High Dimensional Dependent Data Analysis for Neuroimaging

by

Hai Shu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2016

Doctoral Committee:

Professor Bin Nan, Chair
Assistant Professor Veronica Berrocal
Professor Timothy D. Johnson
Professor Elizaveta Levina

To my parents and Zhe

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my advisor Dr. Bin Nan. Without his patient and rigorous instructions as well as thoughtful and insightful suggestions, this dissertation would not have been completed. I have benefited a lot from his philosophy about the integration of theory with practice. I also deeply thank his generous financial support since I came here to study for my Master's degree.

My great thanks also go out to my committee members: Dr. Elizaveta Levina, Dr. Timothy D. Johnson, and Dr. Veronica Berrocal. It has been my great honor to have them on my committee. Thanks for providing valuable comments and suggestions on my research.

I also appreciate Dr. Bhramar Mukherjee, Dr. Veronica Berrocal, Dr. Carlos Mendes de Leon, Dr. Sung Kyun Park, and Dr. Roderick J. Little for their kind financial support in summer and fall of 2014.

I am thankful for Ms. Nicole Fenech and Mr. Chris Scheller from the Department of Biostatistics, and Dr. Kirsten Herold of the SPH Writing Lab for their excellent support.

I also want to thank all my friends for their careful concern and helpful advice. Also thank you to Google and Wikipedia for their wonderful resources.

I must also thank my parents and my girlfriend Zhe, for their absolute faith and constant love for me. This dissertation is dedicated to you.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

APPENDIX

ABSTRACT

This dissertation contains three projects focusing on two major high-dimensional problems for dependent data, particularly neuroimaging data: multiple testing and estimation of large covariance/precision matrices.

Project 1 focuses on the multiple testing problem. Traditional voxel-level false discovery rate (FDR) controlling procedures for neuroimaging data often ignore the spatial correlations among neighboring voxels, thus suffer from substantial loss of efficiency in reducing the false non-discovery rate. We extend the one-dimensional hidden Markov chain based local-significance-index procedure to three-dimensional hidden Markov random field (HMRF). To estimate model parameters, a generalized EM algorithm is proposed for maximizing the penalized likelihood. Simulations show increased efficiency of the proposed approach over commonly used FDR controlling procedures. We apply the method to the comparison between patients with mild cognitive impairment and normal controls in the ADNI FDG-PET imaging study.

Project 2 considers estimating large covariance and precision matrices from temporally dependent observations, in particular, the resting-state functional MRI (rfMRI) data in brain functional connectivity studies. Existing work on large covariance and precision matrices is primarily for i.i.d. observations. The rfMRI data from the Human Connectome Project, however, are shown to have long-range memory. Assuming a polynomial-decay-dominated temporal dependence, we obtain convergence rates for the generalized thresholding estimation of covariance and correlation matrices, and for the constrained ℓ_1 minimization and the ℓ_1 penalized likelihood estimation of precision matrix. Properties of sparsistency and sign-consistency are also established. We apply the considered methods to estimating the functional connectivity from single-subject rfMRI data.

Project 3 extends Project 2 to multiple independent samples of temporally dependent observations. This is motivated by the group-level functional connectivity analysis using rfMRI data, where each subject has a sample of temporally dependent image observations. We use different concentration inequalities to obtain faster convergence rates than those in Project 2 of the considered estimators for multi-sample data. The new proof allows more general within-sample temporal dependence. We also discuss a potential way of improving the convergence rates by using a weighted sample covariance matrix. We apply the considered methods to the functional connectivity estimation for the ADHD-200 rfMRI data.

CHAPTER I

Introduction

High dimensional data refers to cases where the number of variables p is comparable to or larger than the number of observations n, i.e., the so-called "large p, small n" or "large p, large n" scenarios (Donoho et al., 2000; Johnstone and Titterington, 2009). The classical statistical methods built on the "small p, large n" assumption often fail to efficiently handle high dimensional data. This has been called the "curse of dimensionality" (Bellman, 1961). Over the last two decades, significant development has been made in high dimensional data analysis, which is motivated primarily by numerous applications in fields such as neuroscience, genomics, economics and finance (see Fan et al., 2014a).

Neuroimaging data are high dimensional data. The sample size n of images is usually only a few hundred or thousand; however, the variable dimension p can vary from several hundred for brain regions to several hundred thousand for brain voxels. Studies of mental diseases such as Alzheimer's disease, attention deficit hyperactivity disorder (ADHD), schizophrenia and Parkinson's disease can benefit from neuroimaging data analysis. The abnormality found by the analysis is helpful for diagnosing the disease, monitoring disease progression, and understanding the mechanisms underlying the disease. Examples of neuroimaging data are the three-dimensional (3D) ^{18}F-Fluorodeoxyglucose positron emission tomography (FDG-PET) data and the 4D functional magnetic resonance imaging (fMRI) data (temporally observed 3D images), which are involved with spatial and/or temporal dependence.

Statistical methods developed for high dimensional data are largely based on certain independent structures of the data, for which either the p variables are independent or the n observations are independent and even identically distributed (i.i.d.). For example, many multiple testing procedures (Benjamini and Hochberg, 1995, 2000; Genovese and Wasserman, 2004) are built on the former structure, and most large covariance/precision matrix estimating methods (e.g., Rothman et al., 2008, 2009; Cai et al., 2011) assume the latter structure. However, the validity and efficiency of these approaches are questionable for data without such independent structures, in particular, neuroimaging data. Specifically, the first structure is violated when the test statistic obtained at a brain voxel is correlated with the statistics at its neighboring voxels. The second structure fails for the temporally dependent image observations of the fMRI data.

Motivated by the need to analyze neuroimaging data, this dissertation contains three projects focusing on the two major high-dimensional problems for dependent data: multiple testing and estimation of large covariance/precision matrices. In Project 1, an efficient multiple testing procedure is proposed for certain spatially correlated data. In Projects 2 and 3, we study the validity of three widely used estimating methods (Rothman et al., 2008, 2009; Cai et al., 2011), originally developed for i.i.d. observations, under some models of temporal dependence.

In Chapter II, we present Project 1, which focuses on the multiple testing problem. Since it was introduced by Benjamini and Hochberg (1995), the false discovery rate (FDR) has been widely used in multiple testing as an alternative measure of Type I error, specifically for the family-wise error rate (FWER), which is the probability of making at least one Type I error. FDR is defined as the expected proportion of false rejections among the rejected hypotheses. The authors showed that there is a potential gain in power for controlling FDR compared to controlling FWER. The corresponding measure of Type II error to FDR, called the false non-discovery rate (FNR; Genovese and Wasserman, 2002), is the expected proportion of errors among the accepted hypotheses. An FDR controlling procedure is said to be optimal (Sun and Cai, 2009) if it has the smallest FNR among all procedures controlling FDR at a pre-specified level. Traditional FDR procedures (Benjamini and Hochberg, 1995, 2000; Genovese and Wasserman, 2004) theoretically based on independent test statistics may substantially lose the efficiency in reducing FNR under certain dependence structures (Sun and Cai, 2009). To address this problem, Sun and Cai (2009) proposed an optimal FDR procedure built on a new test statistic called the local index of significance (LIS) and a hidden Markov chain (HMC) which models the one-dimensional dependence structure. Wei et al. (2009) extended this procedure to test statistics with different HMC dependence structures.

However, the one-dimensional HMC is not applicable for 3D neuroimaging data. In Chapter II, we extend the LIS-based procedure (Sun and Cai, 2009; Wei et al., 2009) for such data, by using a hidden Markov random field (HMRF) model, in particular, the Ising model (see Brémaud, 1999), to capture the 3D spatial structure. When the HMRF parameters are known, an optimal property is proved for the proposed HMRF-LIS-based procedure. In practice, the HMRF parameters are unknown. To avoid the unboundedness of the original likelihood function, a penalized likelihood approach is applied to the HMRF parameter estimation. A generalized expectation-maximization algorithm is proposed for maximizing the penalized likelihood. Extensive simulations show the superiority of the proposed approach over commonly used FDR procedures in terms of reducing FNR. Using FDG-PET data from the Alzheimer's Disease Neuroimaging Initiative database (`adni.loni.usc.edu`), we apply the method to a comparison between patients with mild cognitive impairment, a disease status with increased risk of developing Alzheimer's or other dementia, and normal controls. More signals are found by the proposed approach than by competing methods, with most discovered signals in regions typically affected by Alzheimer's disease.

Chapter III is devoted to Project 2, on estimating large covariance and precision matrices from temporally dependent observations. This project is motivated by the functional connectivity analysis using resting-state fMRI data. The functional connectivity refers to the statistical associations of activation among brain nodes (regions or voxels; Friston, 2011; Zhou et al., 2009); thus, it can be assessed by either correlations or partial correlations from the covariance matrix or the inverse covariance matrix (a.k.a. precision matrix) respectively. The traditional estimator of the covariance matrix, the sample covariance matrix, is no longer a consistent estimator when the variable dimension p (the number of brain nodes) grows with the sample size n, e.g., $p/n \to c \in (0, \infty)$ in the sense that its eigenvalues may diverge from those of the covariance matrix (Bai and Yin, 1993; Bai and Silverstein, 2010). Moreover, when $p > n$, the sample covariance matrix is not invertible, and thus it cannot be directly applied for estimating the precision matrix by matrix inversion. When the observations are i.i.d., many consistent estimating approaches have been developed, such as the generalized thresholding (Rothman et al., 2009) estimation for covariance matrix, and the constrained ℓ_1 minimization (CLIME; Cai et al., 2011) and the ℓ_1 penalized likelihood estimation (Rothman et al., 2008) for precision matrix. Recently, Chen et al. (2013), Bhattacharjee and Bose (2014), and Zhou (2014) considered the estimation by using temporally dependent observations. But with restrictive conditions, their models do not fit well for the resting-state fMRI data, which may exhibit heterogeneous long-range temporal dependence among the p time series.

To conquer this problem, we consider the aforementioned three estimating approaches under a polynomial-decay-dominated (PDD) temporal dependence. We provide the convergence rates of the considered estimators under both the spectral norm and the Frobenius norm (that is divided by $\sqrt{p}$) which are widely used in the literature (Bickel and Levina, 2008a,b; Rothman et al., 2008, 2009; Cai et al., 2011). Properties of sparsistency and sign-consistency are also established. To reduce the temporal dependence between training and validation datasets, a gap-block cross-validation method is proposed for the tuning parameter selection, which performs well in simulations. We apply the considered approaches to analyzing a single subject's functional connectivity using the resting-state fMRI data obtained from the Human Connectome Project (humanconnectome.org). The discovered functional hubs may be useful for further scientific investigation.

Project 3 is presented in Chapter IV. It is an extension of Project 2 from a single sample of temporally dependent observations to multiple independent samples. This project is motivated by estimating the group-level functional connectivity from multiple subjects each with a sample of temporally dependent image observations. We use the sample covariance matrix obtained from the concatenation of all observations (Smith et al., 2013; Ng et al., 2013) for the estimating methods considered in Project 2. The proof used in Project 2 does not make effective use of the independence among samples. A different proof technique can show improved convergence rates for the multiple samples except the CLIME method for estimating the precision matrix under short-range temporal dependence. Moreover, the new proof allows more general within-sample temporal dependence. We apply the sample-covariance-matrix based methods to estimating the group-level functional connectivity of ADHD patients compared to normal controls using the ADHD-200 resting-state fMRI data (neurobureau.projects.nitrc.org/ADHD200).

At the end of Chapter IV, we also discuss a potential way of improving the convergence rates by using a weighted sample covariance matrix. Accounting for potentially different temporal dependence structures among these samples, a weight assigned for each sample in the proposed matrix aims to be proportional with its effective sample size. Using this matrix as the initial estimator of the covariance matrix can theoretically have faster convergence rates than using the sample covariance matrix, if with appropriate weights. However, to select such weights is difficult in practice.

We leave some future work for discussion in Chapter V.

CHAPTER II

Multiple Testing for Neuroimaging via Hidden Markov Random Field

2.1 Introduction

In a seminal paper, Benjamini and Hochberg (1995) introduced false discovery rate (FDR) as an alternative measure of Type I error in multiple testing problems to the family-wise error rate (FWER). They showed that the FDR is equivalent to the FWER if all null hypotheses are true and is smaller otherwise, thus FDR controlling procedures potentially have a gain in power over FWER controlling procedures. FDR is defined as the expected proportion of false rejections among all rejections. The false nondiscovery rate (FNR; Genovese and Wasserman, 2002), the expected proportion of falsely accepted hypotheses among all acceptances, is the corresponding measure of Type II error. The traditional FDR procedures (Benjamini and Hochberg, 1995, 2000; Genovese and Wasserman, 2004), which are p-value based, are theoretically developed under the assumption that the test statistics are independent. Although these approaches are shown to be valid in controlling FDR under certain dependence assumptions (Benjamini and Yekutieli, 2001; Farcomeni, 2007; Wu, 2008), they may suffer from severe loss of efficiency in reducing FNR when the dependence structure is ignored (Sun and Cai, 2009). By modeling the dependence structure using a hidden Markov chain (HMC), Sun and Cai (2009) proposed an oracle FDR procedure built on a new test statistic, the local index of significance (LIS), and the corresponding asymptotic data-driven procedure, which are optimal in the sense that they minimize the marginal FNR subject to a constraint on the marginal FDR. Following the work of Sun and Cai (2009), Wei et al. (2009) developed a pooled LIS (PLIS) procedure for multiple-group analysis where different groups have different HMC dependence structures, and proved the optimality of the PLIS procedure. Either the LIS procedure or the PLIS procedure only handles the one-dimensional dependency. However, problems with higher dimensional dependence are of particular practical interest in analyzing imaging data.

FDR procedures have been widely used in analyzing neuroimaging data, such as positron emission tomography (PET) imaging and functional magnetic resonance imaging (fMRI) data (Genovese et al., 2002; Chumbley and Friston, 2009; Chumbley et al., 2010, among many others). We extend the work of Sun and Cai (2009) in this chapter by developing an optimal LIS-based FDR procedure for three-dimensional (3D) imaging data using a hidden Markov random field model (HMRF) for the spatial dependency among multiple tests. Existing methods for correlated imaging data, for example, Zhang et al. (2011) are not shown to be optimal, i.e., minimizing FNR.

HMRF model is a generalization of HMC model, which replaces the underlying Markov chain by Markov random field. A well-known classical Markov random field with two states is the Ising model. In particular, the two-parameter Ising model, whose formal definition is given in equation (2.1), reduces to the two-state Markov chain in one-dimension (Brémaud, 1999). The Ising model and its generalization with more than two states, the Potts model, have been widely used to capture the spatial structure in image analysis; see Brémaud (1999), Winkler (2003), Zhang et al. (2008), Huang et al. (2013) and Johnson et al. (2013), among others. In this chapter, we consider a hidden Ising model for each area based on the Brodmann's partition of the cerebral cortex (Garey, 2006) and subcortical regions of the human brain, which provides a natural way of modeling spatial correlations for neuroimaging data. To the best of our knowledge, this is the first work that introduces the HMRF-LIS based FDR procedure to the field of neuroimaging.

We propose a generalized expectation-maximization algorithm (GEM; Dempster et al., 1977) to search for penalized maximum likelihood estimators (Ridolfi, 1997; Ciuperca et al., 2003; Chen et al., 2008) of the hidden Ising model parameters. The penalized likelihood prevents the unboundedness of the likelihood function, and the proposed GEM uses Monte Carlo averages via Gibbs sampler (Geman and Geman, 1984; Roberts and Smith, 1994) to overcome the intractability of computing the normalizing constant in the underlying Ising model. Then the LIS-based FDR procedures can be conducted by plugging in the estimates of the hidden Ising model parameters. In what follows, we use the term "HMRF" to refer to the 3D hidden Ising model.

The chapter is organized as follows. In Section 2.2, we introduce the HMRF model, i.e., the hidden Ising model, for 3D imaging data. We provide the GEM algorithm for the HMRF parameter estimation and the implementation of the HMRF-LIS-based data-driven procedures in Section 2.3. In Section 2.4, we conduct extensive simulations to compare the LIS-based procedures with conventional FDR methods. In Section 2.5, we apply the PLIS procedure to the ^{18}F-Fluorodeoxyglucose PET (FDG-PET) image data of the Alzheimer's Disease Neuroimaging Initiative (ADNI), which finds more signals than conventional methods.

2.2 A Hidden Markov Random Field Model

Let S be a finite lattice of N voxels in an image grid, usually in a 3D space. Let $\boldsymbol{\Theta} = \{\Theta_s \in \{0, 1\} : s \in S\}$ denote the set of latent states on S, where $\Theta_s = 1$ if the null hypothesis at voxel s is false and $\Theta_s = 0$ otherwise. For simplicity, we follow Sun and Cai

(2009) to call hypothesis s to be nonnull if $\Theta_s = 1$ and null otherwise. We also call voxel s to be a signal if $\Theta_s = 1$ and noise otherwise. Let $\boldsymbol{\Theta}$ be generated from a two-parameter Ising model with the following probability distribution

$$P_{\boldsymbol{\varphi}}(\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\varphi})} \exp\{\boldsymbol{\varphi}^T \boldsymbol{H}(\boldsymbol{\theta})\} = \frac{1}{Z(\beta, h)} \exp\left\{\beta \sum_{\langle s,t\rangle} \theta_s\theta_t + h \sum_{s\in S} \theta_s\right\}, \tag{2.1}$$

where $Z(\boldsymbol{\varphi})$ is the normalizing constant, $\boldsymbol{\varphi} = (\beta, h)^T$, $\boldsymbol{H}(\boldsymbol{\theta}) = (\sum_{\langle s,t\rangle} \theta_s\theta_t, \sum_{s\in S} \theta_s)^T$, and $\langle s, t\rangle$ denotes all the unordered pairs in S such that for any s, t is among the six nearest neighbors of voxel s in a 3D setting. This model possesses the Markov property:

$$P_{\boldsymbol{\varphi}}(\theta_s|\boldsymbol{\theta}_{S\setminus\{s\}}) = P_{\boldsymbol{\varphi}}(\theta_s|\boldsymbol{\theta}_{\mathcal{N}(s)}) = \frac{\exp\{\theta_s(\beta \sum_{t\in\mathcal{N}(s)} \theta_t + h)\}}{1 + \exp\{\beta \sum_{t\in\mathcal{N}(s)} \theta_t + h\}},$$

where $S \setminus \{s\}$ denotes the set S after removing s, and $\mathcal{N}(s) \subset S$ is the nearest neighborhood of s in S.

For the above Ising model, it can also be shown that

$$\log\left\{\frac{P(\Theta_s{=}1, \Theta_t{=}1|\boldsymbol{\theta}_{S\setminus\{s,t\}})P(\Theta_s{=}0, \Theta_t{=}0|\boldsymbol{\theta}_{S\setminus\{s,t\}})}{P(\Theta_s{=}1, \Theta_t{=}0|\boldsymbol{\theta}_{S\setminus\{s,t\}})P(\Theta_s{=}0, \Theta_t{=}1|\boldsymbol{\theta}_{S\setminus\{s,t\}})}\right\} = \begin{cases} \beta, & t \in \mathcal{N}(s), \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

Therefore, if s and t are neighbors, β is equal to a log odds ratio that describes the association between Θ_s and Θ_t conditional on all the other state variables being withheld. We can see that β reflects how likely the same-state voxels are clustered together. Similarly,

$$\log\left\{\frac{P(\Theta_s = 1| \sum_{t\in\mathcal{N}(s)} \Theta_t = 0)}{P(\Theta_s = 0|\sum_{t\in\mathcal{N}(s)}\Theta_t = 0)}\right\} = h,$$

which is the log odds for $\Theta_s = 1$ given that $\boldsymbol{\Theta}_{\mathcal{N}(s)}$ are all zero. Thus, that $\beta \geq 0$ and $h \leq 0$ implies the nonnegative dependency of state variables at neighboring voxels. In addition, for a voxel s with m nearest neighbors, we have

$$\log\left\{\left(\frac{P(\Theta_s=1|\sum_{t\in\mathcal{N}(s)}\Theta_t=k)}{P(\Theta_s=0|\sum_{t\in\mathcal{N}(s)}\Theta_t=k)}\right)\Bigg/\left(\frac{P(\Theta_s=0|\sum_{t\in\mathcal{N}(s)}\Theta_t=m-k)}{P(\Theta_s=1|\sum_{t\in\mathcal{N}(s)}\Theta_t=m-k)}\right)\right\}$$

$$= m\beta + 2h, \tag{2.3}$$

where k is an integer satisfying $0 \leq k \leq m$, which reflects the log ratio of the cluster effect of signals (nonnulls) relative to the cluster effect of noises (nulls).

We assume the observed z-values $\boldsymbol{X} = \{X_s : s \in S\}$ are independent given $\boldsymbol{\Theta} = \boldsymbol{\theta}$ with

$$P_{\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{s\in S} P_{\boldsymbol{\phi}}(x_s|\theta_s), \tag{2.4}$$

where $P_{\boldsymbol{\phi}}(x_s|\theta_s)$ denotes the following distribution

$$X_s|\Theta_s \sim (1-\Theta_s)N(\mu_0,\sigma_0^2) + \Theta_s\sum_{l=1}^{L} p_l N(\mu_l,\sigma_l^2) \tag{2.5}$$

with $(\mu_0,\sigma_0^2) = (0,1)$, unknown parameters $\boldsymbol{\phi} = (\mu_1,\sigma_1^2,p_1,...,\mu_L,\sigma_L^2,p_L)^T$, $\sum_{l=1}^{L} p_l = 1$ and $p_l \geq 0$. In particular, the z-value X_s follows the standard normal distribution under the null, and the nonnull distribution is set to be the normal mixture that can be used to approximate a large collection of distributions (Magder and Zeger, 1996; Efron, 2004). The number of components L in the nonnull distribution may be selected by, for example, the Akaike or Bayesian information criterion. Following the recommendation of Sun and Cai (2009), we use $L = 2$ for the ADNI image analysis.

Markov random fields (MRFs; Brémaud, 1999) are a natural generalization of Markov chains (MCs), where the time index of MC is replaced by the space index of MRF. It is well known that any one-dimensional MC is an MRF, and any one-dimensional stationary finite-valued MRF is an MC (Chandgotia et al., 2014). When S is taken to be one-dimensional, the above approach based on (2.1), (2.4) and (2.5) reduces to the HMC method of Sun and Cai (2009).

2.3 Hidden Markov Random Field LIS-Based FDR Procedures

Sun and Cai (2009) developed a compound decision theoretic framework for multiple testing under HMC dependence and proposed LIS-based oracle and data-driven testing procedures that aim to minimize the FNR subject to a constraint on FDR. We extend these procedures under HMRF for image data. The oracle LIS for hypothesis s is defined as $LIS_s(\boldsymbol{x}) = P_{\boldsymbol{\Phi}}(\Theta_s = 0|\boldsymbol{x})$ for a given parameter vector $\boldsymbol{\Phi}$. In our model, $\boldsymbol{\Phi} = (\boldsymbol{\phi}^T, \boldsymbol{\varphi}^T)^T$. Let $LIS_{(1)}(\boldsymbol{x}), ..., LIS_{(N)}(\boldsymbol{x})$ be the ordered LIS values and $\mathcal{H}_{(1)}, ..., \mathcal{H}_{(N)}$ the corresponding null hypotheses. The oracle procedure operates as follows: for a prespecified FDR level α,

$$\text{(2.6)} \qquad \text{let } k = \max\left\{ i : \frac{1}{i}\sum_{j=1}^{i} LIS_{(j)}(\boldsymbol{x}) \leq \alpha \right\}, \text{ then reject all } \mathcal{H}_{(i)},\ i = 1, ..., k.$$

Parameter $\boldsymbol{\Phi}$ is unknown in practice. We can use the data-driven procedure that simply replaces $LIS_{(i)}(\boldsymbol{x})$ in (2.6) with $\widehat{LIS}_{(i)}(\boldsymbol{x}) = P_{\hat{\boldsymbol{\Phi}}}(\Theta_{(i)} = 0|\boldsymbol{x})$, where $\hat{\boldsymbol{\Phi}}$ is an estimate of $\boldsymbol{\Phi}$.

If all the tests are partitioned into multiple groups and each group follows its own HMRF, in contrast to the separated LIS (SLIS) procedure that conducts the LIS-based FDR procedure separately for each group at the same FDR level α and then combines the testing results, we follow Wei et al. (2009) to propose a pooled LIS (PLIS) procedure that is more efficient in reducing the global FNR. The PLIS follows the same procedure as (2.6), but with $LIS_{(1)}, ..., LIS_{(N)}$ being the ordered test statistics from all groups.

Note that the model homogeneity, which is required in Sun and Cai (2009) and Wei et al. (2009) for HMCs, fails to hold for the HMRF model. In other words, $P(\Theta_s = 1)$ for the interior voxels with six nearest neighbors are different to those for the boundary voxels with less than six nearest neighbors. We show the validity and optimality of the oracle HMRF-LIS-based procedures in Appendix A.1.

We now provide details of the LIS-based data-driven procedure for 3D image data, where the parameters of the HMRF model need to be estimated from observed test data.

2.3.1 A Generalized EM Algorithm

We start this subsection by showing the unboundedness of the observed likelihood function of HMRF. For any voxel $t \in S$, define a specific configuration of $\boldsymbol{\Theta}$ by $\boldsymbol{\theta}_{\{t\}} = (\theta_s)_{s \in S}$ with $\theta_t = 1$ and $\theta_s = 0$ if $s \neq t$. Then the observed likelihood function

$$\begin{aligned} L(\boldsymbol{\Phi}|\boldsymbol{x}) &= P_{\boldsymbol{\Phi}}(\boldsymbol{x}) = \sum_{\boldsymbol{\Theta}} P_{\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{\Theta}) P_{\boldsymbol{\varphi}}(\boldsymbol{\Theta}) \\ &\geq P_{\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{\Theta} = \boldsymbol{\theta}_{\{t\}}) P_{\boldsymbol{\varphi}}(\boldsymbol{\Theta} = \boldsymbol{\theta}_{\{t\}}) \\ &= P_{\boldsymbol{\phi}}(x_t|\Theta_t = 1) \prod_{s \in S \setminus \{t\}} P_{\boldsymbol{\phi}}(x_s|\Theta_s = 0) P_{\boldsymbol{\varphi}}(\boldsymbol{\Theta}_{S \setminus \{t\}} = \boldsymbol{0}, \Theta_t = 1) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{ -\frac{(x_t - \mu_1)^2}{2\sigma_1^2} \right\} + \sum_{l=2}^{L} N(x_t; \mu_l, \sigma_l^2) \right) \\ &\qquad \times (2\pi)^{-\frac{N-1}{2}} \exp\left\{ -\frac{1}{2} \sum_{s \in S \setminus \{t\}} x_s^2 \right\} \frac{e^h}{Z(\beta, h)} \\ &\to \infty \end{aligned}$$

if $\mu_1 = x_t$ and $\sigma_1^2 \to 0$ with other parameters fixed. Thus the observed likelihood function is unbounded. The similar unbounded-likelihood phenomenon for Gaussian hidden Markov chain model has been shown in Ridolfi (1997) and Chen et al. (2014).

One solution to avoid the unboundedness is to replace the likelihood by a penalized likelihood (Ridolfi, 1997; Ciuperca et al., 2003)

$$pL(\boldsymbol{\Phi}|\boldsymbol{x}) = L(\boldsymbol{\Phi}|\boldsymbol{x}) \prod_{l=1}^{L} g(\sigma_l^2), \tag{2.7}$$

where $g(\sigma_l^2)$, $l = 1, \ldots, L$, are penalty functions that ensure the boundedness of $pL(\boldsymbol{\Phi}|\boldsymbol{x})$. We follow Ridolfi (1997) and Ciuperca et al. (2003) to choose

$$g(\sigma_l^2) \propto \frac{1}{\sigma_l^{2b}} \exp\left\{-\frac{a}{\sigma_l^2}\right\}, \quad a > 0, b \geq 0,$$

where $x \propto y$ means that $x = cy$ with a positive constant c independent of any parameter. Note that (2.7) reduces to the unpenalized likelihood function when $a = b = 0$. When $a > 0$ and $b > 1$, the penalized likelihood approach is equivalent to setting $g(\sigma_l^2)$ to be the inverse gamma distribution, which is a classical prior distribution for the variance of a normal distribution in Bayesian statistics (Hoff, 2009). We do not impose any prior distribution here. The choice of a and b does not impact the strong consistency of the penalized maximum likelihood estimator (PMLE) based on the same penalty function for a finite mixture of normal distributions (Ciuperca et al., 2003; Chen et al., 2008). Such a penalty performs well in the simulations, though formal proof of the consistency of PMLE for hidden Ising model remains an open question.

We develop an EM algorithm based on the penalized likelihood (2.7) for the estimation of parameters in the HMRF model characterized by (2.1), (2.4) and (2.5). We introduce unobservable categorical variables $\boldsymbol{K} = \{K_s : s \in S\}$, where $K_s = 0$ if $\Theta_s = 0$, and $K_s \in \{1, ..., L\}$ if $\Theta_s = 1$. Hence, $P(K_s{=}0|\Theta_s{=}0) = 1$ and we denote $P(K_s{=}l|\Theta_s{=}1) = p_l$. From (2.5), we let $X_s|K_s \sim N(\mu_{K_s}, \sigma_{K_s}^2)$. To estimate the HMRF parameters $\boldsymbol{\Phi} = (\boldsymbol{\phi}^T, \boldsymbol{\varphi}^T)^T$, $(\boldsymbol{\Theta}, \boldsymbol{K}, \boldsymbol{X})$ are used as the complete data variables to construct the auxiliary function in the $(t{+}1)$st iteration of EM algorithm given the observed data $\boldsymbol{x}$ and the current estimated parameters $\boldsymbol{\Phi}^{(t)}$:

$$Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^{(t)}) = E_{\boldsymbol{\Phi}^{(t)}}[\log P_{\boldsymbol{\Phi}}(\boldsymbol{\Theta}, \boldsymbol{K}, \boldsymbol{X})|\boldsymbol{x}] + \sum_{l=1}^{L} \log g(\sigma_l^2),$$

where $P_{\boldsymbol{\Phi}}(\boldsymbol{\Theta}, \boldsymbol{K}, \boldsymbol{X}) = P_{\boldsymbol{\varphi}}(\boldsymbol{\Theta}) P_{\boldsymbol{\phi}}(\boldsymbol{X}, \boldsymbol{K}|\boldsymbol{\Theta}) = P_{\boldsymbol{\varphi}}(\boldsymbol{\Theta}) \prod_{s \in S} P_{\boldsymbol{\phi}}(X_s, K_s|\Theta_s)$. The Q-

function can be further written as follows

$$Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^{(t)}) = Q_1(\boldsymbol{\phi}|\boldsymbol{\Phi}^{(t)}) + Q_2(\boldsymbol{\varphi}|\boldsymbol{\Phi}^{(t)}),$$

where

$$Q_1(\boldsymbol{\phi}|\boldsymbol{\Phi}^{(t)}) = \sum_{\boldsymbol{\Theta}}\sum_{\boldsymbol{K}} P_{\boldsymbol{\Phi}^{(t)}}(\boldsymbol{\Theta}, \boldsymbol{K}|\boldsymbol{x}) \log P_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{K}|\boldsymbol{\Theta}) + \sum_{l=1}^{L} \log g(\sigma_l^2)$$

and

$$Q_2(\boldsymbol{\varphi}|\boldsymbol{\Phi}^{(t)}) = \sum_{\boldsymbol{\Theta}} P_{\boldsymbol{\Phi}^{(t)}}(\boldsymbol{\Theta}|\boldsymbol{x}) \log P_{\boldsymbol{\varphi}}(\boldsymbol{\Theta}).$$

Therefore, we can maximize $Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^{(t)})$ for $\boldsymbol{\Phi}$ by maximizing $Q_1(\boldsymbol{\phi}|\boldsymbol{\Phi}^{(t)})$ for $\boldsymbol{\phi}$ and $Q_2(\boldsymbol{\varphi}|\boldsymbol{\Phi}^{(t)})$ for $\boldsymbol{\varphi}$, separately.

Maximizing $Q_1(\boldsymbol{\phi}|\boldsymbol{\Phi}^{(t)})$ under the constraint $\sum_{l=1}^{L} p_l = 1$ by the method of Lagrange multipliers yields

$$\begin{aligned}
&(2.8) & p_l^{(t+1)} &= \frac{\sum_{s\in S} w_s^{(t)}(l)}{\sum_{s\in S} \gamma_s^{(t)}(1)}, \\
&(2.9) & \mu_l^{(t+1)} &= \frac{\sum_{s\in S} w_s^{(t)}(l) x_s}{\sum_{s\in S} w_s^{(t)}(l)}, \\
&(2.10) & (\sigma_l^2)^{(t+1)} &= \frac{2a + \sum_{s\in S} w_s^{(t)}(l)(x_s - \mu_l^{(t+1)})^2}{2b + \sum_{s\in S} w_s^{(t)}(l)},
\end{aligned}$$

where

$$w_s(l) = \frac{\gamma_s(1) p_l f_l(x_s)}{f(x_s)},\ \gamma_s(i) = P_{\boldsymbol{\Phi}}(\Theta_s = i|\boldsymbol{x}),\ f_l = N(\mu_l, \sigma_l^2), \text{ and } f = \sum_{l=1}^{L} p_l f_l.$$

For $Q_2(\boldsymbol{\varphi}|\boldsymbol{\Phi}^{(t)})$, taking its first and second derivatives with respect to $\boldsymbol{\varphi}$, we obtain

$$\begin{aligned}
\boldsymbol{U}^{(t+1)}(\boldsymbol{\varphi}) &= \frac{\partial}{\partial \boldsymbol{\varphi}} Q_2(\boldsymbol{\varphi}|\boldsymbol{\Phi}^{(t)}) = E_{\boldsymbol{\Phi}^{(t)}}[\boldsymbol{H}(\boldsymbol{\Theta})|\boldsymbol{x}] - E_{\boldsymbol{\varphi}}[\boldsymbol{H}(\boldsymbol{\Theta})], \\
\boldsymbol{I}(\boldsymbol{\varphi}) &= -\frac{\partial^2}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}^T} Q_2(\boldsymbol{\varphi}|\boldsymbol{\Phi}^{(t)}) = Var_{\boldsymbol{\varphi}}[\boldsymbol{H}(\boldsymbol{\Theta})].
\end{aligned}$$

Maximizing $Q_2(\boldsymbol{\varphi}|\boldsymbol{\Phi}^{(t)})$ is then equivalent to solving the nonlinear equation:

$$(2.11) \qquad \boldsymbol{U}^{(t+1)}(\boldsymbol{\varphi}) = E_{\boldsymbol{\Phi}^{(t)}}[\boldsymbol{H}(\boldsymbol{\Theta})|\boldsymbol{x}] - E_{\boldsymbol{\varphi}}[\boldsymbol{H}(\boldsymbol{\Theta})] = \boldsymbol{0}.$$

It can be shown that equation (2.11) has a unique solution and can be solved by the Newton-Raphson (NR) method (Stoer and Bulirsch, 2002). However, a starting point that is not close enough to the solution may result in divergence of the NR method. Therefore, rather than searching for the solution of equation (2.11) over all $\boldsymbol{\varphi}$, we choose a $\boldsymbol{\varphi}^{(t+1)}$ that increases $Q_2(\boldsymbol{\varphi}|\boldsymbol{\Phi}^{(t)})$ over its value at $\boldsymbol{\varphi} = \boldsymbol{\varphi}^{(t)}$. Together with the maximization of $Q_1(\boldsymbol{\phi}|\boldsymbol{\Phi}^{(t)})$, the approach leads to $Q(\boldsymbol{\Phi}^{(t+1)}|\boldsymbol{\Phi}^{(t)}) \geq Q(\boldsymbol{\Phi}^{(t)}|\boldsymbol{\Phi}^{(t)})$ and thus $pL(\boldsymbol{\Phi}^{(t+1)}|\boldsymbol{x}) \geq pL(\boldsymbol{\Phi}^{(t)}|\boldsymbol{x})$, which is termed a GEM algorithm (Dempster et al., 1977). To find such a $\boldsymbol{\varphi}^{(t+1)}$ that increases the Q_2-function, a backtracking line search algorithm (Nocedal and Wright, 2006) is applied with a set of decreasing positive values λ_m in the following

$$\boldsymbol{\varphi}^{(t+1,m)} = \boldsymbol{\varphi}^{(t)} + \lambda_m \boldsymbol{I}(\boldsymbol{\varphi}^{(t)})^{-1}\boldsymbol{U}^{(t+1)}(\boldsymbol{\varphi}^{(t)}), \tag{2.12}$$

where $m = 0, 1, ...,$ and $\boldsymbol{\varphi}^{(t+1)} = \boldsymbol{\varphi}^{(t+1,m)}$ which is the first one satisfying the Armijo condition (Nocedal and Wright, 2006)

$$Q_2(\boldsymbol{\varphi}^{(t+1,m)}|\boldsymbol{\Phi}^{(t)}) - Q_2(\boldsymbol{\varphi}^{(t)}|\boldsymbol{\Phi}^{(t)}) \geq \alpha\lambda_m \boldsymbol{U}^{(t+1)}(\boldsymbol{\varphi}^{(t)})^T \boldsymbol{I}(\boldsymbol{\varphi}^{(t)})^{-1}\boldsymbol{U}^{(t+1)}(\boldsymbol{\varphi}^{(t)}). \tag{2.13}$$

Since $\boldsymbol{I}(\boldsymbol{\varphi}^{(t)})$ is positive-definite, the Armijo condition guarantees the increase of Q_2-function. In practice, α is chosen to be quite small. We adopt $\alpha = 10^{-4}$, which is recommended by Nocedal and Wright (2006), and halve the Newton-Raphson step length each time by using $\lambda_m = 2^{-m}$.

In the GEM algorithm, Monte Carlo averages are used via Gibbs sampler to approximate the quantities of interest that are involved with the intractable normalizing constant of the Ising model. By the ergodic theorem of the Gibbs sampler (Roberts and Smith,

1994) (see Appendix A.2 for details),

$$\begin{aligned}
\boldsymbol{U}^{(t+1)}(\boldsymbol{\varphi}) &\approx \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{H}(\boldsymbol{\theta}^{(t,i,\boldsymbol{x})}) - \boldsymbol{H}(\boldsymbol{\theta}^{(i,\boldsymbol{\varphi})})\right), \\
\boldsymbol{I}(\boldsymbol{\varphi}) &\approx \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{H}(\boldsymbol{\theta}^{(i,\boldsymbol{\varphi})}) - \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{H}(\boldsymbol{\theta}^{(j,\boldsymbol{\varphi})})\right)^{\otimes 2},
\end{aligned}$$

where $\{\boldsymbol{\theta}^{(t,1,\boldsymbol{x})}, ..., \boldsymbol{\theta}^{(t,n,\boldsymbol{x})}\}$ are large n samples successively generated by the Gibbs sampler from

$$P_{\boldsymbol{\Phi}^{(t)}}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\exp\left\{\beta^{(t)}\sum_{\langle s,r\rangle}\theta_s\theta_r + \sum_{s\in S}h_s^{(t)}\theta_s\right\}}{Z\left(\beta^{(t)}, \{h_s^{(t)}\}_{s\in S}\right)},$$

with

$$\begin{aligned}
h_s^{(t)} = h^{(t)} &- \log\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{(x_s-\mu_0)^2}{2\sigma_0^2}\right\}\right) \\
&+ \log\left(\sum_{l=1}^{L}\frac{p_l^{(t)}}{\sqrt{2\pi\sigma_l^{2(t)}}}\exp\left\{-\frac{(x_s-\mu_l^{(t)})^2}{2\sigma_l^{2(t)}}\right\}\right)
\end{aligned}$$

and $Z\left(\beta^{(t)}, \{h_s^{(t)}\}_{s\in S}\right)$ being the normalizing constant, and $\{\boldsymbol{\theta}^{(1,\boldsymbol{\varphi})}, ..., \boldsymbol{\theta}^{(n,\boldsymbol{\varphi})}\}$ are generated from $P_{\boldsymbol{\varphi}}(\boldsymbol{\theta})$. Here for vector $\boldsymbol{v}$, $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^T$. Similarly,

$$\frac{C}{Z(\boldsymbol{\varphi})} = E_{\boldsymbol{\varphi}}[\exp\{-\boldsymbol{\varphi}^T\boldsymbol{H}(\boldsymbol{\Theta})\}] \approx \frac{1}{n}\sum_{i=1}^{n}\exp\{-\boldsymbol{\varphi}^T\boldsymbol{H}(\boldsymbol{\theta}^{(i,\boldsymbol{\varphi})})\},$$

where C is the number of all possible configurations $\boldsymbol{\theta}$ of $\boldsymbol{\Theta}$. Then the difference between Q_2-functions in the Armijo condition can be approximated by

$$\begin{aligned}
&Q_2(\boldsymbol{\varphi}^{(t+1,m)}|\boldsymbol{\Phi}^{(t)}) - Q_2(\boldsymbol{\varphi}^{(t)}|\boldsymbol{\Phi}^{(t)}) \\
&\quad\approx \frac{1}{n}(\boldsymbol{\varphi}^{(t+1,m)} - \boldsymbol{\varphi}^{(t)})^T\sum_{i=1}^{n}\boldsymbol{H}(\boldsymbol{\theta}^{(t,i,\boldsymbol{x})}) \\
&\qquad + \log\left(\frac{\sum_{i=1}^{n}\exp\{-\boldsymbol{\varphi}^{(t+1,m)T}\boldsymbol{H}(\boldsymbol{\theta}^{(i,\boldsymbol{\varphi}^{(t+1,m)})})\}}{\sum_{i=1}^{n}\exp\{-\boldsymbol{\varphi}^{(t)T}\boldsymbol{H}(\boldsymbol{\theta}^{(i,\boldsymbol{\varphi}^{(t)})})\}}\right).
\end{aligned}$$

Back to $Q_1(\boldsymbol{\phi}|\boldsymbol{\Phi}^{(t)})$, the local conditional probability of $\boldsymbol{\Theta}$ given $\boldsymbol{x}$ can also be approximated by the Gibbs sampler:

$$\gamma_s^{(t)}(i) = P_{\boldsymbol{\Phi}^{(t)}}(\Theta_s = i|\boldsymbol{x}) \approx \frac{1}{n}\sum_{k=1}^{n}\mathbf{1}(\theta_s^{(t,k,\boldsymbol{x})} = i). \tag{2.14}$$

2.3.2 Implementation of the LIS-Based FDR Procedure

The algorithm for the LIS-based data-driven procedure, denoted as LIS for single group analysis, SLIS for separate analysis of multiple groups, and PLIS for pooled analysis for multiple groups, is given below:

1. Set initial values $\mathbf{\Phi}^{(0)} = \{\boldsymbol{\phi}^{(0)}, \boldsymbol{\varphi}^{(0)}\}$ for the model parameters $\mathbf{\Phi}$ of each group;
2. Update $\boldsymbol{\phi}^{(t)}$ from equations (2.8), (2.9) and (2.10);
3. Update $\boldsymbol{\varphi}^{(t)}$ from equations (2.12) and (2.13);
4. Iterate Steps 2 and 3 until convergence, then obtain the estimate $\hat{\mathbf{\Phi}}$ of $\mathbf{\Phi}$;
5. Plug-in $\hat{\mathbf{\Phi}}$ to obtain the test statistics $\widehat{LIS}$ from equation (2.14);
6. Apply the data-driven procedure (LIS, SLIS or PLIS).

The GEM algorithm is stopped when the following stopping rule

$$\max_i \left(\frac{|\Phi_i^{(t+1)} - \Phi_i^{(t)}|}{|\Phi_i^{(t)}| + \epsilon_1} \right) < \epsilon_2, \tag{2.15}$$

where Φ_i is the ith coordinate of vector $\mathbf{\Phi}$, is satisfied for three consecutive regular Newton-Raphson iterations with $m = 0$ in (2.12), or the prespecified maximum number of iterations is reached. Stopping rule (2.15) was applied by Booth and Hobert (1999) to the Monte Carlo EM method, where they set $\epsilon_1 = 0.001$, ϵ_2 between 0.002 and 0.005, and the rule to be satisfied for three consecutive iterations to avoid stopping the algorithm prematurely because of Monte Carlo error. We used $\epsilon_1 = \epsilon_2 = 0.001$ in simulation studies and real-data analysis. Constant $\alpha = 10^{-4}$ is recommended by Nocedal and Wright (2006) for the Armijo condition (2.13), and the Newton-Raphson step length in (2.12) is halved by using $\lambda_m = 2^{-m}$. In practice, the Armijo condition (2.13) might not be satisfied when the step length $\|\boldsymbol{\varphi}^{(t+1,m)} - \boldsymbol{\varphi}^{(t)}\|$ is very small. In this situation, the iteration within Step 3 is stopped by an alternative criterion

$$\max_i \left(\frac{|\varphi_i^{(t+1,m)} - \varphi_i^{(t)}|}{|\varphi_i^{(t)}| + \epsilon_1} \right) < \epsilon_3$$

with $\epsilon_3 < \epsilon_2$, for example, $\epsilon_3 = 10^{-4}$ if $\epsilon_2 = 0.001$. Small a and b should be chosen in (2.10). We choose $a = 1$ and $b = 2$.

2.4 Simulation Studies

The simulation setups are similar to those in Sun and Cai (2009) and Wei et al. (2009), but with 3D data. The performances of the proposed LIS-based oracle (OR) and data-driven procedures are compared with the BH approach (Benjamini and Hochberg, 1995), the q-value procedure (Storey, 2003), and the local FDR (Lfdr) procedure (Sun and Cai, 2007) for single group analysis; and the performances of SLIS and PLIS are compared with BH, q-value, and the conditional Lfdr (CLfdr) procedure (Cai and Sun, 2009) for multiple groups. The Lfdr and CLfdr procedures are shown to be optimal for independent tests (Sun and Cai, 2007; Cai and Sun, 2009). For simulations with multiple groups, all the procedures are globally implemented using all the locally computed test statistics based on each method from each group. The q-values are obtained using the R package `qvalue` (Dabney and Storey, 2014). For the Lfdr or CLfdr procedure, we use the proportion of the null cases generated from the Ising model with given parameters as the estimate of the probability of the null cases $P(\Theta_s = 0)$, together with the given null and nonnull distributions without estimating their parameters. For the LIS-based data-driven procedures, the maximum number of GEM iterations is set to be 1,000 with $\epsilon_1 = \epsilon_2 = 0.001$, $\epsilon_3 = \alpha = 10^{-4}$, $a = 1$ and $b = 2$. For the Gibbs sampler, 5,000 samples are generated from 5,000 iterations after a burn-in period of 1,000 iterations. In all simulations, each HMRF is on a $N = 15{\times}15{\times}15$ cubic lattice S, the number of replications $M = 200$ is the same as that in Wei et al. (2009), and the nominal FDR level is set at 0.10.

2.4.1 Single-Group Analysis

Study 1: $L = 1$

The MRF $\boldsymbol{\Theta} = \{\Theta_s : s \in S\}$ is generated from the Ising model (2.1) with parameters (β, h), and the observations $\boldsymbol{X} = \{X_s : s \in S\}$ are generated conditionally on $\boldsymbol{\Theta}$ from $X_s|\Theta_s \sim (1 - \Theta_s)N(0, 1) + \Theta_s N(\mu_1, \sigma_1^2)$. Note that the MRF $\boldsymbol{\Theta}$ is not observable in practice. Figure 2.1 shows the comparisons of the performance of BH, q-value, Lfdr, OR and LIS. In Figure 2.1(1a-1c), we fix $h = -2.5$, set $\mu_1 = 2$ and $\sigma_1^2 = 1$, and plot FDR, FNR, and the average number of true positives (ATP) yielded by these procedures as functions of β. In Figure 2.1(2a-2c), we fix $\beta = 0.8$, set $\mu_1 = 2$ and $\sigma_1^2 = 1$, and plot FDR, FNR and ATP as functions of h. In Figure 2.1(3a-3c), we fix $\beta = 0.8$ and $h = -2.5$, set $\sigma_1^2 = 1$, and plot FDR, FNR and ATP as functions of μ_1. The corresponding average proportions of the nulls, denoted by P_0, for each Ising model are given in Figure 2.1(1d-3d). The initial values for the numerical algorithm are set at $\beta^{(0)} = h^{(0)} = 0, \mu_1^{(0)} = \mu_1 + 1$ and $\sigma_1^{2(0)} = 2$.

From Figure 2.1(1a-3a), we can see that the FDR levels of all five procedures are controlled around 0.10 except one case of the LIS procedure in Figure 2.1(3a) with the lowest μ_1, whereas the BH and Lfdr procedures are generally conservative. This case of obvious deviation of the LIS procedure is likely caused by the small lattice size N. As a confirmation, additional simulations by increasing the lattice size N to 30×30×30 yield an FDR of 0.1019 for the same setup. From Figure 2.1(1b-3b) and (1c-3c) we can see that the two curves of OR and LIS procedures are almost identical, indicating that the data-driven LIS procedure works equally well as the OR procedure. These plots also show that the LIS procedure outperforms BH, q-value and Lfdr procedures with increased margin of performance in FNR and ATP as β or h increases or μ_1 is at a moderate level. Note that from (2.2) and (2.3), we can see that β controls how likely the same-state cases cluster together,

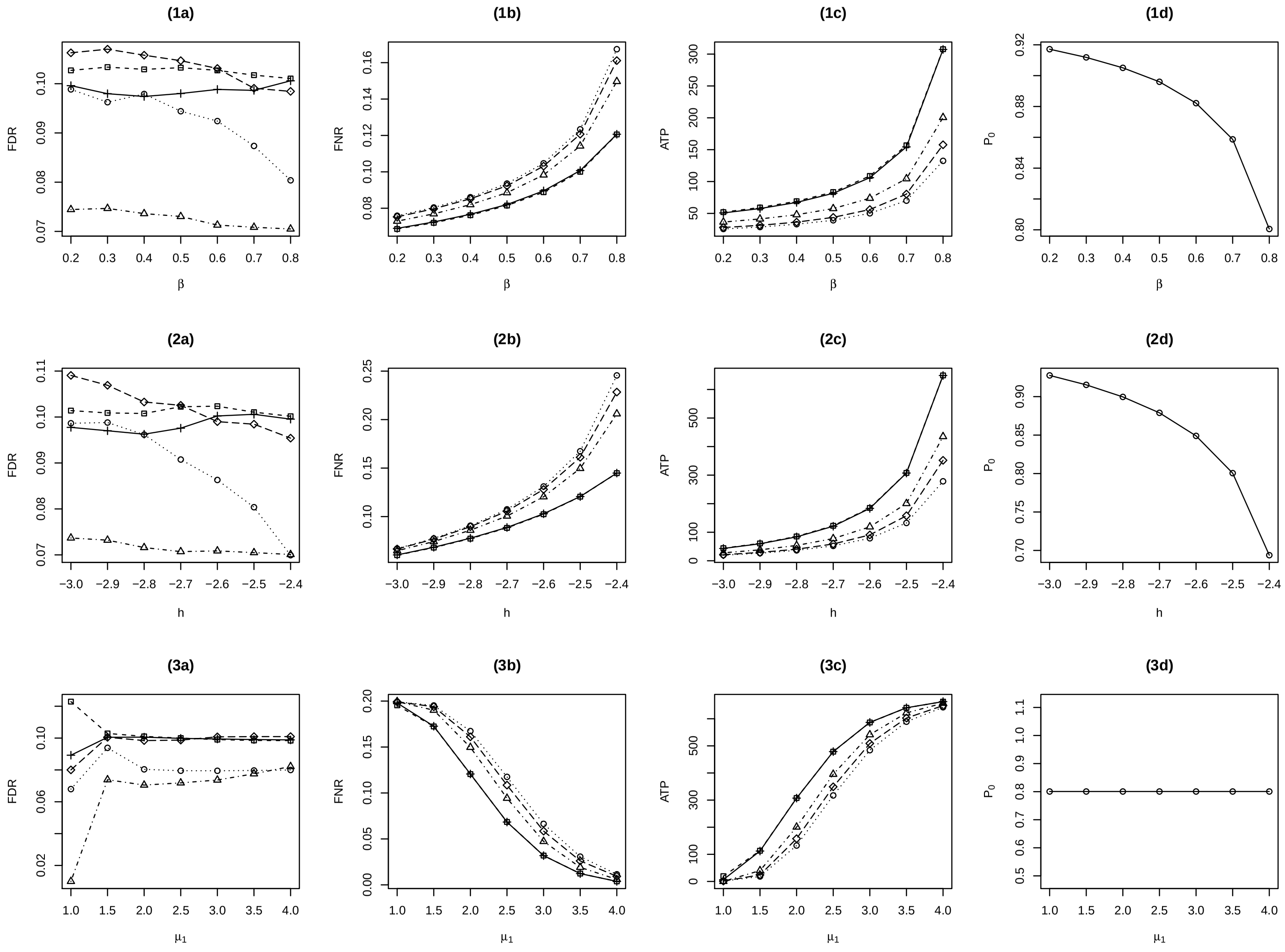


Figure 2.1: Comparison of BH (◯), q-value (◇), Lfdr (△), OR (+) and LIS (□) for a single group with $L = 1$.

and (β, h) together control the proportion of the aggregation of nonnulls relative to that of nulls.

