UNIVERSITY OF CALIFORNIA, SAN DIEGO

Support Recovery and Dictionary Learning for Uncorrelated EEG Sources

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Signal and Image Processing)

by

Ozgur Yigit Balkan

Committee in charge:

Professor Kenneth Kreutz-Delgado, Chair
Professor Sanjoy Dasgupta
Professor Virginia de Sa
Professor Bhaskar Rao
Professor Nuno Vasconcelos

2015

The dissertation of Ozgur Yigit Balkan is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2015

DEDICATION

To my mother Nilgün,

my father Firuz,

and my sister Topik.

TABLE OF CONTENTS

Signature Page . . . iii

Dedication . . . iv

Table of Contents . . . v

List of Figures . . . viii

List of Tables . . . xi

Acknowledgements . . . xii

Vita . . . xv

Abstract of the Dissertation . . . xvi

Chapter 1 Introduction . . . 1
1.1 Sparse Signal Recovery . . . 1
1.1.1 Single Measurement Vector (SMV) Model . . . 1
1.1.2 Multiple Measurement Vector (MMV) Model . . . 3
1.2 Dictionary Learning . . . 4
1.3 EEG Source Identification and Independent Component Analysis . . . 5
1.3.1 EEG Non-stationarity . . . 8
1.4 Our Contributions . . . 9

Chapter 2 Support Recovery of Uncorrelated Sources . . . 10
2.1 Introduction . . . 11
2.2 M-SBL . . . 12
2.3 Analysis . . . 14
2.3.1 Local Minima . . . 14
2.3.2 Exact Support Recovery Conditions . . . 15
2.3.3 Remarks . . . 18
2.4 Experiments . . . 18
2.4.1 Theorem Validation . . . 18
2.4.2 Performance Comparison . . . 19
2.4.3 EEG source localization . . . 23
2.5 Conclusion . . . 25
2.6 Acknowledgements . . . 25

Chapter 3 Covariance-domain Dictionary Learning for EEG Source Identification . . . 26
3.1 Introduction . . . 27
3.2 Related Work . . . 28
3.3 Covariance-Domain Dictionary Learning (Cov-DL) . . . 31
3.3.1 Overcomplete **D** (Cov-DL-1) . . . 32
3.3.2 Undercomplete **D** (Cov-DL-2) . . . 33
3.3.3 Undercomplete **D** (Cov-DL-2) . . . 33
3.3.4 Remarks . . . 34
3.4 Experiments . . . 36
3.4.1 EEG Simulation . . . 36
3.4.2 Experiments on Real EEG . . . 38
3.5 Conclusion . . . 39
3.6 Acknowledgements . . . 39

Chapter 4 Basis Selection for Independent Sources . . . 47
4.1 Introduction . . . 48
4.2 Methods . . . 50
4.2.1 Greedy Methods . . . 50
4.2.2 BASICA . . . 52
4.2.3 BASRICA . . . 57
4.3 Connections to M-SBL . . . 58
4.3.1 BASICA . . . 58
4.3.2 BASRICA and M-SBL . . . 60
4.4 Experiments . . . 61
4.4.1 Simulated Data . . . 61
4.4.2 Experiments on real EEG data . . . 62
4.5 Conclusion . . . 64
4.6 Acknowledgements . . . 65

Chapter 5 Robust Joint-Sparse Recovery On Data with Outliers . . . 67
5.1 Introduction . . . 67
5.2 Sensitivity of M-SBL . . . 69
5.3 LTS and robust-MSBL . . . 70
5.3.1 LTS . . . 70
5.3.2 Robust-MSBL . . . 71
5.4 Experiments . . . 74
5.5 Conclusion . . . 77
5.6 Acknowledgements . . . 78

Chapter 6 Capturing Local Nonstationarities of EEG . . . 79
6.1 Introduction . . . 79
6.2 Shared ICA . . . 81

6.2.1 Algorithm . . . 82
6.2.2 Selection of Parameter k . . . 83
6.2.3 Experiments . . . 84
6.3 Base ICA . . . 85
6.3.1 Algorithm . . . 86
6.3.2 Experiments . . . 88
6.4 Conclusion . . . 89
6.5 Acknowledgements . . . 90

Chapter 7 Source-domain EEG Analysis of Sports-Related Concussion . . . 91
7.1 Introduction . . . 91
7.2 Methods . . . 93
7.2.1 Participants . . . 93
7.2.2 EEG acquisition protocol . . . 93
7.2.3 Data processing . . . 93
7.3 Measure Projection . . . 94
7.3.1 K-means IC clustering . . . 94
7.3.2 Measure Projection . . . 95
7.4 Results . . . 96
7.5 Discussion . . . 98
7.6 Conclusion . . . 100
7.7 Acknowledgements . . . 101

Chapter 8 Conclusion . . . 102

Bibliography . . . 104

LIST OF FIGURES

Figure 1.1: Scalp EEG records a mixture of brain activity generating in many different brain areas, as well as non-brain artifact sources. 6
Figure 1.2: (a) Low-density Emotiv EEG system (b) High-density EEG. 7

Figure 2.1: Validation of Theorem 1. Sample-wise orthogonal sources. The line corresponds to $N = M(M+1)/2$. Below the line we have guaranteed support recovery. Perfect recovery may also observed above the line. 20
Figure 2.2: Comparison of algorithms: Changing resolution N, $M = 20, k = N/5, L = 1024$. 21
Figure 2.3: Comparison of algorithms: Varying the number of sources k, $M = 20, N = 500, L = 1024$. 21
Figure 2.4: Comparison of algorithms: Varying the source duration L. Other parameters are fixed to $M = 20, N = 500, k = 80$. 22
Figure 2.5: Mean failure ratio for support recovery of brain sources. $M = 32, N = 150$. 23
Figure 2.6: Mean failure ratio for support recovery of brain sources. $M = 32, N = 200$. 24
Figure 2.7: Mean failure ratio for support recovery of brain sources. $M = 32, N = 250$. 24

Figure 3.1: The summary of two different strategies of Cov-DL for overcomplete EEG source identification. Cov-DL-1 involves dictionary learning stage thus $k < M(M+1)/2$ sources are assumed to be active at any given segment. Cov-DL-2 does not require sparsity of sources. . . . 40
Figure 3.2: A geometrical explanation of Cov-DL for $M = 2, k = 2$. If $N = 3$, then $\mathbf{A} \in \mathbb{R}^{2\times 3}$, and $\mathbf{D} \in \mathbb{R}^{3\times 3}$. In this case $\mathbf{d_1}, \mathbf{d_2}, \mathbf{d_3}$ are identifiable with a DL algorithm applied on the data of vectorized outer products. Associated $\mathbf{a_1}, \mathbf{a_2}, \mathbf{a_3}$ can then be found via solving Eg. (3.8). 41
Figure 3.3: A geometrical explanation of Cov-DL for $M = 2, k = 2$. If N = 2, then $\mathbf{D} \in \mathbb{R}^{3\times 2}$, and data is not sparse since $k = N = 2$. $\mathbf{D}$ is not identifiable through learning the 2-dimensional subspace. We solve Eq. (3.9) to directly find $\mathbf{A}$ such that $\mathbf{D}$ will span $\mathcal{R}(\mathbf{U})$ (Cov-DL-2). 41
Figure 3.4: (a) Randomly located and oriented $N = 64$ dipoles/sources in the MNI head model that generate the simulated EEG. (b) Some of the scalp maps associated with the dipoles in (a). These constitute columns of true mixing matrix $\mathbf{A}_{\text{true}} \in \mathbb{R}^{32\times 64}$. 42
Figure 3.5: (a) Outer product of the source matrix in a 2sec. segment (sample-covariance) from Scenario 1; $M = 32, N = 64$, $k = 64$. (b) Outer product of the source matrix in a 2sec. segment from Scenario 2; $M = 8, N = 40$. $k = 10$ sources are active at any given segment. . . 43

Figure 3.6: Simulation results for three cases: complete, x2 overcomplete, x5 overcomplete. Complete case: $M = 32, N = 32, k = 32$, Cov-DL-2 is used. Second case: $M = 32, N = 64, k = 64$, Cov-DL-2 is used. Third case: $M = 8, N = 40, k = 10$, Cov-DL-1 is used. 44
Figure 3.7: EEGLAB sample data. $M = 5, N = 30$. AMICA is trained with 6 models. Cov-DL-1 is performed. 45
Figure 3.8: Motor Imagery Task, $M = 5, N = 30$. Cov-DL-1 is performed. . . . 45
Figure 3.9: Arrow Flanker task. $M = 11, N = 30$. AMICA is trained with 3 models. Cov-DL-2 is used. 46

Figure 4.1: $M = 10, N = 20, L = 1000$. gRICA performs the best in both cases. Average computation time for gRICA is 303 seconds for one trial. . 53
Figure 4.2: BASRICA solutions with varying lambda on an experiment with M = 3, N = 64. The sparsity of γ increases with λ. Solution to a complete set with M nonzero γ can be achieved by tuning λ. (Yet, we used $\lambda = 0.05$ in all our experiments, as was used in [52]) . . . 59
Figure 4.3: (a) Comparison of algorithms on synthetically generated data with EEG scalp maps dictionary (b) Perfomance of ICA + column matching on the same type of data. It requires 500x data points compared to BASICA and BASRICA to recover 95% of the true sources . . . 66

Figure 5.1: $n = 100, k_1 = 10, k_2 = 10$. 75
Figure 5.2: $n = 100, k_1 = 10, k_2 = 60$. 76
Figure 5.3: Data with no outliers (ideal case). 77

Figure 6.1: Example source model for shared ICA. Epoch length is $L = 500$, number of shared sources $k = 15$. Each epoch has 5 unique sources. Cov-DL cannot be applied in this model because an epoch-specific source only occurs for a short time. 81
Figure 6.2: $M = 5, N_e = 10, k = 3$. Break point is shown with circle. 85
Figure 6.3: Cost function values at the optimum trained with different k. The circle denotes the optimum break point found with F-test. 86
Figure 6.4: (a) First row consists of scalp maps associated with motor imagery that traditional ICA found. Second row are the corresponding scalp maps that shared ICA found. (b) Some of the brain sources that only shared ICA revealed. 87
Figure 6.5: Recovery of sources with increasing L 89

Figure 7.1: (a) Voxels that show significantly consistent spectra among nearby source locations ($p < 0.05$). (b) Domains created by affinity propagation clustering with maximum similarity threshold across domains 0.9. Domain 4 is the only domain in the frontal part of the brain. . . 97

Figure 7.2: Measure-projected IC log spectra for the concussed and control groups at the maximally centered “exemplar" IC for Domain 4. Shaded regions indicate the standard error of the mean. Black lines correspond to spectral bands with a significant group difference. . . 98
Figure 7.3: Scalp maps of highest contributing 9 sources to Domain 4. 99
Figure 7.4: DIPFIT localized equivalent dipoles for 9 highest contributing IC’s for Domain 4. 100

LIST OF TABLES

Table 2.1: Mean Success Ratio under Different Noise Levels. $M = 20, N = 500, k = 100, L = 100$ 22

Table 4.1: Percentage of 6 second epochs for which Algorithm i (row) produces more MIR than Algorithm j (column). 64

Table 4.2: Percentage of 4 second epochs on which Algorithm i has higher MIR than Algorithm j. Row and column indexes are i, j respectively. . . . 64

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Professor Kenneth Kreutz-Delgado for his overall encouragement, patience and guidance during my PhD study. I thank him for bringing me into the exciting field of sparse recovery and dictionary learning, and introducing me to the possible applications in the field of EEG, to which I will continue to contribute in my career. I will be forever grateful to Dr. Scott Makeig for his intellectual and financial support during my PhD. I would like to thank him for conveying his enthusiasm about the future of EEG imaging, and teaching me how to ask and tackle the important scientific questions.

I would like to extend my thanks to my collaborators and colleagues at Swartz Center for Computational Neuroscience (SCCN), including Nima Bigdely-Shamlo, whose guidance helped me define the problem addressed in this thesis, and Christian Kothe, for his valuable feedbacks. I thank my office mates Yute Wang, Ramon Martinez, Clement Lee, Robert Buffington and other members of SCCN for creating such a warm and friendly lab environment. I would like to thank Dr. Harinath Garudadri for financially supporting part of my study and his overall encouragement.

I would like to thank my parents, sister and family for their continuing support and encouragement not only during my quest for a PhD but throughout my entire education life. My mother devoted her life to her children's happiness and success, made sure that we received the best of everything including the highest quality education. My father introduced me to the beauty of mathematics at an early age and has been the greatest influence on my love for math, engineering, and science. Even though my sister was away, she gave me strength and provided emotional support during tough times of the PhD journey. I am grateful to my girlfriend Natasha Howard and my close friends in San Diego for their emotional support and for acting as my immediate family when I needed encouragement.

Portions of this dissertation are based on papers that are either published, in submission or in preparation for submission in various journals and conferences, which I have co-authored with others. Below are the details along with my contributions.

Chapter 2, in part, is a reprint of the material as it appears in: Balkan, Ozgur; Kreutz-Delgado, Kenneth; Makeig, Scott, "Localization of More Sources Than Sensors via Jointly-Sparse Bayesian Learning", Signal Processing Letters, IEEE 21, no. 2 (2014): 131-134. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in: Balkan, Ozgur; Kreutz-Delgado, Kenneth; Makeig, Scott, "Covariance-domain Dictionary Learning for Overcomplete EEG Source Identification" submitted to IEEE Transactions on Biomedical Engineering. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part, is a reprint of the material as it appears in: Balkan, Ozgur; Bigdely-Shamlo, Nima; Kreutz-Delgado, Kenneth; Makeig, Scott, "Basis Selection for Maximally Independent EEG Sources", In Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, pp. 6639-6642. IEEE, 2014. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in: Balkan, Ozgur; Kreutz-Delgado, Kenneth; Makeig, Scott, "Robust Joint Sparse Recovery On Data with Outliers", In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 3821-3825. IEEE, 2013. The dissertation author was the primary investigator and author of this paper.

Chapter 6, in part, is currently being prepared for submission for publication of the material: Balkan, Ozgur; Kreutz-Delgado, Kenneth; Makeig, Scott, "Capturing Non-Stationary EEG Sources via Shared Source ICA model".

Chapter 7, in full, is a reprint of the material submitted and accepted for pub-

lication as: Balkan, Ozgur; Virji-Babul, Naznin; Miyakoshi, Makoto; Makeig, Scott; Garudadri, Harinath "Source-domain Spectral EEG Analysis of Sports-Related Concussion via Measure Projection Analysis", EMBC 2015, IEEE.

VITA

2009	B. S. in Electrical and Electronic Engineering, Bilkent University, Turkey
2009-2015	Graduate Student Researcher, University of California, San Diego, CA
2011	M. S. in Electrical Engineering, University of California, San Diego, CA
2013	Research Intern, Siemens Corporate Research, Princeton, NJ
2015	Ph. D. in Electrical Engineering, University of California, San Diego, CA

PUBLICATIONS

Ozgur Balkan, Kenneth Kreutz-Delgado, Scott Makeig "Localization of More Sources Than Sensor via Jointly-Sparse Bayesian Learning", *Signal Processing Letters, IEEE*, 21(2) (2014), 131-134.

Ozgur Balkan, Kenneth Kreutz-Delgado, Scott Makeig "Covariance-domain Dictionary Learning for Overcomplete EEG Source Identification", submitted to *Transactions on Biomedical Engineering, IEEE*.

Ozgur Balkan, Nima Bigdely-Shamlo, Kenneth Kreutz-Delgado, Scott Makeig "Basis Selection for Maximally Independent EEG Sources", *In Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, (pp. 6639-6642).

Ozgur Balkan, Kenneth Kreutz-Delgado, Scott Makeig "Robust Joint Sparse Recovery On Data with Outliers" *In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, IEEE.* (pp. 3821-3825).

Ozgur Balkan, Naznin Virji-Babul, Makoto Miyakoshi, Scott Makeig, Harinath Garudadri, "Source-domain Spectral EEG Analysis of Sports-Related Concussion via Measure Projection Analysis", accepted to *IEEE EMBC 2015*

Alican Nalci, Alireza Khodamoradi, Ozgur Balkan, Fatta Nahab, Harinath Garudadri, "A Computer Vision Based Candidate for Functional Balance Test", accepted to *IEEE EMBC 2015*

Arnold Yeung, Harinath Garudadri, Carolyn Van Toen, Patrick Mercier, Ozgur Balkan, Scott Makeig, Naznin Virji-Babul, "Comparison of Foam-Based Dry EEG Electrode and Spring-Loaded Dry EEG Electrodes with Wet Electrodes in Resting and Moving States", accepted to *IEEE EMBC 2015*

ABSTRACT OF THE DISSERTATION

Support Recovery and Dictionary Learning for Uncorrelated EEG Sources

by

Ozgur Yigit Balkan

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California, San Diego, 2015

Professor Kenneth Kreutz-Delgado, Chair

Sparse signal recovery and dictionary learning methods have found a vast number of applications including but not limited to data compression, machine learning and biomedical source localization/separation. The common underlying assumption in domains of application of these methods is that signals of interest are either sparse or can be sparsified in a transform domain. For source localization or identification this implies that the number of coefficients needed to represent the source signals in the transform domain should be less than the number of sensors. This evidently imposes constraints on the types of signals that can be recovered. In this work, we show that these constraints can be relaxed if the source signals are uncorrelated. Our work is inspired by the nature

of electroencephalography (EEG) sources for which the independence assumption has been widely and successfully used.

We focus on the multiple-measurement-vector (MMV) model of the sparse inverse problem. Under the assumption of uncorrelated sources, we first show that the required sparsity conditions for accurate signal support recovery can be relaxed which enables EEG source localization when more sources than sensors are simultaneously active. Later, we show that one can transform the traditional dictionary learning formulation into the covariance-domain to leverage the correlation information of the sources. Our covariance-domain dictionary learning framework can accurately identify the EEG scalp mixing matrix even when sources are not sparse in the traditional sense. This method enables the use of low-cost, low-density systems for high-density EEG brain imaging, which traditionally suffers from poor performance when using constraint-sensitive source separation algorithms like Independent Component Analysis. We also present locally-complete source separation algorithms that tackle the non-stationary nature of EEG sources.

Finally, we present algorithms that targets identification of independent sources given an overcomplete dictionary. Our algorithms differ from the usual MMV sparse recovery algorithms in the sense that they optimize independence of the sources rather than their sparsity. We also a present robust bayesian algorithm for joint-sparse recovery in the MMV formulation.

Chapter 1

Introduction

Techniques of sparse signal recovery from a limited number of measurements in combination with dictionary learning have found a large number of applications in diverse fields of engineering including but not limited to information theory, neural networks, data compression, machine learning, bioinformatics and neuroimaging [8]. In this chapter, we introduce some concepts related to sparse signal recovery and dictionary learning, and provide background on their application to electroencephalography (EEG) source identification and localization.

1.1 Sparse Signal Recovery

1.1.1 Single Measurement Vector (SMV) Model

In the field of sparse signal recovery, the most basic signal model is the following linear model,

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e} \tag{1.1}$$

where $\mathbf{y} \in \mathbb{R}^{M\times 1}$ is the observed vector, $\mathbf{A} = [\mathbf{a_1} \dots \mathbf{a_N}] \in \mathbb{R}^{M\times N}$ is the known dictionary matrix, $\mathbf{x} \in \mathbb{R}^{N\times 1}$ is the unknown source vector, and $\mathbf{e} \in \mathbb{R}^{M\times 1}$ is the noise vector. The goal is to find the original source vector $\mathbf{x}$ from $\mathbf{y}$, given the dictionary $\mathbf{A}$. Given the single observation $\mathbf{y}$, the problem of determining $\mathbf{x}$ is known as the single measurement vector (SMV) problem. Depending on the application, the dictionary $\mathbf{A}$ can have various names such as mixing matrix (source separation), forward model (neuroelectromagnetic source localization), or sensing matrix (compressed sensing). When $\mathbf{A}$ is overcomplete ($M < N$), it has a non-trivial nullspace, and hence there is not a unique solution that satisfies (1.1). However, enforcing sparsity constraints on the unknown solution vector x can reduce the solution set to a unique optimal solution whose cost function can be defined as below

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{x}\|_0 \\ \text{s.t} \quad & \mathbf{y} = \mathbf{A}\mathbf{x}. \end{aligned} \tag{1.2}$$

Here, $\|\mathbf{x}\|_0$ is defined to be the number of nonzero values of the vector $\mathbf{x}$. In this dissertation, we will denote the number of nonzero values in the original unknown vector $\mathbf{x}$ with k. It is known that when k is sufficiently small, the optimal solution to the above cost function is identical to the true unknown x [39]. This optimization, however, is a discrete NP-hard problem. A tractable alternative is often possible because, replacing $\|\mathbf{x}\|_0$ with the efficiently optimizable convex objective ℓ_1 norm $\|x\|_1$, still returns the true solution under certain conditions relating k and dictionary properties such as spark and mutual coherence [18, 17]. In addition to such convex methods, many non-convex methods have been developed to solve the SMV sparse recovery problem such as FOCUSS [39], sparse Bayesian Learning [90, 96, 98], and message passing [29, 80, 10]. The success of all of these convex and non-convex algorithms degrade as k

increases and approaches M.

1.1.2 Multiple Measurement Vector (MMV) Model

In certain applications of sparse signal recovery, such as direction-of-arrival (DOA) estimation and neuroelectromagnetic source localization, the components of $\mathbf{y}$ represent the collected signals on M sensors, the dictionary $\mathbf{A}$ represents the projection of each possible N sources to M sensors, and the components of $\mathbf{x}$ give the source activations. In such applications, source activities are usually captured with multiple snapshots (L) in time and thus producing the sensor signal matrix $\mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_L}]$. It is usually assumed that an active source stays active in the consecutive frames and a silent source stays silent hence preserving the sparsity structure and creating the multiple measurement vector (MMV) sparse linear model,

$$\mathbf{Y} = \mathbf{AX} + \mathbf{E} \tag{1.3}$$

where $\mathbf{E} \in \mathbb{R}^{M \times L}$ is the noise matrix and the rows of the source matrix $\mathbf{X} \in \mathbb{R}^{N \times L}$ carry the time courses of the k sources [21]. In the MMV sparse model, also called the joint-sparse (row-sparse) model, the assumption is that for every column of $\mathbf{X}$, the same row indices are nonzero. In this case, k denotes the number of nonzero rows of the matrix $\mathbf{X}$. Extensions of the sparse recovery algorithms for the SMV case are proposed to tackle the MMV problem such as M-FOCUSS [21, 105], and M-SBL[97, 98]. As in the case of SMV formulation, the sparsity of the source matrix is correlated with the success of these recovery algorithms. It was also observed that as the number of snapshots L increases so does the recovery performance [32, 33, 46]. Indeed, when $L > k$, subspace based algorithms such as [50, 24, 49, 55] can return the true solution for all $1 \leq k \leq M - 1$ with mild conditions on the dictionary.

Both in the SMV and MMV model, if the source signal is sparse it is usually sufficient to find the indices of nonzero rows of $\mathbf{X}$ denoted by the set S, since the nonzero source signals can then be obtained by the pseudo-inverse solution $\hat{\mathbf{X}} = \mathbf{A_S}^{\dagger}\mathbf{Y}$, where $\mathbf{A_S}$ is derived from the dictionary matrix $\mathbf{A}$ by deleting the columns of $\mathbf{A}$ associated with the zero rows of $\mathbf{X}$. The set of nonzero indices of the source signal S is called the *support*. Recovering the support is tantamount to recovering the source signal when $|S| \leq M$. In source localization and DOA estimation, recovering the support also means to identify the locations of sources.

In the case of EEG, the dictionary can represent different things depending on the context. If the goal is to perform source localization onto the cortex, then the dictionary is called the lead-field matrix and stores the pre-computed projections of each source location in the cortex (high-resolution) to the scalp sensors [93, 95]. If the goal is to represent all possible sources contributing to the scalp EEG - not necessarily all brain-related - then the dictionary is called mixing matrix [66, 26]. The lead field matrix is usually larger than the mixing matrix. Indeed, the mixing matrix is traditionally assumed to have $N < M$ [3, 68, 66].

1.2 Dictionary Learning

When the dictionary $\mathbf{A}$ is unknown, it can be learned provided we have multiple measurements $\mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, \ldots \mathbf{y_n}]$ created by $\mathbf{y_i} = \mathbf{A}\mathbf{x_i}$ where $\mathbf{x_i}$ possess different sparsity patterns. Learning is done by jointly solving the following optimization problem on $\mathbf{A}, \mathbf{X}$,

$$\min_{A,X} \frac{1}{2} \sum_{i=1}^{n} \|\mathbf{y_i} - \mathbf{A}\mathbf{x_i}\|_2^2 + \lambda \sum_{i=1}^{n} g(\mathbf{x_i}) \tag{1.4}$$

where $g(\cdot)$ promotes sparsity. The uniqueness of the dictionary is dependent on the sparse recovery conditions, namely the sparsity k of the sources and dictionary properties

[38, 87, 42]. Recovery is not possible if $k \geq M$ since any random dictionary is sufficient to represent data points $\mathbf{Y}$ using only M basis vectors. Some applications of dictionary learning are for image denoising/super-resolution and restoration [31, 103, 104], blind source separation [108, 59], and training deep neural networks [106, 54].

Blind source separation leveraging dictionary learning methods [108, 59, 99, 100, 101] assume that source signals are either sparse in activation or can be sparsified in another domain. This imposes a constraint on the types of sources that can be recovered and limits its use in EEG source identification for which the true nature of sources are not known a priori.

Many algorithms have been proposed for dictionary learning over the last decade [2, 63, 75, 34, 84, 23, 79, 83, 102]. In this work, we do not propose a novel dictionary learning algorithm but rather a framework that incorporates uncorrelatedness of the sources, in which any dictionary learning algorithm can be used.

1.3 EEG Source Identification and Independent Component Analysis

As a non-invasive brain imaging modality, electroencephalography (EEG) provides high temporal resolution, applicability in mobile settings, and direct measurement of electrical brain activity as opposed to BOLD activity measured in fMRI. However, a major issue in EEG signal processing is that signals measured on the scalp surface do not each index a single source of brain activity. Because of the broad point spread function of generated potentials in the brain, EEG data collected on scalp channels is a mixture of simultaneously active brain sources distributed over many different brain areas. See Figure 1.1. In addition, non-brain sources such as eye and muscle movements contribute to the mixing process as well, which makes direct channel-level EEG analysis

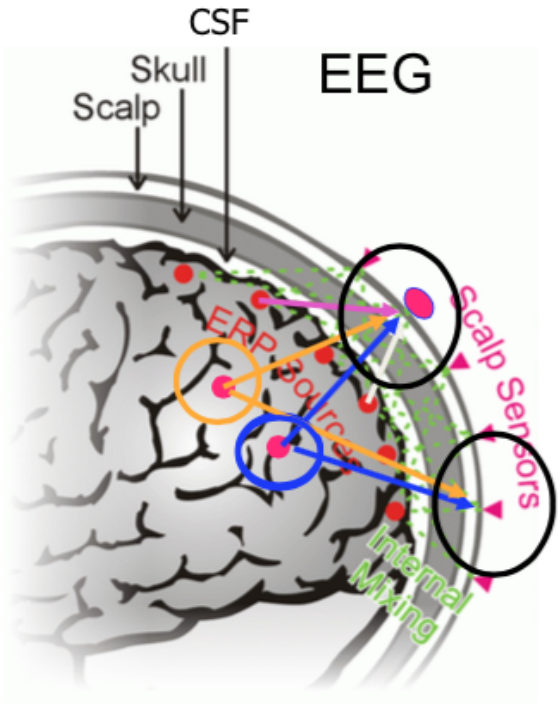


Figure 1.1: Scalp EEG records a mixture of brain activity generating in many different brain areas, as well as non-brain artifact sources.

problematic. For accurate brain activity monitoring, individual sources involved in the mixture have to be identified and extracted from scalp channel data.

Because of the fact that volume conduction and mixing at the sensors is linear, EEG mixing can be formulated as follows

$$\mathbf{Y} = \mathbf{AX} \tag{1.5}$$

where $\mathbf{Y} \in \mathbb{R}^{M \times N_d}$ is the matrix containing collected EEG data at M sensors for N_d data points. $\mathbf{A} \in \mathbb{R}^{M \times N}$ is the unknown mixing matrix, and $\mathbf{X} \in \mathbb{R}^{N \times N_d}$ contains the activations of N sources. The i-th column of $\mathbf{A}$, denoted as $\mathbf{a_i}$, represents the relative projection weights of the i-th source to each channel. The so-called EEG inverse problem is to identify $\mathbf{A}$ and $\mathbf{X}$, given sensor data $\mathbf{Y}$ [64]. Learning the columns of $\mathbf{A}$, namely the scalp maps, can further enable source localization in the cortex through methods such

(a) (b)

Figure 1.2: (a) Low-density Emotiv EEG system (b) High-density EEG.

as DIPFIT [71] or sLORETA [76]. Identifying the rows of $\mathbf{X}$ can enable the computing of time-series measures such as event-related potentials (ERPs), event-related spectral perturbation (ERSPs), and spectral components.

A commonly applied method to solve the EEG inverse problem has been to use independent component analysis (ICA) [64, 65]. Assuming statistical independence between source activities, ICA can separate the scalp mixture into underlying source time-series $\mathbf{X}$ and identify the mixing matrix $\mathbf{A}$. It was shown in [28] that ICA methods are well suited for solving the EEG inverse problem since independence among sources was found to be positively correlated with the number of brain sources that can be extracted from data. ICA has been extensively applied on EEG for artifact rejection and source separation [47, 48] and has been shown to increase accuracy in brain-computer-interface paradigms [92]. However, one major drawback of ICA is that the number of mixed sources is assumed to be less than or equal to the number of sensors ($N \leq M$). This assumption undermines reliability and power of ICA, especially in low-density EEG systems ($M < 32$ number of channels).

