THE UNIVERSITY OF CHICAGO

TWO PROBLEMS IN HIGH DIMENSIONAL INFERENCE: L^2 TEST BY RESAMPLING AND NETWORK ESTIMATION FROM NON-STATIONARY TIME SERIES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
MENGYU XU

CHICAGO, ILLINOIS
AUGUST 2016

Copyright © 2016 by Mengyu Xu

All Rights Reserved

To My Parents

TABLE OF CONTENTS

LIST OF FIGURES v

ACKNOWLEDGMENTS vi

ABSTRACT vii

1 INTRODUCTION 1

2 ASYMPTOTIC THEORY FOR HIGH-DIMENSIONAL L^2 STATISTICS AND RESAMPLING METHODS 3

2.1 Introduction 3

2.2 An Invariance Principle 7

2.2.1 Central and non-Central Limit Theorems 9

2.2.2 High Dimensional Cramér-von Mises Test 13

2.3 Estimating Distributions of Quadratic Forms 15

2.3.1 A Plug-in Procedure 15

2.3.2 A Subsampling Procedure 19

2.3.3 A Half-Sampling Procedure 21

2.4 Extension of Half-Sampling to Two-Sample Problem 24

2.5 Power Analysis 27

2.6 A Simulation Study 29

2.7 A Real Data Application 30

2.8 Proof of the Results 32

3 ESTIMATING HIGH-DIMENSIONAL DYNAMIC GRAPHS 51

3.1 Introduction 51

3.2 Time Series Model and Assumptions 54

3.3 Method: Change Point Detection and Support Recovery 60

3.4 Theoretical Results 63

3.5 A Real Data Application 67

3.6 Proof of the Results 71

3.6.1 Preliminary Lemmas 71

3.6.2 Proof of the Main Results 77

REFERENCES 81

LIST OF FIGURES

2.1 The empirical cumulative distribution function of the p-values based on 1000 simulations from Model 1. The data are generated with $\beta = 2$ and $\beta = 0.6$ in the left and right panels respectively. 49

2.2 The empirical cumulative distribution function of the p-values based on 1000 simulations from Model 2. The data are generated with $a = 0.1$ and $a = 1$ in the left and right panels respectively. 50

3.1 Break size $|D_s|_\infty$. From January 1, 2003 to January 1, 2008. 69

3.2 Estimated networks at time points 813, 828, 888 and 903, corresponding to March 23, 2006, April 13, 2006, July 11, 2006, and August 1, 2006. Networks in the first and second row are estimated before and after the estimated change point at time 858, respectively. Different colors correspond to the nine sections in the S&P dataset. 70

ACKNOWLEDGMENTS

First and foremost, I am deeply grateful to my advisor, Professor Wei-Biao Wu, who has continuously guided, supported and encouraged me throughout my graduate studies with his patience and insight. I am indebted to him and will be forever thankful.

I greatly appreciate Professor Tracy Ke and Professor Per Mykland for being my thesis committee members, for their valuable suggestions and close attention on my research. I thank Professor Xiaofeng Shao and Professor Mei Wang for their great help and support during my job-hunting. I am grateful to Professor Xiaohui Chen for his valuable comments and helpful discussions. I would also like to express my sincere gratitude to all the faculty, staff, and students of the Department of Statistics at the University of Chicago. I have been very fortunate to study in such an intellectually stimulating environment.

I would like to thank my dearest friends who were the source of help and encouragement for all these years. Many thanks to my dear aunts, uncle, grandpa and grandma who support me.

I especially thank my father, Bingzhen Xu; my mother, Yan Zhang; for their unconditional love and endless support through my entire life. Without their love and encouragement, I would not have completed my graduate program. I dedicate this thesis to them.

ABSTRACT

We consider two problems in high-dimensional inference. We first establish an invariance principle for quadratic forms of sample mean vectors under Lyapunov-type conditions that involve a delicate interplay between the dimension p, the sample size n and the moment condition. Under proper normalization, central and non-central limit theorems are obtained. The latter invariance principle is applied to test for mean vectors of high-dimensional data. To obtain cutoff values of our tests, we introduce a plug-in Gaussian multiplier calibration method and normalized consistency, a new matrix convergence criterion. We also propose a sub-sampling and a half-sampling procedures to approximate the distributions of the quadratic forms that do not need estimation of the underlying covariance matrices.

The second part deals with the estimation of time-varying networks from high-dimensional time series. Two types of non-stationarity are investigated: structural breaks and smooth changes. Our approach can achieve consistent detection of the change points and simultaneous estimation of the piece-wisely smooth-varying networks. Rates of convergence for estimating change point and networks are obtained under mild moment and dependence conditions.