Study 2: $L = 2$

We now consider the case where the nonnull distribution is a mixture of two normal distributions. The MRF is generated from the Ising model (2.1) with fixed parameters $\beta = 0.8$ and $h = -2.5$, and the nonnull distribution is a two-component normal mixture $p_1N(\mu_1, \sigma_1^2) + p_2N(\mu_2, \sigma_2^2)$ with fixed $p_1 = p_2 = 0.5$, $\mu_2 = 2$, and $\sigma_2^2 = 1$. In Figure 2.2(1a-1c), σ_1^2 varies from 0.125 to 8, and $\mu_1 = -2$. In Figure 2.2(2a-2c), we fix $\sigma_1^2 = 1$ and vary μ_1 from -4 to -1. The initial values are set at $\beta^{(0)} = h^{(0)} = 0$, $p_1^{(0)} = 1 - p_2^{(0)} = 0.3$, $\mu_l^{(0)} = \mu_l + 1$, and $\sigma_l^{2(0)} = \sigma_l^2 + 1, l = 1, 2$.

Similar to Figure 2.1, we can see that the FDR levels of all the procedures are controlled around 0.10, where BH and Lfdr are conservative, and OR and LIS perform similarly and outperform the other three procedures. In Figure 2.2(2a) at $\mu_1 = -1$, additional simulations yield an FDR of 0.1035 when the lattice size N is increased to $30\times30\times30$ for the same setup.

The results from both simulation studies are very similar to those in Sun and Cai (2009) for the one-dimensional case using HMC. It is clearly seen that, for dependent tests, incorporating dependence structure into a multiple-testing procedure improves efficiency dramatically.

Study 3: misspecified nonnull

Following Sun and Cai (2009), we consider the true nonnull distribution to be the three-component normal mixture $0.4N(\mu, 1) + 0.3N(1, 1) + 0.3N(3, 1)$, but use a misspecified two component normal mixture $p_1N(\mu_1, \sigma_1^2) + p_2N(\mu_2, \sigma_2^2)$ in the LIS procedure. The unobservable states are generated from the Ising model (2.1) with fixed parameters $\beta = 0.8$

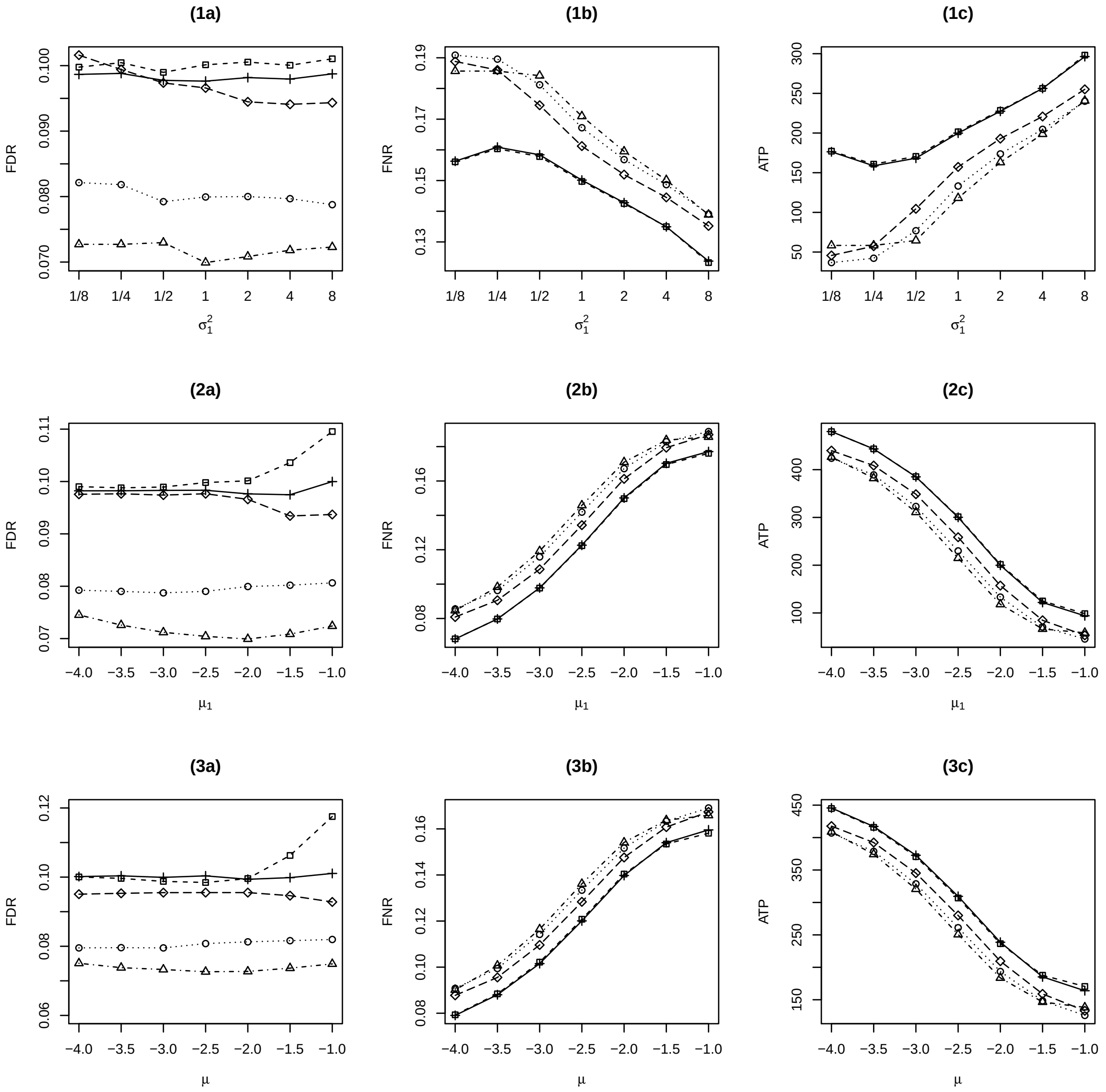


Figure 2.2: Comparison of BH (◯), q-value (◇), Lfdr (△), OR ($+$) and LIS (□) for a single group with $L = 2$ (see 1a-2c), and the one with L being misspecified (see 3a-3c).

and $h = -2.5$. The simulation results are displayed in Figure 2.2(3a-3c), the true μ varies from -4 to -1 with increments of size 0.5. The initial values are set at $\beta^{(0)} = h^{(0)} = 0$, $p_1^{(0)} = p_2^{(0)} = 0.5$, $\mu_1^{(0)} = -\mu_2^{(0)} = -2$, and $\sigma_l^{2(0)} = 2, l = 1, 2$.

Figure 2.2(3a-3c) shows that the LIS procedure performs similarly to OR under misspecified model. Additionally, the obvious biased FDR level by the LIS procedure at $\mu = -1$ reduces to 0.1067 when the lattice size N is increased to $30\times30\times30$.

2.4.2 Multiple-Group Analysis

Voxels in a human brain can be naturally grouped into multiple functional regions. For simulations with grouped multiple tests, we consider two lattice groups each with size $15\times15\times15$. The corresponding MRFs $\boldsymbol{\Theta}_1 = \{\Theta_{1s} : s \in S\}$ and $\boldsymbol{\Theta}_2 = \{\Theta_{2s} : s \in S\}$ are generated from the Ising model (2.1) with parameters $(\beta_1 = 0.2, h_1 = -1)$ and $(\beta_2 = 0.8, h_2 = -2.5)$, respectively. The observations $\boldsymbol{X}_k = \{X_{ks}, s \in S\}$ are generated conditionally on $\boldsymbol{\Theta}_k$, $k = 1, 2$, from $X_{ks}|\Theta_{ks} \sim (1 - \Theta_{ks})N(0, 1) + \Theta_{ks}N(\mu_k, \sigma_k^2)$, where μ_1 varies from 1 to 4 with increments of size 0.5, $\mu_2 = \mu_1 + 1$ and $\sigma_1^2 = \sigma_2^2 = 1$. The initial values are $\beta_1^{(0)} = \beta_2^{(0)} = h_1^{(0)} = h_2^{(0)} = 0$, $\mu_2^{(0)} = \mu_1^{(0)} = \mu_1 + 1$, and $\sigma_1^{2(0)} = \sigma_2^{2(0)} = 2$.

The simulation results are presented in Figure 2.3, which are similar to that in Wei et al. (2009) for the one-dimensional case with multiple groups using HMCs. Figure 2.3(a) shows that all procedures are valid in controlling FDR at the prespecified level of 0.10, whereas BH and CLfdr procedures are conservative. We also plot the within-group FDR levels of PLIS for each group separately. One can see that in order to minimize the global FNR level, the PLIS procedure may automatically adjust the FDRs of each individual group, either inflated or deflated reflecting the group heterogeneity, while the global FDR is appropriately controlled. In Figure 2.3(b) and (c) we can see that both SLIS and PLIS outperform BH, q-value and CLfdr procedures, indicating that utilizing the dependency information can improve the efficiency of a testing procedure, and the improvement is more

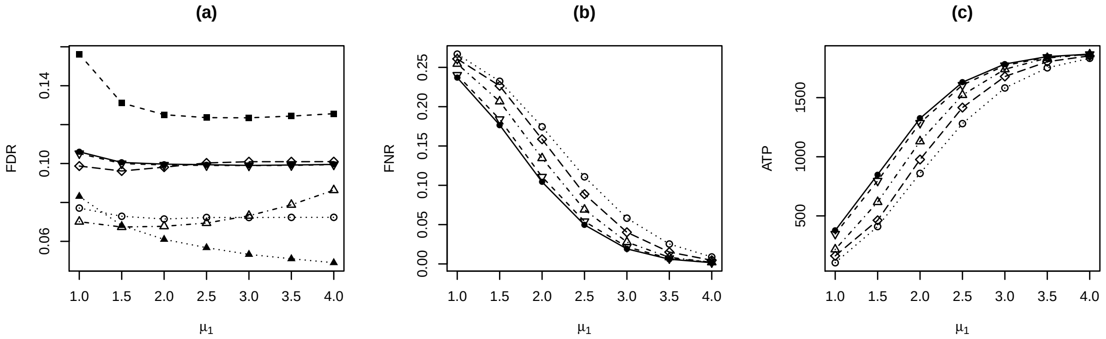


Figure 2.3: Comparison of BH ($\bigcirc$), q-value ($\diamond$), CLfdr ($\triangle$), SLIS ($\bigtriangledown$) and PLIS ($\bullet$) for two groups with $L = 1$. In (a), ■ and ▲ represent the results by PLIS for each individual group; for PLIS, while the global FDR is controlled, individual-group FDRs may vary.

evident for weaker signals (smaller values of μ_1). Between the two LIS-based procedures, PLIS slightly outperforms SLIS, indicating the benefit of ranking the LIS test statistics globally. In particular, ATP is 8.3% higher for PLIS than for SLIS when $\mu_1 = 1$.

2.5 ADNI FDG-PET Image Data Analysis

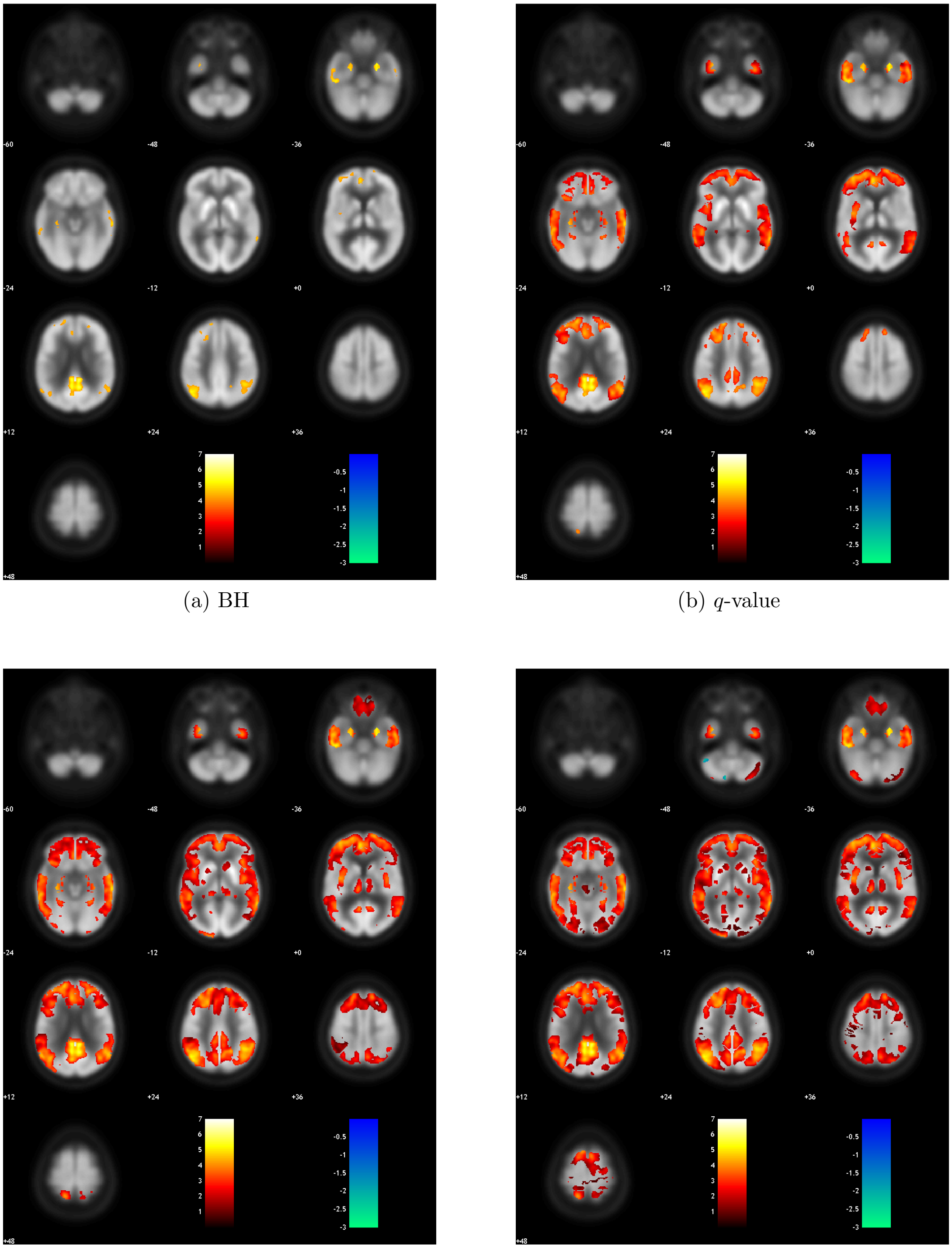
Alzheimer's disease (AD) is the most common cause of dementia in the elderly population. The worldwide prevalence of Alzheimer's disease was 26.6 million in 2006 and is predicted to be 1 in 85 persons by 2050 (Brookmeyer et al., 2007). Much progress has been made in the diagnosis of AD including clinical assessment and neuroimaging techniques. One such extensively used neuroimaging technique is FDG-PET imaging, which is used to evaluate the cerebral metabolic rate of glucose (CMRgl). Numerous FDG-PET studies (Nestor et al., 2003; Mosconi et al., 2005; Langbaum et al., 2009) have demonstrated significant reductions of CMRgl in brain regions in patients with AD and its prodromal stage mild cognitive impairment (MCI), compared with normal control (NC) subjects. These reduction can be used for the early detection of AD. Voxel-level multiple testing methods are common approaches to identify voxels with significant group differences in CMRgl (Alexander et al., 2002; Mosconi et al., 2005; Langbaum et al., 2009). We focus on the comparison between MCI and NC for such a purpose, and consider the FDG-PET image data from the ADNI database (adni.loni.usc.edu) as an illustrative example.

The data set consists of the baseline FDG-PET images of 102 NC subjects and 206 patients with MCI. Each image is normalized by the average of voxel values in pons and cerebellar vermis, which are well preserved regions in Alzheimer's patients. In human brain, the cerebral cortex is segregated into 43 Brodmann areas (BAs) based on the cytoarchitectural organization of neurons (Garey, 2006). We consider 30 of them after removing the BAs that are either too small or not always reliably registered. We also investigate 9 subcortical regions, including hippocampus, which are commonly considered in AD studies. A region is further divided into two if its bilateral parts in the left and right hemispheres are separated completely without a shared border in the middle of the brain. We have considered combining neighboring regions to potentially increase accuracy, but failed to find any pair with similar estimated HMRF model parameters. Finally, 61 regions of interest (ROIs) are included in the analysis, where the number of voxels in each region ranges from 149 to 20,680 with a median of 2,517. The total number of voxels of these 61 ROIs is $N = 251,500$. The goal is to identify voxels with reduced CMRgl in MCI patients comparing to NC.

We apply the HMRF-PLIS procedure to the ADNI data, and compare to BH, q-value and CLfdr procedures. We implement the BH procedure globally for the 61 ROIs, whereas we treat each region as a group for the q-value, CLfdr and PLIS procedures. For the BH and q-value procedures, a total number of N two-sample Welch's t-tests (Welch, 1947) are performed, and their corresponding two-sided p-values are obtained. For the PLIS and CLfdr procedures, z-values are used as the observed data $\boldsymbol{x}$, which are obtained from those t statistics by the transformation $z_i = \Phi^{-1}[G_0(t_i)]$, where Φ and G_0 are the cumulative distribution functions of the standard normal and the t statistic, respectively. The null distribution is assumed to be the standard normal distribution. The nonnull distribution is assumed to be a two-component normal mixture for PLIS. The LIS statistics in the PLIS procedure are approximated by 10^6 Gibbs-sampler samples, and the Lfdr statistics in the CLfdr procedure are computed by using the R code of Sun and Cai (2007). All the four testing procedures are controlled at a nominal FDR level of 0.001. In the GEM algorithm for HMRF estimation, the initial values for β and h in the Ising model are set to be zero. The initial values for the nonnull distributions are estimated from the signals claimed by BH at an FDR level of 0.1. The maximum number of GEM iterations is set to be 5,000 with $\epsilon_1 = \epsilon_2 = 0.001$, $\epsilon_3 = \alpha = 10^{-4}$, $a = 1$ and $b = 2$. For the Gibbs sampler embedded in the GEM, 5,000 samples are generated from 5,000 iterations after a burn-in period of 1,000 iterations. In this data analysis, the GEM algorithm reaches the maximum iteration and is then claimed to be converged for five ROIs. Among all 61 ROIs, the estimates of β have a median of 1.57 with the interquartile range of 0.36, and the estimates of h have a median of -3.71 with the interquartile range of 1.52. Such magnitude of parameter variation supports the multi-region analysis of the ADNI FDG-PET image data because even a 0.1 difference in β or h can result in quite different Ising models, see Figure 2.1(1d) and (2d).

Figure 2.4 shows the z-values (obtained by comparing CMRgl values between NC and MCI) of all the signals claimed by each procedure. Figure 2.5 summarizes the number of voxels that are claimed as signals by each procedure. We can see that PLIS finds the largest number of signals and covers 91.5%, 97.2% and 99.9% of signals detected by CLfdr, q-value and BH, respectively. It is interesting to see that the PLIS procedure finds more than 17 times signals as BH, twice as many signals as q-value, and about 20% more signals than the CLfdr procedure.

Detailed interpretations of the scientific findings are provided in Appendix A.3.



(a) BH

(b) q-value

(c) CLfdr

(d) PLIS

Figure 2.4: Z-values of the signals found by each procedure for the comparison between NC and MCI.

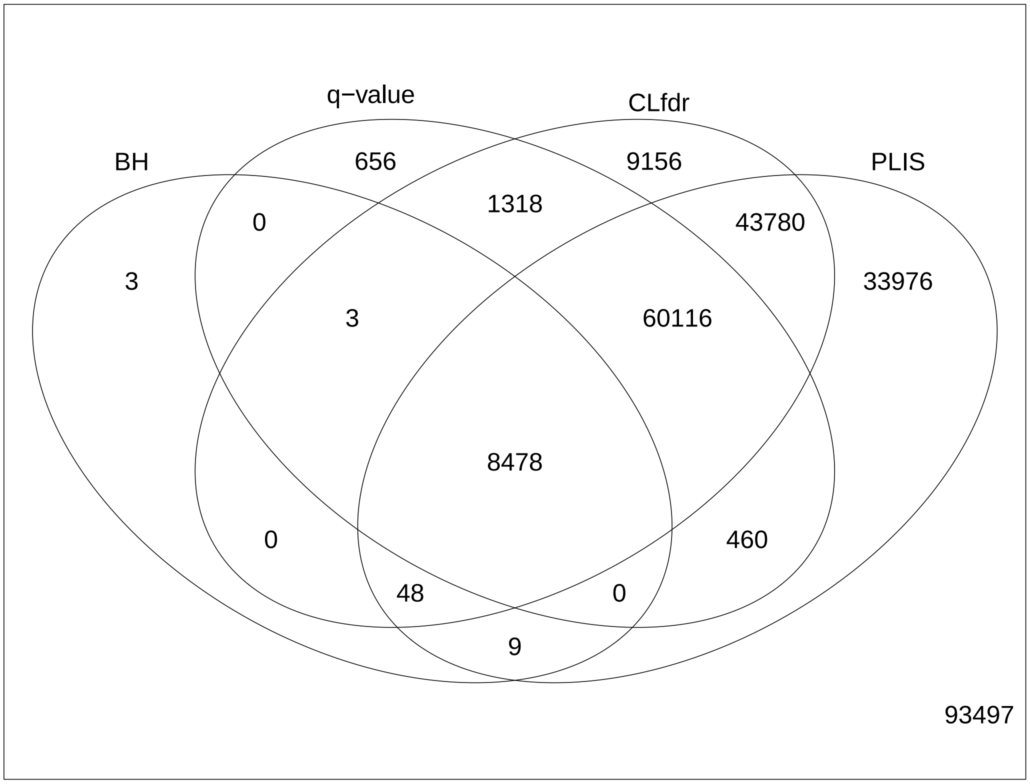


Figure 2.5: Venn diagram for the number of signals found by each procedure for the comparison between NC and MCI. Number of signals discovered by each procedure: BH=8,541, q-value=71,031, CLfdr=122,899, and PLIS=146,867.

CHAPTER III

Estimation of Large Covariance and Precision Matrices from Temporally Dependent Observations

3.1 Introduction

Let $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ be a sample of p-dimensional random vectors, each with the same mean $\boldsymbol{\mu}_p$, covariance matrix $\boldsymbol{\Sigma}$ and precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. It is well known that the sample covariance matrix is not a consistent estimator of $\boldsymbol{\Sigma}$ when p grows with n (Bai and Yin, 1993; Bai and Silverstein, 2010). When the sample observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent and identically distributed (i.i.d.), several regularization methods have been proposed for the consistent estimation of large $\boldsymbol{\Sigma}$, including thresholding (Bickel and Levina, 2008a; El Karoui, 2008; Rothman et al., 2009; Cai and Liu, 2011), block-thresholding (Cai and Yuan, 2012), banding (Bickel and Levina, 2008b) and tapering (Cai et al., 2010). Existing methods also include Cholesky-based method (Huang et al., 2006; Rothman et al., 2010), penalized pseudo-likelihood method (Lam and Fan, 2009) and sparse matrix transform (Cao et al., 2011). Consistent correlation matrix estimation can be obtained similarly from i.i.d. observations (Jiang, 2003; El Karoui, 2008).

The precision matrix $\boldsymbol{\Omega} = (\omega_{ij})_{p \times p}$, when it exists, is closely related to the partial correlations between the pairs of variables in a vector $\boldsymbol{X}$. Specifically, the partial correlation between X_i and X_j given $\{X_k, k \neq i, j\}$ is equal to $-\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}$ (Cramér, 1946, Section 23.4). Zero partial correlation means conditional independence between Gaussian or nonparanormal random variables (Liu et al., 2009). There is also a rich literature on the estimation of large $\boldsymbol{\Omega}$ from i.i.d. observations. Various algorithms for the ℓ_1 penalized maximum likelihood method (ℓ_1-MLE) and its variants have been developed by Yuan and Lin (2007), Banerjee et al. (2008), Friedman et al. (2008) and Hsieh et al. (2014), and related theoretical properties have been investigated by Rothman et al. (2008), Lam and Fan (2009) and Ravikumar et al. (2011). Methods that estimate $\boldsymbol{\Omega}$ column-by-column thus can be implemented with parallel computing include the nodewise Lasso (Meinshausen and Bühlmann, 2006; Van de Geer et al., 2014), graphical Dantzig selector (Yuan, 2010), constrained ℓ_1-minimization for inverse matrix estimation (CLIME; Cai et al., 2011), and adaptive CLIME (Cai et al., 2016).

Recently, researchers become increasingly interested in estimating the large covariance and precision matrices from temporally dependent observations $\{\boldsymbol{X}_t : t = 1, \ldots, n\}$, here t denotes time. Such research is particularly useful in analyzing the resting-state functional magnetic resonance imaging (rfMRI) data to assess the brain functional connectivity (Power et al., 2011; Ryali et al., 2012). In such imaging studies, the number of brain nodes (voxels or regions of interest) p can be greater than the number of images n. The temporal dependence of time series $\boldsymbol{X}_t$ is traditionally dealt with by imposing the so-called strong mixing conditions (Bradley, 2005). To overcome the difficulties in computing strong mixing coefficients and verifying strong mixing conditions, Wu (2005) introduced a new type of dependence measure, the functional dependence measure, and recently applied it to the hard thresholding estimator of large covariance matrix and the ℓ_1-MLE type methods of large precision matrix (Chen et al., 2013). But the functional dependence measure is still difficult to understand and to interpret. Practically, it is straightforward to describe the temporal dependence directly by using cross-correlations (Brockwell and Davis, 1991). By imposing certain weak dependence conditions directly on the cross-correlation matrix of samples $\{\boldsymbol{X}_t\}_{t=1}^n$, Bhattacharjee and Bose (2014) extended the banding and tapering regularization methods for covariance matrix. We consider a family of cross-correlation matrices with much weaker conditions that allow the time series to have long-range temporal dependence (also called long memory), which more reasonably describes, for example, the rfMRI data for brain connectivity studies.

A univariate stationary time series has polynomial decay temporal dependence if its autocorrelation $\rho(t) \sim Ct^{-\alpha}$ as $t \to \infty$ with some constants $C \neq 0$ and $\alpha > 0$. The notation $x_t \sim y_t$ means that $x_t/y_t \to 1$ as $t \to \infty$. This polynomial decay rate is much slower than the exponential rates in autoregressive models. We use a generalized form of such polynomial decay structure to the cross-correlation matrix of multivariate time series. Note that the temporal dependence with $\sum_{t=1}^{\infty} |\rho(t)| = \infty$ is called long memory (Palma, 2007), hence the polynomial decay processes with $0 < \alpha \leq 1$ have long memory. The weak temporal dependence considered by Bhattacharjee and Bose (2014) does not cover the polynomial decay processes with $0 < \alpha \leq 3$, and the short-range temporal dependence assumption of Chen et al. (2013) excludes the case with $0 < \alpha \leq 1$. Moreover, neither of their models covers the long memory processes. Later we argue that the rfMRI data do not meet their restrictive temporal dependence conditions, but well satisfy our model that allows any $\alpha > 0$ (see Figure 3.1(a)).

Note that the estimation of large correlation matrix was not considered by either Chen et al. (2013) or Bhattacharjee and Bose (2014), which is a more interesting problem in, for example, the study of brain functional connectivity. Moreover, they all assumed that $\boldsymbol{\mu}_p = (\mu_{pi})_{1\leq i\leq p}$ is known. But $\boldsymbol{\mu}_p$ is often unknown in practice and needs to be estimated. Although the sample mean $\bar{X}_i = n^{-1}\sum_{j=1}^n X_{ij}$ entrywise converges to μ_{pi} in probability or even almost surely under some dependence conditions (Brockwell and Davis, 1991; Hu et al., 2008), extra care will still be needed when true mean is replaced by sample mean in the estimation of covariance. We consider unknown $\boldsymbol{\mu}_p$ in this chapter. Also note that the estimation of large correlation matrix and its inverse is considered in a recent work by Zhou (2014). However, her method requires that all p time series have the same temporal decay rate, which is rather restrictive and often violated (see Figure 3.1(b) for an example of rfMRI data).

In this chapter, we study the generalized thresholding estimation (Rothman et al., 2009) for covariance and correlation matrices, and the CLIME approach (Cai et al., 2011) and an ℓ_1-MLE type method called SPICE–sparse permutation invariant covariance estimation (Rothman et al., 2008) for precision matrix. The theoretical results of convergence rates, sparsistency and sign-consistency are provided for temporally dependent data, potentially with long memory, which are generated from a class of sub-Gaussian distributions including Gaussian distribution as a special case. A gap-block cross-validation method is proposed for the tuning parameter selection, which shows satisfactory performance for temporally dependent data in simulations. To the best of our knowledge, this is the first work that investigates the estimation of large covariance and precision matrices for temporal data with long memory.

The chapter is organized as follows. In Section 3.2, we introduce a polynomial-decay-dominated model for the temporal dependence, and show that it best describes the rfMRI data comparing to the existing literature (Chen et al., 2013; Bhattacharjee and Bose, 2014; Zhou, 2014). We also introduce the considered sub-Gaussian data generating mechanism. We provide the theoretical results for the estimation of covariance and correlation matrices in Section 3.3 and of precision matrix in Section 3.4 under the considered temporal dependence. In Section 3.5, we introduce a gap-block cross-validation method for the tuning parameter selection, evaluate the estimating performance via simulations, and analyze a rfMRI data set for brain functional connectivity. The proofs of theoretical results are sketched in Section 3.6, with detailed proofs provided in Appendix B.

3.2 Temporal Dependence

We start with a brief introduction of useful notation. For a real matrix $\mathbf{M} = (M_{ij})$, we use the following notation for different norms, see, e.g., Golub and Van Loan (1996):

- spectral norm $\|\mathbf{M}\|_2 = \sqrt{\varphi_{\max}(\mathbf{M}^T\mathbf{M})}$, where $\varphi_{\max}$ denotes the largest eigenvalue, also $\varphi_{\min}$ denotes the smallest eigenvalue;
- Frobenius norm $\|\mathbf{M}\|_F = \sqrt{\sum_i \sum_j M_{ij}^2}$;
- matrix ℓ_1 norm $\|\mathbf{M}\|_1 = \max_j \sum_i |M_{ij}|$;
- elementwise ℓ_1 norm $|\mathbf{M}|_1 = \sum_{i,j} |M_{ij}|$;
- off-diagonal elementwise ℓ_1 norm $|\mathbf{M}|_{1,\text{off}} = \sum_{i \neq j} |M_{ij}|$;
- elementwise ℓ_∞ norm (a.k.a. max norm) $|\mathbf{M}|_\infty = \max_{i,j} |M_{ij}|$.

Define vec$(\mathbf{M})$ = vec$\{M_{ij} : \forall\, i, j\} = \left(\boldsymbol{M}_1^T, \boldsymbol{M}_2^T, \ldots, \boldsymbol{M}_n^T\right)^T$, where $\boldsymbol{M}_j$ is the j-th column of $\mathbf{M}$. Write $\mathbf{M} \succ 0$ when $\mathbf{M}$ is positive definite. Denote the trace and the determinant of a square matrix $\mathbf{M}$ by tr$(\mathbf{M})$ and det$(\mathbf{M})$, respectively. Denote the Kronecker product by $\otimes$. Write $x_n \asymp y_n$ if $x_n = O(y_n)$ and $y_n = O(x_n)$. Define $\lceil x \rceil$ and $\lfloor x \rfloor$ to be the smallest integer $\geq x$ and the largest integer $\leq x$, respectively. Let $\mathbb{I}(A)$ be the indicator function of event A, $(x)_+ = x\mathbb{I}(x \geq 0)$ and sign$(x) = \mathbb{I}(x \geq 0) - \mathbb{I}(x \leq 0)$. Let $A \coloneqq B$ denote that A is defined to be B. Denote $X \overset{d}{=} Y$ if X and Y have the same distribution. Denote $\mathbf{1}_n = (1, 1, \ldots, 1)^T$ with length n and $\mathbf{I}_{n \times n}$ to be the $n \times n$ identity matrix. If without further notification, a constant is independent of n and p. Throughout the rest of the chapter, we assume $p \to \infty$ as $n \to \infty$ and only use $n \to \infty$ in the asymptotic arguments.

3.2.1 Polynomial-Decay-Dominated (PDD) Temporal Dependence

Let $\mathbf{X}_{p\times n} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$ be the data matrix with the covariance matrix $\boldsymbol{\Sigma} = (\sigma_{kl})_{p\times p}$ for each $\boldsymbol{X}_i$. Let $\mathbf{R} = (\rho_{kl})_{p\times p}$ be the correlation matrix for each $\boldsymbol{X}_i$ and $\mathbf{R}^{ij} = (\rho_{kl}^{ij})_{p\times p} = (\mathrm{cov}(X_{ki}, X_{lj})/\sqrt{\sigma_{kk}\sigma_{ll}})_{p\times p}$ be the cross-correlation matrix between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. Clearly, $\mathbf{R} = \mathbf{R}^{ij}$ when $i = j$. We say that $\mathbf{X}_{p\times n}$ has a PDD temporal dependence if its cross-correlation matrices $\{\mathbf{R}^{ij}\}$ belong to

$$\mathcal{B}(C_0, \alpha) = \big\{\{\mathbf{R}^{ij}\} : |\mathbf{R}^{ij}|_\infty \le C_0|i-j|^{-\alpha} \text{ for any } i \ne j\big\} \tag{3.1}$$

with some positive constants C_0 and α. This model allows an individual time series to have the polynomial decay temporal dependence, which is long memory when $0 < \alpha \le 1$. Note that for i.i.d. observations we have $\alpha = \infty$. Our goal is to estimate $\boldsymbol{\Sigma}, \mathbf{R}$ and $\boldsymbol{\Omega}$ while treating $\{\mathbf{R}^{ij}\}_{i\ne j}$ as nuisance parameters that need not be estimated.

3.2.2 Comparisons to Existing Models

For banding and tapering estimators of $\boldsymbol{\Sigma}$, Bhattacharjee and Bose (2014) considered the following weak dependence based on temporal distance. For any $n \ge 1$,

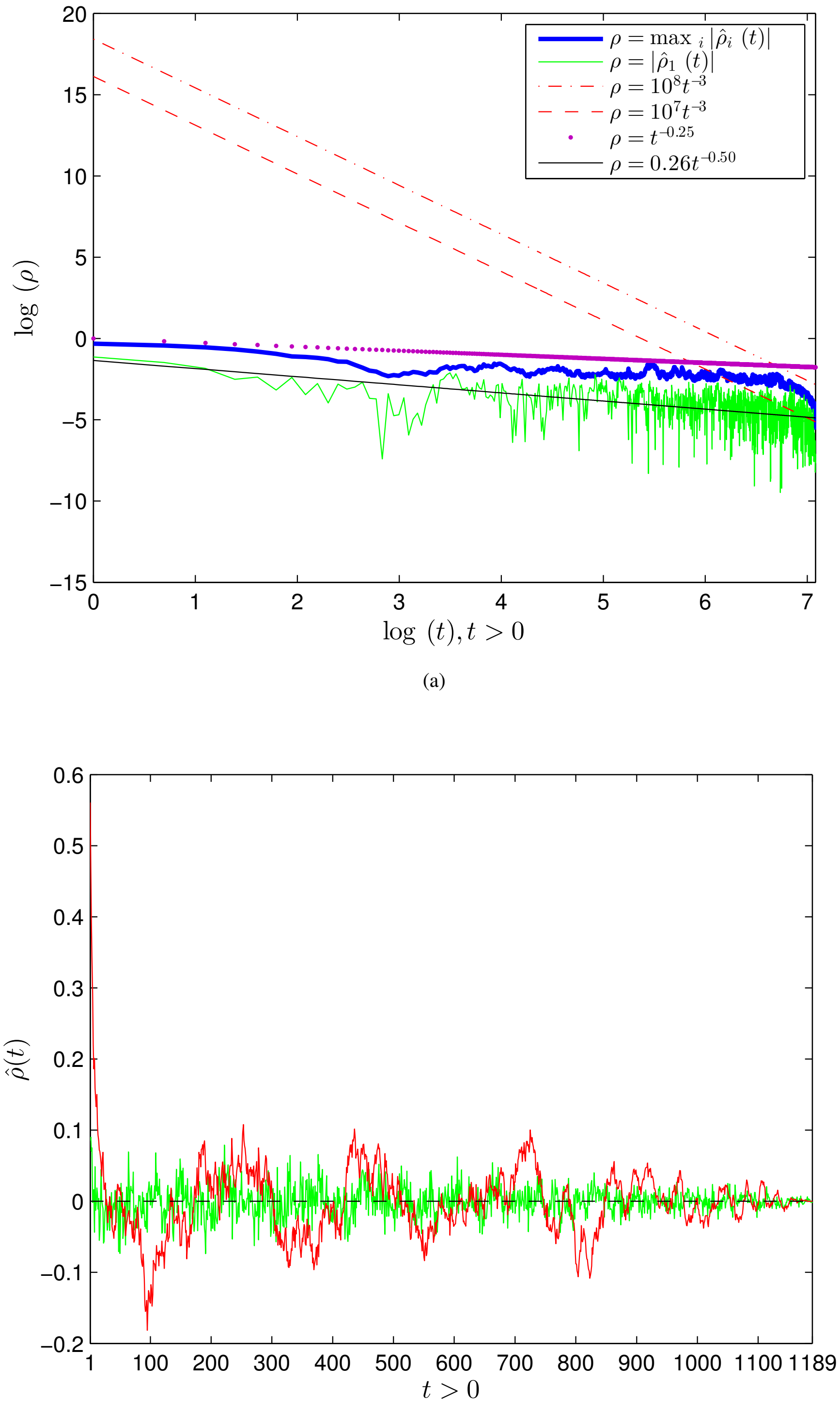
$$\mathcal{A}_n(a_n) = \left\{\{\mathbf{R}^{ij}\} : \max_{a_n\le|i-j|\le n} |\boldsymbol{\Theta}^{ij}|_\infty = O(n^{-2}a_n)\right\},$$

where $\boldsymbol{\Theta}^{ij} = (\theta_{kl}^{ij})_{p\times p}$ with θ_{kl}^{ij} satisfying $\rho_{kl}^{ij} = \theta_{kl}^{ij}\rho_{kl}$, $a_n\sqrt{\log p/n} = o(1)$ and $\{a_n\}_{n\ge1}$ is a non-decreasing sequence of non-negative integers. That $a_n\sqrt{\log p/n} = o(1)$ implies $a_n = o(\sqrt{n})$. Thus, $\big|\boldsymbol{\Theta}^{ij} : |i-j| = a_n\big|_\infty \le \max_{a_n\le|i-j|\le n} |\boldsymbol{\Theta}^{ij}|_\infty = O(n^{-2}a_n) = o(a_n^{-3})$. Then $\sum_{|i-j|=1}^{\infty} |\boldsymbol{\Theta}^{ij}|_\infty < \infty$, which means that their model does not allow any individual time series to be a long memory process. Moreover, $\{\mathbf{R}^{ij}\}$ in model (3.1) is not in the above $\mathcal{A}_n(a_n)$ when $0 < \alpha \le 3$ and $|\boldsymbol{\Theta}^{ij}|_\infty \asymp |i-j|^{-\alpha}$ for any $i \ne j$.

Chen et al. (2013) considered the hard thresholding estimation of $\boldsymbol{\Sigma}$ and an ℓ_1-MLE type estimation of $\boldsymbol{\Omega}$ using the functional dependence measure of Wu (2005). Assume that $\{X_{1t}\}$, the first row of $\mathbf{X}_{p\times n}$, is a stationary process with autocovariance $\gamma_1(t)$, and follow their setup by letting $E(X_{1t}) = 0$, then $\gamma_1(t) = E(X_{11}X_{1,t+1})$. By the argument in the proof of Theorem 1 in Wu and Pourahmadi (2009) together with Lyapunov's inequality (Karr, 1993) and Theorem 1 of Wu (2005), one can see that their model requires $\sum_{t=0}^{\infty} |\gamma_1(t)| < \infty$, which means $\{X_{1t}\}$ cannot be a long memory process. Hence their model does not cover model (3.1) when $0 < \alpha \leq 1$.

Zhou (2014) was interested in estimating a separable covariance $\text{cov}(\boldsymbol{X}_{pn}) = \mathbf{A} \otimes \mathbf{B}$, where $\boldsymbol{X}_{pn} := \text{vec}(\mathbf{X}_{p\times n})$. Her model implies that the autocorrelations $\{\rho_{kk}^{ij}\}_{1\leq i,j\leq n}$ are the same for all k, indicating a rather restrictive model with homogeneous decay rate for all p time series.

Now consider the rfMRI data example of a single subject which will be further analyzed in Subsection 3.5.3. The data set consists of 1190 temporal brain images. We consider 907 functional brain nodes in each image. All node time series have passed the Priestley-Subba Rao test for stationarity (Priestley and Subba Rao, 1969) with a significance level of 0.05 for p-values adjusted by the false discovery rate controlling procedure of Benjamini and Yekutieli (2001). Hence the autocorrelations $\{\rho_{kk}^{ij}\}$ can be approximated by tha sample autocorrelations $\{\hat{\rho}_k(t)\}$ for each k. To save computational cost, we only plot the autocorrelations in Figure 3.1. One may make a mild assumption that the cross-correlations are dominated by the autocorrelations in the sense that $|\rho_{kl}^{ij}| \leq C|\rho_{kk}^{ij}|$ for a fixed constant $C > 0$, thus only need to check the autocorrelations in practice. Figure 3.1(a) shows that $\max_{1\leq i\leq p} |\hat{\rho}_i(t)|$ can be bounded by $10^8 t^{-3}$, but not by $10^7 t^{-3}$. Thus the temporal dependence assumption of Bhattacharjee and Bose (2014) does not seem to fit the data well. For a randomly selected brain node, the least squares fitting for a log-linear model yields $|\hat{\rho}_1(t)| = 0.26t^{-0.50}$, thus the applicability of Chen et al. (2013) is in question. Figure 3.1(b) illustrates the estimated autocorrelations for two randomly se-



(a)

(b)

Figure 3.1: Sample autocorrelations of brain nodes.

lected brain nodes, which clearly have different patterns, indicating that the assumption of homogeneous decay rates for all time series in Zhou (2014) does not hold. On the other hand, Figure 3.1(a) shows that the rfMRI data have the PDD structure with $\hat{\alpha} = 0.25$ since $\max_{1\leq i\leq p} |\hat{\rho}_i(t)| \leq t^{-0.25}$, assuming the cross-correlations are dominated by the autocorrelations.

3.2.3 Sub-Gaussian Data

A random variable Z is called sub-Gaussian if there exists a constant $K \in [0, \infty)$ such that

$$E(\exp\{t[Z - E(Z)]\}) \leq \exp\left\{Kt^2/2\right\}, \text{ for all } t \in \mathbb{R}. \tag{3.2}$$

It can be shown that $K \geq \text{var}(Z)$ (Buldygin and Kozachenko, 2000, Lemma 1.2). We simply call K the parameter of the sub-Gaussian distribution of Z, and call Z standard sub-Gaussian if $E(Z) = 0$ and $\text{var}(Z) = 1$.

Throughout the chapter, we assume that the vectorized data are obtained from the following data generating mechanism

$$\boldsymbol{X}_{pn} = \mathbf{H}\boldsymbol{e} + \boldsymbol{\mu}_{pn}, \tag{3.3}$$

where $\mathbf{H} = (h_{ij})_{pn\times m}$ is a real deterministic matrix, $\boldsymbol{\mu}_{pn} = \mathbf{1}_n \otimes \boldsymbol{\mu}_p$, and the random vector $\boldsymbol{e} = (e_1, \ldots, e_m)^T$ consists of m independent standard sub-Gaussian components with the same parameter $K \geq 1$. We allow $m = \infty$ by requiring that for each i, $\sum_{j=1}^{m} h_{ij}e_j$ converges both almost surely and in mean square when $m \to \infty$. A sufficient and necessary condition for both modes of convergence is $\sum_{j=1}^{\infty} h_{ij}^2 < \infty$ for every i, see Theorem 8.3.4 and its proof in Athreya and Lahiri (2006). Under these two modes of convergence, it can be shown that $E(\mathbf{H}\boldsymbol{e}) = \mathbf{H}E(\boldsymbol{e})$ and $\text{cov}(\mathbf{H}\boldsymbol{e}) = \mathbf{H}\text{cov}(\boldsymbol{e})\mathbf{H}^T$ (Brockwell and Davis, 1991, Proposition 2.7.1). Hence, for either finite or infinite m, we have $E(\boldsymbol{X}_{pn}) = \boldsymbol{\mu}_{pn}$,

$\text{cov}(\boldsymbol{X}_{pn}) = \mathbf{H}\mathbf{H}^T$ with all n submatrices of dimension $p \times p$ on the diagonal equal to $\boldsymbol{\Sigma}$ and temporal correlations, particularly those in (3.1), determined by the off-diagonal submatrices, and moreover,

$$E(\exp\{t[X_{ij} - E(X_{ij})]\}) \leq \exp\{K\sigma_{ii}t^2/2\}, \text{ for all } t \in \mathbb{R}, \tag{3.4}$$

which follows from Fatou's Lemma for $m = \infty$. The advantage of allowing $m = \infty$ is that any case with finite m becomes a special example by adding infinite number of columns of zeros in $\mathbf{H}$. In filtering theory, matrix $\mathbf{H}$ is said to be a linear spatio-temporal coloring filter (Fomin, 1999; Manolakis et al., 2005), which generates the output $\boldsymbol{X}_{pn}$ by introducing both spatial and temporal dependence in the input independent variables $e_1, \ldots, e_m$.

The following are two examples of (3.3) which are widely studied in the literature.

Example III.1 (Gaussian data)**.** Assume that $\boldsymbol{X}_{pn}$ has a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_{pn}, \boldsymbol{\Delta})$. Then $\boldsymbol{\Delta} = \mathbf{H}\mathbf{H}$ with a symmetric real matrix $\mathbf{H}$. If $\boldsymbol{\Delta} \succ 0$, then $\boldsymbol{X}_{pn} = \mathbf{H}\boldsymbol{e} + \boldsymbol{\mu}_{pn}$ with $\boldsymbol{e} = \mathbf{H}^{-1}(\boldsymbol{X}_{pn} - \boldsymbol{\mu}_{pn}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{pn \times pn})$. If $\boldsymbol{\Delta}$ is singular, then $\boldsymbol{X}_{pn}$ has a degenerate multivariate Gaussian distribution, and can be expressed as $\boldsymbol{X}_{pn} \overset{d}{=} \mathbf{H}\boldsymbol{e} + \boldsymbol{\mu}_{pn}$ with any $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{pn \times pn})$. In fact, replacing "$=$" in (3.3) by "$\overset{d}{=}$" does not affect the theoretical results.

Example III.2 (Moving average processes)**.** Consider the following vector moving average processes

$$\boldsymbol{X}_j = \sum_{l=0}^{L} \mathbf{B}_l \boldsymbol{e}_{j-l}, \quad \text{with} \quad 0 \leq L \leq \infty, \tag{3.5}$$

where the case with $L = \infty$ is well-defined in the sense of entrywise almost-sure convergence and mean-square convergence, $\{\mathbf{B}_l\}$ are $p \times p$ real deterministic matrices, $\boldsymbol{e}_j = (e_{1j}, e_{2j} \ldots, e_{pj})^T$ with $\{e_{st} : 1 \leq s \leq p, -\infty \leq t \leq n\}$ being independent standard sub-Gaussian random variables with the same parameter $K \geq 1$. Since every X_{ij} is a linear combination of $\{e_{st}\}$, we always can find a matrix $\mathbf{H}$ such that $\boldsymbol{X}_{pn} = \mathbf{H}\boldsymbol{e}$ with $\boldsymbol{e} = (\boldsymbol{e}_{1-L}^T, \boldsymbol{e}_{2-L}^T, \ldots, \boldsymbol{e}_n^T)^T$. It is well-known that any causal vector autoregressive moving average process of the form

$$\boldsymbol{X}_j - \mathbf{A}_1\boldsymbol{X}_{j-1} - \cdots - \mathbf{A}_a\boldsymbol{X}_{j-a} = \boldsymbol{e}_j + \mathbf{M}_1\boldsymbol{e}_{j-1} + \cdots + \mathbf{M}_b\boldsymbol{e}_{j-b}$$

with finite nonnegative integers a and b, and real deterministic matrices $\{\mathbf{A}_i, \mathbf{M}_k\}$, can be written in the form of (3.5) with $L = \infty$ (Brockwell and Davis (1991), pp. 418).

3.3 Estimation of Covariance and Correlation Matrices

Consider the set of ℓ_q-ball sparse covariance matrices (Bickel and Levina, 2008a; Rothman et al., 2009)

$$\text{(3.6)} \qquad \mathcal{U}(q, c_p, v_0) = \left\{ \mathbf{\Sigma} : \max_{1\leq i\leq p} \sum_{j=1}^{p} |\sigma_{ij}|^q \leq c_p, \max_{1\leq i\leq p} \sigma_{ii} \leq v_0 \right\},$$

and the corresponding set of correlation matrices

$$\text{(3.7)} \qquad \mathcal{R}(q, c_p) = \left\{ \mathbf{R} : \max_{1\leq i\leq p} \sum_{j=1}^{p} |\rho_{ij}|^q \leq c_p \right\},$$

where constants $v_0 > 0$ and $0 \leq q < 1$. For any thresholding parameter $\tau \geq 0$, define a generalized thresholding function (Rothman et al., 2009) by $s_\tau : \mathbb{R} \to \mathbb{R}$ satisfying the following conditions for all $z \in \mathbb{R}$: (i) $|s_\tau(z)| \leq |z|$; (ii) $s_\tau(z) = 0$ for $|z| \leq \tau$; (iii) $|s_\tau(z) - z| \leq \tau$. Such defined generalized thresholding function covers many widely used thresholding functions, including hard thresholding $s_\tau^H(z) = z\mathbb{I}(|z| > \tau)$, soft thresholding $s_\tau^S(z) = \text{sign}(z)(|z| - \tau)_+$, smoothly clipped absolute deviation and adaptive lasso thresholdings. See details about these examples in Rothman et al. (2009). We define the generalized thresholding estimators of $\mathbf{\Sigma}$ and $\mathbf{R}$ respectively by

$$S_\tau(\hat{\mathbf{\Sigma}}) = (s_\tau(\hat{\sigma}_{ij}))_{p\times p} \qquad \text{and} \qquad S_\tau(\hat{\mathbf{R}}) = (s_\tau(\hat{\rho}_{ij})\mathbb{I}(i \neq j) + \mathbb{I}(i = j))_{p\times p},$$

where $\hat{\boldsymbol{\Sigma}} := (\hat{\sigma}_{ij})_{p\times p}$ is the sample covariance matrix defined by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T - \bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T \tag{3.8}$$

with $\bar{\boldsymbol{X}} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i$, and $\hat{\mathbf{R}} := (\hat{\rho}_{ij})_{p\times p} = \left(\hat{\sigma}_{ij}/\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}\right)_{p\times p}$ is the sample correlation matrix. Then we have the following results.

Theorem III.1. *Uniformly on $\mathcal{U}(q, c_p, v_0)$ and $\mathcal{B}(C_0, \alpha)$, for sufficiently large constant $M > 0$, if $\tau = M\tau'$ and $\tau' = o(1)$ with*

$$\tau' := \sqrt{f_0 \log(pf_0)/n} \quad \textit{and} \quad f_0 := \begin{cases} 3C_0(n^{1-\alpha} - \alpha)/(1-\alpha), & 0 < \alpha < 1, \\ 3C_0(1 + \log n), & \alpha = 1, \\ [3C_0(n^{1-\alpha} - \alpha)/(1-\alpha)]^{1/\alpha}, & \alpha > 1, \end{cases} \tag{3.9}$$

then

$$|S_\tau(\hat{\boldsymbol{\Sigma}}) - \boldsymbol{\Sigma}|_\infty = O_P(\tau'),$$

$$\|S_\tau(\hat{\boldsymbol{\Sigma}}) - \boldsymbol{\Sigma}\|_2 = O_P\left(c_p \tau'^{1-q}\right), \tag{3.10}$$

$$\frac{1}{p}\|S_\tau(\hat{\boldsymbol{\Sigma}}) - \boldsymbol{\Sigma}\|_F^2 = O_P\left(c_p \tau'^{2-q}\right). \tag{3.11}$$

Moreover, if $p \geq n^c$ for some constant $c > 0$, then

$$E\left(|S_\tau(\hat{\boldsymbol{\Sigma}}) - \boldsymbol{\Sigma}|_\infty^2\right) = O(\tau'^2),$$

$$E\left(\|S_\tau(\hat{\boldsymbol{\Sigma}}) - \boldsymbol{\Sigma}\|_2^2\right) = O\left(c_p^2 \tau'^{2-2q}\right), \tag{3.12}$$

$$\frac{1}{p}E\left(\|S_\tau(\hat{\boldsymbol{\Sigma}}) - \boldsymbol{\Sigma}\|_F^2\right) = O\left(c_p \tau'^{2-q}\right). \tag{3.13}$$

Remark III.1. The constant 3 in f_0 is chosen for simplicity, which can be replaced by any arbitrary constant greater than 2. It is easily seen that τ' is continuous for $\alpha > 0$, and is monotonically decreasing as α increases, i.e., the temporal dependence decreases.