Low-density EEG systems have the benefit of easy setup time and being low-cost compared to high-density EEG systems. See Figure 1.2.[1] In addition, wireless low-density systems enable the use of EEG in mobile settings and allows for a wide range

[1] Images taken from emotiv.com

of possible applications that are not possible with current high-density EEG systems. However, the applications of low-density EEG in the consumer market remained primitive so far, and is in need of advanced signal processing for overcomplete source identification and source localization.

1.3.1 EEG Non-stationarity

Independent Component Analysis requires stationarity of the sources whereas the nature of true EEG sources might not be so. Namely, EEG experiments can involve complicated tasks during which many separate brain regions might generate signals that are mixed at the scalp. Source dynamics (variance/source distribution) of these separate regions might change either during the progress of the experiment, or at pre-task/during task/post-task durations. Certain artifacts might be present only at some portions of the experiments. The reality is that, the number of total sources N can exceed the number of sensors M in an unpredictable or uncontrollable manner, in which case ICA can give inaccurate results.

Dictionary learning algorithms can recover more sources than sensors ($N > M$), assuming the constraint that at any given time we have $k < M$. Unfortunately, this constraint often prohibits their use for low-density EEG systems, when M is low. To capture the non-stationarity of EEG, the use of mixture model ICA has been proposed [74] for the case of $k = M$. Mixture model ICA has the benefit of capturing distinct changes in short time-frame source stationarity (such as occurs during pre-seizure/seizure/post-seizure states in epilepsy data). However, for most EEG experiments, non-stationarities are subtle and the number of distinct source configuration states can be overwhelmingly larger than the number of mixture models.

1.4 Our Contributions

The contributions of this work can be summarized as follows:

In Chapter 2, we show that if the sources are uncorrelated and are in the form of the MMV problem, the support of the sources can be recovered with Sparse Bayesian Learning for both $k \geq M$ as well as $k < M$. We establish sufficient conditions and provide simulation results for EEG source localization.

In Chapter 3, we show that by incorporating the correlation information into the dictionary learning framework, we can recover the overcomplete dictionary ($N > M$), even when the number of active sources exceed the number of sensors at any given time ($k > M$). We also show the utility of the algorithm on EEG source identification for low-density EEG systems.

In Chapter 4, we develop a basis selection framework to identify independent sources in the MMV model (assuming a known dictionary).

In Chapter 5, we develop a robust algorithm for the MMV model based on Sparse Bayesian Learning.

In Chapter 6, we develop two algorithms that tackle the problem of possible non-stationarities in EEG signals. The proposed methods assume $k = M$ at any given EEG segment yet are more flexible than the mixture model ICA mentioned before.

In Chapter 7, we demonstrate an application of EEG source separation and localization aimed at revealing biomarkers of concussion related injuries.

Chapter 2

Support Recovery of Uncorrelated Sources

In this chapter we analyze the jointly-sparse signal recovery problem (MMV) in the regime where the number of sources k is larger than the number of measurements M. We show that the support set of sources can still be recovered with sparse bayesian learning (M-SBL) even if $k \geq M$. We provide sufficient conditions on the dictionary and sources which theoretically guarantee support set recovery in the noiseless case of M-SBL. We validate our sufficient conditions with experiments and also demonstrate that M-SBL outperforms M-CoSaMP, the algorithm recently used to localize more sources than sensors. Finally, we experimentally show robustness of the approach in the presence of noise.

2.1 Introduction

In the traditional MMV model the source matrix $\mathbf{X} \in \mathbb{R}^{N \times L}$ consists of $k < M$ rows that are nonzero. The data matrix $\mathbf{Y} \in \mathbb{R}^{M \times L}$ is obtained as

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E} \tag{2.1}$$

where $\mathbf{A} \in \mathbb{R}^{M \times N}$ is the dictionary, $\mathbf{E} \in \mathbb{R}^{M \times L}$ is the noise matrix. In this model, L different time samples are collected consecutively in time and the nonzero rows of the source matrix $\mathbf{X}$ carry the time courses of the k sources that are assumed to be active in that time window. The goal is to recover the indices of these nonzero rows from data $\mathbf{Y}$.

Previous efforts to solve the jointly-sparse problem [94, 37, 95] have assumed that the number of sources k is smaller than the number of measurements M, as generally all other joint sparse solvers do [78]. Recently, it was observed that $k < M$ is not a necessary condition, and more sources than sensors can be localized using M-CoSaMP [16, 53].

Here, we show that multiple measurement sparse bayesian learning (M-SBL) [97] can also recover the support set in the noiseless case when $M \leq k \leq N$. Moreover, we provide exact recovery conditions for M-SBL in that regime, conditions which are missing from previous reports on M-CoSaMP and other algorithms. We also experimentally show that M-SBL outperforms the suggested M-CoSaMP algorithm under non-ideal conditions. We assume that mixing at the sensors is instantaneous (no time delay between sources and sensors) and the environment is anechoic. Therefore, we use the time domain representation of the problem instead of the frequency domain.

The sufficient conditions for exact support recovery for M-SBL in the regime $k \geq M$ are twofold. The first condition requires the orthogonality (uncorrelated) of the active sources, that is $\mathbf{X_S}\mathbf{X_S^T} = \Lambda$, where Λ is any diagonal matrix and $\mathbf{X_S} \in \mathbb{R}^{k \times L}$ is the

matrix of active sources. This condition is not very far from ideal when the sources are independent, such as the case of audio sources or brain sources [28], and given enough snapshots L, we will then have $\mathbf{X_S X_S^T} \approx \Lambda$. The second condition imposes a constraint on the sensing dictionary $\mathbf{A}$. For deterministic dictionaries, this condition is easier to check than the spark, or RIP constraints [30, 7]. For random dictionaries, it is equivalent to a dictionary size constraint $N \leq \frac{M(M+1)}{2}$.

The outline of the chapter is as follows: In Section 2.2, we summarize the M-SBL algorithm. In Section 2.3, we provide a theoretical analysis of guaranteed support recovery for any k in the noiseless case. We perform noiseless and noisy experiments in Section 3.4 and provide conclusions in Section 2.5.

2.2 M-SBL

In the M-SBL framework, the i-th row of $\mathbf{X}$, denoted as $\mathbf{x_{i.}}$, has an L-dimensional zero mean gaussian prior whose variance is controlled by a hyperparameter γ_i. That is,

$$p(\mathbf{x_{i.}}; \gamma) \triangleq \mathcal{N}(0, \gamma_i \mathbf{I}). \tag{2.2}$$

We also have

$$\mathbf{p(y_{.j}|x_{.j})} \triangleq \mathcal{N}(\mathbf{Ax_{.j}}, \sigma^2 \mathbf{I}) \tag{2.3}$$

$$\mathbf{p(Y|X)} = \prod_{j=1}^{L} \mathbf{p(y_{.j}|x_{.j})} \tag{2.4}$$

assuming gaussian noise with known variance σ^2, with $\mathbf{x_{.j}}$ denoting the j-th column of $\mathbf{X}$. Integrating out the unknown sources $\mathbf{X}$, we derive $p(\mathbf{Y}; \gamma)$, the marginal likelihood of data $\mathbf{Y}$ given the hyperparameters $\gamma \in \mathbb{R}^N$, which is to be maximized. We apply the $-2\log(\cdot)$

transformation to arrive at the M-SBL cost function to be minimized [97, 98, 90],

$$\mathcal{L}(\gamma) \triangleq -2\log(p(\mathbf{Y};\gamma)) = -2\log\int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X};\gamma)d\mathbf{X}$$
$$\equiv \log|\Sigma| + \frac{1}{L}\sum_{t=1}^{L}\mathbf{y_{.t}}^{T}\Sigma^{-1}\mathbf{y_{.t}} \tag{2.5}$$

with $\Sigma \triangleq (\mathbf{A}\Gamma\mathbf{A^T} + \sigma^{\mathbf{2}}\mathbf{I})$, $\Gamma \triangleq \mathrm{diag}(\gamma)$. A common method to reach a local minimum of (5.1) is to use the EM procedure [97], however the updates may be slow depending on the size of the problem. Reference [98] provides the following fixed point update that is both fast and guaranteed to decrease the likelihood function at every step,

$$\gamma_i^{(k+1)} = \frac{\gamma_i^{(k)}}{\sqrt{\mathbf{a_i^T}(\Sigma^{(k)})^{-1}\mathbf{a_i}}}\frac{\|\mathbf{Y^T}(\Sigma^{(k)})^{-1}\mathbf{a_i}\|_2}{\sqrt{n}} \tag{2.6}$$

Note that with this update scheme, for $\gamma^{(0)} \geq 0$, the converged solution is nonnegative, $\gamma^* \geq 0$. M-SBL can also learn the noise parameter σ^2, however we do not discuss it here since our theoretical analysis considers the the noiseless case, where $\sigma^2 \to 0$. After convergence, support set $\hat{S}$ is extracted from the solution $\hat{\gamma}$ by $\hat{S} = \{i, \hat{\gamma}_i \neq 0\}$.

In the noiseless case, [97] and [98] show that exact sparse reconstruction is guaranteed if the number of active sources $k < M$ and the active set of sources $\mathbf{X_S}$ are sample-wise uncorrelated. In the next section, we show that exact noiseless source localization (support set recovery) with M-SBL is also guaranteed in the regime $k \geq M$ with an additional constraint on the dictionary.

2.3 Analysis

2.3.1 Local Minima

First, we summarize the local minima analysis carried out in [98]. We start by letting $\sigma^2 \to 0$ (as appropriate for the noiseless case) and obtain the limiting cost function,

$$\mathcal{L}(\gamma) = L \log |\mathbf{A}\Gamma\mathbf{A}^{\mathbf{T}}| + \sum_{t=1}^{L} \mathbf{y}_{.\mathbf{t}}{}^{T} (\mathbf{A}\Gamma\mathbf{A}^{\mathbf{T}})^{-1} \mathbf{y}_{.\mathbf{t}} \tag{2.7}$$

We denote the matrix of the support set columns by $\mathbf{A_S}$. Under the condition (i), we construct the following cost function around the neighborhood of a local minimum γ^* as in [98],

$$\begin{aligned} \mathcal{L}(\alpha, \beta) = L \log |\alpha \mathbf{A}\Gamma^{*}\mathbf{A}^{\mathbf{T}} + \beta \mathbf{U}\mathbf{U}^{\mathbf{T}}| + \\ \sum_{t=1}^{L} \mathbf{y}_{.\mathbf{t}}^{\mathbf{T}} (\alpha \mathbf{A}\Gamma^{*}\mathbf{A}^{\mathbf{T}} + \beta \mathbf{U}\mathbf{U}^{\mathbf{T}})^{-1} \mathbf{y}_{.\mathbf{t}}, \end{aligned} \tag{2.8}$$

where $\mathbf{U} = \mathbf{A_S}\Lambda^{\frac{1}{2}}$ and $\Gamma^* = \text{diag}(\gamma^*)$. Note that the overall covariance has the structural form,

$$\exists\, \hat{\gamma} \in \mathbb{R}^N \text{ such that } \alpha \mathbf{A}\Gamma^{*}\mathbf{A}^{\mathbf{T}} + \beta \mathbf{U}\mathbf{U}^{\mathbf{T}} = \mathbf{A}\hat{\Gamma}\mathbf{A}^{\mathbf{T}} \tag{2.9}$$

where $\hat{\Gamma} = \text{diag}(\hat{\gamma})$. We put the constraints $\alpha, \beta \geq 0$ so that the elements of $\hat{\gamma}$ are guaranteed to be nonnegative. Note that when $\alpha = 1 and \beta = 0$, $\hat{\gamma} = \gamma^*$. For $\alpha = 1$ and for small $\beta > 0$, (i.e., when we add a small contribution to the total covariance from $\mathbf{U}$), we expect the cost function not to decrease since γ^* is by definition a local minima of (2.7). Thus the following conditions must hold,

$$\left.\frac{\partial \mathcal{L}(\alpha, \beta)}{\partial \alpha}\right|_{\alpha=1,\beta=0} = 0 \qquad \left.\frac{\partial \mathcal{L}(\alpha, \beta)}{\partial \beta}\right|_{\alpha=1,\beta=0} \geq 0 \tag{2.10}$$

The derivative of the cost w.r.t β at $\beta = 0$ is not necessarily zero since $\beta \geq 0$ due to the nonnegativity constraint that must be satisfied for the elements of $\hat{\gamma}$.

Using the orthogonality constraint $\mathbf{X_S X_S}^{\mathbf{T}} = \Lambda$, the second term in (2.8) can be expressed as,

$$\begin{aligned}
&\sum_{t=1}^{L} \mathbf{y_{.t}^T}(\alpha \mathbf{A}\Gamma^* \mathbf{A^T} + \beta \mathbf{U}\mathbf{U^T})^{-1}\mathbf{y_{.t}} \\
&= \mathrm{tr}(\mathbf{Y}^T(\alpha \mathbf{A}\Gamma^* \mathbf{A^T} + \beta \mathbf{U}\mathbf{U^T})^{-1}\mathbf{Y}) \\
&= \mathrm{tr}(\mathbf{X_S^T}\Lambda^{-\frac{1}{2}}\mathbf{U^T}(\alpha \mathbf{A}\Gamma^* \mathbf{A^T} + \beta \mathbf{U}\mathbf{U^T})^{-1}\mathbf{U}\Lambda^{-\frac{1}{2}}\mathbf{X_S}) \\
&= \mathrm{tr}(\mathbf{U^T}(\alpha \mathbf{A}\Gamma^* \mathbf{A^T} + \beta \mathbf{U}\mathbf{U^T})^{-1}\mathbf{U}\Lambda^{-\frac{1}{2}}\mathbf{X_S X_S^T}\Lambda^{-\frac{1}{2}}) \\
&= \mathrm{tr}(\mathbf{U^T}(\alpha \mathbf{A}\Gamma^* \mathbf{A^T} + \beta \mathbf{U}\mathbf{U^T})^{-1}\mathbf{U}). \qquad (2.11)
\end{aligned}$$

2.3.2 Exact Support Recovery Conditions

Extending the results described in [97, 98], we carry out an analysis for the case where $k \geq M$. First, we define a function $f : \mathbb{R}^{M \times N} \rightarrow \mathbb{R}^{M(M+1)/2 \times N}$:

Definition. $\mathbf{B} = f(\mathbf{A})$ means that the i-th column of $\mathbf{B}$ is given by $\mathbf{b_i} = \mathrm{vech}(\mathbf{a_i a_i^T})$. Vech is the half-vectorization function, which results in a column vector obtained by vectorizing only the lower triangular part of a matrix.

Theorem 1. Given a dictionary $\mathbf{A}$ and a set of observed signals $\mathbf{Y}$, M-SBL recovers the support set of any size k exactly in the noiseless setting, if the following conditions hold.

(i) The active sources $\mathbf{X_S}$ are orthogonal. Namely, $\mathbf{X_S X_S}^{\mathbf{T}} = \Lambda$, Λ is a diagonal matrix.

(ii) $\mathrm{rank}(f(\mathbf{A})) = N$.

Proof. Assume that $k \geq M$[1]. We will show that the minimum of (5.1) is unique if above conditions hold. Moreover, the solution γ^* must have positive values only at indices $i \in S$, thus recovering the support set.

Carrying out the differentiation for the first condition of (2.10) gives,

$$\operatorname{tr}\left(\mathbf{U}^{\mathbf{T}}(\mathbf{A}\Gamma^*\mathbf{A}^{\mathbf{T}})^{-1}\mathbf{U}\right) = LM. \tag{2.12}$$

Similarly, for the second condition in (2.10) we have,

$$\left.\frac{\partial \mathcal{L}(\alpha,\beta)}{\partial \beta}\right|_{\alpha=1,\beta=0} = \sum_{i=1}^{M}\left(L\lambda_i - \lambda_i^2\right) \geq 0, \tag{2.13}$$

where λ_i is the i-th eigenvalue of $\mathbf{U}^{\mathbf{T}}(\mathbf{A}\Gamma^*\mathbf{A}^{\mathbf{T}})^{-\mathbf{1}}\mathbf{U}$. Note that $\lambda_i \geq 0$ since $\mathbf{U}^{\mathbf{T}}(\mathbf{A}\Gamma^*\mathbf{A}^{\mathbf{T}})^{-\mathbf{1}}\mathbf{U}$ is a positive-semidefinite matrix of size $M \times M$ due to the nonnegativity of γ^*. Combining (2.12) and (2.13) we get,

$$\begin{aligned}\left.\frac{\partial \mathcal{L}(\alpha,\beta)}{\partial \beta}\right|_{\alpha=1,\beta=0} &= L^2M - \sum_{i=1}^{M}\lambda_i^2 \\ &\leq L^2M - \sum_{i=1}^{M}\bar{\lambda}^2 \\ &= L^2M - L^2M = 0,\end{aligned} \tag{2.14}$$

with $\bar{\lambda} = \frac{1}{M}\sum_{i=1}^{M}\lambda_i = LM/M = L$.

The inequalities (2.13) and (2.14) together imply

$$\left.\frac{\partial \mathcal{L}(\alpha,\beta)}{\partial \beta}\right|_{\alpha=1,\beta=0} = 0 \tag{2.15}$$

and thus the eigenvalues all satisfy $\lambda_1 = \lambda_2 = \ldots = \lambda_M = L$.

[1]The case $k < M$ is described in reference [97].

We proceed by showing the following: If $\lambda_1 = \lambda_2 = \ldots = \lambda_M = L$, then we must have $\mathbf{A}\Gamma^*\mathbf{A}^\mathbf{T} = \frac{1}{L}\mathbf{U}\mathbf{U}^\mathbf{T}$. This is clear after we write

$$\mathbf{U}^\mathbf{T}(\mathbf{A}\Gamma^*\mathbf{A}^\mathbf{T})^{-\mathbf{1}}\mathbf{U}\mathbf{V} = L\mathbf{V}, \tag{2.16}$$

where $\mathbf{V}$ is a $k \times M$ matrix of M eigenvectors. This equality is equivalent to

$$\left(\mathbf{A}\Gamma^*\mathbf{A}^\mathbf{T}\right)^{-\mathbf{1}}\mathbf{U}\mathbf{V} = L(\mathbf{U}\mathbf{U}^\mathbf{T})^{-\mathbf{1}}\mathbf{U}\mathbf{V}, \tag{2.17}$$

which leads to

$$\mathbf{A}\Gamma^*\mathbf{A}^\mathbf{T} = \frac{1}{L}\mathbf{U}\mathbf{U}^\mathbf{T} = \frac{1}{L}\mathbf{A}_\mathbf{S}\Lambda\mathbf{A}_\mathbf{S}^\mathbf{T} = C. \tag{2.18}$$

where C is a constant matrix.

We now show that there is exactly one solution Γ^* that satisfies (2.18) if (ii) holds. Denoting the j-th column of $\mathbf{A}_\mathbf{S}$ as $\mathbf{A}_{\mathbf{S}j}$, a particular solution is,

$$\gamma_i^* = \begin{cases} \frac{1}{L}\Lambda_{jj}, & \text{if } i \in S, \text{where } \mathbf{a_i} = \mathbf{A}_{\mathbf{S}j} \\ 0, & \text{otherwise} \end{cases}. \tag{2.19}$$

Denoting $\mathbf{B} = f(\mathbf{A})$ as defined above, and vectorizing both sides of (2.18), we get,

$$\begin{aligned} \text{vech}(\sum_{i=1}^{N} \gamma_i^* \mathbf{a_i}\mathbf{a_i}^\mathbf{T}) &= \text{vech}(C) \\ \text{or} \quad \mathbf{B}\gamma^* &= \text{vech}(C) \end{aligned} \tag{2.20}$$

Finally we observe that if (ii) is satisfied, the null space of $\mathbf{B}$ is trivial and thus (2.20) has a unique solution.

□

2.3.3 Remarks

Theoretically, condition (i) of Theorem 1 is satisfied in the limit $L \to \infty$. However, for large L, we can get very close to sample-wise orthogonality ($\mathbf{X_S X_S^T} \approx \Lambda$) for independent sources, which is adequate for exact recovery in practical applications. The fact that perfect, asymptotic sample-wise orthogonality is not a hard requirement for good performance is demonstrated in the experiments.

The second condition can be related to the size of a random dictionary. It can be seen that for (ii) to be satisfied for a random dictionary $\mathbf{A}$, we must have $N \leq M(M+1)/2$. For deterministic dictionaries, the same equality is a necessary condition and can be satisfied even if the columns of $\mathbf{A}$ are coherent. Using this inequality we can estimate the maximum resolution of the source space given the number of sensors M, and vice versa. For instance, in an EEG source localization problem if the desired resolution is $N = 10000$ voxels on the cortex, at least $M = 142$ sensors are needed for condition (ii) to be satisfied. However, as we see in the experiments below the conditions of Theorem 1 are not necessary conditions.

2.4 Experiments

2.4.1 Theorem Validation

In this section, we validate Theorem 1 by experimenting with source activations that satisfy condition (i). For this purpose, we create k random source activations and perform a whitening transformation using eigenvalue decomposition to ensure that $\mathbf{X_S X_S^T} = \mathbf{I}$ at each trial. We create random dictionaries of various sizes and for each trial we randomly choose different support sets (source locations). We assign the number of sources to the values $k = N/4$ and $k = N/2$, which ensures that for almost all of the

parameter settings we have more sources than sensors. We create the data matrix as $\mathbf{Y} = \mathbf{AX}$. We perform 200 iterations of M-SBL and select the k largest elements of the converged hyperparameter vector γ^* as the support set. For a single trial, a success ratio of source localization is calculated as,

$$r = \frac{|s^* \cap s|}{k} \tag{2.21}$$

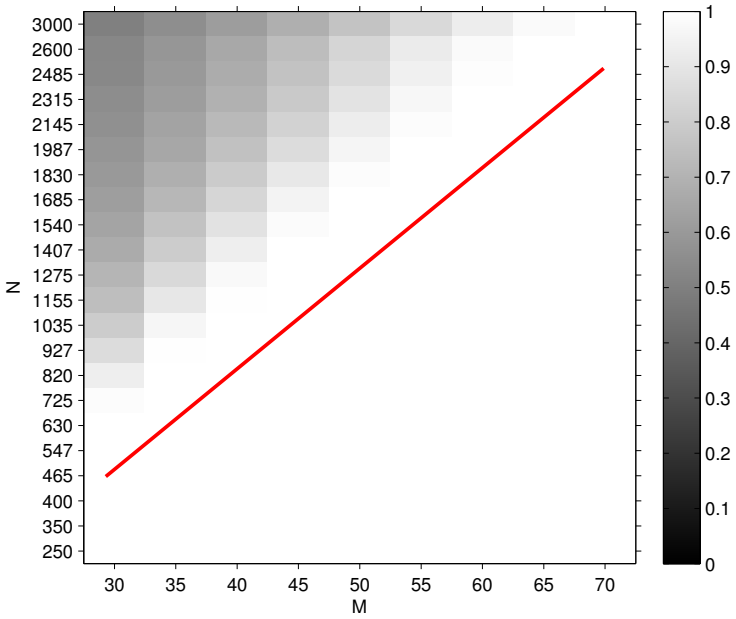
where s^* is the resulting estimated support set and s is the true support set. After performing 100 trials for each parameter setting, we calculate the mean support recovery ratio and plot it in Figure 2.1.

2.4.2 Performance Comparison

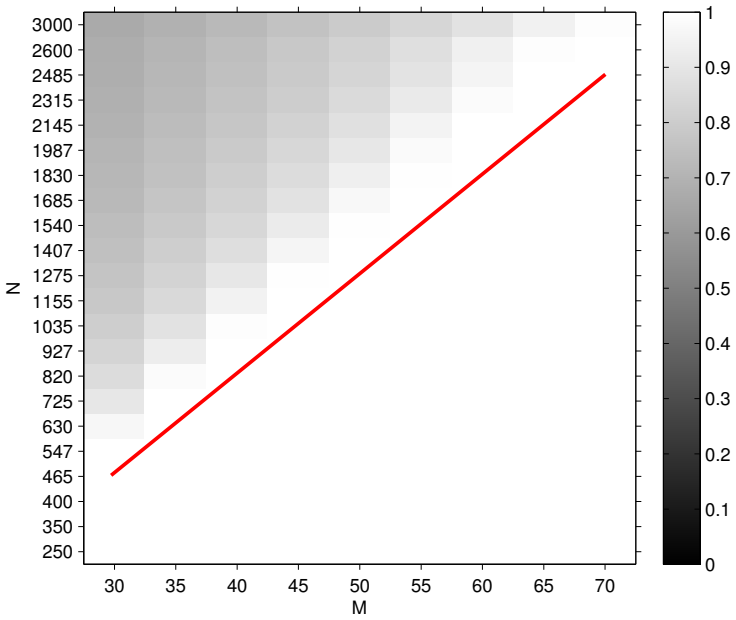
In this part, we compare the performance of M-SBL to M-CoSaMP and M-FOCUSS [21] when the conditions (i) and (ii) of Theorem 1 are violated. We create random sources of length $L = 1024$. The sources are independent yet the outer product is not exactly diagonal ($\mathbf{X_S X_S^T} \neq \Lambda$). The random dictionaries we create also violate (ii), namely $N \geq M(M+1)/2$.

Figure 2.2 shows the source recovery performance of all algorithms with varying resolution number N. The number of sensors $M = 20$ is fixed, and $k = N/5$. Figure 2.3, shows the performance of the algorithms with fixed M, N but varying number of sources k. In both cases we see that M-SBL outperforms the other algorithms. In Figure 2.4, we see the positive effect of having more snapshots L, on the recovery of locations for M-SBL, as $\mathbf{X_S X_S^T}$ approaches to a diagonal matrix.

Finally, we test the algorithms under various noise levels. Although our analysis is based on the noiseless case, we observe that M-SBL still outperforms M-CoSaMP in low noise as well as high noise scenarios. See Table 2.1.



(a) $k = N/4$.



(b) $k = N/2$.

Figure 2.1: Validation of Theorem 1. Sample-wise orthogonal sources. The line corresponds to $N = M(M+1)/2$. Below the line we have guaranteed support recovery. Perfect recovery may also observed above the line.

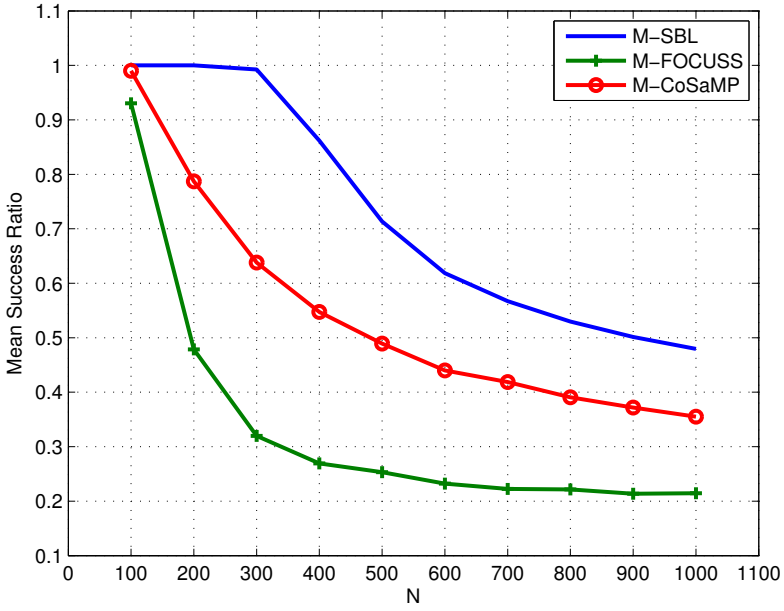


Figure 2.2: Comparison of algorithms: Changing resolution N, $M = 20, k = N/5, L = 1024$.

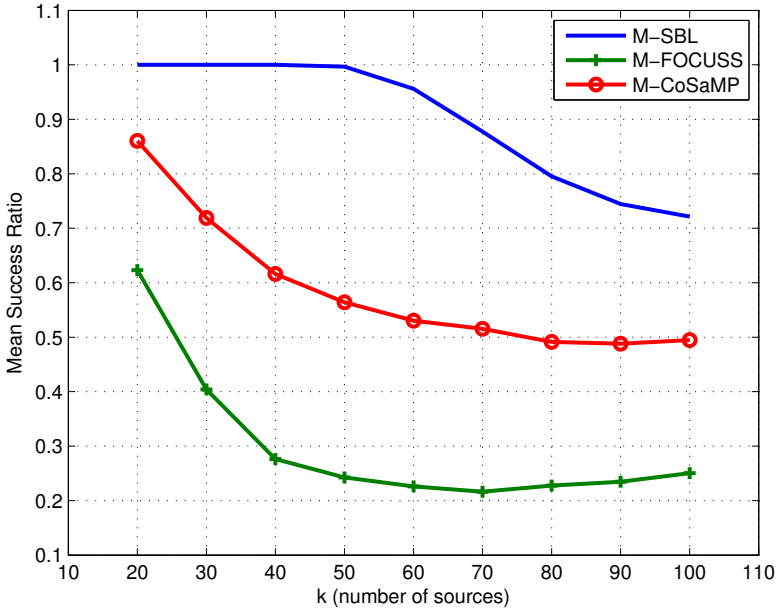


Figure 2.3: Comparison of algorithms: Varying the number of sources k, $M = 20, N = 500, L = 1024$.

