)\\
&\Leftrightarrow && 0 = \mathrm{vec}_{\boldsymbol{\alpha}}^{\mathsf{T}}(\mathbf{S}_m)\mathbf{R}_{m,\alpha}^{-1}\left(\mathbf{I}_V\otimes\mathbf{D}_{\mathbf{a}[\boldsymbol{\alpha}]}\right)\mathrm{vec}_{\boldsymbol{\alpha}}(\mathbf{S}_n) \\
& && \qquad - \mathrm{vec}_{\boldsymbol{\alpha}}^{\mathsf{T}}(\mathbf{S}_n)\mathbf{R}_{n,\alpha}^{-1}\left(\mathbf{I}_V\otimes\mathbf{D}_{\mathbf{b}[\boldsymbol{\alpha}]}\right)\mathrm{vec}_{\boldsymbol{\alpha}}(\mathbf{S}_m) && (2.19)\\
&\Leftrightarrow && 0 = \mathbf{R}_{m,\alpha}^{-1}\left(\mathbf{I}_V\otimes\mathbf{D}_{\mathbf{a}[\boldsymbol{\alpha}]}\right) - \left(\mathbf{I}_V\otimes\mathbf{D}_{\mathbf{b}[\boldsymbol{\alpha}]}\right)\mathbf{R}_{n,\alpha}^{-1} && (2.20)\\
&\Leftrightarrow && \mathbf{R}_{m,\alpha} = \left(\mathbf{I}_V\otimes\mathbf{D}_{\mathbf{a}[\boldsymbol{\alpha}]}\right)\mathbf{R}_{n,\alpha}\left(\mathbf{I}_V\otimes\mathbf{D}_{\mathbf{b}[\boldsymbol{\alpha}]}^{-1}\right) && (2.21)\\
&\Leftrightarrow && \mathbf{R}_{m,\alpha} = (\mathbf{I}_V\otimes\mathbf{D})\mathbf{R}_{n,\alpha}(\mathbf{I}_V\otimes\mathbf{D}), && (2.22)
\end{aligned}$$

where $\mathbf{D}_{\mathbf{a}[\boldsymbol{\alpha}]} \triangleq \mathrm{Diag}(\mathbf{a}[\boldsymbol{\alpha}])$, $\mathbf{D}_{\mathbf{b}[\boldsymbol{\alpha}]} \triangleq \mathrm{Diag}(\mathbf{b}[\boldsymbol{\alpha}])$, $\boldsymbol{\alpha} \in \mathbb{N}^{K_\alpha}$, and $\boldsymbol{\beta} \in \mathbb{N}^{K_\beta}$.

It is straightforward to observe that (2.13), (2.14), (2.15), and (2.16) are equivalent expressions. From the relationship

$$(\mathbf{x} \otimes \mathbf{y})^{\mathsf{T}} (\mathbf{w} \circ \mathbf{z}) = \mathbf{w}^{\mathsf{T}} (\operatorname{Diag}(\mathbf{x}) \otimes \operatorname{Diag}(\mathbf{y}))\, \mathbf{z},$$

the expression in (2.17) holds only when $\boldsymbol{\alpha} = \boldsymbol{\beta}$, i.e., the zero entries of $\mathbf{a}$ and $\mathbf{b}$ are at the same locations. See Lemma 1 below to explain (2.19). Since (2.19) must hold for all possible values of $\operatorname{vec}_{\boldsymbol{\alpha}}(\mathbf{S}_m)$ and $\operatorname{vec}_{\boldsymbol{\alpha}}(\mathbf{S}_n)$, (2.20) must hold. Equation (2.21) is equivalent since all entries of $\mathbf{b}[\boldsymbol{\alpha}]$ are nonzero by (2.18). Lastly, since $\mathbf{R}_{m,\alpha}$ is symmetric we must have that either $\mathbf{R}_{n,\alpha}$ is diagonal or $\mathbf{D}_{\mathbf{a}[\boldsymbol{\alpha}]} = \left(\mathbf{D}_{\mathbf{b}[\boldsymbol{\alpha}]}\right)^{-1}$. In either case (2.22) holds.

□

Lemma 1. *For $m \neq n$,*

$$\operatorname{vec}_{\boldsymbol{\alpha}}^{\mathsf{T}}(\boldsymbol{\Phi}_m) \left(\mathbf{I}_T \otimes \mathbf{D}_{\mathbf{a}[\boldsymbol{\alpha}]}\right) \operatorname{vec}_{\boldsymbol{\alpha}}(\mathbf{S}_n) = \operatorname{vec}_{\boldsymbol{\alpha}}^{\mathsf{T}}(\boldsymbol{\Phi}_n) \left(\mathbf{I}_T \otimes \mathbf{D}_{\mathbf{b}[\boldsymbol{\alpha}]}\right) \operatorname{vec}_{\boldsymbol{\alpha}}(\mathbf{S}_m) \tag{2.23}$$

holds iff

$$\operatorname{vec}_{\boldsymbol{\alpha}}^{\mathsf{T}}(\mathbf{S}_m) \mathbf{R}_m^{-1} \left(\mathbf{I}_T \otimes \mathbf{D}_{\mathbf{a}[\boldsymbol{\alpha}]}\right) \operatorname{vec}_{\boldsymbol{\alpha}}(\mathbf{S}_n) = \operatorname{vec}_{\boldsymbol{\alpha}}^{\mathsf{T}}(\mathbf{S}_n) \mathbf{R}_n^{-1} \left(\mathbf{I}_T \otimes \mathbf{D}_{\mathbf{b}[\boldsymbol{\alpha}]}\right) \operatorname{vec}_{\boldsymbol{\alpha}}(\mathbf{S}_m) \tag{2.24}$$

*and $\mathbf{S}_m$ and $\mathbf{S}_n$ each have an $\boldsymbol{\alpha}$-*Gaussian *component.*

Proof. (⇒) Since the left-hand side of (2.23) is linear in $\operatorname{vec}_{\boldsymbol{\alpha}}(\mathbf{S}_n)$ we must have that $\operatorname{vec}_{\boldsymbol{\alpha}}(\boldsymbol{\Phi}_n)$ is not a function of $\operatorname{vec}_{\boldsymbol{\alpha}^c}(\mathbf{S}_n)$ and it is necessarily linear in $\operatorname{vec}_{\boldsymbol{\alpha}}(\mathbf{S}_n)$, i.e., $\mathbf{S}_n$ must have $\boldsymbol{\alpha}$-*Gaussian* component. By symmetry, the same can be concluded about $\mathbf{S}_m$.

($\Leftarrow$) If $\mathbf{S}_n$ has $\boldsymbol{\alpha}$-*Gaussian* component then $\text{vec}_{\boldsymbol{\alpha}}\left(\boldsymbol{\Phi}_n\right) = \mathbf{R}_n^{-1}\text{vec}_{\boldsymbol{\alpha}}\left(\mathbf{S}_n\right)$. $\square$

It is noteworthy to mention that the IVA identification conditions admit sources for which the distribution can be factored, i.e.,

$$p_n\left(\mathbf{S}_n\right) = \prod_{q=1}^{Q} p_{n_q}\left(\text{vec}_{\mathcal{Q}_q}\left(\mathbf{S}_n\right)\right),$$

where $\{\mathcal{Q}_1, \mathcal{Q}_2, \ldots, \mathcal{Q}_Q\}$, $\mathcal{Q}_q \subset \{1, ..., K\}$, $\mathcal{Q}_q \cap \mathcal{Q}_{q'} = \emptyset\ \forall q \neq q'$, and $\cup_{q=1}^{Q}\mathcal{Q}_q = \{1, ..., K\}$. If, for example $Q = K$, then IVA would produce the same identification conditions as ICA on each dataset individually. Stated differently, identifiability of IVA does not require the sources to possess dependence across datasets.

Recalling that a prime motivation for considering the IVA formulation is to determine when the sources can be aligned in a common way across all datasets, i.e., under what conditions is $\hat{\mathbf{A}}^{[k]} = \left(\mathbf{W}^{[k]}\right)^{-1} = \mathbf{A}^{[k]}\mathbf{P}^{[k]}\boldsymbol{\Lambda}^{[k]}$, where $\boldsymbol{\Lambda}^{[k]}$ is any full rank diagonal matrix and $\mathbf{P}$ is a common permutation matrix such that $\mathbf{P} = \mathbf{P}^{[k]}\ \forall k = 1, \ldots, K$. The common permutation identification condition is given in the next theorem which uses the following definition:

Definition 2 (*$\boldsymbol{\alpha}$-independent*)**.** *A source,* $\mathbf{S} \in \mathbb{R}^{K \times V}$, *is* $\boldsymbol{\alpha}$-independent *when* $\text{vec}_{\boldsymbol{\alpha}}\left(\mathbf{S}\right) \perp\!\!\!\perp \text{vec}_{\boldsymbol{\alpha}^c}\left(\mathbf{S}\right)$.

The $\boldsymbol{\alpha}$-independent definition is used to identify that there exists a subset of rows in a SCM (or SCV) that is independent of the other rows in the SCM (SCV).

Theorem 2 (Common Permutation Matrix for IVA)**.** *Assuming the IVA identification conditions of Theorem 1 are satisfied, i.e., in the limit as* $T \to \infty$ *so that* $\left(\mathbf{W}^{[k]}\right)^{-1} = \mathbf{A}^{[k]}\mathbf{P}^{[k]}\boldsymbol{\Lambda}^{[k]}$*:*

The permutation matrix associated with each dataset is common iff $\forall\, m \neq n \nexists\, \boldsymbol{\alpha} \neq \emptyset$ *such that both* $\mathbf{s}_m$ *and* $\mathbf{s}_n$ *are* $\boldsymbol{\alpha}$-independent*.*

Proof. First, restating the cost function given by (2.6)

$$\mathcal{C}_{\text{IVA}}\left(\boldsymbol{\mathcal{W}}\right) = \sum_{n=1}^{N}\left(\sum_{k=1}^{K}\mathcal{H}_r\left\{y_n^{[k]}\right\} - \mathcal{I}_r\left\{\mathbf{y}_n\right\}\right) - \sum_{k=1}^{K}\log\left|\det \mathbf{W}^{[k]}\right|. \tag{2.25}$$

The above relationship makes it clear that any permutation matrix at most affects the $\mathcal{I}_r\left\{\mathbf{y}_n\right\}$ term. Furthermore, we only need consider permutation matrices that can achieve the global minimum. The proof is by contradiction (in both directions):

$$\exists 1 \leq k_1 \neq k_2 \leq K : \mathbf{P}^{[k_1]} \neq \mathbf{P}^{[k_2]} \tag{2.26}$$

$$\Leftrightarrow \quad \mathcal{I}_r\left\{\mathbf{s}_m^{[\boldsymbol{\alpha}]};\mathbf{s}_m^{[\boldsymbol{\alpha}^c]}\right\} + \mathcal{I}_r\left\{\mathbf{s}_n^{[\boldsymbol{\alpha}]};\mathbf{s}_n^{[\boldsymbol{\alpha}^c]}\right\} = \mathcal{I}_r\left\{\mathbf{s}_m^{[\boldsymbol{\alpha}]};\mathbf{s}_n^{[\boldsymbol{\alpha}^c]}\right\} + \mathcal{I}_r\left\{\mathbf{s}_n^{[\boldsymbol{\alpha}]};\mathbf{s}_m^{[\boldsymbol{\alpha}^c]}\right\} \tag{2.27}$$

$$\Leftrightarrow \quad \mathcal{I}_r\left\{\mathbf{s}_m^{[\boldsymbol{\alpha}]};\mathbf{s}_m^{[\boldsymbol{\alpha}^c]}\right\} + \mathcal{I}_r\left\{\mathbf{s}_n^{[\boldsymbol{\alpha}]};\mathbf{s}_n^{[\boldsymbol{\alpha}^c]}\right\} = 0 \tag{2.28}$$

$$\Leftrightarrow \quad \mathbf{S}_m \text{ and } \mathbf{S}_n \text{ are } \boldsymbol{\alpha}\text{-independent} \tag{2.29}$$

We have used fact that $\mathcal{I}_r\left\{\mathcal{X};\mathcal{Y}\right\} \geq 0$ with equality iff $\mathcal{X} \perp\!\!\!\perp \mathcal{Y}$, which implies by the assumption of IVA that $\mathcal{I}_r\left\{\mathbf{s}_i^{[\boldsymbol{\alpha}_1]};\mathbf{s}_j^{[\boldsymbol{\alpha}_2]}\right\} = 0\ \forall i \neq j, \boldsymbol{\alpha}_1,\ \boldsymbol{\alpha}_2$, where $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are any indexing sets. □

Thus, Theorem 2 provides an additional restriction on the sources (in a pairwise manner) which is required when the estimated dependent sources across all datasets are to be ‘aligned’.

2.5.1 Special Cases

It is now insightful to consider important special cases of IVA with regard to the identification conditions. We begin by considering the case when the T samples are iid. This is equivalent to having $T = 1$, which implies that the identification conditions can be derived as a special case of Theorem 1.

Theorem 3 (IVA Nonidentifiability with iid Samples)**.** *The sources cannot be identified iff* $\exists\, \boldsymbol{\alpha} \neq \emptyset$ *and* $\exists\, m \neq n$ *such that* $\mathbf{s}_m$ *and* $\mathbf{s}_n$ *have* $\boldsymbol{\alpha}$-Gaussian *components and* $\mathbf{R}_{m,\alpha} = \mathbf{D}\mathbf{R}_{n,\alpha}\mathbf{D} \in \mathbb{R}^{K_\alpha \times K_\alpha}$*, where* $\mathbf{D}$ *is any full rank diagonal matrix.*

Another special case of interest is when $K = 1$, yielding the same formulation as ICA assuming sample-to-sample dependence, i.e., not iid samples, the most general form for real-valued ICA.

Theorem 4 (ICA Nonidentifiability [28], [4])**.** *The sources cannot be identified iff* $\exists\, m \neq n$ *such that* $\mathbf{s}_m \in \mathbb{R}^T$ *and* $\mathbf{s}_n \in \mathbb{R}^T$ *are* Gaussian *and* $\mathbf{R}_m = \delta^2 \mathbf{R}_n \in \mathbb{R}^{T \times T}$*, where* $\delta \neq 0$*.*

It can be verified that the identification conditions of Theorem 4 are consistent with the results found in [28, Chapter 4] and [4].

Another special case of interest is when $K = 1$, and assuming iid samples.

Theorem 5 (ICA Nonidentifiability with iid Samples [27])**.** *The sources cannot be identified iff* $\exists\, m \neq n$ *such that* $s_m \in \mathbb{R}$ *and* $s_n \in \mathbb{R}$ *are Gaussian.*

The claim of Theorem 5, originally given in [27], states the well known result for ICA that at most one source can be Gaussian for identification of all iid sources. Algorithms based on the iid assumption using higher-order statistics have been the most widely exploited type of diversity in the derivation of ICA algorithms.

Additional *diversity* can extend the IVA and ICA identification conditions. An example is when data is complex valued. In this work, we only consider the identifability and performance bounds for real-valued IVA. However, we do introduce algorithms for complex-valued IVA in Chapter 4.

In summary, in this section we establish the conditions when the FIM for IVA is invertible. In the next section, we derive the performance bounds under these identifiable (invertible) conditions.

2.6 Cramér-Rao Lower Bound

The variance of the errors in the estimated parameters within a maximum likelihood formulation are lower bounded by the CRLB. Since IVA has been cast in a maximum likelihood framework, we can consider the CRLB of the demixing matrix $\mathbf{W}$. The CRLB associated with the parameter vector $\boldsymbol{\Theta}$ is the inverse of the FIM, i.e., $\operatorname{cov}\left\{\hat{\boldsymbol{\Theta}}\right\} \geq \mathbf{F}^{-1}$, where $\hat{\boldsymbol{\Theta}}$ is an estimator for $\boldsymbol{\Theta}$. Due to the block diagonal structure of (A.9) we have that the inverse (if it exists, see Section 2.5) of the portion of the FIM associated with the mth and nth source, denoted by $\mathbf{F}_{m,n}$

in (2.11), is

$$\mathbf{F}_{m,n}^{-1} = \frac{1}{T}\begin{bmatrix} \left(\boldsymbol{\mathcal{K}}_{m,n} - \boldsymbol{\mathcal{K}}_{n,m}^{-1}\right)^{-1} & * \\ * & \left(\boldsymbol{\mathcal{K}}_{n,m} - \boldsymbol{\mathcal{K}}_{m,n}^{-1}\right)^{-1} \end{bmatrix},\ 1 \leq m < n \leq N,$$

where $*$ denote block matrices of no consequence here.

It yields the following CRLB on the estimates of the demixing matrix quantities,

$$\operatorname{var}\left\{w_{m,n}^{[k]}\right\} \geq \frac{1}{T}\mathbf{e}_k^{\mathsf{T}}\left(\boldsymbol{\mathcal{K}}_{m,n} - \boldsymbol{\mathcal{K}}_{n,m}^{-1}\right)^{-1}\mathbf{e}_k,\ 1 \leq m \neq n \leq N.$$

The expression above is of little direct interest since it only holds when $\mathbf{A} = \mathbf{I}$. Fortunately, we can exploit the results of [119], which makes clear that an iCRLB on the global demixing-mixing matrix, $\mathbf{G} = \mathbf{W}\mathbf{A}$, can be found which is invariant to $\mathbf{A}$, i.e., the CRLB for $\mathbf{G}$ depends only on the nature of the sources. A performance metric that is of direct interest is the interference to source ratio (ISR). The definition of the ISR is the same as in BSS [119, 28], namely[3]:

$$\mathrm{ISR}_{m,n}^{[k]} \triangleq E\left\{\left(g_{m,n}^{[k]}\right)^2\right\}\frac{E\left\{\left|\mathbf{s}_n^{[k]}\right|^2\right\}}{E\left\{\left|\mathbf{s}_m^{[k]}\right|^2\right\}},\ 1 \leq m \neq n \leq N, \tag{2.30}$$

where $g_{m,n}^{[k]} = \mathbf{e}_m^{\mathsf{T}}\mathbf{G}^{[k]}\mathbf{e}_n$ and $\mathbf{G}^{[k]} \triangleq \mathbf{W}^{[k]}\mathbf{A}^{[k]}$ is called the kth global demixing-mixing matrix.

---

[3] For more discussions on ISR see Section 2.7.1.

The iCRLB for ISR is then:

$$\mathrm{ISR}_{m,n}^{[k]} \geq \frac{1}{T}\mathbf{e}_k^{\mathsf{T}}\left(\boldsymbol{\mathcal{K}}_{m,n} - \boldsymbol{\mathcal{K}}_{n,m}^{-1}\right)^{-1}\mathbf{e}_k \frac{E\left\{\left|\mathbf{s}_n^{[k]}\right|^2\right\}}{E\left\{\left|\mathbf{s}_m^{[k]}\right|^2\right\}},\ 1 \leq m \neq n \leq N. \tag{2.31}$$

Since the *sources* are (potentially) multivariate in the IVA formulation, it makes sense to define the ISR according to

$$\mathrm{ISR}_{m,n} \triangleq \sum_{k=1}^{K} \mathrm{ISR}_{m,n}^{[k]},\ 1 \leq m \neq n \leq N.$$

After some simple manipulation, the following compact form for the iCRLB results:

$$\mathrm{ISR}_{m,n} \geq \frac{1}{T}\mathrm{tr}\left(\left(\boldsymbol{\mathcal{K}}_{m,n} - \boldsymbol{\mathcal{K}}_{n,m}^{-1}\right)^{-1} \circ \mathbf{C}_n \oslash \mathbf{C}_m\right),\ 1 \leq m \neq n \leq N,$$

where $\mathbf{C}_n \triangleq E\left\{\mathbf{S}_n \mathbf{S}_n^{\mathsf{T}}\right\} \in \mathbb{R}^{K \times K}$. In what follows, for notational simplicity and without loss of generality, we assume the sources have equal energy within each dataset, i.e., $\mathrm{diag}\left(\mathbf{C}_n\right) = \mathrm{diag}\left(\mathbf{C}_m\right)\ \forall\ 1 \leq m, n \leq N$.

2.6.1 CRLB with iid Samples

When the samples are iid, then the IVA iCRLB simplifies further if we note that:

$$\mathbf{R}_n^{[k_1,k_2]} = E\left\{\mathbf{s}_n^{[k_1]}\left(\mathbf{s}_n^{[k_2]}\right)^{\mathsf{T}}\right\} = E\left\{s_n^{[k_1]}(t)\, s_n^{[k_2]}(t)\right\}\mathbf{I}_T = \sigma_n^{[k_1,k_2]}\mathbf{I}_T, \tag{2.32}$$

$$\mathbf{\Gamma}_m^{[k_1,k_2]} = E\left\{\boldsymbol{\phi}_m^{[k_1]}\left(\boldsymbol{\phi}_m^{[k_2]}\right)^{\mathsf{T}}\right\} = E\left\{\phi_{m,t}^{[k_1]}\phi_{m,t}^{[k_2]}\right\}\mathbf{I}_T = \gamma_m^{[k_1,k_2]}\mathbf{I}_T, \tag{2.33}$$

and

$$\begin{aligned}\mathcal{K}_{m,n}^{[k_1,k_2]} &= \frac{1}{T}\mathrm{tr}\left(\mathbf{\Gamma}_m^{[k_2,k_1]}\mathbf{R}_n^{[k_1,k_2]}\right)\\ &= \frac{1}{T}\mathrm{tr}\left(\gamma_m^{[k_1,k_2]}\mathbf{I}_T\sigma_n^{[k_1,k_2]}\mathbf{I}_T\right)\\ &= \gamma_m^{[k_1,k_2]}\sigma_n^{[k_1,k_2]},\ 1 \le m \ne n \le N, \end{aligned} \tag{2.34}$$

where $\sigma_n^{[k_1,k_2]} \triangleq E\left\{s_n^{[k_1]}(t)\, s_n^{[k_2]}[t]\right\} \in \mathbb{R}$ and $\gamma_m^{[k_1,k_2]} \triangleq E\left\{\phi_m^{[k_1]}(t)\,\phi_m^{[k_2]}(t)\right\} \in \mathbb{R}$ are not dependent on t due to the iid assumption.

For the iid IVA discussion we simplify by replacing the SCM notation with SCV notation, i.e., we define the SCV, $\mathbf{s}_n$, as a random vector with T realizations denoted by $\mathbf{s}_n(t) \in \mathbb{R}^K$. In addition, the multivariate score function is denoted by $\boldsymbol{\phi}_m(\mathbf{s}_m) \in \mathbb{R}^K$. For now, let $\mathbf{R}_n = E\left\{\mathbf{s}_n\mathbf{s}_n^{\mathsf{T}}\right\} \in \mathbb{R}^{K\times K}$ and $\mathbf{\Gamma}_m = E\left\{\boldsymbol{\phi}_m(\mathbf{s}_m)\,\boldsymbol{\phi}_m^{\mathsf{T}}(\mathbf{s}_m)\right\} \in \mathbb{R}^{K\times K}$, from which we observe that $\boldsymbol{\mathcal{K}}_{m,n} = \mathbf{\Gamma}_m \circ \mathbf{R}_n = \mathrm{var}\left\{\boldsymbol{\phi}_m(\mathbf{s}_m)\circ\mathbf{s}_n\right\}$ for $m \ne n$.

The above gives the following iCRLB on the estimates of the demixing matrix entries when the samples are iid,

$$\mathrm{ISR}_{m,n} \ge \frac{1}{T}\mathrm{tr}\left(\left(\mathbf{\Gamma}_m\circ\mathbf{R}_n - \left(\mathbf{\Gamma}_n\circ\mathbf{R}_m\right)^{-1}\right)^{-1}\right), 1 \le m \ne n \le N.$$

The relationship between $\mathbf{\Gamma}$ and $\mathbf{R}$ given in the following lemma is the multivariate extension of the result given by [93, Lemma 1b of Appendix B], which has also been given in [28, Chapter 4].

Lemma 2. $\boldsymbol{\Gamma} \succeq \mathbf{R}^{-1}$, *with equality iff* $\boldsymbol{\phi} = \mathbf{R}^{-1}\mathbf{s}$, *i.e.,* $\mathbf{s}$ *follows the Gaussian distribution.*

Proof. The proof applies the extension of the Cauchy-Schwarz inequality for covariance matrices as given in [63]. Specifically, $\boldsymbol{\Gamma} - E\left\{\boldsymbol{\phi}\mathbf{s}^{\mathsf{T}}\right\}\mathbf{R}^{-1}E\left\{\mathbf{s}\boldsymbol{\phi}^{\mathsf{T}}\right\} \succeq \mathbf{0}$, with equality iff $\boldsymbol{\phi} = E\left\{\boldsymbol{\phi}\mathbf{s}^{\mathsf{T}}\right\}\mathbf{R}^{-1}\mathbf{s}$. By noting that $E\left\{\mathbf{s}\boldsymbol{\phi}^{\mathsf{T}}\right\} = \mathbf{I}$ we arrive at the assertion. □

From this lemma, we see that a measure of non-Gaussianity (or higher-order statistics) is captured by the ‘difference’ between $\boldsymbol{\Gamma}$ and $\mathbf{R}^{-1}$. Next, we show for elliptical distributions—a broad class of source distributions—how this non-Gaussianity measure can be captured by a scalar quantity.

2.6.1.1 Elliptical Distribution

The pdf (assuming it exists) for a zero-mean random vector following the elliptical (contoured) distribution is

$$p\left(\mathbf{x}\right) = \frac{c_K}{\sqrt{\det \boldsymbol{\Sigma}}} h_e\left(\mathbf{x}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}\right), \tag{2.35}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{K\times K}$ is the positive definite matrix frequently termed the dispersion matrix, h_e is some nonnegative function, and c_K denotes the constant that makes (2.35) integrate to one. If the covariance matrix, $E\left\{\mathbf{x}\mathbf{x}^{\mathsf{T}}\right\} = \mathbf{R}$, exists, then for any elliptical distribution it is a scalar multiple of the dispersion matrix, i.e., $\mathbf{R} = \rho\boldsymbol{\Sigma}$, where

$\rho > 0$. Then the score function, $\boldsymbol{\phi}(\mathbf{x}) \triangleq -\partial \log p(\mathbf{x}) / \partial \mathbf{x} = g\left(\mathbf{x}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{x}\right) \boldsymbol{\Sigma}^{-1} \mathbf{x}$, where

$$g(u) = -2 \frac{1}{h_e(u)} \frac{d h_e(u)}{du}.$$

Before proceeding, we show that the score function covariance matrix, $\boldsymbol{\Gamma} = E\left\{\boldsymbol{\phi}\boldsymbol{\phi}^{\mathsf{T}}\right\}$, is a scalar multiple of the inverse of the covariance matrix, for all elliptical distributions defined by (2.35). We begin by letting $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}\mathbf{x}$ so that $p_{\mathbf{z}}(\mathbf{z}) = \left|\det \boldsymbol{\Sigma}^{-1/2}\right|^{-1} p_{\mathbf{x}}\left(\boldsymbol{\Sigma}^{1/2}\mathbf{z}\right) = c_K h_e\left(\mathbf{z}^{\mathsf{T}}\mathbf{z}\right)$, which results in

$$\boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{-1} E\left\{g^2\left(\mathbf{x}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}\right)\mathbf{x}\mathbf{x}^{\mathsf{T}}\right\}\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-1/2} E\left\{g^2\left(\mathbf{z}^{\mathsf{T}}\mathbf{z}\right)\mathbf{z}\mathbf{z}^{\mathsf{T}}\right\}\boldsymbol{\Sigma}^{-1/2}. \tag{2.36}$$

To compute the expectation requires the following multivariate integral to be evaluated:

$$E\left\{g^2\left(\mathbf{z}^{\mathsf{T}}\mathbf{z}\right) z_l z_k\right\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g^2\left(\mathbf{z}^{\mathsf{T}}\mathbf{z}\right) z_l z_k p(\mathbf{z})\, dz_1 \ldots dz_K. \tag{2.37}$$

We use a transformation of variables utilized for similar problems in [88, 14], namely,

$$z_1 = r \prod_{k=1}^{K-1} \sin\theta_k \tag{2.38}$$

$$z_j = r \left(\prod_{k=1}^{K-j} \sin\theta_k\right) \cos\theta_{K-j+1}, 2 \le j \le K-1 \tag{2.39}$$

$$z_K = r\cos\theta_1 \tag{2.40}$$

where $0 < \theta_j \leq \pi$, $j = 1, \ldots, K-2$, $0 < \theta_{K-1} \leq 2\pi$, $0 < r \leq \infty$. By noting that $\mathbf{z}^{\mathsf{T}}\mathbf{z} = r^2$ and the Jacobian of the transformation from $\mathbf{z}$ to $[\theta_1 \ldots \theta_{K-1}\, r]^{\mathsf{T}}$ is $r^{K-1}\sin^{K-2}\theta_1 \sin^{K-3}\theta_2 \cdots \sin\theta_{K-2} = r^{K-1}\prod_{k=1}^{K-2}(\sin\theta_k)^{K-1-k}$, we have $p(r) = c_K h_e(r^2)$.

There are two cases, $l = k$ and $l \neq k$, required to evaluate (2.37). Let us consider the former first,

$$\begin{aligned}
E\left\{g^2\left(\mathbf{z}^{\mathsf{T}}\mathbf{z}\right)z_1^2\right\} &= \int_0^\infty \int_0^{2\pi} \int_0^\pi \cdots \int_0^\pi g^2(r^2)p(r)r^2\left(\prod_{k=1}^{K-1}\sin^2\theta_k\right) \\
&\qquad \left(r^{K-1}\prod_{k=1}^{K-2}(\sin\theta_k)^{K-1-k}\right) d\theta_{K-1}d\theta_{K-2}\cdots d\theta_1 dr \\
&= \int_{0,0,0,\ldots,0}^{\infty,2\pi,\pi,\ldots,\pi} g^2(r^2)p(r)r^{K+1}\sin^2\theta_{K-1}\left(\prod_{k=1}^{K-2}(\sin\theta_k)^{K+1-k}\right) d\theta_{K-1}d\theta_{K-2}\cdots d\theta_1 dr \\
&= \left(\int_0^\infty g^2(r^2)r^{K+1}p(r)\,dr\right)\left(\int_0^{2\pi}\sin^2\theta_{K-1}d\theta_{K-1}\right)\prod_{k=1}^{K-2}\int_0^\pi (\sin\theta_k)^{K+1-k}d\theta_k \\
&= E\left\{g^2(r^2)r^{K+1}\right\}\pi\prod_{k=1}^{K-2}\pi^{1/2}\frac{\Gamma\left((K+2-k)/2\right)}{\Gamma\left((K+3-k)/2\right)} \\
&= E\left\{g^2(r^2)r^{K+1}\right\}\pi\pi^{(K-2)/2}\frac{1}{\Gamma\left((K+2)/2\right)} \\
&= E\left\{g^2(r^2)r^{K+1}\right\}\frac{2\pi^{K/2}}{K\Gamma(K/2)}, \qquad (2.41)
\end{aligned}$$

where we have made use of $\int_0^\pi \sin^n\theta d\theta = \sqrt{\pi}\Gamma\left[(n+1)/2\right]/\Gamma\left[(n+2)/2\right]$ when $n \geq 1$.

Now for the off-diagonal terms, e.g., when $l = K-1$, $k = K$, we have $z_K =$

$r\cos\theta_1$ and $z_{K-1} = r\cos\theta_2\sin\theta_1$ for $K > 2$,

$$E\left\{g^2\left(\mathbf{z}^\mathsf{T}\mathbf{z}\right)z_K z_{K-1}\right\} = \int_0^\infty\int_0^{2\pi}\int_0^\pi\cdots\int_0^\pi g^2(r^2)p(r)r\cos\theta_1 r\cos\theta_2 \left(r^{K-1}\prod_{k=1}^{K-2}\left(\sin\theta_k\right)^{K-1-k}\right)d\theta_{K-1}d\theta_{K-2}\cdots d\theta_1 dr$$

$$= \int_0^\infty\int_0^{2\pi}\int_0^\pi\cdots\int_0^\pi g^2(r^2)p(r)r^{K+1}\cos\theta_1\left(\sin\theta_1\right)^{K-2} \cos\theta_2\left(\prod_{k=2}^{K-2}\left(\sin\theta_k\right)^{K-1-k}\right)d\theta_{K-1}d\theta_{K-2}\cdots d\theta_1 dr$$

$$= 0,$$

and for $K = 2$, $E\left\{g^2\left(\mathbf{z}^\mathsf{T}\mathbf{z}\right)z_1 z_2\right\} = \int_0^\infty\int_0^{2\pi} g^2(r^2)p(r)r\cos\theta\sin\theta d\theta dr = 0$, where we have used the following $\int_0^{n\pi}\cos\left(\theta\right)\sin^n\left(\theta\right)d\theta = 0$ when $n \in \mathbb{N}$. The result holds for the more general case when $l \neq k$ and we arrive at the final expression of

$$\begin{aligned}
\mathbf{\Gamma} &= \mathbf{\Sigma}^{-1/2}E\left\{g^2\left(\mathbf{z}^\mathsf{T}\mathbf{z}\right)\mathbf{z}\mathbf{z}^\mathsf{T}\right\}\mathbf{\Sigma}^{-1/2} \\
&= E\left\{g^2\left(\mathbf{z}^\mathsf{T}\mathbf{z}\right)\mathbf{z}\mathbf{z}^\mathsf{T}\right\}\mathbf{\Sigma}^{-1} \\
&= E\left\{g^2(r^2)r^{K+1}\right\}\frac{2\pi^{K/2}}{K\Gamma\left(K/2\right)}\rho\mathbf{R}^{-1}, K \geq 2 \\
&= \kappa\mathbf{R}^{-1}, K \geq 2, \qquad (2.42)
\end{aligned}$$

where $\kappa \triangleq E\left\{g^2(r^2)r^{K+1}\right\}\frac{2\pi^{K/2}}{K\Gamma(K/2)}\rho$.

By application of Lemma 2 this implies that $\kappa \geq 1$ with equality iff the random variable is normally distributed[4]. Therefore, the iCRLB for ISR with elliptical

---

[4] Under the Gaussian SCV data-model assumption, $E\left\{\boldsymbol{\phi}\boldsymbol{\phi}^\mathsf{T}\right\} = E\left\{\mathbf{R}^{-1}\mathbf{s}\mathbf{s}^\mathsf{T}\mathbf{R}^{-1}\right\} = \mathbf{R}^{-1}$.

sources is

$$\mathrm{ISR}_{m,n} \geq \frac{1}{T}\mathrm{tr}\left(\left(\kappa_m \mathbf{R}_m^{-1} \circ \mathbf{R}_n - \left(\kappa_n \mathbf{R}_n^{-1} \circ \mathbf{R}_m\right)^{-1}\right)^{-1}\right), 1 \leq m \neq n \leq N. \quad (2.43)$$

For this performance bound we provide the following theorem.

Theorem 6. *If two SCVs follow distributions from the elliptical family with covariance matrices,* $\mathbf{R}_m$ *and* $\mathbf{R}_n$*, then* $ISR_{m,n}$ *is less than or equal to the* $ISR_{m,n}$ *associated with Gaussian SCVs having the same covariance matrices.*

Proof. For elliptically distributed sources, via Lemma 2, we have that $\kappa_m \geq 1$ and $\kappa_n \geq 1$, thus

$$\kappa_m \mathbf{R}_m^{-1} \circ \mathbf{R}_n - \kappa_n^{-1}\left(\mathbf{R}_n^{-1} \circ \mathbf{R}_m\right)^{-1} \succeq \mathbf{R}_m^{-1} \circ \mathbf{R}_n - \left(\mathbf{R}_n^{-1} \circ \mathbf{R}_m\right)^{-1}$$
$$\left(\kappa_m \mathbf{R}_m^{-1} \circ \mathbf{R}_n - \kappa_n^{-1}\left(\mathbf{R}_n^{-1} \circ \mathbf{R}_m\right)^{-1}\right)^{-1} \preceq \left(\mathbf{R}_m^{-1} \circ \mathbf{R}_n - \left(\mathbf{R}_n^{-1} \circ \mathbf{R}_m\right)^{-1}\right)^{-1},$$

and since $\mathbf{A} \preceq \mathbf{B}$, it implies $\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} \leq \mathbf{x}^{\mathsf{T}}\mathbf{B}\mathbf{x}, \forall \mathbf{x}$, and thus $\mathrm{tr}(\mathbf{A}) \leq \mathrm{tr}(\mathbf{B})$. □

A special case, which arrives at a form directly analogous to the ICA form, occurs when $\mathbf{R}_m = \mathbf{R}_n = \mathbf{I}_K$:

$$\mathrm{ISR}_{m,n} \geq \frac{K}{T}\frac{\kappa_n}{\kappa_m \kappa_n - 1}, 1 \leq m \neq n \leq N. \quad (2.44)$$

This expression clearly shows how for second-order uncorrelated elliptical sources, the measure of non-Gaussianity expressed by κ, directly determines the source separation performance. The bound is finite as long as $\kappa_m \neq \kappa_n^{-1}$ and thus when

$(\kappa_m, \kappa_n) \neq (1,1)$ holds, i.e., as long at least one source is non-Gaussian. This result is consistent with the identification conditions. In fact, as shown in the following theorem, the same statement holds for second-order correlated elliptical sources.

Theorem 7. *If three SCVs follow distributions from the elliptical family with covariance matrices, $\mathbf{R}_m = \mathbf{R}_{m'}$, and $\mathbf{R}_n$, and $\kappa_m \geq \kappa_{m'}$ then ISR$_{m,n} \leq$ ISR$_{m',n}$.*

Proof. For elliptically distributed sources, via Lemma 2, we have that $\kappa_m \geq \kappa_{m'} \geq 1$ and $\kappa_n \geq 1$, thus

$$\kappa_m \mathbf{R}_m^{-1} \circ \mathbf{R}_n - \kappa_n^{-1} \left(\mathbf{R}_n^{-1} \circ \mathbf{R}_m\right)^{-1} \succeq \kappa_{m'} \mathbf{R}_m^{-1} \circ \mathbf{R}_n - \kappa_n^{-1} \left(\mathbf{R}_n^{-1} \circ \mathbf{R}_m\right)^{-1}$$

$$\left(\kappa_m \mathbf{R}_m^{-1} \circ \mathbf{R}_n - \kappa_n^{-1} \left(\mathbf{R}_n^{-1} \circ \mathbf{R}_m\right)^{-1}\right)^{-1} \preceq \left(\kappa_{m'} \mathbf{R}_m^{-1} \circ \mathbf{R}_n - \kappa_n^{-1} \left(\mathbf{R}_n^{-1} \circ \mathbf{R}_m\right)^{-1}\right)^{-1},$$

and since $\mathbf{A} \preceq \mathbf{B}$ implies $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} \leq \mathbf{x}^\mathsf{T}\mathbf{B}\mathbf{x}, \forall \mathbf{x}$, and thus $\operatorname{tr}(\mathbf{A}) \leq \operatorname{tr}(\mathbf{B})$. □

2.6.1.2 Kotz Distribution

The first IVA algorithm [58], IVA with second-order uncorrelated multivariate Laplace distribution model (IVA-Lap), uses the multivariate score function given by $\boldsymbol{\phi}_n = \mathbf{y}_n / \sqrt{\mathbf{y}_n^\mathsf{T}\mathbf{y}_n}$. Note that the score function used in IVA-Lap requires no distribution parameters. In the following section, we utilize the Kotz distribution family to greatly expand the available multivariate SCV distribution models for use in IVA.