Treating α as a fixed value, we have

$$\tau' \asymp \begin{cases} n^{-\alpha/2}(\log p + \log n)^{1/2}, & 0 < \alpha < 1, \\ n^{-1/2}[(\log n)(\log p + \log\log n)]^{1/2}, & \alpha = 1, \\ n^{-1/2}(\log p)^{1/2}, & \alpha > 1, \end{cases} \tag{3.14}$$

which can be further simplified to

$$\tau' \asymp \begin{cases} n^{-\alpha/2}(\log p)^{1/2}, & 0 < \alpha < 1, \\ n^{-1/2}[(\log n)(\log p)]^{1/2}, & \alpha = 1, \\ n^{-1/2}(\log p)^{1/2}, & \alpha > 1. \end{cases} \tag{3.15}$$

when $p \geq n^c$ with some constant $c > 0$. Thus, for covariance matrix estimation, the rates of convergence in probability given in (3.10) and (3.11) under PDD temporal dependence with fixed $\alpha > 1$ are the same as those under i.i.d. observations given in Bickel and Levina (2008a) and Rothman et al. (2009). The same rates of convergence in probability are also obtained by Basu et al. (2015, Proposition 5.1) for certain short-memory stationary Gaussian data using the hard thresholding method. Moreover, following Cai and Zhou (2012) under the condition that $p \geq n^{c_1}$ and $c_p \leq c_2 n^{(1-q)/2}(\log p)^{-(3-q)/2}$ with some constants $c_1 > 1$ and $c_2 > 0$, it can be shown that the convergence rates in mean-squared norms given in (3.12) and (3.13) for the case with fixed $\alpha > 1$ are minimax optimal, which are the same as the optimal minimax rates for the i.i.d. case.

Theorem III.2 (Sparsistency and sign-consistency)**.** *Under the conditions for the convergence in probability given in Theorem III.1, we have $s_\tau(\hat{\sigma}_{ij}) = 0$ for all (i, j) where $\sigma_{ij} = 0$ with probability tending to 1. If additionally assume that all nonzero elements of $\mathbf{\Sigma}$ satisfy $|\sigma_{ij}| \geq 2\tau$, we then have $\operatorname{sign}(s_\tau(\hat{\sigma}_{ij})) = \operatorname{sign}(\sigma_{ij})$ for all (i, j) where $\sigma_{ij} \neq 0$ with probability tending to 1.*

Corollary III.1. *Theorems III.1 and III.2 hold with* $\hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}, \hat{\sigma}_{ij}, \sigma_{ij}$ *and* $\mathcal{U}(q, c_p, v_0)$ *replaced by* $\hat{\mathbf{R}}, \mathbf{R}, \hat{\rho}_{ij}, \rho_{ij}$ *and* $\mathcal{R}(q, c_p)$*, respectively.*

3.4 Estimation of Precision Matrix

We consider both the CLIME and the SPICE methods for the estimation of $\boldsymbol{\Omega}$, which originally were developed for i.i.d. data.

3.4.1 CLIME Estimation

Following Cai et al. (2011), we consider the following set of precision matrices

$$\mathcal{G}_1(q, c_p, M_p, v_0) = \Big\{ \boldsymbol{\Omega} \succ 0 : \max_{1 \le i \le p} \sum_{j=1}^{p} |\omega_{ij}|^q \le c_p, \|\boldsymbol{\Omega}\|_1 \le M_p,$$

$$\max_{1 \le i \le p} \{\sigma_{ii}, \omega_{ii}\} \le v_0 \Big\}, \tag{3.16}$$

where constants $0 \le q < 1$, $v_0 > 1$, and c_p and M_p are allowed to depend on p. We also assume $\min\{c_p, M_p\} > 1$ for simplicity because it can be shown that $\min\{c_p, M_p\}$ has a positive constant lower bound. The original set considered in Cai et al. (2011) does not contain the condition $\max_i\{\sigma_{ii}, \omega_{ii}\} \le v_0$. But their moment conditions on $\boldsymbol{X}$ (see their (C1) and (C2)) implies $\max_i\{\sigma_{ii}\} \le v_0$. The additional condition $\max_i\{\omega_{ii}\} \le v_0$ facilitates the proof of consistency for the temporally dependent observations, which is easily obtained from the widely used assumption $\varphi_{\max}(\boldsymbol{\Omega}) \le v_0$ (Rothman et al., 2008; Lam and Fan, 2009). Note that the above $\mathcal{G}_1$ contains ℓ_q-ball sparse matrices such as those with exponentially decaying entries from the diagonal, for example, AR(1) matrices. For an invertible band matrix $\boldsymbol{\Sigma}$, its inverse matrix $\boldsymbol{\Omega}$ generally has exponentially decaying entries from the diagonal (Demko et al., 1984).

Let $\hat{\boldsymbol{\Omega}}_\varepsilon^\star$ be a solution of the following optimization problem:

$$\min |\boldsymbol{\Omega}_\varepsilon|_1 \quad \text{subject to} \quad |\tilde{\boldsymbol{\Sigma}}_\varepsilon \boldsymbol{\Omega}_\varepsilon - \mathbf{I}_{p\times p}|_\infty \le \lambda_1, \quad \boldsymbol{\Omega}_\varepsilon \in \mathbb{R}^{p\times p}, \tag{3.17}$$

where $\tilde{\mathbf{\Sigma}}_\varepsilon = \hat{\mathbf{\Sigma}} + \varepsilon \mathbf{I}_{p\times p}$, $\hat{\mathbf{\Sigma}}$ is given in (3.8), $\varepsilon \geq 0$ is a perturbation parameter introduced for the same reasons given in Cai et al. (2011) and can be set to be $n^{-1/2}$ in practice (see Remark III.3 below), and λ_1 is a tuning parameter. The CLIME estimator $\hat{\mathbf{\Omega}}_\varepsilon := (\hat{\omega}_{ij\varepsilon})_{p\times p}$ is then obtained by symmetrizing $\hat{\mathbf{\Omega}}_\varepsilon^\star := (\hat{\omega}_{ij\varepsilon}^\star)_{p\times p}$ with

$$\hat{\omega}_{ij\varepsilon} = \hat{\omega}_{ji\varepsilon} = \hat{\omega}_{ij\varepsilon}^\star \mathbb{I}(|\hat{\omega}_{ij\varepsilon}^\star| \leq |\hat{\omega}_{ji\varepsilon}^\star|) + \hat{\omega}_{ji\varepsilon}^\star \mathbb{I}(|\hat{\omega}_{ij\varepsilon}^\star| > |\hat{\omega}_{ji\varepsilon}^\star|).$$

For $1 \leq i \leq p$, let $\hat{\boldsymbol{\beta}}_{\varepsilon i}$ be a solution of the following convex optimization problem:

$$\min |\boldsymbol{\beta}_{\varepsilon i}|_1 \quad \text{subject to} \quad |\tilde{\mathbf{\Sigma}}_\varepsilon \boldsymbol{\beta}_{\varepsilon i} - \boldsymbol{e}_i|_\infty \leq \lambda_1, \tag{3.18}$$

where $\boldsymbol{\beta}_{\varepsilon i}$ is a real vector and $\boldsymbol{e}_i$ is the vector with 1 in the i-th coordinate and 0 in all other coordinates. Cai et al. (2011) showed that solving the optimization problem (3.17) is equivalent to solving the p optimization problems given in (3.18), i.e., $\{\hat{\mathbf{\Omega}}_\varepsilon^\star\} = \{(\hat{\boldsymbol{\beta}}_{\varepsilon 1}, ..., \hat{\boldsymbol{\beta}}_{\varepsilon p})\}$. This equivalence is useful for both numerical implementation and theoretical analysis. The following theorem gives the convergence results of CLIME under PDD temporal dependence.

Theorem III.3. *Uniformly on $\mathcal{G}_1(q, c_p, M_p, v_0)$ and $\mathcal{B}(C_0, \alpha)$, for sufficiently large constant $M > 0$, if $\lambda_1 = M\lambda'$, $0 \leq \varepsilon \leq M\lambda'/(2v_0)$ and $\lambda' = o(1)$ with*

$$\lambda' := \sqrt{f_1 \log(pf_1)/n} \qquad \textit{and} \qquad f_1 := f_0 \times \begin{cases} (v_0 M_p)^2, & 0 < \alpha \leq 1, \\ (v_0 M_p)^{2/\alpha}, & \alpha > 1, \end{cases} \tag{3.19}$$

then

$$\begin{aligned} |\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}|_\infty &= O_P(M_p \lambda'), \\ \|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_2 &= O_P\left(c_p (M_p \lambda')^{1-q}\right), \\ \frac{1}{p}\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_F^2 &= O_P\left(c_p (M_p \lambda')^{2-q}\right). \end{aligned}$$

Moreover, if $p \geq n^c$ with some constant $c > 0$, then for any constant $C > 0$, there exists a constant $M' > 0$ such that when $M > M'$ and $\min\left\{p^{-C}, M\lambda'/(2v_0)\right\} \leq \varepsilon \leq M\lambda'/(2v_0)$, we have

$$E\left(|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}|_\infty^2\right) = O\left((M_p\lambda')^2\right),$$

$$E\left(\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_2^2\right) = O\left(c_p^2(M_p\lambda')^{2-2q}\right), \tag{3.20}$$

$$\frac{1}{p}E\left(\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_F^2\right) = O\left(c_p(M_p\lambda')^{2-q}\right). \tag{3.21}$$

Remark III.2. The continuity and monotonicity of λ' with respect to α is the same as those of τ' given in Remark III.1. Meanwhile, $\lambda' \asymp M_p^{\mathbb{I}(0<\alpha\leq 1)+\mathbb{I}(\alpha>1)/\alpha}\tau'$. When $\alpha = \infty$, we have $\lambda' \asymp \sqrt{\log p/n}$, and thus for i.i.d data, the convergence rates of CLIME in mean-squared norms given in (3.20) and (3.21) attain the minimax optimal convergence rates of the adaptive CLIME in Cai et al. (2016) under slightly different assumptions. When M_p is constant, then $\lambda' \asymp \tau'$ and the convergence rates are analogous to those for covariance matrix estimation given in Theorem III.1.

Remark III.3. As discussed in Cai et al. (2011), the perturbation parameter $\varepsilon > 0$ is used for a proper initialization of $\{\boldsymbol{\beta}_{\varepsilon i}\}$ in the numerical algorithm, and it also ensures the existence of $E(\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_2^2)$. When $p \geq n^c$, let $M \geq 2v_0$, $C_0 \geq 1/3$ and $C \geq 1/(2c)$, then $M\lambda'/2v_0 \geq n^{-1/2} \geq p^{-1/(2c)} \geq p^{-C}$. Thus, we can simply let $\varepsilon = n^{-1/2}$ in practice, which is also the default setting of the R package `flare` (Li et al., 2015) that implements the CLIME algorithm. A similar choice of ε is given in (10) of Cai et al. (2011) for i.i.d. observations.

To better recover the sparsity structure of $\mathbf{\Omega}$, Cai et al. (2011) introduced additional thresholding on $\hat{\mathbf{\Omega}}_\varepsilon$. Similarly, we may define a hard-thresholded CLIME estimator $\tilde{\mathbf{\Omega}}_\varepsilon = (\tilde{\omega}_{ij\varepsilon})_{p\times p}$ by $\tilde{\omega}_{ij\varepsilon} = \hat{\omega}_{ij\varepsilon}\mathbb{I}(|\hat{\omega}_{ij\varepsilon}| > \xi)$ with tuning parameter $\xi \geq 2M_p\lambda_1$. Although such an estimator enjoys nice theoretical properties given below, how to practically select ξ

remains unknown.

Theorem III.4 (Sparsistency and sign-consistency)**.** *Under the conditions for the convergence in probability given in Theorem III.3, we have* $\tilde{\omega}_{ij\varepsilon} = 0$ *for all* (i, j) *where* $\omega_{ij} = 0$ *with probability tending to 1. If additionally assume all nonzero elements of* $\mathbf{\Omega}$ *satisfy* $|\omega_{ij}| > \xi + 2M_p\lambda_1$*, then we have* $\text{sign}(\tilde{\omega}_{ij\varepsilon}) = \text{sign}(\omega_{ij})$ *for all* (i, j) *where* $\omega_{ij} \neq 0$ *with probability tending to 1.*

3.4.2 SPICE Estimation

For i.i.d. data, Rothman et al. (2008) proposed the SPICE method for estimating the following precision matrix $\mathbf{\Omega}$

$$\mathcal{G}_2(s_p, v_0)=\Big\{\mathbf{\Omega} : \sum_{1\le i\neq j\le p} \mathbb{I}(\omega_{ij} \neq 0) \le s_p, 0<v_0^{-1}\le\varphi_{\min}(\mathbf{\Omega})\le\varphi_{\max}(\mathbf{\Omega})\le v_0\Big\}, \tag{3.22}$$

where s_p determines the sparsity of $\mathbf{\Omega}$ and can depend on p, and v_0 is a constant. Two types of SPICE estimators were proposed:

$$\tilde{\mathbf{\Omega}}_{\lambda_2} = \mathop{\arg\min}_{\tilde{\mathbf{\Omega}}\succ 0,\, \tilde{\mathbf{\Omega}}=\tilde{\mathbf{\Omega}}^T} \Big\{\text{tr}(\tilde{\mathbf{\Omega}}\hat{\mathbf{\Sigma}}) - \log\det(\tilde{\mathbf{\Omega}}) + \lambda_2|\tilde{\mathbf{\Omega}}|_{1,\text{off}}\Big\}, \tag{3.23}$$

and

$$\begin{aligned} \hat{\mathbf{\Omega}}_{\lambda_2} &:= (\hat{\omega}_{ij\lambda_2})_{p\times p} = \hat{\mathbf{W}}^{-1}\hat{\mathbf{K}}_{\lambda_2}\hat{\mathbf{W}}^{-1} \quad \text{with} \\ \hat{\mathbf{K}}_{\lambda_2} &= \mathop{\arg\min}_{\hat{\mathbf{K}}\succ 0,\, \hat{\mathbf{K}}=\hat{\mathbf{K}}^T} \Big\{\text{tr}(\hat{\mathbf{K}}\hat{\mathbf{R}}) - \log\det(\hat{\mathbf{K}}) + \lambda_2|\hat{\mathbf{K}}|_{1,\text{off}}\Big\}, \end{aligned} \tag{3.24}$$

where $\lambda_2 > 0$ is a tuning parameter, and $\hat{\mathbf{W}} = \text{diag}\{\sqrt{\hat{\sigma}_{11}}, \ldots, \sqrt{\hat{\sigma}_{pp}}\}$ is an estimator of $\mathbf{W} = \text{diag}\{\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{pp}}\}$. We can see that $\hat{\mathbf{K}}_{\lambda_2}$ is the SPICE estimator of $\mathbf{K} := \mathbf{R}^{-1}$. The SPICE estimator (3.23) is a slight modification of the graphical Lasso (GLasso) estimator of Friedman et al. (2008). GLasso uses $|\mathbf{\Omega}|_1$ rather than $|\mathbf{\Omega}|_{1,\text{off}}$ in the penalty, but the SPICE estimators (3.23) and (3.24) are more amenable to theoretical analysis (Rothman et al., 2008; Lam and Fan, 2009; Ravikumar et al., 2011), and numerically they give similar results for i.i.d. data (Rothman et al., 2008). It is worth noting that for i.i.d. data, (3.23) requires $\sqrt{(p+s_p)\log p/n} = o(1)$ but (3.24) relaxes it to $\sqrt{(1+s_p)\log p/n} = o(1)$. Similar requirements also hold for temporally dependent observations. Hence in this chapter, we only consider the SPICE estimator given in (3.24).

Theorem III.5. *Uniformly on $\mathcal{G}_2(s_p, v_0)$ and $\mathcal{B}(C_0, \alpha)$, for sufficiently large constant $M > 0$, if $\lambda_2 = M\tau'$ and $\tau' = o(1/\sqrt{1+s_p})$ with τ' defined in* (3.9)*, then*

$$\begin{aligned} \|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_F &= O_P(\tau'\sqrt{s_p}), \\ \|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_2 &= O_P(\tau'\sqrt{1+s_p}), \\ \frac{1}{\sqrt{p}}\|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_F &= O_P\left(\tau'\sqrt{1+s_p/p}\right). \end{aligned} \tag{3.25}$$

If additionally assume $\|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_2 = O_P(\eta)$ with $\eta = O(\tau')$, then with probability tending to 1, $\hat{\omega}_{ij\lambda_2} = 0$ for all (i,j) where $\omega_{ij} = 0$.

For the case with fixed $\alpha > 1$, $\tau' \asymp \sqrt{(\log p)/n}$, so the above results in Theorem III.5 are the same as those given in Rothman et al. (2008, Corollary 1 and Theorem 2) and Lam and Fan (2009, Theorem 4) for i.i.d. observations. By the inequality $\|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_F/\sqrt{p} \le \|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_2 \le \|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_F$ (Golub and Van Loan, 1996) and equation (3.25), the sparsistency result requires that $s_p = O(1)$ if $\eta = \tau'\sqrt{s_p}$, and $s_p = O(p)$ if $\eta = \tau'\sqrt{s_p/p}$. Moreover, the condition $\tau' = o(1/\sqrt{1+s_p})$ implies $s_p = o(\tau'^{-2}) = o(n/\log p)$, meaning that $\mathbf{\Omega}$ needs to be very sparse. Such a condition easily fails for many simple band matrices when $p \ge n$.

Under the irrepresentability condition, however, the sparsity requirement can be relaxed (Ravikumar et al., 2011). In particular, define $\mathbf{\Gamma} = \mathbf{R} \otimes \mathbf{R}$. By (i,j)-th row of $\mathbf{\Gamma}$ we refer to its $[i + (j-1)p]$-th row, and by (k,l)-th column to its $[k + (l-1)p]$-th column. For any two subsets T and T' of $\{1, ..., p\} \times \{1, ..., p\}$, denote $\mathbf{\Gamma}_{TT'}$ be the card$(T)\times$card(T') matrix with rows and columns of $\mathbf{\Gamma}$ indexed by T and T' respectively, where card(T)

denotes the cardinality of set T. Let S be the set of nonzero entries of $\mathbf{\Omega}$ and S^c be the complement of S in $\{1, ..., p\} \times \{1, ..., p\}$. Define $\kappa_{\mathbf{R}} = \|\mathbf{R}\|_1$ and $\kappa_{\mathbf{\Gamma}} = \|\mathbf{\Gamma}_{SS}^{-1}\|_1$. Assume the following irrepresentability condition of Ravikumar et al. (2011):

$$\max_{e \in S^c} \left| \mathbf{\Gamma}_{eS} \mathbf{\Gamma}_{SS}^{-1} \right|_1 \leq 1 - \beta \tag{3.26}$$

for some $\beta \in (0, 1]$. Define d to be the maximum number of nonzeros per row in $\mathbf{\Omega}$. Then we have the following result.

Theorem III.6. *Let $r = (0.5 + 2.5(1 + 8/\beta)\kappa_{\mathbf{\Gamma}}) M\tau' v_0$, where τ' is defined in* (3.9)*. Uniformly on $\mathcal{G}_2(s_p, v_0)$ and $\mathcal{B}(C_0, \alpha)$, for sufficiently large constant $M > 0$, if $\lambda_2 = 8M\tau'/\beta \leq [6(1 + \beta/8)d \max\{\kappa_{\mathbf{R}}\kappa_{\mathbf{\Gamma}}, \kappa_{\mathbf{R}}^3\kappa_{\mathbf{\Gamma}}^2\}]^{-1}$ and $\tau' = o(\min\{1, [(1 + 8/\beta)\kappa_{\mathbf{\Gamma}}]^{-1}\})$, then with probability tending to 1 we have*

$$|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}|_\infty \leq r,$$

$$\|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_2 \leq r \min\left\{d, \sqrt{p + s_p}\right\},$$

$$\frac{1}{\sqrt{p}}\|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_F \leq r\sqrt{1 + s_p/p},$$

and $\hat{\omega}_{ij\lambda_2} = 0$ for all (i, j) with $\omega_{ij} = 0$. If we further assume all nonzero elements of $\mathbf{\Omega}$ satisfy $|\omega_{ij}| > r$, then with probability tending to 1, $\text{sign}(\hat{\omega}_{ij\lambda_2}) = \text{sign}(\omega_{ij})$ *for all (i, j) where $\omega_{ij} \neq 0$.*

Consider the case when β remains constant and $\max\{\kappa_{\mathbf{R}}, \kappa_{\mathbf{\Gamma}}\}$ has a constant upper bound. Then the conditions in Theorem III.6 about λ_2 and τ' reduce to $\lambda_2 = M'\tau'$ and $\tau' = o(1)$ with a constant $M' = 8M/\beta$, and meanwhile we have $\|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_2 = O_P(\tau' d)$. Then the desired result of $\|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_2 = o_P(1)$ is achieved under a relaxed sparsity condition $d = o(\tau'^{-1})$. If $d^2 > 1 + s_p$, then $s_p = o(\tau'^{-2})$ and the condition of Theorem III.5 satisfies. Hence $\|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_2 = O_P(\tau'\sqrt{\min\{d^2, 1 + s_p\}}) = o_P(1)$, which is the better rate between those from Theorems III.5 and III.6.

3.5 Numerical Experiments

3.5.1 Cross-Validation

For tuning parameter selection, we propose a gap-block cross-validation (CV) method that includes the following three steps:

1. Split the data $\mathbf{X}_{p\times n}$ into $H_1 \geq 4$ approximately equal-sized non-overlapping blocks $\mathbf{X}_i^*$, $i = 1, \ldots, H_1$, such that $\mathbf{X}_{p\times n} = (\mathbf{X}_1^*, \mathbf{X}_2^*, \ldots, \mathbf{X}_{H_1}^*)$. For each i, set aside block $\mathbf{X}_i^*$ that will be used as the validation data, and use the remaining data after further dropping the neighboring block at either side of $\mathbf{X}_i^*$ as the training data that are denoted by $\mathbf{X}_i^{**}$.

2. Randomly sample H_2 blocks $\mathbf{X}_{H_1+1}^*, \ldots, \mathbf{X}_{H_1+H_2}^*$ from $\mathbf{X}_{p\times n}$, where $\mathbf{X}_{H_1+j}^*$ consists of $\lceil n/H_1 \rceil$ consecutive columns of $\mathbf{X}_{p\times n}$ for each $j = 1, \ldots, H_2$. Note that these sampled blocks can overlap. For each $i = H_1 + 1, \ldots, H_1 + H_2$, set aside block $\mathbf{X}_i^*$ as the validation data, and use the remaining data by further excluding the $\lceil n/H_1 \rceil$ columns at either side of $\mathbf{X}_i^*$ from $\mathbf{X}_{p\times n}$ as the training data that are denoted by $\mathbf{X}_i^{**}$.

3. Let $H = H_1 + H_2$. For generalized thresholding of covariance matrix estimation, select the optimal tuning parameter τ among a prespecified set of candidates $\{\tau_j\}_{j=1}^J$ and denote it by
$$\tau_s^{\mathbf{\Sigma}} = \operatorname*{arg\,min}_{1\leq j\leq J} \frac{1}{H} \sum_{i=1}^{H} \|S_{\tau_j}(\hat{\mathbf{\Sigma}}_i^{**}) - \hat{\mathbf{\Sigma}}_i^*\|_F^2,$$
where $\hat{\mathbf{\Sigma}}_i^*$ and $\hat{\mathbf{\Sigma}}_i^{**}$ are the corresponding sample covariance matrices based on $\mathbf{X}_i^*$ and $\mathbf{X}_i^{**}$, respectively. For the estimation of correlation matrix, we replace $\hat{\mathbf{\Sigma}}_i^*$ and $\hat{\mathbf{\Sigma}}_i^{**}$ by $\hat{\mathbf{R}}_i^*$ and $\hat{\mathbf{R}}_i^{**}$, respectively. For the estimation of precision matrix, we choose the optimal tuning parameter using the loss function tr$(\hat{\mathbf{\Omega}}_\lambda^{**}\hat{\mathbf{\Sigma}}^*) - \log\det(\hat{\mathbf{\Omega}}_\lambda^{**})$.

In the above CV method, we use gap blocks, each of size $\approx \lceil n/H_1 \rceil$, to separate training and validation datasets so that they are nearly uncorrelated. The idea of using gap blocks has been employed by the hv-block CV of Racine (2000) for linear models with dependent data. Similar to the k-fold CV for i.i.d. data, Step 1 guarantees all observations are used for both training and validation, but is limited due to the constrain of keeping the temporal ordering of the observations. Step 2 allows more data splits. This is particularly useful when Step 1 only allows a small number of data splits due to large-size of the gap block and/or limited sample size n. Step 2 is inspired by the commonly used repeated random subsampling CV for i.i.d. observations (Syed et al., 2012). The above loss functions for selecting tuning parameters are widely used in the literature (Bickel and Levina, 2008a; Rothman et al., 2009; Cai et al., 2011, 2016). The theoretical justification for the gap-block CV remains open. However, our simulation studies show that the method performs well for data with PDD temporal dependence.

3.5.2 Simulation Studies

We evaluate the numerical performance of the hard and soft thresholding estimators for large correlation matrix and the CLIME and SPICE estimators for large precision matrix. We generate Gaussian data with zero mean and covariance matrix $\mathbf{\Sigma}$ or precision matrix $\mathbf{\Omega}$ from one of the following four models:

Model 1: $\sigma_{ij} = 0.6^{|i-j|}$;

Model 2: $\sigma_{ii} = 1$, $\sigma_{i,i+1} = \sigma_{i+1,i} = 0.6$, $\sigma_{i,i+2} = \sigma_{i+2,i} = 0.3$, and $\sigma_{ij} = 0$ for $|i-j| \geq 3$;

Model 3: $\omega_{ij} = 0.6^{|i-j|}$;

Model 4: $\omega_{ii} = 1$, $\omega_{i,i+1} = \omega_{i+1,i} = 0.6$, $\omega_{i,i+2} = \omega_{i+2,i} = 0.3$, and $\omega_{ij} = 0$ for $|i-j| \geq 3$.

Similar models have been considered in Bickel and Levina (2008a), Rothman et al. (2008), Rothman et al. (2009), Cai et al. (2011), and Cai et al. (2016). For the temporal dependence, we set $\rho_{kl}^{ij} = \theta_{kl}^{ij}\rho_{kl}$ with

$$\theta_{kl}^{ij} = (|i-j|+1)^{-\alpha}, \quad 1 \leq i, j \leq n, \tag{3.27}$$

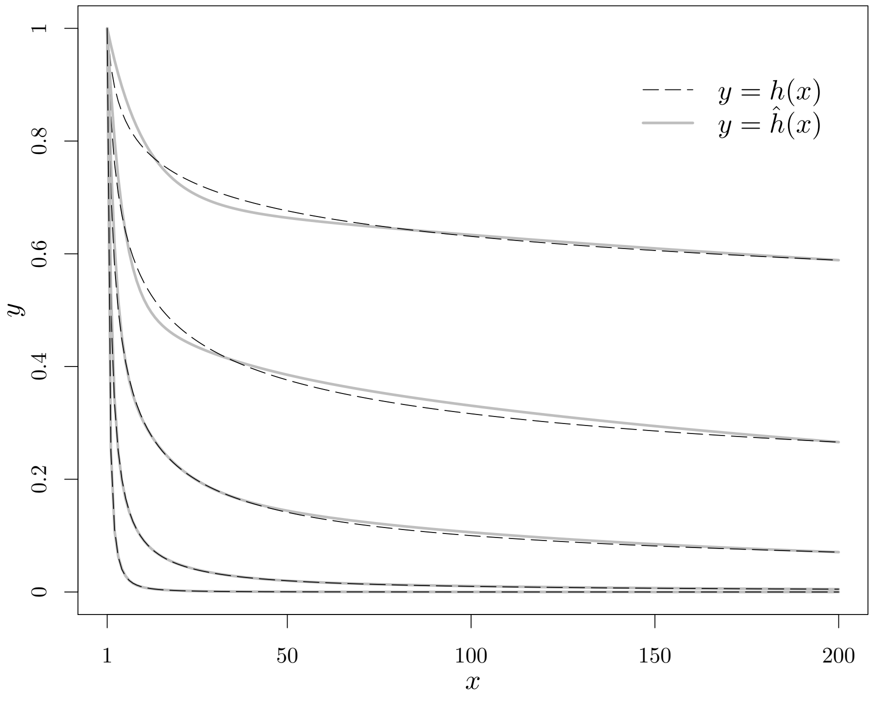


Figure 3.2: Approximation of $h(x) = x^{-\alpha}$ for $\alpha = 0.1, 0.25, 0.5, 1, 2$.

so that $|\rho_{kl}^{ij}| \leq |\theta_{kl}^{ij}| \sim |i-j|^{-\alpha}$. It is computationally expensive to simulate data $\boldsymbol{X}_{pn} := \text{vec}(\mathbf{X}_{p\times n})$ directly from a multivariate Gaussian random number generator because of the large dimension of its covariance matrix $\text{cov}(\boldsymbol{X}_{pn})$. We use an alternative way to simulate the data that approximately satisfy (3.27). Note that $h(x) = x^{-\alpha}$, $x \in [1, n]$ and $\alpha > 0$, can be approximated by $\hat{h}(x) = \sum_{i=0}^{N} a_i \exp(-b_i x)$ with small N and appropriately chosen $\{a_i, b_i\}$ by the method of Bochud and Challet (2007) (see Figure 3.2). Thus, data $\mathbf{X}_{p\times n}$ are simulated as follows: each column of $\mathbf{X}_{p\times n}$ is generated by $\boldsymbol{X}_t = \sum_{i=0}^{N} c_i \boldsymbol{Y}_t^{(i)}$ for $t = 1, ..., n$, where $c_i = \sqrt{a_i \exp(-b_i)}$, $\boldsymbol{Y}_1^{(i)}$ are i.i.d. $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ for all i, and for $t = 2, ..., n$, $\boldsymbol{Y}_t^{(i)} = \rho_i \boldsymbol{Y}_{t-1}^{(i)} + \boldsymbol{e}_t^{(i)}$ with $\rho_i = \exp(-b_i)$ and white noise $(1-\rho_i^2)^{-1/2}\boldsymbol{e}_t^{(i)}$ i.i.d. $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. It is easily seen that $\mathbf{R}^{t,t+j} = \sum_{i=0}^{N} c_i^2 \rho_i^j \mathbf{R} = \sum_{i=0}^{N} a_i \exp\{-b_i(j+1)\}\mathbf{R} \approx (j+1)^{-\alpha}\mathbf{R}$.

Simulations are conducted with sample size $n = 200$, variable dimension p ranging from 100 to 400, and 100 replications under each setting, for which α varies from 0.1 to 2. The i.i.d. case is also considered. For each simulated data set, we choose the optimal tuning parameter from a set of 50 specified values (see Appendix B.3) using the gap-block CV with $H_1 = H_2 = 10$ for the PDD temporal dependence and the ordinary 10-fold CV for the i.i.d. case recommended by Fang et al. (2016). The CLIME and SPICE are computed by the R packages `flare` (Li et al., 2015) and `QUIC` (Hsieh et al., 2014), respectively. For CLIME, we use the default perturbation of `flare` with $\varepsilon = n^{-1/2}$.

The estimation performance is measured by both the spectral norm and the Frobenius norm. True-positive rate (TPR) and false-positive rate (FPR) are used for evaluating sparsity recovering:

$$\text{TPR} = \frac{\#\{(i,j) : s_\tau(\hat{\rho}_{ij}) \neq 0 \text{ and } \rho_{ij} \neq 0, i \neq j\}}{\#\{(i,j) : \rho_{ij} \neq 0, i \neq j\}},$$

and

$$\text{FPR} = \frac{\#\{(i,j) : s_\tau(\hat{\rho}_{ij}) \neq 0 \text{ and } \rho_{ij} = 0, i \neq j\}}{\#\{(i,j) : \rho_{ij} = 0, i \neq j\}}$$

for correlation matrix and similarly for precision matrix. The TPR and FPR are not provided for Models 1 and 3.

Simulation results are summarized in Tables 3.1-3.3. In all setups, the sample correlation matrix and the inverse of sample covariance matrix (whenever possible) perform the worst. It is not surprising that the performance of all the regularized estimators generally is better for weaker temporal dependence or smaller p. The soft thresholding method performs slightly better than the hard thresholding method in terms of matrix losses for small α and slightly worse for large α, and always has higher TPRs but bigger FPRs. The CLIME estimator performs similarly as the SPICE estimator in matrix norms, but generally yields lower FPRs.

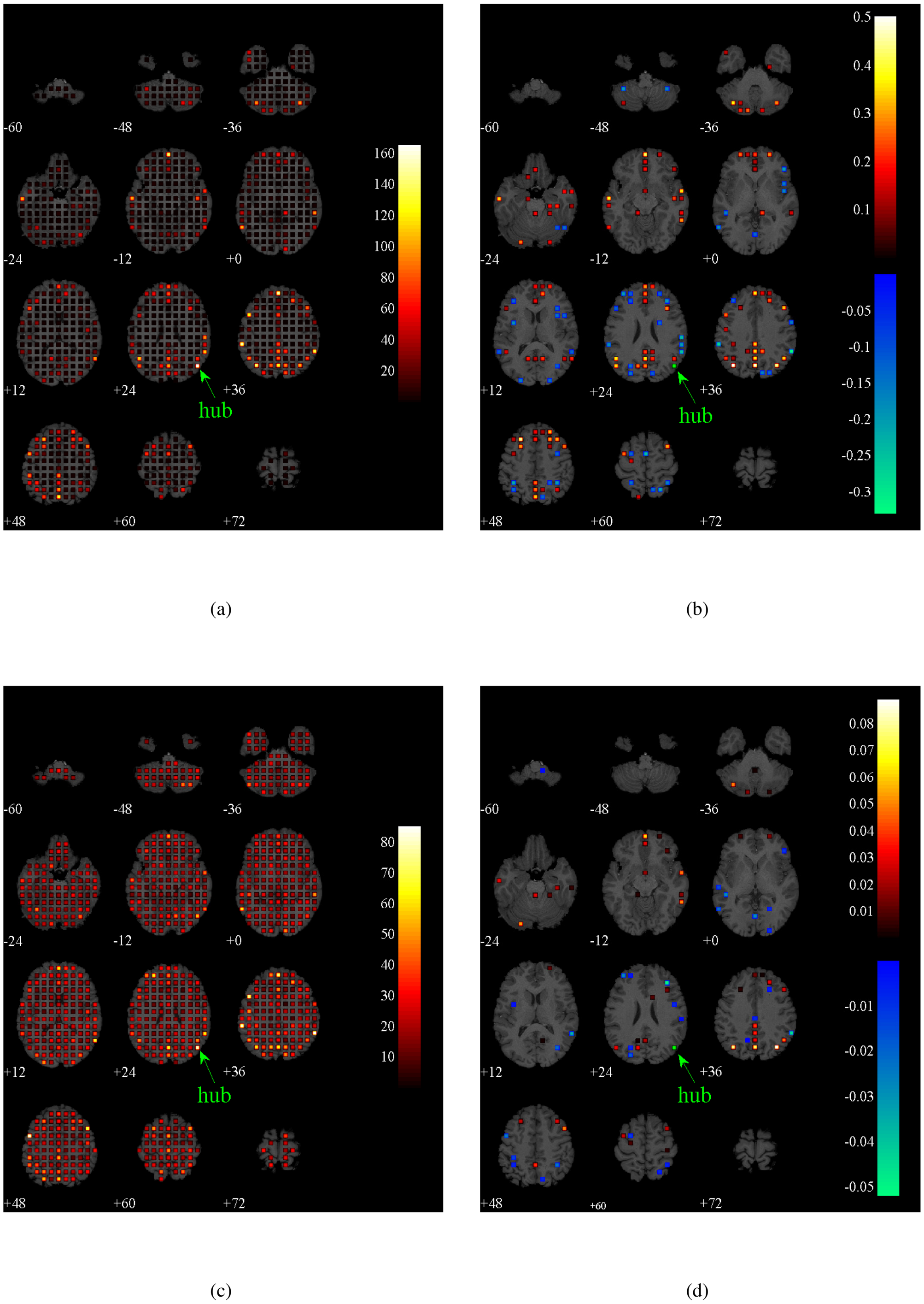
We notice that the SPICE algorithm in the R package `QUIC` is much faster than the

CLIME algorithm in the R package `flare` by using a single computer core. However, the column-by-column estimating nature of CLIME can speed up using parallel computing on multiple cores.

3.5.3 rfMRI Data Analysis

Here we analyze a rfMRI data set for the estimation of brain functional connectivity. The preprocessed rfMRI data of a healthy young woman are provided by the WU-Minn Human Connectome Project (www.humanconnectome.org). The original data consist of 1,200 temporal brain images and each image contains 229,404 brain voxels with size $2{\times}2{\times}2$ mm^3. We discard the first 10 images due to concerns of early nonsteady magnetization, and for the ease of implementation reduce the image dimension using a grid-based method (Sripada et al., 2014) to 907 functional brain nodes that are placed in a regular three-dimensional grid spaced at 12-mm intervals throughout the brain. Each node consists of a 3-mm voxel-center-to-voxel-center radius pseudosphere, which encompasses 19 voxels. The time series for each node is a spatially averaged time series of the 19 voxels within the node. The estimated α from all 907 time series is about 0.25 (see Subsection 3.2.2, Figure 3.1(a)).

The functional connectivity between two brain nodes can be evaluated by either correlation or partial correlation. Here we follow the convention by simply calling them the marginal connectivity and the direct connectivity, respectively. For the marginal connectivity, we only apply the hard thresholding method for estimating the correlation matrix which usually yields less number of false discoveries than the soft thresholding, and find that 1.47$\%$ of all the pairs of nodes are connected with a threshold value of 0.12 to the sample correlations. For the direct connectivity, we calculate the estimated partial correlations $\{-\hat{\omega}_{ij}/\sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}, i \neq j\}$ from the precision matrix estimator $\hat{\mathbf{\Omega}} := (\hat{\omega}_{ij})_{p\times p}$. Both CLIME and SPICE yield similar result, hence we only report the result of CLIME. We find



(a)

(b)

(c)

(d)

Figure 3.3: rfMRI data analysis for brain functional connectivity. (a) Node degrees of marginal connectivity found by hard thresholding. (b) Marginally connected nodes and their estimated correlations to the selected hub. (c) Node degrees of direct connectivity found by CLIME. (d) Directly connected nodes and their estimated partial correlations to the selected hub. The brain is plotted in the Montreal Neurological Institute 152 space with Z-coordinates displayed.

that 2.71% of all the pairs of nodes are connected conditional on all other nodes. Most of the nonzero estimated partial correlations have small absolute values, with the medium at 0.01 and the maximum at 0.45. About 0.62% of all the pairs of nodes are connected both marginally and directly.

Define the degree of a node to be the number of its connected nodes, and a hub to be a high-degree node. We illustrate the node degrees of marginal connectivity and direct connectivity in Figure 3.3 (a) and (c), respectively. The marginal connectivity node degrees range from 0 to 164 with the medium at 2, and the direct connectivity node degrees range from 5 to 85 with the medium at 22. The top 10 hubs found by either method are provided in Appendix B.4 with six overlapping hubs. Seven of the top 10 hubs of marginal connectivity are spatially close to those in Buckner et al. (2009) and Cole et al. (2010) obtained from multiple subjects. Note that they arbitrarily used 0.25 as the threshold value for the sample correlations, whereas our threshold value of 0.12 is selected from cross-validation. As an illustration, we plot the marginal and the direct connectivity of a single hub in Figure 3.3 (b) and (d) respectively. The selected hub has 164 marginally connected nodes and 79 directly connected nodes, where 80% of the directly connected nodes are also marginally connected. It is located in the right inferior parietal cortex, a part of the so-called default mode network (Buckner et al., 2008) that is most active during the resting state.

3.6 Sketched Proofs of Theoretical Results

3.6.1 General Theorems

We first provide theoretical results for the following general model of temporal dependence which includes the PDD temporal dependence as a special case. The proofs of these general results are provided in Appendix B. Then the theoretical results for the PDD temporal dependence given in Sections 3.3 and 3.4 can be obtained directly by specifying the appropriate model parameters, which will be shown in subsection 3.6.2. Consider

$$\mathcal{A}(f(n,p), g(n,p)) = \left\{ \{\mathbf{R}^{ij}\} : \max_{1\le j\le n} \sum_{\substack{i\in\{1\le i\le n: \\ |i-j|=kf, \\ k=1,\ldots,\lfloor n/f\rfloor\}}} |\mathbf{R}^{ij}|_\infty \le g(n,p) \right\} \tag{3.28}$$

where $f(n,p) \in [1,n]$ is an integer-valued function and $g(n,p)$ is a real function. We sometimes drop the dependence of f, g on n, p for notational simplicity. Define $\tau_0 = \sqrt{f\log(pf)/n}$. Then we have the following general theorems.

Theorem III.7. (a). *Uniformly on $\mathcal{U}(q, c_p, v_0)$ and $\mathcal{A}(f,g)$, for sufficiently large constant $M > 0$, if $\tau = M\tau_0$ with $\tau_0 = o(1)$, and $\limsup_{n\to\infty} g(n,p) < 1$, then*

$$|S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}|_\infty = O_P(\tau_0), \tag{3.29}$$

$$\|S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_2 = O_P\left(c_p\tau_0^{1-q}\right), \tag{3.30}$$

$$\frac{1}{p}\|S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_F^2 = O_P\left(c_p\tau_0^{2-q}\right), \tag{3.31}$$

and $s_\tau(\hat{\sigma}_{ij}) = 0$ for all (i,j) where $\sigma_{ij} = 0$ with probability tending to 1. When all nonzero elements of $\mathbf{\Sigma}$ satisfy $|\sigma_{ij}| \ge 2\tau$, then $\text{sign}(s_\tau(\hat{\sigma}_{ij})) = \text{sign}(\sigma_{ij})$ *for all (i,j) where $\sigma_{ij} \ne 0$ with probability tending to 1. Moreover, if $p \ge n^c$ for some constant $c > 0$, then*

$$E\left(|S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}|_\infty^2\right) = O(\tau_0^2), \tag{3.32}$$

$$E\left(\|S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_2^2\right) = O\left(c_p^2\tau_0^{2-2q}\right), \tag{3.33}$$

$$\frac{1}{p}E\left(\|S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_F^2\right) = O\left(c_p\tau_0^{2-q}\right). \tag{3.34}$$

(b). *Part* (a) *holds with $\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}, \hat{\sigma}_{ij}, \sigma_{ij}$ and $\mathcal{U}(q, c_p, v_0)$ replaced by $\hat{\mathbf{R}}, \mathbf{R}, \hat{\rho}_{ij}, \rho_{ij}$ and $\mathcal{R}(q, c_p)$, respectively.*

Theorem III.8. *Uniformly on $\mathcal{G}_1(q, c_p, M_p, v_0)$ and $\mathcal{A}(f,g)$, for sufficiently large constant $M > 0$, if $\lambda_1 = M\tau_0$ and $0 \le \varepsilon \le M\tau_0/(2v_0)$ with $\tau_0 = o(1)$, and $\limsup_{n\to\infty} v_0^2 M_p^2 g < 1$,* *then*

$$|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}|_\infty = O_P(M_p\tau_0), \tag{3.35}$$

$$\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_2 = O_P\left(c_p(M_p\tau_0)^{1-q}\right), \tag{3.36}$$

$$\frac{1}{p}\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_F^2 = O_P\left(c_p(M_p\tau_0)^{2-q}\right), \tag{3.37}$$

and $\tilde{\omega}_{ij\varepsilon} = 0$ for all (i,j) where $\omega_{ij} = 0$ with probability tending to 1. When all nonzero elements of $\mathbf{\Omega}$ satisfy $|\omega_{ij}| > \xi + 2M_p\lambda_1$, then $\operatorname{sign}(\tilde{\omega}_{ij\varepsilon}) = \operatorname{sign}(\omega_{ij})$ *for all (i,j) where $\omega_{ij} \neq 0$ with probability tending to 1. Moreover, if $p \geq n^c$ with some constant $c > 0$, then for any constant $C > 0$, there exists a constant $M' > 0$ such that when $M > M'$ and* $\min\left\{p^{-C}, M\tau_0/(2v_0)\right\} \leq \varepsilon \leq M\tau_0/(2v_0)$*, we have*

$$E\left(|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}|_\infty^2\right) = O\left((M_p\tau_0)^2\right),$$
$$E\left(\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_2^2\right) = O\left(c_p^2(M_p\tau_0)^{2-2q}\right),$$
$$\frac{1}{p}E\left(\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_F^2\right) = O\left(c_p(M_p\tau_0)^{2-q}\right).$$

Theorem III.9. *Uniformly on $\mathcal{G}_2(s_p, v_0)$ and $\mathcal{A}(f,g)$, for sufficiently large constant $M > 0$, if $\lambda_2 = M\tau_0$ with $\tau_0 = o(1/\sqrt{1+s_p})$, and $\limsup_{n\to\infty} g < 1$, then*

$$\|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_F = O_P(\tau_0\sqrt{s_p}), \tag{3.38}$$

$$\|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_2 = O_P(\tau_0\sqrt{1+s_p}), \tag{3.39}$$

$$\frac{1}{\sqrt{p}}\|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_F = O_P\left(\tau_0\sqrt{1+s_p/p}\right). \tag{3.40}$$

When $\|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_2 = O_P(\eta)$ with $\eta = O(\tau_0)$, then with probability tending to 1, we have $\hat{\omega}_{ij\lambda_2} = 0$ for all (i,j) where $\omega_{ij} = 0$. Furthermore, if the conditions $\lambda_2 = M\tau_0$ and $\tau_0 = o(1/\sqrt{1+s_p})$ are replaced by $\lambda_2 = 8M\tau_0/\beta \leq [6(1+\beta/8)d\max\{\kappa_{\mathbf{R}}\kappa_{\mathbf{\Gamma}}, \kappa_{\mathbf{R}}^3\kappa_{\mathbf{\Gamma}}^2\}]^{-1}$ and $\tau_0 = o(\min\{1, [(1+8/\beta)\kappa_{\mathbf{\Gamma}}]^{-1}\})$, let $r = (0.5 + 2.5(1+8/\beta)\kappa_{\mathbf{\Gamma}})\,M\tau_0 v_0$, then under *the irrepresentability condition* (3.26)*, with probability tending to 1,*

$$
\begin{aligned}
|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}|_\infty &\le r,\\
\|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_2 &\le r \min\left\{d, \sqrt{p+s_p}\right\},\\
\frac{1}{\sqrt{p}}\|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_F &\le r\sqrt{1+s_p/p},
\end{aligned}
$$

and $\hat{\omega}_{ij\lambda_2} = 0$ *for all* (i,j) *with* $\omega_{ij} = 0$*, and moreover,* $\operatorname{sign}(\hat{\omega}_{ij\lambda_2}) = \operatorname{sign}(\omega_{ij})$ *for all* (i,j) *where* $\omega_{ij} \neq 0$ *when all nonzero elements of* $\mathbf{\Omega}$ *satisfy* $|\omega_{ij}| > r$.

3.6.2 Proofs of Main Results for PDD Temporal Dependence

We first show that $\mathcal{B}(C_0,\alpha) \subset \mathcal{A}(f,g)$ with suitable choices of f and g. If $\{\mathbf{R}^{ij}\} \in \mathcal{B}(C_0,\alpha)$, then for any $f \in [1,n]$, we have

$$
\begin{aligned}
\max_{1\le j\le n} \sum_{\substack{i\in\{1\le i\le n:\\ |i-j|=kf,\\ k=1,\dots,\lfloor n/f\rfloor\}}} |\mathbf{R}^{ij}|_\infty/(2C_0) &\le (f)^{-\alpha} + (2f)^{-\alpha} + \cdots + (\lfloor n/f\rfloor f)^{-\alpha}\\
&\le \left(1 + \int_1^{\lfloor n/f\rfloor} y^{-\alpha} dy\right)/f^\alpha\\
&\le \begin{cases} f^{-\alpha}[(n/f)^{1-\alpha} - \alpha]/(1-\alpha), & \alpha \neq 1,\\ f^{-1}[1+\log(n/f)], & \alpha = 1.\end{cases}
\end{aligned}
$$

Thus, $\mathcal{B}(C_0,\alpha) \subset \mathcal{A}(f,g)$ with

$$
g = 2C_0 \times \begin{cases} f^{-\alpha}[(n/f)^{1-\alpha} - \alpha]/(1-\alpha), & \alpha \neq 1,\\ f^{-1}[1+\log(n/f)], & \alpha = 1.\end{cases} \tag{3.41}
$$

We then show that all the theoretical results given in Sections 3.3 and 3.4 for $\mathcal{B}(C_0,\alpha)$ can be obtained from the general theorems in Subsection 3.6.1 by specifying appropriate f.

Proofs of Theorems III.1, III.2 and Corollary III.1. These results are for generalized thresholding estimators under $\mathcal{B}(C_0,\alpha)$. We only need to consider the choice of f such that g given in (3.41) also satisfies the assumption $\limsup_{n\to\infty} g < 1$ in Theorem III.7.