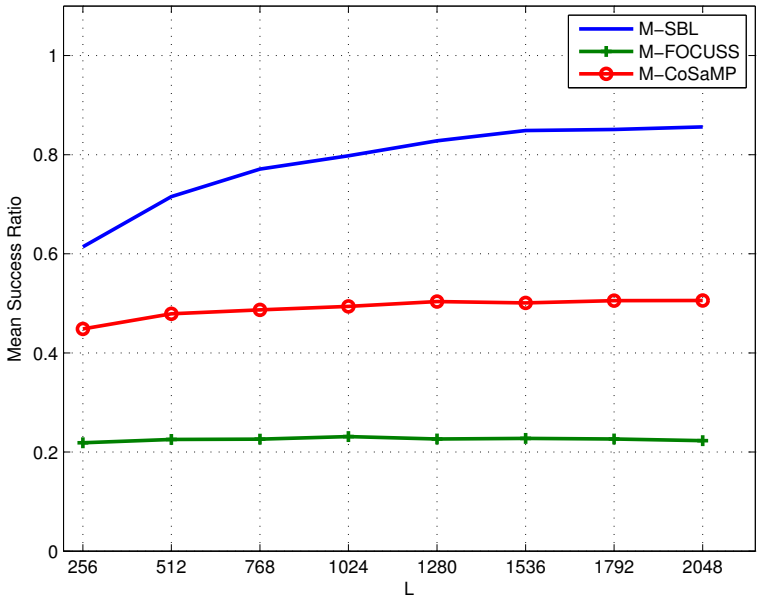


Figure 2.4: Comparison of algorithms: Varying the source duration L. Other parameters are fixed to $M = 20, N = 500, k = 80$.

Table 2.1: Mean Success Ratio under Different Noise Levels. $M = 20, N = 500, k = 100, L = 100$

SNR (dB)	M-CoSaMP	M-SBL
10	0.4601	0.5962
20	0.5018	0.8159
30	0.5028	0.9172
Inf	0.5068	0.9335

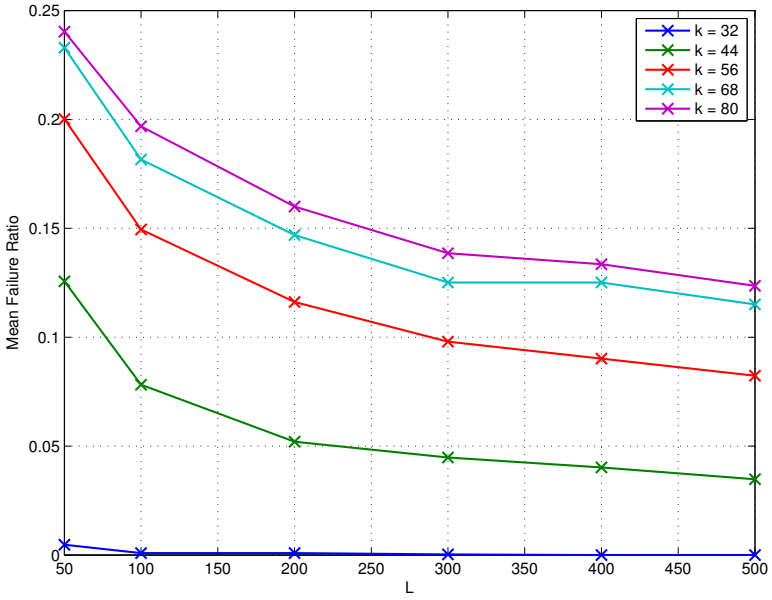


Figure 2.5: Mean failure ratio for support recovery of brain sources. $M = 32, N = 150$.

2.4.3 EEG source localization

Given a segment of EEG data as $\mathbf{Y}$, localizing the contributing sources in the brain can be done via solving the inverse problem $\mathbf{Y} = \mathbf{AX}$ in the MMV form, where a column $\mathbf{a_i}$ of $\mathbf{A}$ is the relative weights of the projection of i-th source to the scalp sensors, namely the lead-field matrix. To construct the dictionary, we sample N possible source locations within the brain (grid), and compute the projection of a dipole activity in each grid to the scalp electrodes. We construct our data by first choosing a subset of size k from N possible sources and assign random Gaussian time series for duration L, thus obtaining uncorrelated source matrix $\mathbf{X}$. We create scalp data by $Y = AX$. We aim to solve the inverse problem and find the support set even when more brain sources are active than the number of channels $k > M$. We test the performance of M-SBL for this type of data under various simulation parameters. See Figure 2.5

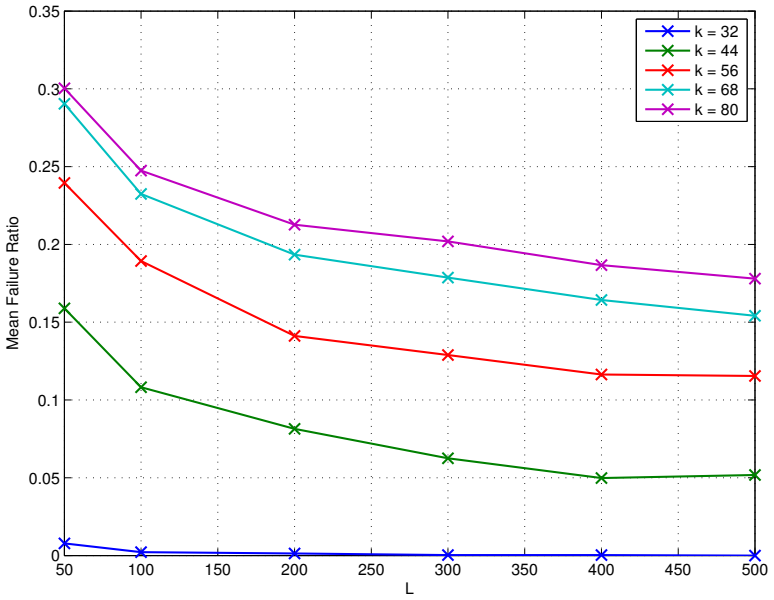


Figure 2.6: Mean failure ratio for support recovery of brain sources. $M = 32, N = 200$.

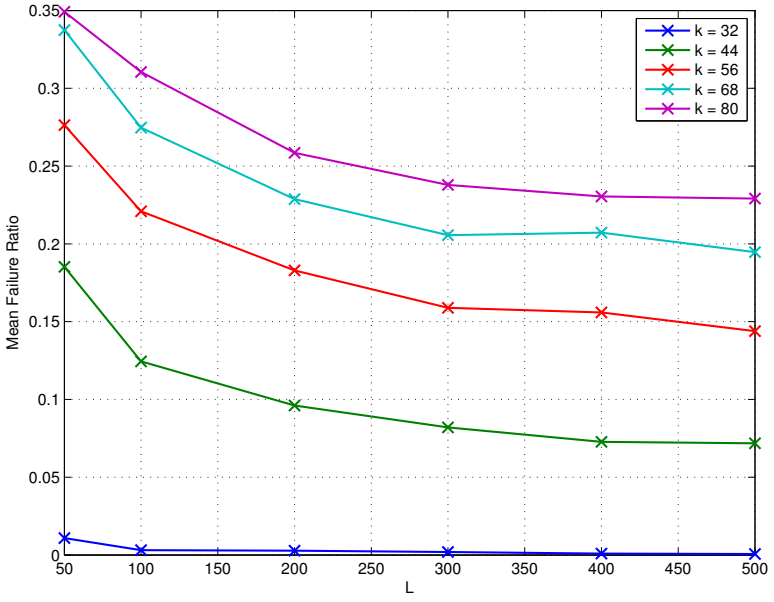


Figure 2.7: Mean failure ratio for support recovery of brain sources. $M = 32, N = 250$.

2.5 Conclusion

In this chapter, we showed that it is possible to perform localization (support recovery) of more sources than sensors using M-SBL applied directly on the time domain data. We theoretically provided sufficient conditions for exact recovery, conditions which were missing for previously suggested joint sparse algorithms in the case of more sources than sensors [53, 21]. Our experiments also demonstrated the superior performance of M-SBL when the conditions of guaranteed recovery of Theorem 1 do not hold.

2.6 Acknowledgements

The text of Chapter 2, in part, is a reprint of the material as it appears in: Balkan, Ozgur; Kreutz-Delgado, Kenneth; Makeig, Scott, "Localization of More Sources Than Sensors via Jointly-Sparse Bayesian Learning", Signal Processing Letters, IEEE 21, no. 2 (2014): 131-134. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Covariance-domain Dictionary Learning for EEG Source Identification

In this chapter, we propose an algorithm targeting the identification of more sources than channels for electroencephalography (EEG). Our overcomplete source identification algorithm leverages dictionary learning methods applied in the covariance-domain (Cov-DL). Assuming that EEG sources are uncorrelated in moving time-windows, and the scalp mixing is linear, forward problem can be transferred to the covariance domain which has higher dimensionality than the original EEG channel domain. This allows for learning the overcomplete mixing matrix that generates the scalp EEG, even when there may be more sources than sensors active at any time segment, i.e. non-sparse sources. This is contrary to straight-forward dictionary learning methods that leverage sparsity, which is not a satisfied condition in the case of low-density EEG systems. We present two different algorithms depending on the overcompleteness of the target dictionary. We demonstrate that Cov-DL outperforms existing overcomplete ICA

algorithms under various scenarios of EEG simulations and real EEG experiments.

3.1 Introduction

There are multiple reasons why an EEG source identification algorithm should be able to handle more sources than sensors. A main motivation is to increase the capabilities of EEG systems to handle large number of artifacts. Depending on the experiment settings and the length of recording, the number of distinct artifact sources could possibly outnumber the brain sources or even exceed the number of channels. In those cases, ICA solution matrix is occupied by artifact sources and only a few brain sources can be extracted from data, which limits further analysis of brain activity. Even in ideal conditions, i.e, when there are no artifacts, higher resolution is desired to better capture true brain dynamics, taking into account the possibility of more than M sources being simultaneously active and/or changing brain source locations throughout the experiment.

It is also desirable to enhance the capabilities of low-density EEG devices that are becoming increasingly popular due to their relative low-cost and ease of use. Low-density EEG allows for a wide range of applications by facilitating EEG recording of mobile and possibly long duration experiments. However, because they are targeted for low-cost research and consumer markets, these systems usually contain about 8-19 channels for which the results of traditional ICA results would be insufficient for reliable brain source monitoring. Extracting more sources than channels may benefit low-cost clinical research and improve consumer-oriented BCI applications.

Here, we propose a covariance-domain dictionary learning algorithm, Cov-DL, that can identify more sources than number of channels for the EEG inverse problem. We note that our algorithm does not learn the explicit source time-series activity $\mathbf{X}$ but

learns the overcomplete mixing matrix **A** (projection of sources to scalp sensors) and the power of individual sources in a given data segment. In this sense, our algorithm is categorically placed between blind source identification and source separation methods.

3.2 Related Work

An important family of blind source identification methods is comprised of cumulant-based algorithms that incorporate second order (SOBI) [12] or fourth order statistics (FOOBI) [25]. In non-EEG settings, it was shown that FOOBI can identify a number of sources that are roughly quadratic in the number of sensors [25]. However, multiple studies [28, 3] showed that cumulant-based methods perform relatively poorly in EEG source separation tasks compared to maximum likelihood based methods such as Infomax [11]. Among all methods, AMICA, an EM-based maximum likelihood ICA framework with flexible source densities, [74], performed best in terms of extracting the most number of plausible brain sources while providing the highest independence among sources [28].

An extension of traditional ICA for the overcomplete case is provided by the ICA mixture model [56, 74]. This approach learns N_{model} mixing matrices, $\mathbf{A}_i \in \mathbb{R}^{M \times M}$, instead of learning one overcomplete mixing matrix **A**, in order to provide tractable computation. An adaptation of this method with AMICA, Multiple Model AMICA, was shown to be successful in identifying more sources than electrodes in some non-stationary EEG paradigms [74]. However, the mixture model has some drawbacks; because it assumes that at most M sources are active at any given time and there are only a few disjoint sets of simultaneously active sources (N_{model}). This is problematic especially when M is low. An ideal algorithm should be able to handle cases where any of $\binom{N}{k}$ sources, $1 \leq k \leq N$, can be jointly active. Our algorithm targets this case.

Another set of overcomplete ICA algorithms [5, 52] model the source estimates as $\hat{\mathbf{X}} = \mathbf{WY}$, where $\mathbf{W} \in \mathbb{R}^{N \times M}$ is a tall unmixing matrix with full column rank. These algorithms optimize $\mathbf{W}$ and return the mixing matrix as $\mathbf{A} = \mathbf{W}^{\mathbf{T}}$. One of the recent algorithms of this type is RICA [52], an efficient method used for unsupervised feature learning in neural networks. We have found that in the complete mixing matrix case ($M = N$), RICA gives almost identical results with Infomax on EEG data. In this paper, we are considering the overcomplete setting for RICA and multiple model AMICA for comparison with our overcomplete method Cov-DL.

Dictionary learning-based sparse coding algorithms are closely related to overcomplete ICA methods. In the dictionary learning framework, the inverse problem is formulated as the following optimization problem,

$$\min_{A,X} \frac{1}{2} \sum_{t=1}^{N_d} \|\mathbf{Y} - \mathbf{AX}\|_F^2 + \lambda \sum_{t=1}^{N_d} g(\mathbf{x_t}) \tag{3.1}$$

where $g(\cdot)$ is a function that promotes sparsity of the source vector $\mathbf{x_t}$ at time index t and λ is the regularization parameter controlling the sparsity of the sources. Optimization is generally performed on $\mathbf{A}$ and $\mathbf{X}$ iteratively, namely learning $\mathbf{X}$ while keeping $\mathbf{A}$ fixed, and vice versa [2, 51]. Given a fixed dictionary $\hat{\mathbf{A}}$, the sources $\hat{\mathbf{X}}$ are learned by solving the following optimization

$$\min_{X} \frac{1}{2} \sum_{t=1}^{N_d} \|\mathbf{Y} - \hat{\mathbf{A}}\mathbf{X}\|_F^2 + \lambda \sum_{t=1}^{N_d} g(\mathbf{x_t}) \tag{3.2}$$

The true dictionary can be recovered if the sources $\mathbf{x_t}$ are sparse ($k_t < M$), where k_t is the number of active sources at time t. The accuracy of recovery is strongly dependent on the level of sparsity as higher accuracy is achieved if $k \ll M$. It was shown for various dictionary learning algorithms that the performance significantly drops as k approaches M [75]. Indeed, when $k \geq M$, any full-row rank dictionary can provide a

source decomposition with sparsity k and zero representation error $\|\mathbf{Y} - \mathbf{AX}\|_{\mathbf{F}}^{\mathbf{2}}$ for (3.1), thus the true mixing matrix becomes unrecoverable. In the case of EEG, this allows at most k = $O(M)$ EEG sources to be simultaneously active which limits direct applicability of dictionary learning to low-density EEG systems.

Recently it was shown that given the true dictionary $\mathbf{A}$, and a data segment $\mathbf{Y_s} \in \mathbb{R}^{M \times L_s}$, where L_s is the length of the segment in data frames, M-SBL (multiple measurement Sparse Bayesian Learning) applied directly on $\mathbf{Y_s}$ can identify active sources under the assumption that sources are uncorrelated in the time segment [6]. The number of sources identified in this case is not limited by the number of channels M, $1 \leq k \leq M(M+1)/2$. This finding is supported by [72], where LASSO is applied on the covariance matrix off the data segment $\mathbf{Y_s}$ to obtain probability bounds on the identification of active sources. Under the assumption of uncorrelated sources $\mathbf{X_s}$, the sample-covariance matrix $\frac{1}{L_s}\mathbf{X_s}\mathbf{X_s^T}$ is assumed to be nearly diagonal ("pseudo-diagonal") and expressible as $\Sigma_{\mathbf{X_s}} = \frac{1}{L_s}\mathbf{X_s}\mathbf{X_s^T} = \Delta + \mathbf{E}$, where Δ is a diagonal matrix composed of diagonal entries of Σ_{X_s}. Hence in [72], $\mathbf{Y_s} = \mathbf{AX_s}$ is modeled as

$$\begin{aligned} \mathbf{Y_s}\mathbf{Y_s^T} &= \mathbf{A}\mathbf{X_s}\mathbf{X_s^T}\mathbf{A^T} \\ \Sigma_{\mathbf{Y_s}} &= \mathbf{A}\Sigma_{\mathbf{X_s}}\mathbf{A^T} \\ \Sigma_{\mathbf{Y_s}} &= \mathbf{A}\Delta\mathbf{A^T} + \mathbf{E} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \Delta_{\mathbf{ii}}\mathbf{a_i}\mathbf{a_i^T} + \mathbf{E}. \end{aligned} \tag{3.3}$$

Since the covariance matrix is symmetric, we can vectorize the lower triangular part of

both sides and obtain,

$$\begin{aligned}
\text{vech}\left(\Sigma_{Y_s}\right) &= \sum_{\mathbf{i=1}}^{\mathbf{N}} \text{vech}\left(\mathbf{a_i a_i^T}\right)\Delta_{\mathbf{ii}} + \text{vech}\left(\mathbf{E}\right) \\
\text{vech}\left(\Sigma_{Y_s}\right) &= \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{d_i}\Delta_{\mathbf{ii}} + \text{vech}\left(\mathbf{E}\right) \\
\text{vech}\left(\Sigma_{Y_s}\right) &= \mathbf{D}\delta + \text{vech}\left(\mathbf{E}\right) \qquad (3.4)
\end{aligned}$$

where $\mathbf{D} = [\mathbf{d_1}, \mathbf{d_2}, \ldots, \mathbf{d_N}]$, $\mathbf{d_i} = \text{vech}\left(\mathbf{a_i a_i^T}\right)$ and vech$(\cdot)$ is a function that maps a symmetric matrix $S \in \mathbb{R}^{M\times M}$ to its vectorized lower triangular matrix, of size $\frac{M(M+1)}{2}$. Here, we also define the inverse function vech$^{-1}(\cdot)$, which takes as an input an $\frac{M(M+1)}{2}$ dimesional vector v and outputs a symmetric matrix of size $M \times M$ whose lower triangular matrix consists of entries in v. Thus, for any vector v, we have $v = \text{vech}\left(\text{vech}^{-1}\left(v\right)\right)$.

It was shown in [72] that this formulation, together with the correlation constraint (3.3) can identify $O(M^2)$ sources given the true dictionary. We leverage this idea to also learn the dictionary $\mathbf{A}$ from EEG data considering multiple segments from the overall recording. We also note that assumption of uncorrelated sources, albeit being a weaker constraint, is implied by the independence of sources, an assumption which was shown to be successful for EEG source separation [28].

3.3 Covariance-Domain Dictionary Learning (Cov-DL)

Here, we describe our covariance based dictionary learning algorithm that leverages the assumed uncorrelated nature of EEG sources. We start by segmenting the overall EEG data matrix $\mathbf{Y} \in \mathbb{R}^{M\times N_d}$, sampled with frequency S_f, into possibly overlapping segments $\mathbf{Y_s} \in \mathbb{R}^{M\times t_s S_f}$ of t_s seconds, where s denotes the index for the corresponding segment. For each segment, the following equation holds under the linear mixture model

of EEG,

$$\mathbf{Y_s} = \mathbf{A}\mathbf{X_s}, \forall s \tag{3.5}$$

and thus, $\mathbf{Y_s}\mathbf{Y_s^T} = \mathbf{A}\mathbf{X_s}\mathbf{X_s^T}\mathbf{A^T}$. Then, we calculate the sample data covariance $\Sigma_{\mathbf{Y_s}} = \frac{1}{L_s}\mathbf{Y_s}\mathbf{Y_s^T}$, for each segment s. We have,

$$\begin{aligned} \Sigma_{\mathbf{Y_s}} &= \mathbf{A}\Delta_{\mathbf{s}}\mathbf{A^T} + \mathbf{E_s} \\ \text{vech}\left(\Sigma_{\mathbf{Y_s}}\right) &= \sum_{\mathbf{i=1}}^{\mathbf{N}} \Delta_{\mathbf{s_{ii}}} \text{vech}\left(\mathbf{a_i}\mathbf{a_i^T}\right) + \text{vech}\left(\mathbf{E_s}\right), \\ \text{vech}\left(\Sigma_{\mathbf{Y_s}}\right) &= \mathbf{D}\delta_{\mathbf{s}} + \text{vech}\left(\mathbf{E_s}\right), \forall \mathbf{s}. \end{aligned} \tag{3.6}$$

where the vector δ_s contains the diagonal entries of the source sample-covariance matrix $\Sigma_{\mathbf{X_s}} = \frac{1}{L_s}\mathbf{X_s}\mathbf{X_s^T}$, and the matrix $\mathbf{D} \in \mathbb{R}^{M(M+1)/2 \times N}$ consists of columns $\mathbf{d_i} = \text{vech}\left(\mathbf{a_i}\mathbf{a_i^T}\right)$. Note that, for each segment, the left hand side of the equations are obtained from data while $\mathbf{D}$ and $\delta_{\mathbf{s}}$ are not known. Our goal is to first learn $\mathbf{D}$ and then find the associated matrix $\mathbf{A}$. We propose two different approaches to recover $\mathbf{D}$ and $\mathbf{A}$ which depend on the relation between the target number of total sources N and the number of channels M. See Fig. 3.1.

3.3.1 Overcomplete D (Cov-DL-1)

When N, the number of total sources to be identified for the whole EEG session, is larger than or equal to $M(M+1)/2$, $\mathbf{D}$ in (3.6) is overcomplete. If we assume that at any given segment s, there are less than $M(M+1)/2$ active sources, namely $\delta_{\mathbf{s}}$ is sparse, then we can learn $\mathbf{D}$ by applying traditional dictionary learning methods on the set of data points $\{\text{vech}\left(\Sigma_{\mathbf{Y_s}}\right), \forall s\}$. Note that, the sparsity constraint imposed here, that is $k < M(M+1)/2$ is much weaker than the traditional sparsity constraint $k < M$ and is

not necessarily violated when $k > M$.

After learning dictionary $\mathbf{D}$, we can find the mixing matrix $\mathbf{A}$ that generated $\mathbf{D}$ through the relation $\mathbf{d_i} = \text{vech}\left(\mathbf{a_i a_i^T}\right)$. For each column of the dictionary we optimize,

$$\min_{\mathbf{a_i}} \|\mathbf{d_i} - \text{vech}\left(\mathbf{a_i a_i^T}\right)\|_2^2 \tag{3.7}$$

or equivalently,

$$\min_{\mathbf{a_i}} \|\text{vech}^{-1}\left(\mathbf{d_i}\right) - \mathbf{a_i a_i^T}\|_F^2 \tag{3.8}$$

The global minimum for this optimization problem is $\hat{\mathbf{a}}_\mathbf{i} = \sqrt{\lambda_1} b_1$, where λ_1 is the largest eigenvalue of $\text{vech}^{-1}\left(\mathbf{d_i}\right)$, and b_1 is the associated eigenvector. For a visualization of the algorithm, see Fig. 3.2.Since there is a sign ambiguity in the learned basis vectors for dictionary learning algorithms, we run the above optimization for $\mathbf{d_i}$ and $-\mathbf{d_i}$ and choose a_i with the lowest cost.

3.3.2 Undercomplete D (Cov-DL-2)

3.3.3 Undercomplete D (Cov-DL-2)

When, $N < M(M+1)/2$, the data points $\{\text{vech}\left(\Sigma_{\mathbf{Y_s}}\right), \forall s\}$ live on or near a subspace of dimension N, which is spanned by the columns of $\mathbf{D}$. We denote this subspace as $\mathcal{R}(\mathbf{D})$. We can learn $\mathcal{R}(\mathbf{D})$ with methods such as Principal Component Analysis (PCA) without imposing any sparsity constraints on δ_s. However, the set of basis vectors $\mathbf{U}$ that a subspace learning algorithm, such as PCA, returns only guarantee $\mathcal{R}(\mathbf{D}) = \mathcal{R}(\mathbf{U})$, not $\mathbf{U} = \mathbf{D}$. Therefore, we can extract $\mathcal{R}(\mathbf{D})$ but there is an ambiguity about the basis vectors $\mathbf{D}$. Note, however, that we can enforce the conditions that the columns of $\mathbf{D}$ satisfy $\mathbf{d_i} = \text{vech}(\mathbf{a_i a_i^T})$ and also span $\mathcal{R}(\mathbf{U})$ as closely as possible. Furthermore, since

the projection operator for a given subspace is unique, namely $\mathcal{R}(\mathbf{D}) = \mathcal{R}(\mathbf{U})$ if and only if $\mathbf{D}(\mathbf{D^T D})^{-1}\mathbf{D^T} = \mathbf{U}(\mathbf{U^T U})^{-1}\mathbf{U^T}$, we can obtain $\mathbf{A}$ by solving the following optimization problem.

$$\min_{\mathbf{a_i}} \|\mathbf{D}(\mathbf{D^T D})^{-1}\mathbf{D^T} - \mathbf{U}(\mathbf{U^T U})^{-1}\mathbf{U^T}\|_F^2$$
$$\text{s.t} \quad \mathbf{d_i} = \text{vech}(\mathbf{a_i a_i^T}) \tag{3.9}$$

where U is learned through a subspace learning algorithm on data points $\{\text{vech}\left(\Sigma_{\mathbf{Y_s}}\right), \forall s\}$. We compute the above cost function's gradient w.r.t $\mathbf{A}$ using the chain rule and can minimize the cost function using quasi-Newton optimization methods. We emphasize that although $\mathbf{D}$ is not overcomplete in this case, the mixing matrix $\mathbf{A}$, which relates the cortical sources to the scalp EEG sensors, can still be complete or overcomplete. For a visualization of the algorithm, see Fig. 3.3.

3.3.4 Remarks

We provide some comments about important aspects of above described algorithms. First, notice that the number of data points that Cov-DL is trained on is substantially reduced because of segmenting and learning in the covariance-domain (there is now effectively one data point per segment). For example, if t_s is 4 seconds and sampling rate is 250Hz, the total number of data points used is $\frac{1}{1000}N_d$ if the segments are non-overlapping. The number of data points for Cov-DL will increase as the overlap ratio increases. However we have found that algorithm performance does not improve when the overlap ratio of consecutive segments increases beyond 0.5. The reduced number of data points in the Cov-DL-1 framework linearly speeds up the dictionary learning computation time and makes its application to EEG feasible.

The segment length t_s is an important parameter that affects the performance of

the algorithms. If the segment length t_s is short, the sample-covariance $\frac{1}{L_s}\mathbf{X_s}\mathbf{X_s^T}$ is no longer pseudo-diagonal and thus the derivation in (3.6) is not accurate. On the other hand, as t_s gets longer, the number of active sources in a segment increases (becomes less sparse), thus the performance of Cov-DL-1 will decrease. We have found that the choice $t_s \in [2,4]$sec. provides a good compromise in our experiments.

We also note that for both algorithms to succeed, the power of the individual sources in segments δ_s should not stay constant throughout the recording. This is required to ensure that $\mathbf{D}$ is identifiable for algorithm Cov-DL-1 and that the data points $\Sigma_{\mathbf{Y_s}}$ obtained by $\mathbf{D}\delta_\mathbf{s}$ fill the space spanned by $\mathbf{D}$ for Cov-DL-2. This requirement holds for most EEG sources, including event-related potentials/oscillations and eye/head movement related artifact sources. To the best of our knowledge, the only EEG source that has constant power across the whole recording is electronic noise/line noise. Yet, the characteristics of this source is available (a 50Hz/60Hz sine wave) and can be filtered in the pre-processing step of EEG analysis.

Finally we note that for algorithm Cov-DL-1, one can choose any dictionary learning algorithm for learning $\mathbf{D}$. Here, we use Bilinear Generalized Approximate Message Passing (BiGAMP-DL) [75], an EM-based bayesian dictionary learning method leveraging approximate message passing. This method has the advantage of automatically learning the sparsity level and signal-to-noise ratio (SNR). For Cov-DL-2, we have used the robust PCA method described in [40] to identify $\mathbf{U}$.

The success of the algorithm Cov-DL-1 is determined by the success of dictionary learning stage, namely if $\mathbf{d_i}$ are identifiable then so are $\mathbf{a_i}$. For Cov-DL-2, direct learning of the subspace spanned by basis vectors of the form $\text{vech}(\mathbf{a_i}\mathbf{a_i^T})$ may not have a unique global solution depending on the true generator basis vectors. However, our simulations showed that solution tends to be unique as N gets lower ($M < N < M(M+1)/2$). Moreover, the cost function (3.9) may have local minima. Therefore, we recommend

using several random initial points. This does not bring a lot of computational burden since this stage of the algorithm is already independent of the original data length and is very efficient.

3.4 Experiments

3.4.1 EEG Simulation

First we test our algorithm on three simulated data scenarios, for which we exactly know the ground truth mixing matrix $\mathbf{A}_{\text{true}}$. We simulate the placement of 32 electrodes on the scalp as shown in Fig. 3.5b. To generate the mixing matrix, we place dipolar sources in the brain using the Montreal Head Institute (MNI) head model. We assign random locations and random orientations for each dipole. Using the FieldTrip toolbox [71], we compute the projection weights of the i-th dipole to each channel (scalp maps) and obtain the true $\mathbf{a_i}$. See Fig. 3.5b. For realistic source activations $\mathbf{X}$, we generate an AR (auto-regressive) model via Source Information Flow Toolbox (SIFT) [27] under EEGLAB [26] and obtain super-Gaussian source activations of duration 66 minutes with 100Hz sampling rate. We choose a segment length t_s = 2sec. (200 frames) and scale the sources in each segment with a random weight uniformly assigned in the continuous interval [1,2] to model the possibly varying power dynamics of brain sources across the recording.