The original introduction of the Kotz distribution [61] has been further analyzed in the more readily accessible work of [89]. The zero-mean K-dimensional real-valued Kotz distribution has the following pdf

$$p\left(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\Sigma}\right)=\frac{\beta\lambda^{\nu}\Gamma\left(K/2\right)\left(\mathbf{x}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}\right)^{\eta-1}}{\sqrt{\pi^{K}}\det\boldsymbol{\Sigma}\Gamma\left(\nu\right)}\exp\left(-\lambda\left(\mathbf{x}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}\right)^{\beta}\right),$$

where $\boldsymbol{\Sigma}\in\mathbb{R}^{K\times K}$ (dispersion matrix) is positive-definite, $\boldsymbol{\theta}=\left[\beta,\eta,\lambda\right]^{\mathsf{T}}$ denotes the scalar Kotz distribution parameters, namely, $\lambda>0$ (kurtosis parameter), $\beta>0$ (shape parameter), $\eta>\left(2-K\right)/2$ (hole parameter), and we use $\nu\triangleq\left(2\eta+K-2\right)/\left(2\beta\right)>0$ for more compact notations. The hole parameter, η, can be used to make the modes of the distribution occur at $\mathbf{x}\neq\mathbf{0}$ when $\eta>1$. We use $\mathbf{x}\sim\operatorname{Kotz}\left(\boldsymbol{\Sigma},\beta,\eta,\lambda\right)$ to say that $\mathbf{x}\in\mathbb{R}^{K}$ follows a K-dimensional Kotz distribution. It is shown in [89] that the dispersion matrix is related to the covariance matrix via $E\left\{\mathbf{x}\mathbf{x}^{\mathsf{T}}\right\}=\rho_{\text{Kotz}}\boldsymbol{\Sigma}$, where

$$\rho_{\text{Kotz}}\triangleq\frac{\lambda^{-1/\beta}}{K}\frac{\Gamma\left(\nu+1/\beta\right)}{\Gamma\left(\nu\right)}. \tag{2.45}$$

Clearly, the Kotz is an elliptical distribution with $h\left(u\right)=u^{\eta-1}e^{-\lambda u^{\beta}}$, $c_K=\frac{\beta\lambda^{\nu}\Gamma(K/2)}{\sqrt{\pi^{K}}}$, $g\left(u\right)=2\left(1-\eta+\lambda\beta u^{\beta}\right)/u$, and multivariate score function

$$\boldsymbol{\phi}\left(\mathbf{x}\right)=2\left(1-\eta+\lambda\beta\left(\mathbf{x}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}\right)^{\beta}\right)\left(\mathbf{x}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}\right)^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{x}. \tag{2.46}$$

Using the results for elliptical distributions given in (2.42), we can compute the non-Gaussianity as measured by κ to be

$$
\begin{aligned}
\kappa_{\text{Kotz}} =& E\left\{g^2(r^2)r^{K+1}\right\}\frac{2\pi^{K/2}}{K\Gamma\left(K/2\right)}\rho_{\text{Kotz}} \\
=& \frac{\rho_{\text{Kotz}}c_K 2\pi^{K/2}}{K\Gamma\left(K/2\right)}\int_0^\infty r^{2(\eta-1)}e^{-\lambda r^{2\beta}}g^2(r^2)r^{K+1}dr \\
=& \frac{8\beta\lambda^{\nu-1/\beta}\Gamma\left(\nu+1/\beta\right)}{K^2\Gamma^2\left(\nu\right)} \\
& \qquad \int_0^\infty r^{K-5+2\eta}e^{-\lambda r^{2\beta}}\left(\left(\eta-1\right)^2+2\left(\eta-1\right)\lambda\beta r^{2\beta}+\lambda^2\beta^2r^{4\beta}\right)dr \\
=& \frac{4\Gamma\left(\nu+1/\beta\right)}{K^2\Gamma^2\left(\nu\right)}\left(\left(1-\eta\right)^2\Gamma\left(\nu-1/\beta\right)+2\beta\left(1-\eta\right)\Gamma\left(\nu-1/\beta+1\right)+\right. \\
& \qquad \left.\beta^2\Gamma\left(\nu-1/\beta+2\right)\right) \\
=& \frac{4\Gamma\left(\nu-1/\beta\right)\Gamma\left(\nu+1/\beta\right)}{K^2\Gamma^2\left(\nu\right)}\left(\left(1-\eta\right)^2+2\beta\left(1-\eta\right)\left(\nu-1/\beta\right)+\right. \\
& \qquad \left.\beta^2\left(\nu-1/\beta\right)\left(\nu-1/\beta+1\right)\right), \qquad (2.47)
\end{aligned}
$$

where the last two equations hold only when $\eta > \left(4-K\right)/2$. The quantity κ_{Kotz} is shown versus β for three values of η in Fig. 2.2.

2.6.1.3 Multivariate Power Exponential Distribution

The set of distributions achieved by varying the Kotz parameters is vast and includes the multivariate power exponential (MPE), which is sometimes called a multivariate generalized Gaussian distribution [43]. The MPE distributions are given when the Kotz parameters $\eta = 1$ and $\lambda = 1/2$, i.e., $\mathbf{x} \sim \text{MPE}\left(\mathbf{\Sigma}, \beta, \eta = 1, \lambda = \frac{1}{2}\right) =$

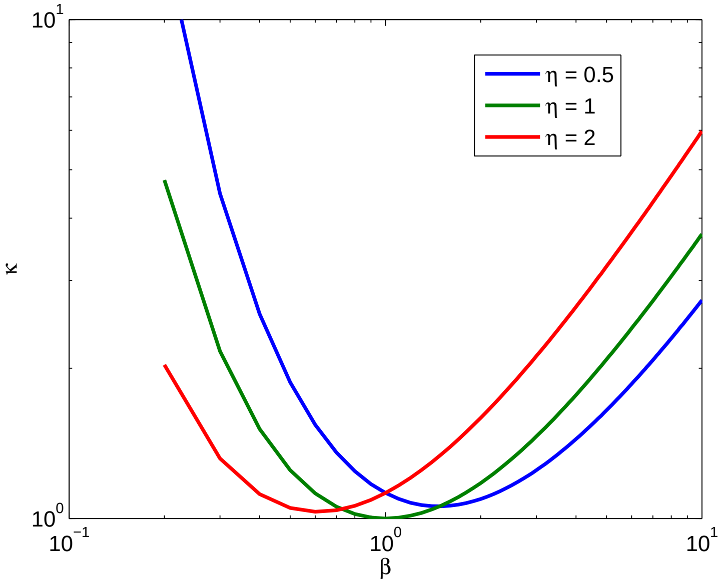


Figure 2.2: The 'degree' of non-Gaussianity as expressed by the κ_{Kotz} computed in (2.47) for the Kotz distribution with $K = 4$.

Kotz $\left(\mathbf{\Sigma}, \beta, \eta = 1, \lambda = \frac{1}{2}\right)$. For the MPE,

$$\boldsymbol{\phi}\left(\mathbf{x}\right) = \beta\left(\mathbf{x}^{\mathsf{T}}\mathbf{\Sigma}^{-1}\mathbf{x}\right)^{\beta-1}\mathbf{\Sigma}^{-1}\mathbf{x}, \tag{2.48}$$

$\nu = K/\left(2\beta\right)$, and

$$\begin{aligned}\kappa_{\text{MPE}} &\triangleq \left(\frac{2\beta}{K\Gamma\left(\frac{K}{2\beta}\right)}\right)^2 \Gamma\left(\frac{K-2+4\beta}{2\beta}\right)\Gamma\left(\frac{K+2}{2\beta}\right), K \geq 2 \\ &= \frac{\Gamma\left(\nu - 1/\beta\right)\Gamma\left(\nu + 1/\beta\right)}{\nu^2\Gamma^2\left(\nu\right)}\left(\nu - 1/\beta\right)\left(\nu - 1/\beta + 1\right), K \geq 2. \end{aligned}\tag{2.49}$$

For κ_{MPE}, we provide an alternative proof of the lower bound using Gurland's ratio [44]. In the proof, we also show that for $K = 2$, κ_{MPE} is monotonically increasing as β goes away from one.

Theorem 8 (MPE lower bound)**.** *If $\beta > 0$ and $K \geq 2$ then*

$$\kappa_{MPE} = \left(\frac{2\beta}{K\Gamma\left(\frac{K}{2\beta}\right)}\right)^2 \Gamma\left(\frac{K-2+4\beta}{2\beta}\right)\Gamma\left(\frac{K+2}{2\beta}\right) \geq 1, \tag{2.50}$$

where the equality holds only when $\beta = 1$.

Proof of κ_{MPE} lower bound. We make use of an inequality first shown in [44] involving Gurland's ratio,

$$T\left(u,v\right) \triangleq \frac{\Gamma\left(u\right)\Gamma\left(v\right)}{\Gamma^2\left(\left(u+v\right)/2\right)}, \quad u,v > 0,$$

namely $T\left(u-\gamma, u+\gamma\right) \geq 1 + \frac{\gamma^2}{u-\gamma}, \quad u > |\gamma|$, where the equality holds only when $\gamma = 1$, [86].

For $K > 2$, let $z = K/2 - 1 > 0$, then

$$\begin{aligned}
\kappa_{\mathrm{MPE}} &= \frac{1}{\nu^2\Gamma^2\left(\nu\right)}\Gamma\left(\nu - 1/\beta + 2\right)\Gamma\left(\nu + 1/\beta\right) \\
&= \frac{\left(\nu - 1/\beta\right)\left(\nu - 1/\beta + 1\right)}{\nu^2}\frac{\Gamma\left(\nu - 1/\beta\right)\Gamma\left(\nu + 1/\beta\right)}{\Gamma^2\left(\nu\right)} \\
&= \frac{\left(\nu - 1/\beta\right)\left(\nu - 1/\beta + 1\right)}{\nu^2}T\left(\nu - 1/\beta, \nu + 1/\beta\right) \\
&\geq \frac{\left(\nu - 1/\beta\right)\left(\nu - 1/\beta + 1\right)}{\nu^2}\left(1 + \frac{\beta^{-2}}{\nu - 1/\beta}\right) \\
&= \frac{\left(\nu - 1/\beta\right)\left(\nu - 1/\beta + 1\right)}{\nu^2}\left(\frac{\nu - 1/\beta + \beta^{-2}}{\nu - 1/\beta}\right) \\
&= \beta^2\frac{z/\beta\left(z/\beta + 1\right)}{\left(z+1\right)^2}\left(\frac{z/\beta + \beta^{-2}}{z/\beta}\right) \\
&= \frac{\left(z\beta^{-1} + 1\right)\left(z\beta + 1\right)}{\left(z+1\right)^2}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{(z+1)^2 + z\left(\beta^{-1} + \beta - 2\right)}{(z+1)^2} \\
&= 1 + \frac{z}{(z+1)^2} \frac{(\beta - 1)^2}{\beta}.
\end{aligned}
$$

Since $\beta > 0$, then $(\beta - 1)^2 \beta^{-1} \geq 0$ and thus $\kappa_{\text{MPE}} \geq 1$. It is clear that the lower bound for κ_{MPE} expressed above is at a minimum when $\beta = 1$ for $\beta > 0$ and the lower bound is one. Furthermore this lower bound is achieved only when $\beta = 1$.

For $K = 2$, we have that

$$
\frac{d\kappa_{\text{MPE}}}{d\beta} = \frac{2\Gamma(2\nu)}{\Gamma^2(\nu)} \left(\Psi(\nu + 1) - \Psi(2\nu)\right),
$$

where $\Psi(x) \triangleq d\ln\Gamma(x)/dx$ is the digamma function. Note that the derivative is positive (negative) when $\Psi(\nu + 1) > (<) \Psi(2\nu)$. Since $\Psi(x)$ is a monotonic nondecreasing function of x [34], then we must have that the arguments have the same relationship, i.e., $\beta \geq (<) 1$ implies $(d\kappa_{\text{MPE}}/d\beta) \geq (<) 0$ with equality only when $\beta = 1$. For any local minima or maxima of κ_{MPE}, the derivative wrt β must be equal to zero, or equivalently $(d\kappa_{\text{MPE}}/d\beta) = 0$, this condition only occurs when $\beta = 1$. Thus for $K = 2$ we have shown that κ_{MPE} is monotonically increasing as β goes away from 1. □

2.6.1.4 Gaussian Distribution

As discussed previously, see Section 2.3.3, the Gaussian source model is important from both a theoretical and practical perspective. The use of the Gaussian distributions is ubiquitous for reasons which are often cited. Let us begin our discussion by considering the iCRLB for Gaussian sources, i.e., (2.43) with $\kappa_m = \kappa_n = 1$,

$$\mathrm{ISR}_{m,n} \geq \frac{1}{T}\mathrm{tr}\left(\left(\boldsymbol{\Sigma}_m^{-1} \circ \boldsymbol{\Sigma}_n - \left(\boldsymbol{\Sigma}_n^{-1} \circ \boldsymbol{\Sigma}_m\right)^{-1}\right)^{-1}\right), 1 \leq m \neq n \leq N, \tag{2.51}$$

where we have used $E\left\{\mathbf{s}_m \mathbf{s}_m^{\mathsf{T}}\right\} = \mathbf{R}_m = \boldsymbol{\Sigma}_m$

The identifiability theorem (see Theorem 3) can be used to make the following claim: $\boldsymbol{\Sigma}_m^{-1} \circ \boldsymbol{\Sigma}_n - \left(\boldsymbol{\Sigma}_n^{-1} \circ \boldsymbol{\Sigma}_m\right)^{-1}$ is positive definite iff there exists no permutation matrix $\mathbf{P}$ and full rank diagonal matrix $\mathbf{D}$ such that $\mathbf{P}\boldsymbol{\Sigma}_n\mathbf{P}^{\mathsf{T}} = \mathbf{P}\mathbf{D}\boldsymbol{\Sigma}_m\mathbf{D}\mathbf{P}^{\mathsf{T}}$ is a block diagonal matrix.

2.6.2 Example of Performance Bounds

We consider a simple case with $N = 2$ sources and $K = 2$ datasets in order to illustrate the performance bounds for IVA. For the first source $\mathbf{s}_1(t) \sim \mathrm{MPE}\left(\boldsymbol{\Sigma} = \rho_{\mathrm{MPE}}^{-1}\mathbf{R}, \beta\right)$, are iid samples, where

$$\mathbf{R} = \begin{bmatrix} 1 & \sigma \\ \sigma & 1 \end{bmatrix}, \quad |\sigma| < 1. \tag{2.52}$$

For the second source, $\mathbf{s}_2(t) = \begin{bmatrix} s_2^{[1]}(t) & s_2^{[2]}(t) \end{bmatrix}^{\mathsf{T}}$, where $s_2^{[1]}(t) \sim \mathcal{N}(0,1)$ with iid samples and $s_2^{[2]}(t) = b s_2^{[2]}(t-1) + z(t)$, $0 \leq b < 1$, and $z(t) \sim \mathcal{N}(0,1)$.

With the above formulation we can demonstrate the three sources of diversity that can be exploited in the IVA framework:

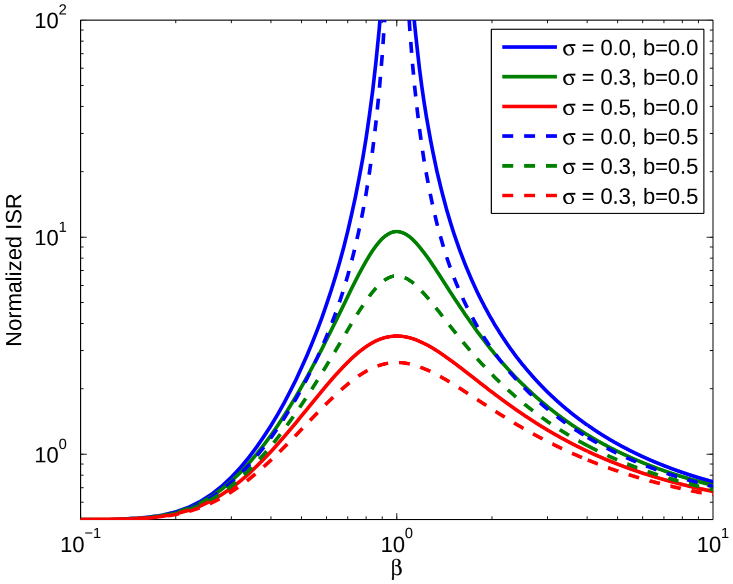


Figure 2.3: The performance bounds for IVA depicting the three types of source diversity that are exploited by real-valued IVA.

- σ: characterizes the dependence captured by the degree of *second-order correlation across datasets*
- β: characterizes the dependence captured by the degree of *non-Gaussianity*
- b: characterizes the dependence captured by the degree of *second-order correlation across samples*

In Fig. 2.3 we see that when the sources are purely Gaussian, $\beta = 1$, we require $\sigma \neq 0$ to achieve a bounded ISR. The ISR is bounded when $\mathbf{s}_1$ is non-Gaussian. The bound on performance improves either as $|b|$ increases or as β goes away from one.

2.6.3 CRLB for ICA

Another special case, which is of particular interest, is when there is only one dataset, i.e., $K = 1$. For this case, the expressions above further simplify to the more extensively studied ICA performance bounds [109, 93, 119, 28]. If $K = 1$, we can replace the SCM notation with source component notation, i.e., let $\mathbf{s}_n \in \mathbb{R}^T$ be the random vector and the multivariate score function be denoted by $\boldsymbol{\phi}_m \in \mathbb{R}^T$. Then, for this section we have $\mathbf{R}_n = E\left\{\mathbf{s}_n \mathbf{s}_n^{\mathsf{T}}\right\} \in \mathbb{R}^{T \times T}$ and $\boldsymbol{\Gamma}_m = E\left\{\boldsymbol{\phi}_m \boldsymbol{\phi}_m^{\mathsf{T}}\right\} \in \mathbb{R}^{T \times T}$, from which we observe that for $m \neq n$,

$$\boldsymbol{\mathcal{K}}_{m,n} = \mathcal{K}_{m,n} = \frac{1}{T} \operatorname{tr}\left(\boldsymbol{\Gamma}_m \mathbf{R}_n\right) = \frac{1}{T} \operatorname{var}\left\{\boldsymbol{\phi}_m^{\mathsf{T}} \mathbf{s}_n\right\}. \tag{2.53}$$

Also:

$$\mathbf{F}_{m,n} = T \begin{bmatrix} \frac{1}{T}\operatorname{var}\left\{\boldsymbol{\phi}_m^{\mathsf{T}} \mathbf{s}_n\right\} & 1 \\ 1 & \frac{1}{T}\operatorname{var}\left\{\boldsymbol{\phi}_n^{\mathsf{T}} \mathbf{s}_m\right\} \end{bmatrix} \in \mathbb{R}^{2 \times 2},\ 1 \leq m < n \leq N. \tag{2.54}$$

Two particular subcases in ICA are of interest. The first case is when the samples are iid, for which $\mathbf{R}_n = \sigma_n^2 \mathbf{I}_T$, $\boldsymbol{\Gamma}_m = E\left\{\phi_m^2\right\} \mathbf{I}_T$, $\mathcal{K}_{m,n} = \sigma_n^2 \kappa_m$, and $\mathbf{C}_{\mathbf{n}} = T\sigma_n^2$ holds, where $\kappa_m \triangleq E\left\{\phi_m^2\right\} \geq \sigma_m^2$. These simplifications result in

$$\mathrm{ISR}_{m,n} \geq \frac{1}{T}\left(\kappa_m \sigma_n^2 - \frac{1}{\kappa_n \sigma_m^2}\right)^{-1} \frac{\sigma_n^2}{\sigma_m^2},\ 1 \leq m \neq n \leq N, \tag{2.55}$$

which for unit variance sources reduces to the same results given in [109, Eq. 38] and [93, Thm. 2], namely:

$$\mathrm{ISR}_{m,n} \geq \frac{1}{T}\left(\kappa_m - \kappa_n^{-1}\right)^{-1} = \frac{1}{T}\frac{\kappa_n}{\kappa_m \kappa_n - 1},\ 1 \leq m \neq n \leq N. \tag{2.56}$$

The second subcase of ICA is for sources with Gaussian sample-to-sample dependence, i.e., $\mathbf{s_n} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{R}_n \in \mathbb{R}^{T \times T}\right)$. Then we have that $\mathbf{\Gamma}_m = \mathbf{R}_m^{-1}$ and $\mathcal{K}_{m,n} = T^{-1}\mathrm{tr}\left(\mathbf{R}_m^{-1}\mathbf{R}_n\right)$, which corresponds to [119, Eq. 19].

2.7 Performance Metrics

In the previous section, we provided a bound on the attainable ISR. In this section, we discuss this performance metric and two additional performance metrics for BSS that are appropriate for IVA. Each metric is described and their respective merits are discussed. All three metrics are based on the global demixing-mixing matrix[5], $\mathbf{G} \triangleq \mathbf{W}\mathbf{A} = \oplus \sum_{k=1}^{K} \mathbf{G}^{[k]}$, where $\mathbf{G}^{[k]} \triangleq \mathbf{W}^{[k]}\mathbf{A}^{[k]}$.

2.7.1 Interference-to-Signal Power Ratio

A common measure of BSS algorithm performance is the ISR. For this JBSS formulation, the definition of the ISR is the same as in BSS [119, 28]. ISR assesses the amount of residual energy from each source that contributes to the energy of

---

[5] The global demixing-mixing matrix has also been termed the global system matrix and global mixing matrix in other references.

each *estimated* source using:

$$\mathrm{ISR}_{m,n}^{[k]} \triangleq E\left\{\left(g_{m,n}^{[k]}\right)^2\right\}\frac{E\left\{\left|\mathbf{s}_n^{[k]}\right|^2\right\}}{E\left\{\left|\mathbf{s}_m^{[k]}\right|^2\right\}},\ 1 \leq m \neq n \leq N, \tag{2.57}$$

where $g_{m,n}^{[k]} = \mathbf{e}_m^\mathsf{T}\mathbf{G}^{[k]}\mathbf{e}_n$. In words, $\mathrm{ISR}_{m,n}^{[k]}$ is the energy of the nth source in the estimate of the mth source in the kth dataset.

ISR is a convenient measure because it is indifferent to the inherent scaling ambiguity in BSS and it is an analytically tractable performance measure that readily admits to derivation of the iCRLB (on ISR) for a wide variety of BSS problems [109, 93, 119].

In order to use the ISR metric, each row of each estimated demixing matrix must be assigned to a source, i.e., *paired*. The pairing can be enforced by applying a permutation matrix to $\mathbf{W}$ to reorder the rows for alignment of sources with their respective estimates. When the source separation performance is sufficiently performed then this pairing is trivial. However, when the source separation is ambiguous a more rigorous approach is required, e.g., see [108].

For IVA, the pairing of sources to their estimates is a more cumbersome task since the permutation matrix applied to each dataset, $\mathbf{P}^{[k]} \in \mathbb{N}^{N\times N}$, has some additional constraints that depend on the nature of the source dependencies across datasets. Two extreme examples help illustrate this:

- If the sources possess no dependencies across datasets, then $\mathbf{P}^{[k]}$ can be different for every dataset.

- If the sources are dependent across all datasets, then the permutation matrix for each dataset must be the same, i.e., $\mathbf{P} = \mathbf{P}^{[k]}$ holds for $k = 1, \ldots, K$.

Obviously, additional examples which lie somewhere in between the two above also exist. In this work, we restrict ourselves to the latter example so that a common permutation matrix is computed and applied to all datasets—otherwise aligning sources would require a permutation algorithm [31, 80]. Thus, for our purposes it will be sufficient to find the *common* permutation matrix which maximizes the sum of traces of the resulting K global demixing-mixing matrices—an admittedly suboptimal procedure. After using this permutation matrix to *globally* permute the order of the sources, we calculate the ISR.

For IVA, an extension of ISR that measures the total energy across the datasets of the nth source in the estimate of the mth source, i.e.,

$$\mathrm{ISR}_{m,n} \triangleq \sum_{k=1}^{K} \mathrm{ISR}_{m,n}^{[k]}, \;\; 1 \leq m \neq n \leq N, \tag{2.58}$$

is particularly useful when computing the iCRLB associated with sources that admit to the common permutation conditions—see Theorem 2.

Lastly, it is useful to define a scalar ISR quantity for plotting and summary performance purposes. We use the average of the ‘off-diagonal’ entries of the ISR matrix, i.e.,

$$\mathrm{ISR}_{\mathrm{norm}} \triangleq \frac{1}{KN\left(N-1\right)} \sum_{\substack{m,n=1 \\ m \neq n}}^{N} \mathrm{ISR}_{m,n}, \tag{2.59}$$

which we call the normalized ISR.

2.7.2 Intersymbol Interference

As discussed above the ISR metric requires the pairing of sources with their estimates due to the inherent source ordering ambiguity of ICA. We now discuss the intersymbol-interference (ISI) performance metric, a metric that does not require the ordering of the sources in order to assess performance. It has been used frequently in ICA performance assessments. The unnormalized version of ISI was adopted from the self-adaptive equalization problem and first introduced to ICA by [87]—ISI has also been called the Amari index due to use of this metric in [6]. A normalization of twice the value used here was then introduced by [26].

$$\mathrm{ISI}\left(\mathbf{G}\right) \triangleq \frac{\sum_{n=1}^{N}\left(\sum_{m=1}^{N}\frac{|g_{n,m}|}{\max_p|g_{n,p}|}-1\right)+\sum_{m=1}^{N}\left(\sum_{n=1}^{N}\frac{|g_{n,m}|}{\max_p|g_{p,m}|}-1\right)}{2N\left(N-1\right)} \tag{2.60}$$

The performance measure is a function of the global demixing-mixing matrix, $\mathbf{G} \triangleq \mathbf{WA}$. The performance measure is normalized based on the number of components, N, so that $0 \leq \mathrm{ISI} \leq 1$, where 0 is optimal. Note that $\mathrm{ISI}\left(\mathbf{G}\right) = \mathrm{ISI}\left(k\mathbf{G}\right)$ for any scalar $k \neq 0$.

The maximum ISI score occurs when $\mathbf{G} = \mathbf{1}$, this is complete and total intersymbol-interference. The minimum ISI score occurs when $\mathbf{W} = \mathbf{DPA}^{-1} = \tilde{\mathbf{P}}\mathbf{A}^{-1}$, where $\mathbf{P}$ is any permutation matrix and $\mathbf{D}$ is any diagonal matrix with nonzero entries along the diagonal. The generalized permutation matrix, $\tilde{\mathbf{P}}$, is such that each row and column of $\tilde{\mathbf{P}}$ contains exactly one nonzero element [111].[6] Then $\mathbf{G} = \tilde{\mathbf{P}}$ and $\mathrm{ISI}\left(\mathbf{G}\right) = \mathrm{ISI}\left(\tilde{\mathbf{P}}\right) = 0$.

Although ISI is a common performance metric for ICA, it does have some disadvantages, most notably a sensitivity to the source scaling. ISI is invariant to a permutation matrix, i.e., $\mathrm{ISI}\left(\mathbf{G}\right) = \mathrm{ISI}\left(\mathbf{PG}\right)$, but it is not invariant to a generalized permutation matrix or $\mathrm{ISI}\left(\mathbf{G}\right) \neq \mathrm{ISI}\left(\tilde{\mathbf{P}}\mathbf{G}\right)$. Thus, to utilize the ISI metric it is essential to fix the scaling ambiguity.

The most common solution to fixing the scaling ambiguity for simulation performance analysis is to scale the mixing and demixing matrices so that both the modeled sources and the estimated sources have unit variance. This is a crucial step for any performance analysis using ISI. *If arbitrary sources are used, then they should first be normalized, so that ISI performance is accurately captured for each arbitrary mixing matrix.* This however is not a limitation of ICA when applied to real world data with sources that do not necessarily have unit variance.

Thus far, we have discussed the ISI metric in the context of ICA. Next, we discuss how this metric has been extended for application in the IVA context.

IVA can be thought of as performing ICA K separate times. Although this perspective is incorrect[7] it motivates the first performance measure, an average of

---

[6] Note that an ambiguity exists in the literature over the properties of a generalized permutation matrix, e.g., see [48]

[7] This perspective is obviously incorrect because ICA cannot solve the case when the sources are Gaussian and iid but IVA can.

K ICA performance measures:

$$\mathrm{ISI}_{\mathrm{AVG}}\left(\mathbf{G}^{[1]}, \ldots, \mathbf{G}^{[K]}\right) \triangleq \frac{1}{K}\sum_{k=1}^{K} \mathrm{ISI}\left(\mathbf{G}^{[k]}\right). \tag{2.61}$$

Since the ISI metric is insensitive to permutations of the rows (or columns) of the global demixing-mixing matrices, $\mathbf{G}$, then the average ISI metric exacts no penalty for *local* permutations across the datasets. Thus average ISI does not assess the ability of the IVA algorithm to correctly align sources across datasets. Rather it is a measure of the average separation performance across all datasets.

To include an assessment of source alignment errors in the performance measure then the absolute value of the entries in the global demixing-mixing matrices are averaged before computing the ISI or:

$$\mathrm{ISI}_{\mathrm{JNT}}\left(\mathbf{G}^{[1]}, \ldots, \mathbf{G}^{[K]}\right) \triangleq \mathrm{ISI}\left(\frac{1}{K}\sum_{k=1}^{K} \left|\mathbf{G}^{[k]}\right|\right) = \mathrm{ISI}\left(\sum_{k=1}^{K} \left|\mathbf{G}^{[k]}\right|\right), \tag{2.62}$$

where $\left|\mathbf{G}^{[k]}\right|$ denotes the matrix where each element is the absolute value of each corresponding entry in $\mathbf{G}^{[k]}$.

This measure of performance, termed here as joint ISI, is suggested in both [55, 79]. Unfortunately, this joint ISI metric is limited to conditions when a common permutation matrix can be applied to all datasets. As we noted previously in the ISR discussion, this common permutation matrix does not always exist when sources do not possess dependence across datasets—see Theorem 2 for exact conditions. Hence, the use of the joint ISI metric must only be applied in simulations for which a common permutation matrix should be found, which is the case in the simulations shown in this work.

Returning to the perspective of IVA as ICA done K times we note that after ICA a clustering (or more specifically a permutation resolving) algorithm can be used to align the components in the K datasets. The resulting joint ISI will be optimized. Then a demonstration of performance benefit provided by IVA over individual ICA can be given (this is a useful benchmark). The joint ISI metric penalizes SCV estimates that are not consistently aligned across datasets and it is normalized so that $0 \leq \mathrm{ISI}_{\mathrm{JNT}} \leq 1$, with 0 implying ideal separation performance. We note that the sensitivity of the ISI metric to the scaling of $\mathbf{G}$ is addressed by scaling the original sources and their estimates to have unit variance.

Next, we argue that essentially, but not exactly, the average ISI is a lower bound on the joint ISI, i.e., $\mathrm{ISI}_{\mathrm{AVG}} \leq \mathrm{ISI}_{\mathrm{JNT}}$. This is strictly true when the absolute value of elements in the global demixing-mixing matrices are one. By this assumption, we have

$$\begin{aligned}
\mathrm{ISI}_{\mathrm{AVG}}\left(\mathbf{G}^{[1]},\ldots,\mathbf{G}^{[K]}\right) &= \frac{1}{K}\sum_{k}\mathrm{ISI}\left(\mathbf{G}^{[k]}\right) \\
&= \frac{\sum_k\left\{\sum_n\left(\sum_m |g_{n,m}^{[k]}| - 1\right) + \sum_m\left(\sum_n |g_{n,m}^{[k]}| - 1\right)\right\}}{2N\left(N-1\right)K} \\
&= \frac{\sum_{n,m}\frac{1}{K}\sum_k |g_{n,m}^{[k]}| + \sum_{m,n}\frac{1}{K}\sum_k |g_{n,m}^{[k]}|}{2N\left(N-1\right)} - \frac{1}{N-1},
\end{aligned}$$

$$
\begin{aligned}
\mathrm{ISI}_{\mathrm{JNT}}\left(\mathbf{G}^{[1]},\ldots,\mathbf{G}^{[K]}\right) &= \mathrm{ISI}\left(\frac{1}{K}\sum_{k}|\mathbf{G}^{[k]}|\right) \\
&= \frac{\sum_n\left(\sum_m \frac{\sum_k |g^{[k]}_{n,m}|}{\max_p \sum_k |g^{[k]}_{n,p}|} - 1\right) + \sum_m\left(\sum_n \frac{\sum_k |g^{[k]}_{n,m}|}{\max_p \sum_k |g^{[k]}_{p,m}|} - 1\right)}{2N\left(N-1\right)} \\
&= \frac{\sum_{m,n} \frac{\sum_k |g^{[k]}_{n,m}|}{\max_p \sum_k |g^{[k]}_{n,p}|} + \sum_{m,n} \frac{\sum_k |g^{[k]}_{n,m}|}{\max_p \sum_k |g^{[k]}_{p,m}|}}{2N\left(N-1\right)} - \frac{1}{N-1},
\end{aligned}
$$

and

$$
\begin{aligned}
\Delta\mathrm{ISI} &= \mathrm{ISI}_{\mathrm{AVG}}\left(\mathbf{G}^{[1]},\ldots,\mathbf{G}^{[K]}\right) - \mathrm{ISI}_{\mathrm{JNT}}\left(\mathbf{G}^{[1]},\ldots,\mathbf{G}^{[K]}\right) \\
&= \frac{1}{K}\sum_{k}\mathrm{ISI}\left(\mathbf{G}^{[k]}\right) - \mathrm{ISI}\left(\frac{1}{K}\sum_{k}|\mathbf{G}^{[k]}|\right) \\
&= \frac{\sum_{m,n}\left(\frac{1}{K}\sum_k |g^{[k]}_{n,m}| - \frac{\sum_k |g^{[k]}_{n,m}|}{\max_p \sum_k |g^{[k]}_{n,p}|}\right) + \sum_{m,n}\left(\frac{1}{K}\sum_k |g^{[k]}_{n,m}| - \frac{\sum_k |g^{[k]}_{n,m}|}{\max_p \sum_k |g^{[k]}_{p,m}|}\right)}{2N\left(N-1\right)} \\
&= \frac{\sum_{k,m,n} |g^{[k]}_{n,m}|\left(\frac{1}{K} - \frac{1}{\max_p \sum_k |g^{[k]}_{n,p}|}\right) + \sum_{k,m,n} |g^{[k]}_{n,m}|\left(\frac{1}{K} - \frac{1}{\max_p \sum_k |g^{[k]}_{p,m}|}\right)}{2N\left(N-1\right)}.
\end{aligned}
$$

The denominators involving the maximum are obviously positive quantities with values less than or equal to K resulting $\mathrm{ISI}_{\mathrm{AVG}} \leq \mathrm{ISI}_{\mathrm{JNT}}$. In general, we do not require the rows of the demixing matrices to have absolute value less than one so this proof does not hold for arbitrary $\mathbf{G}$, yet it becomes “more” true as the average ISI becomes smaller, because then the assumption is nearly true.

2.7.3 Separation Success or Failure

A less frequently used BSS metric is to declare a solution a success or failure based on properties of $\mathbf{G}$. For ICA, an estimate of the demixing matrix is deemed successful if the location of the maximum absolute entry in each row of $\mathbf{G} = \mathbf{WA}$ is unique, [74].

If we assume that the sources could (or should) be identified to within a common permutation matrix across all datasets, then an estimate is deemed successful if the location of the maximum absolute entry in each row of $\mathbf{G}^{[k]} = \mathbf{W}^{[k]}\mathbf{A}^{[k]}$ is unique within each dataset and colocated across the datasets—the former indicates sources are separated within each dataset and the latter indicates if the permutation ambiguity is resolved.

Chapter 3

IVA Algorithms

> "Our greatest glory is not in never falling, but in getting up every time we do." – Confucius

After establishing the IVA objective function(s), the identification conditions, performance metrics, and performance bounds in the previous chapter, we are ready to discuss algorithms. We begin by reviewing existing algorithms and placing the algorithms into one of three categories. The categories differentiate based on the type of dataset dependencies that are assumed or exploited by the algorithms. We then focus on optimization of the IVA cost function when the samples are iid. We proceed by motivating and deriving a decoupling procedure which allows optimization of vectors rather than matrices and permits gradient descent, Newton, and quasi-Newton optimization algorithms. The algorithms are then applied by assuming that the source model follows a multivariate Gaussian distribution. The resulting algorithms are computationally efficient methods for achieving IVA of linearly dependent sources but do not address sources with nonlinear dependency. To do this, we use the Kotz distribution as a source model, which greatly increases the flexibility of the sources that can be accurately separated by IVA algorithms. In fact, the Kotz source distribution model includes the original Laplacian IVA and the Gaussian IVA as special cases. We conclude the chapter by giving simulations. These simulations demonstrate the benefits of the nonorthogonal decoupling procedure in terms of the number of iterations required to converge. The source separation performance achieved by the algorithms are also compared to the theoretical bounds.

3.1 Review of Existing Algorithms

As mentioned previously the origins of algorithms that can be used for IVA date back to pre-BSS times. In fact, the classical CCA [49] algorithm achieves IVA for two datasets with linearly dependent sources. The formulation of CCA can be shown to serve as a basis for all IVA algorithms reviewed here. This is because CCA can be derived from two different, but related principles; maximum likelihood and eigen analysis (diagonalization). Here we choose to separate the approaches into three classes for our review based on the source diversity exploited to achieve BSS. It will be shown that each type of diversity can be utilized—independent of the other two—to achieve IVA.

Our primary interests are in batch algorithms rather than online algorithms. We present some preliminary work for *adaptive* IVA in Appendix D and refer the interested reader to [56, 116] for alternatives.