Since (3.41) gives

$$g = 2C_0 \times \begin{cases} f^{-1}(n^{1-\alpha} - \alpha f^{1-\alpha})/(1-\alpha), & 0 < \alpha < 1 \\ f^{-1}(1 + \log n - \log f), & \alpha = 1, \\ f^{-\alpha}[\alpha - (f/n)^{\alpha-1}]/(\alpha - 1), & \alpha > 1, \end{cases}$$

$$\leq 2C_0 \times \begin{cases} f^{-1}(n^{1-\alpha} - \alpha)/(1-\alpha), & 0 < \alpha < 1, \\ f^{-1}(1 + \log n), & \alpha = 1, \\ f^{-\alpha}(n^{1-\alpha} - \alpha)/(1-\alpha), & \alpha > 1, \end{cases} \tag{3.42}$$

then letting (3.42) be less than 2/3 for convenience (or any constant in $(0, 1)$), we obtain $f > f_0$ with f_0 given in (3.9). Thus, $f = \lfloor f_0 \rfloor + 1$ is an appropriate choice, and then plugging it into $\tau_0 = \sqrt{f \log(pf)/n}$ yields $\tau_0 \asymp \tau'$ that is given in (3.9). Hence, the theoretical results of generalized thresholding for $\mathcal{B}(C_0, \alpha)$ automatically follow from Theorem III.7 with $f = \lfloor f_0 \rfloor + 1$ and g given in (3.41). □

Proofs of Theorems III.3 and III.4. For CLIME, Theorem III.8 requires $\limsup_{n\to\infty} v_0^2 M_p^2 g < 1$. Set $v_0^2 M_p^2 g < 2/3$ for simplicity. Following the same steps shown in the above, we obtain an appropriate choice that $f = \lfloor f_1 \rfloor + 1$ with f_1 given in (3.19). Plugging it into $\tau_0 = \sqrt{f \log(pf)/n}$ yields $\tau_0 \asymp \lambda'$ that is also given in (3.19). Then apply Theorem III.8 to $\mathcal{B}(C_0, \alpha)$ with $f = \lfloor f_1 \rfloor + 1$ and g given in (3.41). □

Proofs of Theorems III.5 and III.6. For SPICE, Theorem III.9 requires $\limsup_{n\to\infty} g < 1$ that is the same condition required by Theorem III.7 for generalized thresholding. Hence, we use the same choice of f, i.e., $f = \lfloor f_0 \rfloor + 1$. Then apply Theorem III.9 to $\mathcal{B}(C_0, \alpha)$ with $f = \lfloor f_0 \rfloor + 1$ and g given in (3.41). □

Table 3.1: Comparison of average (SD) matrix losses for correlation matrix estimation

		Model 1						Model 2					
p	α	$\hat{\mathbf{R}}$	Hard	Soft	$\hat{\mathbf{R}}$	Hard	Soft	$\hat{\mathbf{R}}$	Hard	Soft	$\hat{\mathbf{R}}$	Hard	Soft
		Spectral norm			Frobenius norm			Spectral norm			Frobenius norm		
100	0.1	13.7(1.68)	2.8(0.09)	2.6(0.07)	22.6(1.08)	9.9(0.28)	8.7(0.24)	13.8(1.71)	1.8(0.04)	1.6(0.04)	22.6(1.05)	8.7(0.29)	7.7(0.22)
	0.25	10.5(1.59)	2.4(0.15)	2.4(0.08)	17.4(0.95)	8.1(0.42)	7.5(0.26)	10.5(1.61)	1.5(0.18)	1.4(0.09)	17.5(0.94)	6.7(0.48)	6.5(0.24)
	0.5	7.8(1.14)	2.0(0.15)	2.2(0.08)	14.3(0.69)	6.8(0.33)	6.6(0.23)	7.8(1.10)	1.2(0.17)	1.3(0.07)	14.3(0.66)	5.2(0.34)	5.6(0.21)
	1	4.2(0.45)	1.5(0.10)	1.7(0.08)	9.9(0.29)	5.2(0.23)	5.1(0.20)	4.2(0.40)	0.7(0.09)	1.0(0.05)	10.0(0.27)	4.0(0.17)	4.1(0.16)
	2	2.6(0.24)	1.1(0.09)	1.4(0.08)	7.5(0.17)	3.9(0.15)	4.0(0.19)	2.5(0.18)	0.6(0.05)	0.8(0.04)	7.5(0.14)	2.6(0.25)	3.2(0.13)
	i.i.d.	2.4(0.18)	1.0(0.08)	1.3(0.08)	7.0(0.15)	3.5(0.13)	3.7(0.15)	2.3(0.15)	0.5(0.07)	0.7(0.04)	7.0(0.13)	2.0(0.23)	2.8(0.12)
200	0.1	27.2(2.69)	2.9(0.05)	2.8(0.04)	45.6(1.54)	14.5(0.25)	13.1(0.22)	27.2(2.62)	1.8(0.02)	1.7(0.03)	45.6(1.51)	12.9(0.28)	11.6(0.21)
	0.25	20.6(2.54)	2.5(0.14)	2.5(0.06)	35.0(1.39)	12.2(0.56)	11.4(0.29)	20.6(2.29)	1.6(0.15)	1.5(0.07)	35.0(1.29)	10.3(0.56)	9.9(0.27)
	0.5	15.2(1.77)	2.2(0.12)	2.3(0.06)	28.7(0.99)	10.2(0.40)	10.1(0.25)	15.1(1.58)	1.3(0.14)	1.4(0.05)	28.8(0.88)	7.9(0.43)	8.6(0.21)
	1	7.8(0.64)	1.6(0.08)	1.9(0.06)	20.1(0.35)	7.9(0.24)	7.9(0.21)	7.7(0.57)	0.8(0.10)	1.1(0.04)	20.1(0.34)	5.8(0.15)	6.5(0.20)
	2	4.3(0.24)	1.3(0.08)	1.6(0.06)	15.1(0.15)	5.9(0.19)	6.3(0.18)	4.2(0.18)	0.6(0.05)	0.9(0.04)	15.2(0.14)	4.2(0.30)	5.0(0.13)
	i.i.d.	3.8(0.22)	1.1(0.07)	1.5(0.06)	14.1(0.15)	5.3(0.14)	5.8(0.17)	3.6(0.16)	0.6(0.06)	0.8(0.04)	14.1(0.14)	3.2(0.23)	4.4(0.12)
300	0.1	40.6(3.39)	3.0(0.03)	2.8(0.03)	68.5(1.88)	18.0(0.21)	16.5(0.24)	40.8(3.54)	1.8(0.05)	1.7(0.02)	68.7(1.84)	16.0(0.27)	14.6(0.24)
	0.25	30.9(3.23)	2.6(0.11)	2.6(0.04)	52.6(1.75)	15.4(0.63)	14.5(0.30)	30.8(2.95)	1.7(0.17)	1.6(0.13)	52.6(1.62)	13.2(0.69)	12.5(0.28)
	0.5	22.5(2.16)	2.3(0.12)	2.4(0.04)	43.2(1.16)	12.8(0.47)	12.9(0.27)	22.4(2.04)	1.4(0.12)	1.4(0.09)	43.3(1.10)	10.1(0.57)	10.9(0.25)
	1	11.2(0.79)	1.7(0.05)	2.0(0.05)	30.2(0.42)	9.9(0.21)	10.1(0.25)	11.1(0.73)	0.9(0.08)	1.1(0.03)	30.3(0.41)	7.3(0.16)	8.3(0.20)
	2	5.8(0.27)	1.3(0.08)	1.7(0.05)	22.8(0.16)	7.5(0.25)	8.2(0.19)	5.6(0.22)	0.6(0.04)	0.9(0.04)	22.8(0.14)	5.5(0.29)	6.5(0.18)
	i.i.d.	5.0(0.20)	1.2(0.08)	1.6(0.05)	21.2(0.15)	6.7(0.12)	7.5(0.17)	4.7(0.15)	0.6(0.05)	0.8(0.03)	21.2(0.13)	4.1(0.21)	5.7(0.12)
400	0.1	54.2(4.01)	3.0(0.02)	2.9(0.02)	91.7(2.17)	20.9(0.17)	19.4(0.22)	54.0(3.61)	1.8(0.04)	1.7(0.02)	91.7(1.97)	18.6(0.16)	17.2(0.15)
	0.25	41.0(3.88)	2.7(0.09)	2.7(0.04)	70.1(2.09)	18.4(0.61)	17.1(0.29)	41.1(3.58)	1.7(0.09)	1.7(0.12)	70.2(1.89)	15.8(0.63)	14.9(0.33)
	0.5	29.8(2.62)	2.3(0.12)	2.5(0.04)	57.7(1.38)	15.2(0.59)	15.3(0.30)	29.7(2.53)	1.5(0.17)	1.5(0.08)	57.7(1.29)	12.1(0.62)	13.0(0.24)
	1	14.6(0.91)	1.7(0.05)	2.1(0.04)	40.3(0.48)	11.6(0.22)	12.1(0.20)	14.5(0.86)	0.9(0.08)	1.1(0.03)	40.4(0.46)	8.6(0.16)	10.0(0.23)
	2	7.2(0.26)	1.4(0.07)	1.8(0.04)	30.4(0.16)	9.0(0.27)	9.8(0.23)	7.0(0.26)	0.7(0.04)	0.9(0.03)	30.4(0.14)	6.6(0.26)	7.7(0.15)
	i.i.d.	6.0(0.21)	1.2(0.08)	1.6(0.05)	28.2(0.15)	7.9(0.14)	8.9(0.17)	5.7(0.18)	0.6(0.05)	0.9(0.03)	28.3(0.13)	4.9(0.21)	6.8(0.12)

Table 3.2: Comparison of average (SD) matrix losses for precision matrix estimation

		Model 3						Model 4					
p	α	$\hat{\mathbf{\Sigma}}^{-1}$	CLIME	SPICE	$\hat{\mathbf{\Sigma}}^{-1}$	CLIME	SPICE	$\hat{\mathbf{\Sigma}}^{-1}$	CLIME	SPICE	$\hat{\mathbf{\Sigma}}^{-1}$	CLIME	SPICE
		Spectral norm			Frobenius norm			Spectral norm			Frobenius norm		
100	0.1	381.7(40.07)	4.9(0.26)	5.7(0.53)	850.5(38.22)	28.8(1.54)	27.1(1.46)	355.4(37.62)	4.9(0.40)	5.9(0.72)	829.5(35.78)	28.0(2.05)	26.5(1.68)
	0.25	97.6(9.23)	1.8(0.09)	2.2(0.08)	214.6(9.38)	9.5(0.34)	9.3(0.20)	91.1(8.42)	1.9(0.31)	1.7(0.19)	209.0(8.63)	8.2(1.03)	7.3(0.30)
	0.5	43.3(4.60)	2.4(0.09)	2.7(0.06)	93.9(4.36)	7.7(0.15)	8.6(0.15)	40.7(4.29)	1.1(0.10)	1.4(0.07)	91.6(3.96)	4.7(0.17)	5.8(0.19)
	1	21.8(2.74)	2.6(0.06)	2.9(0.04)	45.4(2.73)	8.0(0.19)	9.2(0.15)	20.5(2.44)	1.3(0.07)	1.6(0.06)	44.4(2.44)	5.1(0.26)	6.2(0.21)
	2	14.1(1.80)	2.7(0.05)	2.9(0.04)	28.9(1.86)	8.0(0.20)	9.1(0.14)	13.3(1.62)	1.4(0.07)	1.6(0.05)	28.3(1.70)	5.3(0.25)	6.3(0.17)
	i.i.d.	12.6(1.66)	2.5(0.06)	2.8(0.04)	25.5(1.56)	7.4(0.20)	8.6(0.15)	11.8(1.44)	1.2(0.06)	1.4(0.05)	25.0(1.37)	4.6(0.24)	5.7(0.18)
200	0.1	N/A	6.2(0.38)	5.8(0.48)	N/A	49.6(2.46)	38.4(1.48)	N/A	5.4(0.50)	5.6(0.61)	N/A	41.4(2.89)	33.9(1.61)
	0.25	N/A	2.1(0.12)	2.4(0.06)	N/A	14.8(0.52)	13.7(0.18)	N/A	1.8(0.19)	1.6(0.14)	N/A	11.5(0.59)	10.5(0.18)
	0.5	N/A	2.6(0.07)	2.8(0.04)	N/A	11.9(0.18)	12.8(0.12)	N/A	1.4(0.11)	1.7(0.04)	N/A	8.5(0.32)	9.6(0.17)
	1	N/A	2.9(0.05)	3.1(0.03)	N/A	12.4(0.23)	13.7(0.14)	N/A	1.6(0.06)	1.8(0.03)	N/A	9.1(0.38)	10.5(0.21)
	2	N/A	2.9(0.04)	3.1(0.02)	N/A	12.6(0.21)	13.8(0.09)	N/A	1.6(0.05)	1.8(0.03)	N/A	9.2(0.32)	10.8(0.17)
	i.i.d.	N/A	2.7(0.04)	3.0(0.02)	N/A	11.6(0.24)	13.3(0.14)	N/A	1.4(0.06)	1.7(0.03)	N/A	7.8(0.34)	9.9(0.17)
300	0.1	N/A	5.3(0.36)	5.9(0.45)	N/A	51.2(2.85)	47.1(1.48)	N/A	6.0(0.54)	5.6(0.67)	N/A	54.7(4.26)	39.8(1.58)
	0.25	N/A	2.4(0.11)	2.4(0.05)	N/A	18.0(0.36)	17.1(0.18)	N/A	1.6(0.12)	1.6(0.14)	N/A	14.0(0.30)	13.2(0.13)
	0.5	N/A	2.8(0.07)	2.9(0.03)	N/A	15.7(0.27)	15.9(0.13)	N/A	1.8(0.07)	1.8(0.04)	N/A	13.1(0.51)	12.5(0.20)
	1	N/A	3.0(0.04)	3.1(0.02)	N/A	15.9(0.28)	17.1(0.12)	N/A	1.9(0.06)	1.9(0.03)	N/A	13.1(0.53)	13.8(0.20)
	2	N/A	3.0(0.03)	3.1(0.01)	N/A	16.1(0.22)	17.3(0.09)	N/A	1.8(0.05)	2.0(0.03)	N/A	12.6(0.39)	14.2(0.20)
	i.i.d.	N/A	2.8(0.04)	3.1(0.02)	N/A	15.0(0.26)	16.8(0.11)	N/A	1.5(0.05)	1.8(0.02)	N/A	10.5(0.38)	13.2(0.19)
400	0.1	N/A	5.8(0.44)	6.0(0.37)	N/A	63.9(4.29)	54.7(1.60)	N/A	5.1(0.46)	5.4(0.62)	N/A	54.4(4.12)	44.6(1.43)
	0.25	N/A	2.6(0.08)	2.5(0.05)	N/A	20.8(0.22)	20.0(0.19)	N/A	1.8(0.09)	1.7(0.14)	N/A	17.5(0.28)	15.5(0.11)
	0.5	N/A	2.9(0.06)	2.9(0.03)	N/A	19.0(0.31)	18.6(0.12)	N/A	2.0(0.06)	1.9(0.03)	N/A	17.3(0.55)	14.9(0.19)
	1	N/A	3.0(0.04)	3.1(0.02)	N/A	19.0(0.32)	19.9(0.13)	N/A	2.0(0.06)	2.0(0.02)	N/A	16.7(0.59)	16.5(0.20)
	2	N/A	3.1(0.03)	3.2(0.01)	N/A	19.0(0.24)	20.2(0.10)	N/A	1.9(0.05)	2.0(0.02)	N/A	15.9(0.50)	17.1(0.20)
	i.i.d.	N/A	2.9(0.04)	3.1(0.01)	N/A	17.9(0.31)	19.7(0.10)	N/A	1.7(0.06)	1.9(0.02)	N/A	13.5(0.48)	16.0(0.20)

Table 3.3: Comparison of average (SD) TPR(%)/FPR(%) for Models 2 & 4

		Model 2		Model 4	
p	α	Hard	Soft	CLIME	SPICE
100	0.1	10.86(4.35)/0.02(0.03)	54.19(4.41)/4.98(1.26)	91.28(2.76)/25.49(2.37)	82.99(2.76)/28.97(1.04)
	0.25	35.16(5.43)/0.07(0.06)	70.72(3.96)/6.10(1.16)	92.65(2.35)/17.82(1.84)	90.93(2.19)/29.68(1.31)
	0.5	48.43(3.76)/0.06(0.06)	80.43(3.19)/6.75(1.19)	95.30(1.73)/17.80(1.47)	96.00(1.54)/31.58(1.49)
	1	60.92(4.25)/0.02(0.03)	94.34(2.12)/7.23(1.39)	98.47(0.90)/14.37(1.21)	99.24(0.66)/30.65(1.49)
	2	83.93(4.08)/0.04(0.05)	99.33(0.73)/7.47(1.57)	99.71(0.36)/11.99(1.27)	99.94(0.17)/27.77(1.34)
	i.i.d.	93.42(2.63)/0.13(0.09)	99.91(0.21)/11.42(1.82)	99.91(0.20)/16.21(1.63)	99.99(0.07)/31.40(1.28)
200	0.1	5.57(2.93)/0.00(0.00)	45.91(3.86)/2.40(0.55)	82.24(2.70)/12.72(0.64)	76.07(1.95)/17.78(0.56)
	0.25	28.31(4.75)/0.02(0.02)	64.71(3.23)/3.20(0.69)	84.83(2.28)/15.70(2.62)	84.75(1.90)/18.87(0.59)
	0.5	44.48(3.02)/0.02(0.02)	74.38(2.42)/3.40(0.59)	89.55(2.39)/13.21(3.00)	91.65(1.45)/20.07(0.64)
	1	57.45(2.14)/0.01(0.01)	91.40(2.11)/3.84(0.81)	93.81(1.52)/7.27(0.58)	97.12(0.97)/19.07(0.85)
	2	79.04(3.66)/0.02(0.01)	98.71(0.67)/3.73(0.58)	97.77(0.97)/4.86(0.55)	99.31(0.42)/16.25(0.81)
	i.i.d.	90.74(2.68)/0.07(0.05)	99.68(0.31)/6.64(0.65)	99.56(0.36)/7.24(0.79)	99.88(0.18)/19.42(0.81)
300	0.1	4.15(2.50)/0.00(0.00)	40.61(3.94)/1.50(0.43)	82.60(3.59)/12.71(2.55)	71.66(1.71)/13.05(0.34)
	0.25	24.28(4.85)/0.01(0.01)	61.27(2.70)/2.13(0.42)	77.62(2.62)/14.39(2.62)	81.09(1.71)/14.06(0.39)
	0.5	41.75(3.51)/0.01(0.01)	71.65(2.51)/2.43(0.47)	82.23(2.48)/14.33(3.57)	88.71(1.44)/14.98(0.42)
	1	55.42(2.10)/0.00(0.00)	89.41(1.80)/2.61(0.44)	86.84(2.58)/4.71(0.67)	94.87(1.02)/14.20(0.54)
	2	74.39(3.23)/0.01(0.01)	98.11(0.69)/2.49(0.57)	94.88(1.38)/2.84(0.41)	98.27(0.68)/11.59(0.65)
	i.i.d.	88.97(2.29)/0.04(0.02)	99.57(0.34)/4.77(0.84)	98.83(0.49)/4.89(0.58)	99.56(0.29)/14.32(0.70)
400	0.1	2.65(1.29)/0.00(0.00)	36.80(2.27)/1.02(0.23)	83.04(2.84)/14.91(2.84)	68.51(1.49)/10.36(0.24)
	0.25	20.81(3.74)/0.01(0.00)	58.30(2.86)/1.54(0.35)	76.76(3.46)/15.11(3.40)	78.50(1.41)/11.41(0.32)
	0.5	40.14(3.58)/0.01(0.01)	68.74(2.06)/1.68(0.35)	78.58(2.35)/15.67(3.64)	86.19(1.44)/12.20(0.35)
	1	53.82(1.65)/0.00(0.00)	87.51(1.87)/1.80(0.40)	79.44(3.05)/4.40(0.77)	92.85(1.09)/11.55(0.41)
	2	72.19(2.58)/0.00(0.00)	97.79(0.66)/1.97(0.22)	90.47(2.32)/1.92(0.35)	96.68(0.85)/8.97(0.55)
	i.i.d.	87.51(1.65)/0.03(0.01)	99.38(0.30)/3.93(0.40)	97.63(0.82)/3.50(0.52)	99.09(0.39)/11.34(0.60)

CHAPTER IV

Estimation of Large Covariance and Precision Matrices from Multiple Independent Samples of Temporally Dependent Observations

4.1 Introduction

Group-level functional connectivity analysis is important for understanding the brain mechanisms underlying mental diseases (see, e.g., Tomson et al., 2015). We are interested in estimating the group-level functional connectivity of p brain nodes (regions or voxels) using n rfMRI images obtained from L subjects in a group of interest. Suppose that the L samples are independent, and each of the n images has the same mean and the same $p \times p$ covariance matrix $\mathbf{\Sigma}$. Our goal is to estimate $\mathbf{\Sigma}$ or the correlation matrix $\mathbf{R}$ for the marginal functional connectivity, and the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ for the direct functional connectivity.

A traditional estimator of $\mathbf{\Sigma}$ is the sample covariance matrix $\hat{\mathbf{\Sigma}}$ for the concatenation of all the n image observations (Smith et al., 2013; Ng et al., 2013). Although $\hat{\mathbf{\Sigma}}$ is not a consistent estimator for $\mathbf{\Sigma}$ when p grows with n (Bai and Yin, 1993; Bai and Silverstein, 2010), we can use it as the initial estimator of $\mathbf{\Sigma}$ in many consistent procedures for estimating $\mathbf{\Sigma}$, $\mathbf{R}$ and $\mathbf{\Omega}$. In this chapter, we focus on the generalized thresholding estimation of $\mathbf{\Sigma}$ and $\mathbf{R}$ as well as the SPICE and CLIME approaches of $\mathbf{\Omega}$ for the multiple independent samples of temporally dependent sub-Gaussian observations. We then apply these approaches to assessing the group-level functional connectivity of patients with attention deficit hyperactivity disorder (ADHD) compared to normal controls using the rfMRI data obtained from the ADHD-200 Preprocessed repository (neurobureau.projects.nitrc.org/ADHD200).

Multiple independent samples provide faster convergence rates. For example, when all the L samples have the same sample size n_1 and the same PDD temporal dependence $\mathcal{B}(C_0, \alpha)$ defined in (3.1) with $\alpha \in (0, 1)$, the convergence rates given in Chapter III by applying the PDD model directly to the total Ln_1 observations are mainly driven by the factor $\sqrt{(\log p)/(Ln_1)^\alpha}$, but we will show that using the independence of the samples, the convergence rates are primarily controlled by the factor $\sqrt{(\log p)/Ln_1^\alpha}$, a faster rate when $L \to \infty$. To achieve such an improvement, we use a different proof technique to that in Chapter III. Recall that in Chapter III, following the grouping idea of Bhattacharjee and Bose (2014), we established the desirable concentration inequality from a set of inequalities obtained for a careful partition of the temporal observations. In this chapter, we establish a different concentration inequality for the independent samples using the large deviation inequalities given in Vershynin (2012). For $L = 1$, this new proof yields faster rates under certain conditions for generalized thresholding estimation of $\mathbf{\Sigma}$ (or $\mathbf{R}$) and the SPICE estimation of $\mathbf{\Omega}$. For CLIME, however, the improvement is not guaranteed. Moreover, the considered family of temporal dependence in this chapter is characterized only by the autocorrelations of each time series without considering the cross-correlations that need some care in Chapter III.

The gap-block cross-validation was proposed in Chapter III for temporally dependent observations. With multiple independent samples, however, we no longer need gap-blocks. The usual k-fold cross validation that partitions independent samples can be applied.

We also discuss a potential way to improve the convergence rates by replacing the sample covariance matrix $\hat{\mathbf{\Sigma}}$ with a weighted sample covariance matrix in the estimation. Each sample is assigned a weight to be proportional with its effective sample size. Given appropriate weights, usually unknown in practice, using the weighted sample covariance matrix can theoretically achieve better convergence rates than using $\hat{\Sigma}$. However, to practically select such weights remains an open question.

In this chapter, we continue using the notation given in Chapter III if without further notification. Also we assume $p \to \infty$ as $n \to \infty$ and only use $n \to \infty$ in the asymptotic arguments. The rest of the chapter is organized as follows. In section 4.2, we introduce the sub-Gaussian data structure for the multiple independent samples. Section 4.3 provides the theoretical results for the considered estimators based on the sample covariance matrix. The performance of the estimators is evaluated by simulations in section 4.4. In section 4.5, we analyze the group-level functional connectivity of a ADHD group compared to a normal control group using the ADHD-200 rfMRI data. We end the chapter with a discussion of using the weighted sample covariance matrix. The detailed proofs for all the theoretical results are provided in Appendix C.

4.2 Data Structure

For each $\ell \in \{1, \dots, L\}$, we observe a p-variate time series $\boldsymbol{X}_1^{(\ell)}, \dots, \boldsymbol{X}_{n_\ell}^{(\ell)}$, where each $\boldsymbol{X}_i^{(\ell)}$ has mean $\boldsymbol{\mu}_p$, covariance matrix $\boldsymbol{\Sigma}$ and precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Write $\mathbf{X}_{p\times n_\ell}^{(\ell)} = (X_{ij}^{(\ell)})_{p\times n_\ell} = (\boldsymbol{X}_1^{(\ell)}, \dots, \boldsymbol{X}_{n_\ell}^{(\ell)})$. We simply call $\mathbf{X}_{p\times n_\ell}^{(\ell)}$ the ℓ-th sample of observations, and assume such L samples are independent. In the application of fMRI data, $\mathbf{X}_{p\times n_\ell}^{(\ell)}$ can be viewed as the n_ℓ images of p prespecified brain nodes for the ℓ-th subject, and n_ℓ are usually equal for all ℓ.

Throughout the chapter, we assume that each sample $\boldsymbol{X}_{pn_\ell}^{(\ell)} = \mathrm{vec}(\mathbf{X}_{p\times n_\ell}^{(\ell)})$ is obtained from the following with its own linear filter $\mathbf{H}^{(\ell)}$:

$$\boldsymbol{X}_{pn_\ell}^{(\ell)} = \mathbf{H}^{(\ell)}\boldsymbol{e} + \mathbf{1}_{n_\ell} \otimes \boldsymbol{\mu}_p, \tag{4.1}$$

where the sub-Gaussian random vector $\boldsymbol{e} = (e_1, e_2, \dots)^T$ with dimension $m = \infty$ is the same for all samples, which is the same as in (3.3).

For the time series $X_{i1}^{(\ell)}, \ldots, X_{i,n_\ell}^{(\ell)}$, let $\mathbf{\Theta}_i^{(\ell)} = (\theta_{i,jk}^{(\ell)})_{n_\ell \times n_\ell}$ be the autocorrelation matrix with $\theta_{i,jk}^{(\ell)} = \text{corr}(X_{ij}^{(\ell)}, X_{ik}^{(\ell)})$. Define

$$g_\ell = \max_{1 \le i \le p} \|\mathbf{\Theta}_i^{(\ell)}\|_2. \tag{4.2}$$

By

$$\|\mathbf{\Theta}_i^{(\ell)}\|_F^2 / n_\ell \le \|\mathbf{\Theta}_i^{(\ell)}\|_2 \le \|\mathbf{\Theta}_i^{(\ell)}\|_1, \tag{4.3}$$

we have $g_\ell \in [1, n_\ell]$. We can see that $g_\ell = 1$ if all the p time series of the ℓ-th sample are white noise processes, and $g_\ell = n_\ell$ if every pair of data points in a univariate time series are perfectly correlated or anti-correlated. The quantity g_ℓ naturally reflects the maximum strength of temporal dependence within the ℓ-th sample. We shall see that $\{g_\ell\}$ are involved in the tuning parameters of the considered estimating procedures.

4.3 Estimating Methods Based on the Sample Covariance Matrix

In this section we study several estimating methods for $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ based on the sample covariance matrix defined by

$$\hat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i=1}^{n_\ell} \left(\boldsymbol{X}_i^{(\ell)} \right)^{\otimes 2} - \left(\frac{1}{n} \sum_{\ell=1}^{L} \sum_{i=1}^{n_\ell} \boldsymbol{X}_i^{(\ell)} \right)^{\otimes 2} \tag{4.4}$$

with $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^T$ for vector $\boldsymbol{v}$. Define $g_{\max} = \max_{1 \le \ell \le L} g_\ell$ and

$$\tau_1 = \max \left\{ \sqrt{\frac{\log p}{n} \sum_{\ell=1}^{L} \frac{n_\ell}{n} g_\ell}, \frac{g_{\max} \log p}{n} \right\}. \tag{4.5}$$

We assume $\tau_1 = o(1)$ in the following. Then $\tau_1 = O(\sqrt{g_{\max}(\log p)/n})$.

4.3.1 Main Results

Theorem IV.1 (Generalized thresholding estimation of $\mathbf{\Sigma}$ and $\mathbf{R}$)**.** (a). *For any data* $\{\mathbf{X}_{p \times n_\ell}^{(\ell)}\}_{\ell=1}^{L}$ *generated from* (4.1)*, uniformly on* $\mathbf{\Sigma} \in \mathcal{U}(q, c_p, v_0)$ *where* $\mathcal{U}$ *is defined* *in* (3.6)*, for sufficiently large constant $M > 0$, if $\tau = M\tau_1$ and $\tau_1 = o(1)$, then*

$$\begin{aligned} |S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}|_\infty &= O_P(\tau_1), \\ \|S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_2 &= O_P\left(c_p\tau_1^{1-q}\right), \\ \frac{1}{p}\|S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_F^2 &= O_P\left(c_p\tau_1^{2-q}\right), \end{aligned}$$

and $s_\tau(\hat{\sigma}_{ij}) = 0$ for all (i, j) where $\sigma_{ij} = 0$ with probability tending to 1. When all nonzero elements of $\mathbf{\Sigma}$ satisfy $|\sigma_{ij}| \geq 2\tau$, then $\mathrm{sign}(s_\tau(\hat{\sigma}_{ij})) = \mathrm{sign}(\sigma_{ij})$ for all (i, j) where $\sigma_{ij} \neq 0$ with probability tending to 1. Moreover, if $p \geq n^c$ for some constant $c > 0$, then

$$\begin{aligned} E\left(|S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}|_\infty^2\right) &= O(\tau_1^2), \\ E\left(\|S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_2^2\right) &= O\left(c_p^2\tau_1^{2-2q}\right), \\ \frac{1}{p}E\left(\|S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_F^2\right) &= O\left(c_p\tau_1^{2-q}\right). \end{aligned}$$

(b). *Part* (a) *holds with $\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}, \hat{\sigma}_{ij}, \sigma_{ij}$ and $\mathcal{U}(q, c_p, v_0)$ replaced by $\hat{\mathbf{R}}, \mathbf{R}, \hat{\rho}_{ij}, \rho_{ij}$ and $\mathcal{R}(q, c_p)$, respectively.*

Theorem IV.2 (SPICE of $\mathbf{\Omega}$)**.** (a). *For any data $\{\mathbf{X}^{(\ell)}_{p\times n_\ell}\}_{\ell=1}^L$ generated from* (4.1)*, uniformly on $\mathbf{\Omega} \in \mathcal{G}_2(s_p, v_0)$ where $\mathcal{G}_2$ is defined in* (3.22)*, for sufficiently large constant $M > 0$, if $\lambda_2 = M\tau_1$ and $\tau_1 = o(1/\sqrt{1 + s_p})$, then*

$$\begin{aligned} \|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_F &= O_P(\tau_1\sqrt{s_p}), \\ \|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_2 &= O_P(\tau_1\sqrt{1 + s_p}), \\ \frac{1}{\sqrt{p}}\|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_F &= O_P\left(\tau_1\sqrt{1 + s_p/p}\right). \end{aligned}$$

When $\|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_2 = O_P(\eta)$ with $\eta = O(\tau_1)$, then with probability tending to 1, we have $\hat{\omega}_{ij\lambda_2} = 0$ for all (i, j) where $\omega_{ij} = 0$.

(b). *If the conditions $\lambda_2 = M\tau_1$ and $\tau_1 = o(1/\sqrt{1 + s_p})$ in part* (a) *are replaced by $\lambda_2 = 8M\tau_1/\beta \leq [6(1 + \beta/8)d\max\{\kappa_{\mathbf{R}}\kappa_{\mathbf{\Gamma}}, \kappa_{\mathbf{R}}^3\kappa_{\mathbf{\Gamma}}^2\}]^{-1}$ and $\tau_1 = o(\min\{1, [(1 +$*

$8/\beta)\kappa_{\mathbf{\Gamma}}]^{-1}\})$, *let* $r = (0.5 + 2.5(1 + 8/\beta)\kappa_{\mathbf{\Gamma}})\, M\tau_1 v_0$, *then under the irrepresentability condition* (3.26)*, with probability tending to 1,*

$$|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}|_\infty \le r,$$

$$\|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_2 \le r \min\left\{d, \sqrt{p + s_p}\right\},$$

$$\frac{1}{\sqrt{p}}\|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_F \le r\sqrt{1 + s_p/p},$$

and $\hat{\omega}_{ij\lambda_2} = 0$ *for all* (i, j) *with* $\omega_{ij} = 0$*, and moreover,* $\text{sign}(\hat{\omega}_{ij\lambda_2}) = \text{sign}(\omega_{ij})$ *for all* (i, j) *where* $\omega_{ij} \neq 0$ *when all nonzero elements of* $\mathbf{\Omega}$ *satisfy* $|\omega_{ij}| > r$.

For CLIME, we consider the following set of precision matrices

$$\mathcal{G}_1^*(q, c_p, M_p, v_0) = \left\{\mathbf{\Omega} \succ 0 : \max_{1\le i\le p} \sum_{j=1}^{p} |\omega_{ij}|^q \le c_p, \|\mathbf{\Omega}\|_1 \le M_p, \max_{1\le i\le p} \sigma_{ii} \le v_0\right\},$$

for $0 \le q < 1$. The set $\mathcal{G}_1^*(q, c_p, M_p, v_0)$ is the original one considered by CLIME in Cai et al. (2011) for i.i.d. observations. It is a modified version of $\mathcal{G}_1(q, c_p, M_p, v_0)$ given in (3.16) without the condition $\max_{1\le i\le p} \omega_{ii} \le v_0$, which was useful for the proof of the consistency of CLIME given in Theorem III.3 for a single sample with PDD dependence. The new proof considered here no longer needs this extra condition. Let the tuning parameter $\xi \ge 4M_p\lambda_1$ for the hard-thresholded CLIME estimator $\tilde{\mathbf{\Omega}}_\varepsilon$. Although how to select an appropriate ξ in practice is unclear, it is still of interest to present the nice properties of sparsistency and sign-consistency for $\tilde{\mathbf{\Omega}}_\varepsilon$.

Theorem IV.3 (CLIME of $\mathbf{\Omega}$)**.** *For any data* $\{\mathbf{X}^{(\ell)}_{p\times n_\ell}\}_{\ell=1}^{L}$ *generated from* (4.1)*, uniformly on* $\mathbf{\Omega} \in \mathcal{G}_1^*(q, c_p, M_p, v_0)$*, for sufficiently large constant* $M > 0$*, if* $\lambda_1 = M\tau_1 M_p$ *and* $0 \le \varepsilon \le \tau_1$ *with* $\tau_1 = o(1)$*, then*

$$|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}|_\infty = O_P(M_p^2\tau_1), \tag{4.6}$$

$$\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_2 = O_P\left(c_p(M_p^2\tau_1)^{1-q}\right), \tag{4.7}$$

$$\frac{1}{p}\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_F^2 = O_P\left(c_p(M_p^2\tau_1)^{2-q}\right), \tag{4.8}$$

and $\tilde{\omega}_{ij\varepsilon} = 0$ *for all* (i, j) *where* $\omega_{ij} = 0$ *with probability tending to 1. When all nonzero elements of* $\mathbf{\Omega}$ *satisfy* $|\omega_{ij}| > \xi + 4M_p\lambda_1$, $\operatorname{sign}(\tilde{\omega}_{ij\varepsilon}) = \operatorname{sign}(\omega_{ij})$ *for all* (i, j) *where* $\omega_{ij} \neq 0$ *with probability tending to 1. Moreover, if* $p \geq n^c$ *with some constant* $c > 0$*, then for any constant* $C > 0$*, there exists a constant* $M' > 0$ *such that when* $M > M'$ *and* $\min\left\{p^{-C}, \tau_1\right\} \leq \varepsilon \leq \tau_1$*, we have*

$$\begin{aligned}
E\left(|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}|_\infty^2\right) &= O\left((M_p^2\tau_1)^2\right), \\
E\left(\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_2^2\right) &= O\left(c_p^2(M_p^2\tau_1)^{2-2q}\right), \\
\frac{1}{p}E\left(\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_F^2\right) &= O\left(c_p(M_p^2\tau_1)^{2-q}\right).
\end{aligned}$$

4.3.2 Comparison to the Results in Chapter III for $L = 1$ under the PDD dependence

In this subsection, we compare the convergence rates of the considered estimators obtained in this chapter with those in Chapter III for the single-sample data (i.e., the case with $L = 1$) under the PDD dependence defined by (3.1). As mentioned in the introduction section, different proof techniques are used for the theoretical results in the two chapters, so we have two slightly different sets of convergence rates.

The PDD dependence $\mathcal{B}(C_0, \alpha)$ satisfies

$$\max_{1\leq i\leq p} |\theta_{i,jk}^{(1)}| \leq C_1(|j-k|+1)^{-\alpha} \quad \text{for } j, k = 1, \ldots, n, \tag{4.9}$$

with a constant C_1 dependent on C_0. Then

$$g_{\max} = g_1 \leq 2C_1 \times \begin{cases} (n^{1-\alpha} - \alpha)/(1-\alpha), & \alpha \neq 1, \\ 1 + \log n, & \alpha = 1, \end{cases}$$

thus $\tau_1 = O(\tau(n,\alpha))$ with

$$\tau(n,\alpha) := \begin{cases} \sqrt{(\log p)/n^{\alpha}}, & 0 < \alpha < 1, \\ \sqrt{(\log n)(\log p)/n}, & \alpha = 1, \\ \sqrt{(\log p)/n}, & \alpha > 1. \end{cases}$$

Remark IV.1. Consider the generalized thresholding estimators of $\mathbf{\Sigma}$ and $\mathbf{R}$ as well as the SPICE estimators of $\mathbf{\Omega}$. Recall that their convergence rates obtained in Chapter III are mainly determined by

$$\tau' \asymp \begin{cases} \sqrt{(\log p + \log n)/n^{\alpha}}, & 0 < \alpha < 1, \\ \sqrt{(\log n)(\log p + \log\log n)/n}, & \alpha = 1, \\ \sqrt{(\log p)/n}, & \alpha > 1. \end{cases}$$

By comparing $\tau_1 = O(\tau(n,\alpha))$ with τ', we see that the convergence rates yielded by Theorems IV.1 and IV.2 are sharper than those given in Chapter III for $\alpha \in (0,1)$ when $\log p = o(\log n)$ and for $\alpha = 1$ when $\log p = o(\log\log n)$, otherwise they are the same.

Remark IV.2. Consider the CLIME estimator of $\mathbf{\Omega}$. The convergence rates given by Theorems III.3 and IV.3 under the same norm can be written by the same function of c_p, q, M_p and λ_1. Recall that $\lambda_1 \asymp M_p^{\mathbb{I}(0<\alpha\leq 1)+\mathbb{I}(\alpha>1)/\alpha}\tau'$ in Theorem III.3, and $\lambda_1 \asymp M_p\tau_1 = O(M_p\tau(n,\alpha))$ in Theorem IV.3. For the scenario when $\alpha > 1$ and M_p grows with p, the convergence rates given in Theorem III.3 are faster than those deduced from Theorem IV.3. Otherwise, similar to Remark IV.1, the rates obtained in Theorem IV.3 can be better than those in Theorem III.3.

4.3.3 Some General Remarks

Remark IV.3. If $g_{\max} < C$ with a constant C, then $\tau_1 \asymp \sqrt{(\log p)/n}$, so all the results given in subsection 4.3.1 except Theorem IV.2 (b) reduce to those for i.i.d. observations given in the literature (Rothman et al., 2009; Cai et al., 2011; Rothman et al., 2008; Lam and Fan, 2009). If additionally assume β is constant and $\max\{\kappa_{\mathbf{R}}, \kappa_{\mathbf{\Gamma}}\}$ has a constant upper bound, then the convergence rates of the correlation-based SPICE given in Theorem IV.2 (b) are the same as those shown in Ravikumar et al. (2011) for the covariance-based SPICE obtained by i.i.d. observations.

The quantities $\{g_\ell\}_{\ell=1}^L$ sometimes are more suitable to characterize the temporal dependence than the PDD model. Examples are given in the following two remarks.

Remark IV.4. A unvariate stationary time series is said to be short-memory if the matrix ℓ_1 norm of its autocorrelation matrix is bounded by a constant. For the single-sample data, i.e., $L = 1$, the PDD model sometimes cannot give suitably theoretical convergence rates of the considered estimators for certain short-memory dependence with spiked autocorrelations. Figure 4.1 illustrates such an example, where the autocorrelation function of a short-memory time series has a finite number of spikes. Fitting a PDD model yields the parameter $\alpha \leq 0.2$. Note that the convergence rates for $\alpha \in (0, 1)$ are mainly driven by the factor $\sqrt{(\log p)/n^\alpha}$. However, if the other $(p - 1)$ univariate time series are also short-memory such that $g_{\max} < C$ with a constant C, from Remark IV.3 we have a much smaller factor $\sqrt{(\log p)/n}$, leading to faster convergence rates.

Remark IV.5. Suppose all the L samples have the same sample size n_1, which is common for fMRI studies, and also satisfy the PDD model in (3.1) with common C_0 and α. We can have $\tau_1 = O(\tau(n_1, \alpha)/\sqrt{L})$. If ignoring the independence among the multiple samples and applying the PDD model to all the Ln_1 observations, we have $\tau_1 = O(\tau(Ln_1, \alpha))$. Note that $\tau(n_1, \alpha)/\sqrt{L} \leq \tau(Ln_1, \alpha)$ with equality only when $\alpha > 1$. When $\alpha \in (0, 1)$, $\tau(n_1, \alpha) = \tau(Ln_1, \alpha)/L^{(1-\alpha)/2} = o(\tau(Ln_1, \alpha))$ if $L \to \infty$, yielding sharper convergence rates.

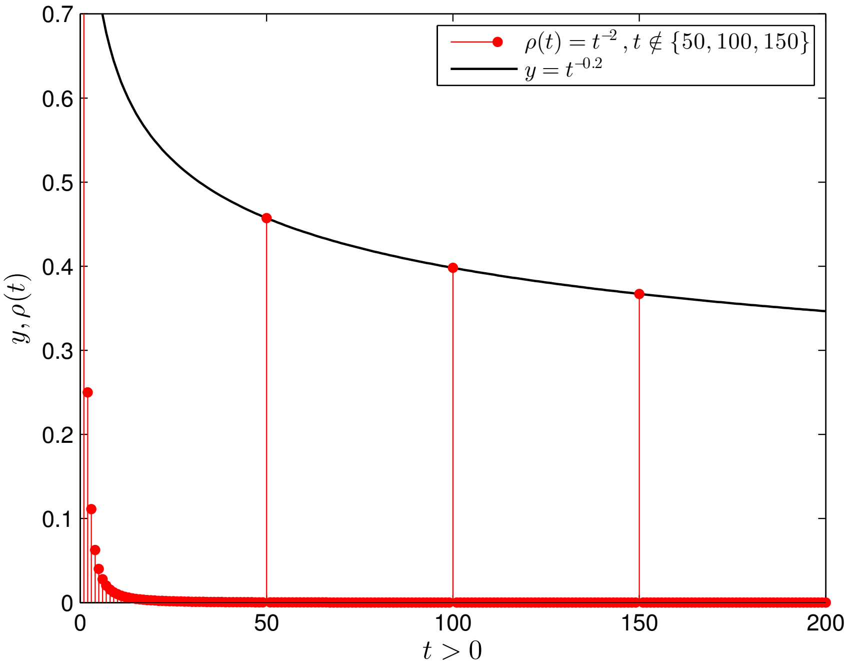


Figure 4.1: Autocorrelation function $\rho(t), t > 0$ of a stationary short-memory time series.

4.4 Simulation Studies

In this section, we evaluate the performance of the sample-covariance-matrix-based estimating methods for the multiple samples of temporally dependent observations. We only consider the hard and soft thresholding estimators of $\mathbf{R}$, and the SPICE estimator of $\mathbf{\Omega}$ computed by the R package `QUIC` (version 1.1; Hsieh et al., 2014). The R package `flare` (version 1.5.0; Li et al., 2015) for computing the CLIME estimator of $\mathbf{\Omega}$ is too slow for simulations, and the R package `fastclime` (version 1.2.5; Pang et al., 2014) is not stable. Hence the CLIME estimator is not considered in our numerical examples.

We generate L=6 samples, each of sample size 200, and thus $n = 1200$. Each setting is simulated with 100 replications. The temporally dependent observations are generated by the same method used in Subsection 3.5.2 from a zero-mean Gaussian distribution with the same model for $\mathbf{\Sigma}$ or $\mathbf{\Omega}$. For each model of $\mathbf{\Sigma}$ or $\mathbf{\Omega}$, we consider the following three

Table 4.1: Comparison of average (SD) matrix losses for correlation matrix estimation

		Spectral norm			Frobenius norm		
p	α	$\widehat{\mathbf{R}}$	Hard	Soft	$\widehat{\mathbf{R}}$	Hard	Soft
				Model 1			
200	0.1	29.28(2.006)	2.934(0.039)	2.834(0.033)	52.87(1.355)	14.56(0.176)	13.53(0.170)
	0.25	16.52(1.392)	2.219(0.138)	2.344(0.052)	30.86(1.073)	10.46(0.294)	10.37(0.229)
	0.5	7.493(0.634)	1.315(0.081)	1.643(0.052)	16.42(0.484)	6.248(0.234)	6.609(0.182)
	1	2.550(0.194)	0.722(0.049)	1.055(0.041)	8.654(0.139)	3.420(0.094)	3.882(0.104)
	2	1.506(0.082)	0.536(0.041)	0.814(0.036)	6.185(0.067)	2.513(0.063)	2.897(0.076)
	Mixed	12.62(1.447)	1.693(0.205)	2.010(0.249)	19.67(1.088)	7.810(0.452)	7.765(0.303)
	i.i.d.	1.366(0.081)	0.507(0.042)	0.772(0.038)	5.732(0.063)	2.327(0.059)	2.716(0.075)
400	0.1	55.64(2.527)	2.978(0.013)	2.901(0.018)	106.1(1.928)	20.88(0.105)	19.74(0.179)
	0.25	31.74(1.745)	2.395(0.125)	2.494(0.043)	62.37(1.372)	15.71(0.525)	15.73(0.254)
	0.5	14.20(0.795)	1.447(0.076)	1.792(0.040)	33.15(0.614)	9.430(0.268)	10.19(0.203)
	1	4.288(0.210)	0.800(0.052)	1.164(0.038)	17.35(0.141)	5.113(0.102)	6.031(0.122)
	2	2.322(0.076)	0.589(0.037)	0.916(0.031)	12.41(0.071)	3.748(0.102)	4.527(0.082)
	Mixed	25.25(1.963)	1.848(0.187)	2.507(0.388)	39.69(1.343)	11.85(0.490)	12.08(0.324)
	i.i.d.	2.072(0.074)	0.554(0.037)	0.855(0.025)	11.51(0.059)	3.512(0.067)	4.213(0.076)
				Model 2			
200	0.1	29.18(1.674)	1.806(0.047)	1.734(0.018)	53.05(1.216)	12.90(0.235)	12.00(0.163)
	0.25	16.40(1.186)	1.411(0.146)	1.384(0.048)	30.92(1.025)	8.275(0.425)	8.847(0.209)
	0.5	7.410(0.573)	0.730(0.105)	0.881(0.041)	16.44(0.464)	4.442(0.282)	5.223(0.170)
	1	2.456(0.161)	0.254(0.071)	0.487(0.024)	8.664(0.123)	1.020(0.075)	2.758(0.068)
	2	1.409(0.057)	0.132(0.016)	0.353(0.017)	6.191(0.061)	0.686(0.029)	1.964(0.048)
	Mixed	12.63(1.344)	0.948(0.175)	1.517(0.279)	19.69(1.008)	5.934(0.302)	6.361(0.313)
	i.i.d.	1.266(0.054)	0.122(0.015)	0.329(0.016)	5.739(0.056)	0.644(0.031)	1.832(0.045)
400	0.1	55.16(2.297)	1.817(0.042)	1.767(0.012)	106.1(1.904)	18.59(0.137)	17.59(0.174)
	0.25	31.55(1.518)	1.671(0.234)	1.521(0.124)	62.44(1.290)	12.72(0.604)	13.46(0.216)
	0.5	14.06(0.704)	0.842(0.191)	0.995(0.078)	33.16(0.636)	7.029(0.278)	8.147(0.167)
	1	4.178(0.198)	0.314(0.050)	0.524(0.019)	17.36(0.135)	1.497(0.087)	4.274(0.067)
	2	2.182(0.059)	0.139(0.014)	0.381(0.014)	12.42(0.062)	0.982(0.032)	3.074(0.054)
	Mixed	25.31(1.926)	1.092(0.220)	2.025(0.322)	39.79(1.293)	8.792(0.222)	9.953(0.291)
	i.i.d.	1.923(0.049)	0.130(0.014)	0.353(0.014)	11.51(0.050)	0.914(0.035)	2.823(0.045)

scenarios for the L=6 samples:

1. Same α: all samples have the same $\alpha \in \{0.1, 0.25, 0.5, 1, 2\}$;
2. Mixed α: $\alpha = 0.25$ for the first and second samples, $\alpha = 0.5$ for the third and fourth samples, and $\alpha = 1$ for the rest two samples;
3. The i.i.d. case.

Two different dimensions are considered: $p = 200$ and $p = 400$. The tuning parameter for each simulated data set is prepared with 50 different candidate values (see Appendix B.3). We run 6-fold cross-validation with data naturally partitioned by the 6 independent samples. The estimation performance is measured by both the spectral norm and the Frobenius

Table 4.2: Comparison of average (SD) matrix losses for precision matrix estimation

		Spectral norm		Frobenius norm	
p	α	$\widehat{\mathbf{\Sigma}}^{-1}$	SPICE	$\widehat{\mathbf{\Sigma}}^{-1}$	SPICE
		Model 3			
200	0.1	78.11(3.163)	2.289(0.134)	273.5(3.776)	16.07(0.332)
	0.25	20.15(0.957)	2.381(0.052)	70.64(1.073)	10.82(0.138)
	0.5	9.102(0.407)	2.640(0.036)	31.35(0.455)	11.28(0.184)
	1	4.499(0.236)	2.542(0.026)	14.84(0.251)	10.65(0.129)
	2	2.961(0.158)	2.249(0.027)	9.657(0.159)	9.130(0.129)
	Mixed	8.937(0.399)	2.523(0.037)	30.46(0.451)	10.69(0.160)
	i.i.d.	2.696(0.158)	2.181(0.027)	8.738(0.163)	8.779(0.126)
400	0.1	190.6(6.810)	2.276(0.139)	870.4(8.239)	22.46(0.356)
	0.25	48.64(1.609)	2.570(0.029)	218.3(2.091)	16.33(0.125)
	0.5	21.61(0.768)	2.798(0.025)	95.40(0.964)	17.13(0.178)
	1	10.66(0.357)	2.718(0.018)	45.22(0.513)	16.42(0.129)
	2	6.995(0.243)	2.538(0.019)	28.96(0.347)	15.03(0.145)
	Mixed	21.80(0.732)	2.700(0.027)	93.99(0.991)	16.38(0.175)
	i.i.d.	6.287(0.224)	2.475(0.021)	25.91(0.300)	14.56(0.152)
		Model 4			
200	0.1	67.94(2.142)	2.511(0.305)	264.0(3.032)	14.49(0.524)
	0.25	17.56(0.649)	1.114(0.063)	68.29(0.841)	6.264(0.134)
	0.5	8.009(0.319)	1.271(0.048)	30.46(0.369)	6.745(0.202)
	1	4.015(0.172)	1.130(0.031)	14.55(0.217)	6.276(0.155)
	2	2.677(0.110)	0.942(0.026)	9.520(0.142)	5.260(0.130)
	Mixed	7.867(0.314)	1.181(0.049)	29.58(0.395)	6.042(0.196)
	i.i.d.	2.453(0.123)	0.897(0.024)	8.617(0.142)	5.005(0.129)
400	0.1	169.6(5.240)	2.306(0.242)	841.5(7.070)	18.61(0.467)
	0.25	43.45(1.252)	1.345(0.056)	211.4(1.765)	10.15(0.183)
	0.5	19.41(0.613)	1.467(0.029)	92.66(0.830)	11.33(0.218)
	1	9.630(0.282)	1.342(0.028)	44.18(0.440)	10.82(0.185)
	2	6.385(0.200)	1.146(0.023)	28.43(0.312)	9.284(0.163)
	Mixed	19.67(0.627)	1.395(0.040)	91.30(0.840)	10.42(0.260)
	i.i.d.	5.783(0.190)	1.092(0.022)	25.47(0.256)	8.826(0.172)

Table 4.3: Comparison of average (SD) TPR(%)/FPR(%) for Models 2 & 4

		Model 2		Model 4
p	α	Hard	Soft	SPICE
200	0.1	5.85(3.03)/0.01(0.01)	39.92(3.52)/2.93(0.68)	95.88(0.96)/26.31(0.67)
	0.25	43.68(3.24)/0.04(0.02)	73.75(2.46)/4.62(0.62)	99.83(0.19)/30.68(0.78)
	0.5	77.88(3.81)/0.05(0.03)	98.24(0.79)/5.82(0.78)	100.00(0.03)/29.95(0.79)
	1	99.85(0.21)/0.00(0.00)	100.00(0.00)/5.70(1.01)	100.00(0.00)/27.25(1.01)
	2	100.00(0.00)/0.00(0.00)	100.00(0.00)/6.13(1.14)	100.00(0.00)/26.95(0.99)
	Mixed	57.11(4.88)/0.02(0.02)	93.85(2.19)/5.36(0.68)	100.00(0.03)/30.71(0.87)
	i.i.d.	100.00(0.00)/0.00(0.00)	100.00(0.00)/5.60(1.02)	100.00(0.00)/26.96(0.92)
400	0.1	3.08(1.15)/0.00(0.00)	32.42(3.43)/1.38(0.47)	92.49(0.98)/15.83(0.40)
	0.25	38.31(3.88)/0.02(0.01)	67.78(2.09)/2.41(0.37)	99.56(0.23)/19.25(0.49)
	0.5	70.63(3.17)/0.02(0.01)	96.92(0.81)/3.15(0.44)	99.99(0.03)/19.55(0.54)
	1	99.77(0.18)/0.00(0.00)	100.00(0.00)/3.15(0.35)	100.00(0.00)/17.69(0.67)
	2	100.00(0.00)/0.00(0.00)	100.00(0.00)/2.91(0.54)	100.00(0.00)/17.42(0.65)
	Mixed	52.73(1.67)/0.01(0.00)	89.86(1.92)/2.74(0.38)	99.99(0.03)/19.81(0.63)
	i.i.d.	100.00(0.00)/0.00(0.00)	100.00(0.00)/3.15(0.27)	100.00(0.00)/17.72(0.73)

norm. The sparsity recovering ability is evaluated by the TPR and FPR defined in Subsection 3.5.2.

The simulation results are summarized in Tables 4.1-4.3. We see that the sample correlation matrix and the inverse of sample covariance matrix have the worst performance. For the considered estimating approaches based on the sample covariance matrix, the overall pattern is similar to what we observed in Subsection 3.5.2 for $L = 1$, with reduced matrix losses due to the larger sample size.

