A Dynamic Approach Solving Undercomplete Noisy Independent Component Analysis

by

Bowen Zhang

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2024

Date of final oral examination: 5/8/2024

The dissertation is approved by the following members of the Final Oral Committee:

Bret Larget, Professor, Statistics

Richard Chappell, Professor, Statistics, Biostatistics and Medical Informatics

Sameer Deshpande, Assistant Professor, Statistics

Keith Levin, Assistant Professor, Statistics

Jun Zhu, Professor, Statistics

To my Family

Acknowledgments

I would like to express my deepest gratitude to my advisor, Professor Bret Larget, whose unwavering support and guidance were indispensable to the completion of this thesis. Professor Larget's insightful advice and constructive criticism helped me navigate a lot of academic challenges I faced. I am incredibly fortunate to have been advised by Professor Larget. His warmth, patience, and dedication greatly enhanced my experience in grad school, and from him, I learned not only academically but also how to be a kind and effective mentor, helping mentees maximize their potential.

I would also like to extend my deepest gratitude to the members of my defense committee. Professor Levin provided constructive and comprehensive feedback on my dissertation, which greatly enhanced my work. Professor Jun Zhu was very supportive while I faced challenges and also offered critical suggestions for my dissertation. Her kind words and persistent encouragement inspired me to overcome obstacles. Professors Rick Chappell and Sameer Deshpande also provided very inspirational feedback on my dissertation, which greatly aided my research.

I am extremely grateful to Professor Zhongdang Pan and Linzhu Tian for being around and supporting me like family.

I would also like to thank Professor Chunming Zhang, for her guidance on this project and for the research assistantship she provided me in the Fall semester of 2023.

I also want to thank University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

Thank all of my peers and friends I met in Madison, Muhong Gao, Yongfeng Wu, Taiyu Ye, Rungang Han, Zheng Liu, Shan Lu, Yuetian Luo, Ting Fung Ma, Xinran Miao, Zijian Ni, Junhyung, Chang, Yanbo Shen, Sheng Wang, Muzhe Zeng, Chengning Zhang, Siyu Zhang, Zihao Zheng, and Xiaobin Zhou. My special thanks go to Yongfeng Wu, Muhong Gao, and Taiyu Ye for their valuable suggestions on my dissertation. Taiyu, in particular, provided comprehensive assistance with checking my proofs. Additionally, they all offered me substantial emotional support during my challenging periods.

Finally, I want to express my deepest appreciation to my parents for their selfless love and unwavering support. I would also like to extend heartfelt thanks to my wife, Liwei Shen. Throughout numerous obstacles and challenges, she has always been there to support me. Her love has been healing and motivating, encouraging me to persevere. She instills confidence in me, completes my life, and inspires me to become my best self.

Contents

List of Tables

List of Figures

Abstract

Independent Component Analysis (ICA) is widely used in various applications. Traditional ICA algorithms typically employ a two-stage approach $(\mathrm{PCA+ICA})$ to handle the noisy ICA model, which encounter two obstacles. The first one lies in determining the optimal number of principal components to retain, while the second is the potential loss of valuable information during the PCA stage, even if the correct number of components is chosen. In this study, we propose a novel methodological framework called dynamic-ICA (dICA). This dICA approach initiates from an augmented matrix and dynamically selects the number of latent components N^* by applying a group regularization technique to this matrix. This method achieves dimension reduction and latent variable estimation simultaneously. In addition, by dynamically estimating N^* instead of pre-specifying it, dICA avoids information loss associated with the two-stage approach. Thirdly, dICA incorporates a contrast function resistant to additive Gaussian noise, enabling dICA to be robust against Gaussian noise. The theoretical underpinnings of the dICA framework are rigorously established, and algorithms for both the population model (infinite sample size) and the sample model (finite sample size) are developed. For the population model, we demonstrate that our algorithm achieves a locally at least linear convergence rate. The effectiveness of the algorithm for the sample model is validated through extensive numerical simulations and real-world applications in EEG analysis. These applications reveal that dICA is capable of uncovering more biologically plausible signals and a broader range of brain activities than traditional methods.

Chapter 1

Introduction

Independent Component Analysis (ICA) is a widely used method for finding a linear representation of multidimensional data. The aim of this representation is to ensure that the components are statistically independent of each other. ICA is categorized under unsupervised learning and finds applications in various fields, including medical imaging (McKeown et al., 2003), image processing (Hyvarinen et al., 1998), econometrics (Gouriéroux et al., 2017; Chen et al., 2019), chemistry (Liu et al., 2018), and climate science (Pfister et al., 2019).

The classical ICA model can be expressed as follows:

$$\boldsymbol{X} = \mathbf{A}\boldsymbol{S}, \tag{1.1}$$

where $\boldsymbol{X}$ is a p-dimensional random observational vector, $\mathbf{A}$ is a $p \times N$ (unknown) mixing matrix, and $\boldsymbol{S}$ is an N-dimensional latent random vector consisting of unknown non-Gaussian independent components (ICs). Non-Gaussianity plays a pivotal role in ensuring the identifiability of the ICA model. In practice, we make the assumption that we have a set of n independent and identically distributed (i.i.d.) realizations $\{\boldsymbol{X}(t_i)\}_{i=1}^{n}$ of the random variable $\boldsymbol{X}$ with corresponding latent variables $\{\boldsymbol{S}(t_i)\}$. The main objective of ICA is to estimate $\mathbf{A}$ and $\{\boldsymbol{S}(t_i)\}$ when only $\boldsymbol{X}$ is available.

Although the noise-free ICA model (1.1) is widely acknowledged, it fails to incorporate

the noise term in its model, even though noise is a realistic factor in practical scenarios. In light of this drawback, researchers propose the following noisy ICA model (Cichocki et al., 1998; Beckmann and Smith, 2004):

$$\boldsymbol{X} = \mathbf{A}^* \boldsymbol{S}^* + \boldsymbol{E}, \tag{1.2}$$

where $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ is a p-dimensional random vector, characterized by the covariance matrix $\Sigma_{\boldsymbol{X}}$. The latent component vector $\boldsymbol{S}^* = (S_1^*, \ldots, S_{N^*}^*)^\top$ is a N^*-dimensional random vector comprising mutually independent, non-Gaussian random variables. Without loss of generality, it is assumed that $\mathrm{cov}(\boldsymbol{S}^*) = \mathbf{I}_{N^*}$. $\mathbf{A}^* = (\boldsymbol{a}_1^*, \ldots, \boldsymbol{a}_{N^*}^*) \in \mathbb{R}^{p \times N^*}$ is an unknown non-random mixing matrix with full column rank, $\boldsymbol{E} = (e_1, \ldots, e_p)^\top$ is a Gaussian noise vector with zero-mean and a diagonal covariance matrix $\Sigma_{\boldsymbol{E}}$. When the noise term $\boldsymbol{E}$ approaches zero or is exactly zero, model (1.2) degenerates to model (1.1). When we have n i.i.d. samples of $\boldsymbol{X}$, the objective is to estimate $\mathbf{A}^*$ and the corresponding $\{\boldsymbol{S}^*(t_i)\}$.

Most ICA algorithms assume either the absence of noise or noise that is small enough to be considered negligible. These algorithms estimate $\mathbf{A}$ (or $\mathbf{A}^*$) by estimating its pseudo-inverse $\mathbf{A}^\dagger \in \mathbb{R}^{N \times p}$ (or $\mathbf{A}^{*\dagger} \in \mathbb{R}^{N^* \times p}$). Since the dICA proposed in this thesis mainly focuses on addressing model (1.2), a simpler notation $\mathbf{W}^*$ is used to represent $\mathbf{A}^{*\dagger}$.

The matrices $\mathbf{A}^\dagger$ and $\mathbf{A}^\dagger$ are commonly referred to as the demixing matrix because multiplying $\boldsymbol{X}$ by them can cancel the mixing matrices, thereby separating the mixed $\boldsymbol{X}$. The demixing matrix should be the matrix $\mathbf{W}$ that maximizes the independence among $\mathbf{W}\boldsymbol{X}$. Typically, ICA algorithms measure independence through non-Gaussianity. The underlying concept is that the sum of ICs tends to resemble a Gaussian distribution more closely than the individual ICs themselves (Hyvärinen and Oja, 2000; Risk et al., 2019).

For simplicity, classical ICA algorithms assume a "square" mixing matrix ($p = N$). When $p > N$ (referred to as undercomplete ICA), a preprocessing step using Principal Component Analysis (PCA) is usually performed to reduce dimensions and meet the square mixing matrix assumption. This two-stage PCA + ICA approach has been popular over the

years and works well when there is minimal or no noise. However, this approach encounters challenges when there is stronger noise present.

In the PCA step, due to the presence of $\boldsymbol{E}$, determining the appropriate number of principal components to retain (ideally equal to N^*) poses a challenge. Overestimating N^* can lead to excessive decomposition of latent components, while underestimating N^* may result in the loss of valuable information. In addition, even if N^* is correctly estimated, PCA can still discard weak yet important components (Porrill and Stone, 1998; Green et al., 2002; Risk et al., 2019). The reason for this is that PCA ranks components according to their variance. In certain scenarios, such as EEG data, it is possible for noise to exhibit a higher variance than certain weak sources.

In the ICA step, it should be noted that the application of those noise-free algorithms introduces noise-induced bias into the estimators. The reason is that Maximizing non-Gaussianity in $\mathbf{W}\boldsymbol{X}$ (with respect to $\mathbf{W}$) is not equivalent to maximizing it in $\mathbf{W}(\boldsymbol{X} - \boldsymbol{E})$.

This thesis proposes a novel framework to solve model (1.2) named dynamic-ICA (dICA). This novel dICA framework is a method that formally considers the noise term and mitigates noise bias using bias removal techniques (Hyvarinen, 1999). Furthermore, dICA manages to simultaneously estimate N^* (hence perform dimension reduction) alongside the entries in demixing matrix $\mathbf{W}^*$ through group regularization technique (Yuan and Lin, 2006; Chen et al., 2019). By integrating dimension reduction and demixing matrix estimation into a single step, dICA achieves a mutual enhancement of both processes — improved dimension reduction contributes to more precise demixing matrix estimation, and conversely, accurate demixing matrix estimation aids in effective dimension reduction. This dynamic approach to perform dimension reduction, as opposed to a predetermined approach, enables dICA to overcome some key limitations commonly associated with the PCA + ICA methods, such as the difficulty in determining N^* and potential information loss during the "PCA" step. The ability to adaptively perform dimension reduction in real-time is a significant advantage of dICA over other methodologies, and it is this characteristic that gives rise to its name, dynamic ICA.

Similar to the noise-free ICA approaches, dICA is based on the non-Gaussian nature of the source signals $\boldsymbol{S}$. It also utilizes higher-order statistics to measure the dependency between ICs and hence requires a large sample size n and a relatively low dimensionality (p) to achieve effective results. This requirement is shared with other noise-free ICA techniques (Xia, 2020; Herrmann and Theis, 2007). Furthermore, when $\boldsymbol{S}$ is drawn from distributions that closely resemble the Gaussian distribution, such as a t-distribution with a high degrees of freedom, dICA demands an even larger sample size to demonstrate its capabilities. Under these circumstances, accurately estimating N^* and $\mathbf{W}^*$ may pose challenges for dICA. Therefore, dICA is most powerful when the distribution of $\boldsymbol{S}$ is sufficiently far away from Gaussian, particularly in comparison to that of $\boldsymbol{E}$, and when a sufficient sample size is available.

In the development of the dICA algorithm and its theoretical properties, we initially make an unrealistic assumption. Specifically, we assume that we have access to $\mathrm{E}\{f(\boldsymbol{X})\}$ for any function f, as long as $\mathrm{E}\{|f(\boldsymbol{X})|\}$ exists. This assumption can be interpreted as having an infinite number of samples available for $\boldsymbol{X}$. The ICA model under this assumption is commonly referred to as the population model in the literature (Wei, 2015). Although this population model may be unrealistic, it commonly serves as a useful starting point in the ICA literature, as it allows for the development of algorithms without the concern for "finite sampling noise" (Hyvärinen and Oja, 2000; Wei, 2015). In practical scenarios where only a finite number of samples are available, the expectations will be replaced by sample averages and the necessary adjustments will be made. In this thesis, we also begin the development of dICA using the population model.

1.1. Motivation for dICA

To motivate the dICA method, we first review the approach solving the noise-free ICA model. For the noise-free ICA model described in (1.1) with $\mathrm{cov}(\boldsymbol{S}) = \mathbf{I}_N$, the typical

method solving it involves solving an optimization problem, which can be formalized as:

$$\begin{aligned} \widehat{\mathbf{A}^{\dagger}} = \quad & \underset{\mathbf{W}_N \in \mathbb{R}^{N \times p}}{\operatorname{argmin}} \quad -\mathcal{J}(\mathbf{W}_N \boldsymbol{X}), \\ & \text{subject to} \quad \mathbf{W}_N \Sigma_{\boldsymbol{X}} \mathbf{W}_N^{\top} = \mathbf{I}_N, \end{aligned} \tag{1.3}$$

where $\mathcal{J}(\cdot)$ represents a non-linear contrast function that measures independence. Any $\mathbf{W}_N$ satisfying constraint $\mathbf{W}_N \Sigma_{\boldsymbol{X}} \mathbf{W}_N^{\top} = \mathbf{I}_N$ differs from $\mathbf{A}^{\dagger}$ by an orthonormal matrix $\mathbf{O} \in \mathbb{R}^{N \times N}$. Hence the constraint is usually referred to as the orthogonal constraint. Extending this optimization framework to the noisy ICA model in (1.2) and considering the fact that $\mathbf{A}^* \boldsymbol{S}^* = \boldsymbol{X} - \boldsymbol{E}$, an analogous form of (1.3) can be posited as follows:

$$\begin{aligned} \widehat{\mathbf{W}} = \quad & \underset{\mathbf{W} \in \mathbb{R}^{N^* \times p}}{\operatorname{argmin}} \quad -\mathcal{J}(\mathbf{W}(\boldsymbol{X} - \boldsymbol{E})), \\ & \text{subject to} \quad \mathbf{W}(\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{E}}) \mathbf{W}^{\top} = \mathbf{I}_{N^*}. \end{aligned} \tag{1.4}$$

However, given the unknowable nature of N^*, $\boldsymbol{E}$, and $\Sigma_{\boldsymbol{E}}$, equation (1.4) is unsolvable. To overcome these obstacles, we formulate dICA as follows:

$$\begin{aligned} \widehat{\widetilde{\mathbf{W}}}, \widehat{\Sigma}_{\boldsymbol{E}} \quad = \quad & \arg \min_{\substack{\widetilde{\mathbf{W}} \in \mathbb{R}^{p \times p}, \\ \operatorname{diag}(\Sigma) \geq \mathbf{0}, \\ \Sigma \text{ is diagonal}}} \quad -\mathcal{J}_{\text{Rob}}(\widetilde{\mathbf{W}} \boldsymbol{X}) + \|\widetilde{\mathbf{W}}\|_F^2, \\ & \text{subject to} \quad (\Sigma_{\boldsymbol{X}} - \Sigma) \widetilde{\mathbf{W}}^{\top} \widetilde{\mathbf{W}} (\Sigma_{\boldsymbol{X}} - \Sigma) = (\Sigma_{\boldsymbol{X}} - \Sigma), \\ & \qquad \operatorname{vec}(\Sigma_{\boldsymbol{X}} - \Sigma) \in \mathcal{A}. \end{aligned} \tag{1.5}$$

Here $\mathcal{J}_{\text{Rob}}(\cdot)$ is a robust-to-noise Gaussian moments function (Hyvarinen, 1999) satisfying $\mathcal{J}_{\text{Rob}}(\widetilde{\mathbf{W}} \boldsymbol{X}) = \mathcal{J}(\widetilde{\mathbf{W}}(\boldsymbol{X} - \boldsymbol{E}))$. And $\widetilde{\mathbf{W}}$ is an augmented matrix in $\mathbb{R}^{p \times p}$. The linear space $\mathcal{A}$ satisfies

$$\begin{aligned} \mathcal{A} &\coloneqq \{\operatorname{vec}(\mathbf{A}^* \mathbf{D} \mathbf{A}^{*\top}) : \mathbf{D} \in \mathbb{R}^{N^* \times N^*} \text{ is any diagonal matrix}\} \\ &= \operatorname{span}\{\operatorname{vec}(\boldsymbol{a}_1^* \boldsymbol{a}_1^{*\top}), \ldots, \operatorname{vec}(\boldsymbol{a}_{N^*}^* \boldsymbol{a}_{N^*}^{*\top})\} \subseteq \mathbb{R}^{p^2}. \end{aligned}$$

And $\mathrm{vec}(\cdot)$ is an operator flattening a matrix into a vector column by column.

The dICA approach starts with considering the optimization problem on the $p \times p$ augmented matrices set $\{\widetilde{\mathbf{W}} : \widetilde{\mathbf{W}} \in \mathbb{R}^{p\times p}\}$ and hence eliminates N^*. Inspired from Lietzén et al. (2020), a $p \times p$ matrix $\widetilde{\mathbf{W}}$ can be considered as an augmented demixing matrix if it satisfies $\widetilde{\mathbf{W}}\mathbf{A}^* = (\mathbf{I}_{N^*}, \mathbf{0})^\top$. When $p > N^*$, there will be infinitely many augmented demixing matrices. For notation simplicity, we define the set containing all the augmented demixing matrices as $\widetilde{\mathcal{W}}^* := \{\widetilde{\mathbf{W}} : \widetilde{\mathbf{W}}\mathbf{A}^* = (\mathbf{I}_{N^*}, \mathbf{0})^\top, \widetilde{\mathbf{W}} \in \mathbb{R}^{p\times p}\}$. Among all those matrices in $\widetilde{\mathcal{W}}^*$, $\widetilde{\mathbf{W}}^* := (\mathbf{W}^{*\top}, \mathbf{0})^\top$ is of most interest to us because it has a sparse structure and we can directly find $\mathbf{W}^*$ if $\widetilde{\mathbf{W}}^*$ becomes available. Through this thesis, we will name the non-zero rows in $\widetilde{\mathbf{W}}^*$ as significant rows and the zero rows in it as nuisance rows. In order to estimate $\widetilde{\mathbf{W}}^*$, dICA does the followings:

- Adopting a robust-to-noise $\mathcal{J}_{\mathrm{Rob}}(\cdot)$.
- Estimating $\Sigma_{\boldsymbol{E}}$ using the relationship $\mathrm{vec}(\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{E}}) \in \mathcal{A}$.
- Developing a hidden orthogonal constraint $(\Sigma_{\boldsymbol{X}} - \Sigma)\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{W}}(\Sigma_{\boldsymbol{X}} - \Sigma) = (\Sigma_{\boldsymbol{X}} - \Sigma)$ such that any $\widetilde{\mathbf{W}}$ satisfying the constraint will differ from $\widetilde{\mathcal{W}}^*$ by an orthonormal matrix. This constraint functions analogously to the constraint in equation (1.4). However, the one in equation (1.5) is specifically tailored for the augmented matrix $\widetilde{\mathbf{W}}$ and is notably independent of the dimension N^*.
- Incorporating a group regularization term (Yuan and Lin, 2006) to encourage sparse structure in $\widehat{\widetilde{\mathbf{W}}}$.

In Section 2, we will show that the resulting matrix $\widehat{\widetilde{\mathbf{W}}}$ equals $\widetilde{\mathbf{W}}^*$.

1.2. Related work

Recognizing the limitations in the noise-free ICA model (1.1), researchers have conducted a series of studies to address them.

These techniques generally fall into three distinct categories. The first category solely aims to determine the N^* in model (1.2). This is usually achieved by PCA. However, there is no universal guideline on how many principle components should be kept. Usually, the number is chosen to be the one that retains 90% to 99% of variance. A statistically solid test is proposed by Virta and Nordhausen (2019), where they develop an asymptotic test for N^* based on the assumption that $\Sigma_{\boldsymbol{E}} = \sigma^2\mathbf{I}$.

The second category aims to identify not only N^*, but also $\mathbf{A}^*$ and $\boldsymbol{S}^*$. This category still performs similarly to the two stage PCA + ICA approach. The difference is that it involves employing more sophisticated preprocessing methods as alternatives to PCA to mitigate noise effects. For instance, Ikeda and Toyama (2000) uses factor analysis as a replacement of PCA, while Beckmann and Smith (2004) utilizes probabilistic PCA. Following the preprocessing of $\boldsymbol{X}$, traditional noise-free ICA is applied to the preprocessed $\boldsymbol{X}$, denoted by $\breve{\boldsymbol{X}}$, to estimate the $\mathbf{W}^*$. While the aforementioned two works achieve satisfactory results in magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) data sets, the designs of their approaches do not formally accommodate the bias caused by the noise term $\boldsymbol{E}$. The reason is that $\breve{\boldsymbol{X}}$, as a linear projection of $\boldsymbol{X}$, will inevitably contain noise no matter how sophisticated the preprocessing step is. Consequently, the effectiveness of these approaches highly relies on the assumption that $\boldsymbol{E}$ has a much smaller magnitude comparing with $\mathbf{A}^*\boldsymbol{S}^*$.

The third category assumes N^* is known and mainly focus on the estimation of $\mathbf{A}^*$ and $\boldsymbol{S}^*$. Most cumulant-based ICA algorithms fall in this category. Yeredor (2000) introduces a method utilizing the second-order derivative of the characteristic function, while De Lathauwer et al. (1996) advocates for a higher-order-only approach that estimates $\mathbf{W}^*$ by jointly diagonalizing cumulant matrices. The Fourth-Order Blind Identification (FOBI) algorithm, developed by Cardoso (1991), leverages higher-order tensor structures. Bonhomme and Robin (2009) develop an algorithm that can identify $\mathbf{W}^*$ using second-to-fourth-order moments. They also show the consistency of their estimator. Methods employing derivatives of the second characteristic function have a lower computational cost but their efficacy critically

depends on the chosen "processing points" for derivative calculations. The selection of these "processing points" is itself a challenging enough issue (Yeredor, 2009). The latter three approaches, reliant on higher order statistics or tensor structures, impose a greater computational load compared to popular noise-free ICA algorithms such as FastICA (Hyvärinen and Oja, 2000).

Another direction within the third category involves modifications of FastICA to accommodate the noise model in Equation (1.2). The FastICA algorithm optimizes $\mathcal{J}(\cdot)$ through an efficient fixed-point algorithm and enjoys a second-order convergence rate. This line of work can be traced back to Hyvarinen (1999) where $\Sigma_{\boldsymbol{E}}$ is assumed to be known. In this approach, the prewhitening is executed using $\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{E}}$ instead of $\Sigma_{\boldsymbol{X}}$ and the subsequent demixing process adopts Gaussian moments functions which remove noise-induced bias. The prior knowledge of $\Sigma_{\boldsymbol{E}}$ is a rather restrictive requirement and Arora et al. (2012) design a approach that relaxes this stringent prerequisite. The quasi-whitening step there only requires that all elements in $\boldsymbol{S}$ have positive kurtosis. Later on, Voss et al. (2013) manages to relax the positive-kurtosis assumption and develops a gradient iteration algorithm enjoying cubic convergence rate. Notably, the quasi-whitening step in both studies depends on the fourth-order cumulant, known for its sensitivity to outliers and the need for a substantial sample size to yield satisfactory outcomes. The simulation results in Voss et al. (2013) also show that their method is outperformed by traditional noise-free ICA methods under a small or moderate sample size. In an effort to bypass the quasi-whitening process, Voss et al. (2015) develops a Pseudo-Euclidean Gradient Iteration (PEGI) approach. Despite this, the methodologies advanced by Arora et al. (2012); Voss et al. (2013, 2015) all employ fourth order statistics as the nonlinear contrast function. Consequently, these methods are particularly sensitive to low-probability outliers in $\boldsymbol{E}$ and need larger sample sizes to affirm their effectiveness.

1.3. Main Contributions

The proposed dICA approach has the following contributions:

- The dICA framework explicitly integrates the noise component within its analytical model, applying bias removal techniques to effectively mitigate the impact of noise. Notably, this is achieved without necessitating prior knowledge of $\Sigma_{\boldsymbol{E}}$, marking a considerable advancement in noise-handling capabilities.

- The dICA approach enables the simultaneous estimation of both $\mathbf{W}^*$ and N^*. This synergistic process enhances the precision in determining N^* and the accuracy of $\mathbf{W}^*$ estimation, resulting in a more cohesive and accurate analytical outcome.

- Employing group regularization, dICA balances the minimization of $-\mathcal{J}_{\mathrm{Rob}}(\cdot)$ and the simplicity of the model. The methodology departs from traditional PCA by prioritizing component retention based on their non-Gaussianity rather than their magnitude. This approach significantly reduces the information loss typically associated with PCA, particularly in scenarios characterized by strong noise and weak signal components, thus producing more meaningful and robust results.

- The dICA approach introduces a more versatile noise model. Contrary to the common assumption in noisy ICA that $\Sigma_{\boldsymbol{E}} = \sigma^2\mathbf{I}$, dICA allows for $\Sigma_{\boldsymbol{E}}$ to be any diagonal positive definite matrix. Davies (2004) shows that $\Sigma_{\boldsymbol{E}}$ is generally unidentifiable without certain structural constraints. Hence, the decision to limit $\Sigma_{\boldsymbol{E}}$ to a diagonal matrix represents a strategic balance between maintaining model flexibility and ensuring identifiability.

This thesis is structured as follows. Chapter 2 introduces the dICA framework for solving the noisy ICA model. The model assumptions are listed in Section 2.1. Sections 2.2 to 2.5 detail the robust-to-noise non-Gaussianity measure, the estimation of $\Sigma_{\boldsymbol{E}}$, the rationale behind the first constraint in equation (1.5), and the functionality of the regularization term $\|\widetilde{\mathbf{W}}\|_F^2$. Collectively, these four subsections provide a comprehensive justification for the

optimization framework proposed in equation (1.5). Section 2.6 presents the dICA algorithm for solving equation (1.5). Chapter 3 discusses the application of the method framework introduced in Chapter 2 to practical scenarios with finitely many samples available, referred to as the sample model. Section 3.1 introduces this sample model. The prewhitening step is explained in Section 3.2, while Section 3.3 delves into the estimation of $\Sigma_{\boldsymbol{E}}$ using finite prewhitened data points. The dICA algorithm tailored for the sample model is presented in Section 3.4. Chapter 4 presents the numerical results. Section 4.1 evaluates the performance of dICA under its Gaussian noise assumption, and Section 4.2 explores its performance when this assumption is violated. The application of dICA to EEG datasets is showcased in Chapter 5. Finally, Chapter 6 discusses some limitations and potential future directions for dICA. The proofs of results from Chapter 2 can be found in Appendix A.

Chapter 2

dICA for the Population Model

This chapter provides a comprehensive introduction to the dICA framework, assuming that we possess knowledge of the probability density function (p.d.f.) of $\boldsymbol{X}$. It is important to note that throughout this chapter, we make the assumption that there is no presence of "finite sampling noise" and that accurate calculations of $\mathrm{E}\{f(\boldsymbol{X})\}$ can be obtained for any given $f(\cdot)$. In this chapter, we shall provide a thorough justification for the rationality of the optimization problem (1.5) and present an algorithm to address it.

$$\begin{aligned}\widehat{\widetilde{\mathbf{W}}}, \widehat{\Sigma}_{\boldsymbol{E}} \;=\; \arg \min_{\substack{\widetilde{\mathbf{W}}\in\mathbb{R}^{p\times p},\\ \mathrm{diag}(\Sigma)\geq \mathbf{0},\\ \Sigma \text{ is diagonal}}} & -\mathcal{J}_{\mathrm{Rob}}(\widetilde{\mathbf{W}}\boldsymbol{X}) + \|\widetilde{\mathbf{W}}\|_F^2, \\ \text{subject to} \quad & (\Sigma_{\boldsymbol{X}} - \Sigma)\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{W}}(\Sigma_{\boldsymbol{X}} - \Sigma) = (\Sigma_{\boldsymbol{X}} - \Sigma), \\ & \mathrm{vec}(\Sigma_{\boldsymbol{X}} - \Sigma) \in \mathcal{A}.\end{aligned}$$

More specifically, Sections 2.2 to 2.5 will offer a comprehensive rationale for the above optimization problem, while Section 2.6 will introduce the dICA algorithm for solving it.

2.1. Model Assumption and Ambiguities in dICA

The noisy ICA model (1.2) is shown again below:

$$\boldsymbol{X} = \mathbf{A}^*\boldsymbol{S}^* + \boldsymbol{E}.$$

To facilitate the subsequent discussion, the assumptions needed in this chapter are listed here.

Assumption 1. *Components $S_1^*, \ldots, S_{N^*}^*$ in $\boldsymbol{S}^*$ are independent and marginally non-Gaussian (at most one of them could be Gaussian), with $\mathrm{E}(\boldsymbol{S}^*) = \mathbf{0}$ and $\mathrm{cov}(\boldsymbol{S}^*) = \mathbf{I}_{N^*}$.*

Assumption 2. *$\boldsymbol{S}^*$ and $\boldsymbol{E}$ are independent.*

Assumption 3. *$\boldsymbol{E}$ follows a normal distribution with zero mean and covariance matrix $\Sigma_{\boldsymbol{E}}$, i.e. $\boldsymbol{E} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{E}})$.*

Assumption 4. *The noise covariance matrix $\Sigma_{\boldsymbol{E}}$ satisfies the following equation:*

$$\Sigma_{\boldsymbol{E}} = \mathrm{diag}(\sigma_{e_1}^2, \ldots, \sigma_{e_p}^2), \quad \sigma_{e_j}^2 \geq 0,\ j = 1, \ldots, p. \tag{2.1}$$

Assumption 5. $\mathrm{rank}(\mathbf{A}^*) = N^*$, *and* $p \geq N^*$.

Assumption 1, Assumption 2, and Assumption 5 are typical assumptions made in the traditional noise-free ICA model (Hyvärinen et al., 2001). The zero-mean assumption is made without loss of generality as $\mathrm{E}(\boldsymbol{X}) = \mathbf{A}^*\mathrm{E}(\boldsymbol{S}^*)$ and the identity covariance assumption includes any scaling ambiguity within $\mathbf{A}^*$. Note that Assumption 3 is common in noisy ICA literature and Assumption 4 presents a more generalized form than the typical assumptions regarding $\Sigma_{\boldsymbol{E}}$ where $\Sigma_{\boldsymbol{E}} = \sigma^2\mathbf{I}$ (Hyvärinen et al., 2001; Beckmann and Smith, 2004).

As an extension of the noise free ICA model (1.1), model (1.2) also inherits ambiguities of sign flips, permutation, and rescaling. For any permutation matrix $\mathbf{P}$ and any non-singular diagonal matrix $\mathbf{D}$, the equality $\mathbf{A}^*\boldsymbol{S}^* = (\mathbf{A}^*\mathbf{P}\mathbf{D})(\mathbf{D}^{-1}\mathbf{P}^{-1}\boldsymbol{S}^*)$ always holds. Consequently,

it is impossible to distinguish between $\mathbf{A}^*$(or $\boldsymbol{S}^*$) and $\mathbf{A}^*\mathbf{P}\mathbf{D}$(or $\mathbf{D}^{-1}\mathbf{P}^{-1}\boldsymbol{S}^*$) when only $\boldsymbol{X}$ is available. Assumption 1 constrains $\{S_k^*\}_{k=1}^{N^*}$ to have unit variances, thus eliminating the rescaling ambiguity. Nevertheless, model (1.2) still suffers from the sign flips and permutation ambiguities. Given the inherent unidentifiability of $\mathbf{P}$ and $\mathbf{D}$ in model (1.2), dICA only aims to identify $\mathbf{A}^*$ up to sign flips and order permutations. In other words, solutions that differ only by sign flips and order permutations are considered equivalent. In subsequent analysis, without loss of generality, this study will omit both $\mathbf{P}$ and $\mathbf{D}$, treating them as identity matrices.

2.2. The Choice of the Robust-to-noise Contrast Function

This paper adopts the Gaussian moments functions proposed by Hyvarinen (1999) as the choices of $\mathcal{J}_{\text{Rob}}(\cdot)$. The $\mathcal{J}_{\text{Rob}}(\cdot)$ is an extension of the $\mathcal{J}(\cdot)$ in equation (1.3). For a smoother transition towards the introduction of $\mathcal{J}_{\text{Rob}}(\cdot)$, a brief introduction of $\mathcal{J}(\cdot)$ is given below.