3.1.1 Linear Dependence

The first class is applicable to problems in which the sources are assumed to have linear dependence across datasets and linearly independent within datasets. The most straightforward and earliest attempts to extend CCA beyond two datasets are lumped to together and termed MCCA. The use of linear dependence in multidataset analysis summarized in [53] provides a number of cost functions based on second-order statistics that result in IVA solutions that can be widely applied. However, the ad hoc nature of the cost functions, the error accumulation of a deflationary procedure for cost optimization, and the orthogonality constraint on the demixing matrix limit the performance of the MCCA implementations given in [53].

Since CCA is derived as generalized eigenvalue decomposition, it can also be posed as a 'generalized joint diagonalization' problem [75]. For IVA of linearly dependent sources the covariance and cross-covariance matrices among the estimated sources in each dataset can be diagonalized. A gradient search solution is used in [76] and a more efficient solution which iteratively solves orthogonal Procustes problems is provided in [75]. The generalized joint diagonalization approach is convenient in that no explicit multivariate source distribution needs to be specified or estimated, the sources are estimated in a symmetric manner, and it is computationally efficient. However, the estimated demixing matrices are orthogonal, which limits the solution space examined. Despite this limitation or constraint on the estimated demixing matrices, the orthogonal solutions in general have better numerical properties in terms of convergence and stability [67].

3.1.2 Nonlinear Dependence

When the sources possess a nonlinear dependence across the datasets then higher order statistics should be utilized, either, explicitly or implicitly. The extension of CCA to nonlinear dependence measures for two datasets dates back to at least 1976 [121]. Extensions to multiple datasets were completed in [32]. Extensions and developments of this early work are summarized in [42].

Another extension for nonlinear CCA of two datasets uses nonparametric univariate and bivariate density estimators in order to maximize the mutual information between two canonical correlation variates, [120]. Kernels have also been used to transform the random vectors into a 'feature-space' where linear CCA is then applied [5, 85]. A different sort of transformation is proposed in [110]. Here measure transform functions are specified for transforming joint probability measures to identify nonlinearly dependent sources. To use either the kernel or measure transform approaches, one must determine the appropriate transform for the problem at hand.

The first introduction of IVA [58, 59] and the similar work of [46] extend the ideas of ICA to multiple datasets so as to solve the permutation ambiguity problem associated with frequency domain ICA [105].

As was the case for linear dependence, diagonalization methods for IVA of nonlinearly dependent sources can be utilized. Specifically, demixing matrices that diagonalize the higher-order statistics (i.e., cumulants of order higher than two) associated with the estimated sources are found [76, 75, 97].

3.1.3 Sample-to-Sample Dependence

Naturally for IVA, as for ICA, algorithms can be developed to exploit sample-to-sample dependence. A *generalization* of joint diagonalization provides such a solution by sampling the vector autocorrelation function at different time lags and finding demixing matrices which minimize correlation between the sources for all time lags considered [76, 75].

3.2 Optimization of IVA Objective Function

A variety of methods are available for optimization of BSS objective functions. Most methods proceed by using an initial (random) guess for the demixing matrix to compute some local derivative information which is then used to revise the previous demixing matrix. Within the set of ICA algorithms that use the likelihood or mutual information as the objective function, two optimization procedures are most prevalent; namely, fixed point algorithm as used in fast independent component analysis (FastICA) [50] and natural (or relative) gradient as used in Infomax [16, 22]. Both procedures have also been used in IVA, see [57, 64]. To begin this section, we describe the original matrix (natural) gradient update procedure originally introduced for IVA. Some issues with this optimization procedure will motivate an alternative optimization procedure which we will utilize to design three algorithms. The new algorithms utilize a 'decoupling trick' first used in the nonorthogonal joint diagonalization procedure of [78]. Variants of this decoupling trick have been subsequently used in [73, 74] for designing ICA algorithms. The decoupling trick provides several advantages over the matrix (natural) gradient and fixed point optimization algorithms, which we highlight in the decoupling derivation to follow. After describing the decoupling it is applied using three approaches we have developed to minimize the iid IVA cost function, (2.7), in a general manner. By *general*, we mean no specific multivariate source distribution model is assumed (other than that a distribution function exists).

3.2.1 Matrix (Natural) Gradient Descent

In [65, 57] the derivative of the cost function wrt each demixing matrix is derived,

$$\begin{aligned}\frac{\partial \mathcal{C}_{\mathrm{IVA}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial \mathbf{W}^{[k]}} &= \sum_{n=1}^{N} \frac{\partial \mathcal{H}\left\{\mathbf{y}_n\right\}}{\partial \mathbf{W}^{[k]}} - \frac{\partial \log \left|\det\left(\mathbf{W}^{[k]}\right)\right|}{\partial \mathbf{W}^{[k]}} \\ &= -\sum_{n=1}^{N} E\left\{\frac{\partial \log p_n\left(\mathbf{y}_n\right)}{\partial \mathbf{W}^{[k]}}\right\} - \left(\mathbf{W}^{[k]}\right)^{-\mathsf{T}} \\ &= -\sum_{n=1}^{N} E\left\{\frac{\partial \log p_n\left(\mathbf{y}_n\right)}{\partial y_n^{[k]}} \frac{\partial y_n^{[k]}}{\partial \mathbf{W}^{[k]}}\right\} - \left(\mathbf{W}^{[k]}\right)^{-\mathsf{T}}, \end{aligned} \tag{3.1}$$

for use in a matrix gradient-based optimization algorithm. The derivative of the first term in the expected value operator depends on the assumed multivariate distribution of the SCVs. The other derivative in (3.1) is readily computed as

$$\left[\frac{\partial y_n^{[k]}}{\partial \mathbf{W}^{[k]}}\right]_{j,i} = \frac{\partial \left(\mathbf{w}_n^{[k]}\right)^{\mathsf{T}} \mathbf{x}^{[k]}}{\partial w_{j,i}^{[k]}} = x_i^{[k]} \delta_{j,n}.$$

Then the nth summand of (3.1) is a matrix of all zeros except for the nth row, which is $\phi_n^{[k]} \left(\mathbf{x}^{[k]}\right)^{\mathsf{T}}$, where $\phi_n^{[k]} \triangleq \phi_n^{[k]}(\mathbf{y}_n) = -\partial \log p_n(\mathbf{y}_n)/\partial y_n^{[k]}$. Throughout, we use the shorthand notation of $\phi_n^{[k]}$ and always recall its dependency on the multivariate source model and the estimated SCV $\mathbf{y}_n$. Finally, the derivative of the IVA cost function wrt $\mathbf{W}^{[k]}$ is,

$$\frac{\partial \mathcal{C}_{\text{IVA}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial \mathbf{W}^{[k]}} = E\left\{\boldsymbol{\phi}^{[k]}\left(\mathbf{x}^{[k]}\right)^{\mathsf{T}}\right\} - \left(\mathbf{W}^{[k]}\right)^{-\mathsf{T}}, \tag{3.2}$$

where,

$$\boldsymbol{\phi}^{[k]} = \left[\phi_1^{[k]}, \ldots, \phi_N^{[k]}\right]^{\mathsf{T}} = -\left[\frac{\partial \log p_1\left(\mathbf{y}_1\right)}{\partial y_1^{[k]}}, \ldots, \frac{\partial \log p_N\left(\mathbf{y}_N\right)}{\partial y_N^{[k]}}\right]^{\mathsf{T}}, \tag{3.3}$$

is formed by selecting the kth entries from each of the N multivariate score functions, $\boldsymbol{\phi}_n = -\partial \log p_n(\mathbf{y}_n)/\partial \mathbf{y}_n$.

Each of the K matrices are updated sequentially by using steepest descent,

$$\mathbf{W}^{[k]} \leftarrow \mathbf{W}^{[k]} - \mu \frac{\partial \mathcal{C}_{\text{IVA}}}{\partial \mathbf{W}^{[k]}}, \tag{3.4}$$

where μ is the scalar positive step size.

Additionally, it is suggested in [65] to use natural gradient updates for faster convergence. The natural gradient is the gradient of (3.2) post-multiplied by $\left(\mathbf{W}^{[k]}\right)^{\mathsf{T}}\mathbf{W}^{[k]}$ and is used to compute the following natural gradient update of the demixing matrix:

$$\mathbf{W}^{[k]} \leftarrow \mathbf{W}^{[k]} - \mu\left(E\left\{\boldsymbol{\phi}^{[k]}\left(\mathbf{y}^{[k]}\right)^{\mathsf{T}}\right\} - \mathbf{I}_N\right)\mathbf{W}^{[k]}. \tag{3.5}$$

3.2.2 Nonorthogonal Decoupling Trick

When we consider the update rule in (3.4) or its natural gradient variation (3.5), a main issue is the use of a single step size to update the demixing matrix. Each row of a demixing matrix corresponds to an estimated source, with a direction for reducing (2.8) specified by the directions of the rows of the matrix (natural) gradient. Thus, using a single step size to update an entire demixing matrix gives equal weight to each direction, when in fact this might be undesirable depending on the shape of the cost function surface. Here, we describe a procedure that permits tailoring the step-size in more flexible yet still efficient manner.

Optimization of cost functions with matrix parameters can occur in many domains, such as signal processing and data mining. Construction of methods that enable each row or column to be individually optimized, i.e., decoupled, are desirable for a number of reasons:

1. The convergence characteristics such as local stability can be improved.
2. Enables density matching in applications such as ICA.
3. Simplifies performance analysis, e.g., derivation of FIM.
4. Efficient Newton algorithms become tractable after decoupling.

The most straightforward method for decoupling rows is to reduce the optimization space to orthogonal matrices. However, the subset of orthonormal matrices might be too restrictive. We present a decoupling procedure—or trick—that uses standard vector optimization procedures while still admitting nonorthogonal solutions, which we now term the nonorthogonal decoupling trick (NDT).

Here we present a distinct derivation of the NDT using basic linear algebra. The NDT can be applied to cost functions that are optimized over the set of full row-rank matrices and have a regularization term based on $\sqrt{\det\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}\right)}$. For full-rank matrices this is equivalent to $|\det\left(\mathbf{W}\right)|$—note this is the regularization term appearing in the objective functions given in Chapter 2, e.g., (2.2). Obviously, such a regularization term appears in cost functions associated with ICA and IVA, but it also occurs in other domains such as nonnegative matrix factorization.

Let the matrix to be estimated, $\mathbf{W}$, be expressed as $\mathbf{W} = \left[\mathbf{w}_1 \ldots \mathbf{w}_M\right]^{\mathsf{T}} \in \mathbb{R}^{M\times N}$, where $M \leq N$. We wish to decouple the estimation of each row in $\mathbf{W}$, $\mathbf{w}_m^{\mathsf{T}}$, $1 \leq m \leq M$. To do so, we will denote the other $M-1$ rows in $\mathbf{W}$ as $\tilde{\mathbf{W}}_m = \left[\mathbf{w}_1 \ldots \mathbf{w}_{m-1}\, \mathbf{w}_{m+1} \ldots \mathbf{w}_M\right]^{\mathsf{T}} \in \mathbb{R}^{(M-1)\times N}$. By using a permutation matrix, $\mathbf{P}_{m,M}$, we can exchange the mth and Mth rows of $\mathbf{W}$ using $\mathbf{P}_{m,M}\mathbf{W}$. This enables us to use the determinant of partitioned matrices given in [47] to write

$$\begin{aligned}
\det\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}\right) &= \det\left(\mathbf{P}_{m,M}\mathbf{W}\mathbf{W}^{\mathsf{T}}\mathbf{P}_{m,M}^{\mathsf{T}}\right)\\
&= \det\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}\right)\\
&= \det\left(\begin{bmatrix} \tilde{\mathbf{W}}_m \\ \mathbf{w}_m^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{W}}_m^{\mathsf{T}} & \mathbf{w}_m \end{bmatrix}\right)\\
&= \det\left(\begin{bmatrix} \tilde{\mathbf{W}}_m\tilde{\mathbf{W}}_m^{\mathsf{T}} & \tilde{\mathbf{W}}_m\mathbf{w}_m \\ \mathbf{w}_m^{\mathsf{T}}\tilde{\mathbf{W}}_m^{\mathsf{T}} & \mathbf{w}_m^{\mathsf{T}}\mathbf{w}_m \end{bmatrix}\right)\\
&= \varrho_m^2\left(\mathbf{w}_m^{\mathsf{T}}\mathbf{w}_m - \mathbf{w}_m^{\mathsf{T}}\tilde{\mathbf{W}}_m^{\mathsf{T}}\left(\tilde{\mathbf{W}}_m\tilde{\mathbf{W}}_m^{\mathsf{T}}\right)^{-1}\tilde{\mathbf{W}}_m\mathbf{w}_m\right)
\end{aligned}$$

$$\begin{aligned} &= \varrho_m^2 \mathbf{w}_m^\mathsf{T} \left(\mathbf{I} - \tilde{\mathbf{W}}_m^\mathsf{T} \left(\tilde{\mathbf{W}}_m \tilde{\mathbf{W}}_m^\mathsf{T} \right)^{-1} \tilde{\mathbf{W}}_m \right) \mathbf{w}_m \\ &= \varrho_m^2 \mathbf{w}_m^\mathsf{T} \tilde{\mathbf{U}}_m \mathbf{w}_m, \end{aligned} \tag{3.6}$$

where $\varrho_m \triangleq \sqrt{\det\left(\tilde{\mathbf{W}}_m \tilde{\mathbf{W}}_m^\mathsf{T}\right)}$ and $\tilde{\mathbf{U}}_m \triangleq \mathbf{I} - \tilde{\mathbf{W}}_m^\mathsf{T} \left(\tilde{\mathbf{W}}_m \tilde{\mathbf{W}}_m^\mathsf{T} \right)^{-1} \tilde{\mathbf{W}}_m \in \mathbb{R}^{N\times N}$. Note that $\mathbf{w}_m^\mathsf{T} \tilde{\mathbf{U}}_m \mathbf{w}_m$ is the Schur complement of $\tilde{\mathbf{W}}_m \tilde{\mathbf{W}}_m^\mathsf{T}$ in $\mathbf{W}\mathbf{W}^\mathsf{T}$.

Recall from linear algebra that the least squares solution to $\mathbf{A}\mathbf{x} = \mathbf{b} \in \mathbb{R}^q$, where $\mathbf{A} \in \mathbb{R}^{q\times r}$ has full column rank (implying that $r \leq q$), can be solved using the normal system of equations, $\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} = \mathbf{A}^\mathsf{T}\mathbf{b}$, or $\mathbf{x} = \left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1}\mathbf{A}^\mathsf{T}\mathbf{b}$. Thus the projection vector, $\mathbf{p} = \mathbf{A}\mathbf{x} = \mathbf{A}\left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1}\mathbf{A}^\mathsf{T}\mathbf{b}$, makes it clear that $\mathbf{A}\left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1}\mathbf{A}^\mathsf{T}$ is a projection matrix that maps vectors onto the column space of $\mathbf{A}$. Orthonormal projection matrices possess many useful properties, one of which is the orthogonal complement, i.e., the projection operator projecting onto the null space of $\mathbf{A}$ is given by $\mathbf{I} - \mathbf{A}\left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1}\mathbf{A}^\mathsf{T}$. With the recollection above, it is clear that $\tilde{\mathbf{U}}_m$ is an orthogonal complement projection matrix for the space spanned by the rows of $\tilde{\mathbf{W}}_m$.

Using the result above one can readily compute derivatives of $\sqrt{\det\left(\mathbf{W}\mathbf{W}^\mathsf{T}\right)}$ wrt each row $\mathbf{w}_m$. It is actually more convenient for our purposes to compute the derivative of $\log\sqrt{\det\left(\mathbf{W}\mathbf{W}^\mathsf{T}\right)}$,

$$\begin{aligned} \frac{\partial \log\sqrt{\det\left(\mathbf{W}\mathbf{W}^\mathsf{T}\right)}}{\partial \mathbf{w}_m} &= \frac{1}{2}\frac{\partial \log \mathbf{w}_m^\mathsf{T}\tilde{\mathbf{U}}_m\mathbf{w}_m}{\partial \mathbf{w}_m} + \frac{\partial \log \varrho_m}{\partial \mathbf{w}_m} \\ &= \frac{\tilde{\mathbf{U}}_m\mathbf{w}_m}{\mathbf{w}_m^\mathsf{T}\tilde{\mathbf{U}}_m\mathbf{w}_m}, \end{aligned} \tag{3.7}$$

and the associated Hessian is

$$\frac{\partial^2 \log \sqrt{\det\left(\mathbf{W}\mathbf{W}^\mathsf{T}\right)}}{\partial \mathbf{w}_m \partial \mathbf{w}_m^\mathsf{T}} = \frac{\tilde{\mathbf{U}}_m}{\mathbf{w}_m^\mathsf{T} \tilde{\mathbf{U}}_m \mathbf{w}_m} - 2\frac{\tilde{\mathbf{U}}_m \mathbf{w}_m \mathbf{w}_m^\mathsf{T} \tilde{\mathbf{U}}_m}{\left(\mathbf{w}_m^\mathsf{T} \tilde{\mathbf{U}}_m \mathbf{w}_m\right)^2} \tag{3.8}$$

The difference between the original derivation and use of [78] with the decoupling presented here is that the former used a 2 × 2 block matrix decomposition of $\mathbf{W}$ rather than the essentially 2 × 1 block matrix decomposition used here. Additionally, with the presented decoupling derivation, the origin of the vector $\mathbf{u}_n$ is clear and consistent with the decoupling used in [73, 74].

We now consider the most common case when $M = N$, i.e., $\mathbf{W}$ is invertible, then by the chosen decomposition of $\mathbf{W}$, we have that $\tilde{\mathbf{U}}_m$ is a rank one matrix. More explicitly, $\tilde{\mathbf{U}}_m = \mathbf{u}_m \mathbf{u}_m^\mathsf{T}$, where $\mathbf{u}_m \in \mathbb{R}^N$ is called the mth decoupling vector for $\mathbf{W}$ and $\|\mathbf{u}_m\| = 1$ due to the requirement that orthonormal projections matrices have eigenvalues of 1 or 0 only. To clarify, $\mathbf{u}_m$ is any vector such that $\tilde{\mathbf{W}}_m \mathbf{u}_m = \mathbf{0} \in \mathbb{R}^{(N-1)}$.

When $M = N$ we have $|\det \mathbf{W}| = \sqrt{\det\left(\mathbf{W}\mathbf{W}^\mathsf{T}\right)}$, thus, in terms of the result in (3.6) we have

$$|\det \mathbf{W}| = \varrho_m \, \left|\mathbf{w}_m^\mathsf{T} \mathbf{u}_m\right|. \tag{3.9}$$

A geometric interpretation of (3.9) is to consider the left hand side as a volumetric term so that A_m is an ‘area’ measure associated with the submatrix $\tilde{\mathbf{W}}_m$ and $\left|\mathbf{w}_m^\mathsf{T} \mathbf{u}_m\right|$ gives a ‘height’ measure.

The derivatives above are simplified for $M = N$ to be:

$$\frac{\partial \log |\det \mathbf{W}|}{\partial \mathbf{w}_m} = \frac{\mathbf{u}_m}{\mathbf{w}_m^{\mathsf{T}} \mathbf{u}_m}, \tag{3.10}$$

and

$$\frac{\partial^2 \log |\det \mathbf{W}|}{\partial \mathbf{w}_m \partial \mathbf{w}_m^{\mathsf{T}}} = \frac{-\mathbf{u}_m \mathbf{u}_m^{\mathsf{T}}}{\left(\mathbf{w}_m^{\mathsf{T}} \mathbf{u}_m\right)^2} = \frac{-\tilde{\mathbf{U}}_m}{\mathbf{w}_m^{\mathsf{T}} \tilde{\mathbf{U}}_m \mathbf{w}_m}. \tag{3.11}$$

A similar decoupling and accompanying gradient and Hessian exists for the case when $\mathbf{W}$ is complex-valued (see Chapter 4).

3.2.3 Vector Gradient Descent

Here, we use the NDT of (3.9) in (2.7) to allow us to use vector derivatives for updating the demixing matrices while preserving the nonorthogonal optimization space. We define $\mathbf{u}_n^{[k]}$ to be the nth decoupling vector for $\mathbf{W}^{[k]}$, i.e., $\widetilde{\mathbf{W}}_n^{[k]} \mathbf{u}_n^{[k]} = \mathbf{0}$, where $\widetilde{\mathbf{W}}_n^{[k]}$ is the $(N-1) \times N$ matrix formed by removing the nth row of the demixing matrix $\mathbf{W}^{[k]}$. Then (2.7) becomes

$$\mathcal{C}_{\text{IVA}}\left(\boldsymbol{\mathcal{W}}\right) = \sum_{m=1}^{N} \mathcal{H}\left\{\mathbf{y}_m\right\} - \sum_{l=1}^{K} \left[\log\left(\left|\left(\mathbf{u}_n^{[l]}\right)^{\mathsf{T}} \mathbf{w}_n^{[l]}\right|\right) + \log \varrho_n^{[l]}\right], \tag{3.12}$$

where we note that $\mathcal{H}\left\{\mathbf{y}_m\right\}$ is independent of $\mathbf{w}_n^{[k]}$ for $m \neq n$, and $\varrho_n^{[l]}$ for $l = 1, \ldots, K$ does not depend $\mathbf{w}_n^{[k]}$. Then, the IVA cost function derivative wrt $\mathbf{w}_n^{[k]}$ is

$$\begin{aligned} \frac{\partial \mathcal{C}_{\mathrm{IVA}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial \mathbf{w}_n^{[k]}} &= -E\left\{\frac{\partial \log p_n\left(\mathbf{y}_n\right)}{\partial y_n^{[k]}} \frac{\partial y_n^{[k]}}{\partial \mathbf{w}_n^{[k]}}\right\} - \frac{\partial \log \left|\left(\mathbf{u}_n^{[k]}\right)^{\mathsf{T}} \mathbf{w}_n^{[k]}\right|}{\partial \mathbf{w}_n^{[k]}} \\ &= E\left\{\phi_n^{[k]} \mathbf{x}^{[k]}\right\} - \frac{\mathbf{u}_n^{[k]}}{\left(\mathbf{u}_n^{[k]}\right)^{\mathsf{T}} \mathbf{w}_n^{[k]}}. \end{aligned} \tag{3.13}$$

The derivative is used to iteratively update each source demixing row and results in nonorthogonal demixing matrices by using a gradient update followed by a normalization step:

$$\left(\mathbf{w}_n^{[k]}\right)^{-} \leftarrow \left(\mathbf{w}_n^{[k]}\right)^{\mathrm{old}} - \mu \frac{\partial \mathcal{C}_{\mathrm{IVA}}}{\partial \mathbf{w}_n^{[k]}} \tag{3.14}$$

$$\left(\mathbf{w}_n^{[k]}\right)^{\mathrm{new}} \leftarrow \frac{\left(\mathbf{w}_n^{[k]}\right)^{-}}{\left\|\left(\mathbf{w}_n^{[k]}\right)^{-}\right\|}. \tag{3.15}$$

The value of the step size, μ, can be fixed to a small number or a line search algorithm can be used to determine the largest μ that results in a decrease in the IVA cost function.

In theory, the normalization step of (3.15) is unnecessary due to the inherent scaling ambiguity that exists in the IVA problem; however in practice it serves several useful purposes. Namely, the normalization step helps maintain acceptable numerical conditioning, conveniently results in unit variance source estimates, and can be exploited to reduce computations. Lastly, we note that the proposed normalization is not unique, although other choices have yet to be considered.

3.2.4 Newton Update

Another, potentially more efficient, optimization approach is to use Newton's method (also known as Newton-Ralphson method). We propose updating each row in all of the demixing matrices concurrently. To do so we denote the demixing vector that estimates the nth SCV as $\mathbf{w}_n^\top = \left[\left(\mathbf{w}_n^{[1]}\right)^\top, \ldots, \left(\mathbf{w}_n^{[K]}\right)^\top\right]$. The gradient of the IVA cost function for the nth SCV demixing vector is

$$\frac{\partial \mathcal{C}_{\text{IVA}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial \mathbf{w}_n} = \left[\left(\frac{\partial \mathcal{C}_{\text{IVA}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial \mathbf{w}_n^{[1]}}\right)^\top, \ldots, \left(\frac{\partial \mathcal{C}_{\text{IVA}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial \mathbf{w}_n^{[K]}}\right)^\top\right]^\top.$$

The corresponding Hessian (matrix), denoted $\mathbf{H}_n \triangleq \partial^2 \mathcal{C}_{\text{IVA}}\left(\boldsymbol{\mathcal{W}}\right) / \partial \mathbf{w}_n \partial \mathbf{w}_n^\top$, is partitioned into K rows and K columns of $N \times N$ matrices with the corresponding matrix in k_1 block row and k_2 block column given by $\mathbf{H}_n^{[k_1,k_2]} = \partial^2 \mathcal{C}_{\text{IVA}}\left(\boldsymbol{\mathcal{W}}\right) / \partial \mathbf{w}_n^{[k_1]} \partial \left(\mathbf{w}_n^{[k_2]}\right)^\top$. The block diagonal entries of the Hessian are

$$\mathbf{H}_n^{[k,k]} = E\left\{\frac{\partial \phi^{[k]}}{\partial \mathbf{w}_n^{[k]}} \left(\mathbf{x}^{[k]}\right)^\top\right\} + \frac{\mathbf{u}_n^{[k]} \left(\mathbf{u}_n^{[k]}\right)^\top}{\left(\left(\mathbf{u}_n^{[k]}\right)^\top \mathbf{w}_n^{[k]}\right)^2} \tag{3.16}$$

and the off-block diagonal entries are

$$\mathbf{H}_n^{[k_1,k_2]} = E\left\{\frac{\partial \phi^{[k_2]}}{\partial \mathbf{w}_n^{[k_1]}} \left(\mathbf{x}^{[k_2]}\right)^\top\right\}, k_1 \neq k_2. \tag{3.17}$$

Depending on the assumed SCV multivariate distribution, $\mathbf{H}_n$ may or may not be strictly positive definite. Later, in Section 3.3.1.3, we show that this Hessian is always strictly positive definite when the source model is a multivariate Gaussian distribution. If the Hessian is ill-conditioned, e.g., singular, then variations of Newton's method can be utilized [19]. Assuming it is positive definite, the Newton algorithm can be directly used and we replace (3.14) with:

$$(\mathbf{w}_n)^{-} \leftarrow (\mathbf{w}_n)^{\mathrm{old}} - \mu \mathbf{H}_n^{-1} \frac{\partial \mathcal{C}_{\mathrm{IVA}}(\boldsymbol{\mathcal{W}})}{\partial \mathbf{w}_n}. \tag{3.18}$$

The step size, μ, for the Newton algorithms can be fixed to any positive quantity less than or equal to one.

3.2.5 Block-Newton Update

Assuming that the Hessian, $\mathbf{H}_n$, is positive definite, its inversion, requires $\mathcal{O}\left((NK)^3\right)$ operations per iteration (per source). Depending on the size of NK, this order of operations might be prohibitive in terms of computation time. We introduce a simple modification to the Newton update that permits the order of operations per iteration (per source) to grow linearly in K and cubically in N. Later in Section 3.4, we show via simulation that this modification provides superior computational performance compared to the previous IVA optimization approaches with only a negligible loss in source separation performance.

The modification we introduce sets all the Hessian off-block diagonal terms of (3.17) to zero. Provided that the Hessian, $\mathbf{H}_n$, is positive definite then all the matrices formed by (3.16), are positive definite, since any principal submatrix of a positive symmetric definite matrix is positive symmetric definite. Then each source in each group can be sequentially updated in each iteration by replacing (3.14) with

$$\left(\mathbf{w}_n^{[k]}\right)^{-} \leftarrow \left(\mathbf{w}_n^{[k]}\right)^{\text{old}} - \mu \left(\mathbf{H}_n^{[k,k]}\right)^{-1} \frac{\partial \mathcal{C}_{\text{IVA}}}{\partial \mathbf{w}_n^{[k]}}. \tag{3.19}$$

Furthermore, in Section 3.3.1.4, we show that after prewhitening the computations per iteration are further decreased because the inverse in (3.19) can be completely avoided via the matrix inversion lemma when using the multivariate Gaussian source model.

3.3 Optimization for Assumed Source Distribution Models

The source distributions are assumed to be known by the algorithms given in the previous section. In practice, the source distribution needs to be assumed or estimated. The estimation of multivariate distributions from data is a challenging problem. Certainly, nonparametric methods exist for this task but this approach is not used here. Rather, we use a parametric approach by assuming a family of distributions with (possibly) some unknown parameters, which will be iteratively updated. In practice, precise knowledge of the source distributions is unknown and so we seek a flexible set of distributions to enable the data to 'speak for itself'. For ICA it is established that precise knowledge of the source distribution is not required in order to achieve source separation. We will demonstrate via simulation that this also holds for IVA.

We start with the Gaussian source model. This model has nice properties that are exploited to achieve efficient optimization of the IVA objective function.

We then give the formulation using the more flexible Kotz source distribution.

3.3.1 Multivariate Gaussian Distribution – IVA-Gauss

Next, we describe the application of the multivariate Gaussian source model to the IVA optimization approaches introduced in the previous section.

3.3.1.1 Matrix Gradient – IVA-G-M

In the IVA-Lap approach of [65, 57], the second-order uncorrelated multivariate Laplace distribution is the assumed form of the SCVs. Therefore, conveniently, there are no unknown parameters for the SCV multivariate distribution model. However, to use the multivariate Gaussian source model in IVA requires an estimate of the SCV covariance matrix, $\hat{\mathbf{\Sigma}}_n$. The *matrix* gradient approach uses the multivariate score function,

$$\boldsymbol{\phi} = -\frac{\partial \log \left\{ p\left(\boldsymbol{y}_n\right) \right\}}{\partial \boldsymbol{y}_n} = \mathbf{\Sigma}_n^{-1} \boldsymbol{y}_n, \tag{3.20}$$

in (3.2) and the maximum likelihood estimate (MLE) of SCV covariance matrix,

$$\hat{\mathbf{\Sigma}}_n = \frac{1}{T} \sum_{t=1}^{\mathsf{T}} \mathbf{y}_n\left(t\right) \mathbf{y}_n^{\mathsf{T}}\left(t\right),$$

from the current demixing solution in (3.21) prior to updating each demixing matrix.

In summary, the procedure for optimizing the IVA cost function using the multivariate Gaussian source model is to use an initial guess for the demixing matrices to generate the estimated sources, then estimate the SCV covariance matrices to compute the multivariate score function, that is then used to update the demixing matrix estimates, and repeat until convergence. We deem that convergence is achieved when

$$\max_k \left(1 - \min \operatorname{diag}\left(\mathbf{W}_{\text{old}}^{[k]} \left(\mathbf{W}_{\text{new}}^{[k]}\right)^{\mathsf{T}}\right)\right) < \epsilon,$$

where ϵ is a small positive number—throughout we use 10^{-6}.

3.3.1.2 Vector Gradient – IVA-G-V

The vector gradient descent algorithm for IVA-Gauss uses the MLE of $\mathbf{\Sigma}$ to compute kth entry of (3.20),

$$\phi_n^{[k]} = \mathbf{e}_k^{\mathsf{T}} \hat{\mathbf{\Sigma}}_n^{-1} \mathbf{y}_n = \mathbf{y}_n^{\mathsf{T}} \hat{\mathbf{\Sigma}}_n^{-1} \mathbf{e}_k, \tag{3.21}$$

in (3.13) to compute the following gradient,

$$\frac{\partial \mathcal{C}_{\text{IVA-G}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial \mathbf{w}_n^{[k]}} = E\left\{\mathbf{x}^{[k]} \mathbf{y}_n^{\mathsf{T}}\right\} \hat{\mathbf{\Sigma}}_n^{-1} \mathbf{e}_k - \frac{\mathbf{u}_n^{[k]}}{\left(\mathbf{u}_n^{[k]}\right)^{\mathsf{T}} \mathbf{w}_n^{[k]}}, \tag{3.22}$$

and using this expression in (3.14).

The operations required in (3.22) grows with the number of samples, T, due to the expectation operation being replaced in practice by an average. However, we can avoid this dependence on the number of samples by observing that

$$E\left\{\mathbf{x}^{[k]}\mathbf{y}_n^{\mathsf{T}}\right\} = \left[\mathbf{R}_x^{[k,1]}\mathbf{w}_n^{[1]},\ldots,\mathbf{R}_x^{[k,K]}\mathbf{w}_n^{[K]}\right], \tag{3.23}$$

where we introduce,

$$\mathbf{R}_x^{[k_1,k_2]} \triangleq E\left\{\mathbf{x}^{[k_1]}\left(\mathbf{x}^{[k_2]}\right)^{\mathsf{T}}\right\}. \tag{3.24}$$

This observation is exploited in practice by using the approximation of (3.24),

$$\hat{\mathbf{R}}_x^{[k_1,k_2]} \triangleq \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}^{[k_1]}(t)\left(\mathbf{x}^{[k_2]}(t)\right)^{\mathsf{T}}, \tag{3.25}$$

in (3.23) to implement (3.22). Obviously, it is true that $\left(\hat{\mathbf{R}}_x^{[k_2,k_1]}\right)^{\mathsf{T}} = \hat{\mathbf{R}}_x^{[k_1,k_2]}$.

Lastly, we note that the (k_1, k_2) entry of the SCV covariance matrix estimate, $\hat{\mathbf{\Sigma}}_n$, is given by

$$\left[\hat{\mathbf{\Sigma}}_n\right]_{k_1,k_2} = \left(\mathbf{w}_n^{[k_1]}\right)^{\mathsf{T}}\hat{\mathbf{R}}_x^{[k_1,k_2]}\mathbf{w}_n^{[k_2]}. \tag{3.26}$$

All the terms required to compute the gradient (and as we will show in the Hessian) can be expressed in terms of the estimated demixing matrices and data mixture covariance matrices. Therefore, the estimated sources do not need to be computed every iteration and the associated computational overhead of doing so is avoided. This property of IVA-Gauss is exploited to develop online learning versions of IVA-Gauss in Appendix D.

3.3.1.3 Newton Update – IVA-G-N

The IVA multivariate Gaussian cost function optimized using the Newton algorithm for each SCV requires evaluation of (3.16) and (3.17), which both use the following derivative:

$$\begin{aligned}
\frac{\partial \phi^{[k_2]}(\mathbf{y}_n)}{\partial \mathbf{w}_n^{[k_1]}} &= \frac{\partial \left[\hat{\boldsymbol{\Sigma}}_n^{-1}\mathbf{y}_n\right]_{k_2}}{\partial \mathbf{w}_n^{[k_1]}} \\
&= \frac{\partial \left(\mathbf{e}_{k_2}^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}_n^{-1}\mathbf{y}_n\right)}{\partial \mathbf{w}_n^{[k_1]}} \\
&= \frac{\partial}{\partial \mathbf{w}_n^{[k_1]}}\left(\sum_{l=1}^{K}\left[\hat{\boldsymbol{\Sigma}}_n^{-1}\right]_{k_2,l}\left(\mathbf{w}_n^{[l]}\right)^{\mathsf{T}}\mathbf{x}^{[l]}\right) \\
&= \sum_{l=1}^{K}\left[\hat{\boldsymbol{\Sigma}}_n^{-1}\right]_{k_2,l}\frac{\partial \left(\mathbf{w}_n^{[l]}\right)^{\mathsf{T}}\mathbf{x}^{[l]}}{\partial \mathbf{w}_n^{[k_1]}} \\
&= \left[\hat{\boldsymbol{\Sigma}}_n^{-1}\right]_{k_2,k_1}\frac{\partial \left(\mathbf{w}_n^{[k_1]}\right)^{\mathsf{T}}\mathbf{x}^{[k_1]}}{\partial \mathbf{w}_n^{[k_1]}} \\
&= \left[\hat{\boldsymbol{\Sigma}}_n^{-1}\right]_{k_1,k_2}\mathbf{x}^{[k_1]}. \qquad (3.27)
\end{aligned}$$

We then find that the expressions in (3.16) and (3.17) can be simplified because:

$$E\left\{\frac{\partial \phi^{[k_2]}(\mathbf{y}_n)}{\partial \mathbf{w}_n^{[k_1]}}\left(\mathbf{x}^{[k_2]}\right)^{\mathsf{T}}\right\} = \left[\hat{\boldsymbol{\Sigma}}_n^{-1}\right]_{k_1,k_2}\hat{\mathbf{R}}_x^{[k_1,k_2]}. \qquad (3.28)$$

The Hessian *for the nth SCV* of the IVA multivariate Gaussian cost function, $\mathbf{H}_n$, can be written as a principal submatrix of a Kronecker product plus a block diagonal matrix or

$$\mathbf{H}_n = \operatorname{psm}\left(\hat{\mathbf{\Sigma}}_n^{-1} \otimes \hat{\mathbf{R}}_x\right) + \oplus \sum_{k=1}^{K} \mathbf{B}^{[k]},$$

where $\operatorname{psm}(\mathbf{A})$ denotes a particular principal submatrix of $\mathbf{A}$,

$$\mathbf{B}^{[k]} \triangleq \frac{\mathbf{u}_n^{[k]} \left(\mathbf{u}_n^{[k]}\right)^{\mathsf{T}}}{\left(\left(\mathbf{u}_n^{[k]}\right)^{\mathsf{T}} \mathbf{w}_n^{[k]}\right)^2},$$

and $\hat{\mathbf{R}}_x$ is the estimate of the complete mixed data covariance matrix, $\mathbf{R}_x \triangleq E\left\{\mathbf{x}\mathbf{x}^{\mathsf{T}}\right\} = \mathbf{A}\mathbf{R}_s\mathbf{A}^{\mathsf{T}} = \mathbf{A}E\left\{\mathbf{s}\mathbf{s}^{\mathsf{T}}\right\}\mathbf{A}^{\mathsf{T}}$.

We now show that the cost function for IVA-G-N will always converge by proving that $\mathbf{H}_{\text{SCV-G}}$ is always positive definite after every iteration, i.e., $\mathbf{H}_{\text{SCV-G}} \succ 0$. First, we note that the Kronecker product of two square matrices has eigenvalues equal to all the possible pairwise products of the eigenvalues associated with each matrix [48]. So $\operatorname{psm}\left(\hat{\mathbf{\Sigma}}_n \otimes \hat{\mathbf{R}}_x\right) \succ 0$ iff $\hat{\mathbf{\Sigma}}_n \succ 0$ and $\hat{\mathbf{R}}_x \succ 0$. The former is estimated after each iteration and the latter is invariant to changes in the demixing matrix.