4.5 Real Data Analysis

Attention deficit hyperactivity disorder (ADHD) is a neurodevelopmental disorder affecting about 7.2% chirden worldwide (Thomas et al., 2015). ADHD can be divided into three different types based on symptom presentation: predominantly inattentive type, predominantly hyperactive-impulsive type, and combined type. The combined type is the most common type of ADHD. We thus analyze the group-level functional connectivity of children with combined type ADHD (ADHD-C) compared with normal controls (NC) using the rfMRI data obtained from the ADHD-200 Preprocessed repository (neurobureau.projects.nitrc.org/ADHD200). The data set contains images with 351 regions of interest (ROIs) from 15 boys with ADHD-C and 15 age-matched healthy boys. All the subjects are medication naïve and right-handed with age between 9 and 15 years. The rfMRI data have been preprocessed by The Neuro Bureau using the Athena pipeline (see details on the above website). Each subject has 232 temporal images. Thus, for either ADHD-C group or NC group, $p = 351$, $L = 15$, and $n_1 = \cdots = n_L = 232$. The time series of each ROI is the spatially averaged time series of all the voxels within the ROI. Following the suggestion of Ng et al. (2013), we normalize each subject's time series by subtracting its sample mean and dividing by its sample standard deviation to reduce the

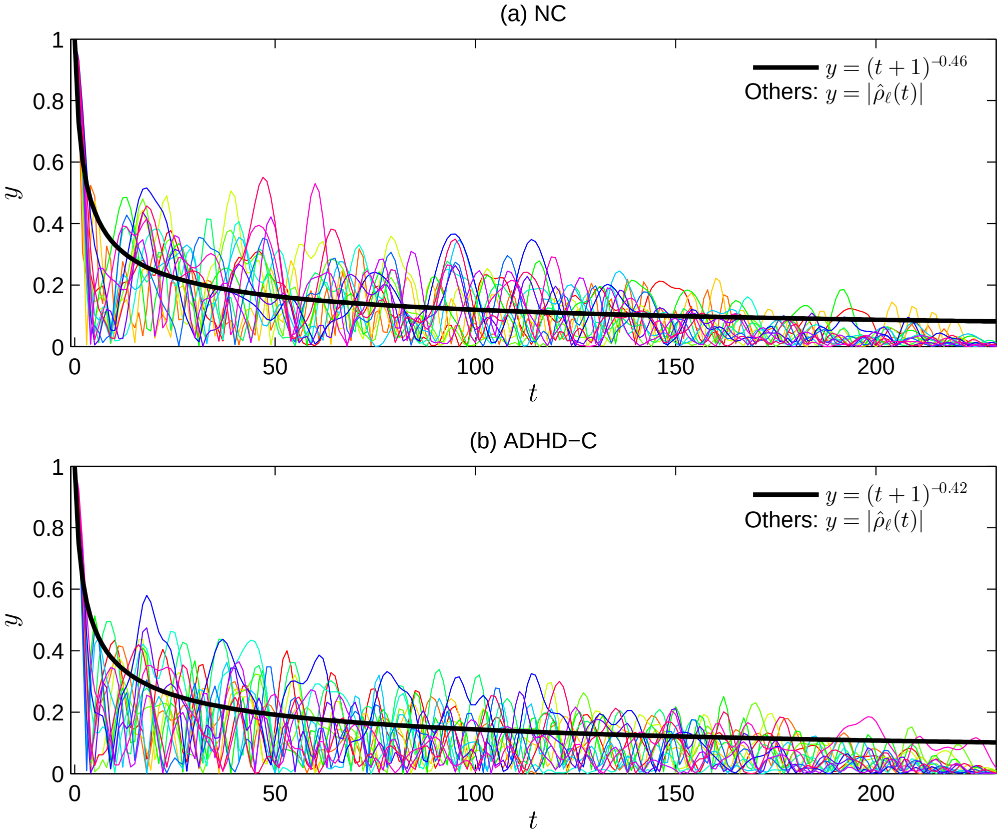


Figure 4.2: Absolute values of sample autocorrelation functions. For each univariate time series, the model $(t+1)^{-\alpha}$ is fitted to its absolute sample autocorrelations. Here $|\hat{\rho}_\ell(t)|$ represents the absolute sample autocorrelation function corresponding to the time series with the smallest fitted α among all the p time series of the ℓ-th subject. The wide solid line is the fitted curve $(t+1)^{-\hat{\alpha}_{\min}}$, where $\hat{\alpha}_{\min}$ is the smallest fitted α among all the L subjects.

inter-subject variability.

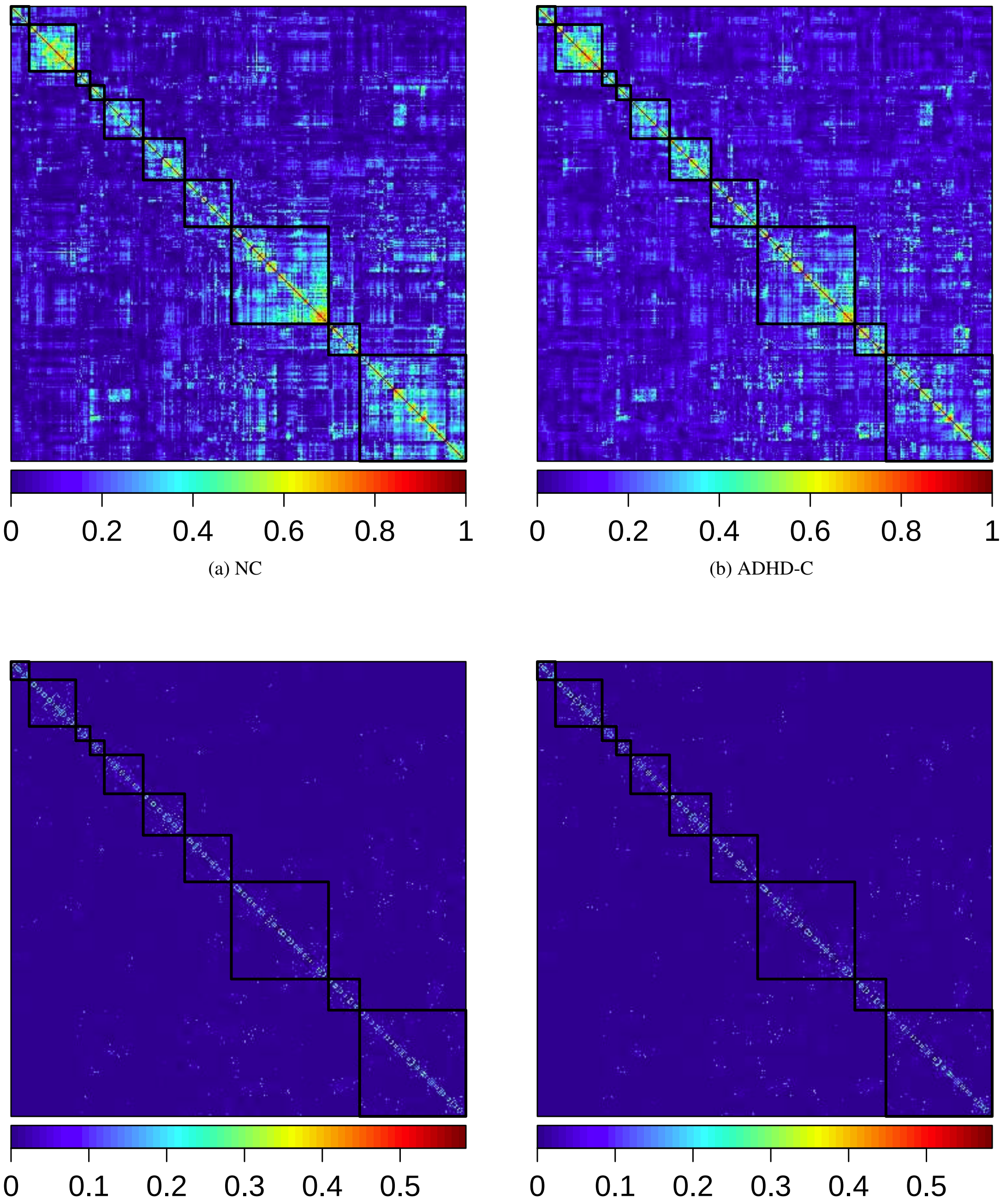
We illustrate the temporal dependence in each group using a rough estimation of the upper bound of $g_{\max}$. Because the sample autocorrelation matrix is not a consistent estimator of the true autocorrelation matrix under the spectral norm (Wu and Pourahmadi, 2009, Theorem 1), it is not appropriate to apply the spectral norm of each sample autocorrelation matrix for the estimation of $g_{\max}$. Instead, we first fit the absolute values of each sample autocorrelation function by $(t+1)^{-\alpha}$ using the nonlinear least-squares method. Denote $\hat{\alpha}_{\min}$ to be the smallest fitted α that is obtained from all the $L \times p$ time series of one of the groups. Then approximate $\max_{\ell \le L} \max_{i \le p} \|\mathbf{\Theta}_i^{(\ell)}\|_1$ by the matrix ℓ_1 norm of $\big((t+1)^{-\hat{\alpha}_{\min}}\big)_{232\times 232}$, which gives an estimated upper bound for $g_{\max}$ following from (4.3). We obtain $\hat{\alpha}_{\min} = 0.46$ for the NC group and $\hat{\alpha}_{\min} = 0.42$ for the ADHD-C group, and the corresponding $g_{\max}$ is roughly bounded by 44.8 and 51.2, respectively. More details are provided in Figure 4.2.

We estimate the correlation matrix and the partial correlation matrix for the marginal and the direct functional connectivities of the 351 ROIs, respectively. The correlation matrix is estimated by the hard thresholding method. We estimate the (i, j)-th off-diagonal entry of the partial correlation matrix by $-\hat{\omega}_{ij}/\sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}$ using the SPICE estimator $\hat{\mathbf{\Omega}} := (\hat{\omega}_{ij})_{p\times p}$ of the precision matrix. The optimal value of each tuning parameter is selected from 50 different candidates (see Appendix B.3 with the largest candidate value to be 1) by using the 5-fold cross-validation that randomly divides the 15 samples into 5 groups, each with 3 samples, together with the one-standard-error rule (Hastie et al., 2009). We find that about 46.5% and 63.1% of pairs of ROIs are marginally connected for the NC and the ADHD-C groups with threshold values of sample correlations around 0.105 and 0.066, respectively. Although the ADHD-C group has a larger number of nonzero estimated correlations, the average of the absolute values of those nonzeros is 0.170 with standard deviation 0.107, which is smaller than 0.215 for the NC group with standard deviation 0.111. In terms of the direct connectivity, about 12.6% of all the pairs of ROIs are connected for the NC group, and 11.2% for the ADHD-C group. The averages of the absolute values of nonzero estimated partial correlations are around 0.034 for both groups with standard deviation 0.055 for NC and 0.056 for ADHD-C.

We reorder the estimated correlation matrix of the NC group using the average linkage hierarchical clustering method (Everitt et al., 2011) based on the dissimilarity measure $d_{ij} = 1 - |s_\tau^H(\hat{\rho}_{ij})|$ for the (i, j)-th entry, so that entries with large absolute values $|s_\tau^H(\hat{\rho}_{ij})|$

are clustered around the diagonal. The resulting order is also applied to the other three estimated matrices for ease of comparison. The heat maps of the reordered estimated matrices are shown in Figure 4.3. In the heat maps, the absolute values of the entries are presented, the diagonals of the estimated partial correlation matrices are set as zero for a better visual effect, and the 10 clusters chosen by visualization are framed with black rectangles.

From Figure 4.3, we see that the ADHD-C group generally has weaker marginal connectivity than the NC group, which can be clearly seen in the largest block on the lower-right corner of the heat maps. However, both groups have very weak and similar direct connectivities. The corresponding cluster of the largest block contains 82 ROIs. The node strength (Barrat et al., 2004) of a brain node (i.e., a ROI here) is defined for marginal connectivity by the sum of its absolute correlations with the other nodes of interest, and similarly defined for direct connectivity by using absolute partial correlations instead. We compare ADHD-C to NC by using ROIs' node strength within the largest cluster. Figures 4.4 shows the difference of estimated node strength of ADHD-C and NC in this cluster. We see that most ROIs in the cluster have reduced node strength of marginal connectivity for ADHD-C. The two most severe losses can be seen in areas at the coordinates $Z = -44$ and $Z = 64$ in Figure 4.4 (a). These two areas are respectively located in the right middle temporal cortex and the left superior parietal cortex, which have been reported with abnormalities for ADHD patients in the literature (Kim et al., 2002; Fan et al., 2014b). For the direct connectivity, the estimated differences in node strength between ADHD-C and NC are very small in this cluster.



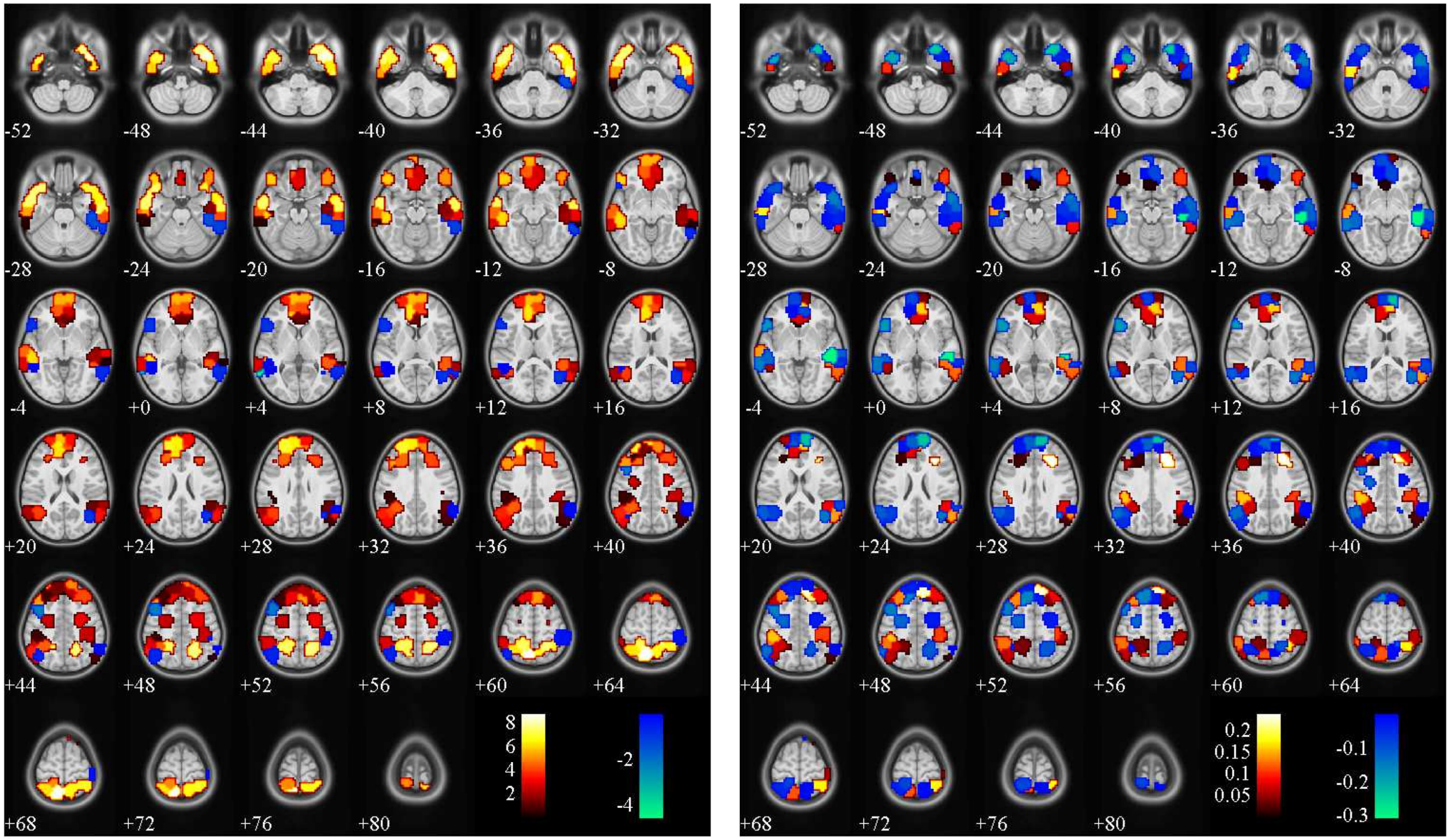
(a) NC

(b) ADHD-C

(c) NC

(d) ADHD-C

Figure 4.3: (a,b) Heat maps of the absolute values of estimated correlation matrices for NC and ADHD-C. (c,d) Heat maps of the absolute values of estimated partial correlation matrices for NC and ADHD-C.



(a) Marginal connectivity

(b) Direct connectivity

Figure 4.4: Estimated reduced node strength in the largest cluster for ADHD-C compared to NC. The brain is plotted in the Montreal Neurological Institute 152 space with Z-coordinates displayed.

4.6 General Results for Estimation Using Weighted Sample Covariance Matrix

In previous sections, we use the sample covariance matrix $\hat{\mathbf{\Sigma}}$ as the initial estimator of $\mathbf{\Sigma}$ in the considered estimating procedures. In fact, we can have consistency results of the final considered estimators by using any given initial estimator, denoted as $\check{\mathbf{\Sigma}}$, if the following concentration inequality

$$P\big[|\check{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty \le Mu\big] = 1 - O(p^{-M'}) \tag{4.10}$$

holds with $u = o(1)$ and some positive constants M, M'. Smaller u yields faster convergence rates. This motivates us to construct an initial estimator of $\mathbf{\Sigma}$ with the max-norm error as small as possible.

An equivalent expression of $\hat{\mathbf{\Sigma}}$ in (4.4) is

$$\hat{\mathbf{\Sigma}} = \sum_{\ell=1}^{L} \frac{n_\ell}{n} \hat{\mathbf{\Sigma}}_0^{(\ell)} - \Big(\sum_{\ell=1}^{L} \frac{n_\ell}{n} \hat{\boldsymbol{\mu}}^{(\ell)}\Big)^{\otimes 2}, \tag{4.11}$$

where $\hat{\mathbf{\Sigma}}_0^{(\ell)} = n_\ell^{-1} \sum_{i=1}^{n_\ell} (\boldsymbol{X}_i^{(\ell)})^{\otimes 2}$ and $\hat{\boldsymbol{\mu}}^{(\ell)} = n_\ell^{-1} \sum_{i=1}^{n_\ell} \boldsymbol{X}_i^{(\ell)}$. In the above expression of $\hat{\mathbf{\Sigma}}$, n_ℓ/n can be viewed as the weight of the ℓ-th sample. For i.i.d. observations, n_ℓ is the effective sample size of the ℓ-th sample. Intuitively, replacing n_ℓ and n by their corresponding effective sample sizes (defined in a certain reasonable way) for the temporally dependent observations may yield a smaller u. Thus, we consider the following weighted sample covariance matrix:

$$\tilde{\mathbf{\Sigma}} := (\tilde{\sigma}_{ab})_{p\times p} = \tilde{\mathbf{\Sigma}}_0 - \tilde{\boldsymbol{\mu}}^{\otimes 2} := \sum_{\ell=1}^{L} \varpi_\ell \hat{\mathbf{\Sigma}}_0^{(\ell)} - \Big(\sum_{\ell=1}^{L} \varpi_\ell \hat{\boldsymbol{\mu}}^{(\ell)}\Big)^{\otimes 2}, \tag{4.12}$$

where the weight $\varpi_\ell := n_\ell f_\ell^{-1} / \sum_i n_i f_i^{-1}$ with any given $f_\ell > 0$ for each $\ell = 1, \ldots, L$. Note that $\{f_\ell\}$ and $\{cf_\ell\}$, with an arbitrary constant $c > 0$, give the same weights $\{\varpi_\ell\}$. The corresponding weighted sample correlation matrix is defined by $\tilde{\mathbf{R}} = (\tilde{\rho}_{ij})_{p\times p} = \big(\tilde{\sigma}_{ij}/\sqrt{\tilde{\sigma}_{ii}\tilde{\sigma}_{jj}}\big)_{p\times p}$. By the following equivalent form of (4.12)

$$\tilde{\mathbf{\Sigma}} = \sum_{\ell=1}^{L} \frac{\varpi_\ell}{n_\ell} \sum_{i=1}^{n_\ell} (\boldsymbol{X}_i^{(\ell)} - \tilde{\boldsymbol{\mu}})^{\otimes 2}, \tag{4.13}$$

it is easily shown that $\tilde{\mathbf{\Sigma}}$ is positive-semidefinite and $|\tilde{\mathbf{R}}|_\infty \leq 1$. Define

$$\tau_2 = \frac{\max\{\sqrt{(\sum_\ell n_\ell g_\ell / f_\ell^2) \log p}, \max_\ell (g_\ell / f_\ell) \log p\}}{\sum_\ell n_\ell / f_\ell}, \tag{4.14}$$

which will become clear in the proof of the following Theorem IV.4. For any given $\{\varpi_\ell\}$, a concentration inequality in the form of (4.10) is given in Theorem IV.4 and the corresponding asymptotic properties of the considered estimators started with $\tilde{\mathbf{\Sigma}}$ are given in Theorem IV.5 that generalizes all the theorems given in Subsection 4.3.1.

Theorem IV.4. *Assume data* $\{\mathbf{X}^{(\ell)}_{p\times n_\ell}\}_{1\leq \ell \leq L}$ *are generated from* (4.1) *with* $|\mathbf{\Sigma}|_\infty \leq v_0$ *for a constant* $v_0 > 0$. *For any constant* $M' > 0$, *there exists a constant* $M > 0$ *such that when* $\tau_2 = o(1)$, *we have*

$$P\left[|\tilde{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty \leq M\tau_2\right] = 1 - O(p^{-M'}). \tag{4.15}$$

Theorem IV.5. *All the statements given in Theorems IV.1–IV.3 hold more generally when* $\tau_1, \hat{\mathbf{\Sigma}}$ *and* $\hat{\mathbf{R}}$ *are replaced by* $\tau_2, \tilde{\mathbf{\Sigma}}$ *and* $\tilde{\mathbf{R}}$ *respectively.*

We can see that τ_2 is a main factor determining the convergence rates. Define two random variables f and g with sample spaces $\{f_\ell\}_{\ell=1}^L$ and $\{g_\ell\}_{\ell=1}^L$ respectively and $P(g/f = g_\ell/f_\ell) = n_\ell g_\ell^{-1} / \sum_{i=1}^L n_i g_i^{-1}$. Then τ_2 can be further written as

$$\tau_2 = \max\left\{\sqrt{\frac{\log p}{\sum_\ell n_\ell g_\ell^{-1}}\left(\frac{\text{var}(g/f)}{[E(g/f)]^2} + 1\right)}, \frac{(\log p)\max(g/f)}{(\sum_\ell n_\ell g_\ell^{-1})E(g/f)}\right\}.$$

Hence, if and only if $f_\ell = cg_\ell$ with an arbitrary constant $c > 0$ for all ℓ, τ_2 attains its minimum. The minimum is

$$\tau_2^* := \sqrt{\frac{\log p}{\sum_\ell n_\ell / g_\ell}} \tag{4.16}$$

when $\tau_2^* = o(1)$. Hence, using the sample covariance matrix $\hat{\mathbf{\Sigma}}$ as the initial estimator yields the optimal τ_2 only when all g_ℓ are equal, e.g., when all the n observations are i.i.d..

In fact, if $g_{\max}/\min_\ell g_\ell = C$ with a constant C, then $\tau_1 \asymp \tau_2^*$ when $\tau_2^* = o(1)$. Moreover, when $f_\ell = 1$ for all ℓ, we have $\tau_2 = \tau_1$, hence Theorem IV.5 reduces to all the theorems with $\hat{\mathbf{\Sigma}}$ as the initial estimator given in Subsection 4.3.1.

4.7 Discussion

In Section 4.6 we see that using the weighted sample covariance matrix in the estimation yields faster convergence rates when $f_\ell = g_\ell$ for all ℓ. However, $\{g_\ell\}$ are unknown in practice and are difficult to be estimated. Even for stationary time series, it is well-known that they cannot be consistently estimated using the sample autocorrelation matrix (Wu and Pourahmadi, 2009, Theorem 1). Developing a procedure for choosing appropriate weights in (4.12) is of great interest.

CHAPTER V

Conclusion and Future Work

Classical statistical methods often fail to handle high dimensional data, for which the variable dimension p is comparable to or larger than the sample size n. Although significant development has been made in high dimensional data analysis over the past two decades, most high dimensional methods are assumed on certain independent structures of the data. There is a great need for statistical methods that are suitable for analyzing large-scale neuroimaging data with spatial and/or temporal dependence.

Motivated by this need, this dissertation focused on two major high dimensional problems for dependent data. We considered (i) the multiple testing problem for spatially correlated data in Chapter II, and (ii) the estimation of large covariance and precision matrices from a single sample of temporally dependent observations in Chapter III and from multiple independent samples in Chapter IV.

In Chapter II, we considered LIS-based FDR procedures based on HMRF for 3D neuroimaging data, where HMRF provides a natural way of modeling spatial correlations. The proposed procedures aim to minimize the FNR while FDR is controlled at a pre-specified level. We found that brain regions are spatially heterogeneous, and hence we modeled each region separately by a single HMRF, and implemented the PLIS procedure to minimize the global FNR. We proposed a GEM algorithm based on the penalized likelihood to obtain the HMRF parameter estimates, which overcomes the unboundedness of the original likelihood function. Numerical analysis showed the superiority of the HMRF-LIS-based procedures over commonly used FDR procedures, illustrating the value of HMRF-LIS-based FDR procedures for spatially correlated image data.

We also proved the validity and optimality of the oracle HMRF-LIS-based procedures, for which the parameters are known. However, when the parameters are unknown, the asymptotic equivalence of the data-driven procedures to the oracle procedures remains an open problem, although they performed similarly in our extensive simulations. Moreover, one can extend the Ising model to more complicated MRFs, but how to examine the model fitness of the selected MRF is unknown. These two points are directions for future research.

In Chapters III and IV, properties of consistency, sparsistency and sign-consistency were established for the generalized thresholding estimation of covariance/correlation matrices and for the CLIME and SPICE estimators of precision matrix using a single sample and multiple independent samples of temporally dependent observations, respectively. A different proof technique to that in Chapter III was used in Chapter IV. They each have their own advantages in terms of the convergence rates.

The results obtained in these two chapters for a single sample apply to the temporal dependence with longer memory than those in Chen et al. (2013) and Bhattacharjee and Bose (2014). As expected, the convergence rates of considered estimators decrease as the temporal dependence increases. Under similar conditions in Cai and Zhou (2012), it can easily be shown that the rates of convergence in mean square are minimax optimal for the covariance/correlation matrix estimation under temporal dependence with $g_{\max} < C$ for some constant $C > 0$. One may consider the minimax optimal rates for the other cases, especially for strong temporal dependence. A gap-block cross-validation method was proposed for the tuning parameter selection, which performed well in simulations by using parameters $H_1 = H_2 = 10$. The theoretical justification of this intuitive cross-validation and its optimal choices of H_1 and H_2 are of future interest.

For multiple independent samples, the results of Chapter IV can give faster convergence rates than those in Chapter III. Compared to using the sample covariance matrix, using the weighted sample covariance matrix in the considered estimating methods can theoretically improve the rates if appropriate weights are given. It is of great interest to develop a procedure for selecting such weights in practice.

A potential way to improve the current convergence rates is incorporating the estimation or modeling of temporal dependence into the estimating procedures of large covariance and precision matrices. This can be an interesting topic.

In conclusion, we proposed an efficient FDR controlling procedure for certain spatially correlated data, and we also showed that several commonly used methods of estimating covariance and precision matrices for independent observations can be applied to a wide family of temporally dependent data. This dissertation makes an innovative contribution to the analysis of high dimensional dependent data, in particular, neuroimaging data.

APPENDICES

APPENDIX A

Supplementary Materials for Chapter II

A.1 Theoretical Results of the Oracle LIS-Based Procedures for HMRF

In this section, we show the theoretical results of the oracle LIS-based procedures originally for HMC model in Sun and Cai (2009) (Theorems 1 to 4 and Corollary 1) and Wei et al. (2009) (Theorems 1 and 2), including the validity and optimality of the procedures, also hold for our HMRF model. Here, an FDR procedure is called *valid* if it controls FDR at a prespecified level α, and is called *optimal* if it minimizes marginal FNR (mFNR) while controlling marginal FDR (mFDR) at the level α. Note that the asymptotic equivalence between FDR and mFDR as well as that between FNR and mFNR hold under certain conditions (Genovese and Wasserman, 2002; Xie et al., 2011), but remain open questions for both HMC and HMRF.

Unless stated otherwise, the notation in this section is the same as in Sun and Cai (2009) to which readers are referred. Define $\pi_{ij} = P(\Theta_i = j), i \in S, j = 0, 1$. The model homogeneity, i.e., $\pi_{ij} = \pi_j^{(k)}$ for all i in k-th HMC, is required in Sun and Cai (2009) and in Wei et al. (2009) but fails to hold for HMRF because the boundary voxels and interior voxels have different numbers of neighbors. However, the theory of the oracle procedures still holds for HMRF if we redefine the average conditional cumulative distribution functions

(CDFs) of the test statistic $\boldsymbol{T}(\boldsymbol{x}) = \{T_i(\boldsymbol{x}) : i \in S\}$ by

$$G^j(t) = \frac{\sum_{i \in S} \pi_{ij} G_i^j(t)}{\sum_{i \in S} \pi_{ij}}, \tag{A.1}$$

where $G_i^j(t) = P(T_i < t | \Theta_i = j)$.

For HMC model, Sun and Cai (2009) proved the optimality of oracle LIS procedure in their Theorems 1 to 3 and Corollary 1, and its validity in their Theorem 4; Wei et al. (2009) showed the validity of oracle SLIS procedure in their Theorem 1, and both validity and optimality of oracle PLIS procedure in their Theorem 2. We modify the statements in these theorems and corollary for HMRF by

(i) replacing HMM by HMRF;

(ii) in Corollary 1 of Sun and Cai (2009), replacing the definition of $G^j(t)$ by (A.1) and the equation $g^1(t)/g^0(t) = (1/t)\pi_0/\pi_1$ by $g^1(t)/g^0(t) = (1/t)\sum_{i \in S} \pi_{i0}/\sum_{i \in S} \pi_{i1}$;

(iii) in Theorem 2 of Wei et al. (2009), more precisely stating the optimality of oracle PLIS procedure based on mFDR and mFNR.

For simplicity, we omit all the modified statements and only provide their proofs in the following.

A.1.1 Modified Theorem 1 of Sun and Cai (2009) for HMRF

Proof. Following the proof of Proposition 1 in Sun and Cai (2007), we have

$$g^0(c)G^1(c) - G^0(c)g^1(c) > 0 \tag{A.2}$$

and

$$g^0(c)[1 - G^1(c)] - g^1(c)[1 - G^0(c)] < 0. \tag{A.3}$$

Additionally, by (A.1),

$$\begin{aligned}
\text{mFDR}(c) &= \frac{E(N_{10})}{E(R)} = \frac{\sum_{i\in S} P(T_i < c, \Theta_i = 0)}{\sum_{i\in S} P(T_i < c)} \\
&= \frac{\sum_{i\in S} \pi_{i0} G_i^0(c)}{\sum_{i\in S}(\pi_{i0} G_i^0(c) + \pi_{i1} G_i^1(c))} \\
&= \frac{G^0(c)\sum_{i\in S}\pi_{i0}}{G^0(c)\sum_{i\in S}\pi_{i0} + G^1(c)\sum_{i\in S}\pi_{i1}},
\end{aligned}$$

and

$$\begin{aligned}
\text{mFNR}(c) &= \frac{E(N_{01})}{E(S)} = \frac{\sum_{i\in S} P(T_i \geq c, \Theta_i = 1)}{\sum_{i\in S} P(T_i \geq c)} \\
&= \frac{\sum_{i\in S} \pi_{i1}[1 - G_i^1(c)]}{\sum_{i\in S}(\pi_{i0}[1 - G_i^0(c)] + \pi_{i1}[1 - G_i^1(c)])} \\
&= \frac{[1 - G^1(c)]\sum_{i\in S}\pi_{i1}}{[1 - G^0(c)]\sum_{i\in S}\pi_{i0} + [1 - G^1(c)]\sum_{i\in S}\pi_{i1}}.
\end{aligned}$$

Then,

$$\begin{aligned}
\frac{d(\text{mFDR}(c))}{dc} &= \Bigg(g^0(c)\sum_{i\in S}\pi_{i0}\left[G^0(c)\sum_{i\in S}\pi_{i0} + G^1(c)\sum_{i\in S}\pi_{i1}\right] \\
&\qquad - G^0(c)\sum_{i\in S}\pi_{i0}\left[g^0(c)\sum_{i\in S}\pi_{i0} + g^1(c)\sum_{i\in S}\pi_{i1}\right]\Bigg) \\
&\qquad \Bigg/ \left[G^0(c)\sum_{i\in S}\pi_{i0} + G^1(c)\sum_{i\in S}\pi_{i1}\right]^2 \\
&= \frac{[g^0(c)G^1(c) - G^0(c)g^1(c)](\sum_{i\in S}\pi_{i0})(\sum_{i\in S}\pi_{i1})}{[G^0(c)\sum_{i\in S}\pi_{i0} + G^1(c)\sum_{i\in S}\pi_{i1}]^2} \\
&> 0
\end{aligned}$$

following from (A.2), and

$$
\begin{aligned}
\frac{d(\text{mFNR}(c))}{dc} &= \Bigg\{ -g^1(c)\sum_{i\in S}\pi_{i1}\left([1-G^0(c)]\sum_{i\in S}\pi_{i0}+[1-G^1(c)]\sum_{i\in S}\pi_{i1}\right) \\
&\quad -\left([1-G^1(c)]\sum_{i\in S}\pi_{i1}\right)\left(-g^0(c)\sum_{i\in S}\pi_{i0}-g^1(c)\sum_{i\in S}\pi_{i1}\right)\Bigg\} \\
&\quad \Bigg/ \left([1-G^0(c)]\sum_{i\in S}\pi_{i0}+[1-G^1(c)]\sum_{i\in S}\pi_{i1}\right)^2 \\
&= \frac{(g^0(c)[1-G^1(c)]-g^1(c)[1-G^0(c)])(\sum_{i\in S}\pi_{i0})(\sum_{i\in S}\pi_{i1})}{([1-G^0(c)]\sum_{i\in S}\pi_{i0}+[1-G^1(c)]\sum_{i\in S}\pi_{i1})^2} \\
&< 0
\end{aligned}
$$

following from (A.3). Hence we obtain part (a) and (b) of the theorem.

For part (c), the classification risk with the loss function

$$L_\lambda(\boldsymbol{\Theta},\boldsymbol{\delta}) = \frac{1}{N}\sum_{i\in S}\{\lambda(1-\Theta_i)\delta_i+\Theta_i(1-\delta_i)\}$$

is

$$
\begin{aligned}
E[L_\lambda(\boldsymbol{\Theta},\boldsymbol{\delta})] &= \frac{1}{N}\sum_{i\in S}\{\lambda P(\Theta_i=0,T_i<c)+P(\Theta_i=1,T_i\geq c)\} \\
&= \frac{1}{N}\sum_{i\in S}\{\lambda\pi_{i0}G_i^0(c)+\pi_{i1}[1-G_i^1(c)]\} \\
&= \frac{1}{N}\left\{\lambda G^0(c)\sum_{i\in S}\pi_{i0}+[1-G^1(c)]\sum_{i\in S}\pi_{i1}\right\}.
\end{aligned}
$$

The optimal cutoff c^* that minimizes this risk satisfies

$$\lambda = \frac{g^1(c^*)\sum_{i\in S}\pi_{i1}}{g^0(c^*)\sum_{i\in S}\pi_{i0}}.$$

Since $\boldsymbol{T}\in\mathcal{T}$, we have $g^1(c^*)/g^0(c^*)$ is monotonically decreasing in c^*. Thus, $\lambda(c^*)$ is monotonically decreasing in c^*. □

A.1.2 Modified Theorem 2 of Sun and Cai (2009) for HMRF

Proof. Suppose there are v_L hypotheses from the null and k_L hypotheses from the nonnull among the r rejected hypotheses when the decision rule $\boldsymbol{\delta}(\boldsymbol{L},c_L)$ is applied with test statistic $\boldsymbol{L}$ and cutoff c_L. We have $v_L = \sum_{i\in S} P(\Theta_i = 0, L_i < c_L)$ and $k_L = \sum_{i\in S} P(\Theta_i = 1, L_i < c_L)$, and the classification risk

$$\begin{aligned}
R_{\lambda(\alpha)} &= E[L_{\lambda(\alpha)}(\boldsymbol{\Theta}, \boldsymbol{\delta}(\boldsymbol{L}, c_L))] \\
&= \frac{1}{N}\sum_{i\in S}\{\lambda(\alpha)P(\Theta_i = 0, L_i < c_L) + P(\Theta_i = 1, L_i \geq c_L)\} \\
&= \frac{1}{N}\left\{\sum_{i\in S}\pi_{i1} + \lambda(\alpha)v_L - k_L\right\}. \qquad \text{(A.4)}
\end{aligned}$$

Then following the proof of Theorem 1 in Sun and Cai (2007) using the expression (A.4) for the classification risk $R_{\lambda(\alpha)}$, we complete the proof. □

A.1.3 Modified Theorems 3 and 4 of Sun and Cai (2009) for HMRF

Proof. The proofs are the same as those of Theorems 3 and 4 in Sun and Cai (2009), thus omitted. □

A.1.4 Modified Corollary 1 of Sun and Cai (2009) for HMRF

Proof. Following the proof of Corollary 1 in Sun and Cai (2009) with the expression of the risk R replaced by

$$\begin{aligned}
R &= \frac{1}{N}\sum_{i\in S}\left\{\frac{1}{t}\pi_{i0}G_i^0(t^*) + \pi_{i1}[1 - G_i^1(t^*)]\right\} \\
&= \frac{1}{N}\left\{\frac{1}{t}G^0(t^*)\sum_{i\in S}\pi_{i0} + [1 - G^1(t^*)]\sum_{i\in S}\pi_{i1}\right\}
\end{aligned}$$

and their equation $g^1(t^*)/g^0(t^*) = (1/t)\pi_0/\pi_1$ substituted by the new equation $g^1(t^*)/g^0(t^*) = (1/t)\sum_{i\in S}\pi_{i0}/\sum_{i\in S}\pi_{i1}$, we complete the proof. □

A.1.5 Modified Theorems 1 and 2 of Wei et al. (2009) for HMRF

Proof. For Theorem 1 and the validity of oracle PLIS procedure in Theorem 2, the proofs are the same as those in Wei et al. (2009). For the optimality of oracle PLIS procedure in Theorem 2, the proof is the same as the proof of the optimality of oracle LIS procedure given above. □

A.2 Gibbs Sampler Approximations

This section presents the approximations of quantities of interest in GEM. Let Ω be the set of all possible configurations of $\boldsymbol{\Theta}$: $\Omega = \{\boldsymbol{\theta} = (\theta_s)_{s \in S} : \theta_s \in \{0,1\}, s \in S\}$. By the ergodic theorem of the Gibbs sampler (see Lemma 1 and Theorem 1 in Roberts and Smith (1994)), for any Gibbs distribution (see definition (4.3) in Geman and Geman (1984)) $\pi(\boldsymbol{\theta})$ and any real-valued function $f(\boldsymbol{\theta})$ on Ω, with probability one,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\theta}^{(i)}) = \int_{\Omega} f(\boldsymbol{\theta}) d\pi(\boldsymbol{\theta}) = E[f(\boldsymbol{\Theta})],$$

where $\boldsymbol{\theta}^{(i)}, i = 1, ..., n$ are samples successively generated using the Gibbs sampler by $\pi(\boldsymbol{\theta})$. For our HMRF, it is easy to see that both the Ising model probability distribution $P_{\boldsymbol{\varphi}}(\boldsymbol{\theta})$ and the conditional probability distribution $P_{\boldsymbol{\Phi}^{(t)}}(\boldsymbol{\theta}|\boldsymbol{x})$ are Gibbs distributions. Thus by the ergodic theorem, the following quantities can be approximated using Monte Carlo averages via Gibbs sampler:

$$\begin{aligned}
\boldsymbol{U}^{(t+1)}(\boldsymbol{\varphi}) &= E_{\boldsymbol{\Phi}^{(t)}}[\boldsymbol{H}(\boldsymbol{\Theta})|\boldsymbol{x}] - E_{\boldsymbol{\varphi}}[\boldsymbol{H}(\boldsymbol{\Theta})] \\
&\approx \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{H}(\boldsymbol{\theta}^{(t,i,\boldsymbol{x})}) - \boldsymbol{H}(\boldsymbol{\theta}^{(i,\boldsymbol{\varphi})})\right), \\
\boldsymbol{I}(\boldsymbol{\varphi}) &= Var_{\boldsymbol{\varphi}}[\boldsymbol{H}(\boldsymbol{\Theta})] \\
&= E_{\boldsymbol{\varphi}}\left[(\boldsymbol{H}(\boldsymbol{\Theta}) - E_{\boldsymbol{\varphi}}[\boldsymbol{H}(\boldsymbol{\Theta})])^{\otimes 2}\right] \\
&\approx \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{H}(\boldsymbol{\theta}^{(i,\boldsymbol{\varphi})}) - \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{H}(\boldsymbol{\theta}^{(j,\boldsymbol{\varphi})})\right)^{\otimes 2}, \\
\gamma_s^{(t)}(i) &= P_{\boldsymbol{\Phi}^{(t)}}(\Theta_s = i|\boldsymbol{x}) = E_{\boldsymbol{\Phi}^{(t)}}[\mathbf{1}(\Theta_s = i)|\boldsymbol{x}] \\
&= E_{\boldsymbol{\Phi}^{(t)}}[\mathbf{1}(\Theta_s = i)\mathbf{1}(\boldsymbol{\Theta} \in \Omega)|\boldsymbol{x}] \\
&\approx \frac{1}{n}\sum_{k=1}^{n}\mathbf{1}(\theta_s^{(t,k,\boldsymbol{x})} = i), \\
\frac{C}{Z(\boldsymbol{\varphi})} &= E_{\boldsymbol{\varphi}}[\exp\{-\boldsymbol{\varphi}^T\boldsymbol{H}(\boldsymbol{\Theta})\}] \\
&\approx \frac{1}{n}\sum_{i=1}^{n}\exp\{-\boldsymbol{\varphi}^T\boldsymbol{H}(\boldsymbol{\theta}^{(i,\boldsymbol{\varphi})})\},
\end{aligned}$$

and

$$
\begin{aligned}
& Q_2(\boldsymbol{\varphi}^{(t+1,m)}|\boldsymbol{\Phi}^{(t)}) - Q_2(\boldsymbol{\varphi}^{(t)}|\boldsymbol{\Phi}^{(t)}) \\
= \;& E_{\boldsymbol{\Phi}^{(t)}}[\log P_{\varphi^{(t+1,m)}}(\boldsymbol{\Theta}) - \log P_{\varphi^{(t)}}(\boldsymbol{\Theta})|\boldsymbol{x}] \\
= \;& E_{\boldsymbol{\Phi}^{(t)}}[(\boldsymbol{\varphi}^{(t+1,m)} - \boldsymbol{\varphi}^{(t)})^T \boldsymbol{H}(\boldsymbol{\Theta})|\boldsymbol{x}] + \log\left(\frac{Z(\varphi^{(t)})}{Z(\varphi^{(t+1,m)})}\right) \\
\approx \;& \frac{1}{n}(\boldsymbol{\varphi}^{(t+1,m)} - \boldsymbol{\varphi}^{(t)})^T \sum_{i=1}^{n} \boldsymbol{H}(\boldsymbol{\theta}^{(t,i,\boldsymbol{x})}) \\
& \quad + \log\left(\frac{\sum_{i=1}^{n} \exp\{-{\boldsymbol{\varphi}^{(t+1,m)}}^T \boldsymbol{H}(\boldsymbol{\theta}^{(i,\boldsymbol{\varphi}^{(t+1,m)})})\}}{\sum_{i=1}^{n} \exp\{-{\boldsymbol{\varphi}^{(t)}}^T \boldsymbol{H}(\boldsymbol{\theta}^{(i,\boldsymbol{\varphi}^{(t)})})\}}\right),
\end{aligned}
$$

where $\{\boldsymbol{\theta}^{(1,\boldsymbol{\varphi})}, ..., \boldsymbol{\theta}^{(n,\boldsymbol{\varphi})}\}$ and $\{\boldsymbol{\theta}^{(t,1,\boldsymbol{x})}, ..., \boldsymbol{\theta}^{(t,n,\boldsymbol{x})}\}$ are large n samples successively generated using the Gibbs sampler by $P_{\boldsymbol{\varphi}}(\boldsymbol{\theta})$ and $P_{\boldsymbol{\Phi}^{(t)}}(\boldsymbol{\theta}|\boldsymbol{x})$ respectively, and C is the cardinality of set Ω.

A.3 ADNI FDG-PET Imaging Data Analysis

We apply the PLIS procedure with HMRFs to the analysis of ADNI FDG-PET imaging data, which is compared with BH, q-value and CLfdr procedures. Since the FDG-PET scans were normalized to the average of pons and cerebellar vermis, areas of the brain known to be least affected in AD, it was not surprising that almost all the signal voxels are found with decreased CMRgl. Both PLIS and CLfdr procedures discovered significant metabolic reduction, with a regional proportion of signals $> 50\%$, in brain regions preferentially affected by AD, including the posterior cingulate (BAs 23, 31; Mosconi et al., 2008; Langbaum et al., 2009), parietal cortex (BAs 7, 37, 39, 40; Minoshima et al., 1995; Matsuda, 2001), temporal cortex (BAs 20 to 22; Alexander et al., 2002; Landau et al., 2011), medial temporal cortex (BAs 28, 34; Karow et al., 2010), frontal cortex (BAs 8 to 11, and 44 to 47; Mosconi, 2005), insular cortex (Perneczky et al., 2007), amygdala (Nestor et al., 2003) and hippocampus (Mosconi et al., 2005). In regions also typically affected in AD, such as anterior cingulate (BAs 24, 32; Fouquet et al., 2009) and occipital cortex (BAs 17 to 19; Langbaum et al., 2009), the proportions of signals found by PLIS are 49.6% and 39.0%, respectively, compared with 35.4% and 11.6% found by CLfdr, 12.2% and 0.94% by q-value, as well as only 1.24% and 0.87% by BH.

With respect to the regions that are relatively spared from AD (Benson et al., 1983; Matsuda, 2001; Ishii, 2002) or rarely reported in the literature of the disease, caudate, thalamus and putamen are found with high proportions of signals by PLIS ($>$ 45%) and CLfdr ($>$ 25%) in each of these regions; signals in medulla, midbrain, cerebellar hemispheres, pre-motor cortex (BA 6) and primary somatosensory cortex (BAs 1, 2, 3, 5) are each claimed with a proportion greater than 20% by PLIS, but very sparse found by the other three procedures. Since MCI as a group consists of a mix of patients, many of them will progress to AD but some will not which may include subjects with corticobasal degeneration (Ishii, 2002), frontotemporal dementia (Jeong et al., 2005), or Parkinsonism (Huang et al., 2007; Zeman et al., 2011; Ishii, 2014), it is not surprising that some areas not typical of AD patients were found to be abnormal in the MCI group.

APPENDIX B

Supplementary Materials for Chapter III

This Supplementary Material contains the detailed proofs of the general theorems given in Subsection 3.6.1, the instructions for selecting the candidates values of tuning parameters, and additional results of the rfMRI data analysis.

B.1 Technical Lemmas

The following lemma is an extension of the "Hanson-Wright inequality" (Rudelson and Vershynin, 2013, Theorem 1.1) and "Hoeffding-type inequality" (Vershynin, 2012, Proposition 5.10) for independent sub-Gaussian data to that for a certain type of uncorrelated sub-Gaussian data.