Choosing $\mathcal{J}(\cdot)$ as a measure of non-Gaussianity is arguably the most popular approach in noise free ICA (Hyvärinen and Oja, 2000; Hyvärinen et al., 2001). The intuition is rooted in the insights drawn from the Central Limit Theorem (CLT), which states that the linear combination of random variables tends to resemble a Gaussian distribution more closely than its individual components. Consequently, the maximization of non-Gaussianity in $\mathbf{W}_N\boldsymbol{X}$ is indicative of a thorough demixing. As a measure of non-Gaussianity, $\mathcal{J}(\cdot)$ takes the following form:

$$\mathcal{J}(\mathbf{W}_N\boldsymbol{X}) = \sum_{j=1}^{N} \left| \mathrm{E}\{G(\boldsymbol{w}_{N_j}^\top \boldsymbol{X})\} - \mathrm{E}\{G(Z)\} \right|, \tag{2.2}$$

where $G(\cdot)$ denotes a non-linear, non-quadratic contrast function, $\boldsymbol{w}_{N_j}^\top$ is the j-th row of $\mathbf{W}_N$, and Z represents a standard normal distribution. Popular choices for $G(\cdot)$ include (i)$G(x) = x^4$; (ii) $G(x) = \log\cosh(x)$; (iii)$G(x) = \exp(-x^2/2)$. One can see from (2.2) that $\mathcal{J}(\cdot)$ measures the discrepancy between $G(\boldsymbol{w}_{N_j}^\top \boldsymbol{X})$ and $G(Z)$. Therefore, $\mathcal{J}(\cdot)$ is conceptualized

as a measure of non-Gaussianity, where a larger discrepancy signifies a stronger departure from Gaussian characteristics in $\boldsymbol{w}_{N_j}^\top \boldsymbol{X}$.

While equation (2.2) aligns intuitively with the principles of non-Gaussianity, it needs a more rigorous theoretical foundation. A notable concern is that the optimizer of (2.2), subject to the orthogonal constraint in equation (1.3), is not necessarily the demixing matrix $\mathbf{A}^\dagger := (\boldsymbol{a}_1^\dagger, \cdots, \boldsymbol{a}_N^\dagger)^\top$. This limitation has been recognized and addressed by Hyvärinen et al. (2001) and Wei (2015). Their works demonstrate that $\boldsymbol{a}_j^\dagger$ is a local maximizer (or minimizer) of $\mathrm{E}\{G(\boldsymbol{w}_{N_j}^\top \boldsymbol{X})\}$ (subject to $\boldsymbol{w}_{N_j}^\top \Sigma_{\boldsymbol{X}} \boldsymbol{w}_{N_j} = 1$) when $\mathrm{E}\{G''(S_j) - G'(S_j)S_j\} < 0$ (or > 0). For the sake of brevity, we denote $\alpha_j := \mathrm{E}\{G''(S_j) - G'(S_j)S_j\}$. Using their results, $\mathbf{A}^\dagger$ qualifies as a local maximizer of equation (2.2) under the orthogonal constraint only if:

$$\mathrm{sign}(\alpha_j) = -\mathrm{sign}(\mathrm{E}\{G(\boldsymbol{a}_j^{\dagger\top} \boldsymbol{X})\} - \mathrm{E}\{G(Z)\}), 1 \le j \le N. \tag{2.3}$$

Encouragingly, equation (2.3) is often valid for a wide range of distributions of S_j. This fact underpins the empirical robustness of equation (2.2) in practical applications of ICA, confirming its utility despite the theoretical nuances. With equation (2.3), a more mathemically rigorous $\mathcal{J}(\cdot)$ should be

$$\mathcal{J}(\mathbf{W}_N \boldsymbol{X}) = \sum_{j=1}^{N} \mathrm{sign}(-\alpha_j) \mathrm{E}\{G(\boldsymbol{w}_{N_j}^\top \boldsymbol{X})\}. \tag{2.4}$$

Returning our focus to the noisy ICA model as outlined in equation (1.2), we need to address two issues in order to extend the $\mathcal{J}(\cdot)$ in equation (2.4) to the $\mathcal{J}_{\text{Rob}}(\cdot)$.

The first issue is to extend the definitions for $\alpha_j, N^* < j \le p$ because $\mathcal{J}_{\text{Rob}}(\cdot)$ is a function being optimized on the augmented matrices set—$\mathbb{R}^{p\times p}$. As aforementioned in Section 1.1, the optimal $\widetilde{\mathbf{W}}^* := (\widetilde{\boldsymbol{w}}_1^*, \cdots, \widetilde{\boldsymbol{w}}_p^*)^\top$ should be $(\mathbf{W}^{*\top}, \mathbf{0}^\top)^\top$. Given that $\widetilde{\boldsymbol{w}}_j^* = \mathbf{0}$ for the rows beyond N^*, we extend the definition of α_j accordingly. For $j > N^*$, $\alpha_j = G''(0) - G'(0) \cdot 0 = G''(0)$.

The second issue would be to eliminate the noise effect. Specifically, we need to under-

stand the relationship between $\mathrm{E}\{G(\widetilde{\boldsymbol{w}}_j^\top \boldsymbol{X})\}$ and $\mathrm{E}\{G(\widetilde{\boldsymbol{w}}_j^\top (\boldsymbol{X}-\boldsymbol{E}))\}$. While this relationship is complicated for a general $G(\cdot)$, Hyvarinen (1999) proposes the use of Gaussian moments functions, which simplifies the relationship considerably. Before delving deep, some notations are needed. Let $\varphi(x)$ denote the probability density function $(\mathrm{p.d.f.})$ of a standard Gaussian distribution $\mathcal{N}(0,1)$. For the $\mathrm{p.d.f.}$ of the Gaussian distribution $\mathcal{N}(0,c^2)$, we define $\varphi_c(x)$ as follows: $\varphi_c(x) = \frac{1}{c}\varphi(\frac{x}{c}) = \frac{1}{\sqrt{2\pi}c}\exp(-\frac{x^2}{2c^2})$ for $c>0$. Here $\varphi_c^{(0)}(x)$ is simply denoted as $\varphi_c(x)$, and for $k \geq 1$, we define $\varphi_c^{(k)}(x)$ as the k-th derivative of $\varphi_c(x)$ with respect to x. Additionally, we introduce $\varphi_c^{(-k)}(x)$ as the k-th integral function of $\varphi_c(x)$, which is obtained by $\varphi_c^{(-k)}(x) = \int_0^x \varphi_c^{(-k+1)}(t)\mathrm{d}t$, for $-\infty < x < \infty$ and $k \geq 1$. Inspired by Hyvarinen (1999), we present the following Lemma 1 and Lemma 2. These Lemmas provide the theoretical underpinning for employing Gaussian moments.

Lemma 1. *Let φ be the* $\mathrm{p.d.f.}$ *of $\mathcal{N}(0,1)$. Then*

$$-\varphi^{(-2)}(x) + x\varphi^{(-1)}(x) = -\varphi(x) + \varphi(0).$$

Lemma 2 (correction version of Theorem 1 in Hyvarinen (1999))**.** *Let V be any random variable independent of a Gaussian noise variable $Z \sim \mathcal{N}(\mu, \sigma^2)$. Then, for any constant $c^2 > \sigma^2$, we have*

$$\begin{aligned}
\mathrm{E}\{\varphi_c(V+\mu)\} &= \mathrm{E}\{\varphi_d(V+Z)\},\\
\mathrm{E}\{\varphi_c^{(-1)}(V+\mu)\} &= \mathrm{E}\{\varphi_d^{(-1)}(V+Z)\},\\
\mathrm{E}\{\varphi_c^{(-2)}(V+\mu)\} &= \mathrm{E}\{\varphi_d^{(-2)}(V+Z)\} + \varphi(0)(d-c),
\end{aligned}$$

where $d = \sqrt{c^2-\sigma^2}$.

Lemma 2 reveals a straightforward approach to eliminate noise bias: simply by adjusting the subscript of the Gaussian moments functions. Motivated from this, choosing $G(\cdot) = \varphi_c^{(-2)}(\cdot)$,

the $\mathcal{J}_{\text{Rob}}(\cdot)$ can be written as following:

$$\begin{aligned}\mathcal{J}_{\text{Rob}}(\widetilde{\mathbf{W}}\boldsymbol{X}) &= \sum_{j=1}^{p} \text{sign}(-\alpha_j)(\text{E}\{\varphi^{(-2)}_{d(\widetilde{\boldsymbol{w}}_j)}(\widetilde{\boldsymbol{w}}_j^\top \boldsymbol{X})\} + \varphi(0)(d(\widetilde{\boldsymbol{w}}_j) - c)) \\ &= \sum_{j=1}^{p} \text{sign}(-\alpha_j)\text{E}\{G(\widetilde{\boldsymbol{w}}_j(\boldsymbol{X} - \boldsymbol{E}))\} \\ &= \mathcal{J}(\widetilde{\mathbf{W}}(\boldsymbol{X} - \boldsymbol{E})), \end{aligned} \tag{2.5}$$

where $\widetilde{\boldsymbol{w}}_j^\top$ is the j-th row of $\widetilde{\mathbf{W}}$ and $d(\widetilde{\boldsymbol{w}}_j) = \sqrt{c^2 - \widetilde{\boldsymbol{w}}_j^\top \Sigma_{\boldsymbol{E}} \widetilde{\boldsymbol{w}}_j}$, with $c^2 > \widetilde{\boldsymbol{w}}_j^\top \Sigma_{\boldsymbol{E}} \widetilde{\boldsymbol{w}}_j$. And

$$\alpha_j = \begin{cases} \text{E}\{\varphi_c(S_j^*) - \varphi_c^{(-1)}(S_j^*)S_j^*\}, & \text{if } 1 \le j \le N^*, \\ \varphi_c(0), & \text{if } N^* < j \le p. \end{cases} \tag{2.6}$$

Inspired from equation (2.5), the following Assumption 6 is crucial to the validity of $\mathcal{J}_{\text{Rob}}(\cdot)$.

Assumption 6. *For any* $1 \le j \le N^*$, $\text{E}\{\varphi_c(S_j^*) - \varphi_c^{(-1)}(S_j^*)S_j^*\} \neq 0$.

The values of $d(\widetilde{\boldsymbol{w}}_j)$ are contingent upon $\Sigma_{\boldsymbol{E}}$. The methodology for estimating $\Sigma_{\boldsymbol{E}}$ will be elaborated upon in the forthcoming Section 2.3.

Remark 1. *While the expression of non-linear contrast functions in equation* (2.2) *is not theoretically guaranteed, it is widely favored in practical applications due to its ease of implementation and alignment with the concept of non-Gaussianity. On the contrary, the formulation of* $\mathcal{J}_{Rob}(\cdot)$ *in* (2.5) *presents implementational challenges, mainly because the* α_j *values are contingent upon the latent sources* S_j^**. Hyvarinen (1999); Hyvärinen et al. (2001) have introduced fixed-point algorithms that converge to* $\mathbf{W}^*$ *without the need for explicit optimization of the non-linear contrast function. This paper adopts a similar approach, developing a fixed-point algorithm that optimizes the* $\mathcal{J}_{Rob}(\cdot)$ *function without a priori knowledge of the* α_j *values.*

Remark 2. *The selection of the constant* c *in* $d(\widetilde{\boldsymbol{w}}_j)$*, as discussed in Lemma 2, is important and*

must be made wisely . According to Lemma 2, the sole constraint on c is expressed as:

$$c^2 > \widetilde{\boldsymbol{w}}_j^\top \Sigma_{\boldsymbol{E}} \widetilde{\boldsymbol{w}}_j, \quad \textit{for all } j \in 1, \ldots, p. \tag{2.7}$$

Furthermore, taking into consideration the constraint presented in (1.5)*, we can observe that* $\widetilde{\mathbf{W}}(\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{E}})\widetilde{\mathbf{W}}^\top$ *forms an idempotent matrix, leading to the following relationship:*

$$\widetilde{\boldsymbol{w}}_j^\top (\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{E}}) \widetilde{\boldsymbol{w}}_j \leq 1, \quad \textit{for any } j \in \{1, \ldots, p\}.$$

Moreover, it is generally the case that the magnitude of the noise is smaller than that of the latent components $\boldsymbol{S}$. Consequently, we can infer that $\Sigma_{\boldsymbol{E}} \preceq \Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{E}}$, and as a result, it is typically true that:

$$\widetilde{\boldsymbol{w}}_j^\top \Sigma_{\boldsymbol{E}} \widetilde{\boldsymbol{w}}_j \leq \widetilde{\boldsymbol{w}}_j^\top (\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{E}}) \widetilde{\boldsymbol{w}}_j \leq 1, \quad \textit{for any } j \in \{1, \ldots, p\}.$$

In practice, choosing $c > 2$ is usually sufficient. In our numerical experiments, we set c to be 10 *to ensure the validity of* (2.7)*.*

2.3. Estimate the Noise Covariance Matrix

Before we explain the details of the estimation of $\Sigma_{\boldsymbol{E}}$, we firstly present two fundamental Lemmas that establish the identifiability of $\Sigma_{\boldsymbol{E}}$.

Lemma 3 (Identifiability of $\mathcal{A}$)**.** *Assume that Assumptions 1, 2, 3, and 5 hold. Suppose that the original sources $S_1^*, \ldots, S_{N^*}^*$ all exhibit non-zero kurtosis. Then the linear subspace $\mathcal{A}$ is identifiable.*

Lemma 4 (Identifiability of $\Sigma_{\boldsymbol{E}}$)**.** *Assume that Assumptions 1, 2, 3, 4, and 5 hold. Suppose that the sources $S_1^*, \ldots, S_{N^*}^*$ all have non-zero kurtosis. Let $\mathcal{D}$ denote a linear subspace consisting of $p \times p$ diagonal matrices, and $\mathrm{vec}(\mathcal{D}) \coloneqq \{\mathrm{vec}(\mathbf{D}) : \mathbf{D} \in \mathcal{D}\}$. Then $\Sigma_{\boldsymbol{E}}$ is identifiable if and only if*

$$\mathrm{vec}(\mathcal{D}) \cap \mathcal{A} = \{\mathbf{0}\}, \tag{2.8}$$

where $\mathbf{0}$ *in* (2.8) *refers to the zero vector in* $\mathbb{R}^{p^2}$.

With the identifiability of $\Sigma_{\boldsymbol{E}}$, now we start the estimation of it. The second constraint in equation (1.5) plays an important role in this procedure. Let $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_{N^*}$ represent the orthonormal bases of $\mathcal{A}$. Additionally, consider $\{\boldsymbol{F}_j\}_{j=1}^{p^2-N^*}$ as the orthonormal bases of the orthogonal complement space $\mathcal{A}^{\perp}$. Consequently, the constraint $\mathrm{vec}(\Sigma_{\boldsymbol{X}} - \Sigma) \in \mathcal{A}$ can be reformulated as:

$$\sum_{j=1}^{p^2-N^*} \langle \boldsymbol{F}_j, \mathrm{vec}(\Sigma_{\boldsymbol{X}} - \Sigma) \rangle^2 = 0, \tag{2.9}$$

where $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ denotes the inner product between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. Under the assumptions of Lemma 3 and Lemma 4, $\Sigma_{\boldsymbol{E}}$ emerges as the only diagonal matrix fulfilling equation (2.9). The remaining challenge is the estimation of $\{\boldsymbol{F}_j\}_{j=1}^{p^2-N^*}$.

The estimation process starts with determining $\{\boldsymbol{H}_j\}_{j=1}^{N^*}$, utilizing the Generalized Covariance Matrix (GCM).[1] Subsequently, $\{\boldsymbol{F}_j\}_{j=1}^{p^2-N^*}$ can be derived via Singular Value Decomposition (SVD) once $\{\boldsymbol{H}_j\}_{j=1}^{N^*}$ are available.

Recalling that the generalized covariance matrix is the second derivative of the moment generating function evaluated at non-zero point (Yeredor, 2000; Podosinnikova et al., 2019). We define the logarithm of the moment generating function of $\boldsymbol{X} \in \mathbb{R}^p$ as $\phi_{\boldsymbol{X}}(\boldsymbol{\tau}) \coloneqq \log\{\mathrm{E}(e^{\boldsymbol{\tau}^\top \boldsymbol{X}})\}$, where $\boldsymbol{\tau} \in \mathbb{R}^p$ is called processing point. The generalized covariance matrix $\mathcal{C}_{\boldsymbol{X}}(\boldsymbol{\tau})$ is defined as the second-order derivative of $\phi_{\boldsymbol{X}}(\boldsymbol{\tau})$ with respect to $\boldsymbol{\tau}$:

$$\mathcal{C}_{\boldsymbol{X}}(\boldsymbol{\tau}) \coloneqq \nabla^2 \phi_{\boldsymbol{X}}(\boldsymbol{\tau}),$$

and it can be expressed as (Podosinnikova et al., 2019):

$$\mathcal{C}_{\boldsymbol{X}}(\boldsymbol{\tau}) = \sum_{k=1}^{N^*} \omega_k(\boldsymbol{\tau}) \boldsymbol{a}_k^* \boldsymbol{a}_k^{*\top} + \Sigma_{\boldsymbol{E}}. \tag{2.10}$$

[1] An alternative option would be to use the higher-order statistics (De Lathauwer et al., 2007; Bonhomme and Robin, 2009; Podosinnikova et al., 2019). The higher-order methods are more accurate when p is small but have larger computational complexity.

Here $\omega_k(\boldsymbol{\tau}) = \frac{\partial^2}{\partial t^2} \log\{\mathrm{E}(e^{S_k^* t})\}\Big|_{t=\boldsymbol{a}_k^{*T}\boldsymbol{\tau}}$ represents the generalized variance of the k-th source S_k^*. The noise covariance matrix $\Sigma_{\boldsymbol{E}}$ associated with Gaussian noise is the unchanged part in $\mathcal{C}_{\boldsymbol{X}}(\boldsymbol{\tau})$ regardless the choice of $\boldsymbol{\tau}$. Consequently, by taking the difference between two generalized covariance matrices at different $\boldsymbol{\tau}$s, $\Sigma_{\boldsymbol{E}}$ can be eliminated, facilitating the extraction of the subspace $\mathcal{A}$. To ensure the identifiability of $\mathcal{A}$, we need to find $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_q$ with $q > N^*$ such that $\mathrm{rank}(\boldsymbol{\omega}(\boldsymbol{\tau}_1) - \boldsymbol{\omega}(\mathbf{0}), \ldots, \boldsymbol{\omega}(\boldsymbol{\tau}_q) - \boldsymbol{\omega}(\mathbf{0})) = N^*$, where $\boldsymbol{\omega}(\boldsymbol{\tau}_\ell) = (\omega_1(\boldsymbol{\tau}_\ell), \ldots, \omega_{N^*}(\boldsymbol{\tau}_\ell))^\top$. In practice, we choose q to be multiple times larger than p to ensure that $\{\mathrm{vec}(\mathcal{C}_{\boldsymbol{X}}(\boldsymbol{\tau}_\ell) - \mathcal{C}_{\boldsymbol{X}}(\mathbf{0}))\}_{\ell=1}^q$ can effectively capture the entire subspace $\mathcal{A}$. More precisely, we define:

$$\mathbf{GC} = (\mathrm{vec}(\mathcal{C}_{\boldsymbol{X}}(\boldsymbol{\tau}_1) - \mathcal{C}_{\boldsymbol{X}}(\mathbf{0})), \ldots, \mathrm{vec}(\mathcal{C}_{\boldsymbol{X}}(\boldsymbol{\tau}_q) - \mathcal{C}_{\boldsymbol{X}}(\mathbf{0}))) \in \mathbb{R}^{p^2 \times q}. \tag{2.11}$$

The $\widehat{N}$ left-singular vectors corresponding to the $\widehat{N}$ non-zero singular values of $\mathbf{GC}$, represented as $\widehat{\boldsymbol{H}}_1, \ldots, \widehat{\boldsymbol{H}}_{\widehat{N}}$, constitute an estimated basis for the subspace $\mathcal{A}$. Once $\{\widehat{\boldsymbol{H}}_j\}_{j=1}^{\widehat{N}}$ is determined, the complementary set of vectors $\{\widehat{\boldsymbol{F}}_j\}_{j=1}^{p^2-\widehat{N}}$ can be derived through SVD. If $\mathrm{rank}(\boldsymbol{\omega}(\boldsymbol{\tau}_1) - \boldsymbol{\omega}(\mathbf{0}), \ldots, \boldsymbol{\omega}(\boldsymbol{\tau}_q) - \boldsymbol{\omega}(\mathbf{0})) = N^*$ and equation (2.8) both hold, the $\Sigma_{\boldsymbol{E}}$ will be the only diagonal matrix fulfilling $\sum_{j=1}^{p^2-\widehat{N}} \langle \widehat{\boldsymbol{F}}_j, \mathrm{vec}(\Sigma_{\boldsymbol{X}} - \Sigma)\rangle^2 = 0$ and $\widehat{N}$ should equal N^*. For the purposes of subsequent analysis, these two assumptions are formally enumerated as Assumption 7 and 8.

Assumption 7. *The q randomly generated processing points satisfy* $\mathrm{rank}(\boldsymbol{\omega}(\boldsymbol{\tau}_1) - \boldsymbol{\omega}(\mathbf{0}), \ldots, \boldsymbol{\omega}(\boldsymbol{\tau}_q) - \boldsymbol{\omega}(\mathbf{0})) = N^*$.

Assumption 8. $\mathrm{vec}(\mathcal{D}) \cap \mathcal{A} = \{\mathbf{0}\}$.

It is worth noting that following equations should hold if Assumptions 1-8 hold.

$$\begin{aligned} \widehat{N} &= N^*, \\ \mathcal{A} &= \mathrm{span}\{\widehat{\boldsymbol{H}}_1, \ldots, \widehat{\boldsymbol{H}}_{\widehat{N}}\} = \mathrm{span}\{\boldsymbol{H}_1, \ldots, \boldsymbol{H}_{N^*}\}, \\ \mathcal{A}^\perp &= \mathrm{span}\{\widehat{\boldsymbol{F}}_1, \ldots, \widehat{\boldsymbol{F}}_{p^2-\widehat{N}}\} = \mathrm{span}\{\boldsymbol{F}_1, \ldots, \boldsymbol{F}_{p^2-N^*}\}. \end{aligned} \tag{2.12}$$

While theoretically it is hard to show how large q should be or how to generate $\boldsymbol{\tau}$s in order for Assumption 7 to hold, this Assumption is empirically valid in our simulations when we generate $\boldsymbol{\tau}$s randomly and let q be 4 to 5 times larger than p.

2.4. On the Hidden Orthogonal Constraint

This section aims to demonstrate that any $\widetilde{\mathbf{W}}$ satisfying constraints in equation (1.5) can be transformed to an element in the $\widetilde{\mathcal{W}}^*$ by an orthogonal transformation. To achieve this, we begin by examining the structure of $\widetilde{\mathbf{W}}$ that satisfies the constraints in (1.5). Reformulate the two constraints in (1.5) using equation (2.9) and equation (2.12), we have

$$(\Sigma_{\boldsymbol{X}} - \Sigma)\widetilde{\mathbf{W}}^\top \widetilde{\mathbf{W}}(\Sigma_{\boldsymbol{X}} - \Sigma) = (\Sigma_{\boldsymbol{X}} - \Sigma), \tag{2.13}$$

$$\sum_{j=1}^{p^2-\widehat{N}} \langle \widehat{\boldsymbol{F}}_j, \mathrm{vec}(\Sigma_{\boldsymbol{X}} - \Sigma)\rangle^2 = 0. \tag{2.14}$$

By employing Lemma 4 and equation (2.12), it can be shown that $\Sigma_{\boldsymbol{E}}$ is the unique diagonal matrix that satisfies equation (2.14), provided that Assumptions 1 through 8 are valid. In light of this finding, subsequent analyses in this section will proceed by substituting $\Sigma_{\boldsymbol{E}}$ for Σ in equation (2.13), thereby omitting (2.14) from further consideration. The focus will now shift to an in-depth examination of equation (2.13). The underlying intuition of (2.13) is encapsulated in the following equation:

$$\begin{aligned}(\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{E}})\widetilde{\mathbf{W}}^{*\top}\widetilde{\mathbf{W}}^*(\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{E}}) &= \mathbf{A}^*\mathbf{A}^{*\top}\widetilde{\mathbf{W}}^{*\top}\widetilde{\mathbf{W}}^*\mathbf{A}^*\mathbf{A}^{*\top} \\ &= \mathbf{A}^*(\mathbf{I}_{N^*}, \mathbf{0}) \begin{pmatrix} \mathbf{I}_{N^*} \\ \mathbf{0} \end{pmatrix} \mathbf{A}^{*\top} \\ &= \mathbf{A}^*\mathbf{A}^{*\top} = \Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{E}}.\end{aligned}$$

Following Lemma 5 establishes the structure of $\widetilde{\mathbf{W}}$ satisfying constraint (2.13).

Lemma 5. *Assume Assumptions 1-8 hold. Let* $\mathbf{A}^* = \mathbf{T}_{N^*}\mathbf{D}_A\mathbf{V}^\top$ *be the economic form of the singular value decomposition of the mixing matrix* $\mathbf{A}^*$*, where* $\mathbf{T}_{N^*} \in \mathbb{R}^{p\times N^*}$ *with* $\mathbf{T}_{N^*}^\top\mathbf{T}_{N^*} = \mathbf{I}_{N^*}$*.* $\mathbf{D}_A \in \mathbb{R}^{N^*\times N^*}$ *is a diagonal matrix, and* $\mathbf{V} \in \mathbb{R}^{N^*\times N^*}$ *with* $\mathbf{V}^\top\mathbf{V} = \mathbf{V}\mathbf{V}^\top = \mathbf{I}_{N^*}$*. Then, the* $\widetilde{\mathbf{W}}$ *that satisfies the constraint* (2.13) *will have the following expression:*

$$\widetilde{\mathbf{W}} = \mathbf{U}_{N^*}\mathbf{D}_A^{-1}\mathbf{T}_{N^*}^\top + \mathbf{\Psi}\mathbf{T}_{p-N^*}^\top \in \mathbb{R}^{p\times p}, \tag{2.15}$$

where $\mathbf{T}_{p-N^*}$ *is a* $p \times (p - N^*)$ *semi-orthogonal matrix satisfying* $\mathbf{T}_{p-N^*}^\top\mathbf{T}_{p-N^*} = \mathbf{I}_{p-N^*}$*,* $\mathbf{T}_{p-N^*}^\top\mathbf{T}_{N^*} = \mathbf{0}$*,* $\mathbf{U}_{N^*}$ *is a* $p \times N^*$ *semi-orthogonal matrix with* $\mathbf{U}_{N^*}^\top\mathbf{U}_{N^*} = \mathbf{I}_{N^*}$*, and* $\mathbf{\Psi}$ *is any* $p \times (p - N^*)$ *matrix.*

Employing the results from Lemma 5, we can deduce that any matrix $\widetilde{\mathbf{W}}$ satisfying equation (2.15) will lead to the relation $\widetilde{\mathbf{W}}\mathbf{A}^* = \mathbf{U}_{N^*}\mathbf{V}^\top$. Then we obtain the following Lemma 6.

Lemma 6. *Assume Assumptions 1-8 hold. For any* $\widetilde{\mathbf{W}}$ *adhering to* (2.13) (*or equivalently* (2.15))*, an orthonormal matrix* $\widehat{\mathbf{O}}$ *exists such that* $\widehat{\mathbf{O}}\widetilde{\mathbf{W}} \in \widetilde{\mathcal{W}}^*$*.*

Lemma 6 proves that $\widetilde{\mathbf{W}}$ satisfying equation (2.15) differs from $\mathcal{W}^*$ by an orthonormal matrix $\widehat{\mathbf{O}}$. The following Lemma 7 demonstrates that for any such $\widetilde{\mathbf{W}}$, the matrix $\widehat{\mathbf{O}}$ emerges as the minimizer of the function $-\mathcal{J}_{\text{Rob}}(\mathbf{O}\widetilde{\mathbf{W}}\boldsymbol{X})$, subject to the condition $\mathbf{O}\mathbf{O}^\top = \mathbf{I}$.

Lemma 7. *Assume Assumptions 1-8 hold. For* $\widetilde{\mathbf{W}}$ *satisfying the constraints in* (2.13)*, denote* $\widetilde{\mathbf{W}}\mathbf{A}^* = \mathbf{U}_{N^*}\mathbf{V}^\top$ *by* $\mathbf{R}$*. Then* $\mathbf{R}$ *is a semi-orthogonal matrix. Furthermore,* $-\mathcal{J}_{Rob}(\widetilde{\mathbf{W}}\boldsymbol{X})$ *achieves its local minimum when* $\mathbf{R} = \mathbf{R}^* := (\mathbf{I}_{N^*}, \mathbf{0})^\top$*, which implies* $\mathbf{U}_{N^*} = (\mathbf{V}^\top, \mathbf{0})^\top$*.*

2.5. Justification of the dICA Optimization Framework

To this end, it is almost clear which roles the constraints and objective function play in (1.5). Constraint $(\Sigma_{\boldsymbol{X}} - \Sigma) \in \mathcal{A}$ helps to identify the $\Sigma_{\boldsymbol{E}}$. With $\Sigma_{\boldsymbol{E}}$, equation (2.13) serves to restrict the structure of $\widetilde{\mathbf{W}}$ as described in equation (2.15). Lemma 6 reveals that any $\widetilde{\mathbf{W}}$ satisfying equation (2.15) differs from $\mathcal{W}^*$ only by an orthonormal matrix $\widehat{\mathbf{O}}$, and Lemma

7 demonstrates that $\widehat{\mathbf{O}}$ can be obtained by optimizing $\mathcal{J}_{\text{Rob}}(\mathbf{O}\widetilde{\mathbf{W}}\boldsymbol{X})$ under the constraint $\mathbf{O}\mathbf{O}^\top = \mathbf{I}$. Noticing the following equation:

$$\widehat{\mathbf{O}}\widetilde{\mathbf{W}} = (\mathbf{V}^\top, \mathbf{0})^\top \mathbf{D}_A^{-1}\mathbf{T}_{N^*}^\top + \mathbf{\Psi}\mathbf{T}_{p-N^*}^\top = \widetilde{\mathbf{W}}^* + \mathbf{\Psi}\mathbf{T}_{p-N^*}^\top,$$

one can see that the only remaining difference between $\widehat{\mathbf{O}}\widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{W}}^*$ is the term $\mathbf{\Psi}\mathbf{T}_{p-N^*}^\top$. Lemma 8 shows the role $\|\widetilde{\mathbf{W}}\|_F^2$ plays—to eliminate $\mathbf{\Psi}$.

Lemma 8. *Assume Assumptions 1-8 hold, and let equation* (2.15) *hold for* $\widetilde{\mathbf{W}}$. *Then* $\|\widetilde{\mathbf{W}}\|_F^2$ *or equivalently* $\sum_{j=1}^p \|\widetilde{\boldsymbol{w}}_j\|_2^2$ *reaches its minimum if and only if* $\mathbf{\Psi} = \mathbf{0}$.

Combining Lemma 4, 5, 6, 7, 8, we have the following Theorem 1.

Theorem 1. *Assume Assumptions 1-8 hold. The local optimizer* $\widehat{\widetilde{\mathbf{W}}}, \widehat{\Sigma}_{\boldsymbol{E}}$ *of the optimization problem* (1.5) *satisfy* $\widehat{\widetilde{\mathbf{W}}} = \widetilde{\mathbf{W}}^*$, $\widehat{\Sigma}_{\boldsymbol{E}} = \Sigma_{\boldsymbol{E}}$.