We can apply Observation 7.1.6 of [47] to show that $\hat{\mathbf{R}}_x = \mathbf{A}\hat{\mathbf{R}}_s\mathbf{A}^{\mathsf{T}} \succ 0$, since by the IVA formulation the mixing matrices, $\mathbf{A}^{[k]}$, are full rank and in practice $\hat{\mathbf{R}}_s \succ 0$ when the number of samples, T, is such that there are N linearly independent samples.

Similarly, we find that $\hat{\mathbf{\Sigma}}_n$ estimated after each iteration is always positive definite. First, we observe from (3.26) that $\hat{\mathbf{\Sigma}}_n = \operatorname{psm}\left(\mathbf{W}\hat{\mathbf{R}}_x\mathbf{W}^{\mathsf{T}}\right)$, where $\mathbf{W} = \oplus \sum_{k=1}^{K} \mathbf{W}^{[k]}$ is full rank. Additionally, any principal submatrix of a positive definite matrix is also positive definite, thus $\hat{\boldsymbol{\Sigma}}_n \succ 0$. Consequently $\text{psm}\left(\hat{\boldsymbol{\Sigma}}_n^{-1} \otimes \hat{\mathbf{R}}_x\right) \succ 0$. Combining this result with the observation that $\oplus \sum_{k=1}^{K} \mathbf{B}^{[k]} \succeq 0$ proves that $\mathbf{H}_{\text{SCV-G}} \succ 0$.

3.3.1.4 Block Newton Update – IVA-G-B

The Newton version presented in the previous subsection requires the inverse of an $NK \times NK$ matrix N times per iteration, which can become computationally burdensome when NK becomes large. The quasi-Newton approach presented in Section 3.2.5 is applied to the multivariate Gaussian and is used to develop an algorithm that avoids any matrix inversion. This modification reduces the computational demands significantly while only slightly degrading the convergence behavior of the algorithm. Additionally, this approach provides greater numerical robustness for large NK by reducing the Hessian dimension to N.

The matrix inverse of (3.19) is avoided by making use of the matrix inversion lemma,

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{C}^{-1} + \mathbf{V}\mathbf{A}^{-1}\mathbf{U}\right)^{-1}\mathbf{V}\mathbf{A}^{-1}.$$

If each dataset is whitened, then $\mathbf{R}_x^{[k,k]} = \mathbf{I}_N$ and the inverse of (3.16), using the multivariate Gaussian source distribution, is

$$\left(\mathbf{H}_n^{[k,k]}\right)^{-1} = \frac{1}{\mathbf{e}_k^{\mathsf{T}} \boldsymbol{\Sigma}_n^{-1} \mathbf{e}_k} \left(\mathbf{I}_N - \frac{\mathbf{u}_n^{[k]} \left(\mathbf{u}_n^{[k]}\right)^{\mathsf{T}}}{\mathbf{e}_k^{\mathsf{T}} \boldsymbol{\Sigma}_n^{-1} \mathbf{e}_k + \left(\left(\mathbf{u}_n^{[k]}\right)^{\mathsf{T}} \mathbf{w}_n^{[k]}\right)^{-2}} \right). \tag{3.29}$$

The prewhitening procedure is not necessary for the previously used optimization strategies but here it is used to avoid any explicit matrix inverse operation in computing the update step given in (3.19).

3.3.2 IVA-Kotz

By appropriately selecting the parameters of the Kotz distribution discussed in Section 2.6.1.2, we arrive at the score function used in

- IVA-Gauss: $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{R}^{-1}\mathbf{x}$ when $\boldsymbol{\theta} = \left[\beta = 1, \eta = 1, \lambda = \frac{1}{2}\right]^{\mathsf{T}}$ and $\boldsymbol{\Sigma} = \mathbf{R}$
- IVA-Lap: $\boldsymbol{\phi}(\mathbf{x}) = \frac{\mathbf{x}}{\sqrt{\mathbf{x}^{\mathsf{T}}\mathbf{x}}}$ when $\boldsymbol{\theta} = \left[\beta = \frac{1}{2}, \eta = 1, \lambda = \frac{1}{2}\right]^{\mathsf{T}}$ and $\boldsymbol{\Sigma} = \mathbf{I}$.

The multivariate Kotz distribution possesses several desirable properties for IVA. It is a distribution which generalizes both IVA-Lap and IVA-Gauss, the score function is readily calculated, and it extends the set of sources for which IVA can be performed. Thus, the Kotz distribution is appealing despite it requiring additional parameters $\boldsymbol{\theta}$ to estimate. The precise estimation of these parameters is difficult at best, yet fortunately for IVA, precise knowledge of these parameters is not critical for successful source separation (as is similarly the case in ICA). One practical approach is to select the $\boldsymbol{\theta}$ from set of $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P]$ that achieves the lowest cost. This approach is naturally accommodated for in the decoupled vector optimization framework. Lastly, the second-order correlation information captured by the dispersion matrix, $\boldsymbol{\Sigma}$, can be estimated suboptimally with the sample average estimate for the covariance matrix, $\hat{\mathbf{R}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}(t)\mathbf{y}^{\mathsf{T}}(t)$, in $\hat{\boldsymbol{\Sigma}} = \rho_{\text{Kotz}}^{-1}\hat{\mathbf{R}}$, where ρ_{Kotz} is given by (2.45).

3.4 Simulations

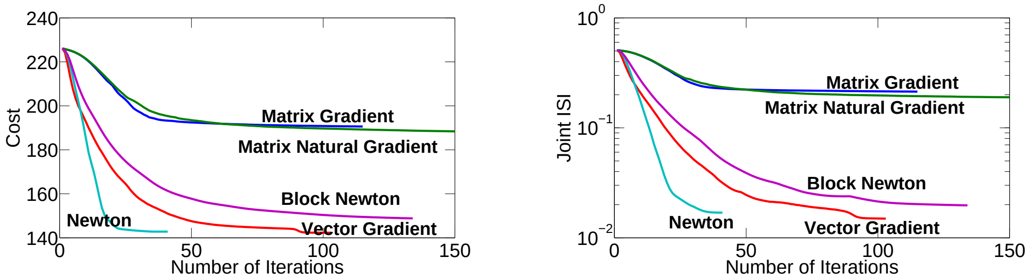
In this section, we consider simulated datasets for assessing the performance of the proposed IVA algorithms and to compare performance versus other algorithms.

3.4.1 Multivariate Gaussian Sources

In this experiment, the nth SCV is a zero-mean, K dimensional, multivariate Gaussian random vector, $\mathbf{s}_n \sim \mathcal{N}(0, \boldsymbol{\Sigma}_n)$, specified by its covariance matrix, $\boldsymbol{\Sigma}_n = \mathbf{U}_n \mathbf{U}_n^{\mathsf{T}}$. The elements of $\mathbf{U}_n$ are from the standard normal distribution, i.e., $[\mathbf{U}_n]_{k_1,k_2} \sim \mathcal{N}(0, 1)$. The kth entry of each SCV is used as a latent source for the kth dataset. Entries of the random mixing matrices are also from the standard normal distribution. It is observed that the results presented are *comparatively* insensitive to sample size, so except where noted otherwise, we only consider one sample size, $T = 10\,000$.

We note that the sample size does need to be large enough to assure that the covariance matrices used in the IVA-Gauss algorithms are positive definite. Theoretically, $T > NK$ is sufficient, however in practice, the minimum number of samples should be greater than this value in certain cases, if for example the observed samples are temporally correlated.

An illustration of the convergence behavior for a single trial is given in Fig. 3.1a of the cost function at each iteration for the multivariate Gaussian based IVA approaches (IVA-Gauss). We define an iteration as an update of all K demixing matrices. We observe that the matrix-based optimization approaches defined by (3.2)



(a) Cost (b) ISI

Figure 3.1: Example of the IVA-Gauss cost (a) and performance measure (b) versus number of iterations for $K = 4$ datasets of $N = 40$ multivariate Gaussian sources, using the matrix gradient, matrix natural gradient extension, vector gradient, Newton, and block-Newton optimization approaches.

and (3.4) using the multivariate Gaussian distribution perform poorly compared to the IVA optimization approaches we introduced based on exploiting the decoupling method of [78] and expressed by (3.9). There is negligible difference between the gradient matrix update and the natural gradient extension, and because of their generally poor performance, we will not consider them further for IVA-Gauss. Amongst the IVA-Gauss decoupled methods, the Newton version of (3.16)–(3.18) converges in the fewest iterations, and the block-Newton approximation of (3.16) and (3.19) converges slower than the vector gradient descent approach of (3.13)–(3.14). The joint ISI performance metric, shown in Fig. 3.1b for the same experiment, indicates that the joint ISI decreases as the cost function decreases.

In Fig. 3.2, purely Gaussian sources are considered and the joint ISI metric is shown versus the number of datasets. The performance for IVA-Gauss is compared with MCCA using the SSQCOR cost function[1] defined in [53] and with the joint diagonalization via second-order statistics (JDIAG-SOS) of [75]. We see that the

---

[1] Our experience suggests that when using MCCA, the SSQCOR cost function generally outperforms the MAXVAR, MINVAR, and GENVAR cost functions for the simulations considered in this work.

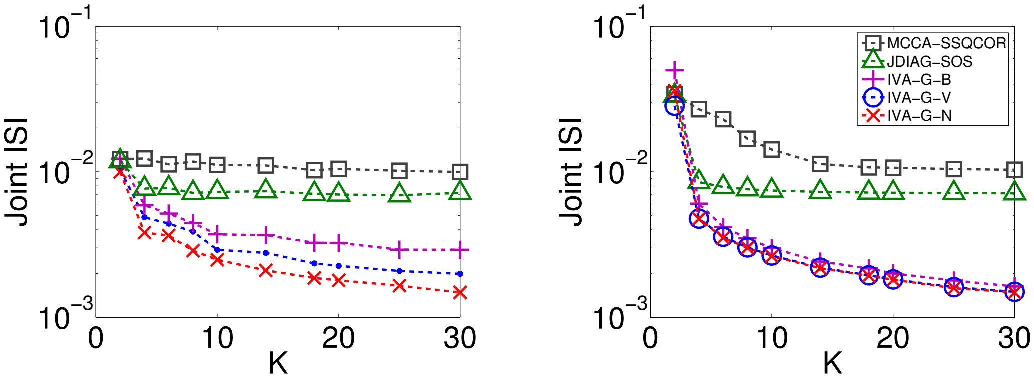


Figure 3.2: Mean joint ISI of 100 trials versus the number of datasets K, using $N = 2$ (a) and $N = 20$ (b) randomly correlated multivariate Gaussian sources.

three IVA-Gauss optimization approaches provide superior performance, with the Newton version (IVA-G-N) being the best, the vector gradient (IVA-G-V) being next best, and the block-Newton (IVA-G-B) being third best. Furthermore, for IVA-Gauss the joint ISI decreases as the number of datasets increases, with the largest rate of decrease occurring for the first few datasets. The difference between the performance of the three IVA-Gauss optimization approaches diminishes for $N = 20$ sources as compared to $N = 2$ sources.

Since the performance of the three decoupled IVA-Gauss approaches are nearly identical when the number of sources increases, then the selection of the preferred algorithm should be based on computation time. Fig. 3.3 shows the average computation time for the algorithms using MATLAB implementations. The results show that IVA-G-V is always the slowest of the three due to the larger number of iterations it requires to converge when compared to the Newton algorithm. When K is small, IVA-G-N is the fastest and IVA-G-B is the fastest of the three for larger K due to the benefit of reducing the number of operations per iteration while main-

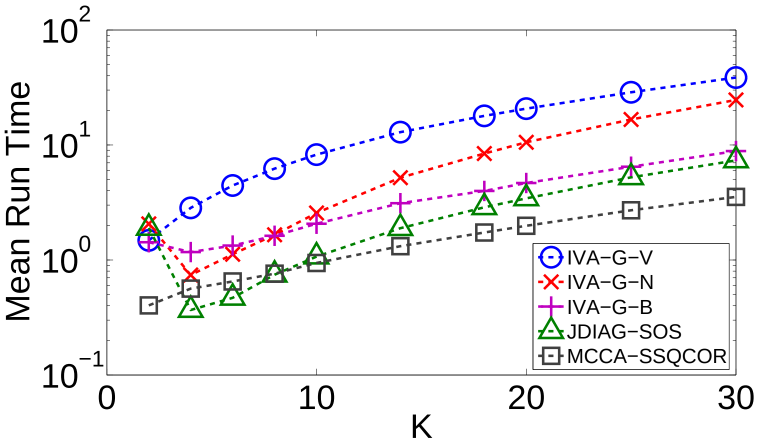


Figure 3.3: Mean run time of 100 trials versus the number of datasets K using $N = 20$ randomly correlated multivariate Gaussian sources.

taining some Newton descent type characteristics. We also note that IVA-G-B is computationally competitive with the orthogonal JDIAG-SOS algorithm for larger K.

Using the multivariate Gaussian sources, we can also compute the iCRLB for ISR using (2.51) and compare this value with the ISR achieved using the IVA-Gauss algorithms. Here we consider a set of $N = 3$ sources each of dimension $K_{\max} = 10$ and each has a random covariance matrix. We then compute the normalized ISR (see (2.59) and accompanying discussion in Section 2.7.1) as the number of datasets is varied from $K = 4$ to $K = K_{\max} = 10$. We compare the theoretical normalized ISR with the average normalized ISR computed from 1000 independent trials of the IVA-G-N algorithm as we vary the number of samples per dataset, T. Since the ISR is invariant wrt the mixing matrices $\mathbf{A}^{[k]}$, the mixing matrices are randomized on each trial. From Fig. 3.4, it is clear that the asymptotic performance of the IVA-G-N approaches the iCRLB when the sample size per dataset is sufficiently large.

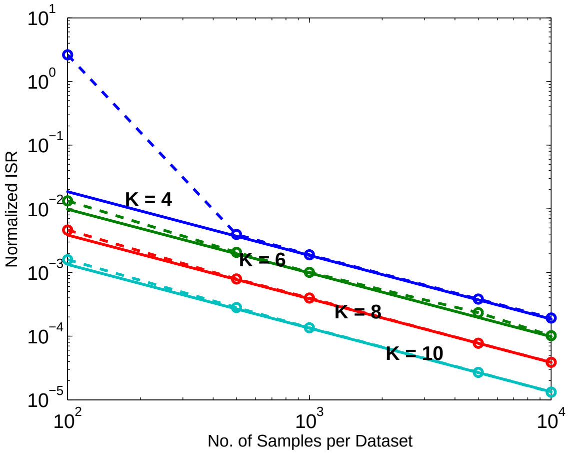
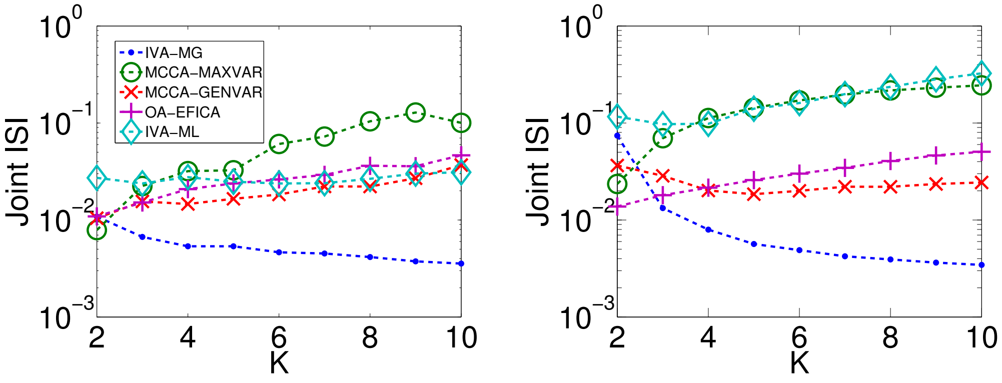


Figure 3.4: Normalized ISR averaged for 1000 trials using IVA-G-N (circle-dashed lines) versus the number of samples per dataset using $N = 3$ sources is compared to iCRLB theory (solid lines). The number of datasets, K, is varied $K = 4, 6, 8, 10$.

3.4.2 Multivariate Non-Gaussian Sources

In many applications, the sources of interest are non-Gaussian. For IVA, the sources must possess second-order (linear) and/or higher-order (nonlinear) dependence across the datasets for unambiguous alignment of sources between datasets.

In certain applications, the multivariate source component vectors may possess marginals that are sub and super Gaussian. We consider sources from generalized Gaussian distribution (GGD) to demonstrate that for non-Gaussian sources, second-order methods are able to perform joint BSS when each source is correlated with a source in each dataset. Specifically, the elements of $\mathbf{z}_n$ are GGD sources with shape parameters selected randomly between $[0.5, 0.75]$ or $[1.25, 1.5]$. Performance in Fig. 3.5 is similar to the performance in Fig. 3.2, indicating reliable performance for IVA-Gauss with second-order correlated non-Gaussian SCV. The same robust behavior holds when the range of the GGD shape parameter is extended to generate either more super or sub-Gaussian sources. For this dataset, the original implemen-



(a) $N = 2$ sources (b) $N = 10$ sources

Figure 3.5: Median joint ISI for 100 trials versus the number of datasets K using linearly combined GGD sources for each SCV.

tation of IVA-Lap [65, 58] using the uncorrelated multivariate Laplace distribution is also evaluated. The performance is poor due to the sources being second-order correlated and the presence of sub and nearly Gaussian sources.

We also examine an alternative approach to joint BSS, using ICA separately K times followed by a source alignment algorithm. We approximate the performance bound of such an alternative using the EFICA algorithm of [60] on each dataset, followed by a permutation algorithm to optimally align estimated sources using *clairvoyant knowledge*. We denote this unrealizable approach, OA-EFICA, for optimally aligned EFICA. We use the EFICA algorithm since it is specifically designed to address GGD sources. Comparing OA-EFICA and IVA-MG in Fig. 3.5 illustrates the benefit of the latter's joint BSS formulation since it is able to separate the nearly Gaussian sources which occur with increasing frequency as the number of datasets increases in this particular data model.

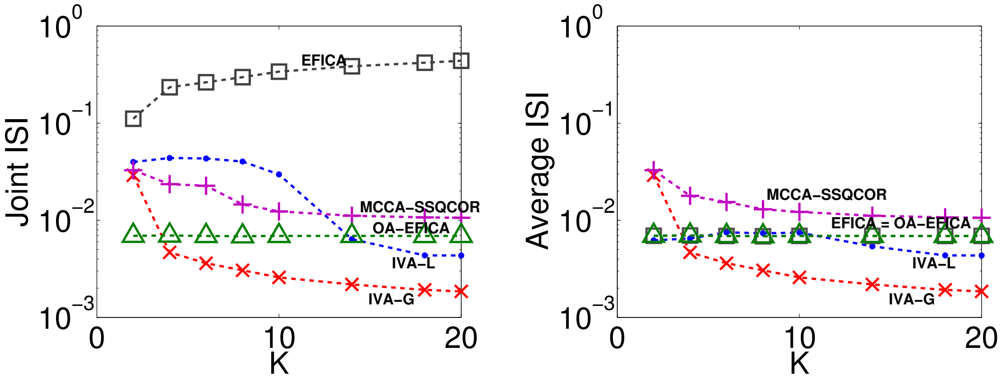
The previous comparison above is somewhat unfair since IVA-Lap is designed to address the super-Gaussian multivariate Laplace source prior and the EFICA algorithm can not separate the nearly Gaussian sources. To permit a more equitable comparison we use the multivariate Laplace source introduced in [35], with a randomly generated covariance *structure* matrix, $\mathbf{\Sigma}_n$. If the sources are second-order uncorrelated then $\mathbf{\Sigma}_n = \mathbf{I}$. The multivariate Laplacian SCVs are generated using

$$\mathbf{s}_n = \sqrt{u}\mathbf{\Sigma}_n^{1/2}\mathbf{z}_n, \tag{3.30}$$

where u is exponentially distributed using a unity rate parameter. The performance of the second-order methods (IVA-Gauss and MCCA) can then be compared with the performance with methods that use the higher-order statistics (IVA-Lap) and the unrealizable BSS solution followed by a permutation resolver (OA-EFICA). In Fig. 3.6a we illustrate that IVA-Gauss can outperform BSS followed by a source alignment algorithm as done in OA-EFICA.

We also show in Fig. 3.6b the *average* ISI performance metric of (2.61)—see Section 2.7.2 for details. Two cases are of interest, when the sources are separated and *aligned* across datasets and two when the sources are separated but *unaligned* across the datasets. For the former, the joint ISI will be about equal to the average ISI and for the latter the average ISI will be small and the joint ISI will be large indicating that the sources are unaligned, i.e., the local permutation problem is not solved. Thus, when we compare the respective curves in the two plots of Fig. 3.6 we observe that IVA-Gauss achieves source alignment, while IVA-Lap does not (for smaller K).

Next, we investigate how the performance of second-order methods compare to



(a) Joint ISI

(b) Average ISI

Figure 3.6: Mean *joint* (a) and *average* (b) ISI for 100 trials for $N = 10$ randomly second-order correlated multivariate Laplacian sources versus the number of datasets, K.

higher-order methods as the amount or degree of second-order correlation increases. We compare the higher-order statistics based method IVA-Lap performance with the IVA-G-N algorithm as the amount of second-order correlation within each SCV is increased using (3.30). We define $\gamma_n \triangleq \det(\mathbf{\Sigma}_n)$ as a measure of the second-order correlation for the nth SCV. The diagonal elements of $\mathbf{\Sigma}_n$ are all unity, and thus $0 \leq \gamma_n \leq 1$, with the upper bound achieved only for second-order *uncorrelated* sources. For this experiment, all of the SCVs have values of γ_n within each trial that differ by less than 1% from each other.

The results of the experiment shown in Fig. 3.7 indicate that only a small amount of second-order correlation is required for IVA-Gauss. The sources being multivariate Laplacian is a model mismatch for IVA-Gauss and the presence of second-order correlation is a model mismatch for IVA-Lap. The performance of IVA-Lap appears invariant wrt the amount of second-order correlation present in this example. The average (mean), however, performance of IVA-Lap is dominated by the approximately 5% to 10% of the runs which fail to achieve JBSS. The cause of the failures is the presence of local minima in the IVA-Lap cost function which exist at permutation ambiguities, as discussed in [31]. Notice that the local minima exist even when $\det(\boldsymbol{\Sigma}) = 1$, which is exactly the IVA-Lap model. We find that the IVA-Lap solution is at a local minimum most frequently when the number of datasets, K, is nominally less than the number of sources, N, but still occurs with a significant frequency when $K > N$ as demonstrated here. The consistency of IVA-Gauss indicates robust source separation and alignment performance despite source distribution model mismatch.

Furthermore, the results suggest a suboptimal approach for reducing the likelihood of permutation problem in IVA-Lap. That is to use IVA-G-N solution to initialize IVA-Lap algorithm. The result of this approach, denoted IVA-GL in Fig. 3.7, suggest that the permutation problem is solved for IVA-Lap when the SCVs possess some minimum degree of second-order correlation.

Now we consider the performance of the decoupled gradient IVA algorithm using the MPE source model (the MPE distribution is a special case of Kotz distribution). The performance of the proposed IVA algorithm is compared with the iCRLB derived in Section 2.6.1.1. In Appendix E, the details on generating Kotz (and MPE) are given along.

For these experiments we consider sources following the MPE distribution, an elliptical distribution with $h_e(u) = \exp\left(-\frac{1}{2}u^{\beta}\right)$ and a pdf normalization constant given by $c_K = \pi^{-K/2}2^{-K/(2\beta)}\beta\Gamma(K/2)$, where $\beta > 0$ is termed the shape parameter. This distribution possesses a score function which includes the score functions used in

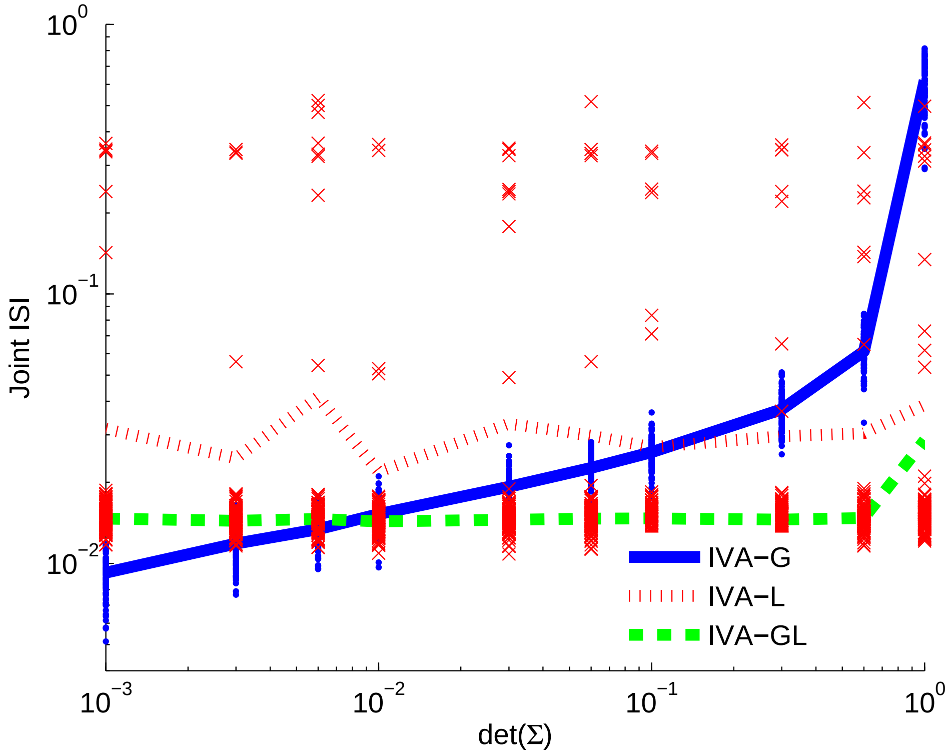


Figure 3.7: Mean joint ISI for 100 trials for $N = 3$ multivariate Laplacian sources in $K = 10$ datasets versus the measure of second order correlation $\det(\mathbf{\Sigma})$ using IVA-Gauss and IVA-Lap. Results from each trial are indicated by dots and x's for IVA-Gauss and IVA-Lap, respectively. The mean joint ISI when IVA-Gauss is used to initialize IVA-Lap, denoted IVA-GL, is shown with dashed line.

both [55] and [11] as special cases. In this section, we consider IVA with multivariate power exponential distribution model (IVA-MPE), where the algorithm is a special case of the IVA with multivariate Kotz distribution model algorithm described in Chapter 3. The datasets are simulated with iid samples from the MPE distribution family. The performance of IVA-MPE is compared with the iCRLB derived in Section 2.6.

For this experiment, there are $N = 3$ MPE SCVs of dimension $K = 5$. All the sources use the same shape parameter, β. The covariance matrix associated with each source is randomly picked for the experiment, yet fixed for all trials in the experiment. The kth entry of each SCV is used as a latent source for the kth dataset. Entries of the random mixing matrices, $\mathbf{A}^{[k]}$, are from the standard normal distribution and are randomly generated for each trial.

We compute the ISR achieved using the MPE algorithm and assuming the correct shape parameter for each source. We then compute the total theoretical ISR using (2.50) in (2.43) and remove the dependency on the number of samples by normalizing the quantities by T. We compare this theoretical ISR with the average ISR computed from 1000 independent trials of the algorithm as we vary the number of samples per dataset.

Due to the presence of local minima in the IVA cost function for non-Gaussian sources [31], the algorithm may converge to local minima. At local minima the sources are separated within a dataset but the SCVs are not successfully identified, i.e., the permutation ambiguity is unresolved. We first compare the iCRLB for the ISR with the mean of the ISR achieved over *successful* trials. A trial is deemed successful if the location of the maximum absolute entry in each row of $\mathbf{G}^{[k]} = \mathbf{W}^{[k]}\mathbf{A}^{[k]}$ is unique within each dataset and colocated across the datasets (the former indicates sources are separated within each dataset and the latter indicates whether the permutation ambiguity is resolved). The fraction of trials which are successful increases as β decreases and/or as the sample size per dataset increases. The lowest success rate was 98%, when $V = 100$ and $\beta = 6$. For all other settings the success rate was greater than 99.5%, see Fig. 3.8. From Fig. 3.9, the performance of the IVA algorithm approaches the iCRLB as the sample size per dataset increases.

We also show in Fig. 3.10—for the same experiment described above—the

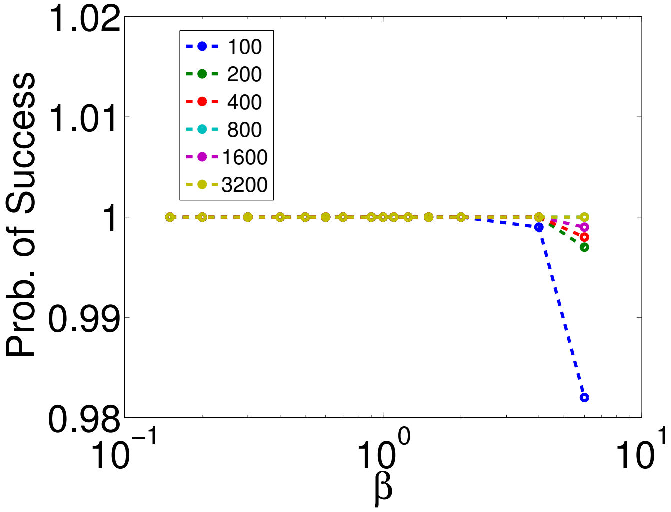


Figure 3.8: The success rate of IVA-MPE algorithm for various numbers of iid samples versus the shape parameter of the simulated SCV.

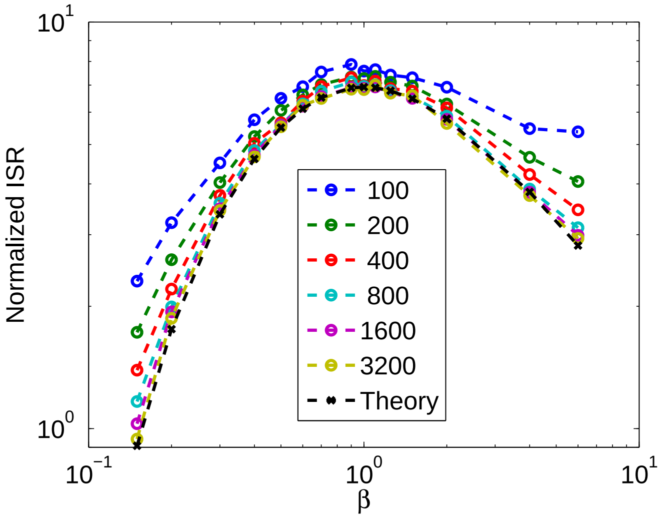


Figure 3.9: The average ISR (of the successful trials) of IVA-MPE algorithm for various numbers of iid samples versus the shape parameter of the simulated SCV in the iid IVA experiment. All results are compared to the iCRLB. Algorithm uses exact knowledge of the shape parameter.

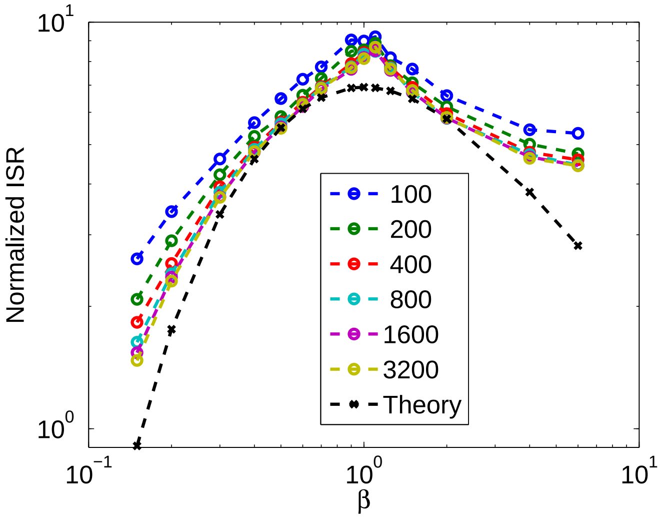


Figure 3.10: The average ISR (of the successful trials) of IVA-MPE algorithm for various numbers of iid samples versus the shape parameter of the simulated SCV in the iid IVA experiment. All results are compared to the iCRLB. Algorithm selects from one of two shape parameters, $\beta \in \{0.5, 2.0\}$.

performance of the IVA-MPE when the algorithm selects between one of two shape parameters ($\beta \in \{0.5, 2.0\}$) according to which shape parameter provides the lowest cost.

In another experiment, we use the same parameters as before except now the SCVs each have identity covariance matrices and the true shape parameters are assumed to be known. For this experiment, there are nonidentifiable conditions when $\beta = 1$, thus we compare the iCRLB for the ISR with the median rather than the mean. From Fig. 3.11, the performance of the IVA algorithm approaches the iCRLB as the sample size per dataset increases. The large degradation in achieved ISR at $\beta = 6$ indicates a sensitivity for nearly *uniform* multivariate sources when the *sample size is small*.

In Fig. 3.9 and Fig. 3.11, the iCRLB follows the behavior predicted by Theorems 3, 6, and 7. Namely, the iCRLB is infinite when sources are Gaussian and

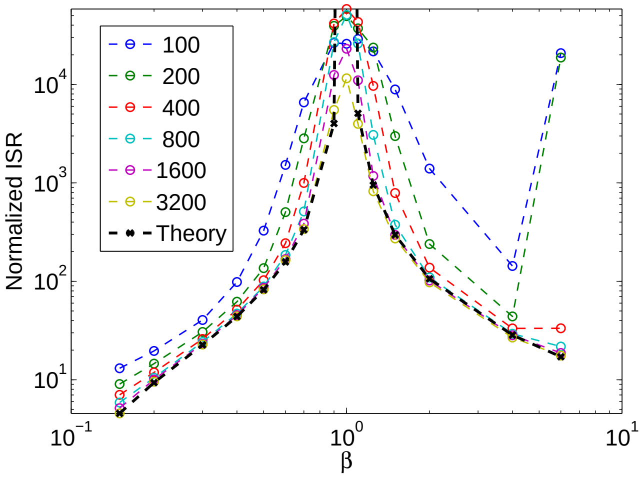


Figure 3.11: The iCRLB theory for ISR as the shape parameter, β, varies is compared to the median ISR of 1000 trials for different numbers of iid samples, T.

$\mathbf{R}_n = \mathbf{I}$ for all sources; the maximum ISR occurs when sources are Gaussian ($\beta = 1$); and as β goes 'away' from one the non-Gaussianity measure κ increases, which yields better source separation, i.e., lower ISR.

3.4.3 Orthogonal Generalized Joint Diagonalization with Second-Order Lags

In this section, we consider the effect of sample dependency. To the best of our knowledge, there is only one algorithm in the IVA framework that accounts for sample-to-sample dependence, namely JDIAG-SOS as given in [75]. The performance of JDIAG-SOS is compared with the iCRLB derived in Section 2.6.

All the sources are a vector moving average of iid Gaussian samples, i.e.,

$$\mathbf{s}_n\left(v\right) = \sum_{l=0}^{L-1} \mathbf{B}_l \mathbf{z}\left(v-l\right), \tag{3.31}$$

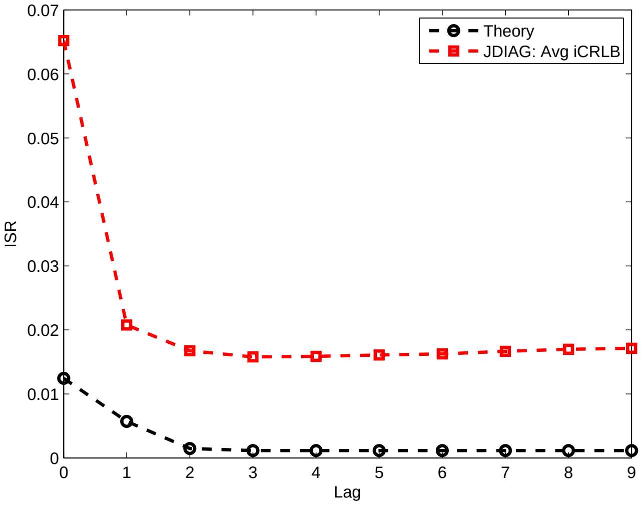


Figure 3.12: The average ISR for 100 trials by JDIAG-SOS. The number of lags used by JDIAG-SOS is varied from 0 to 9 ($L = 1, \ldots, 10$). The iCRLB is shown assuming at most lag $= 3$.

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$ and $[\mathbf{B}_l]_{k_1,k_2} \sim \mathcal{N}(0,1)$. For this experiment there are $N = 3$ sources for $K = 3$ datasets, each with $V = 1000$ samples and $L = 4$. Entries of the random mixing matrices, $\mathbf{A}^{[k]}$, are from the standard normal distribution and are randomly selected each trial.

We compute the theoretical iCRLB for ISR assuming the data was generated with $L = 1, \ldots, 4$. Since $L = 4$ for the data, the performance bound is shown to decrease until the lag is 3. The performance bound for $L = 4$ is shown for lags greater than 3. We compare the performance bounds with the average over 100 independent trials of the ISR achieved using JDIAG-SOS with various lags. Due to JDIAG-SOS estimating orthogonal demixing matrices there exists a noticeable difference between the iCRLB for ISR and the observed ISR.

Chapter 4

Complex-Valued IVA

> "It does not matter how slowly you go as long as you do not stop."
>
> – Confucius

In this chapter, we consider the case when the observations are complex valued. In the previous chapters, the analysis and algorithms assume the observations are real-valued. Complex-valued data is frequently encountered in many applications. It is desirable for the processing of the data to be performed, "…in such a way that the complete information—represented by the interrelationship of the real and imaginary parts…of the signal—can be fully exploited" [2, Chapters 1 and 2]. We achieve such processing by using Wirtinger calculus [117]. A more readily accessible reference for Wirtinger calculus is available in [66]—we also recommend [20] because it provides a succinct summary of the necessary derivatives. Wirtinger calculus enables working completely in the complex domain for the derivation and analysis of IVA. Thus, we avoid unnecessary complications related to transformations mapping the complex to the real domain, i.e., $\mathbb{C}^N \mapsto \mathbb{R}^{2N}$, or having to separate the derivatives wrt the real and imaginary parts. As a result the problem formulation, cost function, and algorithm derivations for complex-valued IVA closely parallel the real-valued case.