For the first scenario, we first test and compare algorithms for the case of a complete mixing matrix ($M = N$). We select $M = N = 32$, for an overcompleteness ratio of $N/M = 1$. We also let $k = N = 32$, so that all the sources are active in any given segment. We generate scalp EEG with $\mathbf{Y} = \mathbf{A}_{\text{true}}\mathbf{X}$ and apply Cov-DL-2 on $\mathbf{Y}$ with t_s = 2sec non-overlapping segments. The accuracy of the result is measured as the ratio of the number of scalp maps that are recovered (having correlation higher than 0.99 with

true scalp maps) to N. We compare our algorithm with the 1-model AMICA [74] and RICA [52].

For the second scenario, we have $M = 32, N = 64$, and overcompleteness ratio $N/M = 2$. We also let $k = N = 64$, and again all the sources are active in any given segment. We generate scalp EEG with $\mathbf{Y} = \mathbf{A}_{\text{true}}\mathbf{X}$ and apply Cov-DL-2 on $\mathbf{Y}$ with $t_s = 2$sec non-overlapping segments. We compare our algorithm with the overcomplete ICA method RICA [52] and concatenated dictionary obtained from multi-model AMICA ($N/M = 2$ models in this case) [74].

For the third scenario, we have $M = 8, N = 40, k = 10$, and overcompleteness ratio $N/M = 5$. In each segment a randomly selected k out of N sources are retained and $N - k$ sources are assigned no activation. At any given segment there are more active sources than channels ($k > M$) and the set of active sources are changing throughout the recording. We select $M = 8$ channels out of the 32 channels shown in Fig. 3.5b such that we uniformly cover the whole head. In this case, since $N \geq M(M+1)/2$ and $k < M(M+1)/2$, we use Cov-DL-1. We compare with the results obtained from RICA and 5-model AMICA ($N/M = 5$).

The results of three scenarios are shown in Fig. 3.6. It can be seen that when $M = N$, single model AMICA shows perfect source identification whereas Cov-DL performs slightly worse but still has accuracy of 0.9687 (recovers 31/32 components). This might be because fewer number of data points (number of segments) are fed to Cov-DL compared to AMICA and AMICA has the ability to model arbitrary source probability densities in an adaptive way. RICA performs the worst with an identification ratio of 0.9375 even under ideal complete conditions. This is likely due to the high coherence of the realistic mixing matrix, since other experiments showed that RICA demonstrates perfect recovery with random mixing matrices (a low coherence situation). In the overcomplete scenarios, we see that there is a drop in the performance of all

algorithms. AMICA and RICA perform poorly due to their differences in modeling the overcompleteness. Multi-model AMICA considers a mixture ICA model which has only few distinct states and can handle at most M sources active at a given time. RICA fits a super-Gaussian distribution to sources obtained as $\mathbf{WY}$ where $\mathbf{W}$ is a tall unmixing matrix. However, sources derived in this form cannot be truly independent simply due to the necessary linear dependence of the rows of a tall matrix. Cov-DL is free of these drawbacks of existing ICA algorithms, and can handle more sources than sensors without requiring sparsity of any form as opposed to traditional dictionary learning algorithms which prohibit their use when $k > M$.

3.4.2 Experiments on Real EEG

Unlike simulated EEG, the true mixing matrix for real EEG is not known beforehand. In order to test our algorithm's performance on real EEG data, we follow the strategy proposed below.

Suppose we have an actual dataset that has M_{orig} channel recordings. After rejection of the artifact windows and contaminated channels, suppose that N channels remain. Then, we apply Extended Infomax ICA and extract N sources and their associated scalp maps. We regard these scalp maps as ground truth mixing matrix and measure how well the proposed algorithms recover these scalp maps from using only a subset of M channels out of N ($M < N$). We choose M channels in a spatially uniform manner as in the previous section. We compare algorithms on 3 different types of datasets; 1) EEGLAB sample data, 2) a Motor Imagery task, 3) a Arrow Flanker task. The results are shown in Figures 3.7,3.8,3.9. The segment length for Cov-DL is t_s = 2sec, with an overlap ratio of 0.5 between consecutive segments. We plot the the sorted correlation values of resulting scalp maps with the best column match in the ground truth mixing matrix. In all 3 datasets, Cov-DL shows consistently higher correlations than multi-model

AMICA and RICA. We also plot the correlation results of complete extended Infomax applied on M channels to show the importance of overcomplete approaches for accurate source identification in low-density EEG systems.

3.5 Conclusion

We proposed a dictionary learning framework, Cov-DL, that incorporates the presumed uncorrelated nature of EEG sources, which is a related but a weaker assumption than EEG source independence [64, 28]. Identification of the mixing matrix is carried to a higher dimensional covariance-domain, which enables source identification even if the number of sources active at any time is larger than the number of sensors - sparsity is not required. We proposed two different algorithms which depend on the relation between the number of sources targeted and number of sensors available. We have shown that the proposed algorithm Cov-DL is more successful than existing overcomplete ICA algorithms for finding the true generating matrix in EEG simulations. We have also demonstrated the power of Cov-DL on real data. The proposed algorithm, because of its ability to provide higher resolution than the number of sensors, can potentially increase the applicability of low-cost, low-density EEG systems in biomedical research.

3.6 Acknowledgements

The text of Chapter 3, in full, is a reprint of the material as it appears in: Balkan, Ozgur; Kreutz-Delgado, Kenneth; Makeig, Scott, "Covariance-domain Dictionary Learning for Overcomplete EEG Source Identification" submitted to IEEE Transactions on Biomedical Engineering. The dissertation author was the primary investigator and author of this paper.

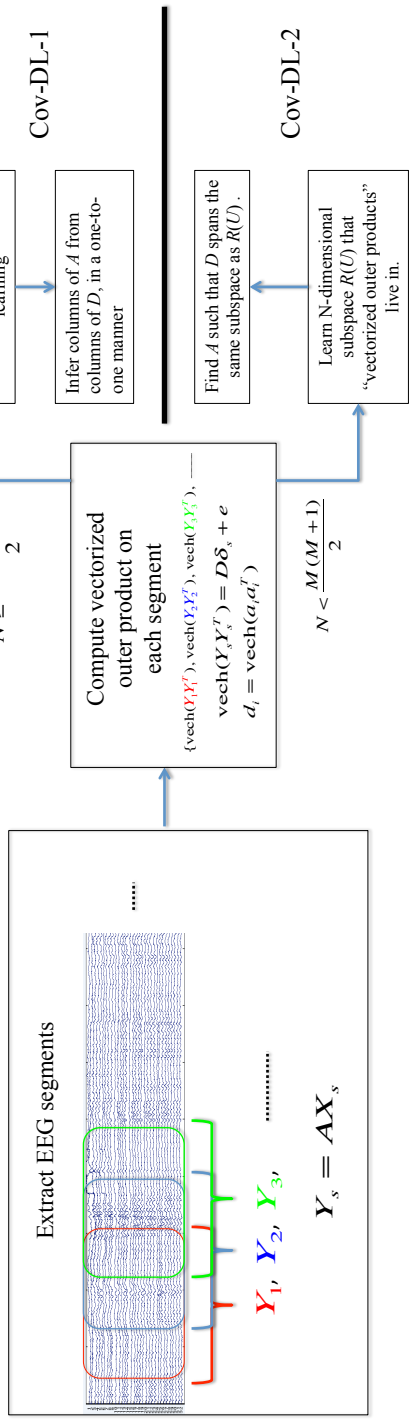


Figure 3.1: The summary of two different strategies of Cov-DL for overcomplete EEG source identification. Cov-DL-1 involves dictionary learning stage thus $k < M(M+1)/2$ sources are assumed to be active at any given segment. Cov-DL-2 does not require sparsity of sources.

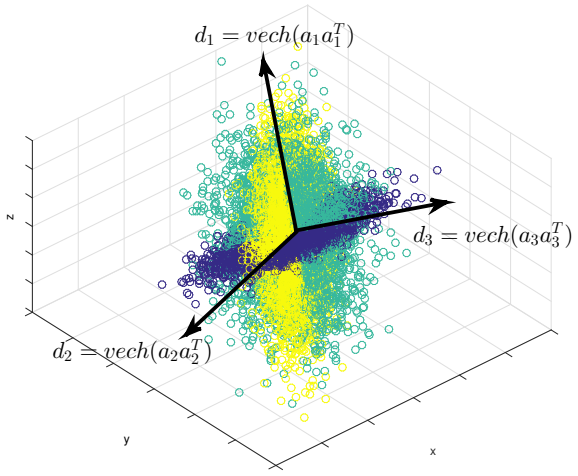


Figure 3.2: A geometrical explanation of Cov-DL for $M = 2, k = 2$. If $N = 3$, then $\mathbf{A} \in \mathbb{R}^{2\times3}$, and $\mathbf{D} \in \mathbb{R}^{3\times3}$. In this case $\mathbf{d_1}, \mathbf{d_2}, \mathbf{d_3}$ are identifiable with a DL algorithm applied on the data of vectorized outer products. Associated $\mathbf{a_1}, \mathbf{a_2}, \mathbf{a_3}$ can then be found via solving Eg. (3.8).

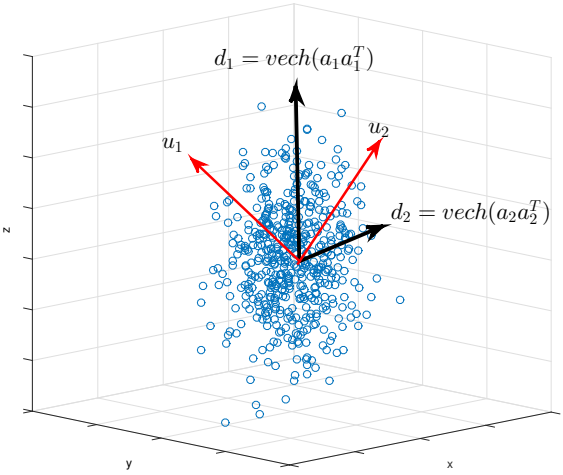
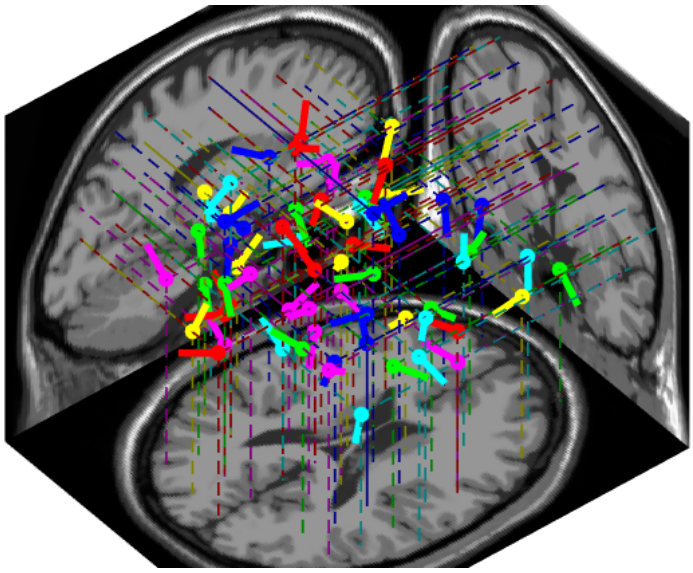
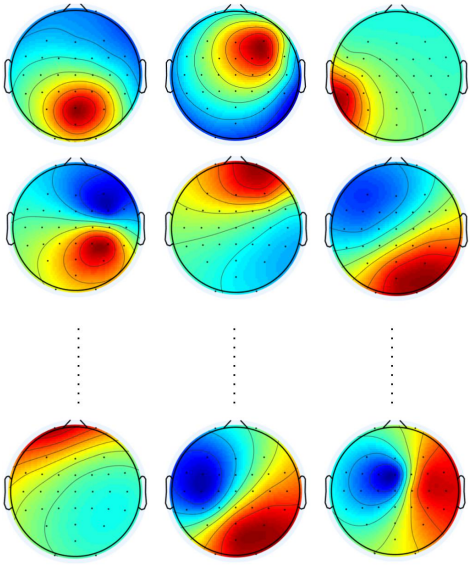


Figure 3.3: A geometrical explanation of Cov-DL for $M = 2, k = 2$. If N = 2, then $\mathbf{D} \in \mathbb{R}^{3\times2}$, and data is not sparse since $k = N = 2$. $\mathbf{D}$ is not identifiable through learning the 2-dimensional subspace. We solve Eq. (3.9) to directly find $\mathbf{A}$ such that $\mathbf{D}$ will span $\mathcal{R}(\mathbf{U})$ (Cov-DL-2).

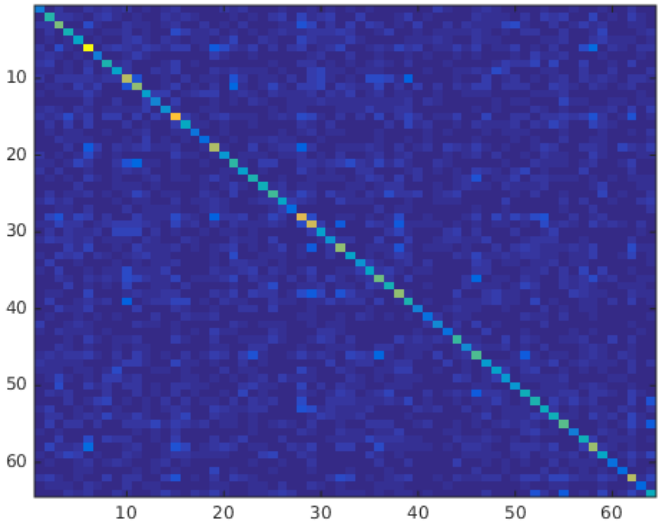


(a)

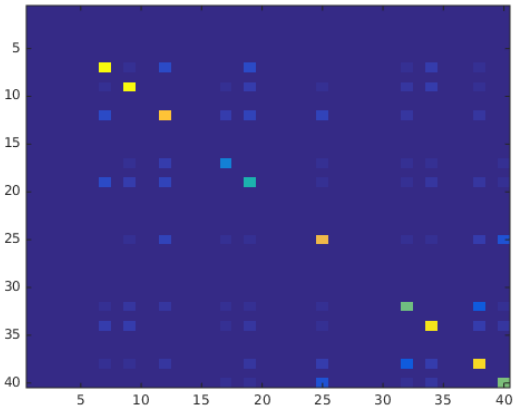


(b)

Figure 3.4: (a) Randomly located and oriented $N = 64$ dipoles/sources in the MNI head model that generate the simulated EEG. (b) Some of the scalp maps associated with the dipoles in (a). These constitute columns of true mixing matrix $\mathbf{A}_{\text{true}} \in \mathbb{R}^{32\times 64}$.



(a)



(b)

Figure 3.5: (a) Outer product of the source matrix in a 2sec. segment (sample-covariance) from Scenario 1; $M = 32, N = 64$, $k = 64$. (b) Outer product of the source matrix in a 2sec. segment from Scenario 2; $M = 8, N = 40$. $k = 10$ sources are active at any given segment.

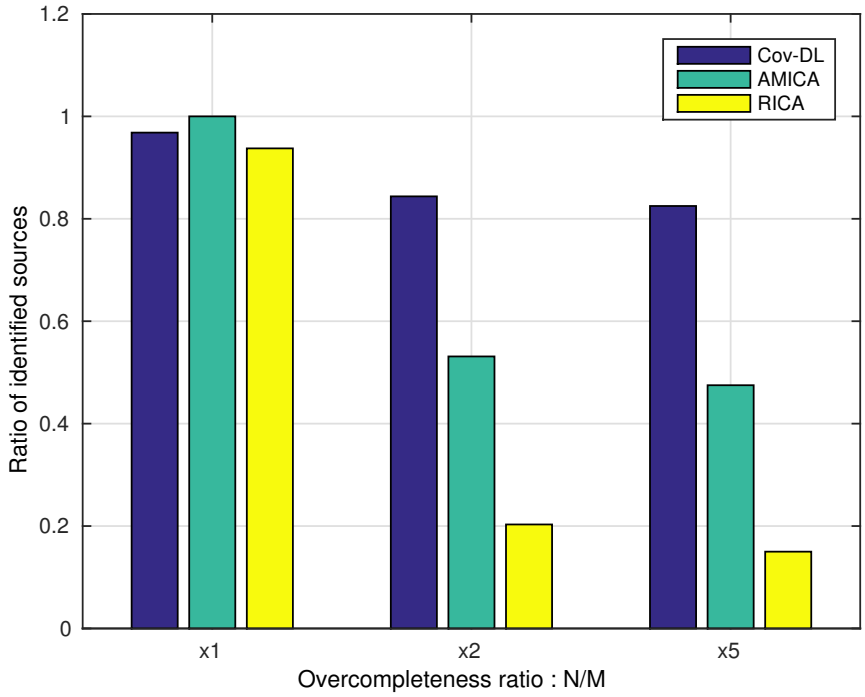


Figure 3.6: Simulation results for three cases: complete, x2 overcomplete, x5 overcomplete. Complete case: $M = 32, N = 32, k = 32$, Cov-DL-2 is used. Second case: $M = 32, N = 64, k = 64$, Cov-DL-2 is used. Third case: $M = 8, N = 40, k = 10$, Cov-DL-1 is used.

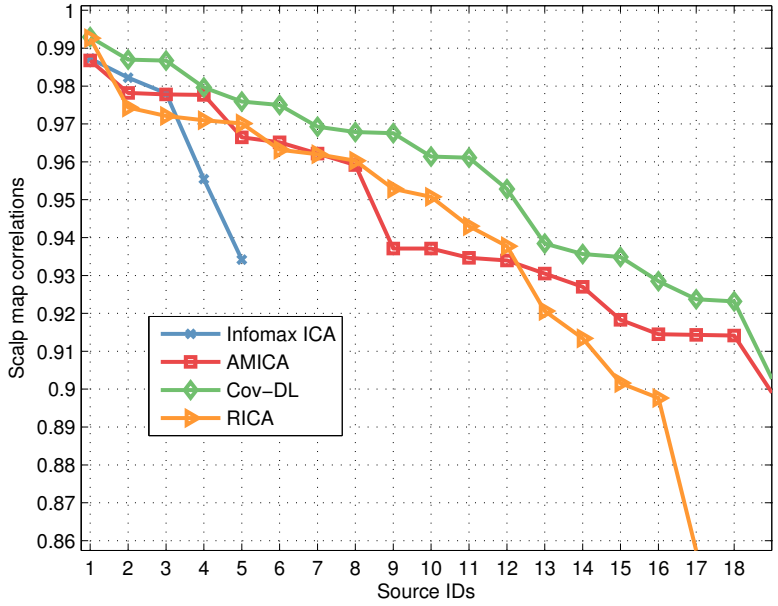


Figure 3.7: EEGLAB sample data. $M = 5, N = 30$. AMICA is trained with 6 models. Cov-DL-1 is performed.

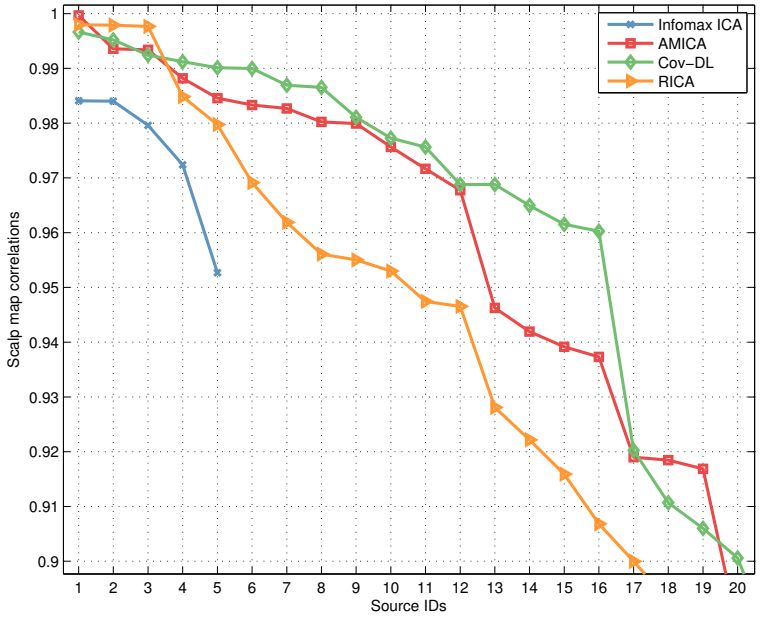


Figure 3.8: Motor Imagery Task, $M = 5, N = 30$. Cov-DL-1 is performed.

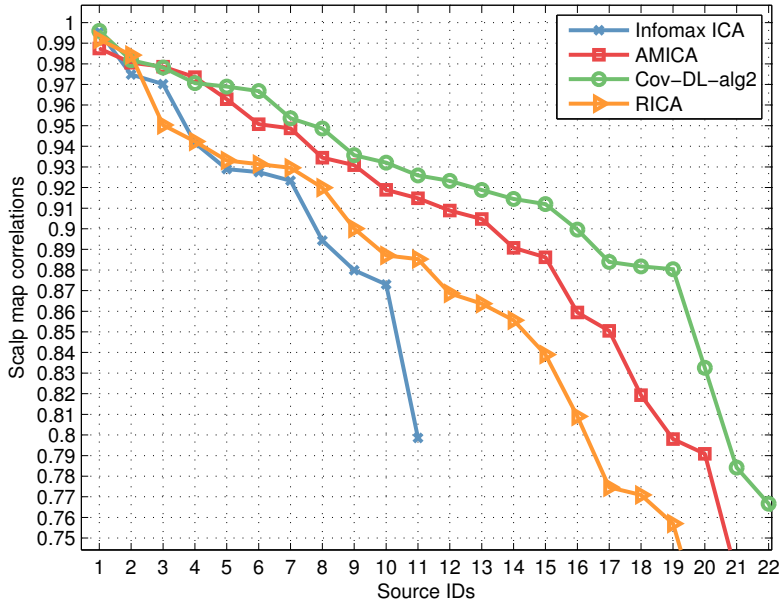


Figure 3.9: Arrow Flanker task. $M = 11, N = 30$. AMICA is trained with 3 models. Cov-DL-2 is used.

Chapter 4

Basis Selection for Independent Sources

In this chapter, we pose and suggest a solution to the following problem: "Given multichannel linear source mixture data of n points $\mathbf{Y} = \{y_1, y_2, \ldots, y_n\} \in \mathbb{R}^M$, and an overcomplete dictionary $\mathbf{A}$ of basis vectors $\mathbf{a_i} \in \mathbb{R}^M$, how can we construct a complete basis $\mathbf{A_0}$ by selecting columns from $\mathbf{A}$ such that the sources $\mathbf{X} = \mathbf{A_0^{-1}Y}$ are statistically as independent as possible from each other?". While conventional independent component analysis (ICA) methods find the mixing matrix $\mathbf{A_0}$ from scratch given $\mathbf{Y}$, we restrict ourselves to selecting basis vectors from a known set of an overcomplete dictionary. Our aim differs from the previous simultaneous sparse signal recovery literature in the sense that we consider solutions with a basis set of size $k = M$ instead of $k < M$. We develop two methods named BASICA and BASRICA which result from the modifications of the maximum likelihood equivalent of the Infomax approach and reconstruction-ICA (RICA), respectively. We show the relation of our algorithms to multiple measurement sparse bayesian learning (M-SBL) in the noiseless case, showing our methods include a larger family of algorithms of which noiseless M-SBL is a special case.

4.1 Introduction

Independent Component Analysis (ICA) has been used in many different contexts and has found a vast number of applications in diverse fields of engineering, including but not limited to blind source separation, neural networks, and biomedical source localization [20]. The common underlying model in these applications is given by

$$\mathbf{Y} = \mathbf{AX}, \tag{4.1}$$

where $\mathbf{Y} \in \mathbb{R}^{\mathbf{M}\times\mathbf{n}}$ is a data matrix in which each column $\mathbf{y_t}$ is a data vector, $\mathbf{A} \in \mathbb{R}^{\mathbf{M}\times\mathbf{N}}$ is the mixing matrix, and $\mathbf{X} \in \mathbb{R}^{\mathbf{N}\times\mathbf{n}}$ is composed of N component (source) activations. ICA aims to find the unknown mixing matrix $\mathbf{A}$ such that the associated source activations (rows of $\mathbf{X}^* = \mathbf{A^{-1}Y} = \mathbf{WY}$) are (maximally) statistically independent of each other. For simplicity, it is usually assumed that the mixing matrix is square, although extentions to other cases are also possible [20, 88]. Learning is usually performed without restrictions on $\mathbf{A}$. Although different constrained ICA algorithms have been considered in the past [61], the constraint we use here differs from previous efforts.

Unlike common ICA algorithms, we restrict ourselves to selecting basis vectors from a pre-designated set with the goal of finding maximally independent sources. To the best of our knowledge, previous efforts for "source separation on a fixed dictionary" focused on finding a sparse solution for source activations $\mathbf{X}$ [94, 58, 108], instead of aiming for the maximally independent solution. The dictionary of basis vectors can be known *a priori*, or can be designed (constructed) in different ways depending on the application, e.g., lead-field matrix for EEG. Although our constraint gives less freedom in learning the mixing matrix, it has some advantages over standard ICA approaches. In particular, since the solution space is finite (although combinatorially large), our approach requires less data than standard ICA to find the true mixing matrix, as we demonstrate in

our experiments.

One of the biggest problems with current simultaneous sparse approximation approaches (joint sparsity or MMV) is that the number of active sources can be assumed to be sparse but the sparsity level k (the number of active sources) is usually unknown, as in EEG/ECoG (electrocorticography) problems. Given an EEG segment and a dictionary of possible sources, solving for a simultaneous sparse solution is therefore problematic. Our approach handles this problem by choosing M columns from the dictionary, and evaluating how much mutual information is reduced by projecting data onto the achieved complete basis set. Such a measure of independence is needed to assess how well we recover the unknown sources when we have no knowledge of the sparsity of the sources other than the assumption of maximal statistical independence among them. For most dictionaries, any randomly selected M columns from the dictionary would span the space that data lives in, but resulting sources would not necessarily be independent.

As in the case of joint sparse signal recovery, the problem that we pose is a non-trivial, NP-hard combinatorial search problem for which one might need to first consider $\binom{N}{M}$ basis sets, and then choose the one that gives maximal independence among the resulting source activations, measurable via mutual information reduction as defined in [28]. Here, we provide polynomial time, sub-optimal algorithms that we show can perform sufficiently well in many cases. It is important to note that the solution we are after selects M columns of the dictionary, and therefore differs from the traditional simultaneous sparse approximation goal to choose $k \ll M$ columns from the dictionary. Still, connections to simultaneous sparse signal recovery algorithms exist as we discuss the relationship of our algorithms to multiple measurement sparse bayesian learning (M-SBL). M-SBL and related automatic relevance determination methods assume a non-sparsity inducing Gaussian prior on the source activation densities yet still achieve sparse solutions [98]. The drawback of these algorithms is that the active source densities

in real data usually do not fit well with the Gaussian assumption. Palmer et al. suggested a variational method to handle non-Gaussian source priors [73], which is, however, tailored for the single data vector case $n = 1$, and thus not applicable to our problem. Our algorithms suggest a direct way to incorporate arbitrary differentiable priors into the formulation as in the standard ICA framework.

The outline of the chapter is as follows: Section 4.2.2 formulates the conventional maximum likelihood (ML) ICA algorithm and modifies it to derive BASICA selection framework. Section 4.2.3 derives BASRICA algorithm using the reconstruction ICA (RICA) formulation. In Section 4.3.1, we point out connections of both algorithms to M-SBL [97]. Section 5.4 performs tests on synthetic and real data.

4.2 Methods

4.2.1 Greedy Methods

First, we suggest a few greedy approaches to tackle the problem of basis selection for independence of the sources. As in the case of greedy methods proposed for sparse signal recovery [67, 70, 91, 22], we will construct our solution $\mathbf{A_0}$ (basis set) by adding basis vectors to the set one by one. We propose 3 methods to do so. But first, we define the measure of independence we will use (Mutual Information Reduction) inside these greedy algorithms.

$$\begin{aligned}\mathrm{MIR}_Y(W) &= I(x) - I(y) \\ &= \sum_{i=1}^{M} h(x_i) - \sum_{i=1}^{M} h(y_i) - h(x) + \dots \\ &\quad \dots + \log|\det W| + h(x) \\ &= \log|\det W| + \sum_{i=1}^{M} h(x_i) - \sum_{i=1}^{M} h(y_i). \end{aligned} \tag{4.2}$$

We describe 3 greedy algorithms that make use of MIR as a decision criteria. MIR is only computable if W and thus A is square. Therefore, unlike greedy sparse recovery methods, at each step we need a complete basis set to measure MIR. Here, we propose 3 different methods to construct the complete mixing matrix at each step.

Greedy RICA (gRICA): Assume that at iteration k, we have already constructed an $M \times k$ basis set $\mathbf{A_0}^{(k-1)}$. We select a basis vector $\mathbf{a_i}$ from the dictionary atoms that is not used in $\mathbf{A_0}^{(k-1)}$ and temporarily construct $\mathbf{A_0}' = [\mathbf{A_0}^{(k-1)}, \mathbf{a_i}, \mathbf{X}]$, where $\mathbf{X} \in \mathbb{R}^{M \times (M-k)}$ is unknown. While keeping the first k columns of the matrix $\mathbf{A_0}'$ constant, we learn best $\mathbf{X}$ using the cost function of an ICA method applied on data $\mathbf{Y}$. (we use RICA [52] for speed/accuracy tradeoff). We do this for all columns of the atom, calculate the MIR (4.2) resulting from each ICA and choose the atom that gives maximum MIR in the k-th step. Then, we move on to the next iteration.