2.6. The dICA Algorithm

Inspired by Lemma 6 and 7, optimizing $-\mathcal{J}_{\text{Rob}}(\widetilde{\mathbf{W}}\boldsymbol{X})$ under the constraints (2.13) and (2.14) is essentially equivalent to finding $\widehat{\mathbf{O}}$ such that $\widehat{\mathbf{O}}\widetilde{\mathbf{W}} \in \widetilde{\mathcal{W}}^*$. Furthermore, it is notable that the constraints (2.13) and (2.14), along with $\|\widetilde{\mathbf{W}}\|_F^2$, remain invariant under orthogonal transformations of $\widetilde{\mathbf{W}}$. Motivated from these facts, this study find the local minimizer of (1.5) by optimizing $-\mathcal{J}_{\text{Rob}}(\widetilde{\mathbf{W}}\boldsymbol{X})$ and $\|\widetilde{\mathbf{W}}\|_F^2$ separately. The first step is to find the $\widehat{\Sigma}_{\boldsymbol{E}}$ by utilizing equation (2.14). The next step is to minimize $\|\widetilde{\mathbf{W}}\|_F^2$ under the constraints (2.13) and to obtain an intermediate $\widetilde{\mathbf{W}}$, denoted as $\widetilde{\mathbf{W}}_{\text{interme}}$. The final step is to identify $\widehat{\mathbf{O}}$ that minimizes $-\mathcal{J}_{\text{Rob}}(\mathbf{O}\widetilde{\mathbf{W}}_{\text{interme}}\boldsymbol{X})$ under the constraint $\mathbf{O}\mathbf{O}^\top = \mathbf{I}$.

$$\widehat{\Sigma}_{\boldsymbol{E}} \quad = \quad \arg \min_{\substack{\Sigma \text{ is diagonal,} \\ \text{diag}(\Sigma) \geq 0}} \sum_{j=1}^{p^2-\widehat{N}} \langle \widehat{\boldsymbol{F}}_j, \text{vec}(\Sigma_{\boldsymbol{X}} - \Sigma) \rangle^2. \tag{2.16}$$

Recall that the $\widehat{\boldsymbol{F}}_j$s in equation (2.16) represent the estimated orthonormal bases of $\mathcal{A}^{\perp}$.

$$\begin{aligned} \widetilde{\mathbf{W}}_{\text{interme}} &= \arg\min_{\widetilde{\mathbf{W}}\in\mathbb{R}^{p\times p}} \|\widetilde{\mathbf{W}}\|_F^2, \\ &\quad \text{subject to} \quad (\Sigma_{\boldsymbol{X}} - \widehat{\Sigma}_{\boldsymbol{E}})\widetilde{\mathbf{W}}^{\top}\widetilde{\mathbf{W}}(\Sigma_{\boldsymbol{X}} - \widehat{\Sigma}_{\boldsymbol{E}}) = (\Sigma_{\boldsymbol{X}} - \widehat{\Sigma}_{\boldsymbol{E}}). \end{aligned} \tag{2.17}$$

$$\begin{aligned} \widehat{\mathbf{O}} &= \arg\min_{\mathbf{O}\in\mathbb{R}^{p\times p}} -\mathcal{J}_{\text{Rob}}(\mathbf{O}\widetilde{\mathbf{W}}_{\text{interme}}\boldsymbol{X}), \\ &\quad \text{subject to } \mathbf{O}\mathbf{O}^{\top} = \mathbf{I}_p. \end{aligned} \tag{2.18}$$

Then $\widehat{\widetilde{\mathbf{W}}} = \widehat{\mathbf{O}}\widetilde{\mathbf{W}}_{\text{interme}}$. Solving equations (2.16) and (2.17) are routine processes. The estimation of $\widehat{\Sigma}_{\boldsymbol{E}}$ can be achieved by solving a nonnegative linear least-squares problem (Lawson and Hanson, 1976). Once $\widehat{\Sigma}_{\boldsymbol{E}}$ is obtained, the intermediate matrix $\widetilde{\mathbf{W}}_{\text{interme}}$ can be readily determined through SVD. Addressing equation (2.18) is more challenging. Inspired by the work of Hyvarinen (1999), this study introduces a novel fixed-point algorithm, which is detailed in Algorithm 1. The following Lemma 9 and Theorem 2 establish the local

Algorithm 1 *Finding* $\widehat{\mathbf{O}}$ *in* (2.18).

Input: $\boldsymbol{X}$, an initial orthonormal matrix $\mathbf{O}_0$ (default is $\mathbf{I}_p$), a weight β, $\widetilde{\mathbf{W}}_{\text{interme}}$, and $\widehat{\Sigma}_{\boldsymbol{E}}$.

1: Compute $\overline{\boldsymbol{X}} = \widetilde{\mathbf{W}}_{\text{interme}}\boldsymbol{X}$ and $\overline{\Sigma}_{\boldsymbol{E}} = \widetilde{\mathbf{W}}_{\text{interme}}\widehat{\Sigma}_{\boldsymbol{E}}\widetilde{\mathbf{W}}_{\text{interme}}^{\top}$.
2: **while** not converged **do**
3: Get $\{\boldsymbol{o}_j\}_{j=1}^{p}$ from $\mathbf{O} = (\boldsymbol{o}_1, ..., \boldsymbol{o}_p)^{\top}$.
4: **for** $j = 1, \ldots, p$ **do**
5:

$$\begin{aligned} \overline{\boldsymbol{o}}_j = \mathrm{E}\ &\{\overline{\boldsymbol{X}}\varphi_{d(\boldsymbol{o}_j)}^{(-1)}(\boldsymbol{o}_j^{\top}\overline{\boldsymbol{X}})\} \\ &- \{\beta\mathrm{cov}(\overline{\boldsymbol{X}}) + (1-\beta)(\mathbf{I} + \overline{\Sigma}_{\mathrm{E}})\}\boldsymbol{o}_j\mathrm{E}\{\varphi_{d(\boldsymbol{o}_j)}(\boldsymbol{o}_j^{\top}\overline{\boldsymbol{X}})\}. \end{aligned} \tag{2.19}$$

6: **end for**
7: Get $\overline{\mathbf{O}} = (\overline{\boldsymbol{o}}_1, \ldots, \overline{\boldsymbol{o}}_p)^{\top}$ and $\mathbf{O} = (\overline{\mathbf{O}}\,\overline{\mathbf{O}}^{\top})^{(-1/2)}\overline{\mathbf{O}}$.
8: **end while**

Output: Orthonormal matrix $\widehat{\mathbf{O}}$.

convergence of Algorithm 1.

Lemma 9 ($\widehat{\mathbf{O}}$ is the fixed point of Algorithm 1)**.** *Assume Assumptions 1-8 hold. Let* $\widetilde{\mathbf{W}}_{interme}$ *be*

defined in (2.17). *Then* $\widehat{\mathbf{O}}$, *satisfying* $\widehat{\mathbf{O}}\widetilde{\mathbf{W}}_{interme} \in \widetilde{\mathcal{W}}^*$, *is the fixed point of Algorithm* 1.

Theorem 2 (Local convergence of Algorithm 1)**.** *Assume Assumptions 1-8 hold. For* $\widetilde{\mathbf{W}}_{interme}$ *defined in* (2.17) *and an initial orthonormal matrix* $\mathbf{O}_0$, *then* $\mathbf{O}\widetilde{\mathbf{W}}_{interme}\mathbf{A}^*$ *converges locally to* $(\mathbf{I}_{N^*}, \mathbf{0})^\top$ *under iterations in Algorithm* 1 *with a linear convergence rate as long as* $(1-\beta) < \min_{1\le j\le N^*} |\alpha_j|/\{2\varphi_c(0)p\}$. *Here* α_j *is defined in equation* (2.6).

Remark 3. *From the proof of Theorem* 2, *it can be observed that setting* $\beta = 1$ *yields a second-order convergence rate when* N^* *is known in advance. However, in the case where* N^* *is unknown, it is necessary for* $1-\beta$ *to be non-zero in order to ensure the invertibility of* $\overline{\mathbf{OO}}^\top$. *For the purposes of our numerical experiments, we have selected* β *to be 0.999.*

Upon obtaining $\widehat{\widetilde{\mathbf{W}}}$, we can proceed to estimate $\mathbf{A}^*$ and $\boldsymbol{S}^*$. Denote $\widehat{\mathbf{W}}$ as the matrix consists of all the non-zero rows in $\widehat{\widetilde{\mathbf{W}}}$. Since $\mathbf{A}^* = \mathbf{A}^*(\mathbf{A}^{*\top}\mathbf{W}^{*\top}) = (\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{E}})\mathbf{W}^{*\top}$. This paper estimates $\mathbf{A}^*$ through the following:

$$\widehat{\mathbf{A}} = (\Sigma_{\boldsymbol{X}} - \widehat{\Sigma}_{\boldsymbol{E}})\widehat{\mathbf{W}}^\top.$$

Estimating $\boldsymbol{S}^*$ in a noisy ICA setting generally poses a complexity due to the intricate non-linear relationship between $\boldsymbol{S}^*$ and $\boldsymbol{X}$ (Hyvärinen, 1998). Moreover, the noise term $\boldsymbol{E}$ remains unidentifiable under the general noisy ICA model, rendering the precise estimation of $\boldsymbol{S}^*$ unattainable in the presence of $\boldsymbol{E}$ (Davies, 2004). In this scenario, we employ the best linear estimator proposed by Koldovsky and Tichavsky (2006) and Koldovskỳ and Tichavskỳ (2007) to approximate the non-linear relationship between $\boldsymbol{X}$ and $\boldsymbol{S}^*$. We derive an estimator for $\boldsymbol{S}^*$ through a linear transformation of $\boldsymbol{X}$. In the presence of $\boldsymbol{E}$, the linear transformation minimizing the mean squared error (MSE) $\mathrm{E}(\|\boldsymbol{S}^* - \widehat{\boldsymbol{S}}\|_2^2)$ between $\boldsymbol{S}^*$ and $\widehat{\boldsymbol{S}}$ is as given by

$$\widehat{\boldsymbol{S}} = \widehat{\mathbf{A}}^\top \Sigma_{\boldsymbol{X}}^{-1} \boldsymbol{X}. \tag{2.20}$$

In this chapter, we explain how dICA estimates $\mathbf{A}^*$ and $\boldsymbol{S}^*$ in the population model through the optimization of equation (1.5). Firstly, we detail the derivation of equations (2.5), (2.13), and (2.14), and unveil the roles they fulfill in the context of equation (1.5). Secondly, we introduce an algorithm designed to solve equation (1.5), and provide a comprehensive proof of the local convergence results for this algorithm. In the next chapter, we will show how to extend the dICA method to the sample model with only finitely many data points.

Chapter 3

dICA for the Finite Sample Model

In finite sample model, the algorithm begins with a prewhitening step to rescale and decorrelate the data matrix. This step can help stabilize the subsequent computation. Additionally, due to the presence of "finite sampling noise," the equality constraints in equations (2.13) and (2.14) may only hold approximately, and $\|\cdot\|_F^2$ cannot guarantee sparsity in the demixing matrix. In this chapter, we will utilize the Group LASSO method (Yuan and Lin, 2006) to promote sparsity in the demixing matrix. Moreover, we will transform the hard constraints in equations (2.13) and (2.14) into soft constraints using the penalty method. Section 3.4 provides a detailed explanation of these approaches.

3.1. Sample Model

In practice, we have n independent and identically distributed (i.i.d.) observations $\boldsymbol{X}(t_1), \ldots, \boldsymbol{X}(t_n)$ of the random vector $\boldsymbol{X}$ in (1.2). The resulting finite sample model can be expressed as:

$$\boldsymbol{X}(t_i) = \mathbf{A}^* \boldsymbol{S}^*(t_i) + \boldsymbol{E}(t_i), \quad i = 1, \ldots, n, \tag{3.1}$$

which leads to

$$\mathbf{X} = \mathbf{A}^* \mathbf{S}^* + \mathbf{E}, \tag{3.2}$$

where $\mathbf{X} = (\boldsymbol{X}(t_1), \ldots, \boldsymbol{X}(t_n))$ is the data matrix, $\mathbf{S}^* = (\boldsymbol{S}^*(t_1), \ldots, \boldsymbol{S}^*(t_n))$ is the component matrix, and $\mathbf{E} = (\boldsymbol{E}(t_1), \ldots, \boldsymbol{E}(t_n))$ is the noise matrix.

3.2. Prewhitening on X

Transitioning to finite samples necessitates several modifications. Firstly, a prewhitening stage is executed prior to the deployment of any optimization algorithms. The rationale for prewhitening in this study is primarily computational. It involves rescaling the abnormally large or small values that appear in the sample covariance matrix and enhancing the stability of subsequent computations. The prewhitening procedure is a two-stage approach. Initially, the data matrix $\mathbf{X}$ is recentered by subtracting the sample mean $\overline{\mathbf{X}}$ from each column of $\mathbf{X}$. Next, the whitening matrix $\mathbf{B} \in \mathbb{R}^{p\times p}$ is determined such that $\mathbf{B}\widehat{\Sigma}_{\boldsymbol{X}}\mathbf{B}^\top = \mathbf{I}_p$, where $\widehat{\Sigma}_{\boldsymbol{X}}$ represents the sample covariance matrix of $\mathbf{X}$. This whitening matrix ensures that the sample covariance matrix of the prewhitened data $\breve{\mathbf{X}}$ becomes an identity matrix, i.e., $\Sigma_{\breve{\mathbf{X}}} = \mathbf{I}_p$. Subsequently, the recentered data matrix undergoes a multiplication operation on both sides by $\mathbf{B}$. This manipulation gives rise to:

$$\breve{\boldsymbol{X}}(t_i) = \breve{\mathbf{A}}^*(\boldsymbol{S}^*(t_i) - \overline{\mathbf{S}}) + \breve{\boldsymbol{E}}(t_i), \quad i = 1, \ldots, n,$$

where $\breve{\boldsymbol{X}}(t_i) = \mathbf{B}(\boldsymbol{X}(t_i) - \overline{\mathbf{X}})$, $\breve{\mathbf{A}}^* = (\breve{\boldsymbol{a}}_1^*, \cdots, \breve{\boldsymbol{a}}_{N^*}^*) = \mathbf{B}\mathbf{A}^*$, and $\breve{\boldsymbol{E}}(t_i) = \mathbf{B}(\boldsymbol{E}(t_i) - \overline{\mathbf{E}})$. Transforming $\breve{\mathbf{X}}$ back to the original scale involves a transformation of $\breve{\mathbf{A}}^*$ without affecting assumptions on the distribution of $\boldsymbol{S}^*$. Throughout this section, all subsequent computations will be conducted on the prewhitened dataset, denoted as $\breve{\mathbf{X}}$, in lieu of the original dataset $\mathbf{X}$, unless explicitly specified otherwise. When we are finished estimating $\breve{\mathbf{A}}^*$, we can estimate $\mathbf{A}^*$ by transforming back to the original scale.

Remark 4. *The prewhitening step, although not a theoretical necessity as demonstrated in Chapter 2, has shown practical benefits in our numerical results. We observe that the accuracy of the results is typically higher with prewhitening as compared to the non-whitening version. Furthermore, the*

convergence of dICA *is faster when applied to the prewhitened data,* $\breve{\mathbf{X}}$*. We provide an example where the estimation error decreases by 12% and the iteration steps decrease by 50%. The graphical results are in Appendix B.*

3.3. Estimate the Noise Covariance Matrix with Whitened Data

This section starts with constructing the space $\breve{\mathcal{A}} := \text{span}\{\text{vec}(\breve{\boldsymbol{a}}_1^* \breve{\boldsymbol{a}}_1^{*\top}), \ldots, \text{vec}(\breve{\boldsymbol{a}}_{N^*}^* \breve{\boldsymbol{a}}_{N^*}^{*\top})\}$. Similar to Section 2.3, we firstly randomly generate $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_q$ from uniform distribution U(-0.05, 0.05), where $q > p$. Then we construct the following matrix:

$$\widehat{\mathbf{GC}} = (\text{vec}(\widehat{\mathcal{C}}_{\breve{\mathbf{X}}}(\boldsymbol{\tau}_1) - \widehat{\mathcal{C}}_{\breve{\mathbf{X}}}(\mathbf{0})), \ldots, \text{vec}(\widehat{\mathcal{C}}_{\breve{\mathbf{X}}}(\boldsymbol{\tau}_q) - \widehat{\mathcal{C}}_{\breve{\mathbf{X}}}(\mathbf{0}))) \in \mathbb{R}^{p^2 \times q}. \tag{3.3}$$

Here $\widehat{\mathcal{C}}_{\breve{\mathbf{X}}}(\boldsymbol{\tau}_l)$ is the sample GCM of $\breve{\mathbf{X}}$ and takes the following form (Yeredor, 2000):

$$\widehat{\mathcal{C}}_{\breve{\mathbf{X}}}(\boldsymbol{\tau}_\ell) = \frac{1}{\sum_{i=1}^n w_{t_i,\boldsymbol{\tau}_\ell}} \sum_{i=1}^n w_{t_i,\boldsymbol{\tau}_\ell} \{\breve{\boldsymbol{X}}(t_i) - \overline{\breve{\mathbf{X}}}(\boldsymbol{.})\}\{\breve{\boldsymbol{X}}(t_i) - \overline{\breve{\mathbf{X}}}(\boldsymbol{.})\}^\top, \tag{3.4}$$

with $w_{t_i,\boldsymbol{\tau}_\ell} = \exp(\boldsymbol{\tau}_\ell^\top \breve{\boldsymbol{X}}(t_i))$ and $\overline{\breve{\mathbf{X}}}(\boldsymbol{.}) = n^{-1} \sum_{i=1}^n \breve{\boldsymbol{X}}(t_i)$ being the sample mean vector of $\breve{\mathbf{X}}$. After obtaining $\widehat{\mathbf{GC}}$, we can construct orthonormal basis of $\text{col}(\widehat{\mathbf{GC}})$. In the finite sample model, it is impossible to obtain N^* by counting non-zero singular values of $\widehat{\mathbf{GC}}$ since $\breve{\mathcal{A}}$ can only be approximately estimated by $\text{col}(\widehat{\mathbf{GC}})$ in the finite sample situation. Consequently, there will only be close-to-zero singular values and it is hard to determine how close-to-zero is close enough to be considered as a zero value. Instead, using the fact that $N^* \leq p$, we construct $\widehat{\breve{\boldsymbol{H}}}_1, \ldots, \widehat{\breve{\boldsymbol{H}}}_p$ corresponding to the p largest singular values of $\widehat{\mathbf{GC}}$. Consequently, $\widehat{\breve{\mathcal{A}}} = \text{span}\{\widehat{\breve{\boldsymbol{H}}}_1, \ldots, \widehat{\breve{\boldsymbol{H}}}_p\}$. Therefore, $\widehat{\breve{\mathcal{A}}}^\perp$ contains only $p^2 - p$ orthonormal vectors $\{\widehat{\breve{\boldsymbol{F}}}_i\}_{i=1}^{p^2-p}$. While $p^2 - p$ is smaller than the dimension of $\breve{\mathcal{A}}^\perp$ (which is $p^2 - N^*$), these $p^2 - p$ orthonormal vectors are sufficient for identifying the p diagonal entries in $\Sigma_{\boldsymbol{E}}$. The algorithm is concisely summarized in Algorithm 2.

Algorithm 2 *Estimation of* $\{\widehat{\breve{\boldsymbol{F}}}_i\}_{i=1}^{p^2-p}$.

Input: The data matrix $\breve{\mathbf{X}} = (\breve{\boldsymbol{X}}(t_1), \ldots, \breve{\boldsymbol{X}}(t_n))$.

1: Randomly generate $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_q$ from U(-0.05, 0.05), where $q > p$.

2: Obtain the matrix

$$\widehat{\mathbf{GC}} = (\mathrm{vec}(\widehat{\mathcal{C}}_{\breve{\mathbf{X}}}(\boldsymbol{\tau}_1) - \widehat{\mathcal{C}}_{\breve{\mathbf{X}}}(0)), \ldots, \mathrm{vec}(\widehat{\mathcal{C}}_{\breve{\mathbf{X}}}(\boldsymbol{\tau}_q) - \widehat{\mathcal{C}}_{\breve{\mathbf{X}}}(0))) \in \mathbb{R}^{p^2 \times q}. \tag{3.5}$$

3: Compute the left-singular vectors $\widehat{\breve{\boldsymbol{H}}}_1, \ldots, \widehat{\breve{\boldsymbol{H}}}_p$ corresponding to the p largest singular values of $\widehat{\mathbf{GC}}$.

4: Obtain $\{\widehat{\breve{\boldsymbol{F}}}_i\}_{i=1}^{p^2-p}$, which are orthogonal to $\mathrm{span}\{\widehat{\breve{\boldsymbol{H}}}_1, \ldots, \widehat{\breve{\boldsymbol{H}}}_p\}$.

Output: orthonormal vectors $\{\widehat{\breve{\boldsymbol{F}}}_i\}_{i=1}^{p^2-p}$.

3.4. The dICA Algorithm for the Sample Model

Motivated from equation (2.17) and iteration in (2.19), dICA for the Sample Model can also be decomposed into solving two subproblems. The first subproblem is optimizing an extension of equation (2.17) into the finite sample model, and the second one is finding the fixed point of an extension of the iteration in (2.19). For a smoother transition, we start with the following preliminary version of the first subproblem.

$$\begin{aligned} \breve{\mathbf{W}}_{\text{interme}}, \widehat{\Sigma}_{\boldsymbol{E}} \;=\; \arg \min_{\substack{\breve{\mathbf{W}} \in \mathbb{R}^{p \times p}, \\ \mathrm{diag}(\Sigma) \geq 0, \\ \Sigma \text{ is diagonal}}} & \|\breve{\mathbf{W}}\|_F^2 + \rho_1 \textstyle\sum_{j=1}^{p} \lambda_j \|\breve{\boldsymbol{w}}_j\|_2, \\ \text{subject to} \quad & \|(\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top)\breve{\mathbf{W}}^\top\breve{\mathbf{W}}(\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top) - (\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top)\|_F^2 \leq \delta_1, \\ & \|\breve{\mathbf{W}}\|_F^2 \textstyle\sum_{j=1}^{p^2-p} |\langle \widehat{\breve{\boldsymbol{F}}}_j, \mathrm{vec}(\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top)\rangle|^2 \leq \delta_2. \end{aligned} \tag{3.6}$$

The parameters δ_1 and δ_2 are introduced to account for the errors attributable to "sampling noise". Specifically, in the finite sample model, $\{\breve{\boldsymbol{F}}_j\}_{j=1}^{p^2-N^*}$, $\Sigma_{\boldsymbol{E}}$, and $\Sigma_{\boldsymbol{X}}$ can only be estimated approximately. As a result, there may not exist any diagonal matrix Σ such that $(\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top)$ has a low-rank structure or $\sum_{j=1}^{p^2-p} |\langle \widehat{\breve{\boldsymbol{F}}}_j, \mathrm{vec}(\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top)\rangle|^2 = 0$. We need δ_1 and δ_2 to accommodate for these differences. In finite sample situation, $\|\breve{\mathbf{W}}\|_F^2$ can shrink the rows of $\breve{\mathbf{W}}$ towards zero but cannot precisely reduce them to zero. Motivating from the (Group) Lasso literature (Yuan and Lin, 2006; Zou, 2006), this study proposes

$\sum_{j=1}^{p} \lambda_j \|\breve{\boldsymbol{w}}_j\|_2$ to encourage the sparsity in $\breve{\mathbf{W}}$ where each λ_j adjusts the penalization weight for the corresponding $\breve{\boldsymbol{w}}_j$.

One may notice that $\sum_{j=1}^{p^2-p} |\langle \widehat{\breve{\boldsymbol{F}}}_j, \mathrm{vec}(\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top)\rangle|^2$ is augmented by a factor $\|\breve{\mathbf{W}}\|_F^2$. This modification warrants justification. The objective of minimizing $\sum_{j=1}^{p^2-p} |\langle \widehat{\breve{\boldsymbol{F}}}_j, \mathrm{vec}(\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top)\rangle|^2$ is to angle $\widehat{\breve{\boldsymbol{F}}}_j$ and $\mathrm{vec}(\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top)$ to be as orthogonal as possible. However, when working with finite samples, the $\widehat{\Sigma}_{\boldsymbol{E}}$ minimizing $\sum_{j=1}^{p^2-p} |\langle \widehat{\breve{\boldsymbol{F}}}_j, \mathrm{vec}(\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top)\rangle|^2$ is not necessarily the one that maximizes the perpendicularity between $\widehat{\breve{\boldsymbol{F}}}_j$ and $\mathrm{vec}(\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top)$, but rather a compromise between maximizing perpendicularity and minimizing the norm of $\mathrm{vec}(\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top)$. A natural resolution would be normalization. However, directly normalizing the objective function by factoring in $1/\|\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top\|_F$ would over-complicate the optimization. A feasible workaround is to use $\|\breve{\mathbf{W}}\|_F$ to replace $1/\|\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top\|_F$. This approximation is based on the observation that $(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)\breve{\mathbf{W}}^\top\breve{\mathbf{W}}(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top) \approx (\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)$.

Rewriting the inequality constraints in equation (3.6) as penalty terms leads us to the final version of the first subproblem.

$$
\begin{aligned}
\breve{\mathbf{W}}_{\text{interme}}, \widehat{\Sigma}_{\boldsymbol{E}} = \quad & \underset{\substack{\breve{\mathbf{W}}\in\mathbb{R}^{p\times p},\\ \mathrm{diag}(\Sigma)\geq 0,\\ \Sigma \text{ is diagonal},\\ \delta_1 \geq 0}}{\operatorname{argmin}} \quad \|\breve{\mathbf{W}}\|_F^2 \Big(\sum_{j=1}^{p^2-p} |\langle \widehat{\breve{\boldsymbol{F}}}_j, \mathrm{vec}(\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top)\rangle|^2\Big) \\
& \qquad\qquad + \rho_1 \textstyle\sum_{j=1}^{p} \lambda_j \|\breve{\boldsymbol{w}}_j\|_2 \\
& \qquad\qquad + \rho_2 \big(\|(\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top)\breve{\mathbf{W}}^\top\breve{\mathbf{W}}(\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top) - (\mathbf{I}_p - \mathbf{B}\Sigma\mathbf{B}^\top)\|_F^2 - \delta_1\big)^2, \\
& \text{subject to} \quad \delta_1 \leq c_1 p^2/n.
\end{aligned}
\tag{3.7}
$$

By comparing (3.6) with (3.7), it becomes evident that the constraints in equation (3.6) have been transformed into penalty functions. This technique is commonly used in constraint optimization (Tibshirani, 1996). Asymptotically, as n grows, the square of the Frobenius norm of the difference between $(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)$ and $(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)\breve{\mathbf{W}}^\top\breve{\mathbf{W}}(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)$ is of the order $O(p^2/n)$. Based on this asymptotic behavior, the value of δ_1 is upper bounded by $c_1 p^2/n$. Section 3.4.1 will discuss the selection of the tuning parameter c_1.

In equation (3.7), there are three tuning parameters: ρ_1, ρ_2, and c_1. Among these parameters, ρ_2 needs to be significantly larger than $\max(\rho_1, 1)$ as the hidden orthogonal constraint ρ_2 corresponding to is the most fundamental constraint in ICA literature and it also protects the significant (non-zero) rows in $\breve{\mathbf{W}}$ from being shrunk by $\|\cdot\|_2$. ρ_1 and c_1 together control the sparsity level in $\breve{\mathbf{W}}$. Among these two, ρ_1 represents the baseline penalty strength on the rows of $\breve{\mathbf{W}}$. It is important not to set ρ_1 too large as it would overshadow the soft constraint regarding Σ. In our numerical experiments, we choose $\rho_1 = 1$. In Chapter 2, it has been demonstrated that the performance of dICA remains consistent regardless of the penalty level on $\|\cdot\|_F^2$ in the population model. Similarly, in the finite sample model, the performance of dICA remains consistent regardless of the choice of ρ_1, as long as it does not excessively distort the scales of the rows in $\breve{\mathbf{W}}$. To simplify matters, a baseline penalty level of 1, denoted as ρ_1, is chosen. The determination of λ_j, which plays a more crucial role in determining the structure of $\breve{\mathbf{W}}$ compared to ρ_1, will be discussed later in this section. One unique characteristic of the proposed dICA is that, due to the presence of the hidden orthogonal constraint, $\|\cdot\|_2$ cannot freely penalize the rows in $\breve{\mathbf{W}}$. In order to achieve the desired sparsity level, it is necessary to adjust the parameter c_1. A very small value of c_1 will result in a tight bound on the difference between $(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)$ and $(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)\breve{\mathbf{W}}^\top\breve{\mathbf{W}}(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)$, thereby failing to account for the rank difference between these two matrices. Consequently, $(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)\breve{\mathbf{W}}^\top\breve{\mathbf{W}}(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)$ will be forced to have a rank close to that of $(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)$, resulting in more non-zero rows in $\breve{\mathbf{W}}$. On the other hand, a very large value of c_1 will create a large gap between $(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)$ and $(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)\breve{\mathbf{W}}^\top\breve{\mathbf{W}}(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)$, causing the scales of the meaningful rows in $\breve{\mathbf{W}}$ to be distorted by $\|\cdot\|_2$. Therefore, c_1 needs to be chosen carefully. It should be large enough to accommodate the rank difference between $(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)\breve{\mathbf{W}}^\top\breve{\mathbf{W}}(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)$ and $(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)$, while also small enough to protect the non-zero rows in $\breve{\mathbf{W}}$ from being overly shrunk by $\|\cdot\|_2$. In Section 3.4.1, this study proposes a data driven method to obtain c_1.

We propose Algorithm 3 to solve (3.7). It updates $\breve{\mathbf{W}}$, Σ, δ_1 in a block-coordinate-descent

Algorithm 3 *Algorithm solving equation* (3.7)

Input: The data matrix $\breve{\mathbf{X}} = (\breve{\boldsymbol{X}}(t_1), \ldots, \breve{\boldsymbol{X}}(t_n))$, $\{\widehat{\breve{\boldsymbol{F}}}_i\}_{i=1}^{p^2-p}$, threshold ξ, initial value $\breve{\mathbf{W}}^{(0)}$, initial value for $\Sigma^{(0)} := \mathrm{diag}\{\sigma_{e_1}^{(0)}, \cdots, \sigma_{e_p}^{(0)}\}$, initial value for $\delta_1^{(0)}$, max iteration number.

for $l = 0, \cdots,$ max iter **do**
  for $j = 1, \ldots, p$ **do**
    if $\breve{\boldsymbol{w}}_j^{(l)\top}(\mathbf{I}_p - \mathbf{B}\Sigma^{(l)}\mathbf{B})\breve{\boldsymbol{w}}_j^{(l)} < \xi$ **then**
      $\lambda_j = 1/\{\breve{\boldsymbol{w}}_j^{(l)\top}(\mathbf{I}_p - \mathbf{B}\Sigma^{(l)}\mathbf{B})\breve{\boldsymbol{w}}_j^{(l)}\}$
    end if
  end for
  for $j = 1, \cdots, p$ **do**
    $\breve{\boldsymbol{w}}_j^{(l+1)} \leftarrow \mathrm{prox}_{\lambda_j, \gamma_j^{(l)}}(\breve{\boldsymbol{w}}_j^{(l)} - \gamma_j^{(l)} \nabla_{\breve{\boldsymbol{w}}_j^{(l)}} \Gamma(\breve{\mathbf{W}}^{(l)}, \Sigma^{(l)}, \delta_1^{(l)}))$.
  end for
  Here $\gamma_j^{(l)}$ is selected through Armijo backtracking (Armijo, 1966; Nocedal and Wright, 2006) and

$$\mathrm{prox}_{\lambda_j, \gamma_j^{(l)}}(\boldsymbol{z}) := \begin{cases} (1 - \rho_2 \gamma_j^{(l)} \lambda_j / \|\boldsymbol{z}\|_2)\boldsymbol{z}, & \text{if } \|\boldsymbol{z}\|_2 > \rho_2 \gamma_j^{(l)} \lambda_j, \\ \mathbf{0}, & \text{if } \|\boldsymbol{z}\|_2 \le \rho_2 \gamma_j^{(l)} \lambda_j. \end{cases}$$

  $\breve{\mathbf{W}}^{(l+1)} = (\breve{\boldsymbol{w}}_1^{(l+1)}, \cdots, \breve{\boldsymbol{w}}_p^{(l+1)})^\top$.
  for $j = 1, \cdots, p$ **do**

$$\sigma_{e_j}^{(l+1)} = \underset{\sigma \ge 0}{\mathrm{argmin}}\, p_j(\sigma). \tag{3.8}$$

  end for
  (3.8) can be solved through Matlab's "root" function in "symbolic" package (The MathWorks, 2019).
  $\Sigma^{(l+1)} = \mathrm{diag}\{\sigma_{e_1}^{(l+1)}, \cdots, \sigma_{e_p}^{(l+1)}\}$
  $\delta_1^{(l+1)} = \min(c_1 p^2/n, \|(\mathbf{I}_p - \mathbf{B}\Sigma^{(l+1)}\mathbf{B}^\top)\breve{\mathbf{W}}^\top\breve{\mathbf{W}}(\mathbf{I}_p - \mathbf{B}\Sigma^{(l+1)}\mathbf{B}^\top) - (\mathbf{I}_p - \mathbf{B}\Sigma^{(l+1)}\mathbf{B}^\top)\|_F^2)$.
end for
$\breve{\mathbf{W}}_{\text{inter}} = \breve{\mathbf{W}}^{(l+1)}$, $\widehat{\Sigma}_{\boldsymbol{E}} = \Sigma^{(l+1)}$.

Output: $\breve{\mathbf{W}}_{\text{inter}}$ and $\widehat{\Sigma}_{\boldsymbol{E}}$.

fashion (Wright, 2015). $\breve{\mathbf{W}}$ is the first block, second block consists of diagonal elements of Σ, and δ_1 is the third block. Within the first block, $\breve{\mathbf{W}}$ is updated using proximal gradient descent (Parikh and Boyd, 2014). In the second block, the diagonal entries $\sigma_{e_1}, \cdots, \sigma_{e_p}$ of Σ are updated individually using coordinate descent (Wright, 2015). Denote the objective function in equation (3.7) as $\Gamma(\breve{\mathbf{W}}, \Sigma, \delta_1)$. For a given σ_{e_j}, with all other variables held constant, the objective function $\Gamma(\cdot)$ in equation (3.7) reduces to an 8-th order polynomial of σ_{e_j}, denoted as $p_j(\cdot)$. The coordinate descent in this block involves minimizing $p_j(\cdot)$ for $j = 1, \cdots, p$. Updating δ_1 is straightforward after the updating of $\breve{\mathbf{W}}$ and Σ.