Lemma B.1. *Let $\boldsymbol{e} = (e_1, e_2, \dots)^T$ be an infinite-dimensional random vector with independent standard sub-Gaussian components, each with the same parameter $K \geq 1$ defined in* (3.2)*. Let $\boldsymbol{Y} = \mathbf{A}\boldsymbol{e}$ be a well-defined random vector with length d in the sense of entrywise almost-sure convergence and mean-square convergence, and $\mathbf{A}\mathbf{A}^T = \mathbf{I}_{d\times d}$. Then for $t \geq 0$, there exists a constant $c > 0$ only dependent on K such that*

$$P\left[\left|\boldsymbol{Y}^T\mathbf{B}\boldsymbol{Y} - E[\boldsymbol{Y}^T\mathbf{B}\boldsymbol{Y}]\right| \geq t\right] \leq 2\exp\left\{-c\min\left(\frac{t^2}{\|\mathbf{B}\|_F^2}, \frac{t}{\|\mathbf{B}\|_2}\right)\right\} \tag{B.1}$$

and

$$P\left[\left|\boldsymbol{b}^T\boldsymbol{Y}\right| \geq t\right] \leq \exp(1)\cdot\exp\left\{-\frac{ct^2}{\|\boldsymbol{b}\|_F^2}\right\}, \tag{B.2}$$

where nonzero matrix $\mathbf{B} = (b_{ij})_{d\times d}$*, and* $\boldsymbol{b}$ *is a* d*-dimensional nonzero vector.*

Proof. Consider the nontrivial case when $t > 0$. Let $\mathbf{A} = (a_{ij})_{d\times\infty}$, $\mathbf{A}_m = (a_{ij})_{d\times m}$ consist of the first m columns of $\mathbf{A}$, $\boldsymbol{e}_m = (e_1, e_2, ..., e_m)^T$ consist of the first m elements of $\boldsymbol{e}$, and $\boldsymbol{Y}_m = (Y_1^m, ..., Y_d^m)^T = \mathbf{A}_m\boldsymbol{e}_m$. For each i, when $m \to \infty$, we have $Y_i^m = \sum_{j=1}^m a_{ij}e_j \overset{P}{\to} Y_i = \sum_{j=1}^\infty a_{ij}e_j$, with $\sum_{j=1}^m a_{ij}^2 \to 1$ and $\sum_{j=1}^m a_{ij}a_{kj} \to 0$ for $i \neq k$ following from $\mathbf{A}\mathbf{A}^T = \mathbf{I}_{d\times d}$. Thus, for dimension d and positive values $\varepsilon_1, \varepsilon_2$ and δ, there exists a number N such that for any $m > N$, we have

$$P\left[|\boldsymbol{Y}^T\mathbf{B}\boldsymbol{Y} - \boldsymbol{Y}_m^T\mathbf{B}\boldsymbol{Y}_m| \geq \varepsilon_1\right] \leq \delta, \tag{B.3}$$

$$P\left[|\boldsymbol{b}^T\boldsymbol{Y} - \boldsymbol{b}^T\boldsymbol{Y}_m| \geq \varepsilon_1\right] \leq \delta, \tag{B.4}$$

$$\left|\sum_{j=1}^m a_{ij}^2 - 1\right| \leq \frac{\varepsilon_2}{d}, \quad \text{and} \left|\sum_{j=1}^m a_{ij}a_{kj}\right| \leq \frac{\varepsilon_2}{d^2} \quad \text{for } i \neq k. \tag{B.5}$$

Since $E(\boldsymbol{Y}_m) = \mathbf{A}_m E(\boldsymbol{e}_m) = \mathbf{0}$ and cov$(\boldsymbol{Y}_m) = \mathbf{A}_m\text{cov}(\boldsymbol{e}_m)\mathbf{A}_m^T = \mathbf{A}_m\mathbf{A}_m^T$, we have $E[\boldsymbol{Y}_m^T\mathbf{B}\boldsymbol{Y}_m] = \sum_{1\leq i,k\leq d} b_{ik}E[Y_i^mY_k^m] = \sum_{1\leq i,k\leq d} b_{ik}\sum_{j=1}^m a_{ij}a_{kj}$. So $E[\boldsymbol{Y}^T\mathbf{B}\boldsymbol{Y}] = \sum_{i=1}^d b_{ii}$. By Lemma 5.5 in Vershynin (2012), there exists a constant c_1 only dependent on K such that

$$\sup_{k\geq 1} k^{-1/2}(E|e_j|^k)^{1/k} \leq c_1 \text{ for all } j = 1, 2, \ldots.$$

Without loss of generality, we assume $c_1 > 1$. Then by Theorem 1.1 in Rudelson and Vershynin (2013) and Proposition 5.10 in Vershynin (2012), for every $t > 0$, there exists an absolute constant $c_2 > 0$ such that

$$
\begin{aligned}
& P\left[\left|\boldsymbol{Y}_m^T\mathbf{B}\boldsymbol{Y}_m - \sum_{1\le i,k\le d} b_{ik}\sum_{j=1}^m a_{ij}a_{kj}\right| \ge t/2\right] \\
& = P\left[\left|\boldsymbol{Y}_m^T\mathbf{B}\boldsymbol{Y}_m - E[\boldsymbol{Y}_m^T\mathbf{B}\boldsymbol{Y}_m]\right| \ge t/2\right] \\
& \le P\left[\left|\boldsymbol{e}_m^T\mathbf{A}_m^T\mathbf{B}\mathbf{A}_m\boldsymbol{e}_m - E[\boldsymbol{e}_m^T\mathbf{A}_m^T\mathbf{B}\mathbf{A}_m\boldsymbol{e}_m]\right| > t/3\right] \\
& \le 2\exp\left\{-c_2\min\left(\frac{t^2}{9c_1^4\|\mathbf{A}_m^T\mathbf{B}\mathbf{A}_m\|_F^2}, \frac{t}{3c_1^2\|\mathbf{A}_m^T\mathbf{B}\mathbf{A}_m\|_2}\right)\right\} \\
\text{(B.6)} \qquad & \le 2\exp\left\{-\frac{c_2}{9c_1^4}\min\left(\frac{t^2}{\|\mathbf{A}_m^T\mathbf{B}\mathbf{A}_m\|_F^2}, \frac{t}{\|\mathbf{A}_m^T\mathbf{B}\mathbf{A}_m\|_2}\right)\right\}
\end{aligned}
$$

and

$$
\text{(B.7)} \quad P\left[|\boldsymbol{b}^T\boldsymbol{Y}_m| \ge t/2\right] = P\left[|\boldsymbol{b}^T\mathbf{A}_m\boldsymbol{e}_m| \ge t/2\right] \le \exp(1)\exp\left\{-\frac{c_2t^2}{4c_1^2\|\boldsymbol{b}^T\mathbf{A}_m\|_F^2}\right\}.
$$

Letting $\varepsilon_2 \le \sqrt{d}$, then by (B.5), we have

$$
\begin{aligned}
& \varphi_{\max}(\mathbf{A}_m^T\mathbf{A}_m) = \varphi_{\max}(\mathbf{A}_m\mathbf{A}_m^T) = \|\mathbf{A}_m\mathbf{A}_m^T\|_2 \\
& \le \|\mathbf{A}_m\mathbf{A}_m^T - \mathbf{A}\mathbf{A}^T\|_2 + \|\mathbf{A}\mathbf{A}^T\|_2 \le \|\mathbf{A}_m\mathbf{A}_m^T - \mathbf{I}_{d\times d}\|_F + 1 \\
& = \sqrt{2\sum_{i=1}^d\sum_{k>i}^d\left(\sum_{j=1}^m a_{ij}a_{kj}\right)^2 + \sum_{i=1}^d\left(\sum_{j=1}^m a_{ij}^2 - 1\right)^2} + 1 \\
& \le \sqrt{(d^2-d)\varepsilon_2^2/d^4 + d\varepsilon_2^2/d^2} + 1 \le \sqrt{\varepsilon_2^2(d^{-2}+d^{-1})} + 1 \le 9,
\end{aligned}
$$

By Lemma 1 in Lam and Fan (2009), we have $\|\mathbf{M}_1\mathbf{M}_2\|_F \le \|\mathbf{M}_1\|_2\|\mathbf{M}_2\|_F$ for real matices $\mathbf{M}_1$ and $\mathbf{M}_2$ of appropriate sizes. Thus,

$$
\begin{aligned}
\|\mathbf{A}_m^T\mathbf{B}\mathbf{A}_m\|_F & \le \|\mathbf{A}_m^T\|_2\|\mathbf{B}\mathbf{A}_m\|_F = \|\mathbf{A}_m^T\|_2\|\mathbf{A}_m^T\mathbf{B}^T\|_F \\
& \le \|\mathbf{A}_m^T\|_2^2\|\mathbf{B}^T\|_F = \varphi_{\max}(\mathbf{A}_m\mathbf{A}_m^T)\|\mathbf{B}\|_F \le 9\|\mathbf{B}\|_F,
\end{aligned}
$$

$$
\begin{aligned}
\|\mathbf{A}_m^T\mathbf{B}\mathbf{A}_m\|_2 & \le \|\mathbf{A}_m^T\|_2\|\mathbf{A}_m\|_2\|\mathbf{B}\|_2 = \sqrt{\varphi_{\max}(\mathbf{A}_m\mathbf{A}_m^T)\varphi_{\max}(\mathbf{A}_m^T\mathbf{A}_m)}\|\mathbf{B}\|_2 \\
& = \varphi_{\max}(\mathbf{A}_m\mathbf{A}_m^T)\|\mathbf{B}\|_2 \le 9\|\mathbf{B}\|_2,
\end{aligned}
$$

and

$$\|\boldsymbol{b}^T\mathbf{A}_m\|_F = \|\mathbf{A}_m^T\boldsymbol{b}\|_F \le \|\mathbf{A}_m^T\|_2\|\boldsymbol{b}\|_F = \sqrt{\varphi_{\max}(\mathbf{A}_m\mathbf{A}_m^T)}\|\boldsymbol{b}\|_F \le 3\|\boldsymbol{b}\|_F.$$

Then from (B.6) and (B.7), we have

$$P\left[\left|\boldsymbol{Y}_m^T\mathbf{B}\boldsymbol{Y}_m - \sum_{1\le i,k\le d} b_{ik}\sum_{j=1}^m a_{ij}a_{kj}\right| \ge t/2\right] \le 2\exp\left\{-c\min\left(\frac{t^2}{\|\mathbf{B}\|_F^2}, \frac{t}{\|\mathbf{B}\|_2}\right)\right\}$$

and

$$P\left[\left|\boldsymbol{b}^T\boldsymbol{Y}_m\right| \ge t/2\right] \le \exp(1)\cdot\exp\left\{-\frac{ct^2}{\|\boldsymbol{b}\|_F^2}\right\}$$

with some constant $c > 0$ only dependent on K. Letting $\varepsilon_1 = t/4$ and $\varepsilon_2 \le \min\{t(8|\mathbf{B}|_\infty)^{-1}, \sqrt{d}\}$, then by (B.3), (B.4) and (B.5), we obtain

$$\begin{aligned}
P\left[\left|\boldsymbol{Y}^T\mathbf{B}\boldsymbol{Y} - E[\boldsymbol{Y}^T\mathbf{B}\boldsymbol{Y}]\right| \ge t\right] &= P\left[\left|\boldsymbol{Y}^T\mathbf{B}\boldsymbol{Y} - \sum_{i=1}^d b_{ii}\right| \ge t\right] \\
&\le P\left[\left|\boldsymbol{Y}^T\mathbf{B}\boldsymbol{Y} - \boldsymbol{Y}_m^T\mathbf{B}\boldsymbol{Y}_m\right| + \left|\sum_{1\le i,k\le d} b_{ik}\sum_{j=1}^m a_{ij}a_{kj} - \sum_{i=1}^d b_{ii}\right| \ge t/2\right] \\
&\quad + P\left[\left|\boldsymbol{Y}_m^T\mathbf{B}\boldsymbol{Y}_m - \sum_{1\le i,k\le d} b_{ik}\sum_{j=1}^m a_{ij}a_{kj}\right| \ge t/2\right] \\
&\le P\Bigg[\left|\boldsymbol{Y}^T\mathbf{B}\boldsymbol{Y} - \boldsymbol{Y}_m^T\mathbf{B}\boldsymbol{Y}_m\right| + \sum_{i=1}^d |b_{ii}|\left|\sum_{j=1}^m a_{ij}^2 - 1\right| \\
&\qquad + \sum_{i=1}^d \sum_{k\ne i,1\le k\le d} |b_{ik}|\left|\sum_{j=1}^m a_{ij}a_{kj}\right| \ge t/2\Bigg] \\
&\quad + 2\exp\left\{-c\min\left(\frac{t^2}{\|\mathbf{B}\|_F^2}, \frac{t}{\|\mathbf{B}\|_2}\right)\right\} \\
&\le P\left[\left|\boldsymbol{Y}^T\mathbf{B}\boldsymbol{Y} - \boldsymbol{Y}_m^T\mathbf{B}\boldsymbol{Y}_m\right| + t/8 + t/8 \ge t/2\right] \\
&\quad + 2\exp\left\{-c\min\left(\frac{t^2}{\|\mathbf{B}\|_F^2}, \frac{t}{\|\mathbf{B}\|_2}\right)\right\} \\
&\le \delta + 2\exp\left\{-c\min\left(\frac{t^2}{\|\mathbf{B}\|_F^2}, \frac{t}{\|\mathbf{B}\|_2}\right)\right\} \qquad \text{(B.8)}
\end{aligned}$$

and

$$
\begin{aligned}
P[|\boldsymbol{b}^T\boldsymbol{Y}| \geq t] &\leq P[|\boldsymbol{b}^T\boldsymbol{Y} - \boldsymbol{b}^T\boldsymbol{Y}_m| \geq t/2] + P[|\boldsymbol{b}^T\boldsymbol{Y}_m| \geq t/2] \\
&\leq \delta + \exp(1) \cdot \exp\left\{-\frac{ct^2}{\|\boldsymbol{b}\|_F^2}\right\}. \qquad \text{(B.9)}
\end{aligned}
$$

Letting $\delta \to 0$ on both sides of inequalities (B.8) and (B.9), we obtain (B.1) and (B.2).

□

Lemma B.2. *Let $\boldsymbol{e} = (e_1, e_2, \dots)^T$ be the same as that in Lemma B.1, and let $\boldsymbol{X} = \mathbf{B}\boldsymbol{e}$ be a well-defined random vector with length d in the sense of entrywise almost-sure convergence and mean-square convergence. Assume the covariance matrix of $\boldsymbol{X}$, denoted as $\boldsymbol{\Sigma}_x$, is positive definite. Then for $u \geq 0$, there exists a constant $c > 0$ only dependent on K such that*

$$
P\left[|\bar{X}|^2 \geq u\right] \leq \exp(1)\exp\{-cdu\} + \exp(1)\exp\left\{-\frac{cdu}{\|\boldsymbol{\Sigma}_x - \mathbf{I}_{d\times d}\|_1}\right\}, \qquad \text{(B.10)}
$$

with the second term on the right hand side (RHS) of (B.10) *being 0 when $\boldsymbol{\Sigma}_x = \mathbf{I}_{d\times d}$.*

Proof. We consider the nontrivial case when $u > 0$. Since $\boldsymbol{\Sigma}_x$ is positive definite, there exists a symmetric positive definite matrix $\boldsymbol{\Sigma}_x^{1/2}$ such that $\boldsymbol{\Sigma}_x = \boldsymbol{\Sigma}_x^{1/2}\boldsymbol{\Sigma}_x^{1/2}$. Let $\boldsymbol{Y} = \boldsymbol{\Sigma}_x^{-1/2}\boldsymbol{X}$ and $\mathbf{A} = \boldsymbol{\Sigma}_x^{-1/2}\mathbf{B}$, then $\boldsymbol{Y} = \boldsymbol{\Sigma}_x^{-1/2}\mathbf{B}\boldsymbol{e} = \mathbf{A}\boldsymbol{e}$. Thus, $\mathbf{A}\mathbf{A}^T = \mathbf{A}\mathrm{cov}(\boldsymbol{e})\mathbf{A}^T = \mathrm{cov}(\mathbf{A}\boldsymbol{e}) = \mathrm{cov}(\boldsymbol{Y}) = \mathrm{cov}(\boldsymbol{\Sigma}_x^{-1/2}\boldsymbol{X}) = \boldsymbol{\Sigma}_x^{-1/2}\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_x^{-1/2} = \mathbf{I}_{d\times d}$, where the second equality holds for the infinite-dimensional $\boldsymbol{e}$ according to Proposition 2.7.1 in Brockwell and Davis (1991). We have

$$
\begin{aligned}
P\left[|\bar{X}|^2 \geq u\right] &= P\left[|\bar{Y} + \bar{X} - \bar{Y}| \geq \sqrt{u}\right] \\
&\leq P\left[|\bar{Y}| \geq \frac{\sqrt{u}}{2}\right] + P\left[\left|\frac{1}{d}\mathbf{1}_d^T(\boldsymbol{\Sigma}_x^{1/2} - \mathbf{I}_{d\times d})\boldsymbol{Y}\right| \geq \frac{\sqrt{u}}{2}\right]. \qquad \text{(B.11)}
\end{aligned}
$$

Consider the nontrivial case when $\boldsymbol{\Sigma}_x \neq \mathbf{I}_{d\times d}$. By (B.2) in Lemma B.1 with a redefined constant $c > 0$ only dependent on K, we have

$$
P\left[|\bar{Y}| \geq \frac{\sqrt{u}}{2}\right] \leq \exp(1)\exp\{-cdu\} \qquad \text{(B.12)}
$$

and

$$
\begin{aligned}
&P\left[\left|\frac{1}{d}\mathbf{1}_d^T(\boldsymbol{\Sigma}_x^{1/2}-\mathbf{I}_{d\times d})\boldsymbol{Y}\right| \geq \frac{\sqrt{u}}{2}\right] \\
&\leq \exp(1)\exp\left\{-\frac{cu}{\|\frac{1}{d}\mathbf{1}_d^T(\boldsymbol{\Sigma}_x^{1/2}-\mathbf{I}_{d\times d})\|_F^2}\right\} \\
&\leq \exp(1)\exp\left\{-\frac{cu}{\|\frac{1}{d}\mathbf{1}_d\|_F^2\|\boldsymbol{\Sigma}_x^{1/2}-\mathbf{I}_{d\times d}\|_2^2}\right\} \\
&= \exp(1)\exp\left\{-\frac{cdu}{\|\boldsymbol{\Sigma}_x^{1/2}-\mathbf{I}_{d\times d}\|_2^2}\right\} \\
&\leq \exp(1)\exp\left\{-\frac{cdu}{\|\boldsymbol{\Sigma}_x-\mathbf{I}_{d\times d}\|_1}\right\} \qquad \text{(B.13)}
\end{aligned}
$$

The second inequality in (B.13) is obtained from Lemma 1 in Lam and Fan (2009). The last inequality in (B.13) follows from

$$
\begin{aligned}
\|\boldsymbol{\Sigma}_x^{1/2}-\mathbf{I}_{d\times d}\|_2^2 &= \max_i |\varphi_i^{1/2}-1|^2 \leq \max_i |(\varphi_i^{1/2}-1)(\varphi_i^{1/2}+1)| \\
&= \max_i |\varphi_i - 1| = \|\boldsymbol{\Sigma}_x-\mathbf{I}_{d\times d}\|_2 \leq \|\boldsymbol{\Sigma}_x-\mathbf{I}_{d\times d}\|_1,
\end{aligned}
$$

where $\varphi_i > 0$, $i = 1,\ldots,d$, are the eigenvalues of $\boldsymbol{\Sigma}_x$. Plugging (B.12) and (B.13) into (B.11) yields (B.10). □

Lemma B.3. *If $\tau_0 = \sqrt{f\log(pf)/n} = o(1)$, then for any positive constants M', c_1, c_2, there exists a constant $M > 0$ such that for sufficiently large n,*

$$
p^{c_1} f \exp\left\{-\frac{c_2 nu}{f}\right\} < p^{c_1} f \exp\left\{-\frac{c_2 nu^2}{f}\right\} \leq p^{-M'},
$$

where $u = M\tau_0$.

Proof. By $\tau_0 = o(1)$, for any constant $M > 0$, there exists a constant $N(M) > 0$ such that when $n > N(M)$, we have $u = M\tau_0 < 1$, thus $p^{c_1} f \exp\{-c_2 nu/f\} <$

$p^{c_1} f \exp\{-c_2 nu^2/f\}$. Since

$$\begin{aligned}
p^{c_1} f \exp\left\{-\frac{c_2 nu^2}{f}\right\} &= \exp\left\{\left(c_1 + \frac{\log f}{\log p} - \frac{c_2 nu^2}{f \log p}\right) \log p\right\} \\
&= \exp\left\{\left(c_1 + \frac{\log f}{\log p} - \frac{c_2 n M^2 \tau_0^2}{f \log p}\right) \log p\right\} \\
&= \exp\left\{\left[c_1 + \frac{\log f}{\log p} - c_2 M^2 \left(1 + \frac{\log f}{\log p}\right)\right] \log p\right\} \\
&= \exp\left\{\left[-(c_2 M^2 - c_1) - (c_2 M^2 - 1) \frac{\log f}{\log p}\right] \log p\right\},
\end{aligned}$$

for any constant $M' > 0$, choosing a constant M such that $c_2 M^2 - c_1 \geq M'$ and $c_2 M^2 - 1 \geq 0$, i.e., $M \geq \sqrt{\max\{(c_1 + M')/c_2, 1/c_2\}}$, we have $p^{c_1} f \exp\{-c_2 nu^2/f\} \leq p^{-M'}$. □

B.2 Proofs of the General Theorems

B.2.1 Proof of Theorem III.7

Proof of Theorem III.7 (a). Similar to the case of i.i.d. data discussed in Bickel and Levina (2008a) and Rothman et al. (2009), the key to the proof is to find a desirable probabilistic bound of $\max_{1 \leq i,j \leq p} |\hat{\sigma}_{ij} - \sigma_{ij}|$ for temporally dependent observations. Once the bound is established, the remaining of the proof for the convergence in probability follows the same steps as those in the aforementioned literature.

Without loss of generality, we assume $\boldsymbol{\mu}_p = \mathbf{0}$. We only consider data generated from (3.3) with $m = \infty$ because any case with finite m can be constructed by adding infinite number of zero columns in $\mathbf{H}$. Since

$$\begin{aligned}
\max_{1 \leq i,j \leq p} |\hat{\sigma}_{ij} - \sigma_{ij}| &\leq \max_{1 \leq i,j \leq p} \left|\bar{X}_i \bar{X}_j\right| + \max_{1 \leq i,j \leq p} \left|\frac{1}{n} \sum_{k=1}^{n} X_{ik} X_{jk} - \sigma_{ij}\right| \\
&\leq \max_{1 \leq i \leq p} \left|\bar{X}_i\right|^2 + \max_{1 \leq i,j \leq p} \left|\frac{1}{n} \sum_{k=1}^{n} X_{ik} X_{jk} - \sigma_{ij}\right|, \qquad \text{(B.14)}
\end{aligned}$$

for any $u > 0$, we have

$$P\left[\max_{1\le i,j\le p} |\hat{\sigma}_{ij} - \sigma_{ij}| \ge 2u\right]$$

$$\le P\left[\max_{1\le i\le p} \left|\bar{X}_i\right|^2 \ge u\right] + P\left[\max_{1\le i,j\le p}\left|\frac{1}{n}\sum_{k=1}^{n} X_{ik}X_{jk} - \sigma_{ij}\right| \ge u\right]. \tag{B.15}$$

Let $Z_{ij} = X_{ij}/\sqrt{\sigma_{ii}}$, $A_{r,f} = \{k \in \mathbb{Z}^+ \cup \{0\} : kf + r \le n\},\ r \in \{1, \ldots, f\}$, and $C_{r,f}$ be the cardinality of $A_{r,f}$. For a fixed integer f and any integer $1 \le j \le n$, we have $j = kf + r$, where $k = \lfloor j/f \rfloor$ if j/f is not an integer, otherwise $k = j/f - 1$. Hence,

$$\sum_{j=1}^{n} X_{ij} = \sum_{r=1}^{f} \sum_{k \in A_{r,f}} X_{i,kf+r} \quad \text{and} \quad n = \sum_{r=1}^{f} C_{r,f}.$$

Moreover, for any $r \in \{1, \ldots, f\}$,

$$n/f - 2 \le \lfloor n/f \rfloor - 1 \le C_{r,f} - 1 \le \lfloor n/f \rfloor \le n/f,$$

thus $n - f \le fC_{r,f} \le 2n$. By $\tau_0 = \sqrt{f\log(pf)/n} = o(1)$, we have $f = o(n)$. Hence, there exists a constant N_1 such that when $n > N_1$, we have

$$n/2 \le fC_{r,f} \le 2n. \tag{B.16}$$

We assume $n > N_1$ in the following.

Now, for the first term on the RHS of (B.15), following from (B.16) and $\sigma_{ii} \le v_0$ we have

$$
\begin{aligned}
P\left[\max_{1\le i\le p}|\bar{X}_i|^2\ge u\right] &\le \sum_{i=1}^{p} P\left[|\bar{X}_i|\ge u^{\frac{1}{2}}\right]\\
&= \sum_{i=1}^{p} P\left[\left|\sum_{j=1}^{n} X_{ij}\right|\ge nu^{\frac{1}{2}}\right]\\
&= \sum_{i=1}^{p} P\left[\left|\sum_{r=1}^{f}\sum_{k\in A_{r,f}} X_{i,kf+r}\right|\ge nu^{\frac{1}{2}}\right]\\
&\le \sum_{i=1}^{p} P\left[\sum_{r=1}^{f}\left|\sum_{k\in A_{r,f}} X_{i,kf+r}\right|\ge nu^{\frac{1}{2}}\right]\\
&\le \sum_{i=1}^{p}\sum_{r=1}^{f} P\left[\left|\sum_{k\in A_{r,f}} X_{i,kf+r}\right|\ge \frac{nu^{1/2}}{f}\right]\\
&= \sum_{i=1}^{p}\sum_{r=1}^{f} P\left[\left|\frac{1}{C_{r,f}}\sum_{k\in A_{r,f}} Z_{i,kf+r}\right|\ge \frac{n}{fC_{r,f}}\sqrt{\frac{u}{\sigma_{ii}}}\right]\\
&\le \sum_{i=1}^{p}\sum_{r=1}^{f} P\left[\left|\frac{1}{C_{r,f}}\sum_{k\in A_{r,f}} Z_{i,kf+r}\right|\ge \frac{1}{2}\sqrt{\frac{u}{v_0}}\right]. \qquad \text{(B.17)}
\end{aligned}
$$

Let $\mathbf{\Delta}^{ifr}$ be the covariance matrix of vec$\{Z_{i,kf+r} : k\in A_{r,f}\}$, then

$$
\begin{aligned}
\|\mathbf{\Delta}^{ifr}-\mathbf{I}_{C_{r,f}\times C_{r,f}}\|_1 &= \max_{l}\sum_{k\ne l}|\rho_{ii}^{kf+r,lf+r}|\\
&\le \max_{1\le b\le n}\sum_{\substack{a\in\{1\le a\le n:\\ |a-b|=kf,\\ k=1,\dots,\lfloor n/f\rfloor\}}}|\mathbf{R}^{ab}|_\infty \le g(n,p) \qquad \text{(B.18)}
\end{aligned}
$$

by assumption (3.28). Since $\limsup_{n\to\infty} g(n,p) < 1$, there exists a constant $c_1 > 0$ such that $\limsup_{n\to\infty} g(n,p) < c_1 < 1$. Then there exists a constant $N_2(c_1) > 0$ such that $g < c_1 < 1$ when $n > N_2(c_1)$. We now assume $n > \max\{N_1, N_2(c_1)\}$. By (B.18), $\mathbf{\Delta}^{ifr}$ is a strictly diagonally dominated matrix, thus positive definite by the Levy-Desplanques theorem (Horn and Johnson, 2013). From equation (3.3), we have

$$
\text{vec}\{Z_{i,kf+r} : k\in A_{r,f}\} = \mathbf{P}^{ifr}\mathbf{H}\boldsymbol{e}, \qquad \text{(B.19)}
$$

where $\mathbf{P}^{ifr}$ is a $C_{r,f} \times pn$ matrix with $\sigma_{ii}^{-1/2}$ in the $\big(k+1, i+(kf+r-1)p\big)$ entries and 0 in all other entries for $k \in A_{r,f}$. By Proposition 2.7.1 in Brockwell and Davis (1991), $\boldsymbol{\Delta}^{ifr} = \mathbf{P}^{ifr}\mathbf{H}(\mathbf{P}^{ifr}\mathbf{H})^T$ holds for the case when $m = \infty$, and since $\boldsymbol{\Delta}^{ifr}$ for all $r \in \{1, \ldots, f\}$ are positive definite, $\mathbf{H}$ has rank no less than $\max_{1 \le r \le f} C_{r,f} = \lfloor (n-1)/f \rfloor + 1$. By (B.19), Lemma B.2, (B.18), (B.16) and $g < 1$, we have

$$\begin{aligned}
&P\left[\left|\frac{1}{C_{r,f}} \sum_{k \in A_{r,f}} Z_{i,kf+r}\right| \ge \frac{1}{2}\sqrt{\frac{u}{v_0}}\right] \\
&\le \exp(1)\exp\left\{-\frac{c_2 C_{r,f} u}{4v_0}\right\} + \exp(1)\exp\left\{-\frac{c_2 C_{r,f} u}{4v_0 g}\right\} \\
&\le 2\exp(1)\exp\left\{-\frac{c_2 nu}{8v_0 f}\right\} \qquad \text{(B.20)}
\end{aligned}$$

with some constant $c_2 > 0$. Plugging (B.20) into (B.17) yields

$$P\left[\max_{1 \le i \le p} |\bar{X}_i|^2 \ge u\right] \le 2pf\exp(1)\exp\left\{-\frac{c_2 nu}{8v_0 f}\right\}. \qquad \text{(B.21)}$$

For the second term on the RHS of (B.15), we use a similar argument in Bhattacharjee and Bose (2014). Note that

$$\begin{aligned}
&P\left[\max_{1 \le i,j \le p}\left|\frac{1}{n}\sum_{k=1}^{n} X_{ik}X_{jk} - \sigma_{ij}\right| \ge u\right] \\
&\le \sum_{1 \le i,j \le p} P\left[\left|\frac{1}{n}\sum_{k=1}^{n} X_{ik}X_{jk} - \sigma_{ij}\right| \ge u\right] \\
&= \sum_{1 \le i,j \le p} P\left[\left|\sum_{k=1}^{n} Z_{ik}Z_{jk} - n\rho_{ij}\right| \ge \frac{nu}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right] \\
&\le \sum_{1 \le i,j \le p} P\left[\left|\sum_{k=1}^{n} (Z_{ik} + C_g Z_{jk})^2 - n\left(1 + C_g^2 + 2C_g\rho_{ij}\right)\right| \ge \frac{2C_g nu}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right] \\
&\quad + \sum_{1 \le i,j \le p} P\left[\left|\sum_{k=1}^{n} (Z_{ik} - C_g Z_{jk})^2 - n\left(1 + C_g^2 - 2C_g\rho_{ij}\right)\right| \ge \frac{2C_g nu}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right], \qquad \text{(B.22)}
\end{aligned}$$

where constant $C_g = \frac{1-\sqrt{c_1}}{1+\sqrt{c_1}} \in (0,1)$. Without loss of generality, we only consider the second term on the RHS of the above inequality. Then

$$
\begin{aligned}
&P\left[\left|\sum_{k=1}^{n}\left(Z_{ik}-C_gZ_{jk}\right)^2-n\left(1+C_g^2-2C_g\rho_{ij}\right)\right|\geq\frac{2C_gnu}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right]\\
&=P\left[\left|\sum_{r=1}^{f}\sum_{k\in A_{r,f}}\left(Z_{i,kf+r}-C_gZ_{j,kf+r}\right)^2\right.\right.\\
&\qquad\left.\left.-\sum_{r=1}^{f}C_{r,f}\left(1+C_g^2-2C_g\rho_{ij}\right)\right|\geq\frac{2C_gnu}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right]\\
&\leq P\left[\sum_{r=1}^{f}\left|\sum_{k\in A_{r,f}}\left(\frac{Z_{i,kf+r}-C_gZ_{j,kf+r}}{\sqrt{1+C_g^2-2C_g\rho_{ij}}}\right)^2-C_{r,f}\right|\right.\\
&\qquad\left.\geq\frac{2C_gnu}{(1+C_g^2-2C_g\rho_{ij})\sqrt{\sigma_{ii}\sigma_{jj}}}\right]\\
&\leq\sum_{r=1}^{f}P\left[\left|\sum_{k\in A_{r,f}}\left(\frac{Z_{i,kf+r}-C_gZ_{j,kf+r}}{\sqrt{1+C_g^2-2C_g\rho_{ij}}}\right)^2-C_{r,f}\right|\right.\\
&\qquad\left.\geq\frac{2C_gnu}{(1+C_g^2-2C_g\rho_{ij})fv_0}\right].
\end{aligned}
\tag{B.23}
$$

Let

$$
\boldsymbol{Z}=\mathrm{vec}\left\{\frac{Z_{i,kf+r}-C_gZ_{j,kf+r}}{\sqrt{1+C_g^2-2C_g\rho_{ij}}}:k\in A_{r,f}\right\},
$$

and $\boldsymbol{\Gamma}:=(\gamma_{kl})_{C_{r,f}\times C_{r,f}}=\mathrm{cov}(\boldsymbol{Z})$, where for $k,l\in A_{r,f}$,

$$
\gamma_{kl}=\begin{cases}\Big[\rho_{ii}^{kf+r,lf+r}-C_g(\rho_{ij}^{kf+r,lf+r}+\rho_{ji}^{kf+r,lf+r})\\ \qquad\qquad+C_g^2\rho_{jj}^{kf+r,lf+r}\Big](1+C_g^2-2C_g\rho_{ij})^{-1}, & k\neq l;\\ 1, & k=l.\end{cases}
$$

Similar to (B.18), we have

$$
\begin{aligned}
\|\boldsymbol{\Gamma}-\mathbf{I}_{C_{r,f}\times C_{r,f}}\|_1&\leq(1+2C_g+C_g^2)(1+C_g^2-2C_g\rho_{ij})^{-1}\max_{1\leq b\leq n}\sum_{\substack{a\in\{1\leq a\leq n:\\|a-b|=kf,\\k=1,\ldots,\lfloor n/f\rfloor\}}}|\mathbf{R}^{ab}|_\infty\\
&\leq(1+C_g)^2(1+C_g^2-2C_g\rho_{ij})^{-1}g(n,p)\\
&\leq\left(\frac{1+C_g}{1-C_g}\right)^2g(n,p)=g(n,p)/c_1<1,
\end{aligned}
\tag{B.24}
$$

thus $\mathbf{\Gamma} \succ 0$. Then there exists a symmetric positive definite matrix $\mathbf{\Gamma}^{1/2}$ such that $\mathbf{\Gamma} = \mathbf{\Gamma}^{1/2}\mathbf{\Gamma}^{1/2}$. Let $\boldsymbol{Y} = \mathbf{\Gamma}^{-1/2}\boldsymbol{Z}$. Then by (B.19), we have

$$\boldsymbol{Y} = \mathbf{A}\boldsymbol{e}, \text{ with } \mathbf{A}\mathbf{A}^T = \mathbf{I}_{C_{r,f}\times C_{r,f}}, \tag{B.25}$$

where

$$\mathbf{A} = \frac{\mathbf{\Gamma}^{-1/2}(\mathbf{P}^{ifr} - C_g\mathbf{P}^{jfr})\mathbf{H}}{\sqrt{1 + C_g^2 - 2C_g\rho_{ij}}},$$

and the second equality in (B.25) is from Proposition 2.7.1 in Brockwell and Davis (1991) which gives $\mathbf{A}\mathbf{A}^T = \mathbf{A}\text{cov}(\boldsymbol{e})\mathbf{A}^T = \text{cov}(\mathbf{A}\mathbf{e}) = \text{cov}(\boldsymbol{Y}) = \mathbf{I}_{C_{r,f}\times C_{r,f}}$. Now we have

$$\begin{aligned}
&P\left[\left|\sum_{k\in A_{r,f}}\left(\frac{Z_{i,kf+r} - C_g Z_{j,kf+r}}{\sqrt{1 + C_g^2 - 2C_g\rho_{ij}}}\right)^2 - C_{r,f}\right| \geq \frac{2C_g nu}{(1 + C_g^2 - 2C_g\rho_{ij})fv_0}\right]\\
&= P\left[|\boldsymbol{Y}^T\mathbf{\Gamma}\boldsymbol{Y} - C_{r,f}| \geq \frac{2C_g nu}{(1 + C_g^2 - 2C_g\rho_{ij})fv_0}\right]\\
&\leq P\left[|\boldsymbol{Y}^T(\mathbf{\Gamma} - \mathbf{I}_{C_{r,f}\times C_{r,f}})\boldsymbol{Y}| \geq \frac{C_g nu}{(1 + C_g^2 - 2C_g\rho_{ij})fv_0}\right]\\
&\quad + P\left[|\boldsymbol{Y}^T\boldsymbol{Y} - C_{r,f}| \geq \frac{C_g nu}{(1 + C_g^2 - 2C_g\rho_{ij})fv_0}\right].
\end{aligned} \tag{B.26}$$

The first term on the RHS of (B.26) obviously equals zero for $u > 0$ when $\mathbf{\Gamma} = \mathbf{I}_{C_{r,f}\times C_{r,f}}$, thus we only consider the case that $\mathbf{\Gamma} \neq \mathbf{I}_{C_{r,f}\times C_{r,f}}$. By the fact that $E[\boldsymbol{Y}^T(\mathbf{\Gamma} - \mathbf{I}_{C_{r,f}\times C_{r,f}})\boldsymbol{Y}] =$

$\mathrm{tr}(\mathbf{\Gamma} - \mathbf{I}_{C_{r,f}\times C_{r,f}}) = 0$, (B.25), (B.1) in Lemma B.1, (B.24), (B.16) and $g < 1$, we have

$$\begin{aligned}
&P\left[|\boldsymbol{Y}^T(\mathbf{\Gamma} - \mathbf{I}_{C_{r,f}\times C_{r,f}})\boldsymbol{Y}| \geq \frac{C_g nu}{(1 + C_g^2 - 2C_g\rho_{ij})fv_0}\right] \\
&\leq 2\exp\Bigg\{ - c_3 \min\Bigg(\frac{(C_g nu)^2}{[(1 + C_g^2 - 2C_g\rho_{ij})fv_0]^2 \|\mathbf{\Gamma} - \mathbf{I}_{C_{r,f}\times C_{r,f}}\|_F^2}, \\
&\qquad\qquad \frac{C_g nu}{(1 + C_g^2 - 2C_g\rho_{ij})fv_0 \|\mathbf{\Gamma} - \mathbf{I}_{C_{r,f}\times C_{r,f}}\|_2} \Bigg) \Bigg\} \\
&\leq 2\exp\Bigg\{ - c_3 \min\Bigg(\frac{(C_g nu)^2}{[(1 + C_g^2 - 2C_g\rho_{ij})fv_0]^2 C_{r,f} \|\mathbf{\Gamma} - \mathbf{I}_{C_{r,f}\times C_{r,f}}\|_1^2}, \\
&\qquad\qquad \frac{C_g nu}{(1 + C_g^2 - 2C_g\rho_{ij})fv_0 \|\mathbf{\Gamma} - \mathbf{I}_{C_{r,f}\times C_{r,f}}\|_1} \Bigg) \Bigg\} \\
&\leq 2\exp\Bigg\{ - c_3 \min\Bigg(\frac{C_g^2 n^2 u^2}{f^2 v_0^2 C_{r,f}(1 + C_g)^4 g^2}, \frac{C_g nu}{fv_0(1 + C_g)^2 g} \Bigg) \Bigg\} \\
&\leq 2\exp\Bigg\{ - c_3 \min\Bigg(\frac{C_g^2 nu^2}{2v_0^2(1 + C_g)^4 f}, \frac{C_g nu}{v_0(1 + C_g)^2 f} \Bigg) \Bigg\} \qquad \text{(B.27)}
\end{aligned}$$

with some constant $c_3 > 0$. Similarly,

$$\begin{aligned}
&P\left[|\boldsymbol{Y}^T\boldsymbol{Y} - C_{r,f}| \geq \frac{C_g nu}{(1 + C_g^2 - 2C_g\rho_{ij})fv_0}\right] \\
&\leq 2\exp\Bigg\{ - c_3 \min\Bigg(\frac{(C_g nu)^2}{[(1 + C_g^2 - 2C_g\rho_{ij})fv_0]^2 C_{r,f}}, \\
&\qquad\qquad \frac{C_g nu}{(1 + C_g^2 - 2C_g\rho_{ij})fv_0} \Bigg) \Bigg\} \\
&\leq 2\exp\Bigg\{ - c_3 \min\Bigg(\frac{C_g^2 nu^2}{2v_0^2(1 + C_g)^4 f}, \frac{C_g nu}{v_0(1 + C_g)^2 f} \Bigg) \Bigg\}. \qquad \text{(B.28)}
\end{aligned}$$

By (B.23), (B.26), (B.27) and (B.28), we have

$$\begin{aligned}
&P\left[\left|\sum_{k=1}^{n} (Z_{ik} - C_g Z_{jk})^2 - n\left(1 + C_g^2 - 2C_g\rho_{ij}\right)\right| \geq \frac{2C_g nu}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right] \\
&\leq 4f \exp\left\{ -c_4 \min\left(\frac{nu^2}{f}, \frac{nu}{f} \right) \right\},
\end{aligned}$$

with some constant $c_4 > 0$. Similarly,

$$P\left[\left|\sum_{k=1}^{n}(Z_{ik}+C_gZ_{jk})^2 - n\left(1+C_g^2+2C_g\rho_{ij}\right)\right| \geq \frac{2C_gnu}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right]$$
$$\leq 4f\exp\left\{-c_4\min\left(\frac{nu^2}{f},\frac{nu}{f}\right)\right\}.$$

Therefore by (B.22), we have

$$P\left[\max_{1\leq i,j\leq p}\left|\frac{1}{n}\sum_{k=1}^{n}X_{ik}X_{jk}-\sigma_{ij}\right| \geq u\right] \leq 8p^2f\exp\left\{-c_4\min\left(\frac{nu^2}{f},\frac{nu}{f}\right)\right\}. \tag{B.29}$$

From (B.15), (B.21) and (B.29), we obtain

$$P\left[\max_{1\leq i,j\leq p}|\hat{\sigma}_{ij}-\sigma_{ij}| \geq 2u\right]$$
$$\leq 2pf\exp(1)\exp\left\{-\frac{c_2nu}{8v_0f}\right\} + 8p^2f\exp\left\{-c_4\min\left(\frac{nu^2}{f},\frac{nu}{f}\right)\right\}. \tag{B.30}$$

By Lemma B.3, for any constant $M' > 0$, there exists a constant $M_1 > 0$ such that when $M \geq M_1$, we have

$$P\left[\max_{1\leq i,j\leq p}|\hat{\sigma}_{ij}-\sigma_{ij}| \geq \tau\right] = O(p^{-M'}), \quad \text{with } \tau = M\tau_0. \tag{B.31}$$

Then following the similar lines of the proof of Theorem 1 after equation (12) in Bickel and Levina (2008a) and the proof of Theorem 1 in Rothman et al. (2009), we obtain that for any constant $M' > 0$, there exists a constant $M_2 \geq M_1$ such that

$$P\left[\|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}\|_2 \geq C_1c_p\tau_0^{1-q}\right] \leq P\left[\|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}\|_1 \geq C_1c_p\tau_0^{1-q}\right]$$
$$= O(p^{-M'}), \tag{B.32}$$

where $\tau = M\tau_0$ with any constant $M \geq M_2$ and some constant $C_1 > 0$ dependent on M. Thus, we obtain (3.30).

By condition (iii) of the generalized thresholding function and (B.31), we have

$$\begin{aligned} P\left[|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}|_\infty \geq 2\tau\right] &= P\left[\max_{1\leq i,j\leq p}|s_\tau(\hat{\sigma}_{ij})-\sigma_{ij}| \geq 2\tau\right] \\ &\leq P\left[\max_{1\leq i,j\leq p}|s_\tau(\hat{\sigma}_{ij})-\hat{\sigma}_{ij}| + \max_{1\leq i,j\leq p}|\hat{\sigma}_{ij}-\sigma_{ij}| \geq 2\tau\right] \\ &\leq P\left[\tau + \max_{1\leq i,j\leq p}|\hat{\sigma}_{ij}-\sigma_{ij}| \geq 2\tau\right] = O(p^{-M'}). \end{aligned} \tag{B.33}$$

Thus, (3.29) holds. By (B.32), (B.33) and the inequality $\|\mathbf{M}\|_F^2 \leq p\|\mathbf{M}\|_1|\mathbf{M}|_\infty$ for any $p \times p$ matrix $\mathbf{M}$, we have

$$\begin{aligned} & P\left[\frac{1}{p}\|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}\|_F^2 \geq 2\tau C_1 c_p \tau_0^{1-q}\right] \\ &\leq P\left[\|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}\|_1|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}|_\infty \geq 2\tau C_1 c_p \tau_0^{1-q}\right] \\ &\leq P\left[\|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}\|_1 \geq C_1 c_p \tau_0^{1-q}\right] + P\left[|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}|_\infty \geq 2\tau\right] \\ &= O(p^{-M'}). \end{aligned} \tag{B.34}$$

Hence, we obtain (3.31).

For the sparsistency and sign-consistency, the proof follows the similar lines of the proof of Theorem 2 in Rothman et al. (2009) by replacing their equation (A.4) with (B.31). Details are hence omitted.

For the convergence in mean square, we additionally assume $p \geq n^c$ for some constant $c > 0$. Now

$$\begin{aligned} & E\|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}\|_2^2 \\ &= E\left[\|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}\|_2^2\mathbb{I}\left(\|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}\|_2 \geq C_1 c_p \tau_0^{1-q}\right)\right] \\ &\qquad + E\left[\|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}\|_2^2\mathbb{I}\left(\|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}\|_2 < C_1 c_p \tau_0^{1-q}\right)\right] \\ &\leq \left(E\|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}\|_2^4\right)^{\frac{1}{2}}\left(P\left[\|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}\|_2 \geq C_1 c_p \tau_0^{1-q}\right]\right)^{\frac{1}{2}} \\ &\qquad + (C_1 c_p \tau_0^{1-q})^2 \\ &\leq \left(E\|S_\tau(\hat{\mathbf{\Sigma}})-\mathbf{\Sigma}\|_F^4\right)^{\frac{1}{2}} O(p^{-\frac{M'}{2}}) + (C_1 c_p \tau_0^{1-q})^2. \end{aligned} \tag{B.35}$$

We want to show $E\|S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_F^4 = O(p^{c_5})$ with a constant $c_5 > 0$, and then choose a sufficiently large M' to obtain desired result. By condition (iii) of the generalized thresholding function, we have $\|S_\tau(\hat{\mathbf{\Sigma}}) - \hat{\mathbf{\Sigma}}\|_F \leq p\tau$, then

$$\begin{aligned}
E\|S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_F^4 &\leq E\left(\|S_\tau(\hat{\mathbf{\Sigma}}) - \hat{\mathbf{\Sigma}}\|_F + \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F\right)^4 \\
&= E\Big[\|S_\tau(\hat{\mathbf{\Sigma}}) - \hat{\mathbf{\Sigma}}\|_F^4 + \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^4 + 4\|S_\tau(\hat{\mathbf{\Sigma}}) - \hat{\mathbf{\Sigma}}\|_F^3\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F \\
&\qquad + 4\|S_\tau(\hat{\mathbf{\Sigma}}) - \hat{\mathbf{\Sigma}}\|_F\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^3 + 6\|S_\tau(\hat{\mathbf{\Sigma}}) - \hat{\mathbf{\Sigma}}\|_F^2\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^2\Big] \\
&\leq p^4\tau^4 + E\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^4 + 4p^3\tau^3 E\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F + 4p\tau E\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^3 \\
&\qquad + 6p^2\tau^2 E\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^2 \\
&\leq p^4\tau^4 + E\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^4 + 4p^3\tau^3\left(E\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^2\right)^{\frac{1}{2}} + 4p\tau\left(E\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^6\right)^{\frac{1}{2}} \\
&\qquad + 6p^2\tau^2 E\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^2. \qquad \text{(B.36)}
\end{aligned}$$

Since $p \geq n^c$, it is easy to see that for $d = 1, 2, 3$,

$$\text{(B.37)} \qquad \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^{2d} = \left\{\sum_{1\leq i,j\leq p}\left(\frac{1}{n}\sum_{k=1}^n X_{ik}X_{jk} - \bar{X}_i\bar{X}_j - \sigma_{ij}\right)^2\right\}^d$$

is a polynomial of variables X_{ij} of degree $4d$, $1 \leq i \leq p$, $1 \leq j \leq n$, the number of its terms is bounded by p^{C_2} with a constant $C_2 > 0$, and all its coefficients are absolutely bounded by a constant C_3 that only depends on v_0. Denote by $P_k^{(d)}$ the k-th term in the corresponding polynomial of X_{ij} in (B.37). Then by the Hölder's inequality (Karr, 1993), there exist positive constants c_6 and c_7 for all k and d such that

$$\text{(B.38)} \qquad E|P_k^{(d)}| \leq C_3\prod_{i,j}(E|X_{ij}|^{c_{ijkd}})^{\frac{1}{C_{ijkd}}}$$

with appropriate choices of integer constants $c_{ijkd} \in [0, c_6]$ and $C_{ijkd} \in [1, c_7]$, and $\sum_{i,j}\mathbb{I}(c_{ijkd} \neq 0) \leq 4d$. By inequality (3.4) and $\sigma_{ii} \leq v_0$, we have

$$\text{(B.39)} \qquad E(\exp\{t[X_{ij} - E(X_{ij})]\}) \leq \exp\{Kv_0t^2/2\}, \text{ for all } t \in \mathbb{R}.$$

Then by (B.39) and Lemma 5.5 in Vershynin (2012), there exists a constant $c_8 > 0$ only dependent on Kv_0 such that $(E|X_{ij} - E(X_{ij})|^k)^{1/k} \leq c_8\sqrt{k}$ for all $k \geq 1, 1 \leq i \leq p$ and $1 \leq j \leq n$, thus

$$(E|X_{ij}|^k)^{1/k} \leq c_8\sqrt{k} \tag{B.40}$$

with the assumption $\boldsymbol{\mu}_p = \mathbf{0}$. Combining (B.37), (B.38) and (B.40), we have

$$E\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^{2d} = O(p^{C_2}), \text{ for } d = 1, 2, 3. \tag{B.41}$$

Then by (B.36), we obtain $E\|S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_F^4 = O(p^{c_5})$ with some constant $c_5 > 0$. Hence, by (B.35), we have

$$E\|S_\tau(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_2^2 \leq O(p^{\frac{c_5 - M'}{2}}) + (C_1 c_p \tau_0^{1-q})^2.$$

Since $\tau_0 = \sqrt{f\log(pf)/n} \geq \sqrt{n^{-1}\log p} \geq \sqrt{p^{-1/c}\log p}$ following from $p \geq n^c$, we can let M' be sufficiently large such that $p^{c_5 - \frac{M'}{2}} = O\left((c_p\tau_0^{1-q})^2\right)$, then (3.33) holds. By (B.33) and (B.34), we can similarly obtain (3.32) and (3.34) respectively. □

Proof of Theorem III.7 (b). The key of the proof is to show that for any constant $M' > 0$, there exists a constant $C' > 0$ such that

$$P\left[\max_{1\leq i,j\leq p} |\hat{\rho}_{ij} - \rho_{ij}| \geq C'\tau_0\right] = O(p^{-M'}). \tag{B.42}$$

Similar to (B.14),

$$\begin{aligned}
\max_{1\leq i,j\leq p}\left|\frac{\hat{\sigma}_{ij} - \sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right| &\leq \max_{1\leq i,j\leq p}\left|\frac{\bar{X}_i\bar{X}_j}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right| + \max_{1\leq i,j\leq p}\left|\frac{1}{n\sqrt{\sigma_{ii}\sigma_{jj}}}\sum_{k=1}^n X_{ik}X_{jk} - \rho_{ij}\right| \\
&\leq \max_{1\leq i,j\leq p}\left|\frac{\bar{X}_i}{\sqrt{\sigma_{ii}}}\right|^2 + \max_{1\leq i,j\leq p}\left|\frac{1}{n\sqrt{\sigma_{ii}\sigma_{jj}}}\sum_{k=1}^n X_{ik}X_{jk} - \rho_{ij}\right| \\
&= \max_{1\leq i,j\leq p}\left|\bar{Z}_i\right|^2 + \max_{1\leq i,j\leq p}\left|\frac{1}{n}\sum_{k=1}^n Z_{ik}Z_{jk} - \rho_{ij}\right|, \qquad \text{(B.43)}
\end{aligned}$$

where $Z_{ik} = X_{ik}/\sqrt{\sigma_{ii}}$. In (B.43), since $\max_{1\le i,j\le p}|\rho_{ij}| \le 1$, we do not need to assume $\max_{1\le i\le p}|\sigma_{ii}| \le v_0$ any more and we impose the ℓ_q-ball sparsity assumption (3.7) on $\mathbf{R}$ instead of (3.6) on $\mathbf{\Sigma}$. Then following the similar lines of the proof of Theorem III.7 (a) up to equation (B.31), we can obtain that for any constant $M_1 > 0$, there exists a constant $C_1 > 0$ such that

$$O(p^{-M_1}) = P\left[\max_{1\le i,j\le p}\left|\frac{\hat{\sigma}_{ij}-\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right| \ge C_1\tau_0\right] \ge P\left[\left|\frac{\hat{\sigma}_{ij}-\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right| \ge C_1\tau_0\right] \tag{B.44}$$

for any $1 \le i,j \le p$. Thus letting $i=j$, we have

$$O(p^{-M_1}) = P\left[\left|\sqrt{\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}}-1\right|\left|\sqrt{\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}}+1\right| \ge C_1\tau_0\right] \ge P\left[\left|\sqrt{\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}}-1\right| \ge C_1\tau_0\right], \tag{B.45}$$

and

$$\begin{aligned}
O(p^{-M_1}) &= P\left[\left|\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}-1\right| \ge C_1\tau_0\right] + P\left[\left|\frac{\hat{\sigma}_{jj}}{\sigma_{jj}}-1\right| \ge C_1\tau_0\right] \\
&\ge P\left[\left|\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}-1\right|\left|\frac{\hat{\sigma}_{jj}}{\sigma_{jj}}-1\right| \ge C_1^2\tau_0^2\right] \\
&= P\left[\left|\sqrt{\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}}-1\right|\left|\sqrt{\frac{\hat{\sigma}_{jj}}{\sigma_{jj}}}-1\right|\left|\sqrt{\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}}+1\right|\left|\sqrt{\frac{\hat{\sigma}_{jj}}{\sigma_{jj}}}+1\right| \ge C_1^2\tau_0^2\right] \\
&\ge P\left[\left|\sqrt{\frac{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}{\sigma_{ii}\sigma_{jj}}}-\sqrt{\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}}-\sqrt{\frac{\hat{\sigma}_{jj}}{\sigma_{jj}}}+1\right| \ge C_1^2\tau_0^2\right] \\
&\ge P\left[\left|\sqrt{\frac{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}{\sigma_{ii}\sigma_{jj}}}-1\right| \ge C_1^2\tau_0^2+\left|\sqrt{\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}}-1\right|+\left|\sqrt{\frac{\hat{\sigma}_{jj}}{\sigma_{jj}}}-1\right|\right]. \qquad \text{(B.46)}
\end{aligned}$$

By (B.45), (B.46) and $\tau_0 = o(1)$, we have

$$\begin{aligned}
&P\left[\left|\sqrt{\frac{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}{\sigma_{ii}\sigma_{jj}}}-1\right| \geq 3C_1\tau_0\right] \\
&\leq P\left[\left|\sqrt{\frac{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}{\sigma_{ii}\sigma_{jj}}}-1\right| \geq 3C_1\tau_0, \left|\sqrt{\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}}-1\right| \leq C_1\tau_0, \left|\sqrt{\frac{\hat{\sigma}_{jj}}{\sigma_{jj}}}-1\right| \leq C_1\tau_0\right] \\
&\qquad + P\left[\left|\sqrt{\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}}-1\right| \geq C_1\tau_0 \text{ or } \left|\sqrt{\frac{\hat{\sigma}_{jj}}{\sigma_{jj}}}-1\right| \geq C_1\tau_0\right] \\
&\leq P\left[\left|\sqrt{\frac{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}{\sigma_{ii}\sigma_{jj}}}-1\right| \geq C_1\tau_0 + \left|\sqrt{\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}}-1\right| + \left|\sqrt{\frac{\hat{\sigma}_{jj}}{\sigma_{jj}}}-1\right|\right] + O(p^{-M_1}) \\
&\leq P\left[\left|\sqrt{\frac{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}{\sigma_{ii}\sigma_{jj}}}-1\right| \geq C_1^2\tau_0^2 + \left|\sqrt{\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}}-1\right| + \left|\sqrt{\frac{\hat{\sigma}_{jj}}{\sigma_{jj}}}-1\right|\right] + O(p^{-M_1}) \\
&= O(p^{-M_1}). \qquad \text{(B.47)}
\end{aligned}$$

Then,

$$\begin{aligned}
&P\left[\max_{1\leq i,j\leq p} |\hat{\rho}_{ij}-\rho_{ij}| \geq 4C_1\tau_0\right] \\
&\leq P\left[\max_{1\leq i,j\leq p}\left|\frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}} - \frac{\hat{\sigma}_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right| + \max_{1\leq i,j\leq p}\left|\frac{\hat{\sigma}_{ij}-\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right| \geq 4C_1\tau_0\right] \\
&\leq P\left[\max_{1\leq i,j\leq p}\left(\left|\frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}\right|\left|\sqrt{\frac{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}{\sigma_{ii}\sigma_{jj}}}-1\right|\right) \geq 3C_1\tau_0\right] + P\left[\max_{1\leq i,j\leq p}\left|\frac{\hat{\sigma}_{ij}-\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right| \geq C_1\tau_0\right] \\
&\leq P\left[\max_{1\leq i,j\leq p}\left|\sqrt{\frac{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}{\sigma_{ii}\sigma_{jj}}}-1\right| \geq 3C_1\tau_0\right] + P\left[\max_{1\leq i,j\leq p}\left|\frac{\hat{\sigma}_{ij}-\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right| \geq C_1\tau_0\right] \\
&= O(p^{-M_1+2}),
\end{aligned}$$

following from (B.44) and (B.47). Equation (B.42) holds by letting $C' = 4C_1$ and $M' = M_1 - 2 > 0$. Then the proof follows similar lines of the proof of Theorem III.7 (a) after equation (B.31), where we simply use $\|S_\tau(\hat{\mathbf{R}}) - \mathbf{R}\|_F^4 \leq 16p^4$ to bound the first term on the RHS of the counterpart of (B.35). $\square$