Greedy PCA (gPCA): Since the number of ICA optimizations gRICA performs is $O(MN)$, it is computationally very costly and can be easily unfeasible when the dictionary gets large. To reduce this drawback, we suggest gPCA which uses PCA to find the complementary basis $\mathbf{X}$ instead of ICA as in gRICA. The speed of this method is substantially faster than gRICA with the cost of decreased accuracy.

Identity Complement Basis (gEye): Instead of optimizing the complementary

basis $\mathbf{X}$, we can assume that it is also fixed, e.g, last $M - k$ columns from the identity matrix. At each step k, we only compute MIR achieved with the possible inclusion of $\mathbf{a_i}$ in the solution set, and choose a_i with maximum MIR.

We will test the effectiveness of these algorithms with simulations. First experiment deals with the case of a random dictionary $\mathbf{A}$. In each trial, we have randomly selected 10 sources from the dictionary and generated the data $\mathbf{Y}$ using these sources. We have generated source activations of 1000 data points sampling from generalized Gaussian densities ($p = 1.6, super - Gaussian$).

In our second experiment, we have replaced second half of the dictionary with the inverse sphering matrix of the data $\mathbf{Y}$. This creates a more coherent dictionary compared to a random dictionary and also confuses algorithms that are do not directly optimize independence but orthogonality (uncorrelatedness). We have performed 100 trials for each experiment and calculated the mean failure ratio for support recovery. See results in Figure 4.1.

gRICA performs very well in both cases. However, one major drawback of this method is its computation time (5min. for a single trial of length $L = 1000$. $M = 10, N = 20$). As M and N increase number of ICA computations will increase with $O(MN)$ but also each ICA computation will be more expensive. Therefore, these greedy methods are not feasible for EEG applications and we need more efficient algortihms. Here, we suggest 2 non-greedy algorithms that are derived by modifying existing ICA algorithms.

4.2.2 BASICA

It was shown by [19] that the Infomax approach to the ICA problem is equivalent to the maximum likelihood formulation of data. The likelihood of the data, which is to

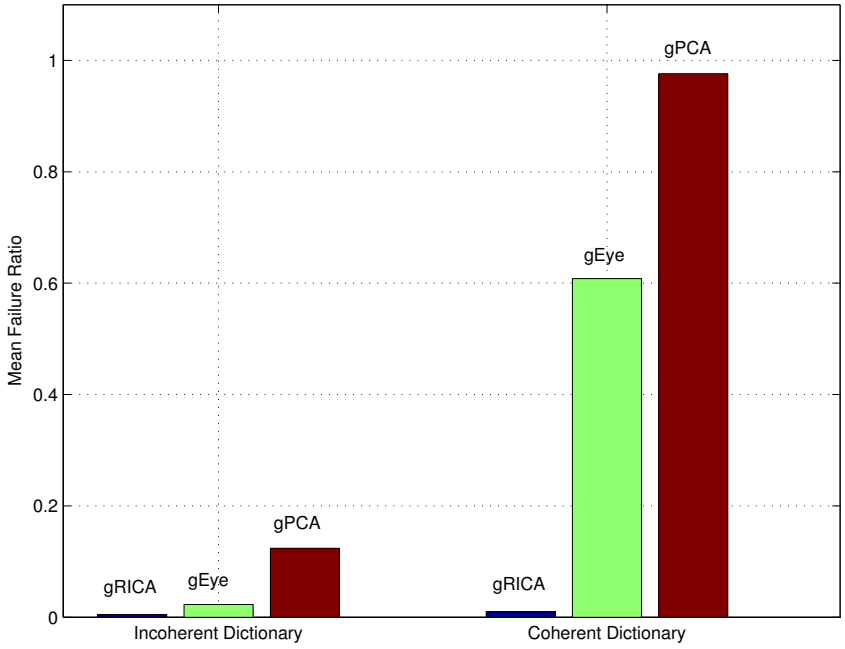


Figure 4.1: $M = 10, N = 20, L = 1000$. gRICA performs the best in both cases. Average computation time for gRICA is 303 seconds for one trial.

be maximized, can be expressed as

$$p(\mathbf{Y}) = \prod_{t=1}^{n} |\det \mathbf{A^{-1}}| p_s(\mathbf{A^{-1}y_t}) = \prod_{t=1}^{n} \frac{1}{|\det \mathbf{A}|} p_s(\mathbf{A^{-1}y_t}) \tag{4.3}$$

where $p_s(.)$ is the vector source density function. When $\mathbf{A}$ or $\mathbf{W}$ is not invertible, the nonexistence of $|\det \mathbf{W}|$ is handled by the substitution of $|\det \mathbf{WW^T}|^{\frac{1}{2}}$ in the undercomplete case [88], which becomes equal to $|\det \mathbf{W}|$ when $\mathbf{W}$ is complete. Similar term for $|\det \mathbf{A}|$ would be $|\det \mathbf{AA^T}|^{\frac{1}{2}}$ when $\mathbf{A}$ is overcomplete. Our goal is to select among columns of $\mathbf{A}$ the basis set which provides independence among resulting source activations. In order to be able to use the general maximum likelihood ICA framework, we model the generating matrix as $\mathbf{A_{gen}} = \mathbf{A}\Gamma^{\frac{1}{2}}$, a weighted selection of columns of $\mathbf{A}$[1]. The weight matrix $\Gamma = \text{diag}(\gamma)$ is a diagonal matrix of size $N \times N$ with nonnegative unknown values on the diagonal which we will learn, with the idea that at the end of the learning phase we are going to achieve M nonzero elements in γ, therefore effectively selecting M columns of the dictionary $\mathbf{A}$. We denote the matrix of nonzero columns of $\mathbf{A_{gen}}$ as $\mathbf{A}_0$.

Adopting the methodology proposed in [88], we substitute $|\det \mathbf{A}|$ in (4.3) with $|\det \mathbf{A}\Gamma^{\frac{1}{2}}(\mathbf{A}\Gamma^{\frac{1}{2}})^{\mathbf{T}}|^{\frac{1}{2}} = |\det \mathbf{A}\Gamma\mathbf{A^T}|^{\frac{1}{2}}$. One should note that when γ has M nonzero values, $|\det \mathbf{A}\Gamma\mathbf{A^T}|^{\frac{1}{2}}$ is equal to the determinant of the underlying forward model $\mathbf{A}_0$, and if γ has more than M nonzero values, the determinant still exists. Here, we assume that $\text{rank}(\mathbf{Y}) = M$, so at least M columns of $\mathbf{A}$ will be needed to explain the data, which ensures at least M elements of γ will be nonzero.

The backward transformation to the source domain is given by

$$\mathbf{x_t} = \mathbf{A_{gen}^{\dagger}y_t} = (\mathbf{A}\Gamma^{\frac{1}{2}})^{\dagger}\mathbf{y_t}, \tag{4.4}$$

where $A^{\dagger}$ denotes the pseudoinverse of the noninvertible matrix A. Here, we should

[1]The square root weighting $\Gamma^{\frac{1}{2}}$ is used instead of Γ to make the connection to M-SBL more obvious

note that although the pseudoinverse mapping to the source domain for overcomplete dictionaries is not always used to perform the ML overcomplete ICA framework [57], it suits well for our problem, which is rather different than the conventional overcomplete ICA. Since our goal is to find a basis set of size $M \times M$ instead of an overcomplete one, $\mathbf{x_t} = (\mathbf{A}\Gamma^{\frac{1}{2}})^{\dagger}\mathbf{y_t}$ serves to compute the regular inverse on the selected M columns of $\mathbf{A}$ in a weighted manner and place it at the corresponding locations in N-dimensional $\mathbf{x_t}$ vector, such that $\mathbf{x_t}$ has $N - M$ zero values. With this source mapping, we can write the data likelihood for our problem as,

$$p(\mathbf{Y}) = \prod_{t=1}^{n} \frac{1}{|\det \mathbf{A}\Gamma\mathbf{A^T}|^{\frac{1}{2}}} p_s((\mathbf{A}\Gamma^{\frac{1}{2}})^{\dagger}\mathbf{y_t}) \tag{4.5}$$

$$p(\mathbf{Y}) = \prod_{t=1}^{n} \frac{1}{|\det \mathbf{A}\Gamma\mathbf{A^T}|^{\frac{1}{2}}} p_s(\Gamma^{\frac{1}{2}}\mathbf{A^T}(\mathbf{A}\Gamma\mathbf{A^T})^{-1}\mathbf{y_t}) \tag{4.6}$$

Note that $p_s(\mathbf{x_t})$ is a vector source distribution and can be decomposed as $p_s(\mathbf{x_t}) = \prod_{i=1}^{N} p_{s_i}(\mathbf{x_{ti}})$ under the independence formulation. Here, we choose an analytical source density function $p_{s_i}(.)$ to be able to explicitly write and optimize (4.6). We use the super-Gaussian density $p_{s_i}(x) = c\,\mathrm{sech}(x)$, which has been shown to be suitable for EEG sources in the past [41]. Moreover, the use of a super-Gaussian source density also enhances the selection property of the algorithm, namely the convergence to sparse γ and equivalently large number of zero rows of sources $\mathbf{X}$. It should also be emphasized that although the actual sources might have arbitrary variances, we fix $p_{s_i}(x) = C\mathrm{sech}\,(x)$ for each source, with zero mean and a fixed variance. The zero mean condition is easy to satisfy by removing the mean of the data, and we are able to allow fixed variance for each source s_i due to the source equation $\mathbf{x_t} = (\mathbf{A}\Gamma^{\frac{1}{2}})^{\dagger}\mathbf{y_t}$ in (4.6), i.e. the actual source variances are embedded in Γ. Sources with higher variance will have higher γ_i.

Separating the vector source distribution in (4.6) into individual scalar source distributions and taking the $-2\log(.)$ transformation of likelihood gives the following,

which is to be minimized:

$$\begin{aligned} L(\gamma) =& n \log |\det \mathbf{A}\Gamma\mathbf{A}^{\mathbf{T}}| \\ & - 2 \sum_{t=1}^{n} \sum_{i=1}^{N} \log p_{s_i}(\gamma_{\mathbf{i}}^{\frac{1}{2}} \mathbf{a}_{\mathbf{i}}^{\mathbf{T}} (\mathbf{A}\Gamma\mathbf{A}^{\mathbf{T}})^{-\mathbf{1}} \mathbf{y}_{\mathbf{t}}). \end{aligned} \tag{4.7}$$

The vector $\mathbf{a_i}$ is the i-th column of $\mathbf{A}$. We minimize the above quantity over γ, which is the only unknown in the model. This approach involves computing the gradient of $L(\gamma)$ with respect to γ and rearranging the terms to achieve the following fixed point update.

Derivation of the update rule for γ starts with taking the gradient of $L(\gamma)$ w.r.t γ_i. For ease of notation, we separate the two terms in (4.7) and write $L(\gamma) = L_1(\gamma) + L_2(\gamma)$, with

$$L_1(\gamma) = n \log |\det \mathbf{A}\Gamma\mathbf{A}^{\mathbf{T}}| \tag{4.8}$$

$$L_2(\gamma) = -2 \sum_{t=1}^{n} \sum_{j=1}^{N} \log p_{s_j}(\gamma_j^{\frac{1}{2}} \mathbf{a}_{\mathbf{j}}^{\mathbf{T}} (\mathbf{A}\Gamma\mathbf{A}^{\mathbf{T}})^{-\mathbf{1}} \mathbf{y}_{\mathbf{t}}) \tag{4.9}$$

Setting $\Sigma = (\mathbf{A}\Gamma\mathbf{A}^{\mathbf{T}})$ and calculating the gradient for each, we get,

$$\frac{\partial L_1}{\partial \gamma_i} = n \mathbf{a}_{\mathbf{i}}^{\mathbf{T}} \Sigma^{-\mathbf{1}} \mathbf{a}_{\mathbf{i}} \tag{4.10}$$

$$\frac{\partial L_2}{\partial \gamma_i} = -2 \sum_{t=1}^{n} \nabla_i(t) \tag{4.11}$$

where

$$\nabla_i(t) = -\left(\sum_{j=1}^{N} \frac{p_s'\left(f(j) y_t\right)}{p_s\left(f(j) y_t\right)} \left(f(j) \mathbf{a_i} \mathbf{a}_{\mathbf{i}}^{\mathbf{T}} \Sigma^{-\mathbf{1}} \mathbf{y_t} \right) \right) \tag{4.12}$$

$$+ \frac{p_s'\left(f(i)\mathbf{y_t}\right)}{p_s\left(f(i)\mathbf{y_t}\right)} \left(\frac{1}{2} \gamma_i^{\frac{-1}{2}} \mathbf{a}_{\mathbf{i}}^{\mathbf{T}} \Sigma^{-\mathbf{1}} \mathbf{y_t} \right) \tag{4.13}$$

with

$$f(j) = \gamma_j^{\frac{1}{2}} \mathbf{a_j^T \Sigma^{-1}}.$$

For the distribution we used here, $\frac{p_s'(x)}{p_s(x)} = -\tanh(x)$ if $p_s(x) = C\text{sech}(x)$.

To achieve the fixed point update we set the gradient equal to 0 and obtain the following equality

$$n a_i^T \Sigma^{-1} a_i = 2 \sum_{t=1}^{n} \nabla_i(t)$$

Following the approach of [62], we form the fixed point update as

$$\gamma_i^{(k+1)} = \gamma_i^{(k)} \frac{2 \sum_{t=1}^{n} \nabla_i(t)}{n a_i^T \Sigma^{-1} a_i}. \tag{4.14}$$

We initialize the algorithm with $\gamma_i^{(0)} = 1, \forall i$, and perform the updates for a fixed number of iterations, or stop once the changes in γ are below a threshold.

4.2.3 BASRICA

In [52], Le et al. proposed an ICA algorithm called Reconstruction ICA (RICA) using a soft reconstruction cost, which is also applicable to the overcomplete ICA case. Instead of finding the mixing matrix, RICA optimizes the tall unmixing matrix $\mathbf{W} \in \mathbb{R}^{N \times M}$ with the following objective function.

$$\min \frac{1}{2} \sum_{t=1}^{n} \|\mathbf{W^T W y_t} - \mathbf{y_t}\|_2^2 + \lambda \sum_{j=1}^{N} \sum_{t=1}^{n} g(\mathbf{W_j} y_t) \tag{4.15}$$

We modify the RICA objective function such that it allows for a basis selection from a known dictionary $\mathbf{A}$. Using the same idea as in Section 4.2.2.A, we regard the mixing matrix as a weighted selection of columns from the dictionary $\mathbf{A}$, namely

$\mathbf{A_{gen}} = \mathbf{A\Gamma}$, with the projection to the source domain as $\mathbf{x_t} = \mathbf{W}'\mathbf{y_t} = (\mathbf{A\Gamma})^{\dagger}\mathbf{y_t}$ [2]. If the data is whitened by the sphering matrix $\mathbf{S}$, such that $\hat{\mathbf{y}}_\mathbf{t} = \mathbf{S}\mathbf{y_t}$, the source equation can be rewritten as $\mathbf{x_t} = \mathbf{W}\hat{\mathbf{y}}_\mathbf{t} = (\mathbf{A\Gamma})^{\dagger}\mathbf{S^{-1}}\hat{\mathbf{y}}_\mathbf{t}$. Plugging $\mathbf{W} = (\mathbf{A\Gamma})^{\dagger}\mathbf{S^{-1}} = \mathbf{\Gamma A}(\mathbf{A\Gamma^2 A^T})^{-1}\mathbf{S^{-1}}$ into (4.15) gives the following function to be optimized for BASRICA.

$$\min_{\gamma} \frac{1}{2}\sum_{t=1}^{n} \|\mathbf{S^{-T}}(\mathbf{A\Gamma^2 A^T})^{-1}\mathbf{S^{-1}}\hat{\mathbf{y}}_\mathbf{t} - \hat{\mathbf{y}}_\mathbf{t}\|_2^2 + \ldots$$
$$\ldots + \lambda \sum_{j=1}^{N}\sum_{t=1}^{n} g(\gamma_\mathbf{j}\mathbf{a_j^T}(\mathbf{A\Gamma^2 A^T})^{-1}\mathbf{S^{-1}}\hat{\mathbf{y}}_\mathbf{t}) \tag{4.16}$$

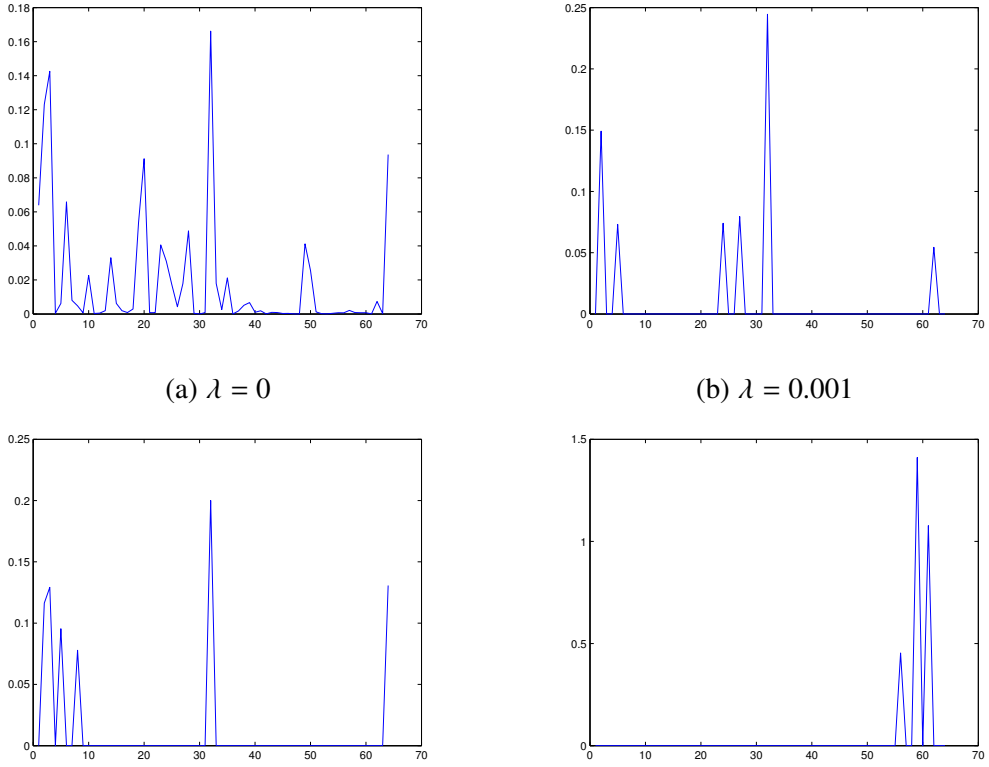
To be consistent with the BASICA formulation we choose $g(.) = \log(\text{sech}(.))$. One of the benefits of this optimization problem is that it allows us to use unconstrained solvers, e.g. L-FBGS. Since RICA results in a degenerate solution $\mathbf{W}$ (only M nonzero rows) without row normalization, we expect $\mathbf{\Gamma A}(\mathbf{A\Gamma^2 A^T})^{-1}$ to converge to $N - M$ zero rows as well, equivalently to a sparse γ. Moreover, the sparsity of the solutions can be altered with the trade-off parameter λ. Figure 4.2 shows the BASRICA solutions for different λ on the same data.

4.3 Connections to M-SBL

4.3.1 BASICA

In this section, we explore the connection of BASICA to M-SBL which is the modification of sparse bayesian learning (SBL) algorithm to the simultaneous sparse approximation [97]. M-SBL is originally aimed at finding a sparse representation of a data matrix $\mathbf{Y}$ given a dictionary $\mathbf{A}$, namely finding $k < M$ columns of $\mathbf{A}$ that may be the generating basis for the data. In order to be able to represent data that lives in an

[2] Without loss of generality, we choose the weighting Γ contrary to $\Gamma^{\frac{1}{2}}$ in BASICA because it will enable us to optimize without the constraint $\gamma \geq 0$.



(a) $\lambda = 0$ (b) $\lambda = 0.001$

(c) $\lambda = 0.01$ (d) $\lambda = 0.05$

Figure 4.2: BASRICA solutions with varying lambda on an experiment with M = 3, N = 64. The sparsity of γ increases with λ. Solution to a complete set with M nonzero γ can be achieved by tuning λ. (Yet, we used $\lambda = 0.05$ in all our experiments, as was used in [52])

M-dimensional space with $k < M$ basis vectors, M-SBL also includes a noise hyper parameter λ that is to be learned as well. However, when λ is forced to be 0 (no noise), and the algorithm is performed on data of rank M, it chooses $k \geq M$ columns of the dictionary to explain the data. In [97], the negative log-likelihood for M-SBL is given as,

$$L_{M-SBL}(\gamma, \lambda) = n \log |\det \left(\lambda \mathbf{I} + \mathbf{A}\Gamma\mathbf{A}^{\mathbf{T}}\right)| + \ldots$$
$$\ldots + \sum_{t=1}^{n} \mathbf{y_t^T} \left(\lambda \mathbf{I} + \mathbf{A}\Gamma\mathbf{A}^{\mathbf{T}}\right)^{-1} \mathbf{y_t}. \tag{4.17}$$

where λ is the noise variance parameter. Plugging $p_{s_i}(x) = \mathcal{N}(0,1)$ into (4.7) for BASICA gives,

$$\begin{aligned} L(\gamma) &\propto n \log|\det \mathbf{A\Gamma A^T}| + \dots \\ &\dots + \sum_{t=1}^{n} \mathbf{y_t^T (A\Gamma A^T)^{-1} A\Gamma^{\frac{1}{2}}\Gamma^{\frac{1}{2}} A^T (A\Gamma A^T)^{-1} y_t} \\ &= n \log|\det \mathbf{A\Gamma A^T}| + \sum_{t=1}^{n} \mathbf{y_t^T (A\Gamma A^T)^{-1} y_t} \\ &\equiv L_{M-SBL}(\gamma, \lambda) \qquad , \lambda \to 0 \end{aligned}$$

We can see that M-SBL in the noiseless limit is a special case of BASICA.

4.3.2 BASRICA and M-SBL

It was previously shown that in the case of strictly sparse ($k < M$) sample-wise orthogonal sources , the solution for M-SBL in the noiseless case is unique and reveals the true sources [97, 98]. Following a similar method, it can be shown that for arbitrary number of sample-wise orthogonal sources, M-SBL solution tries to satisfy $\mathbf{A\Gamma A^T} = \frac{\mathbf{1}}{\mathbf{n}}\mathbf{YY^T}$ [3]. If such a solution $\hat{\Gamma}$ exists, $\hat{\Gamma}^{\frac{1}{2}}$ is a solution of BASRICA with parameter $\lambda = 0$, as shown below.

With $\lambda = 0$, BASRICA minimizes

$$\min \|\mathbf{S^{-T}(A\Gamma^2 A^T)^{-1} S^{-1}\hat{y}_t - \hat{y}_t}\|_2^2 \tag{4.18}$$

which by Lemma 3.1 in [52], is equivalent to minimizing

$$\|\mathbf{S^{-T}(A\Gamma^2 A^T)^{-1} S^{-1} - I}\|_F^2. \tag{4.19}$$

[3] See Chapter 2

Since $\mathbf{SYY^TS^T} = \mathbf{nI}$, we have $\mathbf{YY^T} = \mathbf{n(S^TS)^{-1}}$. Plugging $\hat{\Gamma}^{\frac{1}{2}}$ in BASRICA objective function gives $\|\mathbf{S^{-T}(A\hat{\Gamma}A^T)^{-1}S^{-1}} - \mathbf{I}\|_F^2 = \|\mathbf{S^{-T}(YY^T)^{-1}S^{-1}} - \mathbf{I}\|_F^2 = \|\mathbf{S^{-T}S^TSS^{-1}} - \mathbf{I}\|_F^2$ $= 0$. □

4.4 Experiments

We test our algorithms on synthetic data and real EEG data to compare results with those of M-SBL and reweighted $l_{1,1}$ which induces a Laplacian prior on the sources.

4.4.1 Simulated Data

Here, we compare the performance of our algorithms in a realistic EEG scenario. We construct a coherent dictionary $\mathbf{A}$ of EEG scalp maps of size $M = 32, N = 100$. We obtain this dictionary by selecting a subset of columns that are coherent with each other from the lead-field matrix. The subset we choose creates a highly coherent dictionary with mutual coherence of $\mu = 0.998$, and average spatial map correlation of 0.85.

For each trial, we randomly choose the support set of size M, and obtain M realistic EEG sources from ICA decompositions of earlier EEG studies. We generate data as $\mathbf{Y} = \mathbf{AX}$, and applying different algorithms we try to recover the true support set of sources. After convergence, we extract the support set of size M by choosing M rows of the resulting source matrix that has the highest power. We calculate the success ratio for each algorithm with the below formula after 100 trials

$$r = \frac{1}{100}\sum_{k=1}^{100} |s_k \cap \hat{s}_k|/M. \tag{4.20}$$

where s_k is the true support set for trial k and $\hat{s}_k$ is the support set returned by the algorithm. Figure 4.3a shows the comparison of 4 algorithms. It is seen that BASICA

and BASRICA outperform M-SBL and reweighted $l_{1,1}$ in terms of converging to the true support set in the highly coherent dictionary. Our algorithms require fewer data points to successfully identify the true sources.

Another approach we investigated is performing regular ICA, e.g. Infomax, on data $\mathbf{Y}$ and finding a mixing matrix $\mathbf{A}'$, followed by matching the columns of $\mathbf{A}'$ to the closest columns in dictionary $\mathbf{A}$, in a one-to-one manner. We define the closeness of two vectors $\mathbf{v_1}$ and $\mathbf{v_2}$ as $\mathbf{d}(\mathbf{v_1}, \mathbf{v_2}) = \frac{|\mathbf{v_1^T v_2}|}{\|\mathbf{v_1}\|_2 \|\mathbf{v_2}\|_2}$. Let $\mathbf{a'_i}$ denote the $\mathbf{i}$-th column of the mixing matrix $\mathbf{A}'$. We match $\mathbf{a'_i}$ with $\mathbf{a_j}$ if $\mathbf{j}$ satisfies

$$j = \arg\max_{\mathbf{k}} \mathbf{d}(\mathbf{a'_i}, \mathbf{a_k})$$

We handle the case where more than one vectors in $\mathbf{A}'$ match with the same column $\mathbf{a_j}$ by giving priority to the one who has the highest closeness measure.

Comparing Figure 4.3a and 4.3b, it can be seen that unconstrained ICA requires many more data points to converge to the true mixing matrix. This example shows an important benefit of our proposed approaches for direct basis selection.

4.4.2 Experiments on real EEG data

Given a real EEG data segment and a dictionary of possible sources, it is a challenging task to assess how the algorithms perform, due to the unknown nature of true sources. Yet, for EEG source separation tasks it is widely accepted that the sources are instantaneously statistically independent of each other. Therefore, a measure of independence among the sources, e.g. mutual information reduction (MIR) [28], can serve well to compare the results of different algorithms.

The calculation of MIR in (4.2) requires a square unmixing matrix $\mathbf{W}$ that relates the sources and data as $\mathbf{X} = \mathbf{WY}$. This fits well with our methods, since the support set

s our algorithms select from the dictionary is of size M and we can assign $\mathbf{W} = \mathbf{A}_s^{-1}$, where $\mathbf{A}_s$ is the matrix of selected columns.

We use 32-channel 256-Hz EEG data collected during a rapid serial visual presentation task (RSVP). We first perform an ICA mixture model on the entire dataset using multi-model AMICA [74] with 10 mixture models, returning 10 square mixing matrices $\{\mathbf{A_1}, \mathbf{A_2}, \ldots, \mathbf{A_{10}}\}$. Dictionary $\mathbf{A}$ is then formed by concatenating those individual ICA models and removing the identical scalp maps. Multi-model AMICA was previously shown to capture the possible non-stationarities inherent in EEG data [74], thus forming a tractable way of obtaining more sources than sensors from EEG. The overcomplete scalp maps dictionary we obtain with the above described method is of size $M = 32, N = 63$.

We extract 70 EEG epochs (data segments around events of interest) of length 4 and 6 seconds. Using the dictionary $\mathbf{A}$, we run our algorithms separately on each epoch and compare the resulting MIR values. In Table 4.1 and 4.2, we perform a pairwise comparison of the algorithms and measure the percentage of epochs for which one algorithm results in a larger MIR than the other. In addition to the algorithms examined before, we also compare results with the maximum MIR for the individual ICA models returned by AMICA, namely $\text{MIR}_{\text{AMICA}} = \max_i \text{MIR}(\mathbf{A_i^{-1}})$. It can be seen that BASRICA has the highest likelihood of returning a larger mutual information reduction over all pairwise comparisons. BASRICA obtained a higher MIR than individual AMICA models on $\sim 82\%$ of the epochs ($p < 0.025$ on Wilcoxon signed-rank test). On real EEG, BASRICA performs better than BASICA possibly due to the data representation (error/noise) term in (4.16).

Table 4.1: Percentage of 6 second epochs for which Algorithm i (row) produces more MIR than Algorithm j (column).