After obtaining $\breve{\mathbf{W}}_{\text{interme}}$ and $\widehat{\Sigma}_{\boldsymbol{E}}$, the next step, following the framework in Section 2.6, is to extend equation (2.19) to its empirical version. This involves finding the fixed point $\widehat{\mathbf{O}}$ of the following iterations:

$$\overline{\boldsymbol{o}}_j \leftarrow \mathrm{E}_n\{\overline{\mathbf{X}}\varphi^{(-1)}_{d(\boldsymbol{o}_j)}(\boldsymbol{o}_j^\top \overline{\mathbf{X}})\} - \{\beta \mathrm{cov}(\overline{\mathbf{X}}) + (1-\beta)(\mathbf{I}_p + \overline{\Sigma}_{\boldsymbol{E}})\}\boldsymbol{o}_j \mathrm{E}_n\{\varphi_{d(\boldsymbol{o}_j)}(\boldsymbol{o}_j^\top \overline{\mathbf{X}})\} \quad (3.9)$$

for $1 \le j \le p$,

$$\mathbf{O} \leftarrow (\overline{\mathbf{O}\mathbf{O}}^\top)^{-1/2}\overline{\mathbf{O}}. \quad (3.10)$$

Here, $\mathrm{E}_n\{f(\breve{\mathbf{X}})\} := \frac{1}{n}\sum_{i=1}^n f(\breve{\boldsymbol{X}}(t_i))$, $\mathbf{O}$ (or $\overline{\mathbf{O}}$) $= (\boldsymbol{o}_1, \cdots, \boldsymbol{o}_p)^\top$ (or $(\overline{\boldsymbol{o}}_1, \cdots, \overline{\boldsymbol{o}}_p)^\top$), $\overline{\mathbf{X}} = \breve{\mathbf{W}}_{\text{interme}}\breve{\mathbf{X}}$, $\overline{\Sigma}_{\boldsymbol{E}} = \breve{\mathbf{W}}_{\text{interme}}\mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top\breve{\mathbf{W}}_{\text{interme}}^\top$, and $d(\boldsymbol{o}_j) = \sqrt{c^2 - \boldsymbol{o}_j^\top \overline{\Sigma}_{\boldsymbol{E}}\boldsymbol{o}_j}$. The $\widehat{\mathbf{O}}$ can be obtained by iteratively running iterations (3.9) to (3.10) until convergence is achieved. This process yields $\breve{\mathbf{W}}_{\text{seq}} := \widehat{\mathbf{O}}\breve{\mathbf{W}}_{\text{interme}}$. The subscript "seq" in $\breve{\mathbf{W}}_{\text{seq}}$ indicates that it is obtained by optimizing equation (3.7) and sequentially finding $\widehat{\mathbf{O}}$. This sequential approach is proposed in Section 2.6 and has theoretical guarantees in the population model. However, it is important to note that this sequential approach may not produce a sparse solution in the finite sample model. The intermediate matrix $\breve{\mathbf{W}}_{\text{interme}}$ does not necessarily exhibit a sparse structure. As a result, it is possible that none of its rows can be shrunk to zero. The orthogonal transformation $\widehat{\mathbf{O}}$ is intended to redistribute the norms of the rows in $\breve{\mathbf{W}}_{\text{interme}}$ and produce a sparse solution based on the theoretical results from population model. However, due to the presence of "finite sampling noise", $\widehat{\mathbf{O}}$ in the finite sample model

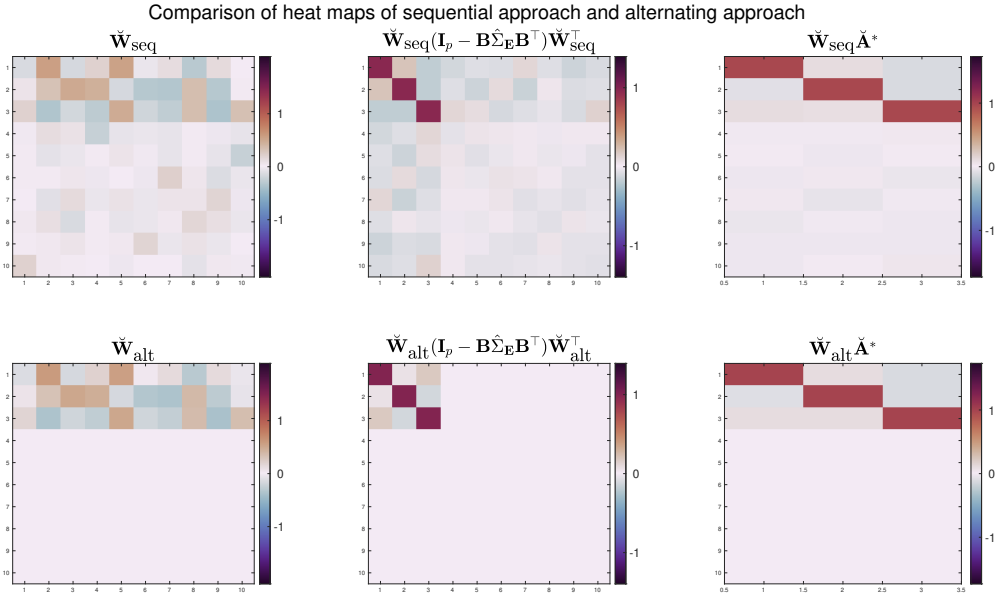

Figure 3.1: Comparison between sequential approach and alternating approach

can only lead to an approximately sparse solution, denoted as $\breve{\mathbf{W}}_{\text{seq}}$. To illustrate this, an example is provided in Figure 3.1 with $(p, N^*) = (10, 3)$. The top left panel of Figure 3.1 clearly shows that the first three rows of $\breve{\mathbf{W}}_{\text{seq}}$ are significantly different from zero, while the remaining seven rows have small norms but are not exactly zero.

To address this issue and obtain a truly sparse solution, we propose the following approach that runs in an alternate manner.

1. Set $\breve{\mathbf{W}}_{\text{interme}}$ and $\widehat{\Sigma}_{\boldsymbol{E}}$ in equation (3.7) as $\breve{\mathbf{W}}_{\text{interme}}^{(0)}$ and $\widehat{\Sigma}_{\boldsymbol{E}}^{(0)}$.

2. Start with $\overline{\mathbf{X}}^{(l)} = \breve{\mathbf{W}}_{\text{interme}}^{(l)}\breve{\mathbf{X}}$, run iterations (3.9) to (3.10) once, and obtain $\widehat{\mathbf{O}}^{(l)}$.

3. Obtain $\breve{\mathbf{W}}^{(l)} = \mathbf{O}^{(l)}\breve{\mathbf{W}}_{\text{interme}}^{(l)}$.

4. Run only one iteration step of Algorithm 3 starting from $\breve{\mathbf{W}}^{(l)}$ and $\widehat{\Sigma}_{\boldsymbol{E}}^{(l)}$, and obtain $\breve{\mathbf{W}}_{\text{interme}}^{(l+1)}$ and $\widehat{\Sigma}_{\boldsymbol{E}}^{(l+1)}$.

5. Update $l = l + 1$ and return to step 2.

In this alternate approach, whenever we apply the transformation to the rows in $\breve{\mathbf{W}}^{(l)}_{\text{interme}}$ using $\mathbf{O}^{(l)}$ in step 3, some rows may become approximately zero. We then immediately proceed to execute one iteration of Algorithm 3 with the objective of reducing some of the nearly-zero rows to exactly zero. As previously mentioned, the λ_js in equation (3.7) play a crucial role in promoting sparsity by controlling the penalizing weights assigned to different rows. In line with the approach taken in Zou (2006), higher penalizing weights are assigned to the potentially nuisance rows in order to encourage their faster shrinkage to zero. As we approach convergence, as illustrated in the top middle of Figure 3.1, the following equation (3.11) will be satisfied, and it aids in determining whether a row $\breve{\boldsymbol{w}}_j^{(l)}$ should be considered as a potential nuisance row or not.

$$\breve{\mathbf{W}}^{(l)}(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)\breve{\mathbf{W}}^{(l)\top} \approx \begin{pmatrix} \mathbf{I}_{N^*} & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.11}$$

From equation (3.11), if $\breve{\boldsymbol{w}}_j^{(l)\top}(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)\breve{\boldsymbol{w}}_j^{(l)} \approx 1$, then $\breve{\boldsymbol{w}}_j^{(l)}$ is likely a non-nuisance row. On the other hand, if $\breve{\boldsymbol{w}}_j^{(l)\top}(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)\breve{\boldsymbol{w}}_j^{(l)} \approx 0$, then $\breve{\boldsymbol{w}}_j^{(l)}$ is a potentially nuisance row. In this study, the initial values of λ_j are set to 1, and once $\breve{\boldsymbol{w}}_j^{(l)\top}(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)\breve{\boldsymbol{w}}_j^{(l)}$ falls below a specific threshold ξ (set at 0.1 in this study), we update λ_j as $1/\{\breve{\boldsymbol{w}}_j^{(l)\top}(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)\breve{\boldsymbol{w}}_j^{(l)}\}$ in order to further shrink $\breve{\boldsymbol{w}}_j^{(l)}$. An alternative option would be to utilize the eigenvalues of $\breve{\mathbf{W}}^{(l)}(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)\breve{\mathbf{W}}^{(l)\top}$. Initially, we identify the r eigenvalues of $\breve{\mathbf{W}}^{(l)}(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)\breve{\mathbf{W}}^{(l)\top}$ that are below the threshold value ξ. Subsequently, the r rows in $\breve{\mathbf{W}}^{(l)}$ that correspond to the r smallest diagonal entries in the matrix $\breve{\mathbf{W}}^{(l)}(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)\breve{\mathbf{W}}^{(l)\top}$ will be assigned λ_j values equal to the reciprocals of those eigenvalues. In certain scenarios where the value of N^* is significantly smaller than p, it may be appropriate to assign a more aggressive penalty weight. In this case, the penalty weights λ_js can be defined as $1/\{\breve{\boldsymbol{w}}_j^{(l)\top}(\mathbf{I}_p - \mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top)\breve{\boldsymbol{w}}_j^{(l)}\}^2$. The purpose of this more aggressive penalty weight is to penalize nuisance rows more severely, thereby facilitating the shrinkage process. It is important to note that our approach is not highly sensitive to the choice of threshold, and a value of 0.1 is selected for simplicity.

By using the proposed approach to adjust λ_j after each iteration cycle, the alternative method can achieve a sparse structure, as shown in the second row of Figure 3.1.

Since the proposed algorithm operates alternately, it is possible that certain rows in $\breve{\mathbf{W}}^{(l)}_{\text{interme}}$ may be reduced to zero before the algorithm converges. These rows, which have been reduced to zero, are considered to be nuisance rows and should not participate in the subsequent computation process. To further decrease the computational load, we have developed Algorithm 4 to determine the fixed point $\widehat{\mathbf{O}}$. This algorithm utilizes an active set method (Nocedal and Wright, 2006) to enhance computational efficiency. Specifically, it updates only the non-nuisance columns and rows in $\mathbf{O}$ that correspond to the non-zero rows in $\breve{\mathbf{W}}^{(l)}_{\text{interme}}$.

Algorithm 4 *Algorithm finding the fixed point* $\widehat{\mathbf{O}}$

Input: $\breve{\mathbf{X}}$, an initial orthonormal matrix $\mathbf{O}_0$ (default is $\mathbf{I}_p$), a weight β, $\breve{\mathbf{W}}_{\text{interme}} := (\breve{\boldsymbol{w}}_{\text{interme},1}, \cdots, \breve{\boldsymbol{w}}_{\text{interme},p})^\top$, and $\widehat{\Sigma}_{\boldsymbol{E}}$.

1: $\mathbb{A} := \{j : \|\breve{\boldsymbol{w}}_{\text{interme},j}\| \neq 0, \quad j = 1, \cdots, p\}$.
2: $\widehat{N} = |\mathbb{A}|$
3: Let $\breve{\mathbf{W}}_{\mathbb{A}}$ consist of non-zero rows in $\breve{\mathbf{W}}_{\text{interme}}$.
4: Let $\mathbb{A}[j]$ be the $j-$th element in $\mathbb{A}$, and create a orthonormal matrix $\mathbf{O}_{\mathbb{A}} \in \mathbb{R}^{\widehat{N}\times\widehat{N}}$.
5: Compute $\overline{\mathbf{X}}_{\mathbb{A}} = \breve{\mathbf{W}}_{\mathbb{A}}\breve{\mathbf{X}}$ and $\overline{\Sigma}_{\boldsymbol{E}} = \breve{\mathbf{W}}_{\mathbb{A}}\mathbf{B}\widehat{\Sigma}_{\boldsymbol{E}}\mathbf{B}^\top\breve{\mathbf{W}}_{\mathbb{A}}^\top$.
6: **for** $l = 1, \cdots,$ max iter **do**
7: Get $\{\boldsymbol{o}_j\}_{j=1}^{\widehat{N}}$ from $\mathbf{O}_{\mathbb{A}} = (\boldsymbol{o}_1, ..., \boldsymbol{o}_{\widehat{N}})^\top$.
8: **for** $j = 1, \ldots, \widehat{N}$ **do**
9: $\overline{\boldsymbol{o}}_j = \mathrm{E}_n\{\overline{\mathbf{X}}_{\mathbb{A}}\varphi^{(-1)}_{d(\boldsymbol{o}_j)}(\boldsymbol{o}_j^\top\overline{\mathbf{X}}_{\mathbb{A}})\} - \{\beta\mathrm{cov}(\overline{\mathbf{X}}_{\mathbb{A}}) + (1-\beta)(\mathbf{I}_{\widehat{N}} + \overline{\Sigma}_{\boldsymbol{E}})\}\boldsymbol{o}_j\mathrm{E}_n\{\varphi_{d(\boldsymbol{o}_j)}(\boldsymbol{o}_j^\top\overline{\mathbf{X}}_{\mathbb{A}})\}$.
10: **end for**
11: Compute $\overline{\mathbf{O}}_{\mathbb{A}} = (\overline{\boldsymbol{o}}_1, \ldots, \overline{\boldsymbol{o}}_{\widehat{N}})^\top$ and $\mathbf{O}_{\mathbb{A}} = (\overline{\mathbf{O}}_{\mathbb{A}}\,\overline{\mathbf{O}}_{\mathbb{A}}^\top)^{(-1/2)}\overline{\mathbf{O}}_{\mathbb{A}}$.
12: **end for**
13: Replace the $\mathbb{A}[i], \mathbb{A}[j]-$th element of $\mathbf{O}_0$ ($\mathbf{O}_0(\mathbb{A}[i], \mathbb{A}[j])$) with $\mathbf{O}_{\mathbb{A}}(i, j)$ and obtain $\widehat{\mathbf{O}}$. Here $1 \leq i, j \leq \widehat{N}$

Output: Orthonormal matrix $\widehat{\mathbf{O}}$.

The dICA approach for the finite sample model, which is a combination of Algorithm 3 and Algorithm 4 is provided in Algorithm 5.

Algorithm 5 dICA *for the finite sample model*

Input: The data matrix $\mathbf{X}$, initial $\breve{\mathbf{W}}^{(0)}$, ρ_1, ρ_2, max iter1, max iter2, threshold ξ, $\{\lambda_j\}_{j=1}^{p} = 1$, $\Delta = \infty$, candidate set $\mathbb{S}$ for c_1.

Prewhiten and obtain $\mathbf{B}$ and $\breve{\mathbf{X}}$ $(= \mathbf{B}\mathbf{X})$.
Run Algorithm 2 or higher-order methods and obtain $\{\widehat{\breve{\boldsymbol{F}}}_i\}_{i=1}^{p^2-p}$.
for c_1 in $\mathbb{S}$ **do**
  $\epsilon = c_1 p^2/n$
  Run Algorithm 3 for max iter1 times and obtain $\breve{\mathbf{W}}_{\text{interme}}$ and $\widehat{\Sigma}_{\boldsymbol{E}}$.
  Reset $\breve{\mathbf{W}}^{(0)} := \breve{\mathbf{W}}_{\text{interme}}$, $\Sigma^{(0)} := \widehat{\Sigma}_{\boldsymbol{E}}$.
  for $l = 0, \ldots,$ max iter2 **do**
    if $\|\breve{\mathbf{W}}^{(l+1)} - \breve{\mathbf{W}}^{(l)}\|_F <$ stopping criterion **then**
      $\widehat{\breve{\mathbf{W}}} = \breve{\mathbf{W}}^{(l)}$, $\widehat{\Sigma}_{\boldsymbol{E}} = \Sigma^{(l)}$
      Break!
    end if
    Run one iteration of Algorithm 3 initiating from $\breve{\mathbf{W}}^{(l)}$ and $\Sigma^{(l)}$ and obtain $\breve{\mathbf{W}}^{(l+1)}_{\text{interme}}$ and $\Sigma^{(l+1)}$.
    Run one iteration of Algorithm 4 and obtain $\widehat{\mathbf{O}}^{(l+1)}$.
    $\breve{\mathbf{W}}^{(l+1)} = \widehat{\mathbf{O}}^{(l+1)}\breve{\mathbf{W}}^{(l+1)}_{\text{interme}}$
  end for
  Let $\mathbf{W}$ consist of the non-zero rows of $\widehat{\breve{\mathbf{W}}}\mathbf{B}$.
  if $\|(\mathbf{W}(\widehat{\Sigma}_{\mathbf{X}} - \widehat{\Sigma}_{\boldsymbol{E}})^2\mathbf{W}^\top)^{-1}\mathbf{W}(\widehat{\Sigma}_{\mathbf{X}} - \widehat{\Sigma}_{\boldsymbol{E}})\mathbf{W}^\top(\mathbf{W}\mathbf{W}^\top)^{-1} - \mathbf{I}\|_F < \Delta$ **then**
    $\widehat{\mathbf{W}}_{\text{sample}} = \mathbf{W}$
    $\Delta = \|(\mathbf{W}(\widehat{\Sigma}_{\mathbf{X}} - \widehat{\Sigma}_{\boldsymbol{E}})^2\mathbf{W}^\top)^{-1}\mathbf{W}(\widehat{\Sigma}_{\mathbf{X}} - \widehat{\Sigma}_{\boldsymbol{E}})\mathbf{W}^\top(\mathbf{W}\mathbf{W}^\top)^{-1} - \mathbf{I}\|_F$.
  end if
end for

Output: $\widehat{\mathbf{W}}_{\text{sample}}$.

3.4.1. Determine c_1

Denote the estimator of $\widetilde{\mathbf{W}}^*$ in dICA for the finite sample model as $\widehat{\widetilde{\mathbf{W}}}_{\text{sample}}$. Let us denote all the non-zero rows in $\widehat{\widetilde{\mathbf{W}}}_{\text{sample}} := \widehat{\breve{\mathbf{W}}}\mathbf{B}$ as $\widehat{\mathbf{W}}_{\text{sample}}$. Since $\widehat{\mathbf{W}}_{\text{sample}}$ depends on c_1, we denote it as $\widehat{\mathbf{W}}_{\text{sample}}(c_1)$. If $\widehat{\mathbf{W}}_{\text{sample}}(c_1)$ is a sufficiently accurate estimate, $\widehat{\mathbf{A}}_{\text{sample}} := (\widehat{\Sigma}_{\boldsymbol{X}} - \widehat{\Sigma}_{\boldsymbol{E}})\widehat{\mathbf{W}}_{\text{sample}}(c_1)^\top$ should be close to $\mathbf{A}^*$. Ideally, we would like to have:

$$\widehat{\mathbf{A}}^{\dagger}_{\text{sample}}\widehat{\mathbf{W}}_{\text{sample}}(c_1)^{\dagger} \approx \mathbf{I},$$

where $\widehat{\mathbf{A}}^{\dagger}_{\text{sample}}$ (and $\widehat{\mathbf{W}}_{\text{sample}}(c_1)^{\dagger}$) are the Moore-Penrose pseudoinverse of the matrix $\widehat{\mathbf{A}}_{\text{sample}}$ (and $\widehat{\mathbf{W}}_{\text{sample}}(c_1)$). We choose c_1 as the value that minimizes the Frobenius norm between $\widehat{\mathbf{A}}^{\dagger}_{\text{sample}}\widehat{\mathbf{W}}_{\text{sample}}(c_1)^{\dagger}$ and $\mathbf{I}$.

Chapter 4

Numerical Results

In the previous chapter, we delve into a detailed discussion on how to expand dICA to the finite sample model. We commence by presenting the definition of the finite sample model and subsequently proceed to the prewhitening step. This step serves to rescale excessively large or small values, thereby enhancing the numerical accuracy of the method. Upon completion of the prewhitening step, we proceed with the subsequent computations based on the prewhitened data matrix $\breve{\mathbf{X}}$. In the context of finite sample situations, the presence of "sampling noise" may result in deviations from the exact constraints outlined in Chapter 2. To address this, we adapt these hard constraints into soft ones and incorporate them into the objective function as penalty terms. Additionally, we offer a practical approach for selecting the tuning parameters in the dICA algorithm.

In this chapter, we will conduct various simulation studies under different settings to demonstrate the performance of the dICA approach.

4.1. Comparison with Competitors under Gaussian Noise

In this section, we scrutinize our algorithm's efficacy across diverse source distributions, focusing on instances where the components consist of $\mathrm{i.i.d.}$ observations. We evaluate the dICA from three distinct perspectives: (I) the accuracy in estimating N^*, (II) the precision

in estimating $\mathbf{A}^*$, (III) the precision in estimating $\mathbf{S}^*$.

In the evaluation of aspect (I) — the accuracy in estimating N^* — our algorithm is benchmarked against a range of competitors, including: (i) the asymptotic test as proposed by Virta and Nordhausen (2019), (ii) PCA retaining 90% of the variance, and (iii) PCA retaining 99% of the variance. For the asymptotic test, we use the implementation by Wu (2023).

For the evaluation of aspect (II), (III) — the precision in estimating $\mathbf{A}^*$ and $\mathbf{S}^*$ — the competitors include: (i) PCA+FastICA using tanh as non-linear contrast function (PFt) (Hyvärinen et al., 2001), (ii) PCA+Infomax (PI)(Bell and Sejnowski, 1995; Lee et al., 1999), (iii) PCA + GIICA (PGIC) (Voss et al., 2015). All of the competitors need an estimated $\widehat{N}$. We provide them with the $\widehat{N}$ obtained from either the asymptotic test or PCA, whichever is closer to N^*. We evaluate the performance of these methods across various situations with following configurations:

- $\mathbf{A}^*$'s dimension $(p, N^*) \in \{(10, 10), (20, 10), (50, 20), (50, 5)\}$.
- $\mathbf{A}^*$ is generated by $\mathbf{U}\mathrm{diag}(\sigma_1, \ldots, \sigma_{N^*})\mathbf{V}$, where $\mathbf{U}$ and $\mathbf{V}$ are $p \times N^*$ and $N^* \times N^*$ randomly generated orthonormal/semi-orthogonal matrices, respectively. $\{\sigma_k\}_{k=1}^{N^*}$ are generated independently from the square root of a uniform distribution, specifically, from $\sqrt{U(2, N^*)}$.
- $\Sigma_{\boldsymbol{E}}$ has 4 settings:
  - **Isotropic moderate noise:** $\Sigma_{\boldsymbol{E}} = \sigma^2\mathbf{I}$ with $\sigma^2 = \sum_{k=1}^{N^*}(\sigma_k^2 - 1)/2p$.
  - **Isotropic strong noise:** $\Sigma_{\boldsymbol{E}} = \sigma^2\mathbf{I}$ with $\sigma^2 = \sum_{k=1}^{N^*}(\sigma_k^2 - 1)/p$.
  - **Random moderate noise:** $\Sigma_{\boldsymbol{E}} = \mathrm{diag}\{\sigma_{e_1}, \ldots, \sigma_{e_p}\}$ with $\{\sigma_{e_j}\}_{j=1}^{p}$ independently randomly generated from $U(1, N^*)N^*/2p$.
  - **Random strong noise:** $\Sigma_{\boldsymbol{E}} = \mathrm{diag}\{\sigma_{e_1}, \ldots, \sigma_{e_p}\}$ with $\{\sigma_{e_j}\}_{j=1}^{p}$ independently randomly generated from $U(1, N^*)N^*/p$.
- Sample size n is chosen from $\{10000, 50000, 250000\}$.

- The components are evenly generated from three different distributions: (a) the Laplacian distribution (a typical example from the super-Gaussian family), (b) the uniform distribution (a typical example from the sub-Gaussian family), (c) Gaussian mixture (density = $0.5\mathcal{N}(-1,1) + 0.5\mathcal{N}(1,1)$). In the numerical simulations, $\lfloor N^*/3 \rfloor$ components are generated from the Laplacian distribution, while an equal number of $\lfloor N^*/3 \rfloor$ components are generated from the uniform distribution. The leftover components are generated from the Gaussian mixture.

- 50 simulations are conducted for each situation.

In terms of the accuracy of estimation of N^*, this study measures it using the percentage of correctly selecting the number N^* and the difference between $\widehat{N}$ and N^*. The results are summarized in Figure 4.1 and Figure 4.2. The dICA approach consistently outperforms all its competitors when $p > N^*$ and tends to underestimate N^* when $p = N^*$ and n is small. In the case where $p = N^*$, it is likely that dICA overestimates the diagonal entries of $\Sigma_{\check{\mathbf{E}}}$. As a result, the eigenvalues of $\Sigma_{\check{\mathbf{X}}} - \widehat{\Sigma}_{\check{\mathbf{E}}}$ are underestimated, leading to an underestimation of N^*. However, as the sample size increases, the estimation of $\Sigma_{\check{\mathbf{E}}}$ becomes more accurate, resulting in a significant improvement in the performance of dICA. On the other hand, PCA generally performs the worst and tends to overestimate N^* in most cases. This is because PCA uses a fixed variance threshold and fails to consider the random nature of the percentage variance retained by components. In our simulation settings, the level of noise can be nearly as strong as the underlying components. When using PCA to retain 90% or 99% of the variance, it becomes apparent, as depicted in Figure 4.2, that N^* is often overestimated. It should be noted that $\mathrm{PCA99}$ has the opportunity to obtain accurate results when $p = N^*$ by retaining all the components. The asymptotic test requires a much larger sample size compared to the simulation dimension for its validity. This test tends to underestimate N^* when n is small and p is large, but its performance steadily improves as n increases. Overall, dICA is the most reliable method for estimating N^*.

In terms of assessing the accuracy in estimating $\mathbf{A}^*$, this study employs the following

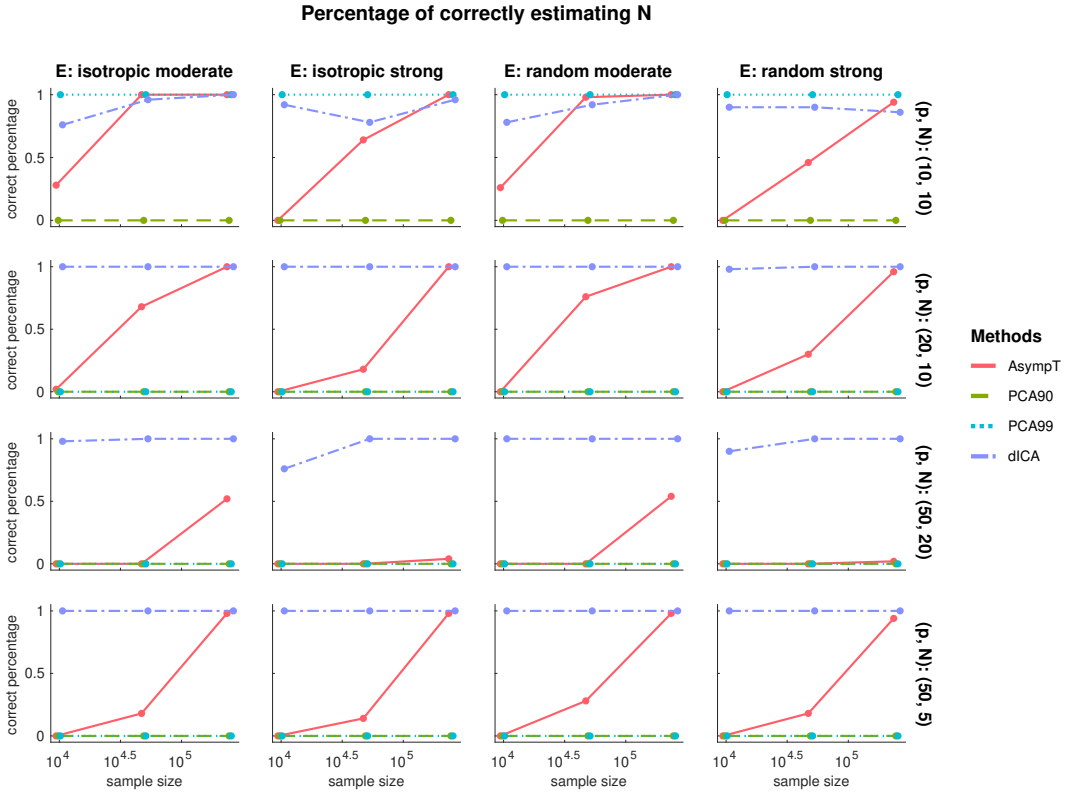

Figure 4.1: Percentage of correctly estimating N^* in 50 simulations

metric: Generalized Minimum Distance Index (GMDI). This metric, unlike its predecessors, is apt for non-square mixing matrices (Ilmonen et al., 2010; Lietzén et al., 2020). This metric quantifies the sign- and permutation-invariant discrepancy between $\widehat{\widetilde{\mathbf{W}}}_{\text{sample}}\mathbf{A}^*$ and $(\mathbf{I}_{N^*}, \mathbf{0})^\top$ as defined below:

$$\text{GMDI}(\widehat{\widetilde{\mathbf{W}}}_{\text{sample}}\mathbf{A}^*) = \sqrt{\frac{N^* - \max_{\mathbf{P}\in\mathcal{P}^{N^*}}\{\text{trace}(\mathbf{P}\overline{\mathbf{M}})\}}{N^*}},$$

where $\mathbf{M} = \widehat{\widetilde{\mathbf{W}}}_{\text{sample}}\mathbf{A}^* = (M_{j,k})$, $\overline{\mathbf{M}} = (\overline{M}_{j,k})$ with $\overline{M}_{j,k} = |M_{j,k}|^2 / \sum_h |M_{j,h}|^2$, and $\mathcal{P}^{N^*}$ represents the set of all possible $N^* \times N^*$ permutation matrices. The GMDI metric ranges from 0 to 1 (Lietzén et al., 2020), where $\text{GMDI} = 0$ indicates that $\widehat{\widetilde{\mathbf{W}}}_{\text{sample}}\mathbf{A}^* = (\mathbf{I}_{N^*}, \mathbf{0})^\top$ up to a sign flip, permutation alteration, and re-scaling, implying a perfect signal separation.

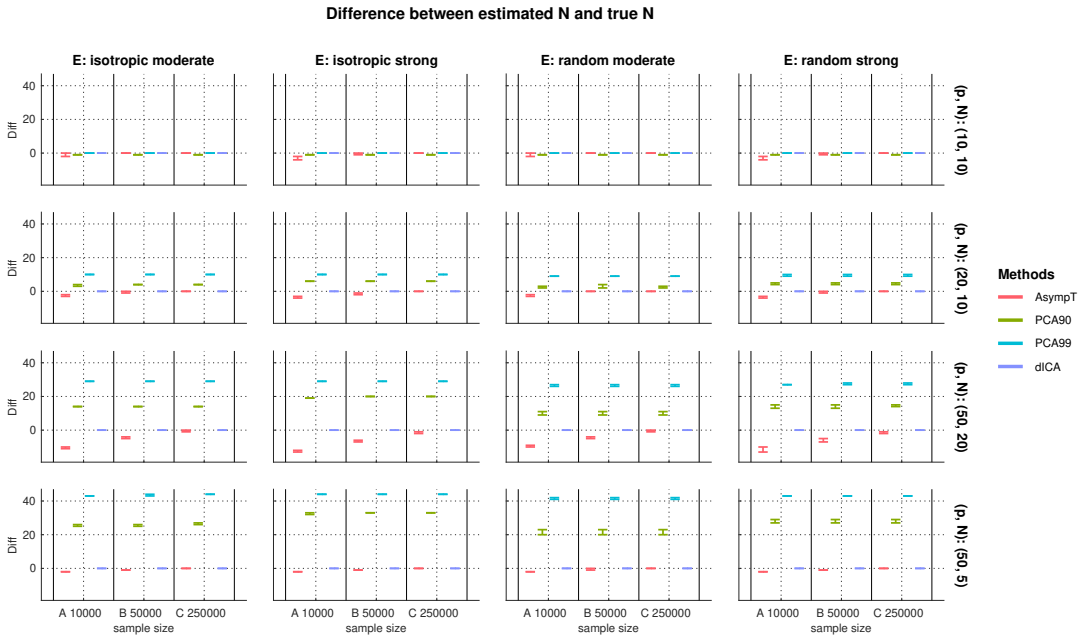

Figure 4.2: Quantile barplot of $\widehat{N} - N^*$ under Gaussian noise for different methods

Conversely, a GMDI value of 1 implies a completely unsuccessful signal separation.