There are variations within the literature on defining the vocabulary for specifying the statistical relationships between the real and imaginary components of complex-valued random variable, see [3]. The covariance matrix is given by $E\left\{\mathbf{x}\mathbf{x}^{\dagger}\right\}$ for a zero-mean random vector. The so called pseudo-covariance matrix is given by $E\left\{\mathbf{x}\mathbf{x}^{\mathsf{T}}\right\}$. A random variable/vector is termed proper or second-order circular if it is uncorrelated with its complex conjugate, i.e., if $E\left\{\mathbf{x}\left(\mathbf{x}^{\dagger}\right)^{*}\right\} = E\left\{\mathbf{x}\mathbf{x}^{\mathsf{T}}\right\} = \mathbf{0}$. In [94], a random variable/vector is defined as (strictly) circular when “it is circularly symmetric about the origin $\mathbf{0}$ in which case its distribution remains invariant under multiplication by any (complex) number on the complex unit circle.”

In the first section, we briefly pose the complex-valued IVA problem. Then we give the likelihood cost function which is optimized using the complex version of the NDT. In particular, we give the gradient and Newton update procedures for IVA with complex-valued sources—both approaches are general in that no particular source model is assumed. By assuming that the sources follow complex noncircular multivariate Gaussian distributions, we can specify computationally efficient methods for achieving IVA of linearly dependent complex-valued sources. The family of noncircular multivariate Gaussians allows the IVA algorithm to exploit the second-order statistics available in the pseudo-covariance matrix. We also show here the connection between CCA and IVA, as was also shown for the real-valued analysis. For the complex case, however, this connection with IVA using noncircular Gaussian source models extends the domain of *linear* CCA to include noncircular sources. We conclude the chapter with simulations to demonstrate performance. We illustrate that the solution is robust though suboptimal when the SCVs are non-Gaussian.

4.1 Problem Formulation

The complex-valued formulation of IVA is the same as that given for real-valued IVA given in Section 2.2, except that the sources and the mixing matrices are complex-valued. In addition, we will assume iid samples, since the focus of this chapter is on algorithms.

The complex-valued IVA solution finds K demixing matrices and the corresponding source estimates for each dataset, with the kth estimates denoted as $\mathbf{W}^{[k]} \in \mathbb{C}^{N\times N}$ and $\mathbf{y}^{[k]}(t) = \mathbf{W}^{[k]}\mathbf{x}^{[k]}(t)$, respectively. The estimate of the nth component from the kth dataset is given by $y_n^{[k]}(t) = \left(\mathbf{w}_n^{[k]}\right)^{\dagger}\mathbf{x}^{[k]}(t) = \sum_{m=1}^{N} w_{n,m}^{[k]} x_m^{[k]}(t)$, where $\left(\mathbf{w}_n^{[k]}\right)^{\dagger}$ is the nth row of $\mathbf{W}^{[k]}$, and $w_{n,m}^{[k]}$ is the element in the nth row and mth column of $\mathbf{W}^{[k]}$. The estimate of the nth SCV is given as $\mathbf{y}_n^{\mathsf{T}}(t) = \left[y_n^{[1]}(t), \ldots, y_n^{[K]}(t)\right]$.

4.2 Likelihood

As was done for real-valued IVA, we can maximize the likelihood function by maximizing the following:

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\mathcal{W}}) &\triangleq -\log p_{\mathbf{X}}(\mathbf{X}) && (4.1)\\
&= \sum_{t=1}^{T} \log p_{\mathbf{x}(t)}(\mathbf{x}(t)) && (4.2)\\
&= \sum_{t=1}^{T} \log \left(p_{\mathbf{x}}(\mathbf{x}(t))\right) && (4.3)\\
&= \sum_{t=1}^{T} \log \left(p_{\mathbf{s}}(\mathbf{W}\mathbf{x}(t)) \left|\det \mathbf{W}\right|^2\right) && (4.4)\\
&= \sum_{t=1}^{T} \log p_{\mathbf{s}}(\mathbf{W}\mathbf{x}(t)) - T \log \left|\det \mathbf{W}\right|^2 && (4.5)\\
&= \sum_{t=1}^{T} \log p_{\mathbf{s}}(\mathbf{W}\mathbf{x}(t)) + T \sum_{k=1}^{K} \log \left|\det \mathbf{W}^{[k]}\right|^2 && (4.6)\\
&= \sum_{t=1}^{T} \sum_{n=1}^{N} \log p_n(\mathbf{y}_n(t)) + T \sum_{k=1}^{K} \log\ \left|\det \mathbf{W}^{[k]}\right|^2, && (4.7)
\end{aligned}
$$

where $p_n(\cdot)$ is the model for the distribution characterizing the multivariate source $\mathbf{s}_n$. The expression holds due to the following properties: statistical independence of samples, identical distribution of samples, linear transformation of complex-valued random vectors (see [2, Eq. 1.48]), determinant of block diagonal matrix given by $\mathbf{W} = \oplus \sum_{k=1}^{K} \mathbf{W}^{[k]}$, and independence of SCVs.

Throughout this chapter, we will use the multivariate score function $\boldsymbol{\phi}_n \triangleq \boldsymbol{\phi}_n(\mathbf{y}_n) = -\partial \log(p_n(\mathbf{y}_n)) / \partial \mathbf{y}_n \in \mathbb{C}^K$ and $\boldsymbol{\phi}_n^{[k]} = \boldsymbol{\phi}_n^{\mathsf{T}} \mathbf{e}_k$.

As was the case for real-valued IVA, minimization of mutual information can be related to maximization of the likelihood objective function above when the model distribution of the sources follows the true source distribution and the number of samples approaches infinity. Thus for completeness we have the mutual information based cost function is simply

$$\mathcal{C}_{\text{IVA}}\left(\boldsymbol{\mathcal{W}}\right) = \sum_{n=1}^{N} \mathcal{H}\left\{\mathbf{y}_n\right\} - \sum_{k=1}^{K} \log\left|\det\left(\mathbf{W}^{[k]}\right)\right|^2 . \tag{4.8}$$

By noting that $\mathcal{H}\left\{\mathbf{y}_n\right\} = \sum_k \mathcal{H}\left\{y_n^{[k]}\right\} - \mathcal{I}\left\{\mathbf{y}_n\right\}$, then (4.8) is

$$\mathcal{C}_{\text{IVA}}\left(\boldsymbol{\mathcal{W}}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathcal{H}\left\{y_n^{[k]}\right\} - \mathcal{I}\left\{\mathbf{y}_n\right\} - \sum_{l=1}^{K} \log\left|\det\left(\mathbf{W}^{[l]}\right)\right|^2 , \tag{4.9}$$

which is the complex-valued counter-part to (2.6).

The score functions of each source in dataset k is given by:

$$\boldsymbol{\phi}^{[k]} \triangleq \left[\phi_1^{[k]}, \ldots, \phi_n^{[k]}\right]^{\mathsf{T}} = -\left[\frac{\partial \log\left\{p_1\left(\mathbf{y}_1\right)\right\}}{\partial y_1^{[k]}}, \ldots, \frac{\partial \log\left\{p_N\left(\mathbf{y}_N\right)\right\}}{\partial y_N^{[k]}}\right]^{\mathsf{T}} , \tag{4.10}$$

4.3 Review of Existing Algorithms

Certainly for complex-valued IVA it is appropriate to consider CCA, MCCA (the multidataset CCA extension), and the generalized joint diagonalization approaches, which were all discussed in Chapter 3 as approaches to minimizing (4.8). We also note that [102] discusses complex-valued CCA but only for the *widely* linear model and not for the strictly linear model which is usually the more appropriate model [77].

IVA as originally derived in [58] was designed to address the BSS problem of data transformed from the real-valued temporal domain to the complex-valued frequency domain. The implementation assumes a multivariate Laplace as the form of the SCV distribution that is isotropic and possesses no second-order correlation, i.e., the covariance matrix of the SCV is assumed to be a scaled identity matrix. Both assumptions allow the assumed SCV distribution to be completely characterized, i.e., there are no distribution parameters to estimate. We denote this particular IVA implementation as IVA-Lap. We note that the optimization strategies of the original IVA derivation [57], denoted the matrix-gradient and natural-matrix-gradient, is originally derived for the circular complex-valued case. The IVA cost function is also optimized for circular complex-valued sources using a fixed point update in [64]. Recent work on noncircular IVA has also been conducted by [122]. This noncircular IVA implementation is effectively the multivariate extension of the complex noncircular Fast ICA algorithm of [91]. The fixed point algorithms of [64] and [122] require the data to be whitened and the demixing matrices to be unitary.

In some applications the degree of second-order correlation across datasets may be minimal as in frequency domain BSS [58, 57], however in other applications the degree of second-order correlation is expected to be much larger as in group fMRI studies [79, 100]. We exploit second-order statistical information across datasets by modeling each SCV as to follow a multivariate Gaussian distribution (IVA-Gauss), rather than the isotropic *uncorrelated* multivariate Laplacian distribution as in IVA-Lap. The complete set of second-order statistics for complex-valued sources, i.e., both the covariance and pseudo-covariance information, is considered by including noncircular multivariate Gaussian sources, (IVA-NC-G).

4.4 Nonunitary Decoupling

The same motivation for decoupling the rows of the demixing matrices while maintaining the parameter space of nonorthogonal[1] matrices used in Section 3.2.2 for real-valued IVA applies to complex-valued IVA. Here, we provide the derivation of the complex NDT, which closely follows the real-valued derivation.

Using notation similar to that for the real-valued NDT, we denote the optimization parameter $\mathbf{W} = [\mathbf{w}_1 \mathbf{w}_2 \dots \mathbf{w}_M]^\dagger \in \mathbb{C}^{M \times N}$ and denoting the matrix with with the mth row removed as $\tilde{\mathbf{W}}_m = [\mathbf{w}_1 \, \dots \mathbf{w}_{m-1} \; \mathbf{w}_{m+1} \dots \mathbf{w}_M]^\dagger \in \mathbb{C}^{(M-1) \times N}$, it is straight-forward to proceed as in Section 3.2.2 to observe that

$$\det\left(\mathbf{W}\mathbf{W}^\dagger\right) = \varrho_m^2 \mathbf{w}_m^\dagger \tilde{\mathbf{U}}_m \mathbf{w}_m, \tag{4.11}$$

where $\varrho_m \triangleq \sqrt{\det\left(\tilde{\mathbf{W}}_m \tilde{\mathbf{W}}_m^\dagger\right)}$ and $\tilde{\mathbf{U}}_m \triangleq \mathbf{I} - \tilde{\mathbf{W}}_m^\dagger \left(\tilde{\mathbf{W}}_m \tilde{\mathbf{W}}_m^\dagger\right)^{-1} \tilde{\mathbf{W}}_m \in \mathbb{C}^{N \times N}$ is a Hermitian matrix, i.e., $\tilde{\mathbf{U}}_m = \tilde{\mathbf{U}}_m^\dagger$. Note that $\mathbf{w}_m^\dagger \tilde{\mathbf{U}}_m \mathbf{w}_m$ is the Schur complement of $\tilde{\mathbf{W}}_m \tilde{\mathbf{W}}_m^\dagger$ in $\mathbf{W}\mathbf{W}^\dagger$.

Using the result above one can readily compute derivatives of $\det\left(\mathbf{W}\mathbf{W}^\dagger\right)$ wrt each row $\mathbf{w}_m$ and its conjugate as

$$\frac{\partial \log \det\left(\mathbf{W}\mathbf{W}^\dagger\right)}{\partial \mathbf{w}_m} = \left(\frac{\partial \log \det\left(\mathbf{W}\mathbf{W}^\dagger\right)}{\partial \mathbf{w}_m}\right)^* = \frac{\tilde{\mathbf{U}}_m^* \mathbf{w}_m^*}{\mathbf{w}_m^\dagger \tilde{\mathbf{U}}_m \mathbf{w}_m} \tag{4.12}$$

---

[1] The role of orthogonal matrices in the real domain is the same as unitary matrices in the complex domain, so more precisely NDT stands for the *nonunitary decoupling trick* applied to the optimization of real-valued cost functions with complex-valued optimization parameters.

and

$$\frac{\partial \log\det\left(\mathbf{W}\mathbf{W}^{\dagger}\right)}{\partial \mathbf{w}_m^{*}} = \frac{\partial \log \mathbf{w}_m^{\dagger}\tilde{\mathbf{U}}_m\mathbf{w}_m}{\partial \mathbf{w}_m^{*}} + \frac{\partial \log \varrho_m^2}{\partial \mathbf{w}_m^{*}} = \frac{\tilde{\mathbf{U}}_m\mathbf{w}_m}{\mathbf{w}_m^{\dagger}\tilde{\mathbf{U}}_m\mathbf{w}_m}, \tag{4.13}$$

respectively. The relevant Hessians are

$$\begin{aligned}\frac{\partial^2 \log\det\left(\mathbf{W}\mathbf{W}^{\dagger}\right)}{\partial \mathbf{w}_m \partial \mathbf{w}_m^{\mathsf{T}}} &= \frac{\partial}{\partial \mathbf{w}_m}\left(\frac{\partial \log\det\left(\mathbf{W}\mathbf{W}^{\dagger}\right)}{\partial \mathbf{w}_m}\right)^{\mathsf{T}} \\ &= \frac{\partial}{\partial \mathbf{w}_m}\left(\frac{\mathbf{w}_m^{\dagger}\tilde{\mathbf{U}}_m}{\mathbf{w}_m^{\dagger}\tilde{\mathbf{U}}_m\mathbf{w}_m}\right) \\ &= \frac{-\tilde{\mathbf{U}}_m^{\mathsf{T}}\mathbf{w}_m^{*}\mathbf{w}_m^{\dagger}\tilde{\mathbf{U}}_m}{\mathbf{w}_m^{\dagger}\tilde{\mathbf{U}}_m\mathbf{w}_m} \end{aligned} \tag{4.14}$$

and

$$\begin{aligned}\frac{\partial^2 \log\det\left(\mathbf{W}\mathbf{W}^{\dagger}\right)}{\partial \mathbf{w}_m \partial \mathbf{w}_m^{\dagger}} &= \frac{\partial}{\partial \mathbf{w}_m}\left(\frac{\partial \log\det\left(\mathbf{W}\mathbf{W}^{\dagger}\right)}{\partial \mathbf{w}_m^{*}}\right)^{\mathsf{T}} \\ &= \frac{\partial}{\partial \mathbf{w}_m}\left(\frac{\mathbf{w}_m^{\mathsf{T}}\tilde{\mathbf{U}}_m^{\mathsf{T}}}{\mathbf{w}_m^{\dagger}\tilde{\mathbf{U}}_m\mathbf{w}_m}\right) \\ &= \frac{-\tilde{\mathbf{U}}_m^{\mathsf{T}}\mathbf{w}_m^{*}\mathbf{w}_m^{\mathsf{T}}\tilde{\mathbf{U}}^{\mathsf{T}}}{\left(\mathbf{w}_m^{\dagger}\tilde{\mathbf{U}}_m\mathbf{w}_m\right)^2} + \frac{\tilde{\mathbf{U}}_m^{\mathsf{T}}}{\mathbf{w}_m^{\dagger}\tilde{\mathbf{U}}_m\mathbf{w}_m}. \end{aligned} \tag{4.15}$$

We now consider the most common case when $M = N$, i.e., $\mathbf{W}$ is invertible, then by the chosen decomposition of $\mathbf{W}$, we have that $\tilde{\mathbf{U}}_m$ is a rank one matrix. More explicitly, $\tilde{\mathbf{U}}_m = \mathbf{u}_m\mathbf{u}_m^{\dagger}$, where $\mathbf{u}_m \in \mathbb{C}^N$ is called the mth decoupling vector for $\mathbf{W}$ and $\|\mathbf{u}_m\| = 1$ due to the requirement that unitary projections matrices have eigenvalues of 1 or 0 only. To clarify, $\mathbf{u}_m$ is any vector such that $\tilde{\mathbf{W}}_m\mathbf{u}_m = \mathbf{0} \in \mathbb{C}^{N-1}$.

When $M = N$, we have that

$$\det \mathbf{W}\mathbf{W}^{\dagger} = \left|\det \mathbf{W}\right|^{2} = \varrho_{m}^{2} \mathbf{w}_{m}^{\dagger} \tilde{\mathbf{U}}_{m} \mathbf{w}_{m} = \varrho_{m}^{2} \left|\mathbf{w}_{m}^{\dagger} \mathbf{u}_{m}\right|^{2} \tag{4.16}$$

and the derivatives simplify as follows:

$$\frac{\partial \log \left|\det \mathbf{W}\right|^{2}}{\partial \mathbf{w}_{m}^{*}} = \frac{\mathbf{u}_{m}}{\mathbf{w}_{m}^{\dagger} \mathbf{u}_{m}}, \tag{4.17}$$

$$\frac{\partial \log \left|\det \mathbf{W}\right|^{2}}{\partial \mathbf{w}_{m}} = \frac{\mathbf{u}_{m}^{*}}{\mathbf{u}_{m}^{\dagger} \mathbf{w}_{m}}, \tag{4.18}$$

$$\frac{\partial^{2} \log \left|\det \mathbf{W}\right|^{2}}{\partial \mathbf{w}_{m} \partial \mathbf{w}_{m}^{\mathsf{T}}} = \frac{-\mathbf{u}_{m}^{*} \mathbf{u}_{m}^{\dagger}}{\left|\mathbf{w}_{m}^{\dagger} \mathbf{u}_{m}\right|^{2}}, \tag{4.19}$$

$$\frac{\partial^{2} \log \left|\det \mathbf{W}\right|^{2}}{\partial \mathbf{w}_{m} \partial \mathbf{w}_{m}^{\dagger}} = \mathbf{0}. \tag{4.20}$$

4.4.1 Vector Gradient Descent

Here, we use the NDT of (4.16) in (4.8) to allow us to use vector derivatives for updating the demixing matrices while preserving the nonunitary optimization space. We define $\mathbf{u}_{n}^{[k]}$ to be the nth decoupling vector for $\mathbf{W}^{[k]}$, i.e., $\tilde{\mathbf{W}}_{n}^{[k]} \mathbf{u}_{n}^{[k]} = \mathbf{0}$, where $\tilde{\mathbf{W}}_{n}^{[k]}$ is the $(N-1) \times N$ matrix formed by removing the nth row of the demixing matrix $\mathbf{W}^{[k]}$. Then (4.8) becomes

$$\mathcal{C}_{\text{IVA}}\left(\boldsymbol{\mathcal{W}}\right)=\sum_{m=1}^{N}\mathcal{H}\left\{\mathbf{y}_m\right\}-\sum_{l=1}^{K}\left[\log\left|\left(\mathbf{u}_n^{[l]}\right)^{\dagger}\mathbf{w}_n^{[l]}\right|^2+2\log\varrho_n^{[l]}\right], \tag{4.21}$$

where we note that $\mathcal{H}\left\{\mathbf{y}_m\right\}$ is independent of $\mathbf{w}_n^{[k]}$ for $m\neq n$, and $\varrho_n^{[l]}$ for $l=1,\ldots,K$ does not depend $\mathbf{w}_n^{[k]}$.

Then, the IVA cost function derivative wrt the conjugate of $\mathbf{w}_n^{[k]}$ (denoted by $\mathbf{w}_n^{*[k]}$) is

$$\begin{aligned}\frac{\partial\mathcal{C}_{\text{IVA}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial\mathbf{w}_n^{*[k]}}&=-E\left\{\frac{\partial\log p_n\left(\mathbf{y}_n\right)}{\partial y_n^{[k]}}\frac{\partial y_n^{[k]}}{\partial\mathbf{w}_n^{*[k]}}+\frac{\partial\log p\left(\mathbf{y}_n\right)}{\partial y_n^{[k]*}}\frac{\partial y_n^{[k]*}}{\partial\mathbf{w}_n^{*[k]}}\right\}-\frac{\partial\log\left|\left(\mathbf{u}_n^{[k]}\right)^{\dagger}\mathbf{w}_n^{[k]}\right|^2}{\partial\mathbf{w}_n^{*[k]}}\\&=E\left\{\phi_n^{[k]}\mathbf{x}^{[k]}\right\}-\frac{\mathbf{u}_n^{[k]}}{\left(\mathbf{u}_n^{[k]}\right)^{\dagger}\mathbf{w}_n^{[k]}},\end{aligned} \tag{4.22}$$

where we have used $\left(y_n^{[k]}\right)^*=y_n^{*[k]}$,

$$\frac{\partial y_n^{[k]}}{\partial\mathbf{w}_n^{*[k]}}=\frac{\partial\left(\mathbf{w}_n^{[k]}\right)^{\dagger}\mathbf{x}^{[k]}}{\partial\mathbf{w}_n^{*[k]}}=\mathbf{x}^{[k]}, \tag{4.23}$$

$$\frac{\partial y_n^{*[k]}}{\partial\mathbf{w}_n^{*[k]}}=\frac{\partial\left(\mathbf{x}^{[k]}\right)^{\dagger}\mathbf{w}_n^{[k]}}{\partial\mathbf{w}_n^{*[k]}}=\mathbf{0}, \tag{4.24}$$

and (4.17). As (4.8) is a real-valued function of complex-valued arguments, we can make use of the complex gradient update in [2, p. 35–36]. The derivative is used to iteratively update each source demixing row and results in nonunitary demixing matrices by using a gradient update followed by a normalization step:

$$\left(\mathbf{w}_n^{[k]}\right)^- \leftarrow \left(\mathbf{w}_n^{[k]}\right)^{\text{old}} - 2\mu\frac{\partial \mathcal{C}_{\text{IVA}}}{\partial \mathbf{w}_n^{[k]}} \tag{4.25}$$

$$\left(\mathbf{w}_n^{[k]}\right)^{\text{new}} \leftarrow \frac{\left(\mathbf{w}_n^{[k]}\right)^-}{\left\|\left(\mathbf{w}_n^{[k]}\right)^-\right\|}. \tag{4.26}$$

The value of the step size, μ, can be fixed to a small number or a line search algorithm can be used to determine the largest μ that results in a decrease in the IVA cost function. The factor of two multiplying the gradient in (4.25) is required to maintain an exact equivalence between the gradient update rule for the real and complex domains (see [66] for more details). Of course, the step-size, μ, can absorb this multiplier and is not of any particular consequence in the subsequent material.

4.4.2 Newton Update

A significant advantage provided by the decoupling of the rows in the demixing matrices is that it readily supports a Newton update optimization procedure. For the Newton update procedure, we propose to update all of the decoupling vectors for the nth SCV simultaneously. The Newton update of the real-valued IVA cost function of complex arguments given by (4.8) is [2, p. 36–37]:

$$\left(\mathbf{w}_n\right)^- \leftarrow \left(\mathbf{w}_n\right)^{\text{old}} - \mu\left(\mathbf{H}_{\text{B},n}^* - \mathbf{H}_{\text{A},n}^*\mathbf{H}_{\text{B},n}^{-1}\mathbf{H}_{\text{A},n}\right)^{-1}\left(\frac{\partial \mathcal{C}_{\text{IVA}}}{\partial \mathbf{w}_n^*} - \mathbf{H}_{\text{A},n}^*\mathbf{H}_{\text{B},n}^{-1}\frac{\partial \mathcal{C}_{\text{IVA}}}{\partial \mathbf{w}_n}\right), \tag{4.27}$$

where, $\mathbf{w}_n^{\mathsf{T}} \triangleq \left[\left(\mathbf{w}_n^{[1]}\right)^{\mathsf{T}} \ldots \left(\mathbf{w}_n^{[K]}\right)^{\mathsf{T}}\right]$, $\mathbf{H}_{\mathrm{A},n} \triangleq \dfrac{\partial^2 \mathcal{C}_{\mathrm{IVA}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial \mathbf{w}_n \partial \mathbf{w}_n^T}$, and $\mathbf{H}_{\mathrm{B},n} \triangleq \dfrac{\partial^2 \mathcal{C}_{\mathrm{IVA}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial \mathbf{w}_n \partial \mathbf{w}_n^H}$

.

The gradient of the IVA cost function for the nth SCV demixing vector is

$$\frac{\partial \mathcal{C}_{\mathrm{IVA}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial \mathbf{w}_n^{\dagger}} = \left[\begin{array}{ccc} \left(\frac{\partial \mathcal{C}_{\mathrm{IVA}}(\boldsymbol{\mathcal{W}})}{\partial \mathbf{w}_n^{[1]}}\right)^{\dagger} & \ldots & \left(\frac{\partial \mathcal{C}_{\mathrm{IVA}}(\boldsymbol{\mathcal{W}})}{\partial \mathbf{w}_n^{[K]}}\right)^{\dagger} \end{array}\right] \in \mathbb{C}^{NK}.$$

The matrices $\mathbf{H}_{\mathrm{A},n}$ and $\mathbf{H}_{\mathrm{B},n}$ can each be partitioned into K rows and K columns with the matrix in k_1 block row and k_2 block column given by

$$\begin{aligned} \mathbf{H}_{\mathrm{A},n}^{[k_1,k_2]} &\triangleq \frac{\partial^2 \mathcal{C}_{\mathrm{IVA}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial \mathbf{w}_n^{[k_1]} \partial \left(\mathbf{w}_n^{[k_2]}\right)^{\mathsf{T}}} \\ &= \frac{\partial}{\partial \mathbf{w}_n^{[k_1]}} \left(E\left\{\phi_n^{*[k_2]} \left(\mathbf{x}^{[k_2]}\right)^{\dagger}\right\} - \frac{\left(\mathbf{u}_n^{[k_2]}\right)^{\dagger}}{\left(\mathbf{u}_n^{[k_2]}\right)^{\dagger} \mathbf{w}_n^{[k_2]}} \right) \\ &= E\left\{\frac{\partial \phi_n^{*[k_2]}}{\partial \mathbf{w}_n^{[k_1]}} \left(\mathbf{x}^{[k_2]}\right)^{\dagger}\right\} + \delta_{k_1,k_2} \frac{\left(\mathbf{u}_n^{[k]}\right)^{*} \left(\mathbf{u}_n^{[k]}\right)^{\dagger}}{\left(\left(\mathbf{u}_n^{[k]}\right)^{\dagger} \mathbf{w}_n^{[k]}\right)^2}, \end{aligned} \tag{4.28}$$

and

$$\begin{aligned} \mathbf{H}_{\mathrm{B},n}^{[k_1,k_2]} &\triangleq \frac{\partial^2 \mathcal{C}_{\mathrm{IVA}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial \mathbf{w}_n^{[k_1]} \partial \left(\mathbf{w}_n^{[k_2]}\right)^{\dagger}} \\ &= \frac{\partial}{\partial \mathbf{w}_n^{[k_1]}} \left(E\left\{\phi_n^{[k_2]} \left(\mathbf{x}^{[k_2]}\right)^{\mathsf{T}}\right\} - \frac{\left(\mathbf{u}_n^{[k_2]}\right)^{\mathsf{T}}}{\left(\mathbf{w}_n^{[k_2]}\right)^{\dagger} \mathbf{u}_n^{[k_2]}} \right) \\ &= E\left\{\frac{\partial \phi_n^{[k_2]}}{\partial \mathbf{w}_n^{[k_1]}} \left(\mathbf{x}^{[k_2]}\right)^{\mathsf{T}}\right\}, \end{aligned} \tag{4.29}$$

respectively.

Depending on the assumed SCV multivariate distribution the equivalent Hessian in $\mathbb{C}^{2\bar{N}\times 2\bar{N}}$ denoted,

$$\mathbf{H}_n \triangleq \begin{bmatrix} \mathbf{H}_{\mathrm{B},n}^* & \mathbf{H}_{\mathrm{A},n}^* \\ \mathbf{H}_{\mathrm{A},n} & \mathbf{H}_{\mathrm{B},n} \end{bmatrix},$$

may or may not be strictly positive definite. If the Hessian is ill-conditioned, e.g., singular, then variations of Newton's method can be utilized [19]. The step size, μ, for the Newton algorithms can be fixed to any positive quantity less than or equal to one.

Finally, we note that one quasi-Newton approach is to let $\mathbf{H}_{\mathrm{A},n}$ be $\mathbf{0}$. By doing so, we obtain the following simplified update equation,

$$(\mathbf{w}_n)^- \leftarrow (\mathbf{w}_n)^{\mathrm{old}} - \mu \left(\mathbf{H}_{\mathrm{B},n}^*\right)^{-1} \frac{\partial \mathcal{C}_{\mathrm{IVA}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial \mathbf{w}_n^*}, \tag{4.30}$$

which requires only one matrix inverse rather than the three required to compute the full Newton update given in (4.27).

4.4.3 Noncircular Gaussian Distribution

For complex-valued sources the complete second-order statistics are captured in the covariance and the pseudo-covariance matrices, $\boldsymbol{\Sigma}_n \triangleq E\left\{\mathbf{s}_n\mathbf{s}_n^\dagger\right\}$ and $\tilde{\boldsymbol{\Sigma}}_n \triangleq E\left\{\mathbf{s}_n\mathbf{s}_n^\mathsf{T}\right\}$, respectively. Hence, we use the zero-mean noncircular multivariate Gaussian distribution [98],

$$p_n\left(\mathbf{y}_n\right) = \pi^{-K}\left(\det \underline{\boldsymbol{\Sigma}}_n\right)^{-1/2} \exp\left(-\frac{1}{2}\underline{\mathbf{y}}_n^{\dagger}\underline{\boldsymbol{\Sigma}}_n^{-1}\underline{\mathbf{y}}_n\right), \tag{4.31}$$

as the assumed form of the K dimensional SCV distribution for the nth source with an *augmented* covariance matrix, $\underline{\boldsymbol{\Sigma}}_n \triangleq E\left\{\underline{\mathbf{s}}_n\underline{\mathbf{s}}_n^{\dagger}\right\}$, where $\underline{\mathbf{s}}_n^{\dagger} \triangleq \left[\mathbf{s}_n^{\dagger}, \mathbf{s}_n^{\mathsf{T}}\right]$. We adopt the convention of using $\underline{(\cdot)}$ to denote all augmented terms. The corresponding entropy of the SCV is

$$\mathcal{H}\left\{\mathbf{y}_n\right\} = -E\left\{\log p_n\left(\mathbf{y}_n\right)\right\} = K\log\left(\pi\right) + \frac{1}{2}\log\det\underline{\boldsymbol{\Sigma}}_n + \frac{1}{2}\mathrm{tr}\left(\underline{\boldsymbol{\Sigma}}_n^{-1}E\left\{\underline{\mathbf{y}}_n\underline{\mathbf{y}}_n^{\dagger}\right\}\right).$$

In the iterative optimization of the IVA-Gauss cost function, the MLE for the SCV augmented-covariance matrix is computed using the SCV estimate of the previous iteration, i.e.,

$$\hat{\underline{\boldsymbol{\Sigma}}}_n = \frac{1}{T}\sum_{t=1}^{T}\underline{\mathbf{y}}_n\left(t\right)\underline{\mathbf{y}}_n^{\dagger}\left(t\right).$$

The estimate of the covariance matrix $\hat{\underline{\boldsymbol{\Sigma}}}_n \rightarrow E\left\{\underline{\mathbf{y}}_n\underline{\mathbf{y}}_n^{\dagger}\right\}$ as $T \rightarrow \infty$. Then $\mathcal{H}\left\{\mathbf{y}_n\right\}$ is estimated by $K\log\left(\pi\exp\left(1\right)\right) + \left(1/2\right)\log\det\hat{\underline{\boldsymbol{\Sigma}}}_n$ and the IVA cost function of (4.8) becomes

$$\mathcal{C}_{\text{IVA-NC-G}}\left(\boldsymbol{\mathcal{W}}\right) = NK\log\left(\pi e\right) + \frac{1}{2}\sum_{n=1}^{N}\log\det\left(\hat{\underline{\boldsymbol{\Sigma}}}_n\right) - 2\sum_{k=1}^{K}\log\left|\det\left(\mathbf{W}^{[k]}\right)\right|.$$

Here we note that the extension of CCA for the case of noncircular sources has been developed for the widely linear model [102], but to the best of our knowledge, not for the linear model. The approach we present here provides a numerical approach for achieving the *linear* CCA of two or more second-order noncircular SCVs.

We calculate the multivariate score function, $\boldsymbol{\phi}_n$, required for the update rules as,

$$\begin{aligned}
\boldsymbol{\phi}_n &\triangleq -\frac{\partial \log p\left(\mathbf{y}_n\right)}{\partial \mathbf{y}_n} \\
&= -\frac{\partial \log \left\{\pi^{-K} \det(\underline{\boldsymbol{\Sigma}}_n)^{-1} \exp\left(-\frac{1}{2}\underline{\mathbf{y}}_n^{\dagger}\underline{\boldsymbol{\Sigma}}_n^{-1}\underline{\mathbf{y}}_n\right)\right\}}{\partial \mathbf{y}_n} \\
&= \frac{1}{2}\frac{\partial\left(\underline{\mathbf{y}}_n^{\dagger}\underline{\boldsymbol{\Sigma}}_n^{-1}\underline{\mathbf{y}}_n\right)}{\partial \mathbf{y}_n}.
\end{aligned} \tag{4.32}$$

For ease of notation, we drop the subscript n for the moment, as it is applied to all variables. We use a partitioning of the augmented covariance matrix inverse [98],

$$\underline{\boldsymbol{\Sigma}}^{-1} = \begin{bmatrix} \mathbf{Q} & \tilde{\mathbf{Q}} \\ \tilde{\mathbf{Q}}^{\dagger} & \mathbf{Q}^{*} \end{bmatrix},$$

where $\mathbf{Q}^{-1} \triangleq \left(\boldsymbol{\Sigma}^{*} - \tilde{\boldsymbol{\Sigma}}^{\dagger}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Sigma}}\right)^{*}$ is a Hermitian matrix and $\tilde{\mathbf{Q}} \triangleq -\left(\tilde{\boldsymbol{\Sigma}}^{\dagger}\boldsymbol{\Sigma}^{-1}\right)^{\dagger}\mathbf{Q}^{-1}$ is a complex symmetric matrix, to express the inner-product in (4.32) explicitly in terms of $\mathbf{y}$,

$$\underline{\mathbf{y}}^{\dagger}\underline{\boldsymbol{\Sigma}}^{-1}\underline{\mathbf{y}} = \begin{bmatrix} \mathbf{y}^{\dagger} & \mathbf{y}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \mathbf{Q} & \tilde{\mathbf{Q}} \\ \tilde{\mathbf{Q}}^{\dagger} & \mathbf{Q}^{*} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{y}^{*} \end{bmatrix} = 2\mathbf{y}^{\dagger}\mathbf{Q}\mathbf{y} + \mathbf{y}^{\mathsf{T}}\tilde{\mathbf{Q}}^{\dagger}\mathbf{y} + \mathbf{y}^{\dagger}\tilde{\mathbf{Q}}\mathbf{y}^{*}. \tag{4.33}$$

Again, using the properties of Wirtinger calculus, the gradient of the terms in (4.33) wrt $\mathbf{y}$ can be directly evaluated as

$$\begin{aligned}
\frac{\partial \mathbf{y}^{\dagger}\mathbf{Q}\mathbf{y}}{\partial \mathbf{y}} &= \mathbf{Q}^{\mathsf{T}}\mathbf{y}^{*} = \mathbf{Q}^{*}\mathbf{y}^{*} \\
\frac{\partial \mathbf{y}^{\mathsf{T}}\tilde{\mathbf{Q}}^{\dagger}\mathbf{y}}{\partial \mathbf{y}} &= \left(\left(\tilde{\mathbf{Q}}^{\dagger}\right)^{\mathsf{T}} + \tilde{\mathbf{Q}}^{\dagger}\right)\mathbf{y} = 2\tilde{\mathbf{Q}}^{\dagger}\mathbf{y} \\
\frac{\partial \mathbf{y}^{\dagger}\tilde{\mathbf{Q}}\mathbf{y}^{*}}{\partial \mathbf{y}} &= \mathbf{0}.
\end{aligned}$$

Thus, returning to the nth SCV, the score function for a noncircular Gaussian model is

$$\boldsymbol{\phi}_n = \frac{1}{2}\frac{\partial\left(\underline{\mathbf{y}}_n^{\dagger}\underline{\boldsymbol{\Sigma}}_n^{-1}\underline{\mathbf{y}}_n\right)}{\partial \mathbf{y}_n} = \mathbf{Q}_n^{*}\mathbf{y}_n^{*} + \tilde{\mathbf{Q}}_n^{\dagger}\mathbf{y}_n. \tag{4.34}$$

We can express the kth entry of the score function as, $\phi_n^{[k]} = \mathbf{e}_k^{\mathsf{T}}\boldsymbol{\phi}_n = \boldsymbol{\phi}_n^{\mathsf{T}}\mathbf{e}_k$, then

$$\begin{aligned}
E\left\{\phi_n^{[k]}\mathbf{x}^{[k]}\right\} &= E\left\{\mathbf{x}^{[k]}\boldsymbol{\phi}_n^{\mathsf{T}}\right\}\mathbf{e}_k \\
&= E\left\{\mathbf{x}^{[k]}\mathbf{y}_n^{\dagger}\mathbf{Q}_n^{\dagger} + \mathbf{x}^{[k]}\mathbf{y}_n^{\mathsf{T}}\tilde{\mathbf{Q}}_n^{\dagger}\right\}\mathbf{e}_k \\
&= E\left\{\mathbf{x}^{[k]}\underline{\mathbf{y}}_n^{\dagger}\right\}\left[\begin{array}{cc} \mathbf{Q}_n & \tilde{\mathbf{Q}}_n \end{array}\right]^{\dagger}\mathbf{e}_k.
\end{aligned} \tag{4.35}$$

Finally, for the complex noncircular multivariate Gaussian source model we can use the general IVA gradient-based update procedure in (4.22),

$$\frac{\partial \mathcal{C}_{\text{IVA-NC-G}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial\left(\mathbf{w}_n^{[k]}\right)^{*}} = E\left\{\mathbf{x}^{[k]}\underline{\mathbf{y}}_n^{\dagger}\right\}\left[\begin{array}{cc} \mathbf{Q}_n & \tilde{\mathbf{Q}}_n \end{array}\right]^{\dagger}\mathbf{e}_k - \frac{\mathbf{u}_n^{[k]}}{\left(\mathbf{w}_n^{[k]}\right)^{\dagger}\mathbf{u}_n^{[k]}}.$$

We can use the data cross-covariance matrices,

$$\mathbf{R}_x^{[k_1,k_2]} \triangleq E\left\{\mathbf{x}^{[k_1]}\left(\mathbf{x}^{[k_2]}\right)^{\dagger}\right\}, \tag{4.36}$$

and pseudo cross-covariance matrices,

$$\mathbf{C}_x^{[k_1,k_2]} \triangleq E\left\{\mathbf{x}^{[k_1]}\left(\mathbf{x}^{[k_2]}\right)^{\mathsf{T}}\right\}, \tag{4.37}$$

to express the following expectations in the gradient expression,

$$E\left\{\mathbf{x}^{[k]}\mathbf{y}_n^{\dagger}\right\} = \left[\begin{array}{ccc} \mathbf{R}_x^{[k,1]}\mathbf{w}_n^{[1]}, & \ldots, & \mathbf{R}_x^{[k,K]}\mathbf{w}_n^{[K]} \end{array}\right], \tag{4.38}$$

and

$$E\left\{\mathbf{x}^{[k]}\mathbf{y}_n^{\mathsf{T}}\right\} = \left[\begin{array}{ccc} \mathbf{C}_x^{[k,1]}\left(\mathbf{w}_n^{[1]}\right)^{*}, & \ldots, & \mathbf{C}_x^{[k,K]}\left(\mathbf{w}_n^{[K]}\right)^{*} \end{array}\right]. \tag{4.39}$$

The terms for the gradient update procedure have all been given for the complex-valued IVA using the noncircular multivariate Gaussian source model. Predictably, the result has a form similar to the real-valued result derived in Chapter 3.