Algorithms	Reweighted $l_{1,1}$	M-SBL	BASICA	BASRICA	$\text{MIR}_{\text{AMICA}}$
Reweighted $l_{1,1}$	·	97.14	0	17.14	2.86
M-SBL	2.86	·	2.86	5.71	2.86
BASICA	100	97.14	·	17.14	14.29
BASRICA	82.86	94.29	82.86	·	82.86
$\text{MIR}_{\text{AMICA}}$	97.14	97.14	85.71	17.14	·

Table 4.2: Percentage of 4 second epochs on which Algorithm i has higher MIR than Algorithm j. Row and column indexes are i, j respectively.

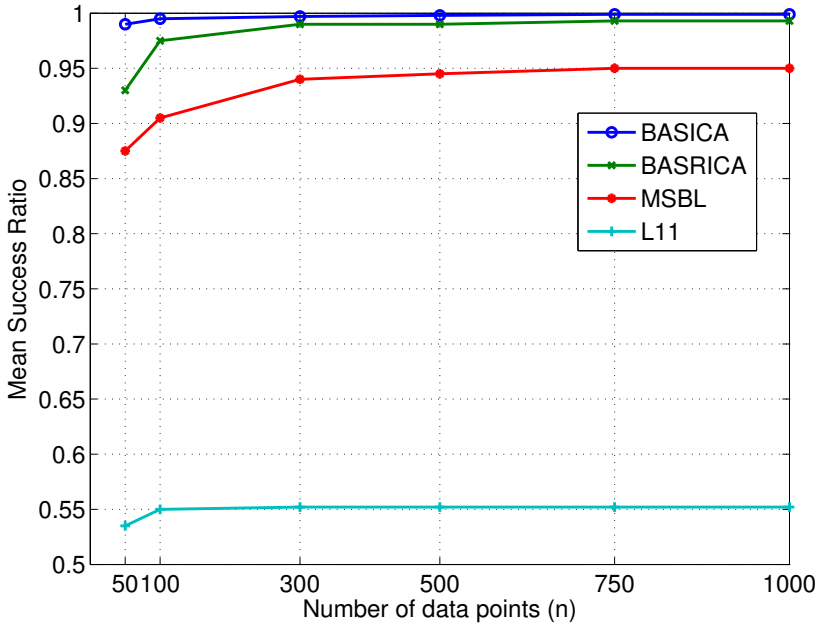
Algorithms	Reweighted $l_{1,1}$	M-SBL	BASICA	BASRICA	$\text{MIR}_{\text{AMICA}}$
Reweighted $l_{1,1}$	·	91.43	2.86	0	2.86
M-SBL	8.57	·	8.57	2.86	8.57
BASICA	97.14	91.43	·	0	20
BASRICA	100	97.14	100	·	100
$\text{MIR}_{\text{AMICA}}$	97.14	91.43	80	0	·

4.5 Conclusion

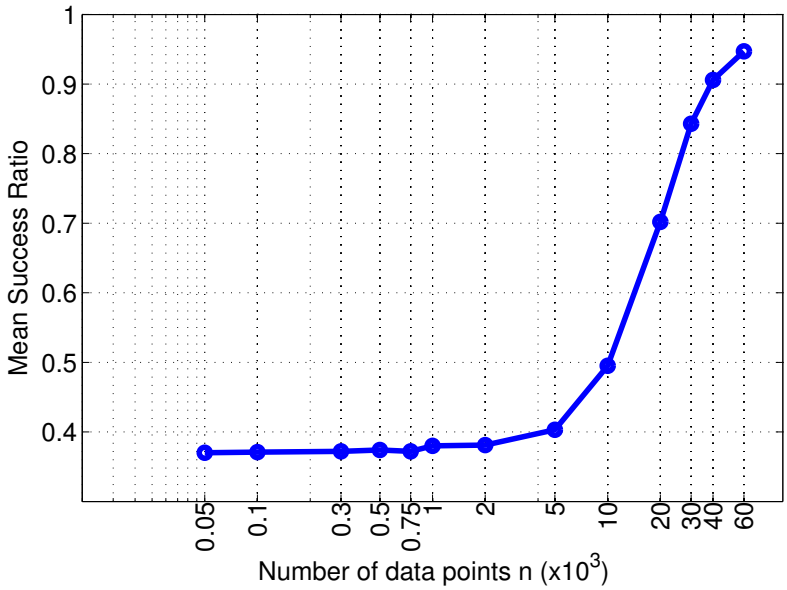
We modify Infomax and RICA to construct two algorithms aimed at finding jointly active sources in the case of a known overcomplete set of possible sources. While previous attempts at underdetermined source recovery problems focus on finding the sparsest solution, our algorithms aim at finding the maximally independent sources. We show that on simulated realistic EEG data our algorithms can recover the true sources in the case of a highly coherent dictionary while requiring relatively fewer data points compared to other algorithms. In real EEG experiments, our algorithms obtain higher mutual information reduction.

4.6 Acknowledgements

The text of Chapter 4, in part, is a reprint of the material as it appears in: Balkan, Ozgur; Bigdely-Shamlo, Nima; Kreutz-Delgado, Kenneth; Makeig, Scott, "Basis Selection for Maximally Independent EEG Sources", In Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, pp. 6639-6642. IEEE, 2014. The dissertation author was the primary investigator and author of this paper.



(a)



(b)

Figure 4.3: (a) Comparison of algorithms on synthetically generated data with EEG scalp maps dictionary (b) Perfomance of ICA + column matching on the same type of data. It requires 500x data points compared to BASICA and BASRICA to recover 95% of the true sources

Chapter 5

Robust Joint-Sparse Recovery On Data with Outliers

In this chapter, we develop a method to solve the multiple measurement vector (MMV) sparse signal recovery problem in a robust manner when data contains outlier points which do not fit the shared sparsity structure otherwise contained in the data. This scenario occurs frequently in the applications of MMV models due to only partially known source dynamics. The algorithm we propose is a modification of MMV-based sparse bayesian learning (M-SBL) by incorporating the idea of least trimmed squares (LTS), which has previously been developed for robust linear regression. Experiments show a significant performance improvement over the conventional M-SBL under different outlier ratios and amplitudes.

5.1 Introduction

In many applications of MMV algorithms, e.g. source localization for EEG and MEG, the data matrix $\mathbf{Y}$ is obtained by taking out a data window of interest from a larger

set of data [94, 37]. If one has a prior knowledge of when the sources turn active and inactive, then the data $\mathbf{Y}$ can be extracted such that the common sparsity assumption holds. However, in most cases this knowledge does not exist and the assumption of common sparsity pattern for 100% of the data becomes far from ideal. Here, we refer to data vectors $\mathbf{y_{.i}}$ as outliers if the associated $\mathbf{x_{.i}}$ do not fit the assumed MMV model, which is the shared sparsity assumption.

If the window size n is expanded for performance improvement, the possibility of the data containing outliers increases which can in turn dramatically decrease the algorithm performance. In [107], the MMV algorithms has been shown to be useful for recovering the sparse sources in the case of time-varying sparsity when it is applied on sliding data windows. However, because of the nature of the problem, and unknown locations of sparsity pattern changes, it is rather likely that sliding windows of data contain outliers.

Due to the non-ideal cases described above, we seek an MMV algorithm that is robust to outliers. To do so, we modify the multiple-Sparse Bayesian Learning (M-SBL) algorithm proposed in [97] such that the new algorithm robust-MSBL captures the support set of the majority of data vectors $\mathbf{y_{.i}}$ ignoring the outliers. We adopt the *least trimmed squares* (LTS) method used for robust linear regression in [82, 81] and apply it to the MMV problem. A similar idea is also followed in [4] for robust sparse linear regression and robust PCA [45]. It should be noted that robustness to noise and high sparsity value k is pursued in [44] for the MMV model, however, our approach differs in the sense that we pursue robustness to outliers that are not jointly sparse with the rest of the data.

The outline of the chapter is as follows: Section 5.2 explains reasons for sensitivity of M-SBL to outliers. Section 5.3 overviews *least trimmed squares* (LTS) and establishes its connection to robust-MSBL. In Section 5.4, we perform tests. And Section 5.5 gives

some conclusions.

5.2 Sensitivity of M-SBL

M-SBL has the following cost function,

$$L(\gamma) \triangleq -2\log(p(\mathbf{Y};\gamma)) = -2\log\int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X};\gamma)d\mathbf{X}$$
$$\equiv \log|\Sigma| + \frac{1}{n}\sum_{t=1}^{n}\mathbf{y_{.t}}^T\Sigma^{-1}\mathbf{y_{.t}} \tag{5.1}$$

where $\Sigma \triangleq (\mathbf{A}\Gamma\mathbf{A^T} + \sigma^2\mathbf{I})$, $\Gamma \triangleq \text{diag}(\gamma)$. Although our modifications in this cahapter could be generalized to every method, we will minimize this cost function by using the fixed point update approach, which has significantly faster convergence rate than the EM update [98].

$$\gamma_i^{(k+1)} = \frac{\gamma_i^{(k)}}{\mathbf{a_i^T}(\Sigma^{(k)})^{-1}\mathbf{a_i}}\frac{\|\mathbf{Y^T}(\Sigma^{(k)})^{-1}\mathbf{a_i}\|_2^2}{n} \tag{5.2}$$

at the $(k+1)$-th iteration and $\gamma_i^{(0)} = 1, \forall i$. The first term $\log|\Sigma|$ in the cost function (5.1) encourages sparsity of γ whereas the second term tries to fit data as pointed out in [94]. It is also possible to learn the noise parameter σ^2, however it was noted before that the best results are achieved using a fixed value (whether estimated by a different method or using prior knowledge) [97]. After convergence, if the sparsity k is known, one extracts the indices of the largest k values of γ in order to recover the support set.

One can see that all data vectors $\mathbf{y_{.t}}$ and thus $\mathbf{x_{.t}}$ are treated equally in this formulation (has the same weight $\frac{1}{n}$ in (5.1)), which makes the cost function sensitive to outliers. A large amplitude source outlier $\mathbf{x_{it'}}$ (at time t' in i-th row) is sufficient to boost γ_i since a zero mean gaussian distribution with variance γ_i is fit for i-th row of $\mathbf{X}$. In a scenario where all source vectors $\mathbf{x_{.t}}$ shares the same sparsity pattern this

phenomenon does not create a problem in terms of recovering the support set because we would already desire γ_i for $i \in S$ to be large. However, if even one $\mathbf{x}_{\cdot\mathbf{t}'}$ that does not share the common sparsity pattern exists (nonzero values at $\mathbf{x}_{\mathbf{it}'}$ for $i \notin S$), resulting γ would contain nonzero values at indices $i \notin S$. Moreover, if $\mathbf{x}_{\mathbf{it}'}$ for $i \notin S$ is large, the associated γ_i would also be large and thus would be falsely regarded as one of the support set indices.

It can also be observed that in the noiseless limit case, as $\sigma^2 \to 0$, if there exists a data vector $\mathbf{y}_{\cdot\mathbf{j}}$ such that $\mathbf{y}_{\cdot\mathbf{j}} \notin \text{span}(\mathbf{A}_\mathbf{S})$, then any sparse γ satisfying $\gamma_i = 0$ for $i \notin S$, cannot be a local minimum of (5.1) since $\mathbf{y}_{\cdot\mathbf{j}}{}^T \Sigma^{-1} \mathbf{y}_{\cdot\mathbf{j}} \to \infty$. Thus, outliers of the likelihood function (5.1) are not only the large amplitude data vectors but also the ones that do not share the sparsity pattern of the majority.

5.3 LTS and robust-MSBL

5.3.1 LTS

One of the most common methods for linear regression is *least squares* (LS), where the regression parameter is fit such that the sum of squared residuals are minimized. Equivalently,

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} r_i^2 \tag{5.3}$$

where θ is the parameter to be optimized and each residual r_i is a function of θ. The weight $\frac{1}{n}$ can be omitted however we keep it to emphasize that what is being minimized is actually the mean of r_i's. Despite its common use, it is known that this method is very sensitive to outliers. Since every point has the same weight $\frac{1}{n}$, a single large outlier can dramatically change the solution. In [82, 81], this problem is analyzed in detail and alternative robust methods are proposed, one of which is *least trimmed squares* (LTS)

with the below function to be optimized.

$$\min_{\theta} \sum_{i=1}^{h} (r^2)_{i:n} \tag{5.4}$$

where $(r^2)_{1:n} \leq (r^2)_{2:n} \leq \ldots \leq (r^2)_{n:n}$ are the ordered residuals, $h \geq \frac{n}{2}$ is the parameter of the LTS estimator determining the number of data points to fit the parameter θ on. LTS therefore allows for some large values of r_i^2 while being able to fit better to the majority of data.

5.3.2 Robust-MSBL

In order to make M-SBL tolerant to outlier data points, we apply an analogue of the LTS formulation to the conventional M-SBL. Using the same notation in 5.3.1, we define the data fit residual for the i-th vector as

$$r_i^2 = \log|\Sigma| + \mathbf{y}_{\mathbf{.i}}^{\mathbf{T}} \Sigma^{-1} \mathbf{y}_{\mathbf{.i}} \tag{5.5}$$

which is a function of γ. It can be seen that the conventional M-SBL formulation in (5.1) is equivalent to the LS estimation when r_i^2 is defined as above. Applying the LTS idea to M-SBL, following cost function is obtained.

$$\begin{aligned} L(\gamma) &= \sum_{t=1}^{h} (r_i^2)_{(t:n)} \\ &= h \log|\Sigma| + \sum_{t=1}^{h} (\mathbf{y}_{\mathbf{.i}}^{\mathbf{T}} \Sigma^{-1} \mathbf{y}_{\mathbf{.i}})_{(t:n)} \\ &\equiv \log|\Sigma| + \frac{1}{h} \sum_{t=1}^{h} (\mathbf{y}_{\mathbf{.i}}^{\mathbf{T}} \Sigma^{-1} \mathbf{y}_{\mathbf{.i}})_{(t:n)} \end{aligned} \tag{5.6}$$

This formulation is equivalent to finding a h-size subset of n columns of $\mathbf{Y}$ that would result in the smallest sum of squared residuals. Given a subset H of size h, we can find γ that minimizes the cost function by conventional M-SBL optimization method given in (5.2). If $L(\gamma, H)$ denotes the M-SBL objective function restricted to the subset H, we have

$$L(\gamma, H) = \log|\Sigma| + \frac{1}{h}\sum_{t\in H} \mathbf{y}_{.\mathbf{t}}^{\mathbf{T}}\Sigma^{-1}\mathbf{y}_{.\mathbf{t}} \tag{5.7}$$

$$\hat{\gamma}_H = \arg\min_{\gamma} L(\gamma, H) \tag{5.8}$$

and the subset of data which will result in global optimum would be given by

$$H^* = \underset{H\subseteq\{1,2,\ldots,n\},|H|=h}{\arg\min} L(\hat{\gamma}_H, H) \tag{5.9}$$

This is yet another combinatorial problem where one needs to consider all subsets of size h and perform M-SBL on each of these subsets of data. However, to optimize (5.6), we follow the iterative method proposed in [81] and also performed in [4]. This method is composed of C-steps which iteratively decrease the objective function value at each step and converge to a local minimum.

C-steps

We start with a random h-size subset of indices $\{1, 2, \ldots, n\}$ and denote this set as H_0. We perform regular M-SBL on this subset of data $\mathbf{Y}$ determined by H_0. With the resulting $\hat{\gamma}_{H_0}$, we compute residuals r_i^2 for all data vectors $\mathbf{y}_{.\mathbf{t}}$ as in (5.5). We find the smallest h of these residuals, assign these indices as the new subset H_1 and keep repeating the same steps until $H_k = H_{k+1}$. As also shown in [4] this method decreases

the cost function at each step, as

$$L(\hat{\gamma}_{H_{k+1}}, H_{k+1}) \leq L(\hat{\gamma}_{H_k}, H_{k+1}) \leq L(\hat{\gamma}_{H_k}, H_k) \tag{5.10}$$

First inequality is valid due to (5.8) and second inequality is true because of the definition of the next subset H_{k+1}. Of course, it is not guaranteed to reach the global minimum with C-steps due to the random initializations and thus it is best to consider different subset initializations and compare the cost function values obtained at the end. Above verifications are adopted from the sparse LTS regression problem in [4] and applies well to the joint sparse recovery problem. Algorithm 1 presents the pseudocode for robust-MSBL.

Input : $\mathbf{A}, \mathbf{Y}, \boldsymbol{\sigma^2}$, num_initializations, h
Output : γ

for $trial \leftarrow 1$ **to** *num_initializations* **do**
  $H_0 \leftarrow$ random h-size subset of $\{1, \ldots, n\}$;
  $k \leftarrow 0$;
  repeat
    $\gamma \leftarrow \text{MSBL}(A, Y(:, H_k), \sigma^2)$;
    $\Sigma \leftarrow A\Gamma A^T + \sigma^2 I$;
    for $i \leftarrow 1$ **to** n **do**
      $r_i^2 \leftarrow \log|\Sigma| + \mathbf{y_{.i}^T}\Sigma^{-1}\mathbf{y_{.i}}$;
    end
    $H_{k+1} \leftarrow$ indices of minimum h of r_i^2 $k \leftarrow k + 1$
  until $H_k = H_{k-1}$;
  costFunc($trial$) $\leftarrow L(\gamma, H_k)$;
  gammas($trial$) $\leftarrow \gamma$;
end
ind $\leftarrow$ index of minimum of costFunc;
$\gamma \leftarrow$ gammas(ind)

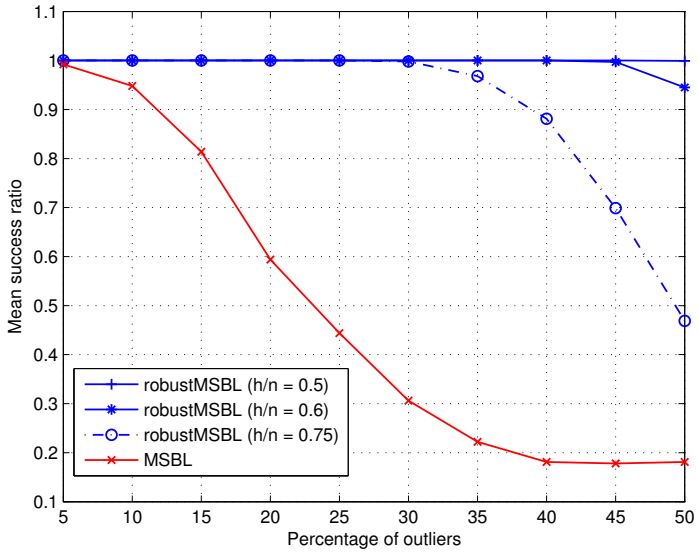
Algorithm 1: robust-MSBL

5.4 Experiments

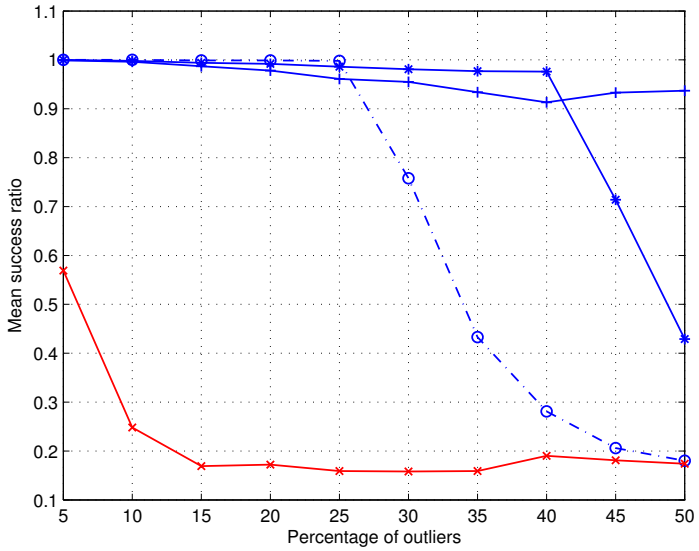
In this section, we perform experiments to compare the performances of MSBL and robust-MSBL under various parameters. We plot the performance with respect to the percentage of the outliers in data, while experimenting with different values of amplitude and sparsity for outliers. At each parameter setting, we perform 100 trials and compute the mean support recovery ratio.

For each trial, we do the following. We create a different random dictionary of size $M = 20$ and $N = 60$ and normalize its columns. We create the source matrix $\mathbf{X}$ of size $N = 60$, $n = 100$ by first separating it into two as $\mathbf{X} = [\mathbf{X_1}; \mathbf{X_2}]$. The first portion of $\mathbf{X}$, i.e. $\mathbf{X_1}$, is the majority of the source vectors which share a common sparsity pattern with sparsity $k_1 = 10$, whereas the second portion of data consists of outliers sharing a different sparsity pattern or not being sparse at all ($k_2 = 10, 60$). Also, note that column permutations of $\mathbf{X}$ or $\mathbf{Y}$ would not affect M-SBL nor robust-MSBL. We randomly select rows with specified parameters k_1 and k_2, and generate source activations from a Gaussian distribution $\mathcal{N}(0, \sigma_1)$ for $\mathbf{X_1}$. The outliers $\mathbf{X_2}$ are sampled from $\mathcal{N}(0, \sigma_2)$. We set the parameter $num\_initializations = 1$. Better performance can be achieved using a higher value however a single initialization was sufficient to show the performance improvement over M-SBL. We add noise to the simulated data such that SNR = 10dB. We recover the support set $\hat{S}$ by extracting the indices of the largest k_1 values of γ's returned by the algorithms and compute the mean success ratio as $\frac{1}{100}\sum_{\text{trial}=1}^{100} |\hat{S} \cap S|/k_1$. Figure 5.1 and Figure 5.2 shows the results when data contains outliers. Figure 5.3 shows the results when there are no outliers.

More indices (k_2) of γ would be nonzero for experiments in Figure 5.2 compared to at most $20 = k_1 + k_2$ in Figure 5.1, thus there is lower chance of true support detection when $k_2 = 60$. The higher the outliers' relative amplitude $\frac{\sigma_2}{\sigma_1}$, the higher γ_i for $i \notin S$,



(a) $k_2 = 10, \frac{\sigma_2}{\sigma_1} = 2$

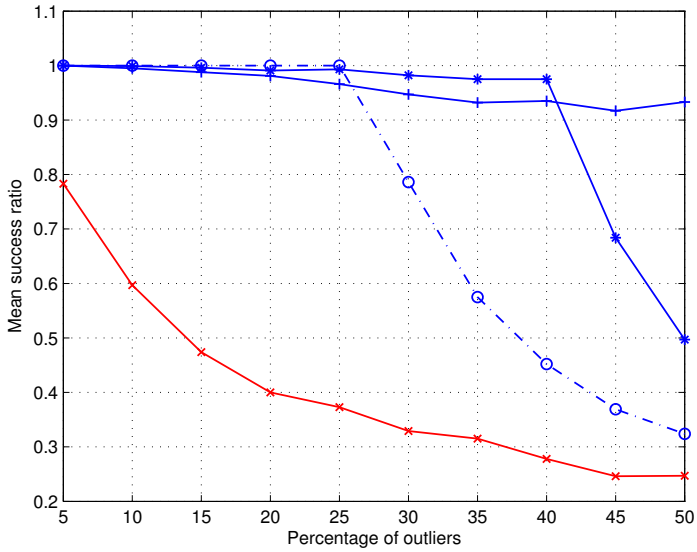


(b) $k_2 = 10, \frac{\sigma_2}{\sigma_1} = 5$

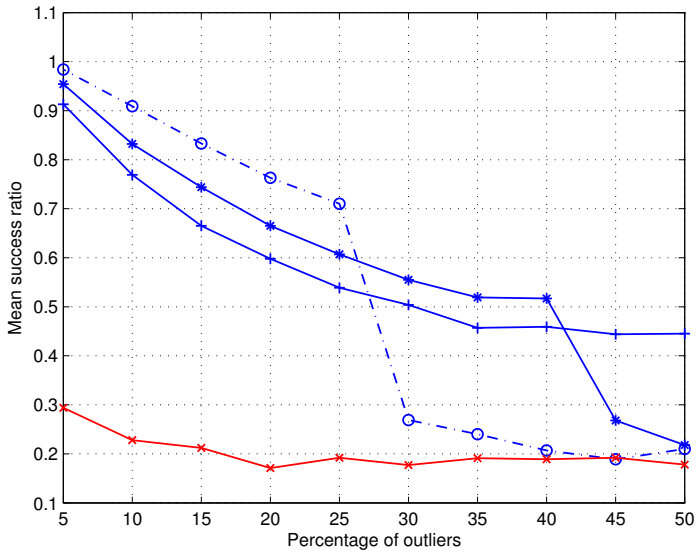
Figure 5.1: $n = 100, k_1 = 10, k_2 = 10$.

hence the performance of correct recovery decreases when $\frac{\sigma_2}{\sigma_1}$ increases. As h increases, the performance of robust-MSBL improves since it finds h points to train γ on. However, number of outliers should be lower than $n - h$.

When there are no outliers, if n is small, M-SBL performs better than robust-



(a) $k_2 = 60, \frac{\sigma_2}{\sigma_1} = 2$



(b) $k_2 = 60, \frac{\sigma_2}{\sigma_1} = 5$

Figure 5.2: $n = 100, k_1 = 10, k_2 = 60$.

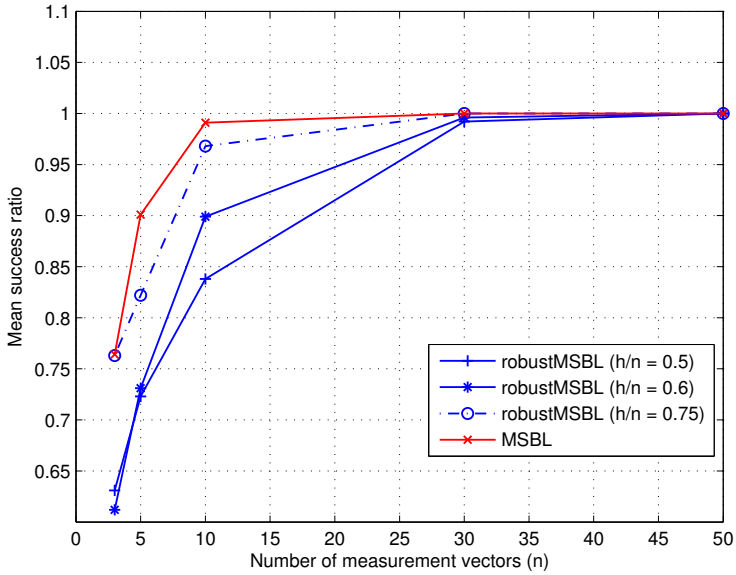


Figure 5.3: Data with no outliers (ideal case).

MSBL when there are no outliers because robust-MSBL finds the best $h < n$ data vectors to optimize γ on, whereas M-SBL uses all n data vectors. However, if the data window n is large enough ($n > 30$ for this experiment), h points become sufficient for robust-MSBL to be as successful as M-SBL.

5.5 Conclusion

In this chapter, we modified M-SBL [97] by exploiting the idea of least trimmed squares. This modification significantly enhances MSBL's robustness to data which contain outliers not sharing the common data sparsity structure. Experiments demonstrate that this approach outperforms M-SBL in recovering the correct support set.

5.6 Acknowledgements

The text of Chapter 5, in full, is a reprint of the material as it appears in: Balkan, Ozgur; Kreutz-Delgado, Kenneth; Makeig, Scott, "Robust Joint Sparse Recovery On Data with Outliers", In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 3821-3825. IEEE, 2013. The dissertation author was the primary investigator and author of this paper.

Chapter 6

Capturing Local Nonstationarities of EEG

In this chapter, we develop methods to identify local, epoch-specific nonstationarities in a given EEG dataset. We assume that most EEG epochs share a common source structure/constellation yet inter-epoch differences may exist. Our method learns the mixing matrix for each segment under the assumption that some portion of the mixing matrix consists of shared components among all epochs. Unlike Chapter 3, a dictionary based method cannot be applied because of the possible uniqueness of epoch-specific sources. We demonstrate our algorithm's performance on simulated data as well as on EEG.

6.1 Introduction

EEG data analysis usually involves extracting data segments of interest by choosing a fixed latency window before and after an event of interest, e.g. [-1000ms, 2000ms] around a target appearing on screen. These segments are called *epochs*, whose EEG time

series are then averaged over multiple trials of the same type and hence an average brain response to a certain event is extracted. Averaging over multiple epochs is a fundamental approach to increase Signal-to-Noise ratio (SNR) and suppress the *variability* among epochs.

Instead of working directly on the signals collected on sensors, one should transfer the problem to the source domain by separating artifacts and brain sources that are mixed at the scalp sensors (for the reasons mentioned in Chapter 3). Traditionally, Independent Component Analysis (ICA) is applied on concatenated EEG epochs for this purpose, and resulting brain source time-series are then processed as described above.

ICA models the overall data $\mathbf{Y}$ as $\mathbf{Y} = \mathbf{AX}$ and thus assumes a single mixing matrix $\mathbf{A}$ that is present in every epoch. It is reasonable to expect certain event-related brain sources to be active in every epoch yet other sources filling the channel space and contributing to the scalp sensor mixing might be different in each epoch.

Our goal is then to find the common sources as well as the epoch-specific differences/nonstationarities. An important example of these epoch-specific sources are the possible variability of the artifacts in each epoch. Being independent from event-related brain responses, muscle/movement/eye artifacts might show countless number of variations and scalp map projections. Moreover, although we might safely assume that a subset of brain activity is responding the same way to certain events at every trial (epoch), there might be epoch-specific variations in the overall source constellations during performing a task. Trying to model these epoch-specific unique sources with a single global mixing matrix, as in the case of ICA, would give erroneous results. Although the common sources might be identified with the traditional ICA approach, a possible lack of success in identifying the remaining sources in the epochs results in inaccurate extraction of time-series for *every* IC.