The results under Gaussian noise are depicted in Figure 4.3. The figure is generated using the "gramm" package in "Matlab" (Morel, 2018). Our findings consistently demonstrate that dICA outperforms its competitors across various configurations when $p > N^*$. When $p = N^*$, dICA exhibits a larger variance compared to its competitors under strong noise but is able to claim its superior performance when a larger sample size is available. Notably, the errors of dICA in estimating $\mathbf{A}^*$ have steady decreasing trends with respect to sample size and hence suggest its potential consistency. When $p = N^*$, both PCA + FastICA and PCA + GIICA exhibit better performance compared to PCA + Infomax. Specifically, PCA + FastICA demonstrates better performance when dealing with small to moderate sample sizes, whereas PCA + GIICA performs better when confronted with large sample sizes. GIICA employs fourth-order statistics to eliminate bias induced by noise, rendering it more susceptible to outliers. Consequently, a larger sample size is typically required for GIICA to mitigate the effects caused by outliers. When $p > N^*$, all competitors suffer

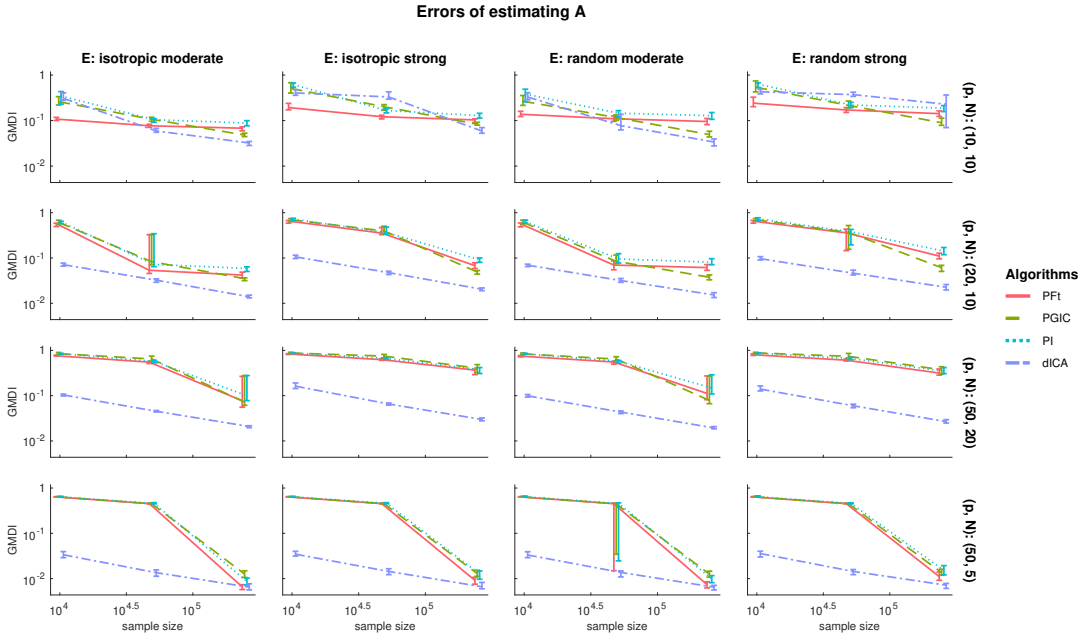

Figure 4.3: Errors of estimating $\mathbf{A}^*$ under Gaussian noise

considerably from the inaccurate estimation of N^*, resulting in unsatisfactory performance for small to moderate sample sizes. However, when the sample size n reaches 250000 and the asymptotic test yields an improved $\widehat{N}$, the performance of the competitors improves significantly. Nevertheless, dICA remains the top performer among all methods. In conclusion, dICA achieves the best performance in estimating $\mathbf{A}^*$.

In terms of assessing the accuracy in estimating $\mathbf{S}^*$, this paper employs the following metric: Mean Squared Error (MSE) to assess $\widehat{\mathbf{S}}$. $\mathrm{MSE}(\widehat{\mathbf{S}}, \mathbf{S}^*)$ can be expressed as $\|\widehat{\mathbf{S}} - \mathbf{S}^*\|_F^2/n$. dICA outperforms its competitors when $p > N^*$ and have comparable performance when $p = N^*$. It is worth noting that the errors in estimating $\mathbf{S}^*$ do not exhibit a consistent decreasing trend in dICA. This phenomenon is attributed to the persistence of noise-induced bias in $\widehat{\mathbf{S}}$, a bias that cannot be mitigated by increasing the sample size.

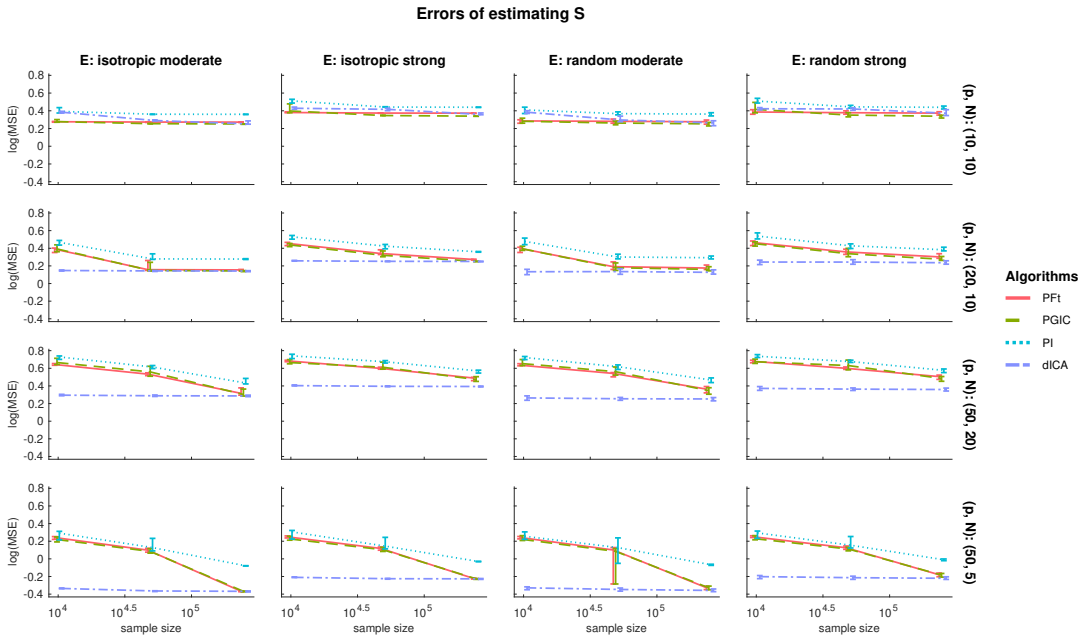

Figure 4.4: Errors of estimating $\mathbf{S}^*$ under Gaussian noise

4.2. Comparison with competitors under non-Gaussian noise

This section evaluates the efficacy of dICA under conditions when Gaussian noise assumption is violated. Specifically, the analysis focuses exclusively on scenarios characterized by random moderate noise in the context of $\Sigma_{\boldsymbol{E}}$. We evaluate the performance of these methods across various situations with following configurations:

- $\mathbf{A}^*$'s dimension $(p, N^*) \in \{(10, 10), (20, 10)\}$.
- $\mathbf{A}^*$ is generated by $\mathbf{U}\mathrm{diag}(\sigma_1, \ldots, \sigma_{N^*})\mathbf{V}$, where $\mathbf{U}$ and $\mathbf{V}$ are $p \times N^*$ and $N^* \times N^*$ randomly generated orthonormal/semi-orthogonal matrices, respectively. $\{\sigma_k\}_{k=1}^{N^*}$ are generated independently from the square root of a uniform distribution, specifically, from $\sqrt{U(2, N^*)}$.
- The investigation encompasses two distinct noise families, namely:
  - t-distribution with degree freedom coming from $\{5, 10, 20\}$.

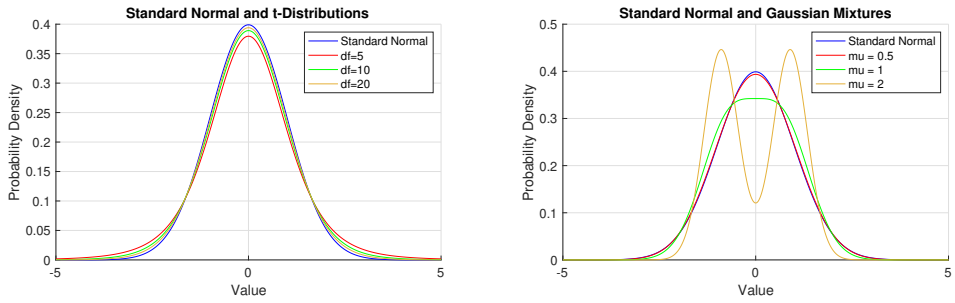

Figure 4.5: The plots of the probability density functions of t-distributions and Gaussian-mixtures. Gaussian-mixtures are standardized to show its sub-Gaussian nature

  - Gaussian mixture with density $0.5\mathcal{N}(-\mu, 1) + 0.5\mathcal{N}(\mu, 1)$ and $\mu \in \{0.5, 1, 2\}$.

  The plots of the noise distributions are in Figure 4.5

- Sample size $n = 50000$.
- The elements in $\mathbf{S}^*$ are generated from 4 different distribution settings:
  - **Laplace:** All the elements in $\mathbf{S}^*$ are generated from standard Laplacian distribution.
  - **mixGaussian:** All the elements in $\mathbf{S}^*$ are generated from Gaussian mixture $0.5\mathcal{N}(-1, 1) + 0.5\mathcal{N}(1, 1)$.
  - **Uniform:** All the elements in $\mathbf{S}^*$ are generated from Uniform distribution $U(-1, 1)$.
  - **Triple:** Elements in $\mathbf{S}^*$ are generated evenly from Laplacian, Uniform, and Gaussian mixture distribution.
- 50 simulations are conducted for each situation.

The results of the accuracy in **estimating** N^* are summarized in Figure 4.6 and Figure 4.7.

When $p > N^*$, dICA still outperforms the other methods. It shows superior performance when the noise deviates slightly or moderately from the Gaussian distribution. However, dICA tends to overestimate N^* when the noise becomes even more non-Gaussian than the sources. This further supports the claim that dICA selects components based on their non-Gaussianity. On the other hand, when $p = N^*$, dICA is outperformed by PCA with 99%

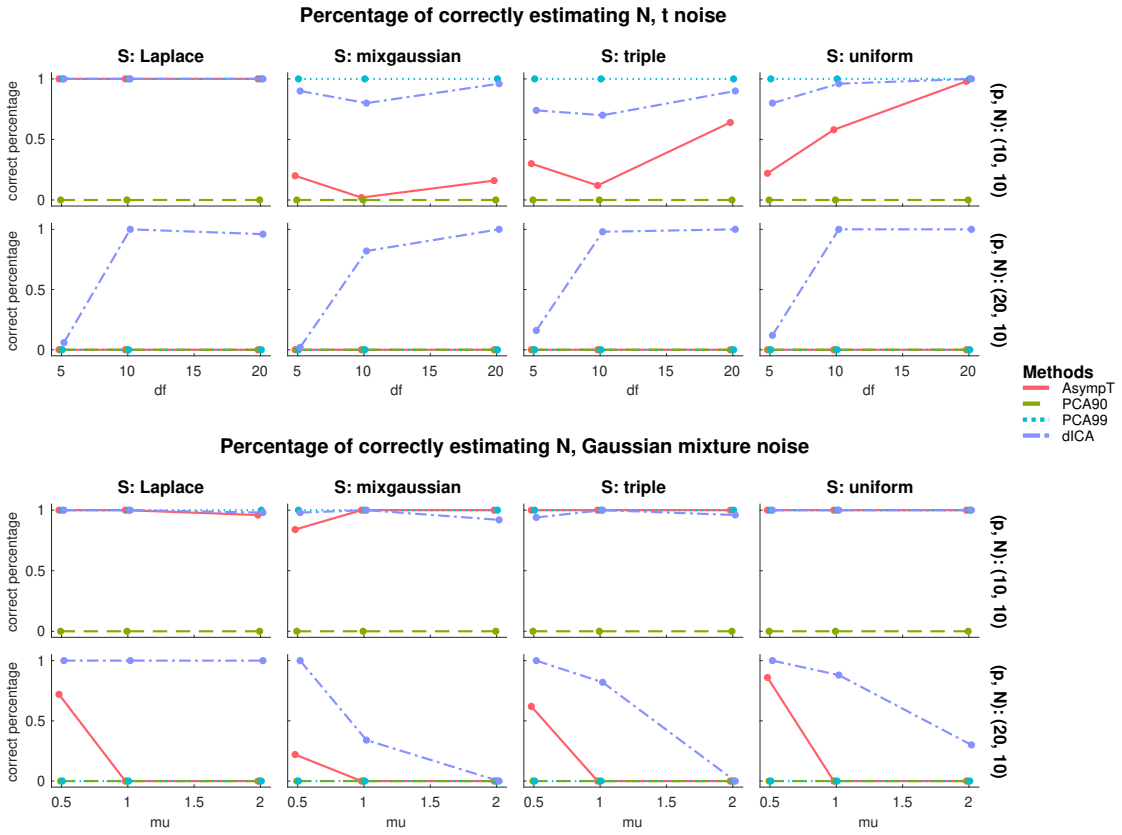

Figure 4.6: Percentage of correctly estimating N^* under non-Gaussian noise in 50 simulations

variance retained, and it occasionally underestimates N^*. PCA relies on a fixed variance threshold and only works when $p = N^*$. The Asymptotic test also performs poorly. It can only identify $\widehat{N}$ when the noise deviates slightly from a Gaussian distribution, but it tends to significantly overestimate $\widehat{N}$ when the noise becomes more non-Gaussian. In summary, dICA still demonstrates greater robustness in the presence of non-Gaussian noise, making it the overall superior method.

The accuracy results for estimating $\mathbf{A}^*$ are summarized in Figure 4.8. When $p > N^*$ and the noise deviates slightly or moderately from Gaussian, dICA exhibits the best performance. However, when the noise becomes more non-Gaussian than the latent components or $p = N^*$, dICA is outperformed by FastICA. This underperformance of dICA can be attributed to its reliance on non-Gaussianity to estimate N^*, making it challenging for dICA to distinguish

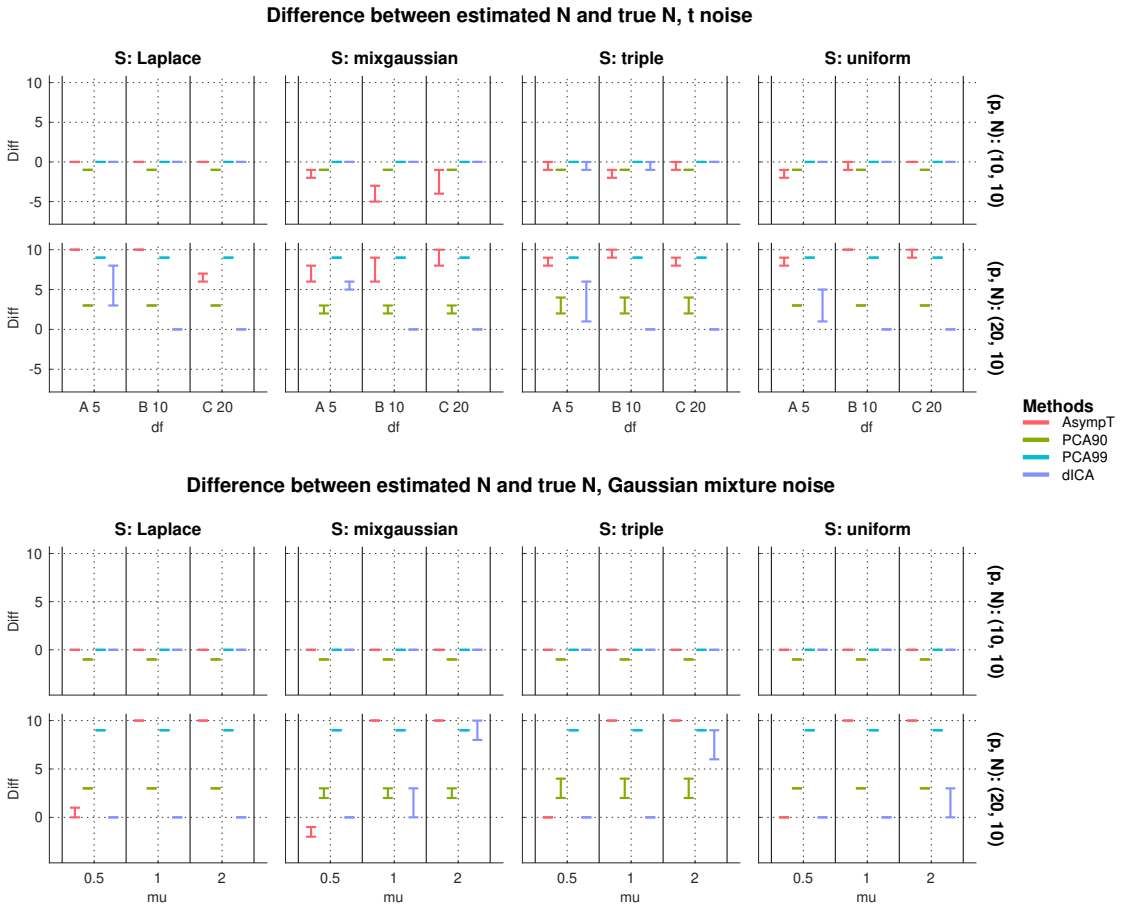

Figure 4.7: Quantile barplot of $\widehat{N} - N^*$ under Gaussian noise for different methods

between noise and components in the presence of highly non-Gaussian noise. It is worth noting that all competing methods are provided with the correct N^* and do not need to estimate it. Furthermore, this section demonstrates that dICA requires a larger sample size to showcase its advantages when $p = N^*$, and $n = 50000$ is not a sufficiently large sample size. Among the competing algorithms, FastICA and Infomax show comparable performance, while GIICA performs the worst.

In terms of estimating $\mathbf{S}^*$, the results are displayed in Figure 4.9. dICA remains the best option when $p > N^*$ and the noise moderately deviates from the Gaussian distribution. When $p = N^*$, GIICA and FastICA perform the best, with FastICA exhibiting a more stable performance. Infomax generally performs the worst.

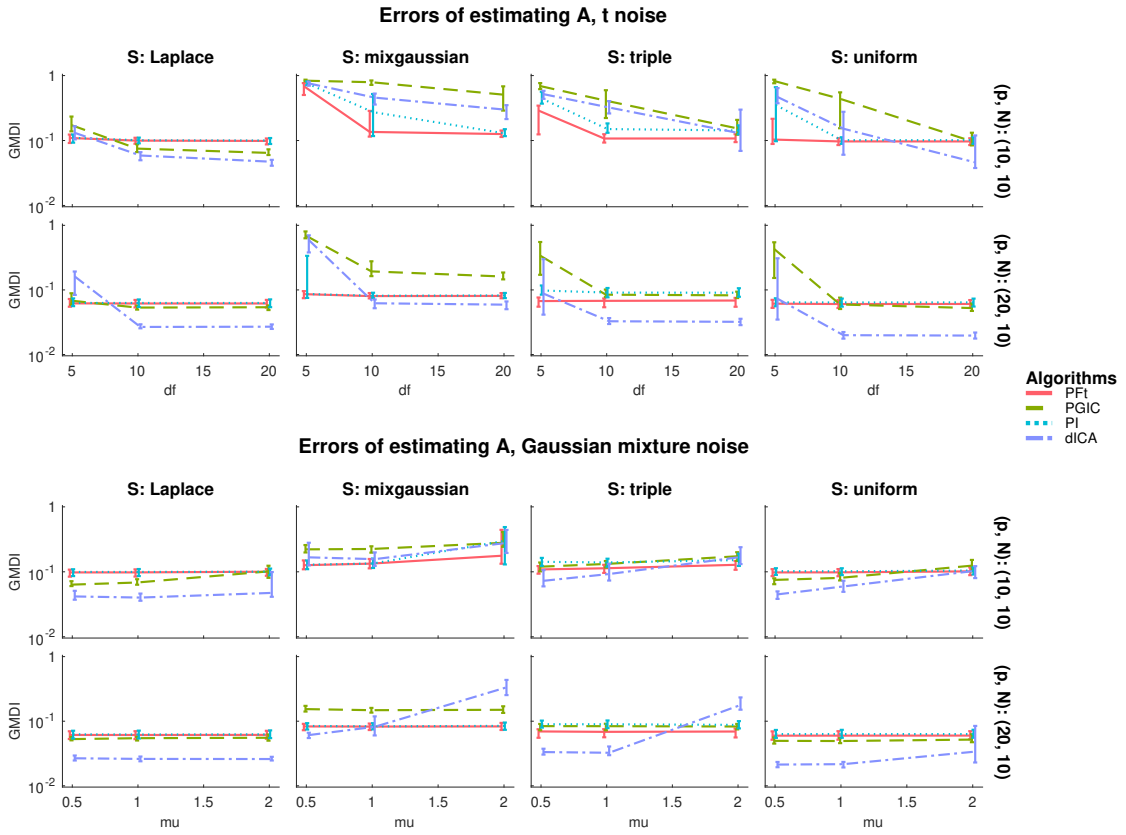

Figure 4.8: Errors of estimating $\mathbf{A}^*$ under non-Gaussian noise

In summary, dICA can still claim its advantage under non-Gaussian noise in situations where $p > N^*$ and the noise does not severely deviate from the Gaussian distribution. However, its performance is not promising when $p = N^*$ or when the noise significantly deviates from Gaussian. The reason dICA does not perform well when $p = N^*$ is that dICA requires a larger sample size to demonstrate its superiority, and the sample size of $n = 50000$ used in Section 4.2 is not large enough. Additionally, dICA does not perform well when the noise term deviates significantly from Gaussian, which is understandable. dICA has two advantages: (i) it simultaneously conducts dimension reduction and latent variable estimation, thus obtaining a more accurate estimation of N^*; (ii) it employs the bias-removal technique to account for the noise effect originating from $\mathbf{X}$. However, both of these advantages are achieved under the assumption that the noise follows a Gaussian

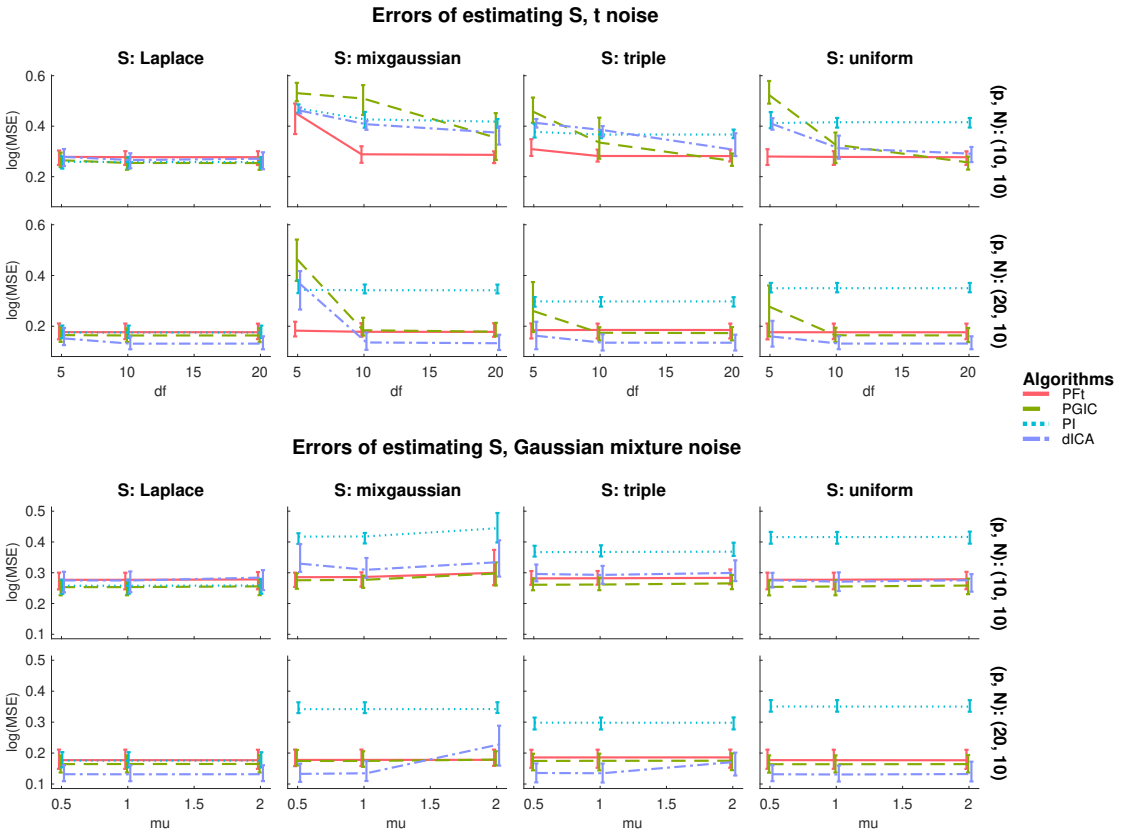

Figure 4.9: Errors of estimating $\mathbf{S}^*$ under non-Gaussian noise

distribution. When all the competing methods are provided with the true N^*, dICA's first advantage is leveled. Furthermore, if the noise deviates severely from the Gaussian distribution, the bias-removal technique will not be able to demonstrate its advantage either. These reasons ultimately lead to the finding that dICA is outperformed when the noise term deviates significantly from the Gaussian distribution.

To summarize, dICA proves to be a promising approach, especially in the presence of Gaussian or quasi-Gaussian noise. Even when the noise deviates from the Gaussian archetype, dICA maintains its performance, particularly when the noise levels are moderate. This highlights its robustness and resilience, making it suitable for situations with slight to moderate deviations from the Gaussian noise assumption. This adaptability ensures that dICA is a versatile tool, providing consistent results across a diverse range of noise profiles.

Chapter 5

Application on EEG datasets

ICA has been widely utilized in Electroencephalogram (EEG) for analyzing brain source activity. One collection of datasets located at http://sccn.ucsd.edu/eeglab/BSSComparison/ (EEGLAB, 2023) were collected from a visual working memory task conducted with 14 participants (7 men and 7 women) aged between 20 and 40 years. The task required participants to memorize a series of letters and determine if subsequent letters were part of the initial set by pressing a button. Auditory feedback was provided to indicate the correctness of their responses. Each participant performed 100-150 trials. EEG data was recorded using 71 channels, including 69 on the scalp and 2 around the eyes. The recordings were sampled at a rate of 250 Hz. Each electrode for each subject comprised approximately 300,000 time points. Since a definitive ground truth was not available, the evaluation of algorithm performance relied on various criteria. This study employs two of these 14 datasets. Each dataset in this study was analyzed using the EEGLab package (Delorme and Makeig, 2004). These datasets first undergo a re-referencing preprocessing step (Delorme et al., 2012). The results of the estimated mixing matrix ($\widehat{\mathbf{A}}$) were visualized through topographic scalp maps. For an estimated mixing matrix $\widehat{\mathbf{A}} \in \mathbb{R}^{p \times \widehat{N}}$, $\widehat{N}$ different scalp maps were obtained, corresponding to the $\widehat{N}$ columns in $\widehat{\mathbf{A}}$. The i, j-th entry in $\widehat{\mathbf{A}}$ represented the projection weights of the j-th IC on the i-th electrode. Therefore, each column of $\widehat{\mathbf{A}}$ indicated the contribution of a specific component to a particular EEG electrode.To generate the scalp

map associated with the j-th component, the EEGLAB software utilizes the j-th column from the $\widehat{\mathbf{A}}$ matrix. This column contains values that indicate the impact of the component on each electrode. Subsequently, these values are plotted on a head-shaped outline, with each value being positioned at the location of its respective electrode on the scalp. The plotting procedure employs color coding to represent the magnitude and polarity of the values: positive values are depicted in shades of red, while negative values are represented in shades of blue. Furthermore, areas on the scalp without an electrode are filled with values using interpolation, thereby creating a continuous representation of the electrical field on the scalp surface. Figure 5.2 and Figure 5.3 display examples of scalp maps. This study will evaluate the quality of $\widehat{\mathbf{A}}$ by examining the scalp maps generated by the columns in $\widehat{\mathbf{A}}$.

The $\widehat{N}$ retrieved components can be classified into two distinct groups according to the dipolarity of their scalp maps resulting in dipolar and non-dipolar groups. Figure 5.1 showcases some instances of each group. Dipolar components are physiologically more plausible (Delorme et al., 2012). In other words, the components are anticipated to possess a scalp map that can be characterized by a singular dipole model. This model postulates that the observed electrical activity can be sufficiently represented by the activity of a single point source ("dipole") with specific attributes, including location, orientation, and magnitude. For each selected column of $\widehat{\mathbf{A}}$, the objective is to identify a dipole whose projected activity onto the scalp exhibits the closest resemblance to the scalp map of the column. To achieve this, a dipole fitting process is employed, which involves (i) employing a head model to simulate the flow of electrical currents through the various tissues of the head, and (ii) iteratively adjusting the parameters of the dipole (location, orientation, and magnitude) to minimize the discrepancy between the predicted scalp map of the dipole and the actual scalp map of the ICA component. The dipole model resulting from this process, under its optimal parameters, is referred to as the best-fitting singular dipole model. A column is deemed to correspond to a dipolar component if the residual variance between its scalp map and the scalp map predicted by the best-fitting singular dipole model is no greater than 10% (Delorme et al., 2012).

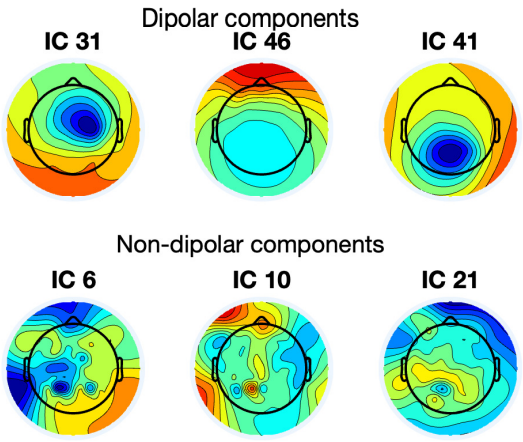

Figure 5.1: Examples of dipolar ICs and non-dipolar ICs

The dipolar components can be subdivided further into brain ICs and non-brain ICs. Brain ICs refer to the components that capture the participants' brain activity, while non-brain ICs encompass muscle ICs, eye ICs, heart ICs, and other ICs (Pion-Tonachini et al., 2019). Among these categories, research interest primarily revolves around brain ICs. To determine the qualification of an IC as a brain source, this study employs the ICLabel plugin within the EEGLab toolbox, developed by Pion-Tonachini et al. (2019). ICLabel utilizes a machine learning model that has been trained on a large number of manually labeled ICs. When applying ICLabel to the ICs obtained from ICA algorithms, it initially extracts features from each component. These features can include scalp maps (columns in $\widehat{\mathbf{A}}$) and spectral power density (rows in $\widehat{\mathbf{S}}$) of the ICs. The extracted features are then input into the trained machine learning model, which assigns each component to one of the predefined categories. The output of ICLabel consists of a set of probabilities for each IC, indicating the likelihood of belonging to each category. If a component exceeds a likelihood of 90% (default threshold) in representing brain signals, it is regarded as a brain IC. In Figure 5.1, IC 46 is an example of an eye IC while IC 31 and IC 41 are examples of Brain ICs.

This study employs three evaluation criteria to assess the performance of algorithms. The first criterion pertains to the number of dipolar ICs. As dipolar components indicate

physiological plausibility, a higher count of dipolar columns in $\widehat{\mathbf{A}}$ corresponds to a superior estimation quality of $\widehat{\mathbf{A}}$. The second criterion focuses on the number of brain ICs. Given the research emphasis on brain ICs, the method's capability to detect a larger quantity of brain ICs reflects a higher quality. Another means of measurement is qualitative. To undertake this qualitative evaluation, we examine the scalp maps of the components (Pfister et al., 2019). This approach demonstrates that the components identified by dICA exhibit neither randomness nor implausibility for EEG recordings and are qualitatively consistent with the findings obtained from competing methods.

In the previous simulations, PCA $+\mathrm{FastICA}$ overall performs the second best in various noise settings. We will compare dICA with PCA+$\mathrm{FastICA}$ in this EEG data application. The typical variance retention in the PCA step fluctuates between 95% and 99%. (Guo et al., 2020; Ofner et al., 2019; Artoni et al., 2019). Notably, PCA with a 99% variance retention has exhibited better performance in these EEG datasets; therefore, this study will adhere to retaining 99% of the variance in the PCA step.