Lastly, we need to compute the terms in (4.27) and (4.29) to complete the Newton updates. For that we need the gradient of the k_2th entry of (4.34) and its conjugate wrt $\mathbf{w}_n^{k_1}$. First note that by (4.23) and (4.24),

$$\frac{\partial \mathbf{e}_{k_2}^{\mathsf{T}}\mathbf{B}\mathbf{y}_n^{*}}{\partial \mathbf{w}_n^{[k_1]}} = \frac{\partial}{\partial \mathbf{w}_n^{[k_1]}}\left(\sum_{l=1}^{K}[\mathbf{B}]_{k_2,l}\left(\mathbf{x}^{[l]}\right)^{\dagger}\mathbf{w}_n^{[l]}\right) = [\mathbf{B}]_{k_2,k_1}\left(\mathbf{x}^{[k_1]}\right)^{*},$$

and

$$\frac{\partial \mathbf{e}_{k_2}^{\mathsf{T}} \mathbf{B} \mathbf{y}_n}{\partial \mathbf{w}_n^{[k_1]}} = \frac{\partial}{\partial \mathbf{w}_n^{[k_1]}} \left(\sum_{l=1}^{K} [\mathbf{B}]_{k_2,l} \left(\mathbf{w}_n^{[l]}\right)^{\dagger} \mathbf{x}^{[l]} \right) = \mathbf{0},$$

for any arbitrary matrix, $\mathbf{B}$, with appropriate dimensions. Then the required gradients are

$$\frac{\partial \phi_n^{[k_2]}}{\partial \mathbf{w}_n^{[k_1]}} = \frac{\partial \mathbf{e}_{k_2}^{\mathsf{T}} \left(\mathbf{Q}_n^* \mathbf{y}_n^* + \tilde{\mathbf{Q}}_n^{\dagger} \mathbf{y}_n \right)}{\partial \mathbf{w}_n^{[k_1]}} = [\mathbf{Q}_n^*]_{k_2,k_1} \left(\mathbf{x}^{[k_1]}\right)^*, \tag{4.40}$$

and

$$\frac{\partial \phi_n^{*[k_2]}}{\partial \mathbf{w}_n^{[k_1]}} = \frac{\partial \mathbf{e}_{k_2}^{\mathsf{T}} \left(\mathbf{Q}_n \mathbf{y}_n + \tilde{\mathbf{Q}}_n \mathbf{y}_n^* \right)}{\partial \mathbf{w}_n^{[k_1]}} = \left[\tilde{\mathbf{Q}}_n\right]_{k_2,k_1} \left(\mathbf{x}^{[k_1]}\right)^*. \tag{4.41}$$

Thus for the complex-valued multivariate noncircular Gaussian,

$$\begin{aligned} E\left\{ \frac{\partial \phi_n^{*[k_2]}}{\partial \mathbf{w}_n^{[k_1]}} \left(\mathbf{x}^{[k_2]}\right)^{\dagger} \right\} &= E\left\{ \left[\tilde{\mathbf{Q}}_n\right]_{k_2,k_1} \left(\mathbf{x}^{[k_1]}\right)^* \left(\mathbf{x}^{[k_2]}\right)^{\dagger} \right\} \\ &= \left[\tilde{\mathbf{Q}}_n\right]_{k_2,k_1} \left(\mathbf{Q}_x^{[k_1,k_2]}\right)^*, \end{aligned} \tag{4.42}$$

and

$$\mathbf{H}_{\mathrm{B},n}^{[k_1,k_2]} = E\left\{ \frac{\partial \phi_n^{[k_2]}}{\partial \mathbf{w}_n^{[k_1]}} \left(\mathbf{x}^{[k_2]}\right)^{\mathsf{T}} \right\} = \left([\mathbf{Q}_n]_{k_2,k_1} \mathbf{R}_x^{[k_1,k_2]} \right)^*. \tag{4.43}$$

The calculations for the terms in the Newton update using noncircular multivariate Gaussian is now complete. In the next subsection, the simplifications to the algorithm when the strictly circular multivariate Gaussian source model is assumed are highlighted.

4.4.4 Circular Multivariate Gaussian Distribution

The results of the previous subsection simplify considerably when the strictly circular multivariate Gaussian source model is used. Namely, we have $\tilde{\mathbf{\Sigma}}_n = \mathbf{0}$, which implies the following $\tilde{\mathbf{Q}}_n = \mathbf{0}$, $\boldsymbol{\phi}_n = \mathbf{Q}_n^* \mathbf{y}_n^*$, $\mathbf{Q}_n = \mathbf{\Sigma}_n^{-1}$, and

$$\frac{\partial \mathcal{C}_{\text{IVA-G}}\left(\boldsymbol{\mathcal{W}}\right)}{\partial \left(\mathbf{w}_n^{[k]}\right)^*} = E\left\{\mathbf{x}^{[k]}\mathbf{y}_n^{\dagger}\right\}\mathbf{\Sigma}_n^{-1}\mathbf{e}_k - \frac{\mathbf{u}_n^{[k]}}{\left(\mathbf{w}_n^{[k]}\right)^{\dagger}\mathbf{u}_n^{[k]}}. \tag{4.44}$$

The form of this gradient has the same form as in the real-valued derivation [11]. Additionally, for the Newton version (4.42) becomes zero, implying that $\mathbf{H}_{\text{B},n} = \mathbf{0}$, i.e., $\mathbf{H}_n$ is block-diagonal.

4.5 Simulations

Next, we compare the performance of the proposed IVA-Gauss algorithms with several other IVA approaches, namely JDIAG-SOS and MCCA.

Performance is assessed using the joint ISI metric (2.62)—see accompanying description in Section 2.7.2.

4.5.1 Multivariate Gaussian Sources

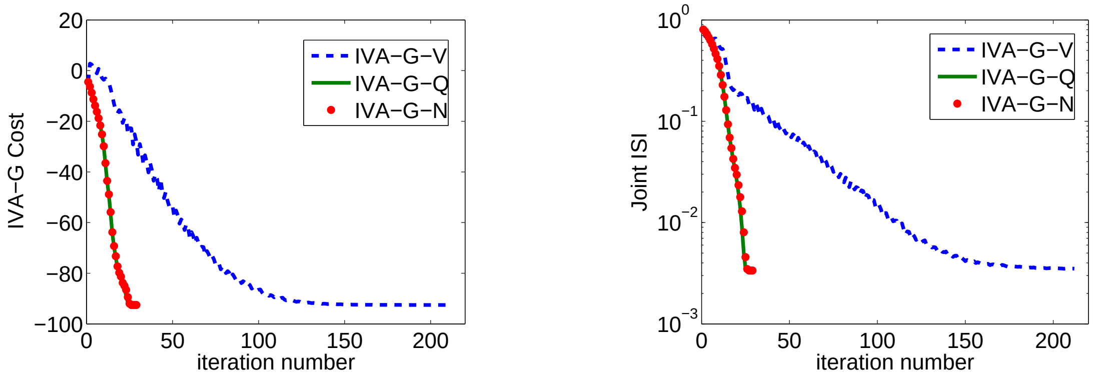
In this experiment, the nth SCV is a zero-mean, K dimensional, complex multivariate Gaussian random vector, $\mathbf{s}_n \sim \mathcal{N}\left(\mathbf{0}, \underline{\boldsymbol{\Sigma}}_n\right)$, specified by its covariance matrix, $\boldsymbol{\Sigma}_n = \mathbf{U}_n \mathbf{U}_n^{\dagger}$ and pseudo-covariance matrix, $\tilde{\boldsymbol{\Sigma}}_n = \mathbf{U}_n \mathrm{Diag}\left(\left[\rho_1, \ldots, \rho_K\right]\right) \mathbf{U}_n^{\mathsf{T}}$. The elements of $\mathbf{U}_n$ are from the standard complex-valued normal distribution, i.e., $\left[\mathbf{U}_n\right]_{k_1,k_2} \sim \mathcal{N}\left(0, 1\right)$, and the nature of the noncircularity coefficients, ρ_k, will vary for the experiments to follow.[2] For instance, when circular sources are considered then $\rho_k = 0$. The kth entry of each SCV is used as a latent source for the kth dataset. Entries of the random mixing matrices are also from the standard complex-valued normal distribution.

It is observed that the results presented are *comparatively* insensitive to sample size, so we only consider one sample size, $T = 10\,000$. We note that the sample size does need to be large enough to assure that the covariance matrices used in the IVA-Gauss algorithms are positive definite.

An illustration of the convergence behavior for a single trial is given in Fig. 4.1a of the cost function at each iteration for the circular multivariate Gaussian based IVA approaches (IVA-Gauss). We define an iteration as an update of all K demixing matrices. Here we illustrate how the rate of convergence is further improved by using Newton-based approaches. The gradient IVA-Gauss decoupled method, is compared to the Newton version. Additionally, in this example the quasi-Newton approach, which sets $\mathbf{H}_{\mathrm{A},n}$ to zero, is shown to be promising as it has nearly identical

---

[2] We note that ρ_k have also been termed circularity coefficients in [37], but we prefer to add the prefix non-, because $\rho_k = 0$ when there is no noncircularity, i.e., when the random variable is (second-order) circular.

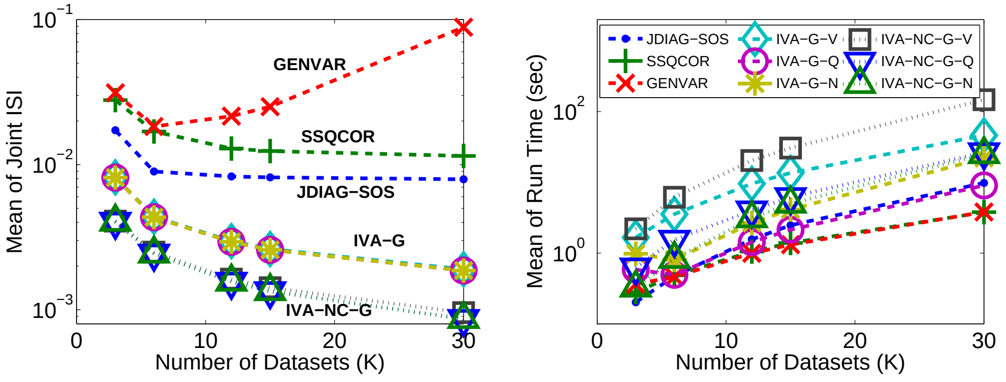


(a) Cost (b) ISI

Figure 4.1: Example of the IVA-Gauss cost (a) and performance measure (b) versus number of iterations for $K = 10$ datasets of $N = 10$ complex-valued circular multivariate Gaussian sources, using the vector gradient, quasi-Newton, and Newton optimization approaches.

convergence performance as that of the exact Newton approach. The joint ISI performance metric, shown in Fig. 4.1b for the same experiment, indicates that the joint ISI decreases as the cost function decreases.

In Fig. 4.2, Gaussian noncircular sources are considered and the joint ISI metric is shown versus the number of datasets. In particular, the noncircularity coefficients are uniformly distributed over the complex unit disc, i.e., $|\rho_k|^2 \sim \mathcal{U}(0,1)$ and $\arg(\rho_k) \sim \mathcal{U}(0,2\pi)$. The performance for IVA-Gauss is compared with MCCA using the SSQCOR and GENVAR cost functions defined in [53] and with the JDIAG-SOS of [75]. We see in Fig. 4.2a that the six IVA-Gauss optimization approaches provide superior joint source separation performance. The best performance is achieved when the noncircular Gaussian cost function, IVA-NC-G, is utilized. The optimization of IVA-Gauss and IVA-NC-G are denoted with postfix V, Q, and N for the vector gradient, quasi-Newton, and Newton algorithms, respectively. For the proposed IVA approaches, the joint ISI decreases as the number of datasets increases

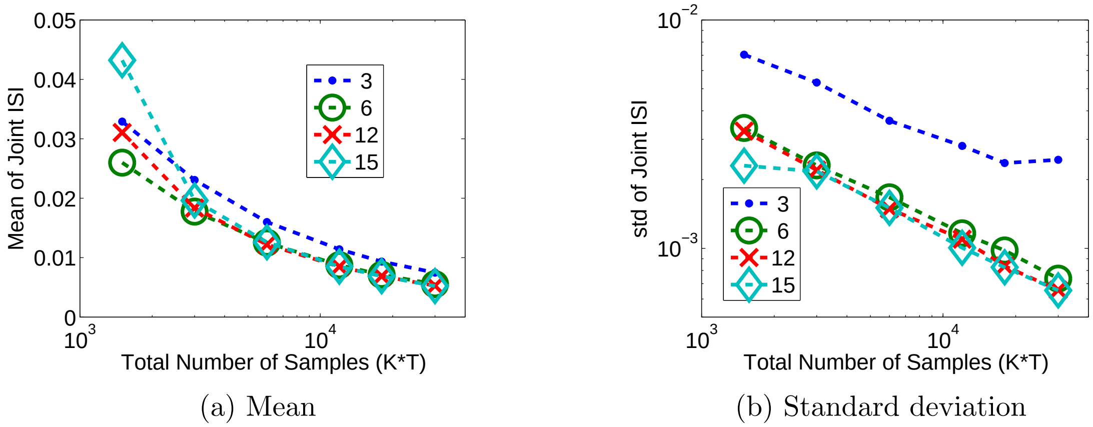


(a) ISI (b) Computation time

Figure 4.2: (a) Mean joint ISI of 100 trials versus the number of datasets K, using $T = 10000$ samples of $N = 10$ randomly correlated noncircular multivariate Gaussian sources. (b) Mean computation time for IVA algorithms.

with the largest rate of decrease occurring for the first few datasets. In Fig. 4.2b the average run time of the respective MATLAB implementations of each algorithm indicates that the Newton-based optimization of the noncircular IVA-Gauss is nearly the same as the circular counter-part IVA-G-N. The MCCA algorithms are computationally efficient but do not provide the same level of source separation performance as the more computationally complex JDIAG-SOS and IVA-Gauss approaches. The JDIAG-SOS approach achieves consistent, but suboptimal source separation performance due to the demixing matrices being constrained to be unitary. The computational efficiency and consistency of JDIAG-SOS source separation is such that we use it to initialize the IVA algorithms for the remaining IVA results.

The observed improvement as we increase the number of datasets is expected in part because of how the experiment above is constructed. As the number of datasets increases the total number of samples supplied to the algorithm also increases, i.e., $T_{\text{total}} \triangleq KT$ is not held constant. Here we demonstrate that there is some benefit



(a) Mean (b) Standard deviation

Figure 4.3: The mean (a) and standard deviation (b) of the joint ISI for 1000 trials are plotted versus the total number of samples for various number of datasets $K = 3$, 6, 12, and 15, with $N = 10$ complex-valued multivariate Gaussian sources.

when K is varied and T_{total} is constant. In Fig. 4.3a, the benefit of using more datasets even when decreasing the number of samples per dataset to keep the total number of samples constant shows a significant benefit from $K = 3$ to $K = 6$, with diminishing returns for larger K in terms of the mean joint ISI. The standard deviation of the joint ISI shown in Fig. 4.3b, exhibits the same behavior as the mean joint ISI in terms of diminishing gains in performance for larger K.

A noteworthy concern in the JBSS formulation is the assumption that the sources have a one-to-one dependence across the datasets. For some applications this assumption might only hold for a subset of the sources. Under such conditions the validity of using JBSS may be questioned. Here we illustrate via simulation that the subset of sources that possess one-to-one dependence may still be identified using the IVA formulation. In this experiment two of the $N = 10$ circular multivariate Gaussian SCVs will not possess any dependence across the datasets, i.e., for these sources we have $\mathbf{\Sigma} = \mathbf{I}$. The JBSS algorithms sort the estimated sources according to

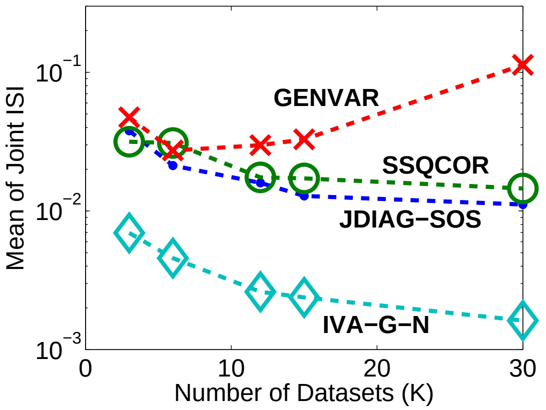


Figure 4.4: The mean of the joint ISI for 100 trials are versus the total number of datasets for various number of datasets K with $T = 10000$ samples and $N = 10$ complex-valued multivariate Gaussian sources. Two of the sources are independent and so the corresponding portions of the global demixing mixing matrices, $\mathbf{G}^{[k]}$, for those two sources are removed prior to computing the joint ISI.

the determinant of the estimated SCV covariance matrix. Then, the normalized joint ISI associated with the first eight estimated SCVs is computed. The performance shown in Fig. 4.4 indeed indicates that the *dependent* sources are reliably identified despite the presence of two SCVs with independent components.

The benefit of extending the source model to the noncircular case is highlighted by considering the following experiment to generate "pseudo-only correlated" SCVs. The covariance matrix of each SCV is $\boldsymbol{\Sigma}_n = \mathbf{I}$ for all sources. The pseudo-covariance matrix for the nth SCV, $\tilde{\boldsymbol{\Sigma}}_n$, varies and is generated using the eigenvectors of $\mathbf{U}_n\mathbf{U}_n^\dagger$, denoted by $\mathbf{F}_n$. To characterize the noncircularity of the sources we use $\kappa = 1 - \prod_{k=1}^{K}(1-\rho_k^2)$ to measure the *total* degree of noncircularity [103]. For this experiment, we only consider the case when all noncircularity coefficients are the same for a given total degree of noncircularity, i.e., $\rho = \rho_k$. We generate noncircular Gaussian sources with the pseudo-covariance matrix $\tilde{\boldsymbol{\Sigma}}_n = \rho\mathbf{F}_n\mathbf{F}_n^{\mathsf{T}}$. For

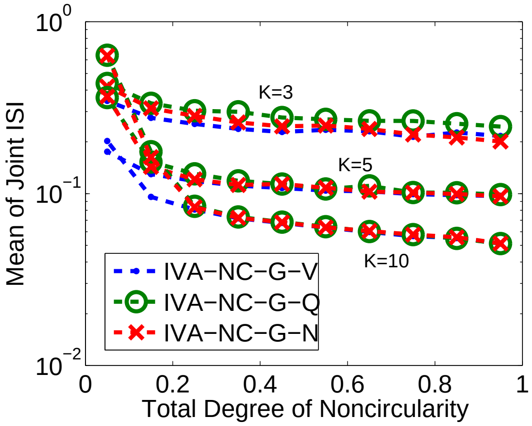


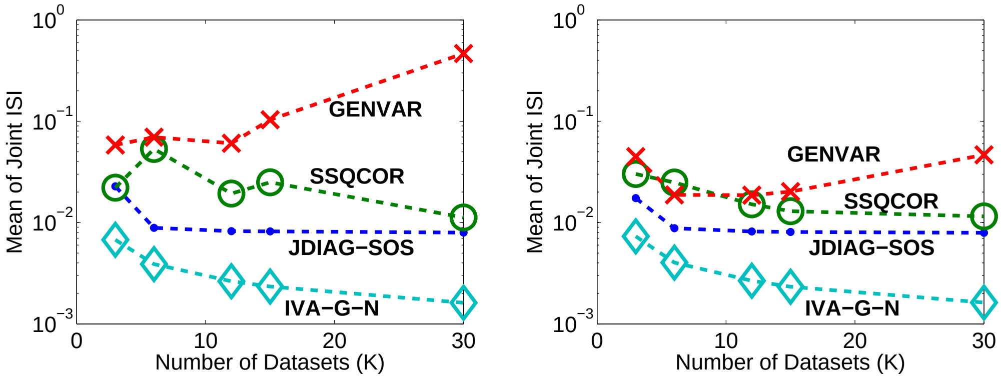
Figure 4.5: Mean joint ISI of 100 trials versus total degree of noncircularity, κ, using $T = 10000$ samples and $N = 5$ "pseudo-only" correlated noncircular multivariate Gaussian sources. The number of datasets, K, shown are 3, 5, and 10.

such "pseudo-only correlated" conditions, there are currently no other approaches for achieving BSS or JBSS, to the best of our knowledge. The results, shown in Fig. 4.5, indicate that performance degrades as the degree of the source noncircularity decreases.

4.5.2 Multivariate Non-Gaussian Sources

In many applications, the sources of interest are non-Gaussian, as is the case for ICA of a single dataset. For JBSS, the sources must possess second-order (linear) and/or higher-order (nonlinear) dependence across the datasets.

In certain applications, the multivariate source component vectors may possess marginals that are sub and super Gaussian. We consider sources from complex GGD [92] to demonstrate that for non-Gaussian sources, second-order methods are able to perform joint BSS when each source is correlated with a source in each dataset. Specifically, the elements of $\mathbf{z}_n$ are complex circular GGD sources with



(a) $N = 2$ sources (b) $N = 20$ sources

Figure 4.6: Average joint ISI for 100 trials versus the number of datasets K using linearly combined GGD sources for each SCV.

shape parameters selected such that the sources are equally likely to be super or sub-Gaussian and are uniformly distributed accordingly between $[0.1, 0.9]$ or $[1.1, 10]$, respectively. The parameterization for the shape parameter is such that a value of one generates Gaussian sources and positive values below and above one generate super and sub-Gaussian sources, respectively. Performance in Fig. 4.6 is similar to the performance in Fig. 4.2a, indicating reliable performance for IVA-Gauss with second-order correlated non-Gaussian SCV. The same robust behavior holds when the range of the GGD shape parameter is extended to generate either more super or sub-Gaussian sources.

In single dataset BSS analysis of complex-valued data, it has been recommended to use second-order methods, namely the strongly uncorrelated transform algorithm [36], as an additional preprocessing step (after whitening) for higher-order complex-valued ICA algorithms when the conditions for identification using only second-order statistics are not entirely satisfied. In a similar manner, the second-order method of IVA-Gauss can be considered as a pre-processing step for higher-order IVA methods, such as IVA-Lap, which are more likely to converge to local minimums as noted in [64].

Chapter 5

Conclusions and Future Work

> "When you know a thing, to hold that you know it, and when you do not know a thing, to allow that you do not know it – this is knowledge."
>
> – Confucius

In this dissertation we provide several important developments that greatly increase the understanding of IVA. In this chapter, we summarize our work and suggest several directions for further research.

5.1 Conclusions

In the motivation section of Chapter 1, we posed the general BSS problem and gave two motivating biomedical applications that have been of interest to the neuroimaging research community. IVA lives within the *general* BSS framework and has been shown to be useful in several applications. Our work has focused on providing a theoretical understanding of IVA. Without understanding when and how IVA works, it is difficult to appropriately explain results observed when applying IVA algorithms to 'real world' data. We began by extending the original IVA formulation from the iid observation assumption to the case when the samples are not iid using a likelihood function. Maximization of the likelihood function is equivalent to the minimization of mutual information among SCVs when the number of observations approaches infinity and the true source distribution is included within the set of modeled sources. Then by examination of the FIM associated with the likelihood function we are able to state the conditions when the sources can be recovered by IVA to within a scaling and permutation ambiguity. We have found that the identification conditions and performance bounds are generalizations of the corresponding theoretical results for ICA. With these results, researchers and practitioners alike can know under what conditions the sources can be recovered using IVA and how well the sources can be separated (if the assumed model is accurate).

A number of powerful algorithms for achieving IVA are now available for researchers to use. We have made several contributions in this dissertation which expand the selection of IVA algorithms and provide a framework for the development of algorithms in the future. We recognized the existence of several algorithms for achieving IVA which existed prior to the development of the IVA concept, and even before ICA. The efficacy of an IVA algorithm depends on the particulars of the problem at hand. Here we exploit a decoupling method that enables algorithms having quasi-Newton like convergence while still preserving the larger optimization space of nonorthgonal (or nonunitary if complex) matrices. The latter feature is desirable in many real-world applications where the data does not precisely follow all of the modeling assumptions. Models for the sources can be constructed by the practitioner and applied using the algorithmic framework given in this work. In this dissertation, we have first utilized the ubiquitous Gaussian distribution as a source model. When the sources follow a Gaussian distribution the algorithms are shown to be efficient and accurate. When considering complex-valued IVA, a Gaussian distribution source model is useful for demonstrating how noncircular sources can be used to achieve source separation. The Gaussian source model also demonstrates the direct connection between ICA and CCA which is provided by IVA. The relationship between the IVA-Gauss cost function and the GENVAR cost function used in MCCA is illuminating, as MCCA is one of many intimately related methods from multivariate analysis [107]. According to the IVA theory we developed for the dissertation, the Gaussian source model provides a lower bound on performance. By considering higher-order statistics and nonlinear dependencies additional separability and improved performance can be had. To enable such performance, we introduce the use of the general Kotz distribution family as a source model—which includes MPE and Gaussian distributions as special cases.

5.2 Future Work

Since the idea of IVA is still relatively new, there are still new theoretical and algorithmic developments to pursue. Below we provide some suggestions on where we plan to continue our research.

- Theory: *Identification conditions for complex-valued data*

  As shown by the identification conditions, several types of source diversity can be exploited to achieve source separation. Our theory however, did not examine the case when the data is complex-valued. For the case of ICA, it has been shown that noncircular sources provide an additional source diversity for identifiability [36, 81], which we have shown to numerically also extend to IVA. We expect that the method of proving identifiability for the real-valued can be extended to complex-valued IVA. We also note that this generalization for IVA will also establish the necessary *and* sufficient identification conditions for complex-valued ICA.

- Algorithms: *Extension of entropy bound minimization (EBM) to IVA*

  In this work, we have chosen to assume a family of multivariate distributions for evaluating and optimizing the IVA cost function(s). This approach is analogous to many of the algorithms used for ICA. EBM is a more recent algorithm for ICA [74]. It utilizes, what we would call, a semiparametric approach to find the lowest upper bound on entropy from a suite of measuring functions. EBM has been shown to be a robust algorithm for achieving source separation of single dataset. Non-Gaussian multivariate distributions are notoriously difficult to estimate, which is why the algorithms thus far have been parametric. We envision that an extension of EBM could be made to multivariate sources that can surmount these difficulties.

- Algorithms: *Accounting for sample-to-sample dependence*

  Only one algorithm, which we are aware of, accounts for sample-to-sample dependence of the sources for achieving IVA—namely, JDIAG-SOS [76, 75]. The current solution is limited to orthogonal demixing matrices and is thus suboptimal theoretically. Approaches such as second-order blind identification [18, 17] and weights-adjusted second-order blind identification [118] account for dependency between samples with a Gaussian random process in order to achieve ICA. These should be able to be extended to IVA. Furthermore, algorithms that exploit non-Gaussianity and sample dependence jointly, such as entropy rate bound minimization [72], are also candidates for extensions to IVA.

- Algorithms: *Constrained IVA*

  The recent work of [99] has made use of the NDT to incorporate prior knowledge in the form of soft-constraints on the demixing matrix and/or sources within the ICA framework. The concepts of constrained ICA are directly applicable to IVA, and will likely improve robustness to noise. The IVA framework does not assume any relationship exists between the mixing matrices. In some applications, it may be reasonable to assume some correlation exists between the corresponding demixing vectors of different datasets. With a constrained IVA framework this assumption could readily be accounted for.

- Algorithms: *Non-Gaussian IVA Online Learning*

  When we consider the update rule for non-Gaussian sources, we encounter difficulties for online learning since the update rule (score function) nonlinearly depends on the data (unlike in IVA-Gauss). To overcome this difficulty we can consider linear approximations of the nonlinear score function. Another possibility is following the work of [56], where a stochastic gradient is used but future work would be for a broader class of sources models, e.g., MPE with second-order correlation.

Appendix A

IVA Fisher Information Matrix

Here we derive the FIM of (2.2) wrt $\boldsymbol{\mathcal{W}}$. The KN^2 parameters result in $KN^2 \times KN^2$ dimension FIM with the entry associated with $w^{[k_1]}_{m_1,n_1}$ and $w^{[k_2]}_{m_2,n_2}$ given by:

$$[\mathbf{F}(\boldsymbol{\mathcal{W}})]^{k_1,m_1,n_1}_{k_2,m_2,n_2} \triangleq E\left\{\frac{\partial \mathcal{L}(\boldsymbol{\mathcal{W}})}{\partial w^{[k_1]}_{m_1,n_1}} \frac{\partial \mathcal{L}(\boldsymbol{\mathcal{W}})}{\partial w^{[k_2]}_{m_2,n_2}}\right\} = -E\left\{\frac{\partial^2 \mathcal{L}(\boldsymbol{\mathcal{W}})}{\partial w^{[k_2]}_{m_2,n_2} \partial w^{[k_1]}_{m_1,n_1}}\right\} \tag{A.1}$$

For the computations to follow it is useful to observe that,

$$\frac{\partial \log\left|\det \mathbf{W}^{[l]}\right|}{\partial w^{[k]}_{m,n}} = \delta_{l,k} \mathbf{e}_m^{\mathsf{T}} \frac{\partial \log\left|\det \mathbf{W}^{[k]}\right|}{\partial \mathbf{W}^{[k]}} \mathbf{e}_n = \delta_{l,k} \mathbf{e}_m^{\mathsf{T}} \left(\mathbf{W}^{[k]}\right)^{-\mathsf{T}} \mathbf{e}_n$$

$$\mathbf{Y}_m^{\mathsf{T}} = \left[\left(\mathbf{X}^{[1]}\right)^{\mathsf{T}} \mathbf{w}^{[1]}_m, \ldots, \left(\mathbf{X}^{[K]}\right)^{\mathsf{T}} \mathbf{w}^{[K]}_m\right] \in \mathbb{R}^{K \times V},$$

$$\frac{\partial \mathbf{Y}_l}{\partial w^{[k]}_{m,n}} = \delta_{l,m} \mathrm{Diag}\left(\mathbf{e}_k\right) \mathbf{X}_n \in \mathbb{R}^{K \times T},$$

and

$$\frac{\partial \log\left(p_m\left(\mathbf{Y}_m\right)\right)}{\partial w_{m,n}^{[k]}} = \operatorname{tr}\left(\frac{\partial \log\left(p_m\left(\mathbf{Y}_m\right)\right)}{\partial \mathbf{Y}_m^{\mathsf{T}}} \frac{\partial \mathbf{Y}_m}{\partial w_{m,n}^{[k]}}\right) \tag{A.2}$$

$$= -\operatorname{tr}\left(\mathbf{\Phi}_m^{\mathsf{T}} \operatorname{Diag}\left(\mathbf{e}_k\right) \mathbf{X}_n\right) \tag{A.3}$$

$$= -\left(\boldsymbol{\phi}_m^{[k]}\right)^{\mathsf{T}} \mathbf{x}_n^{[k]}, \tag{A.4}$$

where $\mathbf{\Phi}_m\left(\mathbf{Y}_m\right) \triangleq -\partial \log\left(p_m\left(\mathbf{Y}_m\right)\right)/\partial \mathbf{Y}_m = \mathbf{\Phi}_m \in \mathbb{R}^{K \times T}$ and $\boldsymbol{\phi}_m^{[k]} = \mathbf{\Phi}_m^{\mathsf{T}} \mathbf{e}_k$ is a column vector given by entries in the kth row of $\mathbf{\Phi}_m$. Note that (A.2) is due to applying the chain rule given in [96, Sect. 2.8.1]. Thus the gradient of the likelihood function in (2.2) is

$$\begin{aligned}
\frac{\partial \mathcal{L}\left(\boldsymbol{\mathcal{W}}\right)}{\partial w_{m,n}^{[k]}} &= \frac{\partial \sum_{i=1}^{N} \log\left(p_i\left(\mathbf{Y}_i\right)\right) + T \sum_{j=1}^{K} \log\left|\det \mathbf{W}^{[j]}\right|}{\partial w_{m,n}^{[k]}} \\
&= \frac{\partial \log\left(p_m\left(\mathbf{Y}_m\right)\right)}{\partial w_{m,n}^{[k]}} + T \frac{\partial \log\left|\det \mathbf{W}^{[k]}\right|}{\partial w_{m,n}^{[k]}} \\
&= \left(\frac{\partial \log\left(p_m\left(\mathbf{Y}_m\right)\right)}{\partial \mathbf{y}_m^{[k]}}\right)^{\mathsf{T}} \frac{\partial \mathbf{y}_m^{[k]}}{\partial w_{m,n}^{[k]}} + T \mathbf{e}_m^{\mathsf{T}} \frac{\partial \log\left|\det \mathbf{W}^{[k]}\right|}{\partial \mathbf{W}^{[k]}} \mathbf{e}_n \\
&= -\left(\boldsymbol{\phi}_m^{[k]}\right)^{\mathsf{T}} \mathbf{x}_n^{[k]} + T \mathbf{e}_m^{\mathsf{T}} \left(\mathbf{W}^{[k]}\right)^{-\mathsf{T}} \mathbf{e}_n \\
&= -\left(\boldsymbol{\phi}_m^{[k]}\right)^{\mathsf{T}} \mathbf{x}_n^{[k]} + T w_{n,m}^{-[k]},
\end{aligned}$$

where $w_{m,n}^{-[k]}$ is the entry in the mth row and nth column $\left(\mathbf{W}^{[k]}\right)^{-1}$.

Letting $\mathbf{A}=\mathbf{I}$ we have

$$\begin{aligned} E\left\{\frac{\partial \mathcal{L}\left(\boldsymbol{\mathcal{W}}\right)}{\partial w_{m_1,n_1}^{[k_1]}}\frac{\partial \mathcal{L}\left(\boldsymbol{\mathcal{W}}\right)}{\partial w_{m_2,n_2}^{[k_2]}}\right\}\Bigg|_{\mathbf{A}=\mathbf{I}} =&E\left\{\left(\boldsymbol{\phi}_{m_1}^{[k_1]}\right)^{\mathsf{T}}\mathbf{s}_{n_1}^{[k_1]}\left(\mathbf{s}_{n_2}^{[k_2]}\right)^{\mathsf{T}}\boldsymbol{\phi}_{m_2}^{[k_2]}\right\}+T^2 w_{n_1,m_1}^{-[k_1]}w_{n_2,m_2}^{-[k_2]} \\ &-TE\left\{\left(\boldsymbol{\phi}_{m_2}^{[k_2]}\right)^{\mathsf{T}}\mathbf{s}_{n_2}^{[k_2]}\right\}w_{n_1,m_1}^{-[k_1]} \\ &-TE\left\{\left(\boldsymbol{\phi}_{m_1}^{[k_1]}\right)^{\mathsf{T}}\mathbf{s}_{n_1}^{[k_1]}\right\}w_{n_2,m_2}^{-[k_2]}, \end{aligned} \tag{A.5}$$

and the FIM of interest with entries given by

$$[\mathbf{F}]_{k_2,m_2,n_2}^{k_1,m_1,n_1}\triangleq\left[\mathbf{F}\left(\boldsymbol{\mathcal{W}}\right)\right]_{k_2,m_2,n_2}^{k_1,m_1,n_1}\Big|_{\mathbf{A}=\mathbf{I},\mathbf{W}=\mathbf{I}} \tag{A.6}$$

$$\begin{aligned} =&E\left\{\left((\boldsymbol{\phi}_{m_1}^{[k_1]}\right)^{\mathsf{T}}\mathbf{s}_{n_1}^{[k_1]}\left(\mathbf{s}_{n_2}^{[k_2]}\right)^{\mathsf{T}}\boldsymbol{\phi}_{m_2}^{[k_2]}\right\}+T^2\delta_{m_1,n_1}\delta_{m_2,n_2} \\ &-TE\left\{\left(\boldsymbol{\phi}_{m_2}^{[k_2]}\right)^{\mathsf{T}}\mathbf{s}_{n_2}^{[k_2]}\right\}\delta_{m_1,n_1}-TE\left\{\left(\boldsymbol{\phi}_{m_1}^{[k_1]}\right)^{\mathsf{T}}\mathbf{s}_{n_1}^{[k_1]}\right\}\delta_{m_2,n_2} \\ =&E\left\{\left(\boldsymbol{\phi}_{m_1}^{[k_1]}\right)^{\mathsf{T}}\mathbf{s}_{n_1}^{[k_1]}\left(\mathbf{s}_{n_2}^{[k_2]}\right)^{\mathsf{T}}\boldsymbol{\phi}_{m_2}^{[k_2]}\right\}-T^2\delta_{m_1,n_1}\delta_{m_2,n_2}, \end{aligned} \tag{A.7}$$

where the following expression holds, $E\left\{\mathbf{s}_{n}^{[k_1]}\left(\boldsymbol{\phi}_{m}^{[k_2]}\right)^{\mathsf{T}}\right\}=\delta_{k_1,k_2}\delta_{m,n}\mathbf{I}_T$, see [28]. Since, by assumption, both $E\left\{\boldsymbol{\phi}_{m}^{[k]}\right\}=\mathbf{0}$ and $E\left\{\mathbf{s}_{m}^{[k]}\right\}=\mathbf{0}$, then it is true that $[\mathbf{F}]_{k_2,m_2,n_2}^{k_1,m_1,n_1}=0$ when one of the entries in (m_1,n_1,m_2,n_2) is unique. It is also zero when $m_1=n_1\neq m_2=n_2$, i.e., $E\left\{\left(\boldsymbol{\phi}_{m_1}^{[k_1]}\right)^{\mathsf{T}}\mathbf{s}_{m_1}^{[k_1]}\right\}E\left\{\left(\mathbf{s}_{m_2}^{[k_2]}\right)^{\mathsf{T}}\boldsymbol{\phi}_{m_2}^{[k_2]}\right\}-T^2=$

$T^2 - T^2 = 0$. Thus, there are only three nonzero cases to consider:

$$[\mathbf{F}]_{k_2,m_2,n_2}^{k_1,m_1,n_1} = \begin{cases} E\left\{\left(\boldsymbol{\phi}_{m_1}^{[k_1]}\right)^{\mathsf{T}} \mathbf{s}_{m_1}^{[k_1]} \left(\mathbf{s}_{m_1}^{[k_2]}\right)^{\mathsf{T}} \boldsymbol{\phi}_{m_1}^{[k_2]}\right\} - T^2 & m_1 = n_2 = m_2 = n_1 \\ \operatorname{tr}\left(E\left\{\boldsymbol{\phi}_{m_1}^{[k_2]} \left(\boldsymbol{\phi}_{m_1}^{[k_1]}\right)^{\mathsf{T}}\right\} E\left\{\mathbf{s}_{n_1}^{[k_1]} \left(\mathbf{s}_{n_1}^{[k_2]}\right)^{\mathsf{T}}\right\}\right) & m_1 = m_2 \neq n_1 = n_2 \\ \operatorname{tr}\left(E\left\{\boldsymbol{\phi}_{m_1}^{[k_1]} \left(\mathbf{s}_{m_1}^{[k_2]}\right)^{\mathsf{T}}\right\} E\left\{\boldsymbol{\phi}_{m_2}^{[k_2]} \left(\mathbf{s}_{m_2}^{[k_1]}\right)^{\mathsf{T}}\right\}\right) & m_1 = n_2 \neq m_2 = n_1 \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} T\left(\mathcal{K}_{m_1,m_1}^{[k_1,k_2]} - T\right) & m_1 = n_2 = m_2 = n_1 \\ T\mathcal{K}_{m_1,n_1}^{[k_1,k_2]} & m_1 = m_2 \neq n_1 = n_2 \\ T\delta_{k_1,k_2} & m_1 = n_2 \neq m_2 = n_1 \\ 0 & \text{otherwise,} \end{cases}$$

where the (k_1, k_2) entry of $\boldsymbol{\mathcal{K}}_{m,n}$ is

$$\begin{aligned} \mathcal{K}_{m,n}^{[k_1,k_2]} &\triangleq \frac{1}{T} E\left\{\left(\boldsymbol{\phi}_{m}^{[k_1]}\right)^{\mathsf{T}} \mathbf{s}_{n}^{[k_1]} \left(\mathbf{s}_{n}^{[k_2]}\right)^{\mathsf{T}} \boldsymbol{\phi}_{m}^{[k_2]}\right\} \\ &= \frac{1}{T} \operatorname{tr}\left(E\left\{\boldsymbol{\phi}_{m}^{[k_2]} \left(\boldsymbol{\phi}_{m}^{[k_1]}\right)^{\mathsf{T}} \mathbf{s}_{n}^{[k_1]} \left(\mathbf{s}_{n}^{[k_2]}\right)^{\mathsf{T}}\right\}\right). \end{aligned} \tag{A.8}$$

The form of the FIM (e.g., see Fig. 2.1) is the block-matrix extension of that for the single dataset FIM given in [81].