In this chapter, we tackle the problem of identifying and separating the globally

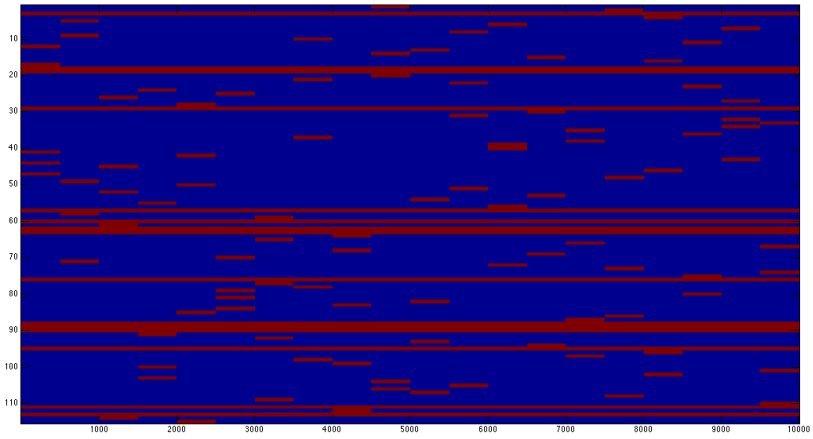


Figure 6.1: Example source model for shared ICA. Epoch length is $L = 500$, number of shared sources $k = 15$. Each epoch has 5 unique sources. Cov-DL cannot be applied in this model because an epoch-specific source only occurs for a short time.

shared sources as well as unique, epoch-dependent sources while still leveraging the independence of the sources.

6.2 Shared ICA

We denote the epochs of EEG as $\mathbf{Y}_e, e = 1, 2, \ldots N_e$, where $\mathbf{Y}_e \in \mathbb{R}^{M \times L_e}$. For each epoch we assume that linear mixing holds because of the nature of cortex/skull/scalp volume conduction. We model the data as

$$\mathbf{Y}_e = \mathbf{A}_e \mathbf{X}_\mathbf{e}, \tag{6.1}$$

where $\mathbf{A_e}$ is the complete mixing matrix present during epoch e. We structure $\mathbf{A}_e$ as,

$$\mathbf{A}_e = [\mathbf{A_c}, \ \mathbf{U_e}] \tag{6.2}$$

where $\mathbf{A_c} \in \mathbb{R}^{M\times k}$ contains the scalp maps of common sources among all epochs. $\mathbf{U_e} \in \mathbb{R}^{M\times M-k}$ represents scalp maps of sources that are unique to epoch e. The number of common sources is denoted by k and is an input to the algorithm described below. Assumed source model is illustrated in Figure 6.1.

6.2.1 Algorithm

Conventional ICA models the data likelihood as

$$p(\mathbf{Y}) = \prod_{t=1}^{n} |\det \mathbf{A}^{-1}| p_s(\mathbf{A^{-1}y_t}) = \prod_{t=1}^{n} \frac{1}{|\det \mathbf{A}|} p_s(\mathbf{A^{-1}y_t}) \tag{6.3}$$

where $p_s(.)$ is the vector source density function. Applying $-\log(\cdot)$ results in the following minimization problem.

$$\mathcal{L}(\mathbf{A}) = \min_{\mathbf{A}} \log|\det \mathbf{A}| + \frac{1}{n} f(\mathbf{x_t}) \tag{6.4}$$

where $\mathbf{x_t} = \mathbf{A^{-1}y_t}$ and $f(\cdot) = -\log(p_s(\cdot))$ The gradient of this cost function is proportional to

$$\begin{aligned} \nabla\mathcal{L}(A) &\propto \mathbf{A}^{-T} + \frac{1}{n}\sum_{t=1}^{n} \mathbf{A}^{-T}\nabla f(\mathbf{x_t})\mathbf{x_t}^T \\ &= \mathbf{A}^{-T}\left(\mathbf{I} - \frac{1}{n}\sum_{t=1}^{n} \mathbf{g_t}\mathbf{x_t^T}\right) \end{aligned} \tag{6.5}$$

with $\mathbf{g_t} = \nabla f(x_t)$. Similarly, we have the following likelihood model for our shared source model:

$$
\begin{aligned}
p(\mathbf{Y}) &= \prod_{e=1}^{N_e} p(\mathbf{Y_e}) = \prod_{e=1}^{N_e} \prod_{t=1}^{L_e} |\det \mathbf{A_e}^{-1}| p_s(\mathbf{A_e^{-1}} \mathbf{Y_e}_t) \\
&= \prod_{e=1}^{N_e} \prod_{t=1}^{L_e} \frac{1}{|\det \mathbf{A_e}|} p_s(\mathbf{A_e^{-1}} \mathbf{Y_e}_t). \\
-\log p(\mathbf{Y}) &\propto \sum_{e=1}^{N_e} \log |\det \mathbf{A_e}| + \frac{1}{L_e} f(\mathbf{X_e}_t)
\end{aligned} \tag{6.6}
$$

Gradient of each epoch can be calculated w.r.t $\mathbf{A_e}$ as in (6.5) and gradient w.r.t $\mathbf{U_e}$ can be extracted. Namely,

$$
\nabla_{\mathbf{U_e}} = \left[\mathbf{A_e}^{-T} \left(\mathbf{I} - \frac{1}{L_e} \sum_{t=1}^{L_e} \mathbf{g_t} \mathbf{X_e}_t^T \right) \right]_{[:,k+1:M]} \tag{6.7}
$$

However, once we have shared sources $\mathbf{A_c}$ in every epoch, overall gradient w.r.t $\mathbf{A_c}$ is the sum of gradients w.r.t $\mathbf{A_c}$ over all epochs.

$$
\nabla_{\mathbf{A_c}} = \sum_{e=1}^{N_e} \left[\mathbf{A_e}^{-T} \left(\mathbf{I} - \frac{1}{L_e} \sum_{t=1}^{L_e} \mathbf{g_t} \mathbf{X_e}_t^T \right) \right]_{[:,1:k]} \tag{6.8}
$$

We have used these gradients with an L-BFGS solver yet other optimization methods may be preferred.

6.2.2 Selection of Parameter k

Shared ICA algorithm has one free parameter k, the number of shared sources among every epoch. In most cases, such as EEG, the true value of k is unknown beforehand. Yet, evaluation of the algorithm for different values of k might reveal insights on to what the true value should be.

As k decreases, since $\mathbf{A_e}$ has more freedom of adapting to the data present in epoch e, we expect that one find a basis set A_e that represents the corresponding epoch better, which results in a lower cost function value at the optimum of the algorithm for every epoch. Therefore, there is a monotonic increase in the optimum value of the cost function as k increases. However, we have observed that the increase is not linear. Moreover, under ideal conditions, i.e. when the true nature of the problem perfectly fits our problem model, there is a clear point of change (elbow) in the trace of cost function values as a function of k. We have created two simulated problems with $M = 5$, $N_e = 10$, $L_e = 1000, 3000$, $k = 3$ and performed Shared ICA for multiple values of k. See Figure 6.2. As expected, total cost increases as k increases, yet the increase is much more apparent after the true value of k. It might be therefore possible to infer the optimal value k by multiple runs of Shared ICA and detecting the point of change. Moreover, it should be noted that although the optimal cost function did not differ much when $1 < k < 3$, the solutions $\mathbf{A_e}$ only matched the ground truth mixing matrices when algorithm parameter was $k = 3$.

6.2.3 Experiments

We implemented Shared ICA on an EEG data collected during the subject is performing a motor imagery task, which was also one of the datasets analyzed in Chapter 3. We divided this 20-minute long 32 channel EEG data into non-overlapping segments of 40 seconds each ($L_e = 4000$) resulting in a total of 30 segments. For parameter section for k, we first performed Shared ICA with values of k in the range $[14, \ldots, 32]$. We plot the Shared ICA optimum cost function value trained for different k's in Figure 6.3. Unlike the case in our simulation, in which the assumed model hold perfectly, we do not have a clear break point in the trace of cost function values when plotted w.r.t k. We suggest finding the break point in this case using linear regressions and F-test. In this

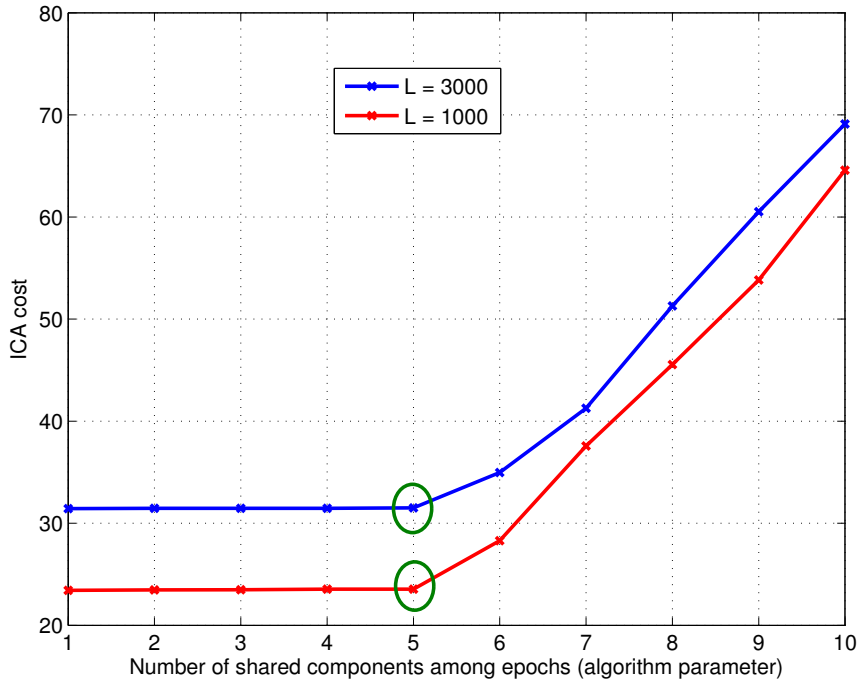


Figure 6.2: $M = 5, N_e = 10, k = 3$. Break point is shown with circle.

dataset, F-test revealed the optimal break point of $k = 28$. This means that each segment would find 4 unique sources and rest of the 28 sources are shared across segments. Some of the sources extracted showed similarity with what regular ICA finds, such as the sources which showed mu-rhythm suppression during imagined movement and located in left and right motor cortex. However, we found that this approach also revealed brain sources that did not exist in the traditional ICA solution. See Figure 6.4

6.3 Base ICA

Here, we are interested in the problem of adapting a given mixing matrix to a data segment. This is an important problem to tackle which can have benefits especially in online brain-computer-interface (BCI) scenarios. In BCI, usually an ICA mixing matrix is learned on training data to extract sources, and the same ICA mixing matrix

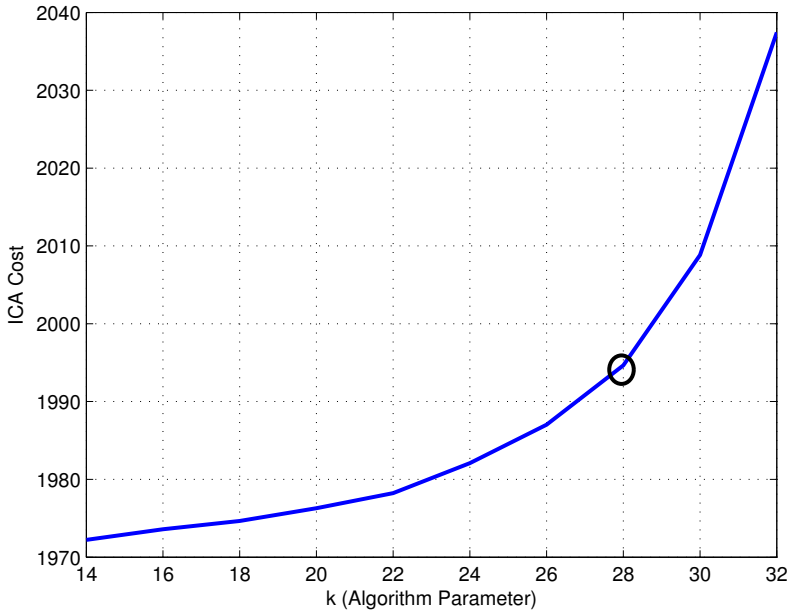
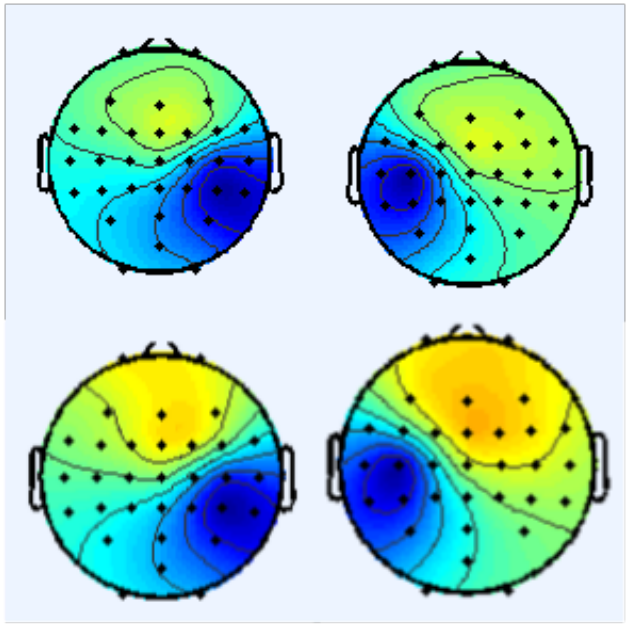


Figure 6.3: Cost function values at the optimum trained with different k. The circle denotes the optimum break point found with F-test.

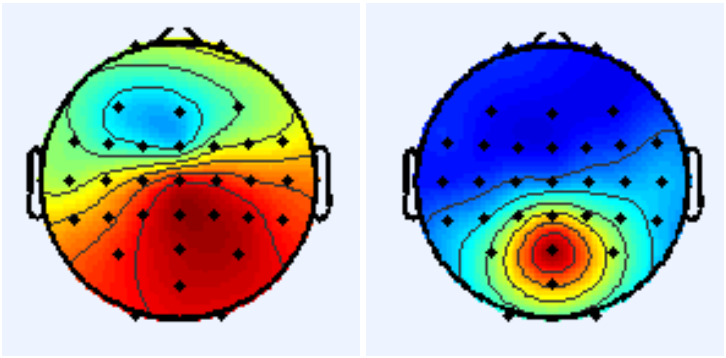
is assumed to hold in the testing phase where new data segments are collected. The underlying idea behind this approach is that, as long as the position of the electrodes do not change between training and test phases, the sources do not change. Yet, this approach is vulnerable when there exists novel sources/artifacts that were not present and not learned in the training phase. Online ICA methods have been proposed [60, 43], yet they fail to capture momentarily appearing/disappearing sources due to their long convergence time.

6.3.1 Algorithm

We are interested in the case where the novel sources are sparse (only a few in number) in a given short data segment. Training ICA from scratch on a given short data segment ($\sim$ 4 sec, with 32 channels) cannot reveal the true sources present due to lack of data, yet it is possible to do so with the *a priori* knowledge that the true mixing matrix



(a)



(b)

Figure 6.4: (a) First row consists of scalp maps associated with motor imagery that traditional ICA found. Second row are the corresponding scalp maps that shared ICA found. (b) Some of the brain sources that only shared ICA revealed.

should be sparsely deviating from the *base* mixing matrix $\mathbf{A}_{\text{base}}$, which was obtained in the training phase.

We impose the assumption on the solution matrix that most of the columns of the mixing matrix $\mathbf{A}_{\text{base}}$ and $\mathbf{A}^*$ are shared yet there could be few columns that differ. This prior assumption on the target variable can be integrated into the likelihood model of

ICA (6.4) facilitating the *maximum-a-posteriori* estimate as

$$\mathcal{L}(\mathbf{A}) = \min_{\mathbf{A}} \log|\det \mathbf{A}| + \frac{1}{n} f(\mathbf{x_t}) + \lambda D\left(\mathbf{A}, \mathbf{A}_{\text{base}}\right). \tag{6.9}$$

where λ is the regularization parameter and $D\left(Q, S\right)$ is the sum of distance between columns of the two matrices Q, S, defined as,

$$D\left(Q, S\right) = \sum_{i=1}^{M} d(q_i, s_i) \tag{6.10}$$

with

$$d(q_i, s_i) = \sqrt{1 - \frac{(q_i^T s_i)^2}{\|q_i\|_2^2 \|s_i\|_2^2}} \tag{6.11}$$

as the positive sine of the angle between two basis vectors. We use angle distance considering the possibility of changes in source powers and thus the scaling of mixing matrix columns. We do not want our algorithm to be sensitive to the scaling difference of the source scalp maps but only to the topographic variations. Equation (6.10) is the sum of absolute values of the sine of the angle and thus in the form of $\|\cdot\|_1$ norm and expected to encourage sparsity with increasing λ, which in turn results in a solution with sparse column deviations from the base matrix $\mathbf{A}_{\text{base}}$.

6.3.2 Experiments

Here, we create short data segments with known ground truth mixing matrix, that is sparsely deviating from a base ICA mixing matrix, multiplied by independent sources. In each experiment, we randomly create the base matrix $\mathbf{A}_{\text{base}}$, and randomly change k columns to obtain $\mathbf{A}^*$. We use 5-fold cross-validation on the data segment to obtain optimal λ using grid search. See Figure 6.5 for $k = 1$ and $L = 100, 500, 1000, 5000$.

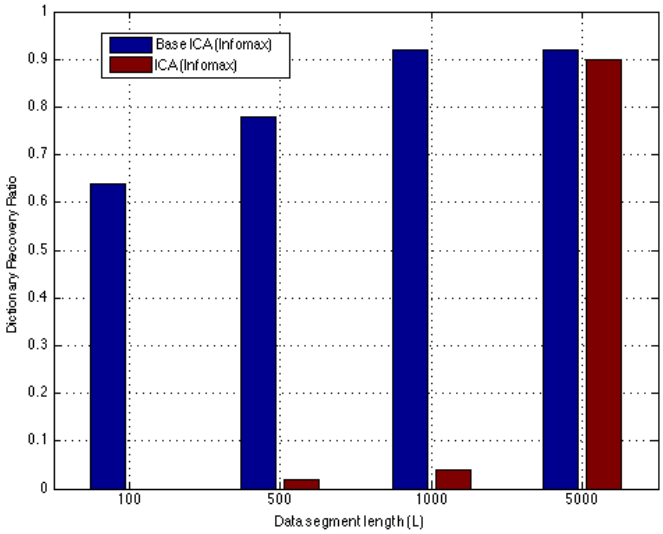


Figure 6.5: Recovery of sources with increasing L

When L is low, ICA does not have enough data to converge to the true solution. Base ICA, on the other hand, makes use of the prior information and stays close to the true solution by automatically learning λ from data. (As $\lambda \rightarrow \infty$, solution $\hat{\mathbf{A}}$ is $\mathbf{A}_{\text{base}}$, if $\lambda \rightarrow 0$, the solution $\mathbf{A}$ is identical to the ICA solution). As L increases, ICA can learn the true mixing matrix without the need for priors.

6.4 Conclusion

In this chapter, we proposed two algorithms that aims at identifying the true EEG sources in the case of non-stationarities. First approach is an algorithm targeted for offline analysis of EEG and assumes that not all but some portion of the active sources are similar throughout the experiment. The algorithm allows to learn unique sources in individual segments/epochs of data. Second algorithm targets the online analysis case and makes use of an existing ICA solution possible obtained on training data. In the online testing mode, it aims at identifying the true sources by automatically learning

sparse deviations from the existing ICA solution. We showed that new sources can be discovered with these methods which were not identifiable with traditional ICA.

6.5 Acknowledgements

The text of Chapter 6, in part, is currently being prepared for submission for publication of the material. Balkan, Ozgur; Kreutz-Delgado, Kenneth; Makeig, Scott, "Capturing Non-Stationary EEG Sources via Shared Source ICA model".

Chapter 7

Source-domain EEG Analysis of Sports-Related Concussion

Here, we show the utility of independent source separation and source localization of EEG in a concussion study to identify concussion related biomarkers. Most work in this field has been through straight-forward sensor-domain analysis. We mention the difficulties source-level analysis brings when assessing group statistics.

7.1 Introduction

Sports activities are a major cause of concussions. It has been estimated that 1.6-3.8 million sports and recreation related concussions occur each year in the United States [35]. A major challenge in the field of neurology is that current neuropsychological, behavioral and standard neuroimaging tools are not sensitive to subtle changes in brain structure and function, thus making the initial diagnosis of concussion difficult. As a result, many adolescents may resume sports activities well before full recovery has occurred, leaving them vulnerable to the potentially serious effects of repeated brain

trauma during their critical period of brain development.

Most studies investigating the EEG correlates of concussion have analyzed the scalp channel data directly [89, 9, 85]. However, because of volume conduction to and source mixing at the electrodes, the EEG sensor signals are simultaneous mixtures of sources located in various parts of the cortex, as mentioned before. In addition, non-brain source processes including eye movements, scalp muscles, head movements, and electrical line noise also contribute to EEG signals. On the assumption that the brain and non-brain source activity time series are statistically independent, independent component analysis (ICA) can separate individual brain and non-brain sources from the scalp mixtures [64]. In this study, we used AMICA [74], that has been shown to provide more dipolar brain sources together with higher independence compared to other algorithms [28].

Although ICA can extract the source activities and facilitate source localization, group-level ICA analysis is non-trivial because of variations across brains and the brain sources extracted for each subject. A recent ICA-based concussion study [77] used k-means clustering of individual subject independent components (ICs) to investigate brain source cluster differences between concussed and non-concussed subjects. Using k-means clustering requires choosing a target number of clusters and relative weights for more than one measure of IC similarity. Here we made use of a recently developed framework, Measure Projection Analysis (MPA), that attempts to avoid these uncertainties[14, 77]. In contrast to [77], which used the standard clinical 19-channel scalp montage, we recorded and analyzed 64-channel EEG data. A lower number of channels limit the number of brain and non-brain sources that can be separated and the accuracy of subsequent source localization [1].

7.2 Methods

7.2.1 Participants

Twenty-one adolescent athletes (all male; mean age, 16.5 years) with a clinical diagnosis of subacute ($\leq$ 3 months previously) sports-related concussion participated in this study. Healthy subjects comprised 33 adolescent soccer players (all male; mean age, 16 years). Subject exclusion criteria included focal neurologic deficits and diagnosis or prescription medications for neurological or psychiatric conditions. All participants were right-handed. Parents of each subject signed an informed consent form that was approved by the University of British Columbia.

7.2.2 EEG acquisition protocol

Resting data were collected for five minutes while subjects had their eyes closed. A 64-channel Hydrogel Geodesic SensorNet (EGI, Eugene, OR) with a Net Amps 300 amplifier was used for EEG recording at a sampling rate of 250 Hz. Electrode impedances were typically below 50 kΩ.

7.2.3 Data processing

Preprocessing

Each subject's EEG signals were band-pass filtered between 0.5 Hz and 45 Hz. Channels whose time series were not consistently correlated with any other channel (r< 0.8) were discarded. On average, 2-3 channels per subject were rejected. Randomly occurring large amplitude artifacts (with power $\geq 3\sigma$ w.r.t clean EEG) were cleaned using ASR [69].

AMICA [74] was used to separate the scalp channel mixtures into maximally

independent brain and non-brain sources. The DIPFIT [1] toolbox in EEGLAB [26] was used to locate the equivalent dipole for each IC (independent component) in the MNI (Montreal Neurological Institute) head model. ICs whose best-fitting dipole accounted for less than %85 of the spatial variance in their scalp projection (column of A), as well as sources whose equivalent dipoles were outside the brain were regarded as non-brain sources and excluded from further analysis. All IC's were initially clustered with respect to their power spectra for the sole purpose of batch rejection of artifact IC's. Visual inspection of power spectra of the resulting cluster centroid and the cluster IC scalp maps identified ICs associated with muscle/movement artifacts which were also excluded from further analysis. In total, 665 ICs were retained, on average ~12 ICs per subject.

7.3 Measure Projection

7.3.1 K-means IC clustering

In most EEG studies, the properties of individual subject ICs, including their scalp projection patterns (scalp maps) and associated dipole locations, are never the same and thus require some form of clustering to identify equivalency classes across subjects. As pointed out in [77, 14] common approaches for multi-subject source-level analysis, such as k-means IC clustering, have some drawbacks. It is not clear which IC features or measures to use for clustering or how different measures should be weighted. In [77], it was shown that clustering on source equivalent dipole locations, on IC spatial extents as estimated by sLORETA, or on IC scalp maps gave component clusters exhibiting different spectral group differences. Assuming some fixed number of clusters is another requirement of k-means clustering that can dramatically affect the nature of the resulting clusters and the numbers of ICs included/excluded.

[1] Available at http://sccn.ucsd.edu/wiki/A08:_DIPFIT

7.3.2 Measure Projection

Measure Projection Analysis (MPA) framework [14] was developed to reduce some drawbacks of other clustering methods. First, MPA divides the (MNI) template head model into a cubic grid with 8-mm spacing comprising 3,908 brain voxels. MPA automatically identifies ICs accounting for eye movement artifact using EyeCatch [13] and excludes them from the analysis. Then, given some activity measure for each IC localized with an equivalent current dipole, each such measure is first smoothed out across neighboring voxels to take into account expected inaccuracy in source dipole locations and between-subject anatomical differences. Here, we used mean-subtracted log spectra as the MPA measure. Subtracting the mean log spectrum is equivalent to inversely scaling each IC time-series by its mean log power.

Denoting the set of voxels in the MNI model by V and one such voxel by v, MPA calculates the projected measure $\mathbb{E}[M(v)]$ at every $v \in V$ as

$$\mathbb{E}[M(v)] = \frac{\sum_{i=1}^{n} P_i(v) M_i}{\sum_{i=1}^{n} P_i(v)} \tag{7.1}$$

where n is the number of ICs in the study. Thus, a spatially-weighted average measure is calculated for each brain voxel, for which the contribution of each IC is determined by the distance between the voxel and the IC equivalent dipole (truncated Gaussian P_i centered at equivalent dipole v_i). Next, brain voxels whose projected measure values are consistent with those at nearby voxels are identified using Ômeasure convergenceÕ $C(v)$, defined in [14] as

$$C(v) = \mathbb{E}[S(v)] = \frac{\sum_{i,j}^{n} P_i(v) P_j(v) S_{i,j}}{\sum_{i,j} P_i(v) P_j(v)} \tag{7.2}$$

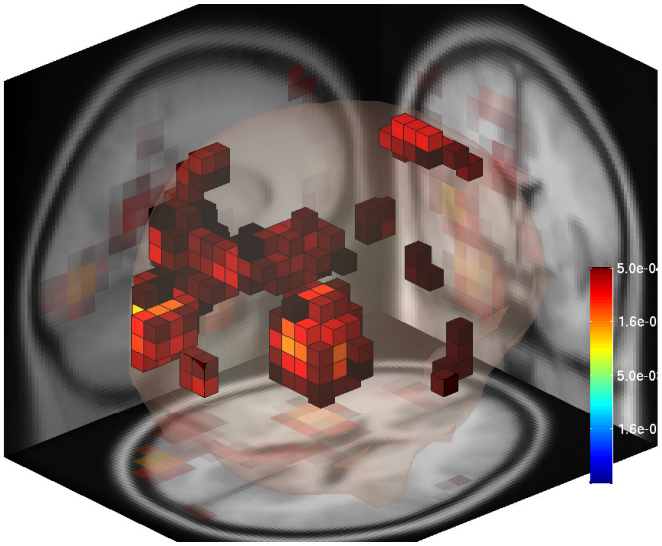
where $S_{i,j}$ is a pairwise IC similarity measure. Estimated measure convergence values

$C(v)$ are assigned significance values after bootstrapping using a surrogate distribution of estimates obtained by randomly re-assigning (with replacement) measure vectors to IC dipole locations. Here, we used a statistical consistency threshold of ($p<0.05$). See Fig. 7.1a. Significantly consistent voxels may then be clustered into "domains" using affinity propagation [36] based on the similarity of their projected measures. Affinity propagation automatically determines the number of clusters and outliers consistent with a given domain disparity threshold. Here, we set the maximum similarity threshold between domains to 0.9. Finally, locations or other properties of the identified regional domains of interest may be investigated to assess group-level measure differences.