B.2.2 Proof of Theorem III.8

Proof. Without loss of generality, we assume $\boldsymbol{\mu}_p = \mathbf{0}$, and $m = \infty$. First, we consider the probabilistic upper bound of $|\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega} - \mathbf{I}_{p\times p}|_\infty$. Note that

$$\begin{aligned}
|\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega} - \mathbf{I}_{p\times p}|_\infty &= \max_{1\le i,j\le p} \left| \sum_{l=1}^{p} \left[\frac{1}{n}\sum_{k=1}^{n} X_{ik}X_{lk} - \bar{X}_i\bar{X}_l \right] \omega_{lj} - \mathbb{I}(i=j) \right| \\
&\le \max_{1\le i,j\le p} \left| \bar{X}_i \frac{1}{n}\sum_{k=1}^{n}\sum_{l=1}^{p} X_{lk}\omega_{lj} \right| + \max_{1\le i,j\le p} \left| \frac{1}{n}\sum_{k=1}^{n} X_{ik}\sum_{l=1}^{p} X_{lk}\omega_{lj} - \mathbb{I}(i=j) \right| \\
&= \max_{1\le i,j\le p} \left| \bar{X}_i \bar{\bar{\tilde{X}}}_j \right| + \max_{1\le i,j\le p} \left| \frac{1}{n}\sum_{k=1}^{n} X_{ik}\tilde{X}_{jk} - \mathbb{I}(i=j) \right| \qquad \text{(B.48)}
\end{aligned}$$

with $\tilde{X}_{jk} := \sum_{l=1}^{p} X_{lk}\omega_{lj}$ and $\bar{\bar{\tilde{X}}}_j := n^{-1}\sum_{k=1}^{n} \tilde{X}_{jk}$. Since $\mathrm{cov}(\boldsymbol{\Omega}\boldsymbol{X}_k) = \boldsymbol{\Omega}$ and $\omega_{jl} = \omega_{lj}$, then we have $\mathrm{var}(\tilde{X}_{jk}) = \mathrm{var}\left(\sum_{l=1}^{p} \omega_{jl}X_{lk}\right) = \omega_{jj}$. Besides, $\mathrm{cov}(X_{ik}, \tilde{X}_{jk}) = E\left[X_{ik}\sum_{l=1}^{p} X_{lk}\omega_{lj}\right] = \sum_{l=1}^{p} \sigma_{il}\omega_{lj} = \mathbb{I}(i=j)$. Let $\tilde{Z}_{jk} = \tilde{X}_{jk}\omega_{jj}^{-1/2}$ and $Z_{ik} = X_{ik}\sigma_{ii}^{-1/2}$, then $\tilde{\rho}_{ij} := \mathrm{corr}(Z_{ik}, \tilde{Z}_{jk}) = \sigma_{ii}^{-1/2}\omega_{jj}^{-1/2}\mathbb{I}(i=j)$. From $\tau_0 = \sqrt{f\log(pf)/n} = o(1)$, (B.16) holds when $n > N_1$ with some constant N_1. In the following proof, we assume $n > N_1$. Now we consider

$$\begin{aligned}
P\left[\max_{1\le i,j\le p} \left|\bar{X}_i\bar{\bar{\tilde{X}}}_j\right| \ge u\right] &\le P\left[\max_{1\le i\le p}|\bar{X}_i| \max_{1\le j\le p}|\bar{\bar{\tilde{X}}}_j| \ge u\right] \\
&\le P\left[\max_{1\le i\le p}|\bar{X}_i| \ge u^{1/2}\right] + P\left[\max_{1\le j\le p}|\bar{\bar{\tilde{X}}}_j| \ge u^{1/2}\right]. \qquad \text{(B.49)}
\end{aligned}$$

We first consider the second term on the RHS of the above inequality. Similar to (B.17),

$$P\left[\max_{1\le j\le p}|\bar{\bar{\tilde{X}}}_j| \ge u^{1/2}\right] \le \sum_{j=1}^{p}\sum_{r=1}^{f} P\left[\left|\frac{1}{C_{r,f}}\sum_{k\in A_{r,f}} \tilde{Z}_{j,kf+r}\right| \ge \frac{1}{2}\sqrt{\frac{u}{v_0}}\right]. \qquad \text{(B.50)}$$

Let $\boldsymbol{\Delta}^{jfr}$ be the covariance matrices of $\mathrm{vec}\{\tilde{Z}_{j,kf+r} : k \in A_{r,f}\}$. Since

$$\begin{aligned}
\mathrm{cov}(\tilde{Z}_{j,kf+r}, \tilde{Z}_{j,lf+r}) &= \omega_{jj}^{-1}\mathrm{cov}(\sum_{s=1}^{p} X_{s,kf+r}\omega_{sj}, \sum_{t=1}^{p} X_{t,lf+r}\omega_{tj}) \\
&= \omega_{jj}^{-1}\sum_{s=1}^{p}\sum_{t=1}^{p} \omega_{sj}\omega_{tj}\sqrt{\sigma_{ss}\sigma_{tt}}\rho_{st}^{kf+r,lf+r}, \qquad \text{(B.51)}
\end{aligned}$$

$$\begin{aligned}\|\mathbf{\Delta}^{jfr}-\mathbf{I}_{C_{r,f}\times C_{r,f}}\|_1 &= \max_{l\in A_{r,f}} \sum_{\substack{k\in A_{r,f}:\\ k\neq l}} |\text{cov}(\tilde{Z}_{j,kf+r},\tilde{Z}_{j,lf+r})| \\ &= \max_{l\in A_{r,f}} \sum_{\substack{k\in A_{r,f}:\\ k\neq l}} \left|\omega_{jj}^{-1}\sum_{s=1}^{p}\sum_{t=1}^{p}\omega_{sj}\omega_{tj}\sqrt{\sigma_{ss}\sigma_{tt}}\rho_{st}^{kf+r,lf+r}\right| \\ &\leq \omega_{jj}^{-1}v_0M_p^2g \leq v_0^2M_p^2g \qquad \text{(B.52)}\end{aligned}$$

following from $\omega_{jj}^{-1} = \tilde{\rho}_{jj}^2\sigma_{jj} \leq v_0$. Since $\limsup_{n\to\infty} v_0^2M_p^2g < 1$, there exists constants $c_1 > 0$ and $N_2(c_1) > 0$ such that $v_0^2M_p^2g < c_1 < 1$ when $n > N_2(c_1)$. We now assume $n > \max\{N_1, N_2(c_1)\}$. By (B.52), $\mathbf{\Delta}^{jfr}$ is strictly diagonally dominant and is thus positive definite. From equation (3.3),

$$\text{vec}\{\tilde{Z}_{j,kf+r} : k\in A_{r,f}\} = \tilde{\mathbf{P}}^{jfr}\mathbf{He}, \qquad \text{(B.53)}$$

where $\tilde{\mathbf{P}}^{jfr}$ is a $C_{r,f}\times pn$ matrix with $\omega_{jj}^{-1/2}\omega_{lj}$ in the $\big(k+1, l+(kf+r-1)p\big), k\in A_{r,f}, l = 1,\ldots,p,$ entries and 0 in all other entries. By (B.53), Lemma B.2, (B.52), (B.50), (B.16), and $v_0^2M_p^2g < 1$, we have

$$\begin{aligned}P\left[\max_{1\leq j\leq p}|\bar{\tilde{X}}_j| \geq u^{1/2}\right] &\leq pf\exp(1)\exp\left\{-\frac{c_2C_{r,f}u}{4v_0}\right\} + pf\exp(1)\exp\left\{-\frac{c_2C_{r,f}u}{4v_0^3M_p^2g}\right\} \\ &\leq 2pf\exp(1)\exp\left\{-\frac{c_2nu}{8v_0f}\right\} \qquad \text{(B.54)}\end{aligned}$$

with some constant $c_2 > 0$. For the first term on the RHS of (B.49), we still have (B.21) here because $\max_i\sigma_{ii} \leq v_0$ and $g < v_0^2M_p^2g < 1$. Thus by (B.21), (B.54) and (B.49), we obtain

$$P\left[\max_{1\leq i,j\leq p}\left|\bar{X}_i\bar{\tilde{X}}_j\right| \geq u\right] \leq 4pf\exp(1)\exp\left\{-\frac{c_2nu}{8v_0f}\right\}. \qquad \text{(B.55)}$$

Now considering the second term in (B.48). Similar to (B.22), we have for any $u > 0$,

$$P\left[\max_{1\le i,j\le p}\left|\frac{1}{n}\sum_{k=1}^{n} X_{ik}\tilde{X}_{jk} - \mathbb{I}(i=j)\right| \ge u\right]$$

$$\le \sum_{1\le i,j\le p} P\left[\left|\sum_{k=1}^{n}\left(Z_{ik}+C_g\tilde{Z}_{jk}\right)^2 - n\left(1+C_g^2+2C_g\tilde{\rho}_{ij}\right)\right| \ge \frac{2C_g nu}{\sqrt{\sigma_{ii}\omega_{jj}}}\right]$$

$$+ \sum_{1\le i,j\le p} P\left[\left|\sum_{k=1}^{n}\left(Z_{ik}-C_g\tilde{Z}_{jk}\right)^2 - n\left(1+C_g^2-2C_g\tilde{\rho}_{ij}\right)\right| \ge \frac{2C_g nu}{\sqrt{\sigma_{ii}\omega_{jj}}}\right], \tag{B.56}$$

where constant $C_g = \frac{1-\sqrt{c_1}}{1+\sqrt{c_1}} \in (0,1)$. Without loss of generality, we only consider the second term on the RHS of the above inequality. Similar to (B.23),

$$P\left[\left|\sum_{k=1}^{n}\left(Z_{ik}-C_g\tilde{Z}_{jk}\right)^2 - n\left(1+C_g^2-2C_g\tilde{\rho}_{ij}\right)\right| \ge \frac{2C_g nu}{\sqrt{\sigma_{ii}\omega_{jj}}}\right]$$

$$\le \sum_{r=1}^{f} P\left[\left|\sum_{k\in A_{r,f}}\left(\frac{Z_{i,kf+r}-C_g\tilde{Z}_{j,kf+r}}{\sqrt{1+C_g^2-2C_g\tilde{\rho}_{ij}}}\right)^2 - C_{r,f}\right| \ge \frac{2C_g nu}{(1+C_g^2-2C_g\tilde{\rho}_{ij})fv_0}\right]. \tag{B.57}$$

Let

$$\boldsymbol{Z} = \mathrm{vec}\left\{\frac{Z_{i,kf+r}-C_g\tilde{Z}_{j,kf+r}}{\sqrt{1+C_g^2-2C_g\tilde{\rho}_{ij}}} : k \in A_{r,f}\right\},$$

and $\boldsymbol{\Gamma} := (\gamma_{kl})_{C_{r,f}\times C_{r,f}} = \mathrm{cov}(\boldsymbol{Z})$. We have $\gamma_{kk} = 1$. For $k \ne l$,

$$\gamma_{kl} = \left(1+C_g^2-2C_g\tilde{\rho}_{ij}\right)^{-1}\Big[\mathrm{cov}(Z_{i,kf+r}, Z_{i,lf+r}) - C_g\mathrm{cov}(Z_{i,kf+r}, \tilde{Z}_{j,lf+r})$$
$$- C_g\mathrm{cov}(Z_{i,lf+r}, \tilde{Z}_{j,kf+r}) + C_g^2\mathrm{cov}(\tilde{Z}_{j,kf+r}, \tilde{Z}_{j,lf+r})\Big],$$

$$\mathrm{cov}(Z_{i,kf+r}, Z_{i,lf+r}) = \rho_{ii}^{kf+r,lf+r},$$

and

$$\mathrm{cov}(Z_{i,kf+r}, \tilde{Z}_{j,lf+r}) = \sigma_{ii}^{-1/2}\omega_{jj}^{-1/2}\mathrm{cov}(X_{i,kf+r}, \sum_{s=1}^{p} X_{s,lf+r}\omega_{sj})$$
$$= \sigma_{ii}^{-1/2}\omega_{jj}^{-1/2}\sum_{s=1}^{p}\omega_{sj}\mathrm{cov}(X_{i,kf+r}, X_{s,lf+r}) = \sigma_{ii}^{-1/2}\omega_{jj}^{-1/2}\sum_{s=1}^{p}\omega_{sj}\sqrt{\sigma_{ii}\sigma_{ss}}\rho_{is}^{kf+r,lf+r}.$$

Then together with (B.51), we obtain

$$\gamma_{kl} = \big(1 + C_g^2 - 2C_g\tilde{\rho}_{ij}\big)^{-1}\Big(\rho_{ii}^{kf+r,lf+r} - C_g\sigma_{ii}^{-1/2}\omega_{jj}^{-1/2}\sum_{s=1}^{p}\omega_{sj}\sqrt{\sigma_{ii}\sigma_{ss}}\rho_{is}^{kf+r,lf+r} - C_g\sigma_{ii}^{-1/2}\omega_{jj}^{-1/2}\sum_{s=1}^{p}\omega_{sj}\sqrt{\sigma_{ii}\sigma_{ss}}\rho_{is}^{lf+r,kf+r} + C_g^2\omega_{jj}^{-1}\sum_{s=1}^{p}\sum_{t=1}^{p}\omega_{sj}\omega_{tj}\sqrt{\sigma_{ss}\sigma_{tt}}\rho_{st}^{kf+r,lf+r}\Big).$$

Hence, similar to (B.52), we have

$$\begin{aligned}
\|\mathbf{\Gamma} - \mathbf{I}_{C_{r,f}\times C_{r,f}}\|_1 &= \max_{l\in A_{r,f}}\sum_{\substack{k\in A_{r,f}:\\ k\neq l}}|\gamma_{kl}| \\
&\le (1 + C_g^2 - 2C_g\tilde{\rho}_{ij})^{-1}\Big(1 + 2C_g\sigma_{ii}^{-1/2}\omega_{jj}^{-1/2}\sum_{s=1}^{p}|\omega_{sj}\sqrt{\sigma_{ii}\sigma_{ss}}| \\
&\quad + C_g^2\omega_{jj}^{-1}\sum_{s=1}^{p}\sum_{t=1}^{p}|\omega_{sj}\omega_{tj}\sqrt{\sigma_{ss}\sigma_{tt}}|\Big)\max_{l\in A_{r,f}}\sum_{\substack{k\in A_{r,f}:\\ k\neq l}}\max_{1\le s,t\le p}|\rho_{st}^{kf+r,lf+r}| \\
&\le (1 + C_g^2 - 2C_g\tilde{\rho}_{ij})^{-1}\Bigg(1 + 2C_g\omega_{jj}^{-1/2}|\mathbf{\Sigma}|_\infty^{1/2}\sum_{s=1}^{p}|\omega_{sj}| \\
&\quad + C_g^2\omega_{jj}^{-1}|\mathbf{\Sigma}|_\infty\big(\sum_{s=1}^{p}|\omega_{sj}|\big)^2\Bigg)\max_{1\le b\le n}\sum_{\substack{a\in\{1\le a\le n:\\ |a-b|=mf\\ m=1,\dots,\lfloor n/f\rfloor\}}}\max_{1\le s,t\le p}|\rho_{st}^{ab}| \\
&\le (1 + C_g^2 - 2C_g\tilde{\rho}_{ij})^{-1}(1 + C_g\omega_{jj}^{-1/2}v_0^{1/2}M_p)^2 g \\
&\le (1 + C_g^2 - 2C_g\tilde{\rho}_{ij})^{-1}(1 + C_g v_0 M_p)^2 g \\
&\le (1 + C_g^2 - 2C_g\tilde{\rho}_{ij})^{-1}(1 + C_g)^2 v_0^2 M_p^2 g \qquad \text{(B.58)} \\
&\le \left(\frac{1 + C_g}{1 - C_g}\right)^2 v_0^2 M_p^2 g = v_0^2 M_p^2 g/c_1 < 1,
\end{aligned}$$

and thus $\mathbf{\Gamma} \succ 0$. Hence, $\mathbf{\Gamma} = \mathbf{\Gamma}^{1/2}\mathbf{\Gamma}^{1/2}$ with a symmetric positive definite matrix $\mathbf{\Gamma}^{1/2}$. Let $\boldsymbol{Y} = \mathbf{\Gamma}^{-1/2}\boldsymbol{Z}$. Then by (B.19) and (B.53),

$$\boldsymbol{Y} = \mathbf{A}\boldsymbol{e} \ \text{ with } \ \mathbf{A}\mathbf{A}^T = \text{cov}(\boldsymbol{Y}) = \mathbf{I}_{C_{r,f}\times C_{r,f}} \ \text{ and } \ \mathbf{A} = \frac{\mathbf{\Gamma}^{-1/2}(\mathbf{P}^{ifr} - C_g\tilde{\mathbf{P}}^{jfr})\mathbf{H}}{\sqrt{(1 + C_g^2 - 2C_g\tilde{\rho}_{ij})}}.$$

Similar to (B.26),

$$
\begin{aligned}
&P\left[\left|\sum_{k\in A_{r,f}}\left(\frac{Z_{i,kf+r}-C_g\tilde{Z}_{j,kf+r}}{\sqrt{1+C_g^2-2C_g\tilde{\rho}_{ij}}}\right)^2-C_{r,f}\right|\geq\frac{2C_gnu}{(1+C_g^2-2C_g\tilde{\rho}_{ij})fv_0}\right]\\
&\leq P\left[|\boldsymbol{Y}^T(\boldsymbol{\Gamma}-\mathbf{I}_{C_{r,f}\times C_{r,f}})\boldsymbol{Y}|\geq\frac{C_gnu}{(1+C_g^2-2C_g\tilde{\rho}_{ij})fv_0}\right]\\
&\quad+P\left[|\boldsymbol{Y}^T\boldsymbol{Y}-C_{r,f}|\geq\frac{C_gnu}{(1+C_g^2-2C_g\tilde{\rho}_{ij})fv_0}\right],
\end{aligned}
\tag{B.59}
$$

and we only consider the nontrivial case when $\boldsymbol{\Gamma}\neq\mathbf{I}_{C_{r,f}\times C_{r,f}}$. Similar to (B.27), by Lemma B.1 and (B.58) we have

$$
\begin{aligned}
&P\left[|\boldsymbol{Y}^T(\boldsymbol{\Gamma}-\mathbf{I}_{C_{r,f}\times C_{r,f}})\boldsymbol{Y}|\geq\frac{C_gnu}{(1+C_g^2-2C_g\tilde{\rho}_{ij})fv_0}\right]\\
&\leq 2\exp\left\{-c_3\min\left(\frac{C_g^2nu^2}{2v_0^2(1+C_g)^4f},\frac{C_gnu}{v_0(1+C_g)^2f}\right)\right\},
\end{aligned}
\tag{B.60}
$$

and similar to (B.28),

$$
\begin{aligned}
&P\left[|\boldsymbol{Y}^T\boldsymbol{Y}-C_{r,f}|\geq\frac{C_gnu}{(1+C_g^2-2C_g\tilde{\rho}_{ij})fv_0}\right]\\
&\leq 2\exp\left\{-c_3\min\left(\frac{C_g^2nu^2}{2v_0^2(1+C_g)^4f},\frac{C_gnu}{v_0(1+C_g)^2f}\right)\right\},
\end{aligned}
\tag{B.61}
$$

with some constant $c_3>0$. From (B.60), (B.61), (B.59) and (B.57), we obtain

$$
\begin{aligned}
&P\left[\left|\sum_{k=1}^{n}\left(Z_{ik}-C_g\tilde{Z}_{jk}\right)^2-n\left(1+C_g^2-2C_g\tilde{\rho}_{ij}\right)\right|\geq\frac{2C_gnu}{\sqrt{\sigma_{ii}\omega_{jj}}}\right]\\
&\leq 4f\exp\left\{-c_4\min\left(\frac{nu^2}{f},\frac{nu}{f}\right)\right\},
\end{aligned}
$$

with some constant $c_4>0$. Then by (B.56),

$$
\begin{aligned}
&P\left[\max_{1\leq i,j\leq p}\left|\frac{1}{n}\sum_{k=1}^{n}X_{ik}\tilde{X}_{jk}-\mathbb{I}(i=j)\right|\geq u\right]\\
&\leq 8p^2f\exp\left\{-c_4\min\left(\frac{nu^2}{f},\frac{nu}{f}\right)\right\}.
\end{aligned}
\tag{B.62}
$$

From (B.48), (B.55) and (B.62), we obtain

$$
\begin{aligned}
&P\left[|\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}-\mathbf{I}_{p\times p}|_\infty\geq 2u\right]\\
&\leq 4pf\exp(1)\exp\left\{-\frac{c_2nu}{8v_0f}\right\}+8p^2f\exp\left\{-c_4\min\left(\frac{nu^2}{f},\frac{nu}{f}\right)\right\}.
\end{aligned}
$$

By Lemma B.3, for any constant $C' > 0$, there exists a constant $M_1 > 0$ such that with $u = M_1\tau_0/4$, $P\left[|\hat{\mathbf{\Sigma}}\mathbf{\Omega} - \mathbf{I}_{p\times p}|_\infty \geq 2u\right] = O(p^{-C'})$. Thus, for any constant $M \geq M_1$,

$$P\left[|\hat{\mathbf{\Sigma}}\mathbf{\Omega} - \mathbf{I}_{p\times p}|_\infty \geq M\tau_0/2\right] \leq P\left[|\hat{\mathbf{\Sigma}}\mathbf{\Omega} - \mathbf{I}_{p\times p}|_\infty \geq M_1\tau_0/2\right] = O(p^{-C'}).$$

Let $0 \leq \varepsilon \leq M\tau_0/(2v_0)$. Then

$$\begin{aligned} P\left[|\tilde{\mathbf{\Sigma}}_\varepsilon\mathbf{\Omega} - \mathbf{I}_{p\times p}|_\infty \geq M\tau_0\right] &\leq P\left[|\hat{\mathbf{\Sigma}}\mathbf{\Omega} - \mathbf{I}_{p\times p}|_\infty + |\varepsilon\mathbf{\Omega}|_\infty \geq M\tau_0\right] \\ = P\left[|\hat{\mathbf{\Sigma}}\mathbf{\Omega} - \mathbf{I}_{p\times p}|_\infty \geq M\tau_0 - |\varepsilon\mathbf{\Omega}|_\infty\right] &\leq P\left[|\hat{\mathbf{\Sigma}}\mathbf{\Omega} - \mathbf{I}_{p\times p}|_\infty \geq M\tau_0/2\right] = O(p^{-C'}). \end{aligned}$$

Let $\lambda_1 = M\tau_0$. Then with probability $1-O(p^{-C'})$, $|\tilde{\mathbf{\Sigma}}_\varepsilon\mathbf{\Omega}-\mathbf{I}_{p\times p}|_\infty \leq \lambda_1$. By the definition of $\hat{\mathbf{\Omega}}_\varepsilon$ and the equivalence between (3.17) and (3.18), on the event $\{|\tilde{\mathbf{\Sigma}}_\varepsilon\mathbf{\Omega}-\mathbf{I}_{p\times p}|_\infty \leq \lambda_1\}$, we have $\|\hat{\mathbf{\Omega}}_\varepsilon\|_1 \leq \|\hat{\mathbf{\Omega}}^\star_\varepsilon\|_1 \leq \|\mathbf{\Omega}\|_1$ and $|\hat{\boldsymbol{\omega}}_{j\varepsilon}|_1 \leq |\hat{\boldsymbol{\omega}}^\star_{j\varepsilon}|_1 \leq |\boldsymbol{\omega}_j|_1$ for $1 \leq j \leq p$, where $\hat{\boldsymbol{\omega}}_{j\varepsilon}$, $\hat{\boldsymbol{\omega}}^\star_{j\varepsilon}$ and $\boldsymbol{\omega}_j$ are j-th columns of $\hat{\mathbf{\Omega}}_\varepsilon$, $\hat{\mathbf{\Omega}}^\star_\varepsilon$ and $\mathbf{\Omega}$ respectively. Thus, on the event $\{|\tilde{\mathbf{\Sigma}}_\varepsilon\mathbf{\Omega} - \mathbf{I}_{p\times p}|_\infty \leq \lambda_1\}$, we have

$$\begin{aligned} |\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}|_\infty \leq |\hat{\mathbf{\Omega}}^\star_\varepsilon - \mathbf{\Omega}|_\infty &= |(\mathbf{\Omega}\tilde{\mathbf{\Sigma}}_\varepsilon - \mathbf{I}_{p\times p})\hat{\mathbf{\Omega}}^\star_\varepsilon + \mathbf{\Omega}(\mathbf{I}_{p\times p} - \tilde{\mathbf{\Sigma}}_\varepsilon\hat{\mathbf{\Omega}}^\star_\varepsilon)|_\infty \\ &\leq \|\hat{\mathbf{\Omega}}^\star_\varepsilon\|_1|\tilde{\mathbf{\Sigma}}_\varepsilon\mathbf{\Omega} - \mathbf{I}_{p\times p}|_\infty + \|\mathbf{\Omega}\|_1|\tilde{\mathbf{\Sigma}}_\varepsilon\hat{\mathbf{\Omega}}^\star_\varepsilon - \mathbf{I}_{p\times p}|_\infty \\ &\leq \lambda_1\|\hat{\mathbf{\Omega}}^\star_\varepsilon\|_1 + \lambda_1\|\mathbf{\Omega}\|_1 \leq 2\lambda_1 M_p, \end{aligned} \tag{B.63}$$

which follows from the inequality $|\mathbf{AB}|_\infty \leq |\mathbf{A}|_\infty\|\mathbf{B}\|_1$ for matrices $\mathbf{A}, \mathbf{B}$ of appropriate sizes, and moreover,

$$\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_2 \leq \|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_1 \leq 12c_p|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}|_\infty^{1-q} \leq 12c_p(2M\tau_0M_p)^{1-q},$$

following from Lemma 7.1 of Cai et al. (2016). Inequality (3.37) follows from the inequality $\|\mathbf{M}\|_F^2 \leq p\|\mathbf{M}\|_1|\mathbf{M}|_\infty$ for any $p \times p$ matrix $\mathbf{M}$.

For the sparsistency and sign-consistency, the proof follows the similar lines of the proof of Theorem 2 in Rothman et al. (2009) by replacing their equation (A.4) with (B.63). Details are hence omitted.

For the convergence in mean square, we additionally assume $p \geq n^c$ with some constant $c > 0$, and for any constant $C > 0$, we let $\min\left\{p^{-C}, M\tau_0/(2v_0)\right\} \leq \varepsilon \leq M\tau_0/(2v_0)$. Since $\varphi_{\min}(\tilde{\mathbf{\Sigma}}_\varepsilon) = \varphi_{\min}(\hat{\mathbf{\Sigma}} + \varepsilon \mathbf{I}_{p\times p}) \geq \varepsilon$, we have $|\tilde{\mathbf{\Sigma}}_\varepsilon^{-1}|_\infty \leq 1/\varphi_{\min}(\tilde{\mathbf{\Sigma}}_\varepsilon) \leq \varepsilon^{-1}$ and $\|\tilde{\mathbf{\Sigma}}_\varepsilon^{-1}\|_1 \leq p|\tilde{\mathbf{\Sigma}}_\varepsilon^{-1}|_\infty \leq p\varepsilon^{-1}$. Then by the definition of $\hat{\mathbf{\Omega}}_\varepsilon$ and the equivalence between (3.17) and (3.18), $\|\hat{\mathbf{\Omega}}_\varepsilon\|_1 \leq \|\hat{\mathbf{\Omega}}_\varepsilon^\star\|_1 \leq \|\tilde{\mathbf{\Sigma}}_\varepsilon^{-1}\|_1 \leq p\varepsilon^{-1}$. In addition with $\|\mathbf{\Omega}\|_1 = \max_{1\leq j\leq p}\sum_{i=1}^p |\omega_{ij}|^q|\omega_{ij}|^{1-q} \leq c_p M_p^{1-q}$, we obtain $\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_1 \leq \|\hat{\mathbf{\Omega}}_\varepsilon\|_1 + \|\mathbf{\Omega}\|_1 \leq p\varepsilon^{-1} + c_p M_p^{1-q}$. Now,

$$\begin{aligned}
E\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_2^2 &\leq E\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_1^2 \\
&= E\left[\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_1^2 \mathbb{I}\left(\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_1 > 12c_p(2M\tau_0M_p)^{1-q}\right)\right] \\
&\quad + E\left[\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_1^2 \mathbb{I}\left(\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_1 \leq 12c_p(2M\tau_0M_p)^{1-q}\right)\right] \\
&\leq \left(E\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_1^4\right)^{1/2}\left(P\left[\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_1 > 12c_p(2M\tau_0M_p)^{1-q}\right]\right)^{1/2} \\
&\quad + \left(12c_p(2M\tau_0M_p)^{1-q}\right)^2 \\
&\leq (p\varepsilon^{-1} + c_pM_p^{1-q})^2 O(p^{-C'/2}) + \left(12c_p(2M\tau_0M_p)^{1-q}\right)^2. \qquad \text{(B.64)}
\end{aligned}$$

Let $M \geq \max(v_0, M_1)$. Since $\tau_0 = \sqrt{f\log(pf)/n} \geq p^{-1/(2c)}$, then we have $\varepsilon^{-1} \leq \max\left(p^C, 2v_0M^{-1}p^{1/(2c)}\right) \leq \max\left(p^C, 2p^{1/(2c)}\right)$. When $C' \geq 2\max(2+2C, 2+1/c)+2/c$, by $\min(c_p, M_p) > 1$ and $M \geq v_0 > 1$, we have

$$\begin{aligned}
(p\varepsilon^{-1} + c_pM_p^{1-q})^2 O(p^{-C'/2}) &\leq O(p^{\max(2+2C,2+1/c)-C'/2}) + O(c_p^2M_p^{2-2q}p^{-C'/2}) \\
&= O(c_p^2M_p^{2-2q}p^{-1/c}) = O\left(\left(12c_p(2M\tau_0M_p)^{1-q}\right)^2\right),
\end{aligned}$$

and thus by (B.64), we have $E\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_2^2 = O\left(c_p^2(\tau_0M_p)^{2-2q}\right)$.

Since $|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}|_\infty \leq \|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_1 \leq \|\hat{\mathbf{\Omega}}_\varepsilon\|_1 + \|\mathbf{\Omega}\|_1 \leq p\varepsilon^{-1} + M_p$ and $p^{-1}\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_F^2 \leq \|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_1|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}|_\infty \leq (p\varepsilon^{-1} + c_pM_p^{1-q})(p\varepsilon^{-1} + M_p)$, similarly to (B.64), we have $E|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}|_\infty^2 \leq (p\varepsilon^{-1} + M_p)^2 O(p^{-C'/2}) + (2M\tau_0M_p)^2$ and $p^{-1}E\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_F^2 \leq (p\varepsilon^{-1} + c_pM_p^{1-q})(p\varepsilon^{-1} + M_p)O(p^{-C'/2}) + 12c_p(2M\tau_0M_p)^{2-q}$. Let C' be sufficiently large, then

$E|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}|_\infty^2 = O\left((\tau_0 M_p)^2\right)$ and $p^{-1}E\|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}\|_F^2 = O\left(c_p(\tau_0 M_p)^{2-q}\right)$. □

B.2.3 Proof of Theorem III.9

Proof. Since $\min_i \sigma_{ii} - \varphi_{\min}(\mathbf{\Sigma}) = \min_i \boldsymbol{e}_i^T(\mathbf{\Sigma} - \varphi_{\min}(\mathbf{\Sigma})\mathbf{I}_{p\times p})\boldsymbol{e}_i \geq 0$ and $\max_i \sigma_{ii} \leq |\mathbf{\Sigma}|_\infty \leq \|\mathbf{\Sigma}\|_2 = \varphi_{\max}(\mathbf{\Sigma})$, we have

$$v_0^{-1} \leq \min_i \sigma_{ii} \leq \max_i \sigma_{ii} \leq v_0. \tag{B.65}$$

Thus,

$$v_0^{-1/2} \leq \|\mathbf{W}\|_2, \|\mathbf{W}^{-1}\|_2 \leq v_0^{1/2}. \tag{B.66}$$

$$\|\mathbf{K}\|_2 \leq \|\mathbf{W}\|_2\|\mathbf{\Omega}\|_2\|\mathbf{W}\|_2 \leq v_0^2, \tag{B.67}$$

$$\|\mathbf{R}\|_2 \leq \|\mathbf{W}^{-1}\|_2\|\mathbf{\Sigma}\|_2\|\mathbf{W}^{-1}\|_2 \leq v_0^2,$$

and

$$v_0^{-2} \leq \|\mathbf{K}\|_2^{-1} = \varphi_{\min}(\mathbf{R}) \leq \varphi_{\max}(\mathbf{R}) = \|\mathbf{R}\|_2 \leq v_0^2. \tag{B.68}$$

Under $\max_i \sigma_{ii} \leq v_0, \limsup_{n\to\infty} g(n,p) < 1$ and $\tau_0 = o(1)$, we can obtain (B.31), (B.42) and (B.44), i.e., for any constant $C' > 0$, there exists a constant $C_1 > 0$ such that with probability $1 - O(p^{-C'})$,

$$\max_{1\leq i,j\leq p} |\hat{\sigma}_{ij} - \sigma_{ij}| \leq C_1\tau_0, \tag{B.69}$$

$$\max_{1\leq i,j\leq p} |\hat{\rho}_{ij} - \rho_{ij}| \leq C_1\tau_0, \tag{B.70}$$

and

$$\max_{1\leq i,j\leq p} \left|\frac{\hat{\sigma}_{ij} - \sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right| \leq C_1\tau_0. \tag{B.71}$$

From (B.69) and (B.65), we obtain that, with probability $1-O(p^{-C'})$, $\max_i \hat{\sigma}_{ii}^{-1/2} \le 2v_0^{1/2}$. Letting $i=j$ in (B.71), we have that, with probability $1-O(p^{-C'})$,

$$\begin{aligned} o(1) = 2C_1\tau_0 v_0^{1/2} &\ge C_1\tau_0 \max_{1\le i\le p} \hat{\sigma}_{ii}^{-\frac{1}{2}} \ge \max_{1\le i\le p} \left|\frac{\hat{\sigma}_{ii}-\sigma_{ii}}{\sigma_{ii}}\right| \max_{1\le i\le p} \hat{\sigma}_{ii}^{-\frac{1}{2}} \\ &\ge \max_{1\le i\le p} \left|\sqrt{\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}}-1\right| \left|\sqrt{\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}}+1\right| \hat{\sigma}_{ii}^{-\frac{1}{2}} \ge \max_{1\le i\le p} \left|\sqrt{\frac{\hat{\sigma}_{ii}}{\sigma_{ii}}}-1\right| \hat{\sigma}_{ii}^{-\frac{1}{2}} \\ &= \max_{1\le i\le p} |\sigma_{ii}^{-\frac{1}{2}} - \hat{\sigma}_{ii}^{-\frac{1}{2}}| = \|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_2, \end{aligned} \tag{B.72}$$

and then by (B.66),

$$\|\hat{\mathbf{W}}^{-1}\|_2 \le \|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_2 + \|\mathbf{W}^{-1}\|_2 = o(1) + v_0^{1/2}. \tag{B.73}$$

Now recall the assumption that $\tau_0 = o(1/\sqrt{1+s_p})$. Following similar lines of the proof of Theorem 1 in Rothman et al. (2008) by replacing their line 10 on page 500 by $r_n = \tau_0\sqrt{s_p} \to 0$, replacing their line 5 on page 501 by (B.70), replacing their inequality (14) by $\mathrm{II} = 0$, replacing their equation (15) by $\lambda_2 = C_1\tau_0/\varepsilon$ with a sufficiently small constant $\varepsilon > 0$, and replacing the last line on their page 501 by $|\Delta_S^-|_1 \le \sqrt{s_p}\|\Delta^-\|_F$, as well as using (B.68) to establish the counterpart of their inequality (18) for $\mathbf{K}$, we can obtain (3.38).

From the proof of Theorem 2 in Rothman et al. (2008), we have

$$\begin{aligned} \|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_2 \le{}& \|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_2(\|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_2^2 + \|\hat{\mathbf{W}}^{-1}\|_2\|\mathbf{W}^{-1}\|_2) \\ &+ \|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_2(\|\hat{\mathbf{K}}_{\lambda_2}\|_2\|\mathbf{W}^{-1}\|_2 + \|\mathbf{K}\|_2\|\hat{\mathbf{W}}^{-1}\|_2) \\ \le{}& \|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_F(\|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_2^2 + \|\hat{\mathbf{W}}^{-1}\|_2\|\mathbf{W}^{-1}\|_2) \\ &+ \|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_2\left[(\|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_F + \|\mathbf{K}\|_2)\|\mathbf{W}^{-1}\|_2 + \|\mathbf{K}\|_2\|\hat{\mathbf{W}}^{-1}\|_2\right]. \end{aligned} \tag{B.74}$$

Plugging (3.38), (B.72), (B.73), (B.66) and (B.67) into (B.74) yields (3.39).

We can obtain (3.40) similarly from

$$
\begin{aligned}
&\|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_F \\
&= \|\hat{\mathbf{W}}^{-1}\hat{\mathbf{K}}_{\lambda_2}\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\mathbf{K}\mathbf{W}^{-1}\|_F \\
&= \|(\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1})(\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K})(\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}) + \mathbf{W}^{-1}\hat{\mathbf{K}}_{\lambda_2}(\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}) \\
&\qquad + (\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1})\mathbf{K}\hat{\mathbf{W}}^{-1} + \hat{\mathbf{W}}^{-1}(\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K})\mathbf{W}^{-1}\|_F \\
&\leq \|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_F(\|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_2^2 + \|\hat{\mathbf{W}}^{-1}\|_2\|\mathbf{W}^{-1}\|_2) \\
&\quad + \|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_F(\|\hat{\mathbf{K}}_{\lambda_2}\|_2\|\mathbf{W}^{-1}\|_2 + \|\mathbf{K}\|_2\|\hat{\mathbf{W}}^{-1}\|_2) \\
&\leq \|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_F(\|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_2^2 + \|\hat{\mathbf{W}}^{-1}\|_2\|\mathbf{W}^{-1}\|_2) \\
&\quad + \sqrt{p}\|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_2\left[(\|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_F + \|\mathbf{K}\|_2)\|\mathbf{W}^{-1}\|_2 + \|\mathbf{K}\|_2\|\hat{\mathbf{W}}^{-1}\|_2\right],
\end{aligned}
$$

where $\|\mathbf{BA}\|_F = \|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_2\|\mathbf{B}\|_F$ for symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ (see Lemma 1 in Lam and Fan, 2009).

If additionally assuming $\|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_2 = O_P(\eta)$ with $\eta = O(\tau_0)$, the proof of the sparsistency property is similar to the proof of Theorem 2 in Lam and Fan (2009) by using the inequality (B.70) and (B.68). Details are hence omitted. Note that our $\eta^2 = \eta_n$ in their notation. Also note that $\hat{\mathbf{K}}_{\lambda_2}$ and $\mathbf{K}$ have the same the sparsity structures as $\hat{\mathbf{\Omega}}_{\lambda_2}$ and $\mathbf{\Omega}$, respectively.

Now, we consider the properties of $\hat{\mathbf{\Omega}}_{\lambda_2}$ under the irrepresentability condition given in (3.26). We replace the original conditions about λ_2 and τ_0 by $\lambda_2 = 8M\tau_0/\beta \leq [6(1 + \beta/8)d\max\{\kappa_{\mathbf{R}}\kappa_{\mathbf{\Gamma}}, \kappa_{\mathbf{R}}^3\kappa_{\mathbf{\Gamma}}^2\}]^{-1}$ and $\tau_0 = o(\min\{1, [(1 + 8/\beta)\kappa_{\mathbf{\Gamma}}]^{-1}\})$. First, we need to show $|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}|_\infty = o_P(1)$, which is similar to the proof of Theorem 1 in Ravikumar et al. (2011). We follow some of their notation for convenience. In the proof, their Θ and Σ are now replaced by our $\mathbf{K}$ and $\mathbf{R}$ respectively. But we keep their W that is our $\hat{\mathbf{R}} - \mathbf{R}$, which should not be confused with our $\mathbf{W}$ in bold. From (B.70), for any constant $\tau > 2$ (note that here we use the notation τ given in Ravikumar et al. (2011) rather than the one defined as $\tau = M\tau_0$ for the thresholding parameter of covariance matrix estimation), there exist constants M_1 and N_1 such that when $M \geq M_1$ and $n > N_1$, we have

$$P(|W|_\infty \leq M\tau_0) \geq P(|W|_\infty \leq M_1\tau_0) \geq 1 - 1/p^{\tau-2}, \tag{B.75}$$

thus we can set their $\bar{\delta}_f(n, p^\tau) = M\tau_0$ and their $1/v_* = \infty$. Then, $\lambda_2 = 8M\tau_0/\beta = 8\bar{\delta}_f(n, p^\tau)/\beta$. From $\lambda_2 \leq [6(1+\beta/8)d \max\{\kappa_{\mathbf{R}}\kappa_{\mathbf{\Gamma}}, \kappa_{\mathbf{R}}^3\kappa_{\mathbf{\Gamma}}^2\}]^{-1}$, we have

$$\bar{\delta}_f(n, p^\tau) \leq [6(1+8/\beta)d \max\{\kappa_{\mathbf{R}}\kappa_{\mathbf{\Gamma}}, \kappa_{\mathbf{R}}^3\kappa_{\mathbf{\Gamma}}^2\}]^{-1}. \tag{B.76}$$

Then following the proof of their Theorem 1 by using (B.75) instead of their Lemma 8, and (B.76) instead of their (15) and (29), with probability $1 - O(p^{2-\tau})$ we have

$$|\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}|_\infty \leq 2(1+8/\beta)\kappa_{\mathbf{\Gamma}}\bar{\delta}_f(n, p^\tau) = 2(1+8/\beta)\kappa_{\mathbf{\Gamma}}M\tau_0 = o(1), \tag{B.77}$$

and all entries of $\hat{\mathbf{K}}_{\lambda_2}$ in S^c are zero. By $|\mathbf{BA}|_\infty = |\mathbf{AB}|_\infty \leq |\mathbf{A}|_\infty \|\mathbf{B}\|_1$ for symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, we have

$$\begin{aligned}
|\hat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}|_\infty &= |\hat{\mathbf{W}}^{-1}\hat{\mathbf{K}}_{\lambda_2}\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\mathbf{K}\mathbf{W}^{-1}|_\infty \\
&= |(\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1})(\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K})(\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}) + \mathbf{W}^{-1}\hat{\mathbf{K}}_{\lambda_2}(\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}) \\
&\qquad + (\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1})\mathbf{K}\hat{\mathbf{W}}^{-1} + \hat{\mathbf{W}}^{-1}(\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K})\mathbf{W}^{-1}|_\infty \\
&\leq |\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}|_\infty \|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_1^2 + |\hat{\mathbf{K}}_{\lambda_2}|_\infty \|\mathbf{W}^{-1}\|_1 \|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_1 \\
&\qquad + |\mathbf{K}|_\infty \|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_1 \|\hat{\mathbf{W}}^{-1}\|_1 + |\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}|_\infty \|\hat{\mathbf{W}}^{-1}\|_1 \|\mathbf{W}^{-1}\|_1 \\
&= |\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}|_\infty \|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_2^2 + |\hat{\mathbf{K}}_{\lambda_2}|_\infty \|\mathbf{W}^{-1}\|_2 \|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_2 \\
&\qquad + \|\mathbf{K}\|_2 \|\hat{\mathbf{W}}^{-1} - \mathbf{W}^{-1}\|_2 \|\hat{\mathbf{W}}^{-1}\|_2 + |\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}|_\infty \|\hat{\mathbf{W}}^{-1}\|_2 \|\mathbf{W}^{-1}\|_2. \qquad \text{(B.78)}
\end{aligned}$$

By inequalities (B.67) and (B.77), with probability $1 - O(p^{2-\tau})$ we have

$$|\hat{\mathbf{K}}_{\lambda_2}|_\infty \leq |\mathbf{K}|_\infty + |\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}|_\infty \leq \|\mathbf{K}\|_2 + |\hat{\mathbf{K}}_{\lambda_2} - \mathbf{K}|_\infty \leq v_0^2 + o(1). \tag{B.79}$$

Plugging (B.77), (B.72), (B.79), (B.66), (B.67), (B.73) into (B.78) and letting $M \geq \max\{M_1, 10C_1v_0^2\}$ yields that, with probability $1 - O(p^{2-\tau})$,

$$\begin{aligned}
|\hat{\boldsymbol{\Omega}}_{\lambda_2} - \boldsymbol{\Omega}|_\infty &\leq 2(1 + 8/\beta)\kappa_{\boldsymbol{\Gamma}} M\tau_0 o(1) + \left(v_0^2 + o(1)\right) v_0^{1/2} 2C_1\tau_0 v_0^{1/2} \\
&\quad + v_0^2 2C_1\tau_0 v_0^{1/2}\left(o(1) + v_0^{1/2}\right) + 2(1 + 8/\beta)\kappa_{\boldsymbol{\Gamma}} M\tau_0 \left(o(1) + v_0^{1/2}\right) v_0^{1/2} \\
&\leq 5C_1\tau_0 v_0^3 + 2.5(1 + 8/\beta)\kappa_{\boldsymbol{\Gamma}} M\tau_0 v_0 \leq \left(0.5 + 2.5(1 + 8/\beta)\kappa_{\boldsymbol{\Gamma}}\right) M\tau_0 v_0 = r, \qquad \text{(B.80)}
\end{aligned}$$

$$\begin{aligned}
\|\hat{\boldsymbol{\Omega}}_{\lambda_2} - \boldsymbol{\Omega}\|_2 \leq \min\{\|\hat{\boldsymbol{\Omega}}_{\lambda_2} - \boldsymbol{\Omega}\|_1, \|\hat{\boldsymbol{\Omega}}_{\lambda_2} - \boldsymbol{\Omega}\|_F\} &\leq \min\{d, \sqrt{p + s_p}\}|\hat{\boldsymbol{\Omega}}_{\lambda_2} - \boldsymbol{\Omega}|_\infty \\
&\leq \min\{d, \sqrt{p + s_p}\}r,
\end{aligned}$$

and

$$\begin{aligned}
p^{-\frac{1}{2}}\|\hat{\boldsymbol{\Omega}}_{\lambda_2} - \boldsymbol{\Omega}\|_F &\leq \min\{\|\hat{\boldsymbol{\Omega}}_{\lambda_2} - \boldsymbol{\Omega}\|_2, p^{-\frac{1}{2}}\sqrt{p + s_p}|\hat{\boldsymbol{\Omega}}_{\lambda_2} - \boldsymbol{\Omega}|_\infty\} \\
&\leq r \min\left\{d, \sqrt{1 + s_p/p}\right\} = r\sqrt{1 + s_p/p},
\end{aligned}$$

where the last equality follows from $\sqrt{1 + s_p/p} \leq \sqrt{1 + (d-1)p/p} = \sqrt{d} \leq d$. For any $(i, j) \in S$, by (B.80) and $|\omega_{ij}| > r$, $\hat{\omega}_{ij\lambda_2}$ cannot differ enough from the nonzero ω_{ij} to change sign. Since $\hat{\boldsymbol{\Omega}}_{\lambda_2}$ has the same sparsity as $\hat{\mathbf{K}}_{\lambda_2}$ and we have shown that, with probability $1 - O(p^{2-\tau})$, all entries of $\hat{\mathbf{K}}_{\lambda_2}$ in S^c are zero, then $\hat{\boldsymbol{\Omega}}_{\lambda_2}$ also has this sparsistency result. □

B.3 Candidate Values for Tuning Parameters

In this section, we introduce the method selecting candidate values for the tuning parameter of each considered estimating approach. We use η as the general notation of considered tuning parameters such that $\eta = \tau$ for generalized thresholding, $\eta = \lambda_1$ for CLIME, and $\eta = \lambda_2$ for SPICE. The ordered candidate values $\eta_1, \ldots, \eta_N$ of η are chosen from a logarithmic spaced grid. Specifically, $\log\eta_1, \ldots, \log\eta_N$ are equally spaced values with $\eta_1 = r\eta_N$ and a ratio number $r \in (0, 1)$. In numerical examples, we use $N = 50$ and $r = 0.01$.

For the generalized trhesholding estimation of correlation matrix, we let η_N be the largest absolute value in the off-diagonal of the sample correlation matrix so that the thresholding estimator with η_N is a diagonal matrix.

For CLIME, we use the same η_N generated by the R package `flare` (version 1.5.0; see the function `sugm`) based on the following formula

$$\eta_N = \mathbb{I}(\eta^* \neq 0)\eta^* + I(\eta^* = 0)\eta^{**}$$

with

$$\begin{aligned}
\eta^* &= \min\left\{\max_{1\le i,j\le p} s_{ij}, -\min_{1\le i,j\le p} s_{ij}\right\} \\
\eta^{**} &= \max\left\{\max_{1\le i,j\le p} s_{ij}, -\min_{1\le i,j\le p} s_{ij}\right\} \\
S &:= (s_{ij})_{p\times p} = \hat{\mathbf{\Sigma}} - \text{diag}\{\hat{\sigma}_{11}, \ldots, \hat{\sigma}_{pp}\}.
\end{aligned}$$

For SPICE, we generate its η_N using the same approach implemented in the R package `huge` (version 1.2.7; see the function `huge.glasso`; Zhao et al., 2012) for GLasso. Thus η_N is the largest absolute value in the off-diagonal of the sample correlation matrix. Note that SPICE is a slight modification of GLasso.

B.4 Additional Results of the rfMRI Data Analysis

The top 10 hubs for marginal connectivity and the top 10 hubs for direct connectivity are listed in the following two tables. The coordinates of the center of each hub is given in the Montreal Neurological Institute (MNI) 152 space. The hubs with MNI coordinates listed in bold numbers are spatially close to those found in Buckner et al. (2009) and Cole et al. (2010) from studies with multiple subjects. The hub illustrated in Subsection 3.5.3 is ranked No. 1 in degree of marginal connectivity and No. 4 in degree of direct connectivity.

Table B.1: Top 10 hubs for marginal connectivity found by hard thresholding

Rank	Location	MNI coordinates	Degree	Direct rank	Direct degree
1	Inferior parietal	**48, -72, 24**	164	4	79
2	Supramarginal	-60, -36, 36	151	3	82
3	Superior frontal	**0, 48, 36**	150	6	73
4	Medial orbitofrontal	**0, 60, -12**	140	20	53
5	Inferior parietal	**-36, -72, 36**	137	15	61
6	Supramarginal	**60, -48, 36**	131	1	85
7	Precuneus	0, -72, 48	128	16	58
8	Precuneus	0, -72, 36	125	10	64
9	Rostral middle frontal	**-48, 12, 36**	121	5	74
10	Inferior parietal	**-48, -60, 24**	109	37	48

Table B.2: Top 10 hubs for direct connectivity found by CLIME

Rank	Location	MNI coordinates	Degree	Marginal rank	Marginal degree
1	Inferior parietal	60, -48, 36	85	6	131
2	Precentral	-48, 0, 48	82	18	98
3	Supramarginal	-60, -36, 36	82	2	151
4	Inferior parietal	48, -72, 24	79	1	164
5	Rostral middle frontal	-48, 12, 36	74	9	121
6	Superior frontal	0, 48, 36	73	3	150
7	Caudal middle frontal	48, 12, 48	68	29	87
8	Middle temporal	60, -60, 12	66	19	96
9	Precuneus	0, -72, 24	65	14	101
10	Precuneus	0, -72, 36	64	8	125

APPENDIX C

Supplementary Materials for Chapter IV

In this appendix, we prove Theorems IV.4 and IV.5 with the weighted sample covariance matrix as the initial estimator. As special cases of these two theorems, all the theorems given in Subsection 4.3.1 using the sample covariance matrix can be immediately obtained by letting $f_1 = \cdots = f_L = 1$.

Before proceeding to the proofs, we introduce a technical lemma.