There exists a permuted FIM in which there are $N + N\left(N-1\right)/2$ nonzero matrices along the diagonal, i.e.,

$$\mathbf{F} = \begin{bmatrix} \oplus_{n=1}^{N}\mathbf{F}_n & \mathbf{0} \\ \mathbf{0} & \oplus_{m=1,n=m+1}^{N,N}\mathbf{F}_{m,n} \end{bmatrix}. \tag{A.9}$$

The submatrices are given by

$$\mathbf{F}_n \triangleq \mathrm{var}\left\{\mathrm{diag}\left(\mathbf{\Phi}_n\mathbf{S}_n^{\mathsf{T}} - \mathbf{I}_T\right)\right\} = T\left(\boldsymbol{\mathcal{K}}_{n,n} - T\mathbf{1}_{K\times K}\right) \in \mathbb{R}^{K\times K} \tag{A.10}$$

and

$$\mathbf{F}_{m,n} \triangleq \mathrm{cov}\left\{\begin{bmatrix} \mathrm{diag}\left(\mathbf{\Phi}_m\mathbf{S}_n^{\mathsf{T}}\right) \\ \mathrm{diag}\left(\mathbf{\Phi}_n\mathbf{S}_m^{\mathsf{T}}\right) \end{bmatrix}\right\} = T\begin{bmatrix} \boldsymbol{\mathcal{K}}_{m,n} & \mathbf{I}_K \\ \mathbf{I}_K & \boldsymbol{\mathcal{K}}_{n,m} \end{bmatrix} \in \mathbb{R}^{2K\times 2K}. \tag{A.11}$$

It is also useful to note that for $1 \leq m \neq n \leq N$ we have $\mathcal{K}_{m,n}^{[k_1,k_2]} = \frac{1}{T}\mathrm{tr}\left(\mathbf{\Gamma}_m^{[k_2,k_1]}\mathbf{R}_n^{[k_1,k_2]}\right)$, where $\mathbf{R}_n^{[k_1,k_2]} \triangleq E\left\{\mathbf{s}_n^{[k_1]}\left(\mathbf{s}_n^{[k_2]}\right)^{\mathsf{T}}\right\} \in \mathbb{R}^{T\times T}$ and $\mathbf{\Gamma}_n^{[k_1,k_2]} \triangleq E\left\{\boldsymbol{\phi}_n^{[k_1]}\left(\boldsymbol{\phi}_n^{[k_2]}\right)^{\mathsf{T}}\right\} \in \mathbb{R}^{T\times T}$.

To be complete we confirm that the regularity condition holds, i.e., $E\left\{\frac{\partial\mathcal{L}(\boldsymbol{\mathcal{W}})}{\partial w_{m,n}^{[k]}}\right\} =$

0, around the optimal solution,

$$E\left\{\frac{\partial\mathcal{L}\left(\boldsymbol{\mathcal{W}}\right)}{\partial w_{m,n}^{[k]}}\right\}\Bigg|_{\mathbf{A}=\mathbf{I},\mathbf{W}=\mathbf{I}} = -E\left\{\left(\boldsymbol{\phi}_m^{[k]}\right)^{\mathsf{T}}\mathbf{x}_n^{[k]}\right\}\Bigg|_{\mathbf{A}=\mathbf{I},\mathbf{W}=\mathbf{I}} + Tw_{n,m}^{-[k]}\Big|_{\mathbf{W}=\mathbf{I}} \tag{A.12}$$

$$= -E\left\{\left(\boldsymbol{\phi}_m^{[k]}\right)^{\mathsf{T}}\mathbf{s}_n^{[k]}\right\} + T\delta_{n,m} \tag{A.13}$$

$$= \begin{cases} -\mathrm{tr}\left(\mathbf{I}_T\right) + T & m = n \\ -E\left\{\boldsymbol{\phi}_m^{[k]}\right\}^{\mathsf{T}} E\left\{\mathbf{s}_n^{[k]}\right\} & m \neq n \end{cases} \tag{A.14}$$

$$= 0. \tag{A.15}$$

Appendix B

Decoupled ICA

We utilize the NDT to develop a new decoupled ICA algorithm that uses Newton optimization enabling superior performance when the sample size is limited. Our formulation is the special case of IVA with one dataset of iid samples and the proposed algorithm can be derived starting from either (2.7) or (2.8). We use the latter, with $K = 1$ to define the ICA cost function:

$$\mathcal{C}_{\text{ICA}}\left(\mathbf{W}\right) \triangleq \sum_{n=1}^{N} \mathcal{H}\left\{y_n\right\} - \log\left|\det \mathbf{W}\right|. \tag{B.1}$$

The role of the $\log\left|\det\left(\mathbf{W}\right)\right|$ portion of the cost function is to act as a regularization term. Since entropy is not scale invariant, i.e., $\mathcal{H}\left\{z\right\} \neq \mathcal{H}\left\{az\right\}$ for $a \neq 1$, then without the regularization term the cost function could be minimized by scaling the estimated sources. A simplification of the cost function can be had by restricting the set of permissible demixing matrices to those that achieve the following *necessary condition* for estimating independent sources, namely requiring that the estimated sources are second-order uncorrelated, or $E\left\{\mathbf{y}\mathbf{y}^{\mathsf{T}}\right\} = \mathbf{D}$, where $\mathbf{D}$ is a full rank diagonal matrix. This necessary condition strictly holds as $T \to \infty$, and is achieved by a decorrelation or whitening matrix, $\mathbf{Q}$, such that $E\left\{\mathbf{z}\mathbf{z}^{\mathsf{T}}\right\} = \mathbf{I}$, where $\mathbf{z} = \mathbf{U}\mathbf{x}$. The set of all demixing matrices which meet this second-order requirement can then be expressed as $\mathbf{W} = \mathbf{V}\mathbf{Q}$, where $\mathbf{V}$ is an orthonormal matrix to be estimated via optimization of this simplified ICA cost function,

$$\mathcal{C}_{\mathrm{ICA}_o}(\mathbf{V}) = \sum_{n=1}^{N} \mathcal{H}\{y_n\} - C_1. \tag{B.2}$$

We have used $C_1 \triangleq \log|\det(\mathbf{W})|$, since this is now a constant wrt the orthogonal optimization parameter $\mathbf{V}$. It should be clear that prewhitening data using $\mathbf{U}$ does not make (B.1) and (B.2) equivalent. Data can be whitened prior to using (2.8) but the data *must* be whitened to use (B.2). In the above and throughout the sequel, we restrict our discussion to the real domain and note that similar results can be achieved with proper analysis for the complex domain.

The more restrictive cost function given in (B.2) is widely used as it allows the estimation of each source component using standard vector optimization procedures. If T is not sufficiently large then the accuracy of the whitening matrix, $\mathbf{Q}$, is degraded and the restriction on decomposing $\mathbf{W}$ as the product of an orthogonal matrix and a whitening matrix may degrade source estimation performance. This is the primary motivation for considering the more general cost function of (B.1) rather than the simpler cost function of (B.2). For simplicity, we will refer to (B.1) and (B.2) as the nonorthogonal and orthogonal ICA cost functions, respectively.

B.1 Decoupled Maximum Likelihood ICA

In this section, we use the decoupling procedure derived in Section 3.2.2 to develop a new decoupled maximum likelihood ICA algorithm. To do so, we use the more general ICA cost function given in (B.1), whose minimization corresponds to minimizing the mutual information of the estimated components. The demixing matrix estimate can be optimized by minimizing the cost function using either gradient or Newton-based optimization methods. Previous optimization approaches compute the derivative of (B.1) wrt the entire demixing matrix. To achieve faster convergence and to avoid computing the inverse of the demixing matrix the natural (relative) gradient algorithm can be used [28]. Although Newton optimization procedures for updating the demixing matrix are theoretically possible, they are generally not practical for moderate to large values of N because the dimension of the Hessian grows as N^2. Here, we achieve a Newton-based algorithm by using the NDT.

The gradient of the ICA cost function (B.1) wrt the nth row of $\mathbf{W}$ is

$$\frac{\partial \mathcal{C}_{\text{ICA}}(\mathbf{W})}{\partial \mathbf{w}_n} = E\{\phi_n \mathbf{x}\} - \frac{\mathbf{u}_n}{\mathbf{w}_n^{\mathsf{T}} \mathbf{u}_n}, \tag{B.3}$$

where we have applied the chain rule and used $\phi_n \triangleq -d \log p_n(y_n)/dy_n$, termed the score function. For a Newton update, the Hessian can be computed using

$$\frac{\partial^2 \mathcal{C}_{\text{ICA}}(\mathbf{W})}{\partial \mathbf{w}_n \partial \mathbf{w}_n^{\mathsf{T}}} = E\left\{\phi_n' \mathbf{x}\mathbf{x}^{\mathsf{T}}\right\} + \frac{1}{\left(\mathbf{w}_n^{\mathsf{T}} \mathbf{u}_n\right)^2} \tilde{\mathbf{U}}_n, \tag{B.4}$$

where $\phi_n' \triangleq d\phi(y_n)/dy_n$ is defined provided the pdf is twice differentiable. Given such a score function, a Newton update of the nth demixing vector is

$$\mathbf{w}_n^{\text{new}} \leftarrow \mathbf{w}_n^{\text{old}} - \mu \left(\frac{\partial^2 \mathcal{C}_{\text{ICA}}(\mathbf{W})}{\partial \mathbf{w}_n \partial \mathbf{w}_n^{\mathsf{T}}}\right)^{-1} \frac{\partial \mathcal{C}_{\text{ICA}}(\mathbf{W})}{\partial \mathbf{w}_n}, \tag{B.5}$$

where $\mu > 0$ is a step-size parameter, which is one for a truly Newton algorithm but its value can be adjusted to control the convergence speed.

The computations required to compute the Hessian directly as expressed in (B.4) and the Hessian inverse as required in (B.5) are potentially computationally burdensome. To reduce the computational operations required to iteratively update $\mathbf{w}_n$ we consider prewhitened data. Then a simplifying assumption used in [50] to derive the FastICA algorithm can be considered, namely $E\left\{\phi_n' \mathbf{x}\mathbf{x}^{\mathsf{T}}\right\} \approx E\left\{\phi'\left(y_n\right)\right\}\mathbf{I}$. Using this simplification provides the following Hessian approximation,

$$\frac{\partial^2 \mathcal{C}_{\mathrm{ICA}}\left(\mathbf{W}\right)}{\partial \mathbf{w}_n \partial \mathbf{w}_n^{\mathsf{T}}} \approx E\left\{\phi_n'\right\}\mathbf{I} + \frac{1}{\left(\mathbf{w}_n^{\mathsf{T}}\mathbf{u}_n\right)^2}\tilde{\mathbf{U}}_n. \tag{B.6}$$

Furthermore, by the matrix inversion lemma [47], we have the following expression for the Hessian inverse,

$$\begin{aligned}\left(\frac{\partial^2 \mathcal{C}_{\mathrm{ICA}}\left(\mathbf{W}\right)}{\partial \mathbf{w}_n \partial \mathbf{w}_n^{\mathsf{T}}}\right)^{-1} &\approx \left(E\left\{\phi_n'\right\}\mathbf{I} + \frac{1}{\left(\mathbf{w}_n^{\mathsf{T}}\mathbf{u}_n\right)^2}\tilde{\mathbf{U}}_n\right)^{-1} \\ &= \gamma_n \mathbf{I} - \frac{\gamma_n^2}{\alpha_n + \gamma_n}\tilde{\mathbf{U}}_n, \end{aligned} \tag{B.7}$$

where $\gamma_n^{-1} \triangleq E\left\{\phi_n'\right\}$ and $\alpha_n \triangleq \left(\mathbf{w}_n^{\mathsf{T}}\mathbf{u}_n\right)^2$. Using the approximation for the Hessian inverse of (B.7) in (B.5), we have the following computationally efficient quasi-Newton update rule for decoupled ICA (D-ICA)

$$\mathbf{w}_n^{\mathrm{new}} \leftarrow \mathbf{w}_n^{\mathrm{old}} - \mu\left(\gamma_n \mathbf{I} - \frac{\gamma_n^2}{\alpha_n + \gamma_n}\tilde{\mathbf{U}}_n\right)\frac{\partial \mathcal{C}_{\mathrm{ICA}}\left(\mathbf{W}\right)}{\partial \mathbf{w}_n}. \tag{B.8}$$

B.1.1 Algorithm Details

In the previous subsection, the two fundamental update equations for D-ICA are given in (B.5) and (B.8). In this subsection, we provide implementation details.

A prudent practice is to initialize nonorthogonal algorithms with solutions from (faster) orthogonal algorithms. For this work, we have chosen to use the solution of the popular orthogonal FastICA [50] to initialize the estimate of the demixing matrix. By doing so, the nonorthogonal D-ICA algorithm refines the orthogonal FastICA solution so that only a few iterations by D-ICA are necessary to converge. Thus concerns about the additional computational cost of the decoupling procedure can be reduced considerably. Additionally, a fast recursive method for computing $\mathbf{u}_n$ can be used [74].

In each iteration, all the rows of the demixing matrix are updated using (B.8). After each update the demixing row vectors are normalized to have unit length. As the algorithm converges measures of changes in the demixing vector estimates between iterations, such as $\theta_n \triangleq 1 - \left|\mathbf{w}_{n,\mathrm{old}}^{\mathsf{T}} \mathbf{w}_{n,\mathrm{new}}\right|$, become very small. We deem that convergence is achieved when $\max\left(\theta_1, \ldots, \theta_N\right) < \epsilon$, where ϵ is a small positive number with a typical value of 10^{-6}. The quasi-Newton update of (B.8) is used until the last iteration, which uses the more computationally expensive Newton update of (B.5). For both the exact and quasi-Newton update rules we have used $\mu = 1$ for the step-size parameter.

Lastly, to implement D-ICA we specify an a distribution model, namely a member of the inverse-cosh family of distributions, $p\left(y\right) \propto 1/\cosh^{1/\beta}\left(\beta y\right)$. In par-

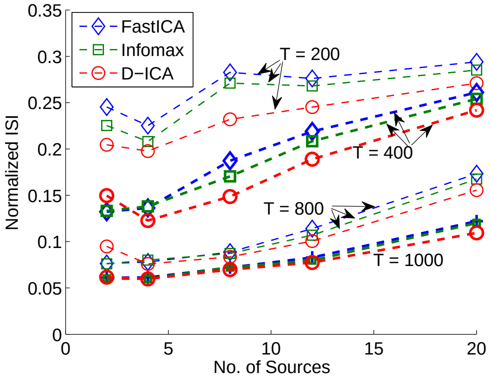


Figure B.1: Average normalized ISI of 50 trials for 200, 400, 800, and 1000 samples versus number of sources.

ticular, we let $\beta = 1/2$, then $\phi(y) = \tanh(y/2)$ and $\phi'(y) = \left(1 - \phi^2(y)\right)/2$.

B.1.2 Algorithm Performance

To demonstrate the performance, we simulate sources as iid samples of the inverse-cosh distribution described above. For comparison, we compare the performance of D-ICA with FastICA and Infomax [16]. Both Infomax and D-ICA use the symmetric FastICA solution for initialization. All three algorithms are using the same source density matching tangent hyperbolic score function. To compare performance we consider the normalized ISI, see (2.60) for definition.

In each experiment, the elements of the mixing matrix are drawn from the standard normal distribution. The average normalized ISI of 50 trials for various number of sources and sample sizes are shown in Fig. B.1. For this example, the D-ICA algorithm provides equal or better performance than FastICA and Infomax for all experimental settings and provides the largest benefit when the sample size is small.

B.2 Conclusions

The ability to decouple the rows of nonorthogonal matrix optimization parameters can be preferred to the more restrictive orthogonal decoupling. The benefits of decoupling can be exhibited in terms of simplified algorithm design and improved optimization performance as demonstrated here. Additionally, decoupling enables density matching for ICA algorithms.

Appendix C

Canonical Correlation Analysis

In this appendix, we review CCA and the multiset extension denoted MCCA. The main point of the review is to allow the reader to understand CCA and MCCA in terms of the notation used in this work.

C.1 CCA

The idea of CCA is simply to find the pairs of weighting vectors, $\mathbf{w}_1^{[1]}$ and $\mathbf{w}_1^{[2]}$, that when applied to the two data sets, $\mathbf{x}^{[1]} \in \mathbb{R}^{N^{[1]}}$ and $\mathbf{x}^{[2]} \in \mathbb{R}^{N^{[2]}}$, maximize the correlation coefficient between the two resulting inner products, $y_1^{[1]} \triangleq \left(\mathbf{w}_1^{[1]}\right)^{\mathsf{T}} \mathbf{x}^{[1]}$ and $y_1^{[2]} \triangleq \left(\mathbf{w}_1^{[2]}\right)^{\mathsf{T}} \mathbf{x}^{[2]}$. We denote the first correlation coefficient as

$$\begin{aligned}\rho_1\left(\mathbf{w}_1^{[1]}, \mathbf{w}_1^{[2]}\right) &\triangleq \frac{E\left\{y_1^{[1]} y_1^{[2]}\right\}}{\sqrt{E\left\{\left(y_1^{[1]}\right)^2\right\} E\left\{\left(y_1^{[2]}\right)^2\right\}}}, \\ &= \frac{\left(\mathbf{w}_1^{[1]}\right)^{\mathsf{T}} \mathbf{R}_x^{[1,2]} \mathbf{w}_1^{[2]}}{\sqrt{\left(\mathbf{w}_1^{[1]}\right)^{\mathsf{T}} \mathbf{R}_x^{[1,1]} \mathbf{w}_1^{[1]} \left(\mathbf{w}_1^{[2]}\right)^{\mathsf{T}} \mathbf{R}_x^{[2,2]} \mathbf{w}_1^{[2]}}}\end{aligned}$$

and we want to maximize this quantity.

The derivation of CCA provided here follows closely that given in [45] with the notation modified to readily accommodate the MCCA description that follows.

The scaling of either or both weighting vectors does not effect the absolute value of the correlation coefficient. So we can fix the quantities in the denominator to be a constant, i.e.,

$$\left(\mathbf{w}_1^{[1]}\right)^{\mathsf{T}} \mathbf{R}_x^{[1,1]} \mathbf{w}_1^{[1]} = 1$$
$$\left(\mathbf{w}_1^{[2]}\right)^{\mathsf{T}} \mathbf{R}_x^{[2,2]} \mathbf{w}_1^{[2]} = 1,$$

and maximize the following Lagrangian,

$$\mathcal{L}\left(\boldsymbol{\lambda}, \mathbf{w}_1^{[1]}, \mathbf{w}_1^{[2]}\right) = \left(\mathbf{w}_1^{[1]}\right)^{\mathsf{T}} \mathbf{R}_x^{[1,2]} \mathbf{w}_1^{[2]} - \sum_{k=1}^{K=2} \frac{\lambda_1^{[k]}}{2} \left(\left(\mathbf{w}_1^{[k]}\right)^{\mathsf{T}} \mathbf{R}_x^{[k]} \mathbf{w}_1^{[k]} - 1\right).$$

To maximize, we take the derivative with respect to each weighting vector and set it equal to 0:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_1^{[1]}} = \mathbf{R}_x^{[1,2]} \mathbf{w}_1^{[2]} - \lambda_1^{[1]} \mathbf{R}_x^{[1]} \mathbf{w}_1^{[1]} = \mathbf{0} \tag{C.1}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_1^{[2]}} = \mathbf{R}_x^{[2,1]} \mathbf{w}_1^{[1]} - \lambda_1^{[2]} \mathbf{R}_x^{[2]} \mathbf{w}_1^{[2]} = \mathbf{0}. \tag{C.2}$$

We then pre-multiply each equation by $\left(\mathbf{w}_1^{[1]}\right)^{\mathsf{T}}$ and $\left(\mathbf{w}_1^{[2]}\right)^{\mathsf{T}}$ to exploit the constraints. The resulting two expressions imply that $\lambda_1^{[1]} = \lambda_1^{[2]}$. Letting $\lambda_1 = \lambda_1^{[1]} = \lambda_1^{[2]}$. Then assuming that $\mathbf{R}_x^{[2]}$ is invertible, from (C.2) we have $\mathbf{w}_1^{[2]} = \frac{\left(\mathbf{R}_x^{[2]}\right)^{-1} \mathbf{R}_x^{[1,2]} \mathbf{w}_1^{[1]}}{\lambda_1}$. This result is then substituted into (C.1) and rearranged,

$$\mathbf{R}_x^{[1,2]} \left(\mathbf{R}_x^{[2]}\right)^{-1} \mathbf{R}_x^{[2,1]} \mathbf{w}_1^{[1]} = \lambda_1^2 \mathbf{R}_x^{[1]} \mathbf{w}_1^{[1]}.$$

We can recognize this as a generalized eigenproblem, i.e., form is $\mathbf{Ax} = \lambda\mathbf{Bx}$. Of course this can be transformed into a standard symmetric eigenproblem if desired by whitening the datasets. Before summarizing the algorithm we need to state that the subscript 1 in the description of CCA is just for the first correlation coefficient and there are up to $N = \min\left(N^{[1]}, N^{[2]}\right)$ such coefficients. Hence, we can recognize that in the whitened domain the solutions are orthogonal but that the resulting demixing matrices, $\mathbf{W}^{[1]}$ and $\mathbf{W}^{[2]}$, are not necessarily orthogonal.

C.2 MCCA

We now proceed to an overview of multiset CCA, an extension of CCA for when more than two datasets are considered simultaneously. Several extensions of CCA were developed by several authors and these extensions were summarized and connected by Kettenring [53]. Here we simply provide descriptions of the objective functions possible in the MCCA framework and do not discuss the methods for optimizing these objective functions as this is not necessary for our purposes.

The five extensions of CCA, which we generally refer to as MCCA, are all based on assessing the correlation matrix of the estimated SCV (using the concepts of JBSS). Recalling, $y_n^{[k]} = \left(\mathbf{w}_n^{[k]}\right)^{\mathsf{T}} \mathbf{x}^{[k]}$ and $\mathbf{y}_n = \left[y_n^{[1]}, \ldots, y_n^{[K]}\right]$, then we wish to find the set of weights which optimize some objective function based on $\mathbf{\Sigma}_n \triangleq E\left\{\mathbf{y}_n\mathbf{y}_n^{\mathsf{T}}\right\}$. All but one use the eigenvalues of $\mathbf{\Sigma}_n$, which we denote $\boldsymbol{\lambda}_n$ and is assume it is sorted such that $\left[\boldsymbol{\lambda}_n\right]_1 \geq \ldots \geq \left[\boldsymbol{\lambda}_n\right]_K$. The five objective functions are enumerated and described in the following list:

1. SUMCOR: Maximize the sum of elements in $\mathbf{\Sigma}_n$, $\mathbf{1}^\mathsf{T}\mathbf{\Sigma}_n\mathbf{1}$
2. MAXVAR: Maximize the largest eigenvalue of $\mathbf{\Sigma}_n$, $[\boldsymbol{\lambda}_n]_1$
3. SSQCOR: Maximize the sum of $\mathbf{\Sigma}_n$ eigenvalues squared, $\displaystyle\sum_{k=1}^{K}([\boldsymbol{\lambda}_n]_k)^2$
4. MINVAR: Minimize the smallest eigenvalue of $\mathbf{\Sigma}_n$, $[\boldsymbol{\lambda}_n]_K$
5. GENVAR: Minimize the product of $\mathbf{\Sigma}_n$ eigenvalues, $\displaystyle\prod_{k=1}^{K}[\boldsymbol{\lambda}_n]_k$

Reasons for picking one of the MCCA objective functions versus another is given some discussion in [53], but it is safe to say that as presented there can be no theoretical basis for preferring one objective over any of the others. In fact, the basis for the objective function to use is typically empirical.

Appendix D

Online IVA

The algorithms presented in Chapter 3 are designed for batch processing. If the sources and/or the mixing matrix are nonstationary or if real-time processing is required, then online learning or stochastic algorithms for IVA are more appropriate.

In this appendix, we give an online algorithm for IVA-Gauss. We note that the IVA-Gauss gradient and Hessian are functions of the previous demixing estimates and the estimated data covariance matrix,

$$\hat{\mathbf{R}}_x = \frac{1}{T}\sum_{t=1}^{\mathsf{T}} \mathbf{x}\left(t\right)\mathbf{x}^{\mathsf{T}}\left(t\right) \approx E\left\{\mathbf{x}\mathbf{x}^{\mathsf{T}}\right\},$$

where $\mathbf{x}^{\mathsf{T}} = \left[\left(\mathbf{x}^{[1]}\right)^{\mathsf{T}}, \ldots, \left(\mathbf{x}^{[K]}\right)^{\mathsf{T}}\right]$. This is the essential feature that readily yields an online IVA-Gauss algorithm.

D.1 Algorithm

For online learning we wish to update the previously estimated demixing matrices using t samples with the information provided by the $t+1$ sample without storing all of the previous samples, i.e., fixed memory size.

We can denote the algorithm functionally, $\boldsymbol{\mathcal{W}}_{\text{IVA-G}} = \text{IVA-G}\left(\hat{\mathbf{R}}_x, \boldsymbol{\mathcal{W}}_0\right)$, where $\mathbf{W}_0$ is an optional initial guess for the demixing matrix. We assume that some finite memory limitation admits a batch solution for the IVA-Gauss problem. To process any additional samples we only need memory to hold the previously estimated demixing matrices, $\mathbf{W}^{[k]}(t)$ and the data covariance matrix, $\hat{\mathbf{R}}_x(t)$, where $1 \leq k \leq K$ and t is the number of samples used in the previous estimate. We have used the following notation: $\hat{\mathbf{R}}_x(t)$.

The new data to process may contain one or more samples. We can either batch the new data or process via a buffer. Either approach is captured here by setting the size of new samples to process in the next update to be p. Then we define a new data covariance matrix estimate of the additional data:

$$\hat{\mathbf{P}}_x = \frac{1}{p} \sum_{i=t+1}^{t+p} \mathbf{x}(i)\,\mathbf{x}^{\mathsf{T}}(i).$$

We can update the covariance matrix with the new data by:

$$\hat{\mathbf{R}}_x(t+p) = \lambda(t)\,\hat{\mathbf{R}}_x(t) + (1-\lambda(t))\,\hat{\mathbf{P}}_x,\ 0 < \lambda(t) < 1,$$

where $\lambda(t)$ is the memory factor, i.e., the larger it is the more heavily the past is weighted. If the IVA model is assumed stationary then all the samples should be weighted equally and therefore $\lambda(t) = t/(t+p)$. The memory factor can also be set to a constant value. This is useful when either the demixing matrices or source distributions are slowly varying, then the weighting of the more recent samples can be increased by decreasing λ to something small. Then then online algorithm is just,

$$\boldsymbol{\mathcal{W}}_{\text{IVA-G}}\left(t+p\right)=\text{IVA-G}\left(\hat{\mathbf{R}}_{x}\left(t+p\right),\boldsymbol{\mathcal{W}}_{\text{IVA-G}}\left(t\right)\right). \tag{D.1}$$

The last detail of the online algorithm to discuss is the initialization. For IVA-Gauss we need to have a data covariance matrix estimate that is supported with as many samples as the dimensions of the covariance matrix, i.e., $t > NK$. To attempt to have a solution at initialization when $t < NK$ we can diagonally load a covariance matrix estimate until we determine that the covariance matrix estimate is full rank. To avoid corrupting the data covariance matrix with the diagonal loading we can store an uncorrupted copy and use the loaded covariance matrix until the uncorrupted version is full rank.

D.2 Simulation Results

We evaluate the performance of the online algorithm for a few cases.

We start with N multivariate Gaussian sources with random covariance matrices. Each SCV sample is of dimension $K \times 1$, so that the kth dimension is the source used in the kth dataset, or $\mathbf{s}_n = \left[s_n^{[1]}, \ldots, s_n^{[K]}\right]^{\mathsf{T}} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_n\right) \in \mathbb{R}^K$. The mixing matrices, $\mathbf{A}^{[k]}, k = 1, \ldots, K$ have entries following the standard normal distribution. To examine performance we look at the joint ISI metric (see Section 2.7.2 for definition) and the number of iterations required per update. We expect both quantities to decrease as either the number of samples, T, or the number of datasets, K, increase.

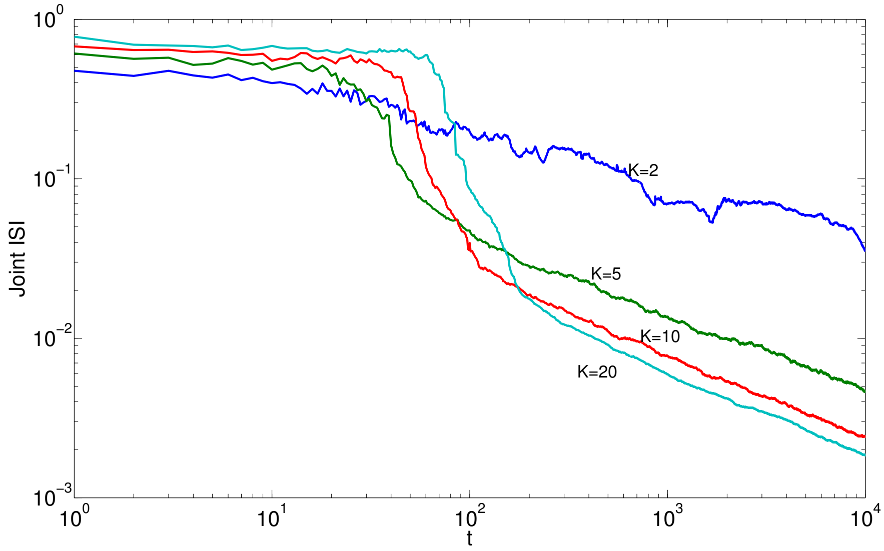


Figure D.1: Example of joint ISI for online IVA-G versus last sample processed, t. Using $N = 10$ MG sources with varying number of datasets.

For illustration, we have $N = 10$ multivariate Gaussian sources with $K = 2$, 5, 10, and 20 datasets. In this first experiment we increment with a single sample each time and use the sample size dependent memory factor, i.e., $\lambda(t) = t/(t + p)$ with $p = 1$. In Fig. D.1 we see that the joint ISI does decrease with increasing samples, t. Additionally it is apparent that for the algorithm to perform reliably the number of samples needs to be about NK. After the number of samples exceeds NK we have that the joint ISI is improved with more datasets, i.e., larger K. For the same simulation, we show in Fig. D.2 that for $K = 2$ the number of iterations is generally reducing with increasing t but appears to spike frequently. However, the number of iterations to update the demixing matrices reduces quickly once $t > NK$ for $K \geq 5$, and can often only require one or two iterations to incorporate the new data. This is the desired and expected behavior.

The results for the second experiment shown in Fig. D.3 and Fig. D.4, use

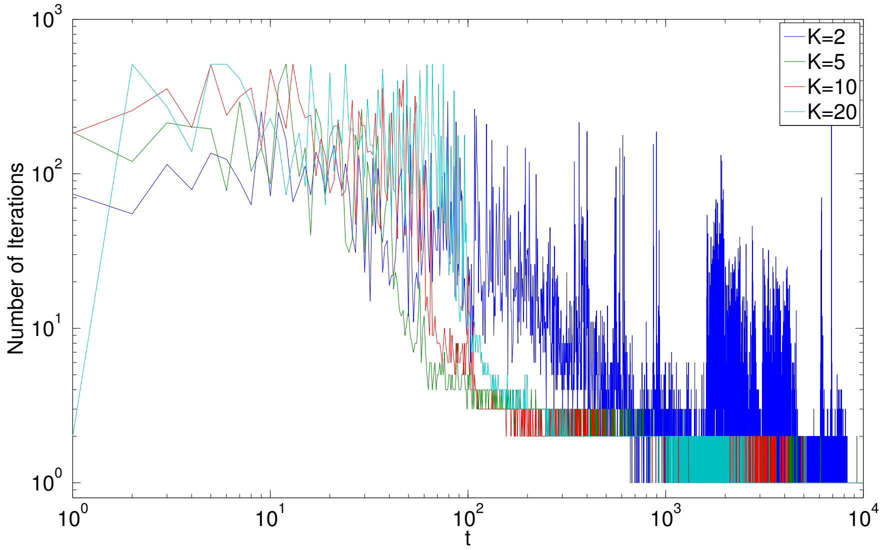


Figure D.2: Number of iterations for online IVA-G versus last sample processed, t. Using $N = 10$ MG sources with varying number of datasets.

$N = 10$ and $K = 20$ and consider various choices for the memory factor. For stationary datasets the optimum choice is $\lambda(t) = t/(t+1)$. We will also consider $\lambda = 0.999$, 0.99, and 0.95. As the memory factor decreases the joint ISI achieved at steady state increases and the number of iterations required per sample update at steady state increases. This loss in performance might be an acceptable trade-off when the data is nonstationary. In the experiment, we initialize the solution using the first $t = 100$ samples so as to speed up the experiment.

Lastly, we show an example to show the potential benefit of memory loading. Every 1000 samples a new set of random mixing matrices are generated. The results in Fig. D.5 show by choosing an appropriate fading factor a reasonable joint ISI can be achieved after each transition. This benefit seems to come at the cost of more iterations per update, Fig. D.6.

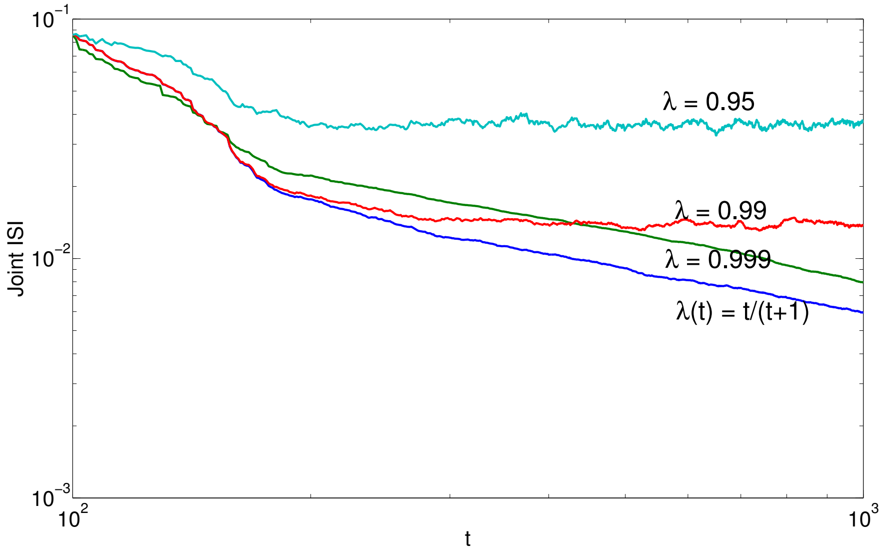


Figure D.3: Joint ISI for online IVA-Gauss versus last sample processed, t, for $N = 10$ MG sources and $K = 20$ datasets for different memory factors, λ.