The study proceeds to evaluate the two algorithms on the two distinct datasets derived from different participants. Below is the table summarizing the results for the number of dipolar ICS and the number of brain ICs and $\widehat{N}$. The dICA algorithm is implemented with three different parameter settings. Among these, the optimal choice is $(\rho_1, c_1) = (1, 1.6)$, while $(\rho_1, c_1) = (100, 120)$ and $(\rho_1, c_1) = (100, 160)$ are specifically designed to encourage sparsity. However, the latter two choices are not ideal, as the large values of ρ_1 and c_1 have caused some overshrinking and distortion in the results. The reason for choosing such large values is to test the robustness of dICA to parameter selection. Table 5.1 demonstrates that as ρ_1 and c_1 increase, $\widehat{N}$ decreases significantly. This aligns with the claim made in Chapter 3 that larger ρ_1 and c_1 result in a sparser solution. As $\widehat{N}$ decreases, the number of dipolar ICs also decreases due to overshrinking. However, two observations are worth highlighting. First, the number of brain ICs remains consistent regardless of the choice of ρ_1 and c_1, indicating the robustness of dICA in detecting the most important components. Second, even with a smaller $\widehat{N}$, dICA is still able to detect more dipolar ICs and brain ICs

than the $\mathrm{PCA} + \mathrm{FastICA}$ approach. This further demonstrates the effectiveness of dICA in avoiding information loss caused by PCA, particularly in low signal-to-noise ratio scenarios such as EEG. When comparing $\mathrm{dICA}(\rho_1 = 1, c_1 = 1.6)$ with $\mathrm{PCA} + \mathrm{FastICA}$, it is observed that dICA consistently identifies approximately 50% more near-dipolar components. Additionally, dICA surpasses $\mathrm{PCA} + \mathrm{FastICA}$ in identifying 40%-50% more brain sources, thus demonstrating its superior performance in this aspect. It is worth noting that the number of dipolar ICs and $\widehat{N}$ identified by dICA remains consistent across various datasets, which underscores the reliability of this method.

Data set		Method			
		dICA $\rho_1 = 1$ $c_1 = 1.6$	dICA $\rho_1 = 100$ $c_1 = 120$	dICA $\rho_1 = 100$ $c_1 = 160$	PCA + FastICA
ds80	$\widehat{N}$	59	47	29	40
	# of dipolar ICs	27	23	19	16
	# of Brain ICs	13	13	12	9
km81	$\widehat{N}$	56	42	40	23
	# of dipolar ICs	30	25	22	18
	# of Brain ICs	17	16	16	12

Table 5.1: Number of Dipolar ICs and Brain ICs

The analysis then progresses to examining the topographic map similarities between the $\mathrm{PCA} + \mathrm{FastICA}$ and dICA methods. This is achieved by evaluating the brain ICs identified by $\mathrm{PCA} + \mathrm{FastICA}$ and assessing whether dICA identifies similar ICs. The findings, summarized in Figures 5.2 and 5.3, reveal that almost all brain ICs pinpointed by $\mathrm{PCA} + \mathrm{FastICA}$ are also detected by dICA. Moreover, dICA succeeds in identifying additional brain ICs, thereby facilitating a more comprehensive analysis of the EEG signals.

In summary, dICA demonstrates the ability to identify a consistent number of components and brain ICs. Furthermore, dICA exhibits the capability to detect a significantly greater number of physiologically plausible components and brain ICs than $\mathrm{FastICA}$. Consequently, dICA emerges as a superior method to $\mathrm{FastICA}$ for these EEG datasets.

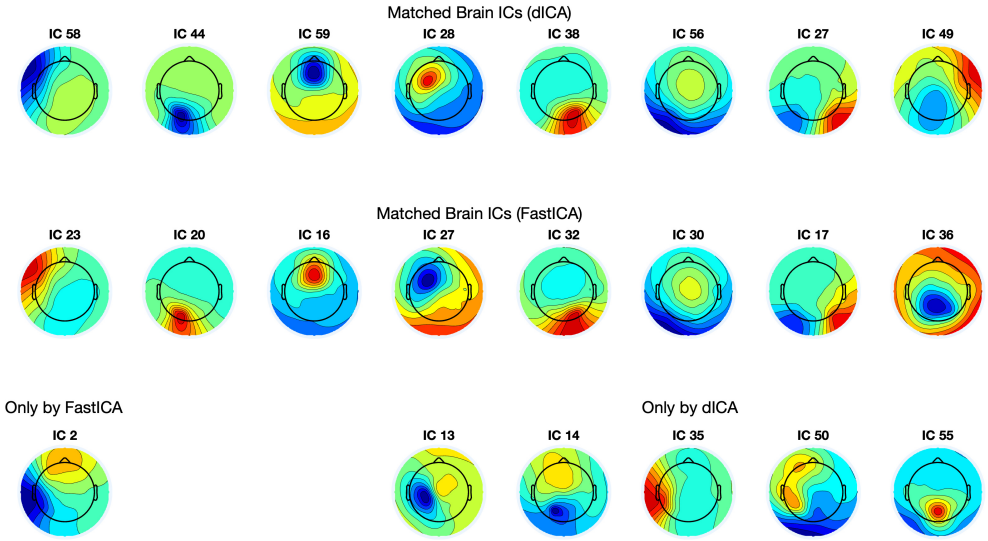

Figure 5.2: **Brain ICs scalp maps of data set ds80**: the first and second row consists of matched Brain ICs detected by dICA and FastICA respectively. The third row displays the unmatched Brain ICs that only detected either by FastICA or dICA

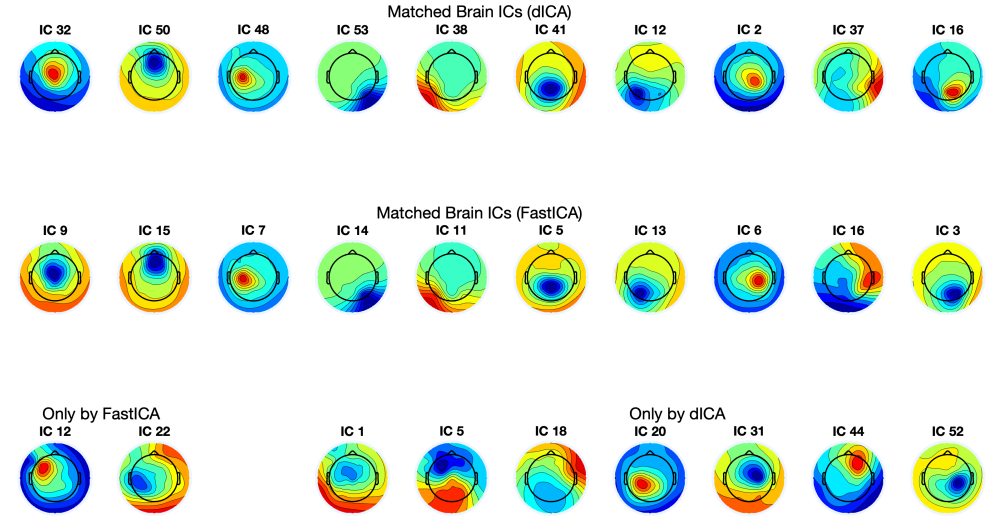

Figure 5.3: **Brain ICs scalp maps of data set km81**: the first and second row consists of matched Brain ICs detected by dICA and FastICA respectively. The third row displays the unmatched Brain ICs that only detected either by FastICA or dICA

Chapter 6

Discussion

We present a novel algorithm called dICA for under-complete noisy Independent Component Analysis, capable of automatically determining the appropriate number of components while estimating the mixing matrix. Our approach incorporates group regularization to shrink nuisance rows in the demixing matrix, ensuring accurate results. To enable the automatic selection of the optimal number of components, we introduce a new constraint that parallels the classical constraint in equation (1.3), but does not rely on knowing N^*. By formulating these ideas as a constraint optimization problem, we develop an efficient fixed-point algorithm to solve it. Theoretical analysis confirms that our algorithm achieves at least a linear convergence rate. The dICA approach outperforms competitors when applied to both simulated and real EEG dataset, provided that the under-complete assumption is satisfied.

Several issues need further exploration. First, in some application scenarios, we may have information regarding the range of N^*. How to incorporate that information with our automatic selection procedure is worth further investigation. Second, like $\mathrm{FastICA}$, dICA can be trapped in a saddle point, preventing it from converging to the correct local optimizer. When the algorithm is trapped in a saddle point, it also has the risk of incorrectly estimating N^*, which further deteriorates the performance. Tichavsky et al. (2006) proposes a way to check whether $\mathrm{FastICA}$ is trapped in a saddle point in the noise-free case. Extending

this saddle point checking approach in the noisy ICA model is a potential future direction of research. Furthermore, the choice of function $G(\cdot)$ in $\mathcal{J}(\cdot)$ is uniform over all latent components. Hastie and Tibshirani (2002) have shown that choosing $G(\cdot)$ dynamically can result in better performance in the noiseless case. Research on extending this approach to the noisy case remains open. Moreover, this thesis primarily focused on the development of point estimators for $\mathbf{A}^*$ and $\{\boldsymbol{S}^*(t_i)\}$. It would be intriguing to investigate the uncertainties associated with these point estimators and formulate statistical inference methods for them. Furthermore, it is worth noting that this thesis assumes $\{\boldsymbol{X}(t_i)\}$ to be i.i.d. observations, which may not accurately capture the time dependency present in certain types of datasets. For instance, time series models may be more suitable for accurately representing EEG data. Although this i.i.d. assumption might not impact the validity of the point estimators (Risk et al., 2019), it is important to further explore this assumption when quantifying the uncertainties associated with these point estimators.

Appendix A

Proof of Main Results

A.1. Notations and symbols

The superscript $\top$ denotes transpose. Let $\mathbf{I}$ or $\mathbf{I}_p$ denote a square (or $p \times p$) identity matrix, and $\boldsymbol{e}_j$ is the jth column vector of an identity matrix, whereas $\mathbf{I}_{n\times p}$ denotes the $n \times p$ matrix with ones along the diagonal. Let $\mathbf{0}$ denote a vector of zeros, $\mathbf{0}_p$ denote a square (or $p \times p$) zero matrix, whereas $\mathbf{0}_{n\times p}$ denotes the $n \times p$ zero matrix. For a matrix M, $M_{j,k}$ denotes the entry at the jth row and kth column. The trace of a square matrix $M \in \mathbb{R}^{p\times p}$ is denoted by $\mathrm{trace}(M) = \sum_{j=1}^{p} M_{j,j}$. The Frobenius norm of a matrix M is defined as $\|M\|_F = \{\mathrm{trace}(M^\top M)\}^{1/2}$. The notation $\mathrm{col}(M)$ represents the column space of the matrix M. The notation $M \succ 0$ denotes that a square matrix M is positive definite.

For matrices $A = (A_{i,j})$ and $B = (B_{i,j})$, the inner product of A and B is denoted by $\langle A, B\rangle = \sum_{i,j} A_{i,j}B_{i,j} = \mathrm{trace}(A^\top B)$, and the Hadamard multiplication is denoted by $A \circ B = (A_{i,j}B_{i,j})$. The Khatri-Rao product $A \odot B$ of two matrices $A \in \mathbb{R}^{n\times k}$ and

$B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k) \in \mathbb{R}^{m \times k}$, is an $nm \times k$ matrix of the form:

$$A \odot B \coloneqq \begin{pmatrix} A_{1,1}\boldsymbol{b}_1 & A_{1,2}\boldsymbol{b}_2 & \ldots & A_{1,k}\boldsymbol{b}_k \\ A_{2,1}\boldsymbol{b}_1 & A_{2,2}\boldsymbol{b}_2 & \ldots & A_{2,k}\boldsymbol{b}_k \\ \vdots & \vdots & \ldots & \vdots \\ A_{n,1}\boldsymbol{b}_1 & A_{n,2}\boldsymbol{b}_2 & \ldots & A_{n,k}\boldsymbol{b}_k \end{pmatrix}.$$

The outer product between two vectors $\boldsymbol{a} \in \mathbb{R}^{p_1}$ and $\boldsymbol{b} \in \mathbb{R}^{p_2}$ is given by

$$\boldsymbol{a} \otimes \boldsymbol{b} \coloneqq \boldsymbol{a}\boldsymbol{b}^\top = (a_i b_j)_{1 \le i \le p_1;\, 1 \le j \le p_2}.$$

The outer product among four vectors $\boldsymbol{a} \in \mathbb{R}^{p_1}$, $\boldsymbol{b} \in \mathbb{R}^{p_2}$, $\boldsymbol{c} \in \mathbb{R}^{p_3}$, and $\boldsymbol{d} \in \mathbb{R}^{p_4}$ is given by

$$\boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c} \otimes \boldsymbol{d} \coloneqq (a_i b_j c_k d_l)_{1 \le i \le p_1;\, 1 \le j \le p_2; 1 \le k \le p_3; 1 \le l \le p_4}.$$

For random quantities V_1 and V_2, $V_1 V_2$ denotes that V_1 and V_2 are independent. The kurtosis of a univariate random variable S is defined as

$$\kappa_S = \mathrm{E}\{(S - \mathrm{E}(S))^4\} - 3\{\mathrm{var}(S)\}^2.$$

A.2. Proof of Lemmas in Section 2.2

Proof of Lemma 1**.** Using integration by parts, we have:

$$\begin{aligned} \int_0^x t\varphi(t)\mathrm{d}t &= x\varphi^{(-1)}(x) - \int_0^x \varphi^{(-1)}(t)\mathrm{d}t \\ &= x\varphi^{(-1)}(x) - \varphi^{(-2)}(x) + \varphi^{(-2)}(0) \\ &= x\varphi^{(-1)}(x) - \varphi^{(-2)}(x). \end{aligned}$$

The above equation, together with a direct evaluation $\int_0^x t\varphi(t)\mathrm{d}t = -\varphi(x) + \varphi(0)$, shows equation (2.4). ■

Proof of Lemma 2. Without loss of generality, we only prove the case when $Z \sim \mathcal{N}(0, \sigma^2)$. The results can be generalized to the case when $Z \sim \mathcal{N}(\mu, \sigma^2)$ by using the equation below:

$$\mathrm{E}\{\varphi_d^{(-k)}(V+Z)\} = \mathrm{E}\{\varphi_d^{(-k)}((V+\mu)+(Z-\mu))\}, \quad k = 0, 1, 2.$$

We start by proving $\mathrm{E}\{\varphi_c(V)\} = \mathrm{E}\{\varphi_d(V+Z)\}$:

$$\begin{aligned}\mathrm{E}\{\varphi_d(V+Z)\} &= \mathrm{E}\{\mathrm{E}\{\varphi_d(V+Z)|V\}\} = \mathrm{E}\{\int \varphi_d(V+z)\varphi_\sigma(z)\mathrm{d}z \mid V\} \\ &= \mathrm{E}\{\varphi_c(V)\}. \end{aligned} \tag{A.1}$$

Now we move on to prove $\mathrm{E}\{\varphi_c^{(-1)}(V)\} = \mathrm{E}\{\varphi_d^{(-1)}(V+Z)\}$. Given that the derivation in (A.1) does not rely on the distribution of V, for any $t \in \mathbb{R}$, the following equations always hold:

$$\begin{aligned}\mathrm{E}\{\varphi_c(V+t)\} &= \mathrm{E}\{\varphi_d(V+Z+t)\}, \\ \int_{-\infty}^{\theta} \mathrm{E}\{\varphi_c(V+t)\}\mathrm{d}t &= \int_{-\infty}^{\theta} \mathrm{E}\{\varphi_d(V+Z+t)\}\mathrm{d}t, \quad \theta \in \mathbb{R}.\end{aligned}$$

By the Fubini Theorem,

$$\int_{-\infty}^{\theta} \mathrm{E}\{\varphi_c(V+t)\}\mathrm{d}t = \mathrm{E}\Big\{\int_{-\infty}^{\theta} \varphi_c(V+t)\mathrm{d}t\Big\} = \mathrm{E}\{\varphi_c^{(-1)}(V+\theta) + 1/2\}. \tag{A.2}$$

Following a similar argument to (A.2), we have:

$$\int_{-\infty}^{\theta} \mathrm{E}\{\varphi_d(V+Z+t)\}\mathrm{d}t = \mathrm{E}\{\varphi_d^{(-1)}(V+Z+\theta) + 1/2\}. \tag{A.3}$$

Combining (A.2) and (A.3) and setting $\theta = 0$, we have:

$$\mathrm{E}\{\varphi_c^{(-1)}(V)\} = \mathrm{E}\{\varphi_d^{(-1)}(V+Z)\}.$$

Finally, we prove $\mathrm{E}\{\varphi_c^{(-2)}(V)\} = \mathrm{E}\{\varphi_d^{(-2)}(V+Z)\} + \varphi(0)(d-c)$. Again, since all the

above derivations do not depend on V, we have:

$$\int_{\theta_2}^{\theta_1} \mathrm{E}\{\varphi_c^{(-1)}(V+t)\}\mathrm{d}t = \int_{\theta_2}^{\theta_1} \mathrm{E}\{\varphi_d^{(-1)}(V+Z+t)\}\mathrm{d}t, \quad \theta_1, \theta_2 \in \mathbb{R}. \tag{A.4}$$

Applying the Fubini's Theorem to (A.4), we obtain,

$$\mathrm{E}\{\varphi_c^{(-2)}(V+\theta_1) - \varphi_c^{(-2)}(V+\theta_2)\} = \mathrm{E}\{\varphi_d^{(-2)}(V+Z+\theta_1) - \varphi_d^{(-2)}(V+Z+\theta_2)\}. \tag{A.5}$$

Setting $\theta_1 = 0$ in (A.5) and after some algebraic rearrangement, we have the following:

$$\mathrm{E}\{\varphi_c^{(-2)}(V)\} - \mathrm{E}\{\varphi_d^{(-2)}(V+Z)\} = \mathrm{E}\{\varphi_c^{(-2)}(V+\theta_2)\} - \mathrm{E}\{\varphi_d^{(-2)}(V+Z+\theta_2)\}.$$

The derivation of (A.5) does not rely on the distribution of V, hence for any θ_2,

$$\mathrm{E}\{\varphi_c^{(-2)}(0)\} - \mathrm{E}\{\varphi_d^{(-2)}(Z)\} = \mathrm{E}\{\varphi_c^{(-2)}(\theta_2)\} - \mathrm{E}\{\varphi_d^{(-2)}(Z+\theta_2)\}.$$

This implies that

$$\begin{aligned}
\mathrm{E}\{\varphi_c^{(-2)}(0)\} - \mathrm{E}\{\varphi_d^{(-2)}(Z)\} &= \mathrm{E}\{\mathrm{E}\{\varphi_c^{(-2)}(0) - \varphi_d^{(-2)}(Z) \mid V\}\} \\
&= \mathrm{E}\{\mathrm{E}\{\varphi_c^{(-2)}(V) - \varphi_d^{(-2)}(Z+V) \mid V\}\} \\
&= \mathrm{E}\{\varphi_c^{(-2)}(V)\} - \mathrm{E}\{\varphi_d^{(-2)}(V+Z)\}.
\end{aligned}$$

Hence, we have:

$$\mathrm{E}\{\varphi_c^{(-2)}(V)\} - \mathrm{E}\{\varphi_d^{(-2)}(V+Z)\} = \text{Constant} = -\mathrm{E}\{\varphi_d^{(-2)}(Z)\} = \varphi(0)(d-c).$$

This completes the proof. ■

A.3. Proof of Lemmas in Section 2.3

Proof of Lemma 3. The definition of the fourth-order cumulant tensor $\mathcal{C}_{\boldsymbol{X}}^{(4)} \in \mathbb{R}^{p\times p\times p\times p}$ is given by:

$$\mathcal{C}_{\boldsymbol{X}}^{(4)} \coloneqq (\mathrm{cum}(X_i, X_j, X_k, X_l))_{1\le i\le p; 1\le j\le p; 1\le k\le p; 1\le l\le p},$$

where

$$\begin{aligned}\mathrm{cum}(X_i, X_j, X_k, X_l) = \mathrm{E}(X_iX_jX_kX_l) - \mathrm{E}(X_iX_j)\mathrm{E}(X_kX_l)\\ -\mathrm{E}(X_iX_k)\mathrm{E}(X_jX_l) - \mathrm{E}(X_iX_l)\mathrm{E}(X_jX_k).\end{aligned}$$

Since $\boldsymbol{X}$ is observable, the fourth-order cumulant tensor $\mathcal{C}_{\boldsymbol{X}}^{(4)}$ can be calculated.

Using results from Podosinnikova et al. (2019) (see Appendix B.2 there for details), we have:

$$\mathcal{C}_{\boldsymbol{X}}^{(4)} = \sum_{k=1}^{N^*} \kappa_{S_k^*} \boldsymbol{a}_k^* \otimes \boldsymbol{a}_k^* \otimes \boldsymbol{a}_k^* \otimes \boldsymbol{a}_k^*.$$

The flattened version of $\mathcal{C}_{\boldsymbol{X}}^{(4)}$ is denoted by $\mathbf{C} \in \mathbb{R}^{p^2\times p^2}$. The relationship between $\mathbf{C}$ and $\mathcal{C}_{\boldsymbol{X}}^{(4)}$ is defined as follows:

$$\mathbf{C}_{(i-1)p+j,(k-1)p+l} \coloneqq \mathcal{C}_{\boldsymbol{X}}^{(4)}{}_{ijkl} = \mathrm{cum}(X_i, X_j, X_k, X_l), \quad 1 \le i, j, k, l \le p.$$

Furthermore, from Podosinnikova et al. (2019) (see Appendix B.2 there for details), we have $\mathbf{C} = (\mathbf{A}^* \odot \mathbf{A}^*)\mathrm{diag}(\kappa_{S_1^*}, \ldots, \kappa_{S_{N^*}^*})(\mathbf{A}^* \odot \mathbf{A}^*)^\top$. Since $\mathrm{col}(\mathbf{A}^* \odot \mathbf{A}^*) = \mathcal{A}$, we only need to prove that $\mathrm{col}(\mathbf{A}^* \odot \mathbf{A}^*)$ is identifiable. Given that $\mathrm{col}(\mathbf{C}) \subset \mathrm{col}(\mathbf{A}^* \odot \mathbf{A}^*)$ and $\mathrm{rank}(\mathbf{C}) = \mathrm{rank}(\mathbf{A}^* \odot \mathbf{A}^*)$, we have $\mathrm{col}(\mathbf{C}) = \mathrm{col}(\mathbf{A}^* \odot \mathbf{A}^*)$. The column space $\mathrm{col}(\mathbf{C})$ of $\mathbf{C}$ is completely determined by the distribution of $\boldsymbol{X}$, and as a consequence, $\boldsymbol{X}$ is subject to changes in $\mathrm{col}(\mathbf{C})$. This implies that $\mathrm{col}(\mathbf{C})$ is identifiable. This completes the proof. ■

Proof of Lemma 4. **Sufficiency:** When $\mathrm{vec}(\mathcal{D}) \cap \mathcal{A} = \{\mathbf{0}\}$: Suppose there exist $\Sigma_{\boldsymbol{E}_1} \in \mathcal{D}$ and $\Sigma_{\boldsymbol{E}_2} \in \mathcal{D}$ both satisfying

$$\mathrm{vec}(\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{E}_i}) \in \mathcal{A}, \quad i = 1, 2.$$

Then, we have

$$\mathrm{vec}(\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{E}_1} - \Sigma_{\boldsymbol{X}} + \Sigma_{\boldsymbol{E}_2}) \in \mathcal{A}.$$

Since $\mathrm{vec}(\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{E}_1} - \Sigma_{\boldsymbol{X}} + \Sigma_{\boldsymbol{E}_2}) \in \mathrm{vec}(\mathcal{D})$, we can conclude that $\Sigma_{\boldsymbol{E}_2} = \Sigma_{\boldsymbol{E}_1}$.

Necessity: When $\mathrm{vec}(\mathcal{D}) \cap \mathcal{A} \neq \{\mathbf{0}\}$: There exists a $\mathbf{D} \in \mathcal{D}$ satisfying $\mathrm{vec}(\mathbf{D}) \in \mathrm{vec}(\mathcal{D}) \cap \mathcal{A}$ and $\mathbf{D} \neq \mathbf{0}$. Then,

$$\mathrm{vec}(\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{E}}) \in \mathcal{A} \Rightarrow \mathrm{vec}(\Sigma_{\boldsymbol{X}} - (\Sigma_{\boldsymbol{E}} - \mathbf{D})) \in \mathcal{A}.$$

Hence, $\Sigma_{\boldsymbol{E}}$ is unidentifiable (we cannot distinguish between $\Sigma_{\boldsymbol{E}}$ and $\Sigma_{\boldsymbol{E}} - \mathbf{D}$). This completes the proof. ■

A.4. Proof of Lemmas in Section 2.4

Proof of Lemma 5. Let $\mathbf{A}^* = \mathbf{T}_{N^*}\mathbf{D}_A\mathbf{V}^\top$ be the SVD of $\mathbf{A}^*$. Then $(\mathbf{T}_{N^*}, \mathbf{T}_{p-N^*})$ will be a matrix whose columns are orthonormal bases of $\mathbb{R}^p$. $\widetilde{\mathbf{W}} \in \mathbb{R}^{p\times p}$ can be expressed as follows:

$$\widetilde{\mathbf{W}} = \boldsymbol{\Psi}_A\mathbf{T}_{N^*}^\top + \boldsymbol{\Psi}\mathbf{T}_{p-N^*}^\top,$$

where $\boldsymbol{\Psi}_A \in \mathbb{R}^{p\times N^*}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{p\times(p-N^*)}$.

When $\widetilde{\mathbf{W}}$ satisfies equation (2.13), this implies

$$\mathbf{A}^*\mathbf{A}^{*\top}\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{W}}\mathbf{A}^*\mathbf{A}^{*\top} = \mathbf{A}^*\mathbf{A}^{*\top}. \tag{A.6}$$

Plugging $\widetilde{\mathbf{W}} = (\mathbf{\Psi}_A, \mathbf{\Psi})(\mathbf{T}_{N^*}, \mathbf{T}_{p-N^*})^\top$ and $\mathbf{A}^* = \mathbf{T}_{N^*}\mathbf{D}_A\mathbf{V}^\top$ into (A.6), we have the following:

$$\mathbf{T}_{N^*}\mathbf{D}_A^2\mathbf{\Psi}_A^\top\mathbf{\Psi}_A\mathbf{D}_A^2\mathbf{T}_{N^*}^\top = \mathbf{T}_{N^*}\mathbf{D}_A^2\mathbf{T}_{N^*}^\top. \tag{A.7}$$

Since columns in $\mathbf{T}_{N^*}$ are linearly independent and $\mathbf{D}_A$ is a positive definite matrix, (A.7) implies that $\mathbf{\Psi}_A^\top\mathbf{\Psi}_A = \mathbf{D}_A^{-2}$. Setting $\mathbf{U}_{N^*} = \mathbf{\Psi}_A\mathbf{D}_A$, we obtain $\mathbf{U}_{N^*}^\top\mathbf{U}_{N^*} = \mathbf{I}_{N^*}$. This completes the proof. ■

Proof of Lemma 6**.** From Lemma 5, any $\widetilde{\mathbf{W}}$ satisfying equation (2.13) can be expressed as the form in equation (2.15). One can verify that $\widetilde{\mathbf{W}}\mathbf{A}^* = \mathbf{U}_{N^*}\mathbf{V}^\top = \mathbf{R}$. Then consider the following $\widehat{\mathbf{O}}$:

$$\widehat{\mathbf{O}} = (\mathbf{R}, \mathbf{R}_\perp)^\top,$$

where $\mathbf{R}_\perp$ could be any semi-orthogonal matrix satisfying $\mathbf{R}_\perp^\top\mathbf{R} = \mathbf{0}$. Then we have the following equations:

$$\begin{aligned}\widehat{\mathbf{O}}\widetilde{\mathbf{W}}\mathbf{A}^* &= (\mathbf{R}, \mathbf{R}_\perp)^\top\mathbf{R} = (\mathbf{I}_{N^*}, \mathbf{0})^\top,\\ \widehat{\mathbf{O}}\widehat{\mathbf{O}}^\top &= \mathbf{I}_p.\end{aligned}$$

This completes the proof. ■

Proof of Lemma 7**.** Recall that

$$\begin{aligned}S_j^* &= \widetilde{\boldsymbol{w}}_j^{*\top}\mathbf{A}^*\boldsymbol{S}^*,\\ \alpha_j &= \mathrm{E}\{\varphi_c(S_j^*) - S_j^*\varphi_c^{(-1)}(S_j^*)\},\end{aligned}$$

and for simplicity, define

$$\Phi(\mathbf{R}) \coloneqq \sum_{j=1}^{p} \operatorname{sign}(\alpha_j)\varphi_c^{(-2)}(\boldsymbol{r}_{j,\cdot}^{\top}\boldsymbol{S}^*) = \sum_{j=1}^{p} \operatorname{sign}(\alpha_j)\varphi_c^{(-2)}(\widetilde{\boldsymbol{w}}_j^{\top}\mathbf{A}^*\boldsymbol{S}^*).$$

The minimization of the function $\Phi(\mathbf{R})$ subject to the constraint $\mathbf{R}^{\top}\mathbf{R} = \mathbf{I}_{N^*}$ constitutes an optimization problem on the following Stiefel manifold:

$$\mathcal{M} \coloneqq \{\mathbf{R} \in \mathbb{R}^{p\times N^*} : \mathbf{R}^{\top}\mathbf{R} = \mathbf{I}_{N^*}\}.$$

In optimization within ordinary Euclidean Space, the minimizer is verified by ensuring that its gradient equals zero and its Hessian is positive definite. Here, to prove that $\mathbf{R}^* = (\mathbf{I}_{N^*}, \mathbf{0})^{\top}$ is an minimizer of $\Phi(\cdot)$ on manifold $\mathcal{M}$, we examine the Riemannian gradient $\mathcal{G}$ and Riemannian Hessian $\mathcal{H}$ of $\Phi(\cdot)$ at $\mathbf{R}^*$, ensuring they are zero and "positive definite," respectively. The verification of $\mathcal{G} = 0$ is referred to as the first-order necessary condition, also known as the KKT condition. Ensuring the positive definiteness of $\mathcal{H}$ is referred to as the verification of the second-order sufficient condition. To prove Lemma 7, it's required to demonstrate that both the KKT condition and the second-order sufficient condition are satisfied at $\mathbf{R}^*$ on the Stiefel manifold $\mathcal{M}$ (see Hu et al. (2020), pages 220–221 for details). Before delving deeper, some mathematical notation is necessary. Let $T_{\mathbf{R}^*}\mathcal{M}$ denote the tangent space of Stiefel manifold $\mathcal{M}$ at point $\mathbf{R}^*$, Hu et al. (2020) shows that:

$$T_{\mathbf{R}^*}\mathcal{M} = \{(\mathbf{A}, \mathbf{B})^{\top} : \mathbf{A} + \mathbf{A}^{\top} = \mathbf{0}, \mathbf{B} \in \mathbb{R}^{N^*\times(p-N^*)}, \mathbf{A} \in \mathbb{R}^{N^*\times N^*}\} \subseteq \mathbb{R}^{p\times N^*}.$$

$\mathcal{H}$ is essentially a tensor mapping $T_{\mathbf{R}^*}\mathcal{M}$ to $T_{\mathbf{R}^*}\mathcal{M}$ (see Hu et al. (2020), pages 215-217 for details). Subsequently, the KKT condition and second-order sufficient condition can be expressed as follows:

$$\begin{aligned} &\text{KKT condition:} && \mathcal{G} = 0, \\ &\text{Second-order sufficient condition:} && \langle \mathbf{Z}, \mathcal{H}[\mathbf{Z}] \rangle > 0, \quad \forall \mathbf{Z} \neq 0, \mathbf{Z} \in T_{\mathbf{R}^*}\mathcal{M}, \end{aligned}$$

where $\mathcal{H}[\mathbf{Z}] \in T_{\mathbf{R}^*}\mathcal{M}$ represents the transformed matrix after applying $\mathcal{H}$ to $\mathbf{Z}$. $\mathcal{G}$ and $\mathcal{H}[\mathbf{Z}]$ can be computed using $\nabla\Phi(\mathbf{R}^*)$ and $\nabla^2\Phi(\mathbf{R}^*)$ via the ensuing projection operator $\mathbb{P}_{\mathbf{R}^*}(\cdot)$ at $\mathbf{R}^*$:

$$\mathbb{P}_{\mathbf{R}^*}(\mathbf{Y}) \coloneqq \mathbf{Y} - \mathbf{R}^*(\mathbf{R}^{*\top}\mathbf{Y}/2 + \mathbf{Y}^\top\mathbf{R}^*/2), \quad \forall \mathbf{Y} \in \mathbb{R}^{p\times N^*}.$$

Plug in $\mathbf{R}^* = (\mathbf{I}_{N^*}, \mathbf{0})^\top$ into $\mathbb{P}_{\mathbf{R}^*}(\mathbf{Y})$, and let $\mathbf{Y} = (\underbrace{\mathbf{Y}_1^\top}_{N^*\times N^*}, \underbrace{\mathbf{Y}_2^\top}_{N^*\times p-N^*})^\top$, one can verify that

$$\mathbb{P}_{\mathbf{R}^*}(\mathbf{Y}) = (\mathbf{Y}_1^\top/2 - \mathbf{Y}_1/2, \mathbf{Y}_2^\top)^\top.$$

In essence, $\mathbb{P}_{\mathbf{R}^*}(\cdot)$ transforms $\mathbf{Y}_1$ into a skew-symmetric matrix while leaving $\mathbf{Y}_2$ unaffected. With this understanding of $\mathbb{P}_{\mathbf{R}^*}(\cdot)$, we can proceed with the computation of $\mathcal{G}$ and $\mathcal{H}[\mathbf{Z}]$.

$$\begin{aligned}
\mathcal{G} &= \mathbb{P}_{\mathbf{R}^*}(\nabla\Phi(\mathbf{R}^*)),\\
\mathcal{H}[\mathbf{Z}] &= \mathbb{P}_{\mathbf{R}^*}(\nabla^2\Phi(\mathbf{R}^*)[\mathbf{Z}] - \mathbf{Z}(\mathbf{R}^{*\top}\nabla\Phi(\mathbf{R}^*)/2 + \nabla\Phi(\mathbf{R}^*)^\top\mathbf{R}^*/2)),
\end{aligned}$$

where $\nabla^2\Phi(\mathbf{R}^*)[\mathbf{Z}]_{jk} = \sum_{l=1}^{p}\sum_{m=1}^{N^*}\nabla^2\Phi(\mathbf{R}^*)_{jklm}\mathbf{Z}_{lm}$. The mathematical details of $\mathcal{G}$ and $\mathcal{H}[\mathbf{Z}]$ can be found in Hu et al. (2020). Equipped with these notations and insights, we can proceed to verify the specified conditions.