Lemma C.1. *Let $\boldsymbol{e} = (e_1, e_2, \dots)^T$ be an infinite-dimensional random vector with independent standard sub-Gaussian components, each with the same parameter $K \geq 1$ defined in* (3.2)*. Let $\boldsymbol{X} = \mathbf{A}\boldsymbol{e}$ and $\boldsymbol{Y} = \mathbf{B}\boldsymbol{e}$ be two well-defined random vector with length d in the sense of entrywise almost-sure convergence and mean-square convergence. Then for $t > 0$, there exists a constant $c > 0$ only dependent on K such that*

$$\begin{aligned} & P\left[\left|\boldsymbol{X}^T\boldsymbol{Y} - E(\boldsymbol{X}^T\boldsymbol{Y})\right| \geq t\right] \\ \text{(C.1)} \qquad & \leq 2\exp\left\{-c\min\left(\frac{t^2}{\|\mathbf{A}\mathbf{A}^T\|_F\|\mathbf{B}\mathbf{B}^T\|_F}, \frac{t}{\sqrt{\|\mathbf{A}\mathbf{A}^T\|_2\|\mathbf{B}\mathbf{B}^T\|_2}}\right)\right\}, \end{aligned}$$

and for a d-dimensional vector $\boldsymbol{b}$,

$$P\left[|\boldsymbol{b}^T\boldsymbol{X}| \geq t\right] \leq \exp(1)\exp\left\{-\frac{ct^2}{\|\boldsymbol{b}\|_F^2\|\mathbf{A}\mathbf{A}^T\|_2}\right\}, \tag{C.2}$$

where the right hand sides of the above inequalities are zero if $\mathbf{A}^T\mathbf{B}$ and $\boldsymbol{b}^T\mathbf{A}$ are zero, respectively.

Proof. Consider the nontrivial case when both $\mathbf{A}^T\mathbf{B}$ and $\boldsymbol{b}^T\mathbf{A}$ are not zero. Let $\mathbf{A} = (a_{ij})_{d\times\infty}$ and $\mathbf{B} = (b_{ij})_{d\times\infty}$. Let $\mathbf{A}_m = (a_{ij})_{d\times m}$ and $\mathbf{B}_m = (b_{ij})_{d\times m}$ consist of the first m columns of $\mathbf{A}$ and $\mathbf{B}$ respectively, $\boldsymbol{e}_m = (e_1, e_2, ..., e_m)^T$ consist of the first m elements of $\boldsymbol{e}$, $\boldsymbol{X}_m = (X_1^m, ..., X_d^m)^T = \mathbf{A}_m\boldsymbol{e}_m$, and $\boldsymbol{Y}_m = (Y_1^m, ..., Y_d^m)^T = \mathbf{B}_m\boldsymbol{e}_m$. By the entrywise almost-sure convergence and mean-square convergence, for each i, when $m \to \infty$, we have $X_i^m = \sum_{j=1}^m a_{ij}e_j \xrightarrow{P} X_i = \sum_{j=1}^\infty a_{ij}e_j$, $Y_i^m = \sum_{j=1}^m b_{ij}e_j \xrightarrow{P} Y_i = \sum_{j=1}^\infty b_{ij}e_j$, $\sum_{j=1}^\infty a_{ij}^2 < \infty$ and $\sum_{j=1}^\infty b_{ij}^2 < \infty$. Thus, for any positive $d, \varepsilon_1, \varepsilon_2$ and δ, there exists a number N such that for any $m > N$, we have

$$P\left[|\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}_m^T\boldsymbol{Y}_m| \geq \varepsilon_1\right] \leq \delta, \tag{C.3}$$

$$P\left[|\boldsymbol{b}^T\boldsymbol{X} - \boldsymbol{b}^T\boldsymbol{X}_m| \geq \varepsilon_1\right] \leq \delta, \tag{C.4}$$

and for each $1 \leq i, j \leq d$,

$$\left|\sum_{k=1}^m a_{ik}b_{ik} - \sum_{k=1}^\infty a_{ik}b_{ik}\right| \leq \varepsilon_2/d, \tag{C.5}$$

$$\left|\sum_{k=1}^m a_{ik}a_{jk} - \sum_{k=1}^\infty a_{ik}a_{jk}\right| \leq \delta/d, \tag{C.6}$$

$$\left|\sum_{k=1}^m b_{ik}b_{jk} - \sum_{k=1}^\infty b_{ik}b_{jk}\right| \leq \delta/d. \tag{C.7}$$

The convergence of $\sum_{k=1}^m a_{ik}b_{ik}$ given in (C.5) holds because

$$\sum_{k=1}^\infty |a_{ik}b_{ik}| \leq \sqrt{\sum_{k=1}^\infty |a_{ik}|^2 \sum_{k=1}^\infty |b_{ik}|^2} < \infty.$$

By the similar argument, we obtain (C.6) and (C.7). Then we have

$$\begin{aligned}
\|\mathbf{A}_m\mathbf{A}_m^T\|_F &\leq \|\mathbf{A}\mathbf{A}^T\|_F + \|\mathbf{A}_m\mathbf{A}_m^T - \mathbf{A}\mathbf{A}^T\|_F \\
&= \|\mathbf{A}\mathbf{A}^T\|_F + \sqrt{\sum_{1\leq i,j\leq d}\left|\sum_{k=1}^m a_{ik}a_{jk} - \sum_{k=1}^\infty a_{ik}a_{jk}\right|^2} \\
&\leq \|\mathbf{A}\mathbf{A}^T\|_F + \sqrt{d^2(\delta/d)^2} = \|\mathbf{A}\mathbf{A}^T\|_F + \delta, \qquad \text{(C.8)}
\end{aligned}$$

and

$$
\begin{aligned}
\|\mathbf{A}_m\mathbf{A}_m^T\|_2 &\leq \|\mathbf{A}\mathbf{A}^T\|_2 + \|\mathbf{A}_m\mathbf{A}_m^T - \mathbf{A}\mathbf{A}^T\|_2 \\
&\leq \|\mathbf{A}\mathbf{A}^T\|_2 + \|\mathbf{A}_m\mathbf{A}_m^T - \mathbf{A}\mathbf{A}^T\|_F \leq \|\mathbf{A}\mathbf{A}^T\|_2 + \delta. \qquad \text{(C.9)}
\end{aligned}
$$

Similarly,

$$
\|\mathbf{B}_m\mathbf{B}_m^T\|_F \leq \|\mathbf{B}\mathbf{B}^T\|_F + \delta \qquad \text{and} \qquad \|\mathbf{B}_m\mathbf{B}_m^T\|_2 \leq \|\mathbf{B}\mathbf{B}^T\|_2 + \delta. \qquad \text{(C.10)}
$$

By Lemma 5.5 in Vershynin (2012), there exists a constant c_1 only dependent on K such that

$$
\sup_{k\geq 1} k^{-1/2}(E|e_j|^k)^{1/k} \leq c_1 \text{ for all } j = 1, 2, \dots.
$$

Then by Theorem 1.1 in Rudelson and Vershynin (2013) and Proposition 5.10 in Vershynin (2012), for every $t > 0$, there exists a constant $c > 0$ only dependent on c_1, i.e., only dependent on K, such that

$$
P\left[|\boldsymbol{X}_m^T\boldsymbol{Y}_m - E(\boldsymbol{X}_m^T\boldsymbol{Y}_m)| \geq t/4\right] \leq 2\exp\left\{-c\min\left(\frac{t^2}{\|\mathbf{A}_m^T\mathbf{B}_m\|_F^2}, \frac{t}{\|\mathbf{A}_m^T\mathbf{B}_m\|_2}\right)\right\}
$$

and

$$
P\left[|\boldsymbol{b}^T\boldsymbol{X}_m| \geq t/2\right] \leq \exp(1)\exp\left\{-\frac{ct^2}{\|\boldsymbol{b}^T\mathbf{A}_m\|_F^2}\right\}.
$$

Since

$$
\begin{aligned}
\|\mathbf{A}_m^T\mathbf{B}_m\|_F^2 &= \text{tr}(\mathbf{A}_m^T\mathbf{B}_m\mathbf{B}_m^T\mathbf{A}_m) = \text{tr}(\mathbf{A}_m\mathbf{A}_m^T\mathbf{B}_m\mathbf{B}_m^T) \\
&\leq \sqrt{\text{tr}(\mathbf{A}_m\mathbf{A}_m^T\mathbf{A}_m\mathbf{A}_m^T)\text{tr}(\mathbf{B}_m\mathbf{B}_m^T\mathbf{B}_m\mathbf{B}_m^T)} = \|\mathbf{A}_m\mathbf{A}_m^T\|_F\|\mathbf{B}_m\mathbf{B}_m^T\|_F,
\end{aligned}
$$

$$
\begin{aligned}
\|\mathbf{A}_m^T\mathbf{B}_m\|_2 &\leq \|\mathbf{A}_m^T\|_2\|\mathbf{B}_m\|_2 = \sqrt{\varphi_{\max}(\mathbf{A}_m\mathbf{A}_m^T)\varphi_{\max}(\mathbf{B}_m^T\mathbf{B}_m)} \\
&= \sqrt{\varphi_{\max}(\mathbf{A}_m\mathbf{A}_m^T)\varphi_{\max}(\mathbf{B}_m\mathbf{B}_m^T)} = \sqrt{\|\mathbf{A}_m\mathbf{A}_m^T\|_2\|\mathbf{B}_m\mathbf{B}_m^T\|_2},
\end{aligned}
$$

and

$$\|\boldsymbol{b}^T\mathbf{A}_m\|_F^2 = \|\mathbf{A}_m^T\boldsymbol{b}\|_F^2 \leq \|\mathbf{A}_m^T\|_2^2\|\boldsymbol{b}\|_F^2 = \varphi_{\max}(\mathbf{A}_m\mathbf{A}_m^T)\|\boldsymbol{b}\|_F^2 = \|\mathbf{A}_m\mathbf{A}_m^T\|_2\|\boldsymbol{b}\|_F^2$$

which is obtained by Lemma 1 in Lam and Fan (2009), then

$$\begin{aligned} & P\left[\left|\boldsymbol{X}_m^T\boldsymbol{Y}_m - E(\boldsymbol{X}_m^T\boldsymbol{Y}_m)\right| \geq t/4\right] \\ & \leq 2\exp\left\{-c\min\left(\frac{t^2}{\|\mathbf{A}_m\mathbf{A}_m^T\|_F\|\mathbf{B}_m\mathbf{B}_m^T\|_F}, \frac{t}{\sqrt{\|\mathbf{A}_m\mathbf{A}_m^T\|_2\|\mathbf{B}_m\mathbf{B}_m^T\|_2}}\right)\right\} \end{aligned} \tag{C.11}$$

and

$$P\left[|\boldsymbol{b}^T\boldsymbol{X}_m| \geq t/2\right] \leq \exp(1)\exp\left\{-\frac{ct^2}{\|\boldsymbol{b}\|_F^2\|\mathbf{A}_m\mathbf{A}_m^T\|_2}\right\}. \tag{C.12}$$

Let $\varepsilon_1 = t/2$ and $\varepsilon_2 = t/4$, then by (C.3), (C.5), (C.11), (C.8), (C.9) and (C.10) we have

$$\begin{aligned} & P\left[\left|\boldsymbol{X}^T\boldsymbol{Y} - E(\boldsymbol{X}^T\boldsymbol{Y})\right| \geq t\right] \\ & \leq P\left[|E(\boldsymbol{X}_m^T\boldsymbol{Y}_m) - E(\boldsymbol{X}^T\boldsymbol{Y})| + |\boldsymbol{X}_m^T\boldsymbol{Y}_m - E(\boldsymbol{X}_m^T\boldsymbol{Y}_m)| \geq t/2\right] \\ & \qquad + P\left[\left|\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}_m^T\boldsymbol{Y}_m\right| \geq t/2\right] \\ & = P\left[\left|\sum_{i=1}^{d}\sum_{k=1}^{m} a_{ik}b_{ik} - \sum_{i=1}^{d}\sum_{k=1}^{\infty} a_{ik}b_{ik}\right| + |\boldsymbol{X}_m^T\boldsymbol{Y}_m - E(\boldsymbol{X}_m^T\boldsymbol{Y}_m)| \geq t/2\right] \\ & \qquad + P\left[\left|\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}_m^T\boldsymbol{Y}_m\right| \geq \varepsilon_1\right] \\ & \leq P\left[\sum_{i=1}^{d}\left|\sum_{k=1}^{m} a_{ik}b_{ik} - \sum_{k=1}^{\infty} a_{ik}b_{ik}\right| + |\boldsymbol{X}_m^T\boldsymbol{Y}_m - E(\boldsymbol{X}_m^T\boldsymbol{Y}_m)| \geq t/2\right] + \delta \\ & \leq P\left[|\boldsymbol{X}_m^T\boldsymbol{Y}_m - E(\boldsymbol{X}_m^T\boldsymbol{Y}_m)| \geq t/2 - \varepsilon_2\right] + \delta \\ & \leq 2\exp\left\{-c\min\left(\frac{t^2}{\|\mathbf{A}_m\mathbf{A}_m^T\|_F\|\mathbf{B}_m\mathbf{B}_m^T\|_F}, \frac{t}{\sqrt{\|\mathbf{A}_m\mathbf{A}_m^T\|_2\|\mathbf{B}_m\mathbf{B}_m^T\|_2}}\right)\right\} + \delta \\ & \leq 2\exp\Bigg\{ - c\min\Bigg(\frac{t^2}{(\|\mathbf{A}\mathbf{A}^T\|_F + \delta)(\|\mathbf{B}\mathbf{B}^T\|_F + \delta)}, \\ & \qquad\qquad \frac{t}{\sqrt{(\|\mathbf{A}\mathbf{A}^T\|_2 + \delta)(\|\mathbf{B}\mathbf{B}^T\|_2 + \delta)}}\Bigg)\Bigg\} + \delta, \end{aligned} \tag{C.13}$$

and by (C.4), (C.12) and (C.9) we obtain

$$
\begin{aligned}
P\left[|\boldsymbol{b}^T\boldsymbol{X}| \geq t\right] &\leq P\left[|\boldsymbol{b}^T\boldsymbol{X}_m| \geq t/2\right] + P\left[|\boldsymbol{b}^T\boldsymbol{X} - \boldsymbol{b}^T\boldsymbol{X}_m| \geq t/2\right] \\
&\leq \exp(1)\exp\left\{-\frac{ct^2}{\|\boldsymbol{b}\|_F^2\|\mathbf{A}_m\mathbf{A}_m^T\|_2}\right\} + \delta \\
&\leq \exp(1)\exp\left\{-\frac{ct^2}{\|\boldsymbol{b}\|_F^2(\|\mathbf{A}\mathbf{A}^T\|_2 + \delta)}\right\} + \delta. \qquad \text{(C.14)}
\end{aligned}
$$

Letting $\delta \to 0$ on both sides of inequalities (C.13) and (C.14), we obtain (C.1) and (C.2).

□

Proof of Theorem IV.4. From (4.13) we see that $\tilde{\boldsymbol{\Sigma}}$ is invariant with any mean $\boldsymbol{\mu}_p$, so we assume $\boldsymbol{\mu}_p = \mathbf{0}$ without loss of generality.

Define $\boldsymbol{Z}_i^{(\ell)} = (Z_{i1}^{(\ell)}, \ldots, Z_{in_\ell}^{(\ell)})^T$ with $Z_{ij}^{(\ell)} = X_{ij}^{(\ell)}/\sqrt{\sigma_{ii}}$, then by (4.1), $\boldsymbol{Z}_i^{(\ell)} = \mathbf{P}_i^{(\ell)}\mathbf{H}^{(\ell)}\boldsymbol{e}$, where $\mathbf{P}_i^{(\ell)}$ is a $n_\ell \times pn_\ell$ matrix with $\sigma_{ii}^{-1/2}$ in the $\big(j, i + (j-1)p\big)$ entries and 0 in all other entries for $j = 1, \ldots, n_\ell$. From Proposition 2.7.1 in Brockwell and Davis (1991), we have $\text{corr}(\boldsymbol{Z}_i^{(\ell)}) = \text{cov}(\boldsymbol{Z}_i^{(\ell)}) = \text{cov}(\mathbf{P}_i^{(\ell)}\mathbf{H}^{(\ell)}\boldsymbol{e}) = \mathbf{P}_i^{(\ell)}\mathbf{H}^{(\ell)}\text{cov}(\boldsymbol{e})\mathbf{H}^{(\ell)^T}\mathbf{P}_i^{(\ell)^T} = \mathbf{P}_i^{(\ell)}\mathbf{H}^{(\ell)}\mathbf{H}^{(\ell)^T}\mathbf{P}_i^{(\ell)^T}$. Since

$$
|\tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}|_\infty \leq |\tilde{\boldsymbol{\mu}}^{\otimes 2}|_\infty + |\tilde{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}|_\infty \leq |\tilde{\boldsymbol{\mu}}|_\infty^2 + |\tilde{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}|_\infty, \qquad \text{(C.15)}
$$

then for any $u > 0$, by $|\mathbf{\Sigma}|_\infty \leq v_0$ we have

$$
\begin{aligned}
P\left[|\tilde{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty \geq 2u\right] &\leq P\left[|\tilde{\boldsymbol{\mu}}|_\infty^2 \geq u\right] + P\left[|\tilde{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}|_\infty \geq u\right] \\
&\leq \sum_{i=1}^{p} P\left[\left|\sum_{\ell=1}^{L} \frac{\varpi_\ell}{n_\ell} \sum_{k=1}^{n_\ell} X_{ij}^{(\ell)}\right| \geq u^{1/2}\right] \\
&\quad + \sum_{1\leq i,j\leq p} P\left[\left|\sum_{\ell=1}^{L} \frac{\varpi_\ell}{n_\ell} \sum_{k=1}^{n_\ell} X_{ik}^{(\ell)} X_{jk}^{(\ell)} - \sigma_{ij}\right| \geq u\right] \\
&= \sum_{i=1}^{p} P\left[\left|\sum_{\ell=1}^{L} \frac{\varpi_\ell}{n_\ell} \sum_{j=1}^{n_\ell} Z_{ij}^{(\ell)}\right| \geq \sqrt{\frac{u}{\sigma_{ii}}}\right] \\
&\quad + \sum_{1\leq i,j\leq p} P\left[\left|\sum_{\ell=1}^{L} \frac{\varpi_\ell}{n_\ell} \sum_{k=1}^{n_\ell} Z_{ik}^{(\ell)} Z_{jk}^{(\ell)} - \rho_{ij}\right| \geq \frac{u}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right] \\
&\leq \sum_{i=1}^{p} P\left[\left|\sum_{\ell=1}^{L} \frac{\varpi_\ell}{n_\ell} \sum_{j=1}^{n_\ell} Z_{ij}^{(\ell)}\right| \geq \sqrt{\frac{u}{v_0}}\right] \\
&\quad + \sum_{1\leq i,j\leq p} P\left[\left|\sum_{\ell=1}^{L} \frac{\varpi_\ell}{n_\ell} \sum_{k=1}^{n_\ell} Z_{ik}^{(\ell)} Z_{jk}^{(\ell)} - \rho_{ij}\right| \geq \frac{u}{v_0}\right]. \qquad \text{(C.16)}
\end{aligned}
$$

Now, consider the first term on the RHS of (C.16). For $i = 1, \ldots, p$, $\ell = 1, \ldots, L$, and $t > 0$, by (C.2) in Lemma C.1, we have

$$
\begin{aligned}
&P\left[\frac{1}{\sqrt{n_\ell g_\ell}}\left|\sum_{j=1}^{n_\ell} Z_{ij}^{(\ell)}\right| \geq \frac{t}{\sqrt{n_\ell g_\ell}}\right] \\
&= P\left[\left|\sum_{j=1}^{n_\ell} Z_{ij}^{(\ell)}\right| \geq t\right] = P\left[|\mathbf{1}_{n_\ell}^T \mathbf{P}_i^{(\ell)} \mathbf{H}^{(\ell)} \boldsymbol{e}| \geq t\right] \\
&\leq \exp(1)\exp\left\{-\frac{c_1 t^2}{\|\mathbf{1}_{n_\ell}\|_F^2 \|\mathbf{P}_i^{(\ell)} \mathbf{H}^{(\ell)} \mathbf{H}^{(\ell)T} \mathbf{P}_i^{(\ell)^T}\|_2}\right\} \\
&= \exp(1)\exp\left\{-\frac{c_1 t^2}{n_\ell \|\mathrm{corr}(\boldsymbol{Z}_i^{(\ell)})\|_2}\right\} \leq \exp(1)\exp\left\{-\frac{c_1 t^2}{n_\ell g_\ell}\right\}
\end{aligned}
$$

with some constant $c_1 > 0$ dependent on K. Obviously for $t = 0$, we still have the above inequality. Thus, $\sum_{j=1}^{n_\ell} Z_{ij}^{(\ell)}/\sqrt{n_\ell g_\ell}$ is a sub-Gaussian random variable from Definition 5.7

in Vershynin (2012), and then by their Proposition 5.10 we have

$$
\begin{aligned}
P\left[\left|\sum_{\ell=1}^{L}\frac{\varpi_\ell}{n_\ell}\sum_{j=1}^{n_\ell}Z_{ij}^{(\ell)}\right|\geq\sqrt{\frac{u}{v_0}}\right] &\leq \exp(1)\exp\left\{-\frac{c_2 u}{\sum_\ell(\varpi_\ell\sqrt{g_\ell/n_\ell})^2}\right\}\\
&= \exp(1)\exp\left\{-\frac{c_2(\sum_\ell n_\ell/f_\ell)^2 u}{\sum_\ell n_\ell g_\ell/f_\ell^2}\right\} \qquad \text{(C.17)}
\end{aligned}
$$

with some constant $c_2 > 0$ dependent on c_1 and v_0.

Next, consider the second term on the RHS of (C.16). For $i, j = 1, \ldots, p$, $\ell = 1, \ldots, L$, and $t > 0$, by (C.1) in Lemma C.1 we have

$$
\begin{aligned}
P&\left[\left|\sum_{k=1}^{n_\ell}Z_{ik}^{(\ell)}Z_{jk}^{(\ell)}-n_\ell\rho_{ij}\right|\geq t\right] = P\left[\left|\boldsymbol{Z}_i^{(\ell)^T}\boldsymbol{Z}_j^{(\ell)}-E(\boldsymbol{Z}_i^{(\ell)^T}\boldsymbol{Z}_j^{(\ell)})\right|\geq t\right]\\
&\leq 2\exp\Bigg\{-c_3\min\Bigg(\frac{t^2}{\|\mathbf{P}_i^{(\ell)}\mathbf{H}^{(\ell)}\mathbf{H}^{(\ell)^T}\mathbf{P}_i^{(\ell)^T}\|_F\|\mathbf{P}_j^{(\ell)}\mathbf{H}^{(\ell)}\mathbf{H}^{(\ell)^T}\mathbf{P}_j^{(\ell)^T}\|_F},\\
&\qquad\qquad \frac{t}{\sqrt{\|\mathbf{P}_i^{(\ell)}\mathbf{H}^{(\ell)}\mathbf{H}^{(\ell)^T}\mathbf{P}_i^{(\ell)^T}\|_2\|\mathbf{P}_j^{(\ell)}\mathbf{H}^{(\ell)}\mathbf{H}^{(\ell)^T}\mathbf{P}_j^{(\ell)^T}\|_2}}\Bigg)\Bigg\}\\
&= 2\exp\Bigg\{-c_3\min\Bigg(\frac{t^2}{\|\mathrm{corr}(\boldsymbol{Z}_i^{(\ell)})\|_F\|\mathrm{corr}(\boldsymbol{Z}_j^{(\ell)})\|_F},\\
&\qquad\qquad \frac{t}{\sqrt{\|\mathrm{corr}(\boldsymbol{Z}_i^{(\ell)})\|_2\|\mathrm{corr}(\boldsymbol{Z}_j^{(\ell)})\|_2}}\Bigg)\Bigg\}\\
&\leq 2\exp\left\{-c_3\min\left(\frac{t^2}{n_\ell g_\ell},\frac{t}{g_\ell}\right)\right\} \qquad \text{(C.18)}
\end{aligned}
$$

with some constant $c_3 > 0$ dependent on K, where the last inequality follows from (4.2) and

$$
\begin{aligned}
\frac{1}{n_\ell}\|\mathrm{corr}(\boldsymbol{Z}_i^{(\ell)})\|_F^2 &= \frac{1}{n_\ell}\mathrm{tr}([\mathrm{corr}(\boldsymbol{Z}_i^{(\ell)})]^2) = \frac{1}{n_\ell}\sum_{k=1}^{n_\ell}\varphi_k^2(\mathrm{corr}(\boldsymbol{Z}_i^{(\ell)}))\\
&\leq \varphi_{\max}(\mathrm{corr}(\boldsymbol{Z}_i^{(\ell)}))\frac{1}{n_\ell}\sum_{k=1}^{n_\ell}\varphi_k(\mathrm{corr}(\boldsymbol{Z}_i^{(\ell)}))\\
&= \|\mathrm{corr}(\boldsymbol{Z}_i^{(\ell)})\|_2\frac{1}{n_\ell}\mathrm{tr}(\mathrm{corr}(\boldsymbol{Z}_i^{(\ell)})) = \|\mathrm{corr}(\boldsymbol{Z}_i^{(\ell)})\|_2 \leq g_\ell.
\end{aligned}
$$

Obviously for $t = 0$, we still have (C.18). Let $Y_{ij}^{(\ell)} = \sum_{k=1}^{n_\ell}Z_{ik}^{(\ell)}Z_{jk}^{(\ell)} - n_\ell\rho_{ij}$, $Y_{ij,1}^{(\ell)} =$

$Y_{ij}^{(\ell)}\mathbb{I}(|Y_{ij}^{(\ell)}| \le n_\ell)$, and $Y_{ij,2}^{(\ell)} = Y_{ij}^{(\ell)}\mathbb{I}(|Y_{ij}^{(\ell)}| > n_\ell)$. Then for $t \ge 0$,

$$\begin{aligned} P\left[\frac{|Y_{ij,1}^{(\ell)}|}{\sqrt{n_\ell g_\ell}} \ge \frac{t}{\sqrt{n_\ell g_\ell}}\right] &= P\left[|Y_{ij,1}^{(\ell)}| \ge t\right] \\ &= P\left[|Y_{ij,1}^{(\ell)}| \ge t\right]\mathbb{I}(t \le n_\ell) + P\left[|Y_{ij,1}^{(\ell)}| \ge t\right]\mathbb{I}(t > n_\ell) \\ &\le P\left[|Y_{ij}^{(\ell)}| \ge t\right]\mathbb{I}(t \le n_\ell) + 0 \le 2\exp\left\{-\frac{c_3 t^2}{n_\ell g_\ell}\right\} \end{aligned}$$

and

$$P\left[\frac{|Y_{ij,2}^{(\ell)}|}{g_\ell} \ge \frac{t}{g_\ell}\right] = P\left[|Y_{ij,2}^{(\ell)}| \ge t\right] = P\left[|Y_{ij,2}^{(\ell)}| \ge t\right]\mathbb{I}(t \le n_\ell) + P\left[|Y_{ij,2}^{(\ell)}| \ge t\right]\mathbb{I}(t > n_\ell)$$

$$\le \mathbb{I}(t=0) + P\left[|Y_{ij}^{(\ell)}| > n_\ell\right]\mathbb{I}(0 < t \le n_\ell) + P\left[|Y_{ij}^{(\ell)}| \ge t\right]\mathbb{I}(t > n_\ell) \le 2\exp\left\{-\frac{c_3 t}{g_\ell}\right\}.$$

Thus, by Definitions 5.7 and 5.13 in Vershynin (2012), $Y_{ij,1}^{(\ell)}/\sqrt{n_\ell g_\ell}$ and $Y_{ij,2}^{(\ell)}/g_\ell$ are sub-Gaussian and sub-exponential random variables, respectively. Then by Propositions 5.10 and 5.16 in Vershynin (2012) and $g_\ell \le n_\ell$, we have

$$P\left[\left|\sum_{\ell=1}^{L}\frac{\varpi_\ell Y_{ij,1}^{(\ell)}}{n_\ell}\right| \ge \frac{u}{2v_0}\right] \le \exp(1)\exp\left\{-\frac{c_4(\sum_\ell n_\ell/f_\ell)^2 u^2}{\sum_\ell n_\ell g_\ell/f_\ell^2}\right\}$$

and

$$\begin{aligned} P\left[\left|\sum_{\ell=1}^{L}\frac{\varpi_\ell Y_{ij,2}^{(\ell)}}{n_\ell}\right| \ge \frac{u}{2v_0}\right] &\le 2\exp\left\{-c_4\min\left[\frac{u^2}{\sum_\ell(\varpi_\ell g_\ell/n_\ell)^2}, \frac{u}{\max_\ell(\varpi_\ell g_\ell/n_\ell)}\right]\right\} \\ &= 2\exp\left\{-c_4\min\left[\frac{(\sum_\ell n_\ell/f_\ell)^2u^2}{\sum_\ell(g_\ell/f_\ell)^2}, \frac{(\sum_\ell n_\ell/f_\ell)u}{\max_\ell(g_\ell/f_\ell)}\right]\right\} \\ &\le 2\exp\left\{-c_4\min\left[\frac{(\sum_\ell n_\ell/f_\ell)^2u^2}{\sum_\ell n_\ell g_\ell/f_\ell^2}, \frac{(\sum_\ell n_\ell/f_\ell)u}{\max_\ell(g_\ell/f_\ell)}\right]\right\} \end{aligned}$$

with some constant $c_4 > 0$ dependent on c_3 and v_0. Hence,

$$\begin{aligned} P\left[\left|\sum_{\ell=1}^{L}\frac{\varpi_\ell}{n_\ell}\sum_{k=1}^{n_\ell}Z_{ik}^{(\ell)}Z_{jk}^{(\ell)} - \rho_{ij}\right| \ge \frac{u}{v_0}\right] &= P\left[\left|\sum_{\ell=1}^{L}\frac{\varpi_\ell Y_{ij}^{(\ell)}}{n_\ell}\right| \ge \frac{u}{v_0}\right] \\ &\le P\left[\left|\sum_{\ell=1}^{L}\frac{\varpi_\ell Y_{ij,1}^{(\ell)}}{n_\ell}\right| \ge \frac{u}{2v_0}\right] + P\left[\left|\sum_{\ell=1}^{L}\frac{\varpi_\ell Y_{ij,2}^{(\ell)}}{n_\ell}\right| \ge \frac{u}{2v_0}\right] \\ &\le [\exp(1)+2]\exp\left\{-\frac{c_4(\sum_\ell n_\ell/f_\ell)^2u^2}{\sum_\ell n_\ell g_\ell/f_\ell^2}\right\} + 2\exp\left\{-\frac{c_4(\sum_\ell n_\ell/f_\ell)u}{\max_\ell(g_\ell/f_\ell)}\right\}. \end{aligned} \tag{C.19}$$

Plugging (C.17) and (C.19) into (C.16), we obtain

$$\begin{aligned}
&P\left[|\tilde{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty \geq 2u\right] \\
\leq\, & p\exp(1)\exp\left\{-\frac{c_2(\sum_\ell n_\ell/f_\ell)^2 u}{\sum_\ell n_\ell g_\ell/f_\ell^2}\right\} + p^2[\exp(1)+2]\exp\left\{-\frac{c_4(\sum_\ell n_\ell/f_\ell)^2 u^2}{\sum_\ell n_\ell g_\ell/f_\ell^2}\right\} \\
&+ 2p^2\exp\left\{-\frac{c_4(\sum_\ell n_\ell/f_\ell)u}{\max_\ell(g_\ell/f_\ell)}\right\} \\
\leq\, & [p\exp(1) + p^2\exp(1) + 2p^2]\exp\left\{-\min(c_2,c_4)\frac{(\sum_\ell n_\ell/f_\ell)^2 u^2}{\sum_\ell n_\ell g_\ell/f_\ell^2}\right\} \\
&+ 2p^2\exp\left\{-\frac{c_4(\sum_\ell n_\ell/f_\ell)u}{\max_\ell(g_\ell/f_\ell)}\right\}
\end{aligned}$$

for $0 < u < 1$. By $\tau_2 = o(1)$, we have $u = o(1)$ when $u = M\tau_2/2$ with a constant $M > 0$. Then plugging $u = M\tau_2/2$ into the above inequality yields (4.15) for any given constant $M' > 0$ by choosing sufficiently large M. □

Proof of Theorem IV.5. The proofs for generalized thresholding and SPICE are identical to the proof of Theorem III.7 after (B.31) and the proof of Theorem III.9, respectively, with corresponding notational changes. The proof for the consistency of the CLIME estimator is identical to the proofs of Theorems 2, 5 and 6 in Cai et al. (2011) following (4.15), where we also obtain $|\hat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}|_\infty \leq 4M_p\lambda_1$ with probability tending to 1. Then the proof for the sparsistency and sign-consistency of the thresholded CLIME estimator follows the same arguments for the proof of Theorem 2 in Rothman et al. (2009). Details are hence omitted. □

BIBLIOGRAPHY

BIBLIOGRAPHY

Alexander, G. E., Chen, K., Pietrini, P., Rapoport, S. I., and Reiman, E. M. (2002), "Longitudinal PET evaluation of cerebral metabolic decline in dementia: a potential outcome measure in Alzheimer's disease treatment studies," *American Journal of Psychiatry*, 159, 738–745.

Athreya, K. B. and Lahiri, S. N. (2006), *Measure Theory and Probability Theory*, New York: Springer.

Bai, Z. D. and Silverstein, J. W. (2010), *Spectral Analysis of Large Dimensional Random Matrices*, New York: Springer, 2nd ed.

Bai, Z. D. and Yin, Y. Q. (1993), "Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix," *Annals of Probability*, 21, 1275–1294.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008), "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *The Journal of Machine Learning Research*, 9, 485–516.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2004), "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences of the United States of America*, 101, 3747–3752.

Basu, S., Michailidis, G., et al. (2015), "Regularized estimation in sparse high-dimensional time series models," *The Annals of Statistics*, 43, 1535–1567.

Bellman, R. E. (1961), *Adaptive Control Processes: A Guided Tour*, Princeton University Press.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 289–300.

— (2000), "On the adaptive control of the false discovery rate in multiple testing with independent statistics," *Journal of educational and Behavioral Statistics*, 25, 60–83.

Benjamini, Y. and Yekutieli, D. (2001), "The control of the false discovery rate in multiple testing under dependency," *Annals of Statistics*, 29, 1165–1188.

Benson, D. F., Kuhl, D. E., Hawkins, R. A., Phelps, M. E., Cummings, J. L., and Tsai, S. (1983), "The fluorodeoxyglucose 18F scan in Alzheimer's disease and multi-infarct dementia," *Archives of neurology*, 40, 711–714.

Bhattacharjee, M. and Bose, A. (2014), "Consistency of large dimensional sample covariance matrix under weak dependence," *Statistical Methodology*, 20, 11–26.

Bickel, P. J. and Levina, E. (2008a), "Covariance regularization by thresholding," *Annals of Statistics*, 36, 2577–2604.

— (2008b), "Regularized estimation of large covariance matrices," *Annals of Statistics*, 36, 199–227.

Bochud, T. and Challet, D. (2007), "Optimal approximations of power laws with exponentials: application to volatility models with long memory," *Quantitative Finance*, 7, 585–589.

Booth, J. G. and Hobert, J. P. (1999), "Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61, 265–285.

Bradley, R. C. (2005), "Basic properties of strong mixing conditions. A survey and some open questions," *Probab. Surv.*, 2, 107–144, update of, and a supplement to, the 1986 original.

Brémaud, P. (1999), *Markov Chains: Gibbs fields, Monte Carlo Simulation, and Queues*, New York: Springer.

Brockwell, P. J. and Davis, R. A. (1991), *Time Series: Theory and Methods*, New York: Springer-Verlag, 2nd ed.

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., and Arrighi, H. M. (2007), "Forecasting the global burden of Alzheimer's disease," *Alzheimer's & dementia*, 3, 186–191.

Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008), "The brain's default network," *Annals of the New York Academy of Sciences*, 1124, 1–38.

Buckner, R. L., Sepulcre, J., Talukdar, T., Krienen, F. M., Liu, H., Hedden, T., Andrews-Hanna, J. R., Sperling, R. A., and Johnson, K. A. (2009), "Cortical hubs revealed by intrinsic functional connectivity: Mapping, assessment of stability, and relation to Alzheimer's disease," *The Journal of Neuroscience*, 29, 1860–1873.

Buldygin, V. V. and Kozachenko, Y. V. (2000), *Metric Characterization of Random Variables and Random Processes*, Providence, RI: American Mathematical Society.

Cai, T. T. and Liu, W. D. (2011), "Adaptive thresholding for sparse covariance matrix estimation," *Journal of the American Statistical Association*, 106, 672–684.

Cai, T. T., Liu, W. D., and Luo, X. (2011), "A constrained ℓ_1 minimization approach to sparse precision matrix estimation," *Journal of the American Statistical Association*, 106, 594–607.

Cai, T. T., Liu, W. D., and Zhou, H. H. (2016), "Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation," *Annals of Statistics*, 44, 455–488.

Cai, T. T. and Sun, W. (2009), “Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks,” *Journal of the American Statistical Association*, 104, 1467–1481.

Cai, T. T. and Yuan, M. (2012), “Adaptive covariance matrix estimation through block thresholding,” *Annals of Statistics*, 40, 2014–2042.

Cai, T. T., Zhang, C. H., and Zhou, H. H. (2010), “Optimal rates of convergence for covariance matrix estimation,” *Annals of Statistics*, 38, 2118–2144.

Cai, T. T. and Zhou, H. H. (2012), “Optimal rates of convergence for sparse covariance matrix estimation,” *Annals of Statistics*, 40, 2389–2420.

Cao, G., Bachega, L. R., and Bouman, C. A. (2011), “The sparse matrix transform for covariance estimation and analysis of high dimensional signals,” *IEEE Transactions on Image Processing*, 20, 625–640.

Chandgotia, N., Han, G., Marcus, B., Meyerovitch, T., and Pavlov, R. (2014), “One-dimensional Markov random fields, Markov chains and topological Markov fields,” *Proceedings of the American Mathematical Society*, 142, 227–242.

Chen, J., Huang, Y., and Wang, P. (2014), “Composite likelihood under hidden Markov model,” *Statistica Sinica*, doi:10.5705/ss.2013.084t.

Chen, J., Tan, X., and Zhang, R. (2008), “Inference for normal mixtures in mean and variance,” *Statistica Sinica*, 443–465.

Chen, X., Xu, M., and Wu, W. B. (2013), “Covariance and precision matrix estimation for high-dimensional time Series,” *Annals of Statistics*, 41, 2994–3021.

Chumbley, J., Worsley, K., Flandin, G., and Friston, K. (2010), “Topological FDR for neuroimaging,” *Neuroimage*, 49, 3057–3064.

Chumbley, J. R. and Friston, K. J. (2009), “False discovery rate revisited: FDR and topological inference using Gaussian random fields,” *NeuroImage*, 44, 62–70.

Ciuperca, G., Ridolfi, A., and Idier, J. (2003), “Penalized maximum likelihood estimator for normal mixtures,” *Scandinavian Journal of Statistics*, 30, 45–59.

Cole, M. W., Pathak, S., and Schneider, W. (2010), “Identifying the brain’s most globally connected regions,” *NeuroImage*, 49, 3132–3148.

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton, NJ: Princeton University Press.

Dabney, A. and Storey, J. D. (2014), “qvalue: Q-value estimation for false discovery rate control,” *R package*, version 1.36.0.

Demko, S., Moss, W. F., and Smith, P. W. (1984), “Decay rates for inverses of band matrices,” *Mathematics* *of Computation*, 43, 491–499.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the royal statistical society. Series B (methodological)*, 39, 1–38.

Donoho, D. L. et al. (2000), “High-dimensional data analysis: The curses and blessings of dimensionality,” *AMS Math Challenges Lecture*, 1–32.

Efron, B. (2004), “Large-scale simultaneous hypothesis testing: the choice of a null hypothesis,” *J. Amer. Statist. Assoc.*, 99, 96–104.

El Karoui, N. (2008), “Operator norm consistent estimation of large-dimensional sparse covariance matrices,” *Annals of Statistics*, 36, 2717–2756.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011), *Cluster Analysis*, Hoboken: Wiley, 5th ed.

Fan, J., Han, F., and Liu, H. (2014a), “Challenges of big data analysis,” *National science review*, 1, 293–314.

Fan, L., Gau, S. S., and Chou, T. (2014b), “Neural correlates of inhibitory control and visual processing in youths with attention deficit hyperactivity disorder: a counting Stroop functional MRI study,” *Psychological Medicine*, 44, 2661–2671.

Fang, Y., Wang, B., and Feng, Y. (2016), “Tuning-parameter selection in regularized estimations of large covariance matrices,” *Journal of Statistical Computation and Simulation*, 86, 494–509.

Farcomeni, A. (2007), “Some results on the control of the false discovery rate under dependence,” *Scand. J. Statist.*, 34, 275–297.

Fomin, V. (1999), *Optimal Filtering. Volume II: Spatio-Temporal Fields*, Dordrecht: Kluwer Academic Publishers.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), “Sparse inverse covariance estimation with the graphical Lasso,” *Biostatistics*, 9, 432–441.

Friston, K. J. (2011), “Functional and effective connectivity: a review,” *Brain connectivity*, 1, 13–36.

Garey, L. J. (2006), *Brodmann’s Localisation in the Cerebral Cortex*, Springer.

Geman, S. and Geman, D. (1984), “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Genovese, C. and Wasserman, L. (2002), “Operating characteristics and extensions of the false discovery rate procedure,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64, 499–517.

— (2004), “A stochastic process approach to false discovery control,” *Ann. Statist.*, 32, 1035–1061.

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002), "Thresholding of statistical maps in functional neuroimaging using the false discovery rate," *NeuroImage*, 15, 870–878.

Golub, G. H. and Van Loan, C. F. (1996), *Matrix Computations*, Baltimore, MD: Johns Hopkins University Press, 3rd ed.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning*, New York: Springer, 2nd ed.

Hoff, P. D. (2009), *A First Course in Bayesian Statistical Methods*, New York: Springer.

Horn, R. A. and Johnson, C. R. (2013), *Matrix Analysis*, Cambridge: Cambridge University Press, 2nd ed.

Hsieh, C. J., Sustik, M. A., Dhillon, I. S., and Ravikumar, P. K. (2014), "QUIC: Quadratic approximation for sparse inverse covariance estimation," *Journal of Machine Learning Research*, 15, 2911–2947.

Hu, T.-C., Rosalsky, A., and Volodin, A. (2008), "On convergence properties of sums of dependent random variables under second moment and covariance restrictions," *Statistics & Probability Letters*, 78, 1999–2005.

Huang, C., Tang, C., Feigin, A., Lesser, M., Ma, Y., Pourfar, M., Dhawan, V., and Eidelberg, D. (2007), "Changes in network activity with the progression of Parkinson's disease," *Brain*, 130, 1834–1846.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006), "Covariance matrix selection and estimation via penalised normal likelihood," *Biometrika*, 93, 85–98.

Huang, L., Goldsmith, J., Reiss, P. T., Reich, D. S., and Crainiceanu, C. M. (2013), "Bayesian scalar-on-image regression with application to association between intracranial DTI and cognitive outcomes," *NeuroImage*, 83, 210–223.

Ishii, K. (2002), "Clinical application of positron emission tomography for diagnosis of dementia," *Annals of nuclear medicine*, 16, 515–525.

— (2014), "PET approaches for diagnosis of dementia," *American Journal of Neuroradiology*, 35, 2030–2038.

Jeong, Y., Cho, S. S., Park, J. M., Kang, S. J., Lee, J. S., Kang, E., Na, D. L., and Kim, S. E. (2005), "18F-FDG PET findings in frontotemporal dementia: an SPM analysis of 29 patients," *Journal of Nuclear Medicine*, 46, 233–239.

Jiang, T. (2003), "The limiting distributions of eigenvalues of sample correlation matrices," *Sankhyā*, 66, 35–48.

Johnson, T. D., Liu, Z., Bartsch, A. J., and Nichols, T. E. (2013), "A Bayesian non-parametric Potts model with application to pre-surgical FMRI data," *Statistical Methods in Medical Research*, 22, 364–381.

Johnstone, I. M. and Titterington, D. M. (2009), "Statistical challenges of high-dimensional data," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367, 4237–4253.

Karr, A. F. (1993), *Probability*, New York: Springer-Verlag.

Kim, B., Lee, J., Shin, M., Cho, S., and Lee, D. (2002), "Regional cerebral perfusion abnormalities in attention deficit/hyperactivity disorder," *European Archives of Psychiatry and Clinical Neuroscience*, 252, 219–225.

Lam, C. and Fan, J. (2009), "Sparsistency and rates of convergence in large covariance matrix estimation," *Annals of Statistics*, 37, 4254–4278.

Langbaum, J. B., Chen, K., Lee, W., Reschke, C., Bandy, D., Fleisher, A. S., Alexander, G. E., Foster, N. L., Weiner, M. W., Koeppe, R. A., et al. (2009), "Categorical and correlational analyses of baseline fluorodeoxyglucose positron emission tomography images from the Alzheimer's Disease Neuroimaging Initiative (ADNI)," *NeuroImage*, 45, 1107–1116.

Li, X., Zhao, T., Yuan, X., and Liu, H. (2015), "The flare package for high dimensional linear regression and precision matrix estimation in R," *Journal of Machine Learning Research*, 16, 553–557.

Liu, H., Lafferty, J., and Wasserman, L. (2009), "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs," *Journal of Machine Learning Research*, 10, 2295–2328.

Magder, L. S. and Zeger, S. L. (1996), "A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians," *J. Amer. Statist. Assoc.*, 91, 1141–1151.

Manolakis, D. G., Ingle, V. K., and Kogon, S. M. (2005), *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering, and Array Processing*, Norwood, MA: Artech House.

Matsuda, H. (2001), "Cerebral blood flow and metabolic abnormalities in Alzheimer's disease," *Annals of Nuclear Medicine*, 15, 85–92.

Meinshausen, N. and Bühlmann, P. (2006), "High-dimensional graphs and variable selection with the Lasso," *The Annals of Statistics*, 34, 1436–1462.

Mosconi, L., Tsui, W.-H., De Santi, S., Li, J., Rusinek, H., Convit, A., Li, Y., Boppana, M., and De Leon, M. (2005), "Reduced hippocampal metabolism in MCI and AD Automated FDG-PET image analysis," *Neurology*, 64, 1860–1867.

Nestor, P. J., Fryer, T. D., Smielewski, P., and Hodges, J. R. (2003), "Limbic hypometabolism in Alzheimer's disease and mild cognitive impairment," *Annals of Neurology*, 54, 343–351.

Ng, B., Varoquaux, G., Poline, J. B., and Thirion, B. (2013), "A novel sparse group Gaussian graphical model for functional connectivity estimation," in *Information Processing in Medical Imaging*, Springer, pp. 256–267.

Nocedal, J. and Wright, S. J. (2006), *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer, New York, 2nd ed.

Palma, W. (2007), *Long-Memory Time Series: Theory and Methods*, Hoboken, NJ: Wiley-Interscience.

Pang, H., Liu, H., and Vanderbei, R. (2014), "The FASTCLIME package for linear programming and large-scale precision matrix estimation in R," *The Journal of Machine Learning Research*, 15, 489–493.

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., and Petersen, S. E. (2011), "Functional network organization of the human brain," *Neuron*, 72, 665–678.

Priestley, M. B. and Subba Rao, T. (1969), "A test for non-stationarity of time-series," *JRSSB*, 31, 140–149.

Racine, J. (2000), "Consistent cross-validatory model-selection for dependent data: *hv*-block cross-validation," *Journal of Econometrics*, 99, 39–61.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011), "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electronic Journal of Statistics*, 5, 935–980.

Ridolfi, A. (1997), "Maximum likelihood estimation of hidden Markov model parameters, with application to medical image segmentation," Tech. rep., Milan, Italy.

Roberts, G. O. and Smith, A. F. M. (1994), "Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms," *Stochastic Process. Appl.*, 49, 207–216.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008), "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, 2, 494–515.

Rothman, A. J., Levina, E., and Zhu, J. (2009), "Generalized thresholding of large covariance matrices," *Journal of the American Statistical Association*, 104, 177–186.

— (2010), "A new approach to Cholesky-based covariance regularization in high dimensions," *Biometrika*, 97, 539–550.

Rudelson, M. and Vershynin, R. (2013), "Hanson-Wright inequality and sub-Gaussian concentration," *Electronic Communications in Probability*, 18, 1–9.

Ryali, S., Chen, T., Supekar, K., and Menon, V. (2012), "Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty," *NeuroImage*, 59, 3852–3861.

Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D. A., Griffanti, L., Harms, M. P., et al. (2013), "Resting-state fMRI in the human connectome project," *Neuroimage*, 80, 144–168.

Sripada, C., Angstadt, M., Kessler, D., Phan, K. L., I., L., Evans, G. W., Welsh, R. C., Kim, P., and Swain, J. E. (2014), "Volitional regulation of emotions produces distributed alterations in connectivity between visual, attention control, and default networks," *NeuroImage*, 89, 110–121.

Stoer, J. and Bulirsch, R. (2002), *Introduction to Numerical Analysis*, vol. 12 of *Texts in Applied Mathematics*, Springer-Verlag, New York, 3rd ed., translated from the German by R. Bartels, W. Gautschi and C. Witzgall.

Storey, J. D. (2003), "The positive false discovery rate: a Bayesian interpretation and the q-value," *Ann. Statist.*, 31, 2013–2035.

Sun, W. and Cai, T. T. (2007), "Oracle and adaptive compound decision rules for false discovery rate control," *J. Amer. Statist. Assoc.*, 102, 901–912.

— (2009), "Large-scale multiple testing under dependence," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71, 393–424.

Syed, M. N., Principe, J. C., and Pardalos, P. M. (2012), "Correntropy in data classification," in *Dynamics of Information Systems: Mathematical Foundations*, eds. Sorokin, A., Murphey, R., Thai, M. T., and Pardalos, P. M., New York: Springer, pp. 81–117.

Thomas, R., Sanders, S., Doust, J., Beller, E., and Glasziou, P. (2015), "Prevalence of attention-deficit/hyperactivity disorder: a systematic review and meta-analysis," *Pediatrics*, 135, e994–e1001.

Tomson, S. N., Schreiner, M. J., Narayan, M., Rosser, T., Enrique, N., Silva, A. J., Allen, G. I., Bookheimer, S. Y., and Bearden, C. E. (2015), "Resting state functional MRI reveals abnormal network connectivity in neurofibromatosis 1," *Human brain mapping*, 36, 4566–4581.

Van de Geer, S., Bühlmann, P., Ritov, Y. A., and Dezeure, R. (2014), "On asymptotically optimal confidence regions and tests for high-dimensional models," *The Annals of Statistics*, 42, 1166–1202.

Vershynin, R. (2012), "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing*, eds. Eldar, Y. C. and Kutyniok, G., Cambridge: Cambridge University Press, pp. 210–268.

Wei, Z., Sun, W., Wang, K., and Hakonarson, H. (2009), "Multiple testing in genome-wide association studies via hidden Markov models," *Bioinformatics*, 25, 2802–2808.

Welch, B. L. (1947), "The generalization of 'Student's' problem when several different population variances are involved," *Biometrika*, 34, 28–35.

Winkler, G. (2003), *Image analysis, random fields and Markov chain Monte Carlo methods*, vol. 27 of *Applications of Mathematics (New York)*, Springer-Verlag, Berlin, 2nd ed., a mathematical introduction, With 1 CD-ROM (Windows), Stochastic Modelling and Applied Probability.

Wu, W. B. (2005), "Nonlinear system theory: Another look at dependence," *Proceedings of the National Academy of Sciences*, 102, 14150–14154.

— (2008), “On false discovery control under dependence,” *Ann. Statist.*, 36, 364–380.

Wu, W. B. and Pourahmadi, M. (2009), “Banding sample autocovariance matrices of stationary processes,” *Statistica Sinica*, 19, 1755–1768.

Xie, J., Cai, T. T., Maris, J., and Li, H. (2011), “Optimal false discovery rate control for dependent data,” *Statistics and Its Interface*, 4, 417–430.

Yuan, M. (2010), “High dimensional inverse covariance matrix estimation via linear programming,” *The Journal of Machine Learning Research*, 11, 2261–2286.

Yuan, M. and Lin, Y. (2007), “Model selection and estimation in the Gaussian graphical model,” *Biometrika*, 94, 19–35.

Zeman, M. N., Carpenter, G. M., and Scott, P. J. (2011), “Diagnosis of dementia using nuclear medicine imaging modalities,” in *12 Chapters on Nuclear Medicine*, ed. Gholamrezanezhad, A., Croatia: InTech, pp. 199–230.

Zhang, C., Fan, J., and Yu, T. (2011), “Multiple testing via FDR_L for large-scale image data,” *Annals of statistics*, 39, 613–642.

Zhang, X., Johnson, T. D., Little, R. J. A., and Cao, Y. (2008), “Quantitative magnetic resonance image analysis via the EM algorithm with stochastic variation,” *Ann. Appl. Stat.*, 2, 736–755.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012), “The huge Package for high-dimensional undirected graph estimation in R,” *The Journal of Machine Learning Research*, 13, 1059–1062.

Zhou, D., Thompson, W. K., and Siegle, G. (2009), “MATLAB toolbox for functional connectivity,” *Neuroimage*, 47, 1590–1607.

Zhou, S. (2014), “Gemini: Graph estimation with matrix variate normal instances,” *Annals of Statistics*, 42, 532–562.