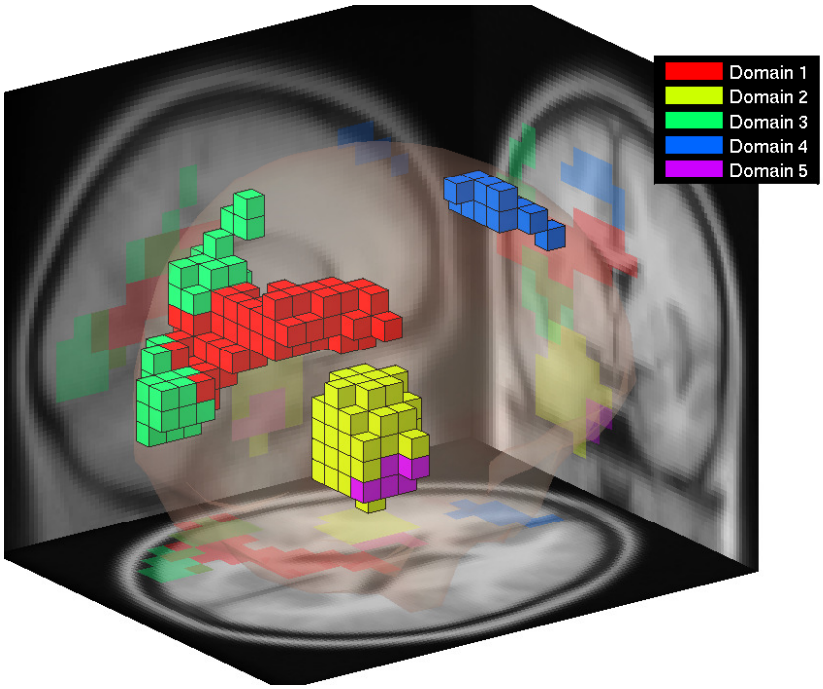
7.4 Results

Affinity propagation clustering initially produced 15 domains. Only 5 domains remained after removal of very small domains including less than 6 voxels. See Fig. 7.1b. The previous IC-based spectral analysis of concussion reported significant group differences in some frontal brain areas [77]. Similarly, we found a cluster located in frontal cortex (Domain 4, centered in the superior frontal gyrus), that contained 48 ICs, 13 ICs from 10 or the 21 concussed subjects (%47.61), and 36 ICs from 21 (%63.63) of the control subjects. Scalp maps of some of the ICs contributing to the domain, as well as their associated equivalent dipole locations, are shown in Figs. 7.3 and 7.4. This domain includes voxels from both the right and left hemispheres of the MNI model, although the majority (%85) are on the left side. Mean projected domain spectra for both groups are shown in Fig. 7.2.

We divided the spectrum into frequency bands Ð delta (2-4Hz), theta (4-8Hz), alpha (8-13Hz), beta (13-30Hz), gamma (30-45Hz) Ð and calculated mean log power in these bands for the ICs contributing to each domain. Bootstrap statistics (using 5,000



(a)



(b)

Figure 7.1: (a) Voxels that show significantly consistent spectra among nearby source locations ($p < 0.05$). (b) Domains created by affinity propagation clustering with maximum similarity threshold across domains 0.9. Domain 4 is the only domain in the frontal part of the brain.

iterations) revealed significant group differences for Domain 4 in the delta, theta and beta bands. Contributing ICs from the concussed group had significantly less delta and

theta band power ($p < 0.01$, $p < 0.05$ respectively) but higher beta power ($p < 0.005$) than contributing ICs from the control group.

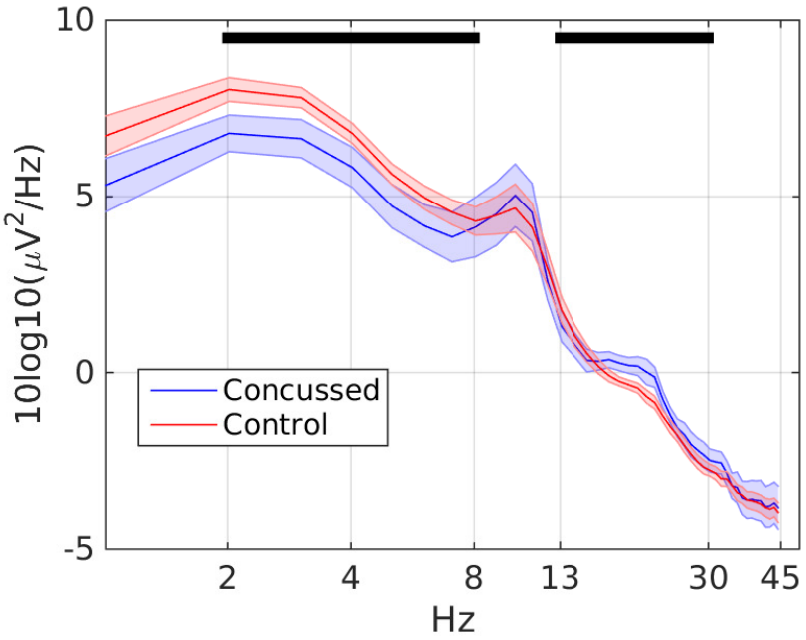


Figure 7.2: Measure-projected IC log spectra for the concussed and control groups at the maximally centered “exemplar" IC for Domain 4. Shaded regions indicate the standard error of the mean. Black lines correspond to spectral bands with a significant group difference.

7.5 Discussion

Our findings partly overlap results of the previous ICA-based concussion study [77]. We also found significantly more beta band power in or near frontopolar cortex in the concussed group during resting state EEG. However, unlike the previous study, we observed significantly less delta and theta power in the concussed group, although EEG spectral slowing has been a commonly reported finding in traumatic brain injury patients. Our finding of power increases and decreases in specific bands in the frontal cortex parallels previous fMRI studies on concussed adolescents that have shown both increases

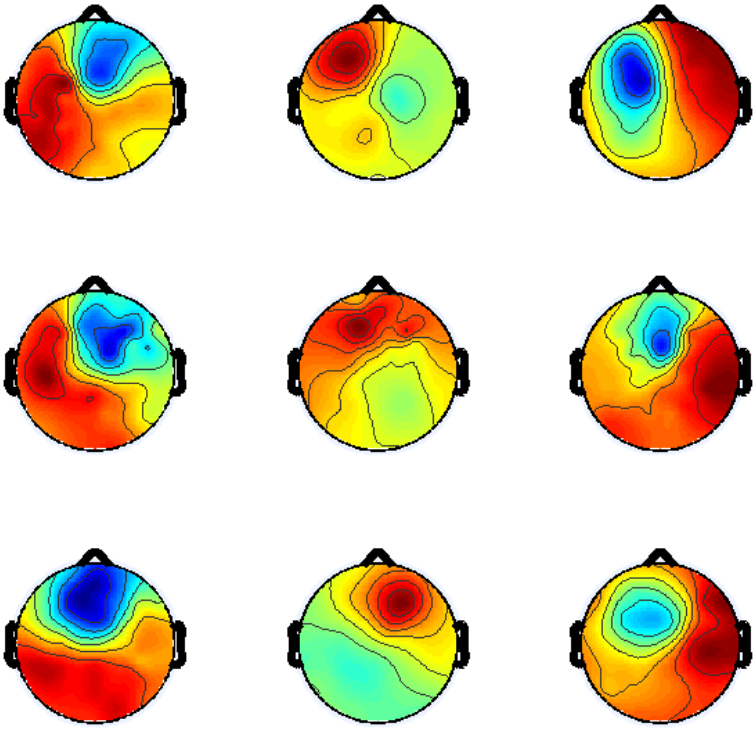


Figure 7.3: Scalp maps of highest contributing 9 sources to Domain 4.

and decreases in functional connectivity within the frontal regions of the brain [15]. These region specific changes in the frontal areas may reflect parallel processes of response to injury and recovery and may be signature of acute concussion. Decreased power at the lower end of the EEG spectrum, together with higher beta power are associated with increased alertness and attentional focus. As pointed out in [86], this can be explained by a mechanism that mTBI patients use to compensate for cognitive deficits arising from their brain injury. Possibly this frequency profile might be used to define a single measure that could differentiate individuals with concussion from still-healthy controls at time of injury, and/or might be used as an index of recovery from mild traumatic brain injury (mTBI).

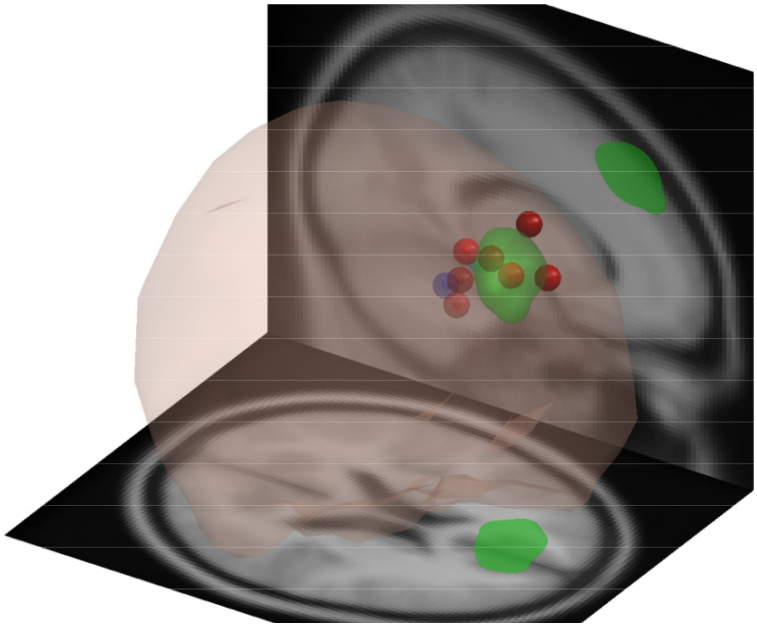


Figure 7.4: DIPFIT localized equivalent dipoles for 9 highest contributing IC's for Domain 4.

7.6 Conclusion

In this study, we investigated the effects of sports-related concussion on source-resolved EEG spectral measures during eyes-closed rest. Unlike most studies in this field, we used ICA decomposition of high-density scalp data to clean EEG data of artifacts and to transform our analysis domain from scalp sensors to brain sources. Rather than using k-means clustering methods for group-level source analysis, we used a probabilistic framework, Measure Projection Analysis, that takes into account expected inaccuracy in source location estimates and inter-subject differences in brain anatomy. Our results suggested that the concussed group had significantly less delta and theta band power and higher beta power in or near a medial frontal brain area.

7.7 Acknowledgements

The text of Chapter 7, in full, is a reprint of the material submitted to Balkan, Ozgur; Virji-Babul, Naznin; Miyakoshi, Makoto; Makeig, Scott; Garudadri, Harinath "Source-domain Spectral EEG Analysis of Sports-Related Concussion via Measure Projection Analysis", EMBC 2015, IEEE.

Chapter 8

Conclusion

In this work, we investigated the domains of support recovery and dictionary learning in sparse linear inverse models in the presence of uncorrelated/independent sources. We found that multiple-measurement Sparse Bayesian Learning (M-SBL) is able to recover the support set for uncorrelated sources even though there may be more sources active than sensors. We theoretically provided exact support recovery conditions for this case. In the MMV model, we developed a robust algorithm as a result of the modification of Sparse Bayesian Learning. Derived algorithm is able to find the support recovery in the presence of outliers.

We modeled the EEG source identification problem as consisting of multiple MMV problems with same unknown dictionary. Since most EEG sources are assumed to be independent and thus uncorrelated, our new dictionary learning framework can reveal the mixing matrix even when more sources than sensors are active in any given EEG data segment. This is done via first transforming the problem to the covariance-domain and learning a transformed dictionary. Actual dictionary is then derived through an inverse transformation. We showed our method's superiority to other overcomplete ICA models on simulated EEG and real EEG. We provided locally-complete methods to identify EEG

local non-stationarities that can be used in low-density as well as high-density EEG.

Later, we provided algorithms to extract independent sources given a multi-channel mixture of a data segment and an overcomplete dictionary to choose the sources from. This problem is different than the MMV sparse model, since we desire that sources carry maximal independence rather than maximal sparsity. The idea of seeking independent sources is again motivated by the nature of most EEG sources and unknown level of sparsity. We have shown that our algorithms can identify independent sources in overcomplete setting through simulations and on real EEG. We showed our algorithms connections to M-SBL.

Bibliography

[1] Zeynep Akalin Acar and Scott Makeig. Effects of forward model errors on eeg source localization. *Brain topography*, 26(3):378–396, 2013.

[2] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322, 2006.

[3] Laurent Albera, Amar Kachenoura, Pierre Comon, Ahmad Karfoul, Fabrice Wendling, Lotfi Senhadji, and Isabelle Merlet. Ica-based eeg denoising: a comparative analysis of fifteen methods. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 60(3):407–418, 2012.

[4] A. Alfons, C. Croux, and S. Gelper. Sparse least trimmed squares regression. *Available at SSRN 1967418*, 2011.

[5] Shun-Ichi Amari. Natural gradient learning for over-and under-complete bases in ica. *Neural Computation*, 11(8):1875–1883, 1999.

[6] Ozgur Balkan, Kenneth Kreutz-Delgado, and Scott Makeig. Localization of more sources than sensors via jointly-sparse bayesian learning. *Signal Processing Letters, IEEE*, 21(2):131–134, 2014.

[7] A.S. Bandeira, E. Dobriban, D.G. Mixon, and W.F. Sawin. Certifying the restricted isometry property is hard. *Information Theory, IEEE Transactions on*, 59(6):3448–3450, 2013.

[8] R.G. Baraniuk, E. Candes, M. Elad, and Y. Ma. Applications of sparse representation and compressive sensing [scanning the issue]. *Proceedings of the IEEE*, 98(6):906–909, 2010.

[9] William B Barr, Leslie S Prichep, Robert Chabot, Matthew R Powell, and Michael McCrea. Measuring brain electrical activity to track recovery from sport-related concussion. *Brain Injury*, 26(1):58–66, 2012.

[10] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *Information Theory, IEEE Transactions on*, 57(2):764–785, 2011.

[11] Anthony J. Bell and Terrence J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *NEURAL COMPUTATION*, 7:1129–1159, 1995.

[12] Adel Belouchrani, Karim Abed-Meraim, J-F Cardoso, and Eric Moulines. A blind source separation technique using second-order statistics. *Signal Processing, IEEE Transactions on*, 45(2):434–444, 1997.

[13] Nima Bigdely-Shamlo, Ken Kreutz-Delgado, Christian Kothe, and Scott Makeig. Eyecatch: Data-mining over half a million eeg independent components to construct a fully-automated eye-component detector. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 5845–5848. IEEE, 2013.

[14] Nima Bigdely-Shamlo, Tim Mullen, Kenneth Kreutz-Delgado, and Scott Makeig. Measure projection analysis: a probabilistic approach to eeg source comparison and multi-subject inference. *Neuroimage*, 72:287–303, 2013.

[15] Michael Borich, Aliya-Nur Babul, Po Hsiang Yuan, Lara Boyd, and Naznin Virji-Babul. Alterations in resting-state brain networks in concussed adolescent athletes. *Journal of neurotrauma*, 2014.

[16] Petros T Boufounos, Paris Smaragdis, and Bhiksha Raj. Joint sparsity models for wideband array processing. In *SPIE Optical Engineering+ Applications*, pages 81380K–81380K. International Society for Optics and Photonics, 2011.

[17] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[18] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[19] Jean-FranÂĎois Cardoso. Infomax and maximum likelihood for blind source separation, 1997.

[20] Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 1st edition, 2010.

[21] Shane F. Cotter, Bhaskar D. Rao, Kjersti Engan, Kenneth Kreutz-delgado, and Senior Member. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Trans. Signal Processing*, pages 2477–2488, 2005.

[22] Wei Dai and Olgica Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *Information Theory, IEEE Transactions on*, 55(5):2230–2249, 2009.

[23] Wei Dai, Tao Xu, and Wenwu Wang. Simultaneous codeword optimization (simco) for dictionary update and learning. *Signal Processing, IEEE Transactions on*, 60(12):6340–6353, 2012.

[24] Mike E Davies and Yonina C Eldar. Rank awareness in joint sparse recovery. *Information Theory, IEEE Transactions on*, 58(2):1135–1146, 2012.

[25] Lieven De Lathauwer, Joséphine Castaing, and Jean-François Cardoso. Fourth-order cumulant-based blind identification of underdetermined mixtures. *Signal Processing, IEEE Transactions on*, 55(6):2965–2973, 2007.

[26] Arnaud Delorme and Scott Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21, 2004.

[27] Arnaud Delorme, Tim Mullen, Christian Kothe, Zeynep Akalin Acar, Nima Bigdely-Shamlo, Andrey Vankov, and Scott Makeig. Eeglab, sift, nft, bcilab, and erica: new tools for advanced eeg processing. *Computational intelligence and neuroscience*, 2011:10, 2011.

[28] Arnaud Delorme, Jason Palmer, Julie Onton, Robert Oostenveld, and Scott Makeig. Independent eeg sources are dipolar. *PLoS ONE*, 7, 02 2012.

[29] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[30] Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.

[31] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, 15(12):3736–3745, 2006.

[32] Yonina C Eldar and Moshe Mishali. Robust recovery of signals from a structured union of subspaces. *Information Theory, IEEE Transactions on*, 55(11):5302–5316, 2009.

[33] Yonina C Eldar and Holger Rauhut. Average case analysis of multichannel sparse recovery using convex relaxation. *Information Theory, IEEE Transactions on*, 56(1):505–519, 2010.

[34] Kjersti Engan, Sven Ole Aase, and J Hakon Husoy. Method of optimal directions for frame design. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 5, pages 2443–2446. IEEE, 1999.

[35] M Faul, L Xu, M.M. Wald, and V.G. Coronado. *Traumatic brain injury in the United States: emergency department visits, hospitalizations and deaths 2002âĂŞ2006*. Centers for Disease Control, National Center for Injury Prevention and Control: Atlanta, GA, 2010.

[36] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[37] Karl Friston, Lee Harrison, Jean Daunizeau, Stefan Kiebel, Christophe Phillips, Nelson Trujillo-Barreto, Richard Henson, Guillaume Flandin, and JÃŇÂl'rÃŇÂl'mie Mattout. Multiple sparse priors for the m/eeg inverse problem. *NeuroImage*, 39(3):1104 – 1120, 2008.

[38] Quan Geng and John Wright. On the local correctness of âĎŞ 1-minimization for dictionary learning. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 3180–3184. IEEE, 2014.

[39] Irina F Gorodnitsky and Bhaskar D Rao. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *Signal Processing, IEEE Transactions on*, 45(3):600–616, 1997.

[40] Soren Hauberg, Aasa Feragen, and Michael J Black. Grassmann averages for scalable robust pca. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3810–3817. IEEE, 2014.

[41] S. Haufe, R. Tomioka, G. Nolte, K.-R. MÅÿ andller, and M. Kawanabe. Modeling sparse connectivity between underlying brain sources for eeg/meg. *Biomedical Engineering, IEEE Transactions on*, 57(8):1954 –1963, aug. 2010.

[42] Christopher Hillar and Friedrich T Sommer. When can dictionary learning uniquely recover sparse data from subsamples? *arXiv preprint arXiv:1106.3616*, 2011.

[43] Sheng-Hsiou Hsu, Tim Mullen, Tzyy-Ping Jung, and Gert Cauwenberghs. Online recursive independent component analysis for real-time source separation of high-density eeg. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 3845–3848. IEEE, 2014.

[44] M.M. Hyder and K. Mahata. A robust algorithm for joint-sparse recovery. *Signal Processing Letters, IEEE*, 16(12):1091 –1094, dec. 2009.

[45] D.A. Jackson and Y. Chen. Robust principal component analysis and outlier detection with ecological data. *Environmetrics*, 15(2):129–139, 2004.

[46] Yuzhe Jin and Bhaskar D Rao. Support recovery of sparse signals in the presence of multiple measurement vectors. *arXiv preprint arXiv:1109.1895*, 2011.

[47] Tzyy-Ping Jung, Scott Makeig, Colin Humphries, Te-Won Lee, Martin J Mckeown, Vicente Iragui, and Terrence J Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(02):163–178, 2000.

[48] Tzyy-Ping Jung, Scott Makeig, Martin J McKeown, Anthony J Bell, Te-Won Lee, and Terrence J Sejnowski. Imaging brain dynamics using independent component analysis. *Proceedings of the IEEE*, 89(7):1107–1122, 2001.

[49] Jong Min Kim, Ok Kyun Lee, and Jong Chul Ye. Compressive music: revisiting the link between compressive sensing and array signal processing. *Information Theory, IEEE Transactions on*, 58(1):278–301, 2012.

[50] Jong Min Kim, Ok Kyun Lee, and Jong Chul Ye. Improving noise robustness in subspace-based joint sparse recovery. *Signal Processing, IEEE Transactions on*, 60(11):5799–5809, 2012.

[51] Kenneth Kreutz-Delgado, Joseph F Murray, Bhaskar D Rao, Kjersti Engan, Te-Won Lee, and Terrence J Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396, 2003.

[52] Quoc V Le, Alexandre Karpenko, Jiquan Ngiam, and Andrew Y Ng. Ica with reconstruction cost for efficient overcomplete feature learning. In *Advances in Neural Information Processing Systems*, pages 1017–1025, 2011.

[53] Jonathan Le Roux, Petros T Boufounos, Kang Kang, and John R Hershey. Source localization in reverberant environments using sparse optimization. *ICASSP*, 2013.

[54] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 609–616. ACM, 2009.

[55] Kiryung Lee, Yoram Bresler, and Marius Junge. Subspace methods for joint sparse recovery. *Information Theory, IEEE Transactions on*, 58(6):3613–3641, 2012.

[56] Te-Won Lee, Michael S Lewicki, and Terrence J Sejnowski. Ica mixture models for unsupervised classification of non-gaussian classes and automatic context switching in blind signal separation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(10):1078–1089, 2000.

[57] Michael S. Lewicki, Terrence J. Sejnowski, and Howard Hughes. Learning overcomplete representations. *Neural Computation*, 12:337–365, 1998.

[58] Yuanqing Li, Andrzej Cichocki, and Shun-ichi Amari. Analysis of sparse representation and blind source separation. *Neural Comput.*, 16(6):1193–1234, June 2004.

[59] Yuanqing Li, Andrzej Cichocki, Shun-ichi Amari, Sergei Shishkin, Jianting Cao, and Fanji Gu. Sparse representation and its applications in blind source separation. *Advances in neural information processing systems*, 16:241, 2004.

[60] Jui-Chieh Liao, Wei-Yeh Shih, Kuan-Ju Huang, and Wai-Chi Fang. An online recursive ica based real-time multichannel eeg system on chip design with automatic eye blink artifact rejection. In *VLSI Design, Automation, and Test (VLSI-DAT), 2013 International Symposium on*, pages 1–4. IEEE, 2013.

[61] Wei Lu and Jagath C. Rajapakse. Approach and applications of constrained ica. *IEEE Trans. Neural Netw*, 16(1):203–212, 2005.

[62] David J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1991.

[63] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.

[64] Scott Makeig, Anthony J Bell, Tzyy-Ping Jung, Terrence J Sejnowski, et al. Independent component analysis of electroencephalographic data. *Advances in neural information processing systems*, pages 145–151, 1996.

[65] Scott Makeig, Tzyy-Ping Jung, Anthony J Bell, Dara Ghahremani, and Terrence J Sejnowski. Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences*, 94(20):10979–10984, 1997.

[66] Scott Makeig, Julie Onton, et al. Erp features and eeg dynamics: an ica perspective. *Oxford Handbook of Event-Related Potential Components. New York, NY: Oxford*, 2009.

[67] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.

[68] Brenton W McMenamin, Alexander J Shackman, Jeffrey S Maxwell, David RW Bachhuber, Adam M Koppenhaver, Lawrence L Greischar, and Richard J Davidson. Validation of ica-based myogenic artifact correction for scalp and source-localized eeg. *Neuroimage*, 49(3):2416–2432, 2010.

[69] Tim Mullen, Christian Kothe, Yu Mike Chi, Alejandro Ojeda, Trevor Kerth, Scott Makeig, Gert Cauwenberghs, and Tzyy-Ping Jung. Real-time modeling and

3d visualization of source dynamics and connectivity using wearable eeg. In *Conference proceedings:... IEEE EMBC, EMBS*, volume 2013, page 2184. NIH Public Access, 2013.

[70] Deanna Needell and Joel A Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[71] Robert Oostenveld, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011, 2010.

[72] Piya Pal and PP Vaidyanathan. Pushing the limits of sparse support recovery using correlation information. *Signal Processing, IEEE Transactions on*, 2015.

[73] Jason Palmer, David Wipf, Kenneth Kreutz-Delgado, and Bhaskar Rao. Variational em algorithms for non-gaussian latent variable models. *Advances in neural information processing systems*, 18:1059, 2006.

[74] Jason A. Palmer, Scott Makeig, Kenneth Kreutz-Delgado, and Bhaskar D. Rao. Newton method for the ica mixture model. In *ICASSP*, pages 1805–1808. IEEE, 2008.

[75] Jason T Parker, Philip Schniter, and Volkan Cevher. Bilinear generalized approximate message passing. *arXiv preprint arXiv:1310.2632*, 2013.

[76] Roberto Domingo Pascual-Marqui et al. Standardized low-resolution brain electromagnetic tomography (sloreta): technical details. *Methods Find Exp Clin Pharmacol*, 24(Suppl D):5–12, 2002.

[77] VA Ponomarev, OE Gurskaia, IuD Kropotov, LV Artiushkova, and A Muller. [the comparison of clustering methods of eeg independent components in healthy subjects and patients with post concussion syndrome after traumatic brain injury]. *Fiziologiia cheloveka*, 36(2):5–14, 2009.

[78] A. Rakotomamonjy. Surveying and comparing simultaneous sparse approximation (or group-lasso) algorithms. *Signal Processing*, 91(7):1505 – 1526, 2011.

[79] Alain Rakotomamonjy. Direct optimization of the dictionary learning problem. *Signal Processing, IEEE Transactions on*, 61(22):5495–5506, 2013.

[80] Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 2168–2172. IEEE, 2011.

[81] Peter J. Rousseeuw and Katrien Driessen. Computing LTS regression for large data sets. *Data Min. Knowl. Discov.*, 12(1):29–45, January 2006.

[82] P.J. Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

[83] Mostafa Sadeghi, Massoud Babaie-Zadeh, and Christian Jutten. Learning over-complete dictionaries based on atom-by-atom updating. *Signal Processing, IEEE Transactions on*, 62(4):883–891, 2014.

[84] Karl Skretting and Kjersti Engan. Recursive least squares dictionary learning algorithm. *Signal Processing, IEEE Transactions on*, 58(4):2121–2130, 2010.

[85] Semyon Slobounov, Wayne Sebastianelli, and Mark Hallett. Residual brain dysfunction observed one year post-mild traumatic brain injury: combined eeg and balance study. *Clinical Neurophysiology*, 123(9):1755–1761, 2012.

[86] Semyon M Slobounov and Wayne Sebastianelli. *Concussions in Athletics: From Brain to Behavior*. Springer Science & Business Media, 2014.

[87] Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 3087–3090. AAAI Press, 2013.

[88] James V. Stone and John Porrill. Undercomplete independent component analysis for signal separation and dimension reduction. Technical report, 1997.

[89] RW Thatcher, C Biver, JF Gomez, D North, R Curtin, RA Walker, and A Salazar. Estimation of the eeg power spectrum using mri t 2 relaxation time in traumatic brain injury. *Clinical Neurophysiology*, 112(9):1729–1745, 2001.

[90] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *The journal of machine learning research*, 1:211–244, 2001.

[91] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655–4666, 2007.

[92] Yijun Wang and Tzyy-Ping Jung. Improving brain–computer interfaces using independent component analysis. In *Towards Practical Brain-Computer Interfaces*, pages 67–83. Springer, 2013.

[93] David Wipf and Srikantan Nagarajan. A unified bayesian framework for meg/eeg source imaging. *NeuroImage*, 44(3):947 – 966, 2009.

[94] David P. Wipf, Julia P. Owen, Hagai Attias, Kensuke Sekihara, and Srikantan S. Nagarajan. Estimating the location and orientation of complex, correlated neural activity using meg. In *NIPS*, pages 1777–1784, 2008.

[95] David P. Wipf, Julia P. Owen, Hagai T. Attias, Kensuke Sekihara, and Srikantan S. Nagarajan. Robust bayesian estimation of the location, orientation, and time course of multiple correlated neural sources using meg. *NeuroImage*, 49(1):641 – 655, 2010.

[96] David P Wipf and Bhaskar D Rao. Sparse bayesian learning for basis selection. *Signal Processing, IEEE Transactions on*, 52(8):2153–2164, 2004.

[97] David P Wipf and Bhaskar D Rao. An empirical bayesian strategy for solving the simultaneous sparse approximation problem. *Signal Processing, IEEE Transactions on*, 55(7):3704–3716, 2007.

[98] David Paul Wipf. *Bayesian methods for finding sparse representations*. ProQuest, 2006.

[99] Tao Xu and Wenwu Wang. A compressed sensing approach for underdetermined blind audio source separation with sparse representation. In *Statistical Signal Processing, 2009. SSP'09. IEEE/SP 15th Workshop on*, pages 493–496. IEEE, 2009.

[100] Tao Xu and Wenwu Wang. A block-based compressed sensing method for underdetermined blind speech separation incorporating binary mask. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 2022–2025. IEEE, 2010.

[101] Tao Xu and Wenwu Wang. Methods for learning adaptive dictionary in underdetermined speech separation. In *Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on*, pages 1–6. IEEE, 2011.

[102] Mehrdad Yaghoobi, Thomas Blumensath, and Mike E Davies. Dictionary learning for sparse approximations with the majorization method. *Signal Processing, IEEE Transactions on*, 57(6):2178–2191, 2009.

[103] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang. Coupled dictionary training for image super-resolution. *Image Processing, IEEE Transactions on*, 21(8):3467–3478, 2012.

[104] Ming-Chun Yang, Chao-Tsung Chu, and Y-CF Wang. Learning sparse image representation with support vector regression for single-image super-resolution. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 1973–1976. IEEE, 2010.

[105] Rafal Zdunek and Andrzej Cichocki. Improved m-focuss algorithm with overlapping blocks for locally smooth sparse signals. *Signal Processing, IEEE Transactions on*, 56(10):4752–4761, 2008.

[106] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Robert Fergus. Deconvolutional networks. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2528–2535. IEEE, 2010.

[107] Z. Zhang and B. D. Rao. Exploiting Correlation in Sparse Signal Recovery Problems: Multiple Measurement Vectors, Block Sparsity, and Time-Varying Sparsity. *ArXiv e-prints*, May 2011.

[108] Michael Zibulevsky and Barak A Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural computation*, 13(4):863–882, 2001.