KKT condition: Recall that $\mathcal{G}$ is the Riemannian gradient of the function $\Phi(\mathbf{R})$ at the point $\mathbf{R}^* \coloneqq (\mathbf{I}_{N^*}, \mathbf{0})^\top$ and $\mathcal{G} = \mathbb{P}_{\mathbf{R}^*}(\nabla\Phi(\mathbf{R}^*))$. One can compute $\nabla\Phi(\mathbf{R}^*)$ as follows:

$$\nabla\Phi(\mathbf{R}^*)_{jk} = \begin{cases} \operatorname{sign}(\alpha_j)\mathrm{E}\{\varphi_c^{(-1)}(S_j^*)S_k^*\}, & \text{if } j \le N^*,\\ 0, & \text{if } j > N^*, \end{cases}$$

where $1 \leq j \leq p$ and $1 \leq k \leq N^*$. As a result, $\mathcal{G}$ is a $p \times N^*$ matrix, where the j,k-th entry takes the form:

$$\mathcal{G}_{jk} = \mathbb{P}_{\mathbf{R}^*}(\nabla\Phi(\mathbf{R}^*))_{jk} = \begin{cases} \operatorname{sign}(\alpha_j)\mathrm{E}\{\varphi_c^{(-1)}(S_j^*)S_k^*\}- & \\ \operatorname{sign}(\alpha_j)\mathrm{E}\{\varphi_c^{(-1)}(S_j^*)S_k^*\}/2- & \text{if } j \leq N^*, \\ \operatorname{sign}(\alpha_k)\mathrm{E}\{\varphi_c^{(-1)}(S_k^*)S_j^*\}/2, & \\ 0, & \text{if } j > N^*, \end{cases}$$

where $1 \leq j \leq p$ and $1 \leq k \leq N^*$.

From Assumption A1, $S_j^* S_k^*$ and $\mathrm{E}(S_k^*) = 0$ for any $1 \leq k \leq N^*$. Therefore, we have:

$$\operatorname{sign}(\alpha_j)\mathrm{E}\{\varphi_c^{(-1)}(S_j^*)S_k^*\} = \operatorname{sign}(\alpha_k)\mathrm{E}\{\varphi_c^{(-1)}(S_k^*)S_j^*\} = 0,\ \forall j \neq k. \tag{A.8}$$

Equation (A.8) implies that $\mathcal{G}$ is a $p \times N^*$ zero matrix, and hence the KKT condition is met at $\mathbf{R}^*$.

Second-order sufficient condition: One can compute that

$$\nabla^2\Phi(\mathbf{R}^*)[\mathbf{Z}]_{jk} = \sum_{l=1}^{p}\sum_{m=1}^{N^*}\nabla^2\Phi(\mathbf{R}^*)_{jklm}\mathbf{Z}_{lm} = \begin{cases} \operatorname{sign}(\alpha_j)\mathrm{E}\{\varphi_c(S_j^*)S_k^{*2}\}\mathbf{Z}_{jk}, & \text{if } j \leq N^*, \\ \mathrm{E}\{\varphi_c(0)\}\mathbf{Z}_{jk}, & \text{if } j > N^*, \end{cases}$$

and

$$(\mathbf{Z}(\mathbf{R}^{*\top}\nabla\Phi(\mathbf{R}^*)/2 + \nabla\Phi(\mathbf{R}^*)^\top\mathbf{R}^*/2))_{jk} = \operatorname{sign}(\alpha_k)\mathrm{E}\{S_k^*\varphi_c^{(-1)}(S_k^*)\}\mathbf{Z}_{jk},$$

where $1 \leq j \leq p, 1 \leq k \leq N^*$. As a result, we can express $[\mathcal{H}[\mathbf{Z}]]_{jk}$ as follows:

$$[\mathcal{H}[\mathbf{Z}]]_{jk} = \mathbb{P}_{\mathbf{R}^*}(\nabla^2\Phi(\mathbf{R}^*)[\mathbf{Z}] - \mathbf{Z}(\mathbf{R}^{*\top}\nabla\Phi(\mathbf{R}^*)/2 + \nabla\Phi(\mathbf{R}^*)^\top\mathbf{R}^*/2))_{jk}$$

$$= \begin{cases} (\text{sign}(\alpha_j)\alpha_j + \text{sign}(\alpha_k)\alpha_k)\mathbf{Z}_{jk}/2, & j \neq k, j \leq N^*, \\ 0, & \text{if } j = k \leq N^*, \\ \{\varphi_c(0) - \text{sign}(\alpha_k)\text{E}(S_k^* \varphi_c^{(-1)}(S_k^*))\}\mathbf{Z}_{jk}, & \text{if } j > N^*. \end{cases}$$

Therefore, we have:

$$\begin{aligned} \langle \mathcal{H}[\mathbf{Z}], \mathbf{Z} \rangle &= \sum_{j=1}^{p} \sum_{k=1}^{N^*} [\mathcal{H}[\mathbf{Z}]]_{jk} \mathbf{Z}_{jk} \\ &= \sum_{j \neq k, j \leq N^*} \sum (\text{sign}(\alpha_j)\alpha_j + \text{sign}(\alpha_k)\alpha_k)\mathbf{Z}_{jk}^2/2 \\ &\quad + \sum_{k=1}^{N^*} (\varphi_c(0) - \text{sign}(\alpha_k)\text{E}(S_k^* \varphi_c^{(-1)}(S_k^*))) \sum_{j=N^*+1}^{p} Z_{jk}^2. \end{aligned}$$

One can clearly see that $(\text{sign}(\alpha_j)\alpha_j + \text{sign}(\alpha_k)\alpha_k) > 0, \quad j, k \leq N^*$, we move on to evaluate $\varphi_c(0) - \text{sign}(\alpha_k)\text{E}(S_k^* \varphi_c^{(-1)}(S_k^*))$. By the Lagrangian mean value theorem, we have the following inequality:

$$\begin{aligned} |\varphi_c^{(-1)}(S_k^*)| &= |\varphi_c^{(-1)}(S_k^*) - \varphi_c^{(-1)}(0)| \\ &\leq \varphi_c(\xi)|S_k^* - 0| < \varphi_c(0)|S_k^*|, \end{aligned} \tag{A.9}$$

where ξ is located between 0 and S_k^*. Using (A.9), we can write:

$$\begin{aligned} &\{\varphi_c(0) - \text{sign}(\alpha_k)\text{E}(S_k^* \varphi_c^{(-1)}(S_k^*))\} \\ &\geq \varphi_c(0) - \text{E}(|S_k^*||\varphi_c^{(-1)}(S_k^*)|) \\ &> \varphi_c(0) - \varphi_c(0)\text{E}(S_k^{*2}) = 0. \end{aligned}$$

Therefore,

$$\langle \mathcal{H}[\mathbf{Z}], \mathbf{Z} \rangle = \sum_{j=1}^{p} \sum_{k=1}^{N^*} [\mathcal{H}[\mathbf{Z}]]_{jk} \mathbf{Z}_{jk} > 0, \quad \forall \mathbf{Z} \neq \mathbf{0}, \mathbf{Z} \in T_{\mathbf{R}^*}\mathcal{M}.$$

This completes the proof. ■

A.5. Proof of the Lemmas in Section 2.5

Proof of Lemma 8**.**

$$\begin{aligned}\mathrm{trace}(\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^\top) &= \mathrm{trace}((\mathbf{T}_{N^*}, \mathbf{T}_{p-N^*})(\boldsymbol{\Psi}_A, \boldsymbol{\Psi})^\top(\boldsymbol{\Psi}_A, \boldsymbol{\Psi})(\mathbf{T}_{N^*}, \mathbf{T}_{p-N^*})^\top) \\ &= \mathrm{trace}(\boldsymbol{\Psi}_A\boldsymbol{\Psi}_A^\top) + \mathrm{trace}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top) \geq \mathrm{trace}(\mathbf{D}_A^{-2}).\end{aligned}$$

The equality holds if and only if $\mathrm{trace}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top) = 0$. This completes the proof. ■

A.6. Proof of the Lemmas in Section 2.6

Proof of Lemma 9**.** In the proof below, let us denote $\mathbf{R} \coloneqq \widetilde{\mathbf{W}}_{\text{interme}}\mathbf{A}^* = \mathbf{U}_{N^*}\mathbf{V}^\top$. Then $\mathbf{R}$ is a $p \times N^*$ semi-orthogonal matrix. Next, let $\mathbf{R}_\perp$ be another $p \times (p - N^*)$ semi-orthogonal matrix that belongs to the orthogonal complement of $\mathbf{R}$. Our goal is to prove that $\widehat{\mathbf{O}} = (\mathbf{R}, \mathbf{R}_\perp)^\top$ is a fixed point of Algorithm 1. We divide the proof into two steps: (i) Firstly, we prove the following equation holds:

$$\begin{aligned}\overline{\boldsymbol{o}}_j &= \mathrm{E}\{\overline{\boldsymbol{X}}\varphi_{d(\boldsymbol{o}_j)}^{(-1)}(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}})\} - \{\beta\mathrm{cov}(\overline{\boldsymbol{X}}) + (1-\beta)(\mathbf{I} + \overline{\Sigma}_{\boldsymbol{E}})\}\boldsymbol{o}_j\mathrm{E}\{\varphi_{d(\boldsymbol{o}_j)}(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}})\} \\ &= \mathbf{R}\mathrm{E}\{\boldsymbol{S}^*\varphi_c^{(-1)}(\boldsymbol{o}_j^\top\mathbf{R}\boldsymbol{S}^*)\} - \{\beta\mathbf{R}\mathbf{R}^\top + (1-\beta)\mathbf{I}\}\boldsymbol{o}_j\mathrm{E}\{\varphi_c(\boldsymbol{o}_j^\top\mathbf{R}\boldsymbol{S}^*)\}; \qquad \text{(A.10)}\end{aligned}$$

(ii) Secondly, we prove that $\widehat{\mathbf{O}}$ is a fixed point of equation (A.10).

Step (i): For simplicity, let's denote $d(\boldsymbol{o}_j) \coloneqq \sqrt{c^2 - \boldsymbol{o}_j^\top\overline{\Sigma}_{\boldsymbol{E}}\boldsymbol{o}_j}$ by $\boldsymbol{d}$. We have:

$$\begin{aligned}\overline{\boldsymbol{o}}_j &= \mathbf{R}\mathrm{E}\{\boldsymbol{S}^*\varphi_c^{(-1)}(\boldsymbol{o}_j^\top\mathbf{R}\boldsymbol{S}^*)\} - \underbrace{\{\beta\mathbf{R}\mathbf{R}^\top + (1-\beta)\mathbf{I}\}\boldsymbol{o}_j\mathrm{E}\{\varphi_c(\boldsymbol{o}_j^\top\mathbf{R}\boldsymbol{S}^*)\}}_{\text{(I)}} \\ &= \mathrm{E}\{\nabla_{\boldsymbol{o}_j}\varphi_c^{(-2)}(\boldsymbol{o}_j^\top\mathbf{R}\boldsymbol{S}^*)\} - \text{(I)} \\ &= \mathrm{E}\{\nabla_{\boldsymbol{o}_j}(\varphi_{\boldsymbol{d}}^{(-2)}(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}}) + \varphi(0)(\boldsymbol{d} - c))\} - \text{(I)} \\ &= \mathrm{E}\{\overline{\boldsymbol{X}}\varphi_{\boldsymbol{d}}^{(-1)}(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}}) + \overline{\Sigma}_{\boldsymbol{E}}\boldsymbol{o}_j(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}})\varphi_{\boldsymbol{d}}^{(-1)}(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}})/\boldsymbol{d}^2 - \overline{\Sigma}_{\boldsymbol{E}}\boldsymbol{o}_j\varphi_{\boldsymbol{d}}^{(-2)}(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}})/\boldsymbol{d}^2\end{aligned}$$

$$-\overline{\Sigma}_{\boldsymbol{E}}\boldsymbol{o}_j\varphi(0)/\boldsymbol{d}\} - (\mathrm{I}) \tag{A.11}$$

$$= \mathrm{E}\{\overline{\boldsymbol{X}}\varphi_{\boldsymbol{d}}^{(-1)}(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}}) - \overline{\Sigma}_{\boldsymbol{E}}\boldsymbol{o}_j\varphi_{\boldsymbol{d}}(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}})\} - (\mathrm{I}) \tag{A.12}$$

$$= \mathrm{E}\{\overline{\boldsymbol{X}}\varphi_{\boldsymbol{d}}^{(-1)}(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}}) - \overline{\Sigma}_{\boldsymbol{E}}\boldsymbol{o}_j\varphi_{\boldsymbol{d}}(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}})\} - \{\beta\mathbf{R}\mathbf{R}^\top + (1-\beta)\mathbf{I}_p\}\boldsymbol{o}_j\mathrm{E}\{\varphi_{\boldsymbol{d}}(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}})\}$$

$$= \mathrm{E}\{\overline{\boldsymbol{X}}\varphi_{\boldsymbol{d}}^{(-1)}(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}})\} - \{\beta(\mathbf{R}\mathbf{R}^\top + \overline{\Sigma}_{\boldsymbol{E}}) + (1-\beta)(\mathbf{I}_p + \overline{\Sigma}_{\boldsymbol{E}})\}\boldsymbol{o}_j\mathrm{E}\{\varphi_{\boldsymbol{d}}(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}})\}$$

$$= \mathrm{E}\{\overline{\boldsymbol{X}}\varphi_{d(\boldsymbol{o}_j)}^{(-1)}(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}})\} - \{\beta\mathrm{cov}(\overline{\boldsymbol{X}}) + (1-\beta)(\mathbf{I}_p + \overline{\Sigma}_{\boldsymbol{E}})\}\boldsymbol{o}_j\mathrm{E}\{\varphi_{d(\boldsymbol{o}_j)}(\boldsymbol{o}_j^\top\overline{\boldsymbol{X}})\}. \tag{A.13}$$

Step (ii): Let

$$(\widehat{\boldsymbol{o}}_1, \ldots, \widehat{\boldsymbol{o}}_p)^\top = \widehat{\mathbf{O}} = (\mathbf{R}, \mathbf{R}_\perp)^\top = (\boldsymbol{r}_{\boldsymbol{\cdot},1}, \ldots, \boldsymbol{r}_{\boldsymbol{\cdot},p})^\top.$$

Plug $\{\widehat{\boldsymbol{o}}_j\}_{j=1}^p$ into (A.10) and notice the orthogonality among $(\widehat{\boldsymbol{o}}_1, \ldots, \widehat{\boldsymbol{o}}_p)$, we have

$$\overline{\widehat{\boldsymbol{o}}}_j = \begin{cases} \widehat{\boldsymbol{o}}_j\mathrm{E}\{S_j^*\varphi_c^{(-1)}(S_j^*)\} - \widehat{\boldsymbol{o}}_j\mathrm{E}\{\varphi_c(S_j^*)\}, & \text{if } j \le N^*, \\ (1-\beta)\widehat{\boldsymbol{o}}_j\mathrm{E}\{\varphi_c(0)\}, & \text{if } N^* < j \le p. \end{cases}$$

Hence, $\overline{\widehat{\boldsymbol{o}}}_j \propto \widehat{\boldsymbol{o}}_j$ which implies that

$$(\overline{\widehat{\mathbf{O}}\widehat{\mathbf{O}}}^\top)^{(-1/2)}\overline{\widehat{\mathbf{O}}} = \mathrm{diag}\{\mathrm{sign}(\alpha_1), \ldots, \mathrm{sign}(\alpha_{N^*}), \mathrm{sign}(\alpha_{N^*+1}), \ldots, \mathrm{sign}(\alpha_p)\}\widehat{\mathbf{O}},$$

where the $\{\alpha_j\}_{j=1}^p$ have been defined in (2.6). Hence, $(\overline{\widehat{\mathbf{O}}\widehat{\mathbf{O}}}^\top)^{(-1/2)}\overline{\widehat{\mathbf{O}}}$ differs from $\widehat{\mathbf{O}}$ only by sign flips. However, sign flips are inherited ambiguities in the ICA model and will not affect the validity of the results. Thus, $\widehat{\mathbf{O}}$ can still be treated as a fixed point.

Justification of (A.11) **and** (A.12)**:** Equation (A.11) can be justified by using the chain rule and the following facts:

$$\begin{aligned} \varphi_{\boldsymbol{d}}(x) &= \varphi(x/\boldsymbol{d})/\boldsymbol{d}, \\ \varphi_{\boldsymbol{d}}^{(-1)}(x) &= \int_0^x \varphi_{\boldsymbol{d}}(x) = \varphi^{(-1)}(x/\boldsymbol{d}), \\ \varphi_{\boldsymbol{d}}^{(-2)}(x) &= \int_0^x \varphi_{\boldsymbol{d}}^{(-1)}(x) = \boldsymbol{d}\varphi^{(-2)}(x/\boldsymbol{d}). \end{aligned}$$

Equation (A.12) can be justified by using the following facts:

$$\begin{aligned}
&-\varphi^{(-2)}(x) + x\varphi^{(-1)}(x) - \varphi(0) = -\varphi(x) \\
\Rightarrow\ &-\varphi^{(-2)}(x/\boldsymbol{d}) + x/\boldsymbol{d}\varphi^{(-1)}(x/\boldsymbol{d}) - \varphi(0) = -\varphi(x/\boldsymbol{d}) \\
\Rightarrow\ &-\boldsymbol{d}\varphi^{(-2)}(x/\boldsymbol{d}) + x\varphi^{(-1)}(x/\boldsymbol{d}) - \boldsymbol{d}\varphi(0) = -\boldsymbol{d}\varphi(x/\boldsymbol{d}) = -\boldsymbol{d}^2\varphi_{\boldsymbol{d}}(x).
\end{aligned}$$

This completes the proof. ■

Proof of Theorem 2. We prove this theorem by introducing zero sources $S^*_{N^*+1}, \ldots, S^*_p$ ($S^*_{N^*+1} = \cdots = S^*_p = 0$) and the semi-orthogonal matrix $\mathbf{R}_\perp$, as defined in the proof of Lemma 9. Subsequently, let the extended source vector $(S^*_1, \ldots, S^*_{N^*}, S^*_{N^*+1}, \ldots, S^*_p)^\top$ be denoted by $\boldsymbol{S}^+$. Let the augmented matrix $(\mathbf{R}, \mathbf{R}_\perp)$ be represented as $\widehat{\mathbf{O}}^\top$ following the definition in the proof of Lemma 9. The definitions of $\mathbf{R}$ and $\mathbf{R}_\perp$ remain consistent with those presented in Lemma 9. It is important to note that neither $S^*_{N^*+1}, \ldots, S^*_p$ nor $\mathbf{R}_\perp$ are present in model (1.2). We create them in order to consider the proof of convergence on a $p \times p$ square matrix $(\mathbf{R}, \mathbf{R}_\perp)$. Introducing these zero sources and the matrix $\mathbf{R}_\perp$ does not impact the outcome, as evident from the subsequent equations:

$$\begin{aligned}
\widehat{\mathbf{O}}^\top \boldsymbol{S}^+ &= \mathbf{R}\boldsymbol{S}^* + \mathbf{R}_\perp \mathbf{0} &&= \mathbf{R}\boldsymbol{S}^*, \\
\mathbf{O}\widehat{\mathbf{O}}^\top \boldsymbol{S}^+ &= \mathbf{O}\mathbf{R}\boldsymbol{S}^* + \mathbf{0} &&= \mathbf{O}\mathbf{R}\boldsymbol{S}^*.
\end{aligned}$$

Next, let η denote the following vector:

$$\begin{aligned}
\eta &\coloneqq (\underbrace{1, \ldots, 1}_{N^*}, \underbrace{(1-\beta), \ldots, (1-\beta)}_{p-N^*})^\top, \\
\eta_0 &\coloneqq (\underbrace{0, \ldots, 0}_{N^*}, \underbrace{(1-\beta), \ldots, (1-\beta)}_{p-N^*})^\top.
\end{aligned}$$

Multiplying both sides of the iteration in (A.10) by $\widehat{\mathbf{O}}$ and denoting $\mathbf{O}\widehat{\mathbf{O}}^\top \coloneqq \mathbf{U} =$

$\begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{pmatrix}$, we can rewrite the iteration in (A.10) using $\mathbf{U}$ as follows:

$$\begin{aligned}
\overline{\boldsymbol{u}}_{i,\boldsymbol{\cdot}} &:= \widehat{\mathbf{O}}\overline{\boldsymbol{o}}_i = \widehat{\mathbf{O}}[\mathbf{R}\mathrm{E}\{\boldsymbol{S}^*\varphi_c^{(-1)}(\boldsymbol{o}_i^\top \mathbf{R}\boldsymbol{S}^*)\} - \{\beta\mathbf{R}\mathbf{R}^\top + (1-\beta)\mathbf{I}\}\boldsymbol{o}_i\mathrm{E}\{\varphi_c(\boldsymbol{o}_i^\top \mathbf{R}\boldsymbol{S}^*)\}] \\
&= \widehat{\mathbf{O}}[\widehat{\mathbf{O}}^\top\mathrm{E}\{\boldsymbol{S}^+\varphi_c^{(-1)}(\boldsymbol{o}_i^\top \widehat{\mathbf{O}}^\top\boldsymbol{S}^+)\} - \{\beta\mathbf{R}\mathbf{R}^\top + (1-\beta)\mathbf{I}\}\boldsymbol{o}_i\mathrm{E}\{\varphi_c(\boldsymbol{o}_i^\top \widehat{\mathbf{O}}^\top\boldsymbol{S}^+)\}] \\
&= \mathrm{E}\{\boldsymbol{S}^+\varphi_c^{(-1)}(\boldsymbol{u}_{i,\boldsymbol{\cdot}}^\top\boldsymbol{S}^+)\} - \{\beta[\mathbf{I}_{N^*}, \mathbf{0}_{N^*\times(p-N^*)}]^\top\mathbf{R}^\top + (1-\beta)\widehat{\mathbf{O}}\}\boldsymbol{o}_i\mathrm{E}\{\varphi_c(\boldsymbol{u}_{i,\boldsymbol{\cdot}}^\top\boldsymbol{S}^+)\}, \\
\mathbf{U} &:= \mathbf{O}\widehat{\mathbf{O}}^\top = (\overline{\mathbf{O}\mathbf{O}}^\top)^{-1/2}\overline{\mathbf{O}}\widehat{\mathbf{O}}^\top = (\overline{\mathbf{O}}\widehat{\mathbf{O}}^\top\widehat{\mathbf{O}}\overline{\mathbf{O}}^\top)^{-1/2}\overline{\mathbf{O}}\widehat{\mathbf{O}}^\top = (\overline{\mathbf{U}\mathbf{U}}^\top)^{-1/2}\overline{\mathbf{U}}.
\end{aligned}$$

Using the definition of $\widehat{\mathbf{O}}^\top$, $\mathbf{U}$ and η, we can rewrite the above equation as:

$$\begin{aligned}
\overline{\boldsymbol{u}}_{i,\boldsymbol{\cdot}} &= \mathrm{E}\{\boldsymbol{S}^+\varphi_c^{(-1)}(\boldsymbol{u}_{i,\boldsymbol{\cdot}}^\top\boldsymbol{S}^+)\} - (\boldsymbol{u}_{i,\boldsymbol{\cdot}}\circ\eta)\mathrm{E}\{\varphi_c(\boldsymbol{u}_{i,\boldsymbol{\cdot}}^\top\boldsymbol{S}^+)\}, \quad 1\le i\le p, \\
\mathbf{U} &= (\overline{\mathbf{U}}\,\overline{\mathbf{U}}^\top)^{-1/2}\overline{\mathbf{U}}.
\end{aligned} \tag{A.14}$$

Our goal is to demonstrate that the product $\mathbf{O}\mathbf{R} = [\mathbf{U}_{11}^\top, \mathbf{U}_{21}^\top]^\top$ converges to $[\mathbf{I}_{N^*}, \mathbf{0}_{N^*\times(p-N^*)}]^\top$. To validate this, it is sufficient to establish that matrix $\mathbf{U}$ converges to $\mathbf{U}^* := \begin{pmatrix} \mathbf{I}_{N^*} & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_{22}^* \end{pmatrix}$, where $\mathbf{U}_{22}^* \in \mathbb{R}^{(p-N^*)\times(p-N^*)}$ could be any orthonormal matrix since $\mathbf{U}_{22}^*$ corresponds to the zero sources $S_{N^*+1}^*, \ldots, S_p^*$ and it will not affect the outcome.

Assuming that we start from an initial point $\mathbf{U}$ within the vicinity of $\mathbf{U}^*$ and denote their discrepancy by $\boldsymbol{\varepsilon}$. Then, $\mathbf{U}$ can be expressed as $\mathbf{U}^* + \boldsymbol{\varepsilon}$. Here $\boldsymbol{\varepsilon} := (\varepsilon_{i,j})_{1\le i\le p; 1\le j\le p} := \begin{pmatrix} \boldsymbol{\varepsilon}_{11} & \boldsymbol{\varepsilon}_{12} \\ \boldsymbol{\varepsilon}_{21} & \boldsymbol{\varepsilon}_{22} \end{pmatrix}$, and we assume that entries $\varepsilon_{i,j}$ are small for all i, j. Specifically, we have $|\varepsilon_{i,j}| \le \epsilon$ for a small positive number ϵ. The $i-$row of the matrix $\mathbf{I}_p$ are denoted by $\boldsymbol{e}_i^\top$.

As mentioned above, the expression of $\mathbf{U}_{22}^*$ is not unique because the only constraint on $\mathbf{U}_{22}^*$ is that it should be an orthonormal matrix. By applying Taylor expansion to the functions $\varphi_c^{(-1)}(\cdot)$ and $\varphi_c(\cdot)$ at the point S_i^*, we obtain:

$$\begin{aligned}
\varphi_c^{(-1)}(S_i^* + \boldsymbol{e}_i^\top\boldsymbol{\varepsilon}\boldsymbol{S}^+) &= \varphi_c^{(-1)}(S_i^*) + \varphi_c(S_i^*)\boldsymbol{e}_i^\top\boldsymbol{\varepsilon}\boldsymbol{S}^+ + O(\epsilon^2), \\
\varphi_c(S_i^* + \boldsymbol{e}_i^\top\boldsymbol{\varepsilon}\boldsymbol{S}^+) &= \varphi_c(S_i^*) + \varphi_c^{(1)}(S_i^*)\boldsymbol{e}_i^\top\boldsymbol{\varepsilon}\boldsymbol{S}^+ + O(\epsilon^2).
\end{aligned} \tag{A.15}$$

Plugging (A.15) into (A.14), we have:

$$\begin{aligned}\overline{\boldsymbol{u}}_{i,\boldsymbol{\cdot}} &= \mathrm{E}\{\boldsymbol{S}^{+}\varphi_c^{(-1)}(\boldsymbol{u}_{i,\boldsymbol{\cdot}}^{\top}\boldsymbol{S}^{+})\} - (\boldsymbol{u}_{i,\boldsymbol{\cdot}}\circ\eta)\mathrm{E}\{\varphi_c(\boldsymbol{u}_{i,\boldsymbol{\cdot}}^{\top}\boldsymbol{S}^{+})\} \\ &= \underbrace{\mathrm{E}\{\boldsymbol{S}^{+}\varphi_c^{(-1)}((\boldsymbol{u}_{i,\boldsymbol{\cdot}}^{*\top}+\boldsymbol{e}_i^{\top}\boldsymbol{\varepsilon})\boldsymbol{S}^{+})\}}_{\text{(I)}} - \underbrace{(\boldsymbol{u}_{i,\boldsymbol{\cdot}}\circ\eta)\mathrm{E}\{\varphi_c((\boldsymbol{u}_{i,\boldsymbol{\cdot}}^{*\top}+\boldsymbol{e}_i^{\top}\boldsymbol{\varepsilon})\boldsymbol{S}^{+})\}}_{\text{(II)}}.\end{aligned}$$

Using the assumptions that $S_i^*S_j^*$ when $1\le i\ne j\le p$ and $\mathrm{E}\{S_i^*\}=0$ when $1\le i\le p$, we have:

$$\begin{aligned}\text{(I)} &= \mathrm{E}\{\boldsymbol{S}^{+}[\varphi_c^{(-1)}(S_i^*)+\varphi_c(S_i^*)\boldsymbol{e}_i^{\top}\boldsymbol{\varepsilon}\boldsymbol{S}^{+}]\}+O(\epsilon^2) \\ &= \mathrm{E}\{\boldsymbol{S}^{+}\varphi_c^{(-1)}(S_i^*)\}+\mathrm{E}\{\boldsymbol{S}^{+}\varphi_c(S_i^*)\boldsymbol{e}_i^{\top}\boldsymbol{\varepsilon}\boldsymbol{S}^{+}\}+O(\epsilon^2) \\ &= \boldsymbol{e}_i\mathrm{E}\{S_i^*\varphi_c^{(-1)}(S_i^*)\}+(\boldsymbol{\varepsilon}^{\top}\boldsymbol{e}_i)\circ\mathrm{E}\{\boldsymbol{S}^{+}\circ\boldsymbol{S}^{+}\varphi_c(S_i^*)\}+O(\epsilon^2), \\ \text{(II)} &= ((\boldsymbol{u}_{i,\boldsymbol{\cdot}}^{*}+\boldsymbol{\varepsilon}^{\top}\boldsymbol{e}_i)\circ\eta)\mathrm{E}\{\varphi_c(S_i^*)+\varphi_c^{(1)}(S_i^*)\boldsymbol{e}_i^{\top}\boldsymbol{\varepsilon}\boldsymbol{S}^{+}\}+O(\epsilon^2).\end{aligned}$$

In the subsequent analysis, we will derive the definitive expression of $\overline{\boldsymbol{u}}_{i,\boldsymbol{\cdot}}$ as presented in (A.16). The assumptions employed during this derivation differ depending on whether $i\le N^*$ or $i>N^*$. Consequently, we will examine these two scenarios independently.