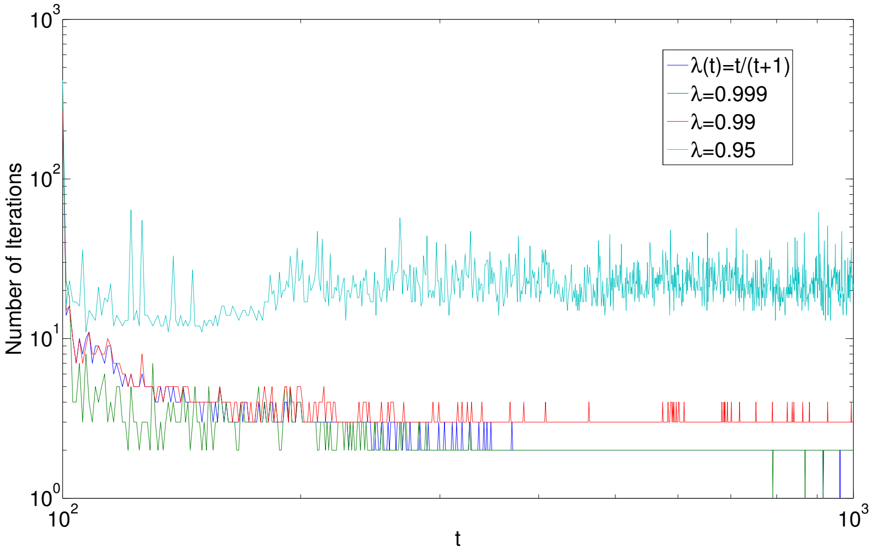


Figure D.4: Number of iterations for online IVA-G versus last sample processed, t, for $N = 10$ MG sources and $K = 20$ datasets for different memory factors, λ.

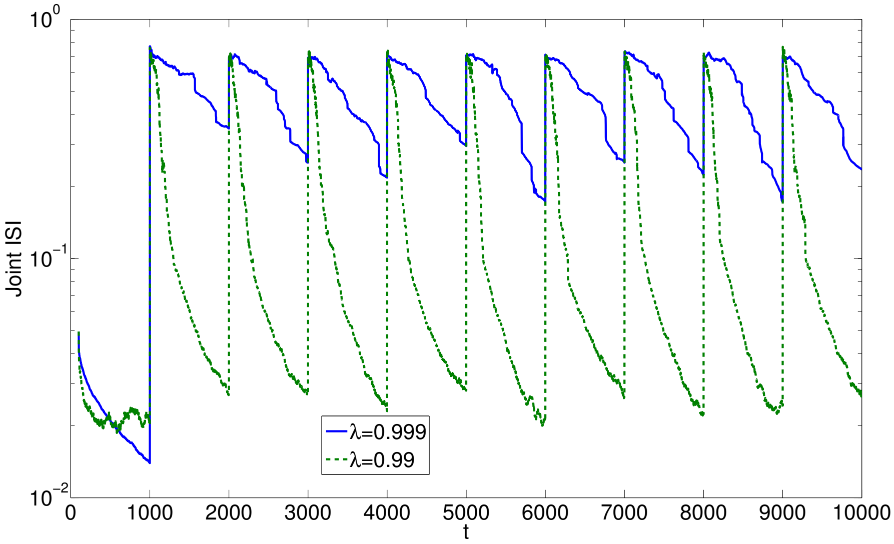


Figure D.5: Example of joint ISI for online IVA-Gauss versus last sample processed, t with random mixing matrices applied every 1000th sample. Using $N = 10$ MG sources and $K = 10$ datasets for different memory factors, λ.

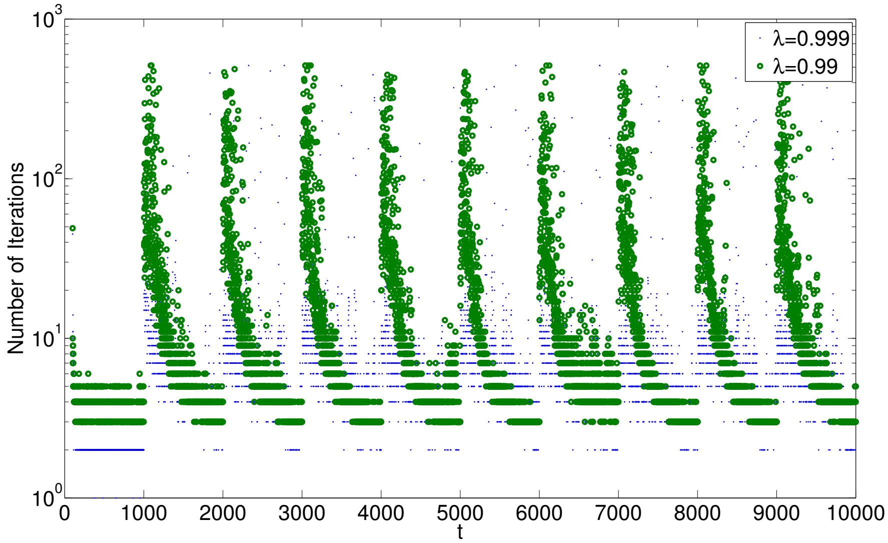


Figure D.6: Number of iterations for online IVA-G versus last sample processed, t with random mixing matrices applied every 1000th sample. Using $N = 10$ MG sources and $K = 10$ datasets for different memory factors, λ.

Appendix E

Kotz Sample Generation

In this appendix, we briefly describe the procedure used to generate samples from the general multivariate Kotz distribution. The basis of the procedure is given in [38, 89]. Here we succinctly describe the relevant background material, define the procedure in general, and provide implementation details for completeness. After which we give some examples of the empirical distribution compared to the desired pdf to demonstrate the effectiveness of the proposed procedure.

Intuitively, samples from the centralized (zero mean) spherically symmetric set of distributions can quite naturally be described in total by some function of a scalar variable. It readily follows that samples from spherical distributions can be generated using $\mathbf{x} = \sqrt{r^2}\mathbf{u} \in \mathbb{R}^K$, where r is independent of $\mathbf{u}$. The $\mathbf{u}$ is the uniform base of the spherical distribution, i.e., $\mathbf{u} = \mathbf{z}/\|\mathbf{z}\|$ with $\mathbf{z} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)$. The square of the generating variate, $v = r^2$, has the following pdf,

$$f\left(v\right) = \frac{\pi^{K/2}}{\Gamma\left(K/2\right)} v^{K/2-1} g\left(v\right), \quad 0 \leq v \leq \infty \tag{E.1}$$

where $g\left(\cdot\right)$ is density generator function, [38, Eq. 2.21].

As one would expect, the samples from the family of elliptically symmetric distributions can be generated with $\mathbf{x} = \boldsymbol{\mu} + \sqrt{r^2}\boldsymbol{\Sigma}^{1/2}\mathbf{u}$, where $\text{rank}\left(\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^\mathsf{T}\right) = K$. For the Kotz distribution the density generator function is given in [38, Eq.

3.15] as

$$g(v) = Cv^{\eta-1} \exp\left(-\kappa v^{\beta}\right). \tag{E.2}$$

Thus, all that remains is to generate samples from the Kotz distribution is to generate samples of v that are in accord with the pdf of (E.1) using (E.2).

We have chosen to generate samples of v via the inverse (or percentile) transformation method [95] with the cumulative distribution function (cdf), $F(v)$, calculated via numerical integration of (E.1) using (E.2). The only difficulty remaining is to determine the maximum value of the random variable, $V_{\max}$, to evaluate the cdf over. We choose to specify the max value of $F(v)$, $F_{\max}$. Then the value of $V_{\max}$ is increased by some factor until the numeric integration results in $F(V_{\max}) > F_{\max}$. Then via numerical integration, we have $F(v_m)$ for a fixed sampling, $v_m = m\Delta, m = 0, 1, \ldots, M$, where $\Delta = V_{\max}/M$. The percentile transformation method is completed using linear interpolation of the $F^{-1}(v_m)$ with the random samples $u \sim \mathcal{U}(0, F_{\max})$.

The Matlab code implementing the above procedure is provided at `http://mlsp.umbc.edu/resources.html`. The relevant code are given below.

Next, we evaluate the efficacy of the approach for various members of the two-dimensional Kotz distribution via visualization of the pdf versus the empirical histogram. For the experiments the base case will be the zero-mean two-dimensional Gaussian with correlation coefficient 0.25, i.e., $K = 2$, $\beta = 1$, $[P]_{1,1}$, $[P]_{2,2} = 1$, $[P]_{1,2} = 0.25$, $\eta = 1$, $\kappa = 0.5$, $\boldsymbol{\mu} = \mathbf{0}$. We generate one million samples for each distribution evaluated and bin results into using 100 equally spaced and size bins in both dimensions, i.e., 10000 bins.

First we vary the shape parameter, $\beta = 0.5, 1, 2, 4$ in Fig. E.1, and the hole parameter, $\eta = 0.5, 1, 2, 4$ in Fig. E.2. The excellent correlation between the histograms and pdfs indicates the Kotz generation procedure outline above is performing as desired. The extremes of the Kotz family, e.g., $\beta \to 0$, can cause numeric difficulties and should be avoided.

```
function [Z]=randmv_kotz(T,dim,s,cov_mx,N,kurt_param,mu)
% [Z]=randmv_kotz(T,dim,s,cov_mx,N,kurt_param,mu)
%
% Input:
%  T -- number of samples
%  dim -- dimension of multivariate distribution (p)
%  s -- shape-like dispersion parameter (s)
%  cov_mx -- pos. def. covariance matrix
%         = scalar times dispersion matrix, [dim x dim]
%  N -- denominator exponent, regulates the volume of hole,
%       where (N > (2-p)/2)
%  kurt_param -- kurtosis parameter (r)
%  mu -- mean vector, [dim x 1]
%
% Output:
%  Z -- [dim x T] random samples of specified Kotz-type distribution
maxF_desired=0.999;
v_max=10;

p=dim;
r=kurt_param;
covScal=r^(-1/s)/p*gamma((2*N+p)/(2*s)) / gamma((2*N+p-2)/(2*s));
disp_mx=1/covScal*cov_mx;

fh_pdf_v=@(v) (pdf_v(p,N,s,r,v));
maxF=quadl(fh_pdf_v,0,v_max);

while maxF<maxF_desired
   v_max=v_max*3;
   maxF=quadl(fh_pdf_v,0,v_max);
end

v=linspace(0,v_max,1e4);
Fv=NaN(size(v)); Fv(1)=0;
for iv=2:length(v)
   dFv=quadl(fh_pdf_v,v(iv-1),v(iv));
   if dFv > eps
      Fv(iv)=Fv(iv-1)+dFv;
   end
end
iNaN=find(isnan(Fv)); v(iNaN)=[]; Fv(iNaN)=[];

u_samp=maxF_desired*rand(1,T);
v_samp=interp1(Fv,v,u_samp);

U=randn(p,T);
Umag=sqrt(sum(U.*U));
U=bsxfun(@times,1./Umag,U);
Z=bsxfun(@plus,mu,bsxfun(@times,sqrt(v_samp),sqrtm(disp_mx)*U));
```

```
function [f_v,v]=pdf_v(p,N,s,r,v)
% [f_v,v]=pdf_v(p,N,s,r,v)
%
% Evaluate Eq. 3 (using Eqs. 2 and 4) of Nadarajah 2003 paper.
%
% p=dimension, N>(2-p)/2, s>0, r>0

if nargin<5
   v=linspace(0,40,10000);
end
arg=(2*N+p-2)/(2*s);
Cp=s*gamma(p/2) / (pi^(p/2) * gamma(arg)) * r^arg;
g_v=Cp*v.^(N-1).*exp(-r*v.^s);
f_v=pi^(p/2)/gamma(p/2) * v.^(p/2-1) .* g_v;
```

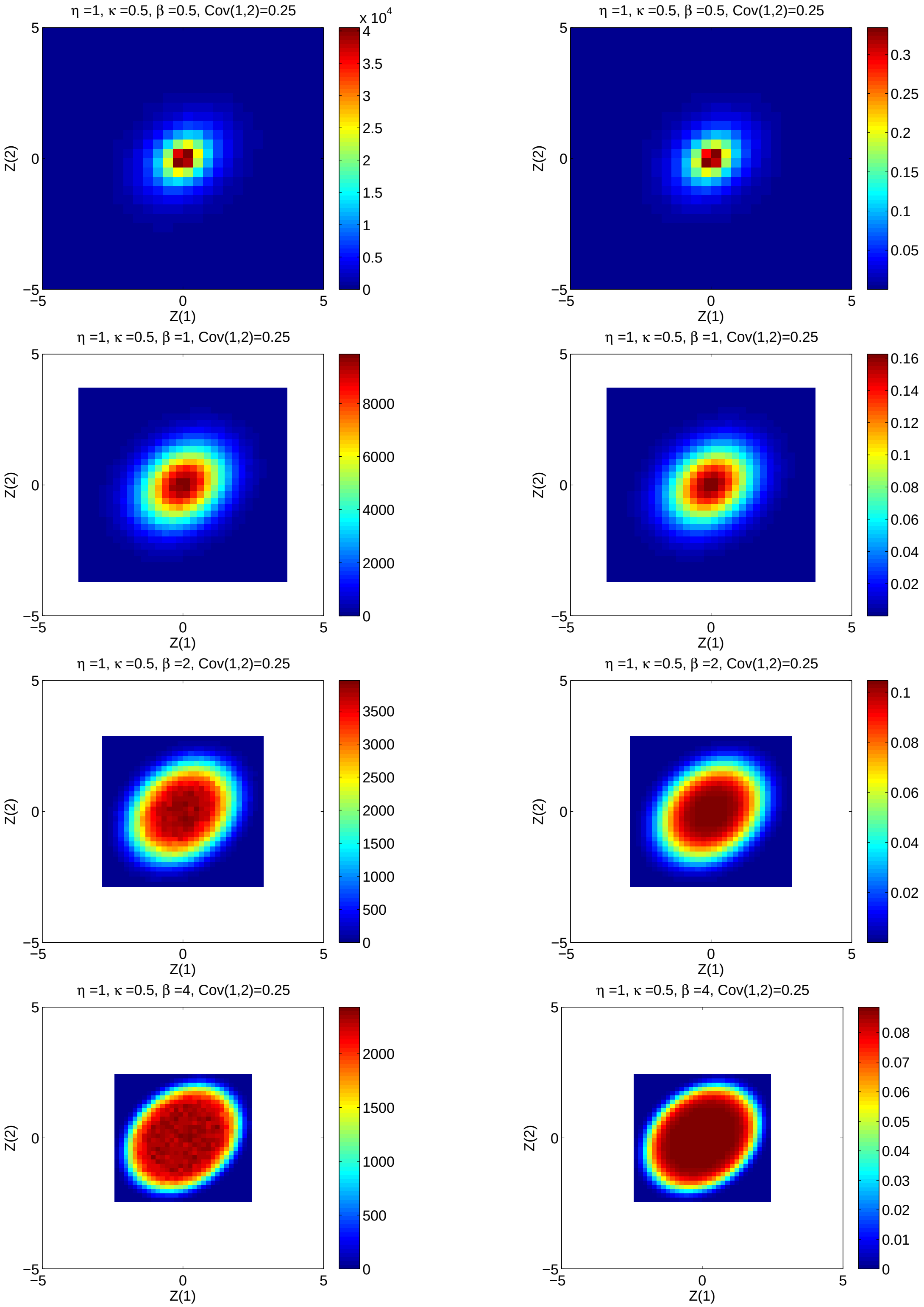


Figure E.1: Empirical histograms and associated pdfs with varying shape parameters.

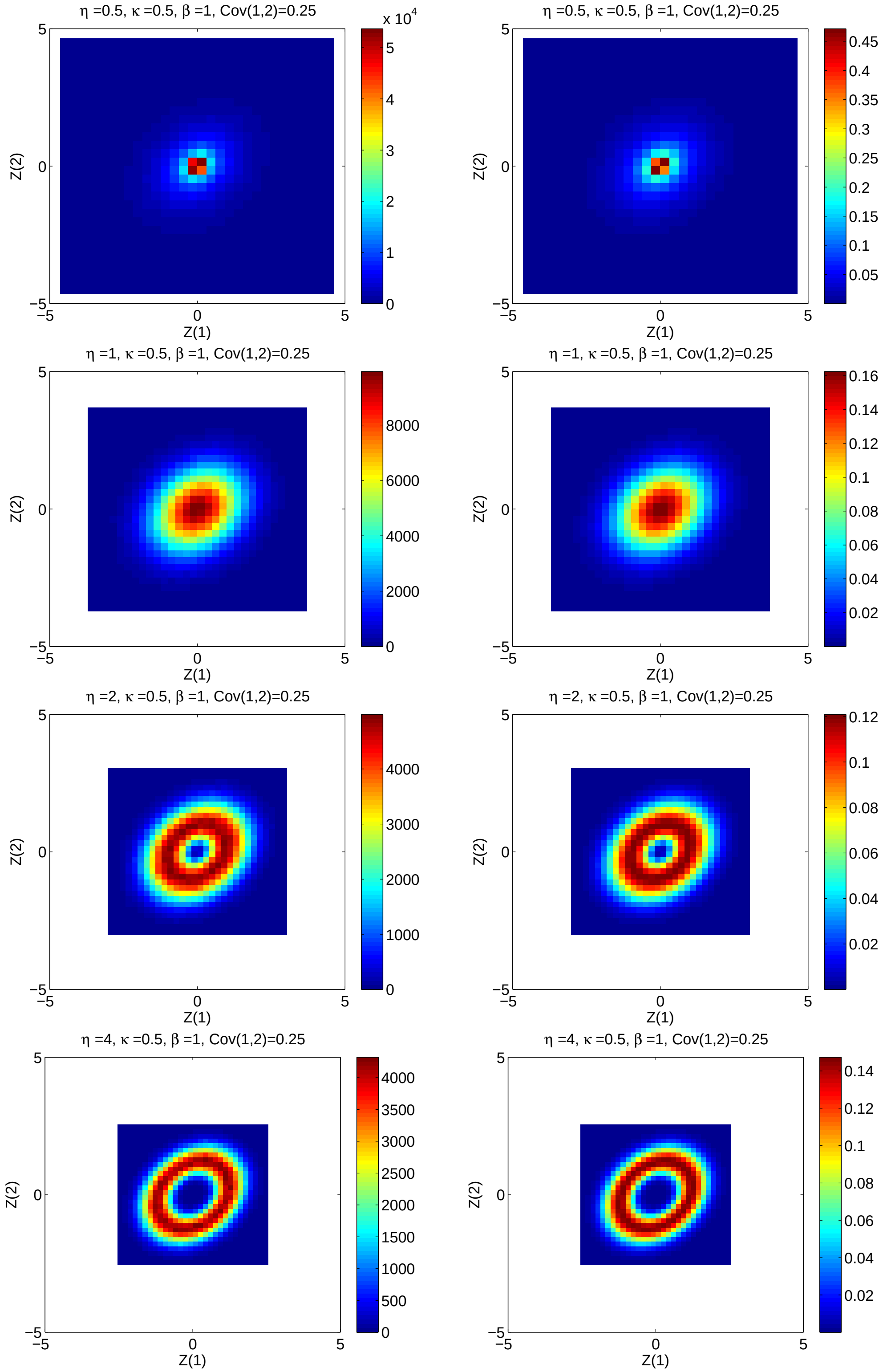


Figure E.2: Empirical histograms and associated pdfs with varying hole parameters.

Bibliography

[1] T. Adalı, M. Anderson, and G. Fu. IVA and ICA: Use of diversity in independent decompositions. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 61–65, Aug. 2012.

[2] T. Adalı and S. Haykin, editors. *Adaptive Signal Processing: Next Generation Solutions.* Wiley-IEEE Press, 2010.

[3] T. Adalı, P. J. Schreier, and L. L. Scharf. Complex-valued signal processing: The proper way to deal with impropriety. *IEEE Trans. Signal Process.*, 59(11):5101–5125, 2011.

[4] B. Afsari. Sensitivity analysis for the problem of matrix joint diagonalization. *SIAM J. Matrix Anal. Appl.*, 30(3):1148–1171, Sept. 2008.

[5] S. Akaho. A kernel method for canonical correlation analysis. In *International Meeting on Psychometric Society (IMPS2001)*, 2001.

[6] S. Amari, A. Cichocki, and H. H. Yang. Advances in neural information processing systems. In *Advances in Neural Information Processing Systems*, volume 8, pages 757–763, Cambridge, MA, 1996. MIT Press.

[7] M. Anderson and T. Adalı. A general approach for robustification of ICA algorithms. In *Latent Variable Analysis and Signal Separation*, volume 6365 of *Lecture Notes in Computer Science*, pages 295–302. Springer Berlin / Heidelberg, 2010.

[8] M. Anderson, T. Adalı, and X.-L. Li. Joint blind source separation of multivariate Gaussian sources: Algorithms and performance analysis. *IEEE Trans. Signal Process.*, 60(4):1672–1683, Apr. 2012.

[9] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adalı. Independent vector analysis: Identification conditions and performance bounds. *IEEE Trans. Signal Process.* Submitted.

[10] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adalı. Independent vector analysis, the Kotz distribution, and performance bounds. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2013. Accepted.

[11] M. Anderson, X.-L. Li, and T. Adalı. Nonorthogonal independent vector analysis using multivariate Gaussian model. In *Latent Variable Analysis and Signal Separation*, volume 6365 of *Lecture Notes in Computer Science*, pages 354–361. Springer Berlin / Heidelberg, 2010.

[12] M. Anderson, X.-L. Li, and T. Adalı. Complex-valued independent vector analysis: Application to multivariate Gaussian model. *Signal Process.*, (8):1821–1831, 2012. Latent Variable Analysis and Signal Separation.

[13] M. Anderson, X.-L. Li, P. Rodriguez, and T. Adalı. An effective decoupling method for matrix optimization and its application to the ICA problem. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pages 1885–1888, Mar. 2012.

[14] G. Aulogiaris and K. Zografos. A maximum entropy characterization of symmetric Kotz type and Burr multivariate distributions. *TEST*, 13:65–83, 2004. 10.1007/BF02603001.

[15] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, Mar. 2003.

[16] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[17] A. Belouchrani, K. Abed-Meraim, J. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Trans. Signal Process.*, 45(2):434–444, 1997.

[18] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. Second-order blind separation of temporally correlated sources. In *Proc. Int. Conf. Digital Signal Processing*, pages 346–351. Citeseer, 1993.

[19] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, 2nd edition, Sept. 1999.

[20] D. Brandwood. A complex gradient operator and its application in adaptive array theory. *IEE Proceedings. H, Microwaves, Optics and Antennas*, 130(1):11–16, Feb. 1983.

[21] V. D. Calhoun and T. Adalı. Feature-based fusion of medical imaging data. *IEEE Trans. Inf. Technol. Biomed.*, 13(5):711–720, Sept. 2009.

[22] J.-F. Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal Process. Lett.*, 4(4):112–114, Apr. 1997.

[23] J.-F. Cardoso. Multidimensional independent component analysis. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, volume 4, pages 1941–1944, May 1998.

[24] J.-F. Cardoso and B. H. Laheld. Equivariant adaptive source separation. *IEEE Trans. Signal Process.*, 44(12):3017–3030, Dec. 1996.

[25] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. *Radar and Signal Processing, IEE Proceedings F*, 140(6):362–370, Dec. 1993.

[26] S. Choi, A. Cichocki, L. Zhang, and S. Amari. Approximate maximum likelihood source separation using the natural gradient. In *Wireless Communications, 2001. (SPAWC '01). 2001 IEEE Third Workshop on Signal Processing Advances in*, pages 235–238, 2001.

[27] P. Comon. Independent component analysis, a new concept? *Signal Process.*, 36(3):287–314, 1994.

[28] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications.* Academic Press, 1st edition, 2010.

[29] N. M. Correa, T. Adalı, Y.-O. Li, and V. D. Calhoun. Canonical correlation analysis for data fusion and group inferences. *IEEE Signal Process. Mag.*, 27(4):39–50, 2010.

[30] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, 2006.

[31] M. Davies. Audio source separation. *Math. Signal Process. V*, pages 57–68, 2002.

[32] J. de Leeuw. The Gifi-system of nonlinear multivariate analysis. *Data Analysis and Informatics*, III:415–424, 1984.

[33] J. T. Dea, M. Anderson, E. Allen, V. D. Calhoun, and T. Adalı. IVA for multi-subject FMRI analysis: A comparative study using a new simulation toolbox. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sept. 2011.

[34] S. S. Dragomir, R. P. Agarwal, and N. S. Barnett. Inequalities for beta and gamma functions via some classical and new integral inequalities. *Journal of Inequalities and Applications*, 5:103–165, 2000.

[35] T. Eltoft, T. Kim, and T.-W. Lee. On the multivariate Laplace distribution. *IEEE Signal Process. Lett.*, 13(5):300–303, May 2006.

[36] J. Eriksson and V. Koivunen. Complex-valued ICA using second order statisitcs. In *Proceedings of the 2004 14th IEEE Signal Processing Society Workshop*, pages 183–192. Machine Learning for Signal Processing, 2004.

[37] J. Eriksson and V. Koivunen. Complex random vectors and ica models: identifiability, uniqueness, and separability. *IEEE Trans. Inf. Theory*, 52(3):1017–1029, 2006.

[38] K.-T. Fang, S. Kotz, and K. W. Ng. *Symmetric Multivariate and Related Distributions.* Chapman and Hall, 1990.

[39] O. Friman, M. Borga, P. Lundberg, and H. Knutsson. Exploratory fMRI analysis by autocorrelation maximization. *NeuroImage*, 16(2):454–464, 2002.

[40] G. Fu, H. Li, M. Anderson, and T. Adalı. Order detection for dependent samples using entropy rate. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pages 2161–2164, 2012.

[41] G.-S. Fu, M. Anderson, R. Phlypo, and T. Adalı. Algorithms for Markovian source separation by entropy rate minimization. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2013. Accepted.

[42] A. Gifi. *Nonlinear multivariate analysis.* Wiley, New York, 1990.

[43] E. Gomez, M. Gomez-Villegas, and J. M. Marin. A multivariate generalization of the power exponential family of distributions. *Communications in Statistics-Theory and Methods*, 27(3):589–600, 1998.

[44] J. Gurland. An inequality satisfied by the gamma function. *Scandinavian Actuarial Journal*, 29:171–172, 1956.

[45] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis; an overview with application to learning methods. Technical report, Department of Computer Science, Royal Holloway, University of London, May 2003.

[46] A. Hiroe. Solution of permutation problem in frequency domain ICA, using multivariate probability density functions. In J. Rosca, D. Erdogmus, J. C. Príncipe, and S. Haykin, editors, *Independent Component Analysis and Blind Signal Separation*, volume 3889 of *Lecture Notes in Computer Science*, pages 601–608. Springer Berlin Heidelberg, 2006.

[47] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge, 1985.

[48] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, Cambridge, 1991.

[49] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[50] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626–634, May 1999.

[51] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* Wiley-Interscience, 2001.

[52] A. Hyvärinen, E. Oja, P. Hoyer, and J. Hurri. Image feature extraction by sparse coding and independent component analysis. In *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, volume 2, pages 1268–1273, Aug. 1998.

[53] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

[54] K. Kim and G. Shevlyakov. Why Gaussianity? *Signal Processing Magazine, IEEE*, 25(2):102–113, 2008.

[55] T. Kim. *Independent Vector Analysis.* PhD thesis, Department of BioSystems Korea Advanced Institute of Science and Technology, 2006.

[56] T. Kim. Real-time independent vector analysis for convolutive blind source separation. *Circuits and Systems I: Regular Papers, IEEE Transactions on*, 57(7):1431–1438, July 2010.

[57] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee. Blind source separation exploiting higher-order frequency dependencies. *IEEE Trans. Audio Speech Lang. Process.*, 15(1):70–79, Jan. 2007.

[58] T. Kim, T. Eltoft, and T.-W. Lee. Independent vector analysis: an extension of ICA to multivariate components. In *Independent Component Analysis and Blind Signal Separation*, volume 3889 of *Lecture Notes in Computer Science*, pages 165–172. Springer Berlin / Heidelberg, 2006.

[59] T. Kim, I. Lee, and T.-W. Lee. Independent vector analysis: Definition and algorithms. In *Proc. of 40th Asilomar Conference on Signals, Systems, and Computers*, pages 1393–1396, Oct. 2006.

[60] Z. Koldovský, P. Tichavský, and E. Oja. Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér-Rao lower bound. *IEEE Trans. Neural Netw.*, 17(5), Sept. 2006.

[61] S. Kotz. Multivariate distributions at a cross-road. *Statistical Distributions in Scientific Work*, 1975.

[62] D. Lahat, J.-F. Cardoso, and H. Messer. Second-order multidimensional ica: Performance analysis. *IEEE Trans. Signal Process.*, 60(9):4598–4610, Sept. 2012.

[63] P. Lavergne. A Cauchy-Schwarz inequality for expectation of matrices. Discussion papers, Department of Economics, Simon Fraser University, 2008.

[64] I. Lee, T. Kim, and T.-W. Lee. Fast fixed-point independent vector analysis algorithms for convolutive blind source separation. *Signal Process.*, 87(8):1859–1871, 2007. Independent Component Analysis and Blind Source Separation.

[65] J.-H. Lee, T.-W. Lee, F. A. Jolesz, and S.-S. Yoo. Independent vector analysis (IVA): Multivariate approach for fMRI group study. *NeuroImage*, 40(1):86–109, 2008.

[66] H. Li and T. Adalı. Complex-valued adaptive signal processing using nonlinear functions. *EURASIP J. Adv. Signal Process.*, pages 122:1–122:9, January 2008.

[67] H. Li and T. Adalı. Algorithms for complex ML ICA and their stability analysis using Wirtinger calculus. *IEEE Trans. Signal Process.*, 58(12):6156–6167, Dec. 2010.

[68] H. Li, M. Anderson, and T. Adalı. Kernel least mean pth power adaptive algorithm for nonlinear adaptive filtering. *IEEE Trans. Neural Netw. Learn. Syst.* Submitted.

[69] H. Li, X.-L. Li, M. Anderson, and T. Adalı. A class of adaptive algorithms based on entropy estimation achieving CRLB for linear non-Gaussian filtering. *IEEE Trans. Signal Process.*, 60(4):2049–2055, Apr. 2012.

[70] X.-L. Li, , T. Adalı, and M. Anderson. Detection of circular and noncircular signals in the presence of circular white Gaussian noise. In *Proc. of 44th Asilomar Conference on Signals, Systems, and Computers*, 2010.

[71] X.-L. Li and T. Adalı. A novel entropy estimator and its application to ICA. *Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on*, pages 1–6, Sept. 2009.

[72] X.-L. Li and T. Adalı. Blind spatiotemporal separation of second and/or higher-order correlated sources by entropy rate minimization. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pages 1934–1937, Mar. 2010.

[73] X.-L. Li and T. Adalı. Complex independent component analysis by entropy bound minimization. *Circuits and Systems I: Regular Papers, IEEE Transactions on*, 57(7):1417–1430, July 2010.

[74] X.-L. Li and T. Adalı. Independent component analysis by entropy bound minimization. *IEEE Trans. Signal Process.*, 58(10):5151–5164, Oct. 2010.

[75] X.-L. Li, T. Adalı, and M. Anderson. Joint blind source separation by generalized joint diagonalization of cumulant matrices. *Signal Process.*, 91(10):2314–2322, Oct. 2011.

[76] X.-L. Li, M. Anderson, and T. Adalı. Second and higher-order correlation analysis of multiset multidimensional variables by joint diagonalization. In *Latent Variable Analysis and Signal Separation*, volume 6365 of *Lecture Notes in Computer Science*, pages 197–204. Springer Berlin / Heidelberg, 2010.

[77] X.-L. Li, M. Anderson, and T. Adalı. Noncircular principal component analysis and its application to model selection. *IEEE Trans. Signal Process.*, 59(10):4516–4528, Oct. 2011.

[78] X.-L. Li and X.-D. Zhang. Nonorthogonal joint diagonalization free of degenerate solution. *IEEE Trans. Signal Process.*, 55(5):1803–1814, May 2007.

[79] Y.-O. Li, T. Adalı, W. Wang, and V. D. Calhoun. Joint blind source separation by multiset canonical correlation analysis. *IEEE Trans. Signal Process.*, 57(10):3918–3929, Oct. 2009.

[80] B. Loesch, F. Nesta, and B. Yang. On the robustness of the multidimensional state coherence transform for solving the permutation problem of frequency-domain ICA. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pages 225–228, Mar. 2010.

[81] B. Loesch and B. Yang. Cramer-rao bound for circular and noncircular complex independent component analysis. *IEEE Trans. Signal Process.*, 61(2):365–379, Jan. 2013.

[82] S. Makeig, A. J. Bell, T. ping Jung, and T. J. Sejnowski. Independent component analysis of electroencephalographic data. In *in Advances in Neural Information Processing Systems*, pages 145–151. MIT Press, 1996.

[83] MATLAB. *version 7.10.0 (R2010a).* The MathWorks Inc., Natick, Massachusetts, 2010.

[84] M. J. Mckeown, S. Makeig, G. G. Brown, T.-P. Jung, S. S. Kindermann, R. S. Kindermann, A. J. Bell, and T. J. Sejnowski. Analysis of fmri data by blind separation into independent spatial components. *Human Brain Mapping*, 6(3):160–188, 1998.

[85] T. Melzer, M. Reiter, and H. Bischof. Nonlinear feature extraction using generalized canonical correlation analysis. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Artificial Neural Networks - ICANN 2001*, volume 2130 of *Lecture Notes in Computer Science*, pages 353–360. Springer Berlin Heidelberg, 2001.

[86] M. Merkle. Gurland's ratio for the gamma function. *Computers & Mathematics with Applications*, 49(2–3):389–406, 2005.

[87] E. Moreau and O. Macchi. A one stage self-adaptive algorithm for source separation. *Acoustics, Speech, and Signal Processing*, 3:49–52, 1994.

[88] R. J. Muirhead. *Aspects of Multivariate Statistical Theory.* Wiley-In, 2005.

[89] S. Nadarajah. The Kotz-type distribution with applications. *Statistics*, pages 341–358, 2003.

[90] A. Nielsen. Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE Trans. Image Process.*, 11(3):293–305, Mar. 2002.

[91] M. Novey and T. Adalı. On extending the complex fastica algorithm to noncircular sources. *IEEE Trans. Signal Process.*, 56(5):2148–2154, May 2008.

[92] M. Novey, T. Adalı, and A. Roy. A complex generalized Gaussian distribution; characterization, generation, and estimation. *IEEE Trans. Signal Process.*, 58(3):1427–1433, Mar. 2010.

[93] E. Ollila, K. Hyon-Jung, and V. Koivunen. Compact Cramér-Rao bound expression for independent component analysis. *IEEE Trans. Signal Process.*, 56(4):1421–1428, Apr. 2008.

[94] E. Ollila, D. E. Tyler, V. Koivunen, and H. Poor. Complex elliptically symmetric distributions: Survey, new results and applications. *IEEE Trans. Signal Process.*, 60(11):5597–5625, 2012.

[95] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, 1991.

[96] K. B. Petersen and M. S. Pedersen. The matrix cookbook, Nov. 2008.

[97] R. Phlypo. Jacobi iterations for canonical dependence analysis. *Signal Processing*, 93(1):185–197, 2013.

[98] B. Picinbono. Second-order complex random vectors and normal distributions. *IEEE Trans. Signal Process.*, 44(10):2637–2640, Oct. 1996.

[99] P. Rodriguez, M. Anderson, X.-L. Li, and T. Adalı. General non-orthogonal constrained ICA. *IEEE Trans. Signal Process.* Submitted.

[100] I. Rustandi, M. A. Just, and T. M. Mitchell. Integrating multiple-study multiple-subject fMRI datasets using canonical correlation analysis. In *Proc. MICCAI 2009 Workshop: Statist. Model. Detection Issues in Intra- and Inter-Subject Functional MRI Data Anal.*, Sept. 2009.

[101] G. E. Sarty. *Computing Brain Activity Maps from fMRI Time-Series Images.* Cambridge University Press, 2007.

[102] P. J. Schreier. A unifying discussion of correlation analysis for complex random vectors. *IEEE Trans. Signal Process.*, 56(4):1327–1336, Apr. 2008.

[103] P. J. Schreier and L. L. Scharf. *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals.* Cambridge University Press, 2010.

[104] G. A. Seber. *A matrix handbook for statisticians.* Wiley-Int, 2008.

[105] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22(1–3):21–34, 1998.

[106] C. M. Teixeira, M. Anderson, and C. H. Jaffe. Training data non-stationarity mitigation for STAP. In *MSS Tri-Service Radar Symposium*, June 2006.

[107] J. M. ten Berge. *Least Squares Optimization in Multivariate Analysis.* DSWO Press, 2005.

[108] P. Tichavský and Z. Koldovský. Optimal pairing of signal components separated by blind techniques. *IEEE Signal Process. Lett.*, 11(2):119–122, feb. 2004.

[109] P. Tichavský, Z. Koldovský, and E. Oja. Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis. *IEEE Trans. Signal Process.*, 54(4):1189–1203, Apr. 2006.

[110] K. Todros and A. O. Hero. On measure transformed canonical correlation analysis. *IEEE Trans. Signal Process.*, 60(9):4570–4585, Sept. 2012.

[111] L. Tong, V. Soon, Y. Inouye, Y. Huang, and R. Liu. Waveform-preserving blind estimation of multiple sources. In *Decision and Control, 1991., Proceedings of the 30th IEEE Conference on*, volume 3, pages 2388–2393, Dec. 1991.

[112] L. Tong, R. wen Liu, V. C. Soon, and Y.-F. Huang. Indeterminacy and identifiability of blind identification. *Circuits and Systems, IEEE Transactions on*, 38(5):499–509, May 1991.

[113] 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 Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2011.

[114] 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)*, Beijing, China, Sept. 2011.

[115] J. Vía and I. Santamaría. Adaptive blind equalization of SIMO systems based on canonical correlation analysis. In *Signal Processing Advances in Wireless Communications, 2005 IEEE 6th Workshop on*, pages 318–322, June 2005.

[116] J. Vía, I. Santamaría, and J. Pérez. A learning algorithm for adaptive canonical correlation analysis of several data sets. volume 20, pages 139–152, Oxford, UK, 2007. Elsevier Science Ltd.

[117] W. Wirtinger. Zur formalen Theorie der Funktionen von mehr komplexen Veränderlichen. *Mathematische Annalen*, 97:357–375, 1927.

[118] A. Yeredor. Blind separation of Gaussian sources via second-order statistics with asymptotically optimal weighting. *IEEE Signal Process. Lett.*, 7(7):197–200, July 2000.

[119] A. Yeredor. Blind separation of Gaussian sources with general covariance structures: Bounds and optimal estimation. *IEEE Trans. Signal Process.*, 58(10):5057–5068, Oct. 2010.

[120] X. Yin. Canonical correlation analysis based on information theory. *Journal of Multivariate Analysis*, 91(2):161–176, 2004.

[121] F. W. Young, J. De Leeuw, and Y. Takane. Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 41:505–529, 1976.

[122] H. Zhang, L. Li, and W. Li. Independent vector analysis for convolutive blind noncircular source separation. *Signal Processing*, 92(9):2275–2283, 2012.