When $\boldsymbol{i\le N^*}$: Using the facts that $S_i^*S_j^*$ when $1\le i\ne j\le p$, $\mathrm{E}\{S_i^*\}=0$ when $1\le i\le p$, $\mathrm{E}\{S_i^{*2}\}=1$ when $1\le i\le N^*$, and $\boldsymbol{u}_{i,\boldsymbol{\cdot}}^{*}=\boldsymbol{u}_{i,\boldsymbol{\cdot}}^{*}\circ\eta=\boldsymbol{e}_i$ when $1\le i\le N^*$, we have:

$$\begin{aligned}\text{(I)} &= \boldsymbol{e}_i\mathrm{E}\{S_i^*\varphi_c^{(-1)}(S_i^*)\} \\ &\quad+[\underbrace{\varepsilon_{i,1}\mathrm{E}\{\varphi_c(S_i^*)\},\cdots,\varepsilon_{i,i-1}\mathrm{E}\{\varphi_c(S_i^*)\},\varepsilon_{i,i}\mathrm{E}\{S_i^{*2}\varphi_c(S_i^*)\},\cdots,\varepsilon_{i,N^*}\mathrm{E}\{\varphi_c(S_i^*)\}}_{N^*},\underbrace{0,\cdots,0}_{p-N^*}]^{\top} \\ &\quad+O(\epsilon^2), \\ \text{(II)} &= \boldsymbol{e}_i\mathrm{E}\{\varphi_c(S_i^*)\}+\boldsymbol{e}_i\mathrm{E}\{\varphi_c^{(1)}(S_i^*)\sum_{j=1}^{p}\varepsilon_{ij}S_j^*\}+(\boldsymbol{\varepsilon}^{\top}\boldsymbol{e}_i)\circ\eta\mathrm{E}\{\varphi_c(S_i^*)\}+O(\epsilon^2) \\ &= \boldsymbol{e}_i\mathrm{E}\{\varphi_c(S_i^*)\}+\boldsymbol{e}_i\varepsilon_{ii}\mathrm{E}\{\varphi_c^{(1)}(S_i^*)S_i^*\}+(\boldsymbol{\varepsilon}^{\top}\boldsymbol{e}_i)\circ\eta\mathrm{E}\{\varphi_c(S_i^*)\}+O(\epsilon^2).\end{aligned}$$

Combine (I) and (II), group terms without $\boldsymbol{\varepsilon}$ distinctively from those with $\boldsymbol{\varepsilon}$, when $1\le i\le$

N^*, we have:

$$\begin{aligned}\overline{\boldsymbol{u}}_{i,\boldsymbol{\cdot}} &= \text{(I)} - \text{(II)} \\ &= \boldsymbol{e}_i \mathrm{E}\{S_i^* \varphi_c^{(-1)}(S_i^*) - \varphi_c(S_i^*)\} + \boldsymbol{e}_i \varepsilon_{ii} \mathrm{E}\{S_i^{*2} \varphi_c(S_i^*) - \varphi_c^{(1)}(S_i^*) S_i^* - \varphi_c(S_i^*)\} \\ &\quad -(\boldsymbol{\varepsilon}^\top \boldsymbol{e}_i) \circ \eta_0 \mathrm{E}\{\varphi_c(S_i^*)\} + O(\epsilon^2).\end{aligned}$$

When $\boldsymbol{i > N^*}$: Using the fact that $\varphi_c^{(-1)}(0) = \varphi_c^{(1)}(0) = 0$, $\mathrm{E}\{S_i^{*2}\} = 1$ when $1 \le i \le N^*$, $\mathrm{E}\{S_i^{*2}\} = 0$ when $N^* < i \le p$.

$$\begin{aligned}\text{(I)} &= (\boldsymbol{\varepsilon}^\top \boldsymbol{e}_i) \circ \mathrm{E}\{\boldsymbol{S}^+ \circ \boldsymbol{S}^+\} \varphi_c(0) + O(\epsilon^2) \\ &= (\boldsymbol{\varepsilon}^\top \boldsymbol{e}_i) \circ (\eta - \eta_0) \varphi_c(0) + O(\epsilon^2), \\ \text{(II)} &= (\boldsymbol{u}_{i,\boldsymbol{\cdot}}^* \circ \eta) \varphi_c(0) + (\boldsymbol{\varepsilon}^\top \boldsymbol{e}_i) \circ \eta \varphi_c(0) + O(\epsilon^2).\end{aligned}$$

Combining (I) and (II), grouping terms without $\boldsymbol{\varepsilon}$ distinctively from those with $\boldsymbol{\varepsilon}$, using the fact that $\boldsymbol{u}_{i,\boldsymbol{\cdot}}^* \circ \eta = \boldsymbol{u}_{i,\boldsymbol{\cdot}}^* \circ \eta_0$ when $N^* < i \le p$, we have:

$$\begin{aligned}\overline{\boldsymbol{u}}_{i,\boldsymbol{\cdot}} &= \text{(I)} - \text{(II)} \\ &= -(\boldsymbol{u}_{i,\boldsymbol{\cdot}}^* \circ \eta) \varphi_c(0) - (\boldsymbol{\varepsilon}^\top \boldsymbol{e}_i) \circ \eta_0 \varphi_c(0) + O(\epsilon^2) \\ &= -(\boldsymbol{u}_{i,\boldsymbol{\cdot}}^* \circ \eta_0) \varphi_c(0) - (\boldsymbol{\varepsilon}^\top \boldsymbol{e}_i) \circ \eta_0 \varphi_c(0) + O(\epsilon^2).\end{aligned}$$

Combine the expressions of $\overline{\boldsymbol{u}}_{i,\boldsymbol{\cdot}}$ when $1 \le i \le N^*$ and $N^* < i \le p$, we have:

$$\overline{\boldsymbol{u}}_{i,\boldsymbol{\cdot}} = \begin{cases} \boldsymbol{e}_i \mathrm{E}\{S_i^* \varphi_c^{(-1)}(S_i^*) - \varphi_c(S_i^*)\} + \boldsymbol{e}_i \varepsilon_{ii} \mathrm{E}\{S_i^{*2} \varphi_c(S_i^*) - \varphi_c^{(1)}(S_i^*) S_i^* - \varphi_c(S_i^*)\} & \text{if } i \le N^*, \\ -(\boldsymbol{\varepsilon}^\top \boldsymbol{e}_i) \circ \eta_0 \mathrm{E}\{\varphi_c(S_i^*)\} + O(\epsilon^2), & \\ -(\boldsymbol{u}_{i,\boldsymbol{\cdot}}^* \circ \eta_0) \varphi_c(0) - (\boldsymbol{\varepsilon}^\top \boldsymbol{e}_i) \circ \eta_0 \varphi_c(0) + O(\epsilon^2), & \text{if } i > N^*. \end{cases} \tag{A.16}$$

Rewriting the above equation in matrix form, we have the following:

$$\overline{\mathbf{U}} \coloneqq (\overline{\boldsymbol{u}}_{1,\boldsymbol{.}}, \ldots, \overline{\boldsymbol{u}}_{p,\boldsymbol{.}})^\top = \begin{pmatrix} \mathbf{D}_1 & \mathbf{f}(\boldsymbol{\varepsilon}) \\ 0 & \overline{\mathbf{U}}_{22} \end{pmatrix} + O(\epsilon^2),$$

where $\mathbf{D}_1 = \mathrm{diag}\{-\alpha_1 + O(\epsilon), \ldots, -\alpha_{N^*} + O(\epsilon)\}$ is a diagonal matrix, $\overline{\mathbf{U}}_{22} = (1-\beta)\varphi_c(0)\mathbf{U}_{22}$, and $\mathbf{f}(\boldsymbol{\varepsilon}) = (f_{i,j}(\boldsymbol{\varepsilon}))_{1\le i\le N^*; 1\le j\le p-N^*} = (\varepsilon_{i,j+N^*}(1-\beta)\mathrm{E}\{\varphi_c(S_i^*)\})_{1\le i\le N^*; 1\le j\le p-N^*}$ is a $N^* \times (p-N^*)$ matrix which satisfies $\max_{1\le i\le N^*; 1\le j\le p-N^*} |f_{i,j}(\boldsymbol{\varepsilon})| = \max_{1\le i\le N^*; 1\le j\le p-N^*} |\varepsilon_{i,j+N^*}(1-\beta)\mathrm{E}\{\varphi_c(S_i^*)\}| \le (1-\beta)\varphi_c(0)\epsilon$. Then,

$$\overline{\mathbf{U}}\,\overline{\mathbf{U}}^\top = \begin{pmatrix} \mathbf{D}_1^2 & \mathbf{f}(\boldsymbol{\varepsilon})\overline{\mathbf{U}}_{22}^\top \\ \overline{\mathbf{U}}_{22}\mathbf{f}(\boldsymbol{\varepsilon})^\top & \overline{\mathbf{U}}_{22}\overline{\mathbf{U}}_{22}^\top \end{pmatrix} + O(\epsilon^2).$$

Then we will use the inversion formula of block matrices to analyze $(\overline{\mathbf{U}}\,\overline{\mathbf{U}}^\top)^{-1}$. For the purpose of notation simplicity, we define:

$$\boldsymbol{\Omega} = \mathbf{D}_1^2 - \mathbf{f}(\boldsymbol{\varepsilon})\overline{\mathbf{U}}_{22}^\top(\overline{\mathbf{U}}_{22}\overline{\mathbf{U}}_{22}^\top)^{-1}\overline{\mathbf{U}}_{22}\mathbf{f}(\boldsymbol{\varepsilon})^\top = \mathbf{D}_1^2 - \mathbf{f}(\boldsymbol{\varepsilon})\mathbf{f}(\boldsymbol{\varepsilon})^\top = \mathbf{D}_1^2 + O(\epsilon^2).$$

Then, we have:

$$\begin{aligned}(\overline{\mathbf{U}}\,\overline{\mathbf{U}}^\top)^{-1} &= \begin{pmatrix} \boldsymbol{\Omega}^{-1} & -\boldsymbol{\Omega}^{-1}\mathbf{f}(\boldsymbol{\varepsilon})\overline{\mathbf{U}}_{22}^\top(\overline{\mathbf{U}}_{22}\overline{\mathbf{U}}_{22}^\top)^{-1} \\ -(\overline{\mathbf{U}}_{22}\overline{\mathbf{U}}_{22}^\top)^{-1}\overline{\mathbf{U}}_{22}\mathbf{f}(\boldsymbol{\varepsilon})^\top\boldsymbol{\Omega}^{-1} & (\overline{\mathbf{U}}_{22}\overline{\mathbf{U}}_{22}^\top)^{-1} + \overline{\mathbf{U}}_{22}^{-\top}\mathbf{f}(\boldsymbol{\varepsilon})^\top\boldsymbol{\Omega}^{-1}\mathbf{f}(\boldsymbol{\varepsilon})\overline{\mathbf{U}}_{22}^{-1} \end{pmatrix} + O(\epsilon^2) \\ &= \begin{pmatrix} \mathbf{D}_1^{-2} & -\mathbf{D}_1^{-2}\mathbf{f}(\boldsymbol{\varepsilon})\overline{\mathbf{U}}_{22}^{-1} \\ -\overline{\mathbf{U}}_{22}^{-\top}\mathbf{f}(\boldsymbol{\varepsilon})^\top\mathbf{D}_1^{-2} & (\overline{\mathbf{U}}_{22}\overline{\mathbf{U}}_{22}^\top)^{-1} \end{pmatrix} + O(\epsilon^2). \end{aligned} \tag{A.17}$$

The derivation of (A.17) uses the fact that both $\mathbf{D}_1$ and $\overline{\mathbf{U}}_{22}$ are invertible as long as ϵ is small enough.

Next, we will use the Daleckii-Krein theorem (see Carlsson (2018), page 1 for details) from matrix perturbation theory to analyze $(\overline{\mathbf{U}}\,\overline{\mathbf{U}}^\top)^{-1/2}$. Some notations is needed before using

the Daleckii-Krein theorem. Consider the following eigenvalue decomposition:

$$\begin{pmatrix} \mathbf{D}_1^{-2} & \mathbf{0} \\ \mathbf{0} & \overline{\mathbf{U}}_{22}^{-\top}\overline{\mathbf{U}}_{22}^{-1} \end{pmatrix} := \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_l \end{pmatrix} \begin{pmatrix} \mathbf{D}_1^{-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}^{-2} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_l^\top \end{pmatrix},$$

where $\mathbf{\Sigma} = \mathrm{diag}\{\sigma_1, \ldots, \sigma_{p-N^*}\}$ consists of the singular values of $\overline{\mathbf{U}}_{22}$, and $\mathbf{V}_l \in \mathbb{R}^{(p-N^*)\times(p-N^*)}$ consists of the left singular vectors of $\overline{\mathbf{U}}_{22}$.

Using the Daleckii-Krein theorem, we obtain:

$$(\overline{\mathbf{U}}\,\overline{\mathbf{U}}^\top)^{-1/2} = \begin{pmatrix} \mathbf{D}_1^{-1} & \mathbf{0} \\ \mathbf{0} & (\overline{\mathbf{U}}_{22}\overline{\mathbf{U}}_{22}^\top)^{-1/2} \end{pmatrix} + \begin{pmatrix} \mathbf{0} & \widetilde{\mathbf{f}}(\boldsymbol{\varepsilon}) \\ \widetilde{\mathbf{f}}(\boldsymbol{\varepsilon})^\top & \mathbf{0} \end{pmatrix} + O(\epsilon^2),$$

where $\widetilde{\mathbf{f}}(\boldsymbol{\varepsilon})$ takes the following form:

$$\begin{aligned} \widetilde{\mathbf{f}}(\boldsymbol{\varepsilon}) &= (\mathcal{B} \circ (-\mathbf{D}_1^{-2}\mathbf{f}(\boldsymbol{\varepsilon})\overline{\mathbf{U}}_{22}^{-1}\mathbf{V}_l))\mathbf{V}_l^\top, \\ [\mathcal{B}]_{ij} &:= \frac{1}{1/|\mathbf{D}_1(i,i)| + 1/\sigma_j} < |\mathbf{D}_1(i,i)|, \quad 1 \le i \le N^*, 1 \le j \le p - N^*. \end{aligned}$$

Multiplying $\overline{\mathbf{U}}$ by $(\overline{\mathbf{U}}\,\overline{\mathbf{U}}^\top)^{-1/2}$, we obtain

$$\begin{aligned} \mathbf{U}^+ := \begin{pmatrix} \mathbf{U}_{11}^+ & \mathbf{U}_{12}^+ \\ \mathbf{U}_{21}^+ & \mathbf{U}_{22}^+ \end{pmatrix} &:= (\overline{\mathbf{U}}\,\overline{\mathbf{U}}^\top)^{-1/2}\overline{\mathbf{U}} = \begin{pmatrix} \mathbf{D}_1^{-1} & \widetilde{\mathbf{f}}(\boldsymbol{\varepsilon}) \\ \widetilde{\mathbf{f}}(\boldsymbol{\varepsilon})^\top & (\overline{\mathbf{U}}_{22}\overline{\mathbf{U}}_{22}^\top)^{-1/2} \end{pmatrix} \begin{pmatrix} \mathbf{D}_1 & \mathbf{f}(\boldsymbol{\varepsilon}) \\ 0 & \overline{\mathbf{U}}_{22} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I}_N & \mathbf{D}_1^{-1}\mathbf{f}(\boldsymbol{\varepsilon}) + \widetilde{\mathbf{f}}(\boldsymbol{\varepsilon})\overline{\mathbf{U}}_{22} \\ \widetilde{\mathbf{f}}(\boldsymbol{\varepsilon})^\top\mathbf{D}_1 & (\overline{\mathbf{U}}_{22}\overline{\mathbf{U}}_{22}^\top)^{-1/2}\overline{\mathbf{U}}_{22} \end{pmatrix} + O(\epsilon^2). \end{aligned} \tag{A.18}$$

To prove linear convergence, we only need to show $\|\mathbf{U}^+ - \mathbf{U}^*\|_\infty < \epsilon$. From Equation (A.18), we obtain

$$\begin{aligned} \|\mathbf{U}_{11}^+ - \mathbf{I}_{N^*}\|_\infty &= O(\epsilon^2) \ll \epsilon, \\ \|\mathbf{U}_{22}^+ - \mathbf{U}_{22}^*\|_\infty &= O(\epsilon^2) \ll \epsilon. \end{aligned}$$

The only remanning task is to show $\|\mathbf{U}_{12}^{+}-\mathbf{0}\|_{\infty}<\epsilon$ and $\|\mathbf{U}_{21}^{+}-\mathbf{0}\|_{\infty}<\epsilon$. Noticing that $\mathbf{U}^{+}$ is an orthonormal matrix, we have $\|\mathbf{U}_{12}^{+}\|_F=\|\mathbf{U}_{21}^{+}\|_F$. In the following proof, we will firstly bound $\|\mathbf{U}_{12}^{+}\|_F$ and then bound $\|\mathbf{U}_{12}^{+}\|_{\infty}$ and then bound $\|\mathbf{U}_{21}^{+}\|_{\infty}$ using the fact that $\|\mathbf{U}_{12}^{+}\|_{\infty}\leq\|\mathbf{U}_{12}^{+}\|_F$.

We bound $\|\mathbf{U}_{12}^{+}\|_F$ by first bounding $\|\widetilde{\mathbf{f}}(\boldsymbol{\varepsilon})\|_F$. Define $\alpha_{\min}:=\min\limits_{1\leq i\leq N^*}|\mathbf{D}_1(i,i)|$, then

$$\begin{aligned}
\|\widetilde{\mathbf{f}}(\boldsymbol{\varepsilon})\|_F &= \|(\mathcal{B}\circ(-\mathbf{D}_1^{-2}\mathbf{f}(\boldsymbol{\varepsilon})\overline{\mathbf{U}}_{22}^{-1}\mathbf{V}_l))\mathbf{V}_l^{\top}\|_F=\|\mathcal{B}\circ(-\mathbf{D}_1^{-2}\mathbf{f}(\boldsymbol{\varepsilon})\overline{\mathbf{U}}_{22}^{-1}\mathbf{V}_l)\|_F\\
&\leq\frac{1}{\alpha_{\min}}\|\mathbf{f}(\boldsymbol{\varepsilon})\overline{\mathbf{U}}_{22}^{-1}\mathbf{V}_l\|_F=\frac{1}{\alpha_{\min}}\|\mathbf{f}(\boldsymbol{\varepsilon})\overline{\mathbf{U}}_{22}^{-1}\|_F\\
&=\frac{1}{\alpha_{\min}}\|(1-\beta)\mathrm{diag}\{\mathrm{E}\{\varphi_c(S_1^*),\cdots,\mathrm{E}\{\varphi_c(S_{N^*}^*)\}\}\boldsymbol{\varepsilon}_{12}((1-\beta)\varphi_c(0)\mathbf{U}_{22})^{-1}\|_F\\
&\leq\frac{1}{\alpha_{\min}}\|\mathrm{diag}\{\mathrm{E}\{\varphi_c(S_1^*)\}/\varphi_c(0),\ldots,\mathrm{E}\{\varphi_c(S_{N^*}^*)\}/\varphi_c(0)\}\boldsymbol{\varepsilon}_{12}\mathbf{U}_{22}^{*\top}\|_F+O(\epsilon^2)\\
&\leq\frac{1}{\alpha_{\min}}\|\boldsymbol{\varepsilon}_{12}\|_F+O(\epsilon^2)\\
&\leq\frac{\epsilon}{\alpha_{\min}}\sqrt{N^*(p-N^*)}+O(\epsilon^2).
\end{aligned}$$

Secondly, we bound the Frobenius norm of $\mathbf{D}_1^{-1}\mathbf{f}(\boldsymbol{\varepsilon})+\widetilde{\mathbf{f}}(\boldsymbol{\varepsilon})\overline{\mathbf{U}}_{22}$.

$$\begin{aligned}
\|\mathbf{D}_1^{-1}\mathbf{f}(\boldsymbol{\varepsilon})+\widetilde{\mathbf{f}}(\boldsymbol{\varepsilon})\overline{\mathbf{U}}_{22}\|_F &\leq \|\mathbf{D}_1^{-1}\mathbf{f}(\boldsymbol{\varepsilon})\|_F+\|\widetilde{\mathbf{f}}(\boldsymbol{\varepsilon})\overline{\mathbf{U}}_{22}\|_F\\
&\leq\frac{(1-\beta)\varphi_c(0)}{\alpha_{\min}}\|\boldsymbol{\varepsilon}_{12}\|_F+(1-\beta)\varphi_c(0)\|\widetilde{\mathbf{f}}(\boldsymbol{\varepsilon})\mathbf{U}_{22}\|_F+O(\epsilon^2)\\
&=\frac{(1-\beta)\varphi_c(0)}{\alpha_{\min}}\|\boldsymbol{\varepsilon}_{12}\|_F+(1-\beta)\varphi_c(0)\|\widetilde{\mathbf{f}}(\boldsymbol{\varepsilon})\mathbf{U}_{22}^{*}\|_F+O(\epsilon^2)\\
&=\frac{(1-\beta)\varphi_c(0)}{\alpha_{\min}}\|\boldsymbol{\varepsilon}_{12}\|_F+(1-\beta)\varphi_c(0)\|\widetilde{\mathbf{f}}(\boldsymbol{\varepsilon})\|_F+O(\epsilon^2)\\
&\leq\frac{(1-\beta)\varphi_c(0)}{\alpha_{\min}}\|\boldsymbol{\varepsilon}_{12}\|_F+\frac{(1-\beta)\varphi_c(0)}{\alpha_{\min}}\|\boldsymbol{\varepsilon}_{12}\|_F+O(\epsilon^2)\\
&\leq\frac{2(1-\beta)\varphi_c(0)\epsilon}{\alpha_{\min}}\sqrt{N^*(p-N^*)}+O(\epsilon^2).
\end{aligned}\tag{A.19}$$

From (A.18) and (A.19), we have:

$$\begin{aligned}
\|\mathbf{U}_{12}^{+}\|_F^2 &\leq\Big\{\frac{2(1-\beta)\varphi_c(0)\epsilon}{\alpha_{\min}}\Big\}^2N^*(p-N^*)+O(\epsilon^3)\\
&\leq(1-\beta)^2\varphi_c^2(0)\epsilon^2p^2/\alpha_{\min}^2+O(\epsilon^3).
\end{aligned}\tag{A.20}$$

From (A.20), when $(1-\beta) < \alpha_{\min}/(2\varphi_c(0)p)$, we have $\|\mathbf{U}_{12}^+\|_F^2 < \epsilon^2/4 + O(\epsilon^3)$. Given that $\mathbf{U}^+$ is an orthonormal matrix, we have $\|\mathbf{U}_{12}^+\|_F^2 = \|\mathbf{U}_{21}^+\|_F^2$. This implies that:

$$\|\mathbf{U}_{12}^+\|_F = \|\mathbf{U}_{21}^+\|_F < \epsilon/2, \quad \text{if } 1-\beta < \alpha_{\min}/(2\varphi_c(0)p) \text{ and } \epsilon \text{ small enough.}$$

Hence, $\|\mathbf{U}^+ - \mathbf{U}^*\|_\infty$ can be strictly bounded by $\epsilon/2$ if we choose $(1-\beta)$ small enough. This implies a linear convergence rate, and this completes the proof. ■

Appendix B

Supplementary Files

B.1. More graphical results

From Figure B.1, one can see that the prewhitened dICA has smaller estimation error. Meanwhile, the prewhitened dICA reaches the fixed point within 400 steps while the non-

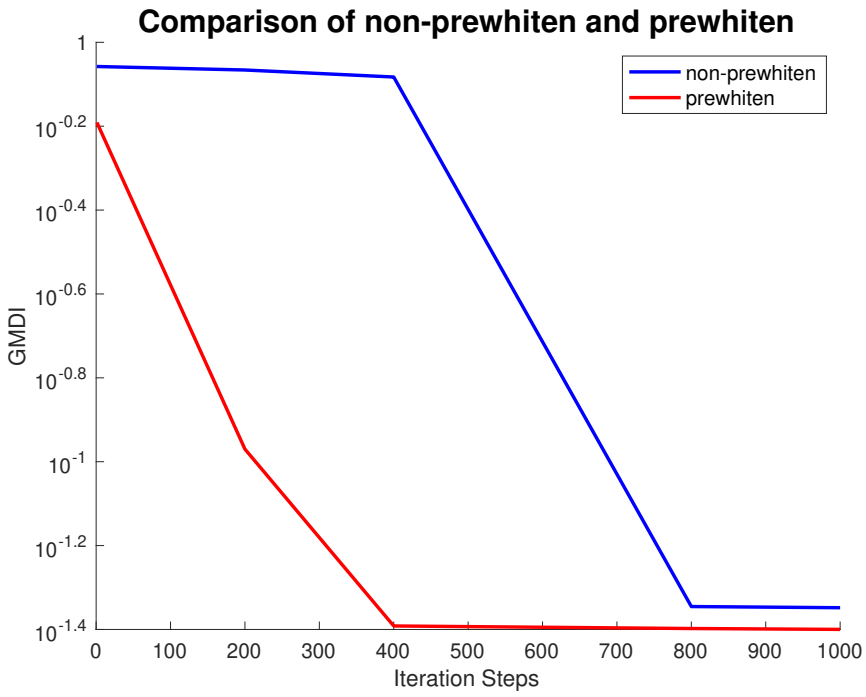

Figure B.1: The Error of estimating $\mathbf{A}^*$ with respect to the iteration steps

prewhitened dICA takes 800 steps to achieve that.

Bibliography

Armijo, L. (1966). Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics 16*, 1–3.

Arora, S., R. Ge, A. Moitra, and S. Sachdeva (2012). Provable ICA with unknown gaussian noise, with implications for Gaussian mixtures and autoencoders. *Advances in Neural Information Processing Systems 25*.

Artoni, F., A. Delorme, and S. Makeig (2019). A visual working memory dataset collection with bootstrap Independent Component Analysis for comparison of electroencephalographic preprocessing pipelines. *Data in brief 22*, 787–793.

Beckmann, C. F. and S. M. Smith (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE transactions on medical imaging 23*(2), 137–152.

Bell, A. J. and T. J. Sejnowski (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation 7*(6), 1129–1159.

Bonhomme, S. and J.-M. Robin (2009). Consistent noisy independent component analysis. *Journal of Econometrics 149*(1), 12–25.

Cardoso, J.-F. (1991). Super-symmetric decomposition of the fourth-order cumulant tensor. Blind identification of more sources than sensors. In [*Proceedings*] *ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, pp. 3109–3112 vol.5.

Carlsson, M. (2018). Perturbation theory for the matrix square root and matrix modulus. *arXiv preprint arXiv:1810.01464*.

Chen, Y., L. Niu, R.-B. Chen, and Q. He (2019). Sparse-Group Independent Component Analysis with application to yield curves prediction. *Computational Statistics & Data Analysis 133*, 76–89.

Cichocki, A., S. Douglas, and S. Amari (1998). Robust techniques for independent component analysis (ICA) with noisy data. *Neurocomputing 22*(1), 113–129.

Davies, M. (2004). Identifiability issues in noisy ICA. *IEEE Signal processing letters 11*(5), 470–473.

De Lathauwer, L., J. Castaing, and J.-F. Cardoso (2007). Fourth-order cumulant-based blind identification of underdetermined mixtures. *IEEE Transactions on Signal Processing 55*(6), 2965–2973.

De Lathauwer, L., B. De Moor, and J. Vandewalle (1996). Independent component analysis based on higher-order statistics only. In *Proceedings of 8th Workshop on Statistical Signal and Array Processing*, pp. 356–359. IEEE.

Delorme, A. and S. Makeig (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods 134*(1), 9–21.

Delorme, A., J. Palmer, J. Onton, R. Oostenveld, and S. Makeig (2012). Independent EEG sources are dipolar. *PloS one 7*(2), e30135.

EEGLAB (2023). EEGLAB Wiki. `https://sccn.ucsd.edu/wiki/EEGLAB`. Accessed: 2023-12-19.

Gouriéroux, C., A. Monfort, and J.-P. Renne (2017). Statistical inference for independent component analysis: Application to structural VAR models. *Journal of Econometrics 196*(1), 111–126.

Green, C. G., R. R. Nandy, and D. Cordes (2002). PCA-preprocessing of fmri data adversely affects the results of ICA. In *Proceedings of international society of magnetic resonance in medicine*, pp. 10.

Guo, R., C. Zhang, and Z. Zhang (2020). Maximum Independent Component Analysis with Application to EEG Data. *Statistical Science 35*(1), 145 – 157.

Hastie, T. and R. Tibshirani (2002). Independent components analysis through product density estimation. *Advances in neural information processing systems 15*.

Herrmann, J. M. and F. J. Theis (2007). Statistical analysis of sample-size effects in ica. In *International Conference on Intelligent Data Engineering and Automated Learning*, pp. 416–425. Springer.

Hu, J., X. Liu, Z.-W. Wen, and Y.-X. Yuan (2020). A brief introduction to manifold optimization. *Journal of the Operations Research Society of China 8*, 199–248.

Hyvärinen, A. (1998). Independent component analysis in the presence of gaussian noise by maximizing joint likelihood. *Neurocomputing 22*(1-3), 49–67.

Hyvarinen, A. (1999). Gaussian moments for noisy independent component analysis. *IEEE signal processing letters 6*(6), 145–147.

Hyvärinen, A., J. Karhunen, and E. Oja (2001). *Independent Component Analysis*. New York: Wiley.

Hyvärinen, A. and E. Oja (2000). Independent component analysis: algorithms and applications. *Neural networks 13*(4-5), 411–430.

Hyvarinen, A., E. Oja, P. Hoyer, and J. Hurri (1998). Image feature extraction by sparse coding and independent component analysis. In *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*, Volume 2, pp. 1268–1273. IEEE.

Ikeda, S. and K. Toyama (2000). Independent component analysis for noisy data—MEG data analysis. *Neural Networks 13*(10), 1063–1074.

Ilmonen, P., K. Nordhausen, H. Oja, and E. Ollila (2010). A new performance index for ICA: properties, computation and asymptotic analysis. In *Latent Variable Analysis and Signal Separation: 9th International Conference, LVA/ICA 2010, St. Malo, France, September 27-30, 2010. Proceedings 9*, pp. 229–236. Springer.

Koldovsky, Z. and P. Tichavsky (2006). Methods of fair comparison of performance of linear ICA techniques in presence of additive noise. In *2006 ieee international conference on acoustics speech and signal processing proceedings*, Volume 5, pp. V–V. IEEE.

Koldovskỳ, Z. and P. Tichavskỳ (2007). Blind instantaneous noisy mixture separation with best interference-plus-noise rejection. In *Independent Component Analysis and Signal Separation: 7th International Conference, ICA 2007, London, UK, September 9-12, 2007. Proceedings 7*, pp. 730–737. Springer.

Lawson, C. L. and R. J. Hanson (1976). Solving least squares problems. In *Classics in applied mathematics*.

Lee, T.-W., M. Girolami, and T. J. Sejnowski (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural computation 11*(2), 417–441.

Lietzén, N., J. Virta, K. Nordhausen, and P. Ilmonen (2020). Minimum distance index for BSS, generalization, interpretation and asymptotics. *Aust. J. Stat 49*(4), 57–68.

Liu, Y., H. Xu, Z. Xia, and Z. Gong (2018). Multi-spectrometer calibration transfer based on independent component analysis. *Analyst 143*(5), 1274–1280.

McKeown, M. J., L. K. Hansen, and T. J. Sejnowsk (2003). Independent component analysis of functional MRI: what is signal and what is noise? *Current opinion in neurobiology 13*(5), 620–629.

Morel, P. (2018, mar). Gramm: grammar of graphics plotting in matlab. *The Journal of Open Source Software 3*(23), 568.

Nocedal, J. and S. J. Wright (2006). *Numerical Optimization* (2e ed.). New York, NY, USA: Springer.

Ofner, P., A. Schwarz, J. Pereira, D. Wyss, R. Wildburger, and G. R. Müller-Putz (2019). Attempted arm and hand movements can be decoded from low-frequency EEG from persons with spinal cord injury. *Scientific reports 9*(1), 7134.

Parikh, N. and S. Boyd (2014, jan). Proximal algorithms. *Found. Trends Optim. 1*(3), 127–239.

Pfister, N., S. Weichwald, P. Bühlmann, and B. Schölkopf (2019). Robustifying independent component analysis by adjusting for group-wise stationary noise. *Journal of Machine Learning Research 20*(147), 1–50.

Pion-Tonachini, L., K. Kreutz-Delgado, and S. Makeig (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage 198*, 181–197.

Podosinnikova, A., A. Perry, A. S. Wein, F. Bach, A. dâAspremont, and D. Sontag (2019). Overcomplete independent component analysis via SDP. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2583–2592. PMLR.

Porrill, J. and J. V. Stone (1998). Undercomplete independent component analysis for signal separation and dimension reduction. *report, Citeseer*.

Risk, B. B., D. S. Matteson, and D. Ruppert (2019). Linear non-Gaussian component analysis via maximum likelihood. *Journal of the American Statistical Association 114*(525), 332–343.

The MathWorks, I. (2019). *Symbolic Math Toolbox*. Natick, Massachusetts, United State.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Tichavsky, P., Z. Koldovsky, and E. Oja (2006). Performance analysis of the FastICA algorithm and Crame/spl acute/r-rao bounds for linear independent component analysis. *IEEE transactions on Signal Processing 54*(4), 1189–1203.

Virta, J. and K. Nordhausen (2019). Estimating the number of signals using principal component analysis. *Stat 8*(1), e231.

Voss, J. R., M. Belkin, and L. Rademacher (2015). A pseudo-euclidean iteration for optimal recovery in noisy ICA. *Advances in neural information processing systems 28*.

Voss, J. R., L. Rademacher, and M. Belkin (2013). Fast algorithms for Gaussian noise invariant independent component analysis. *Advances in neural information processing systems 26*.

Wei, T. (2015). A convergence and asymptotic analysis of the generalized symmetric FastICA algorithm. *IEEE transactions on signal processing 63*(24), 6445–6458.

Wright, S. J. (2015). Coordinate descent algorithms. *Mathematical programming 151*(1), 3–34.

Wu, Y. (2023). *Topics in Independent Component Analysis With Noises*. Ph. D. thesis. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-07-19.

Xia, Y. (2020). Chapter Eleven - Correlation and association analyses in microbiome study integrating multiomics in health and disease. In J. Sun (Ed.), *The Microbiome in Health and Disease*, Volume 171 of *Progress in Molecular Biology and Translational Science*, pp. 309–491. Academic Press.

Yeredor, A. (2000). Blind source separation via the second characteristic function. *Signal Processing 80*(5), 897–902.

Yeredor, A. (2009). On optimal selection of correlation matrices for matrix-pencil-based separation. In *International Conference on Independent Component Analysis and Signal Separation*, pp. 187–194. Springer.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 49–67.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.