CHAPTER 1

INTRODUCTION

The advances in data collection tools and computing technologies have given rise to the high dimensional data in many scientific fields. A wide range of classical statistical tools no longer work in the modern problems. For example, when the dimension is large compared to the sample size, the sample covariance matrix may not be a consistent estimator of the true covariance (see for example Marčenko and Pastur [1967], Bai and Silverstein [2010]); the inverse of the sample covariance matrix may not exist; the multivariate central limit theorem can be invalid (Portnoy [1986]). Despite the tremendous exploration and ever-increasing development in the high-dimensional models, real-world applications still induce considerable challenges to statistical methods and theory. In this thesis we deal with two problems in high-dimensional inference: the L^2-type test without assumptions on the covariance matrix, and the simultaneous estimation of dynamic networks using time-series data.

The classical Hotelling's T^2 multivariate mean test loses power in the high-dimensional context (c.f. Bai and Saranadasa [1996]) and can be even undefined. In seeking of solutions, researchers established asymptotic normality for various L^2 statistics under conditions that the largest eigenvalue of the covariance matrix should be much smaller than the Frobenius norm; see Bai and Saranadasa [1996], Chen and Qin [2010], Chen et al. [2011], to name a few. This rules out a remarkable amount of real-world applications such as the factor model and long-range dependence among the covariates. In Chapter 2, we develop an invariance principle for the distribution of the quadratic form of the sample mean, which imposes no condition of the true covariance. In linear process models, our asymptotic theory holds with no limit on the dimension, given that the sample size is sufficiently large.

In addition, we investigate into a plug-in calibration, a subsampling procedure and a half-sampling procedure to estimate the distribution, which involves the true covariance matrix as an unknown high-dimensional parameter. The latter two procedures avoid the estimation of the covariance matrix and are consistent as long as the conditions for the invariance principle holds.

In Chapter 3, we consider the joint estimation problem of time-varying networks, which is central in visualizing the relational information in many complicated high-dimensional systems. It is well known that edges in the undirected graphs have meaningful statistical interpretation corresponding to non-zero partial correlations, which corresponds to the non-zero entries in the inverse covariance matrix (precision matrix) (c.f. Lauritzen [1996], Peng et al. [2009]). A huge volume of literature is dedicated on estimating the high-dimensional static network from independent observations; see Meinshausen and Bühlmann [2006], Friedman et al. [2008], Yuan and Lin [2007], Banerjee et al. [2008] among many others. However, the assumption of time-invariance and time-independence is restrictive in practice. In fact, data collected from many fields such as economics, finance, geography and genetics is of non-stationary nature and has both cross-sectional and serial dependence. In addition, it can be heavy tailed, which is also neglected by most of the previous study.

Recent successful attempts allow the more flexible time-varying models (see Zhou et al. [2010], Kolar and Xing [2011], Kolar et al. [2010], Kolar and Xing [2014], Qiu et al. [2015], Lu et al. [2015], Ahmed and Xing [2009]). However, there are still a number of major challenges in the current high-dimensional literature such as abrupt breaks, temporal dependence and heavy tails.

we focus on recovering the time-varying undirected graphs based on regularized estimation of the precision matrices in Chapter 3. Our study allows two types of dynamics: unknown number of abrupt changes and gradual varies between the change points. In particular, we study high-dimensional *piecewise locally stationary processes* in a general nonlinear temporal dependency framework, where the observations are allowed to have only polynomial moments up to a finite order. Consistency of the change points estimation and uniform convergence rate of support recovery are developed.

CHAPTER 2

ASYMPTOTIC THEORY FOR HIGH-DIMENSIONAL L^2 STATISTICS AND RESAMPLING METHODS

2.1 Introduction

Multiple hypothesis testing is a basic problem in statistics. Let $X, X_i, i \in \mathbb{Z}$, be independent and identically distributed (i.i.d.) p-dimensional random vectors with mean $\mathbb{E}X_i = \mu$ and covariance matrix $\mathrm{Cov}(X_i) = \Sigma$. Given the sample $X_1, \ldots, X_n$ and a pre-specified vector μ_0, we aim to test the hypothesis

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0. \tag{2.1}$$

As a popular classical approach, one uses the Hotelling T^2 statistic

$$T_n = n(\bar{X}_n - \mu_0)^\top \hat{\Sigma}_n^{-1}(\bar{X}_n - \mu_0), \tag{2.2}$$

where $\bar{X}_n = \sum_{i=1}^n X_i/n$, $\hat{\Sigma}_n = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)^\top$ is the sample covariance matrix. In the high dimensional setting with $p > n$, $\hat{\Sigma}_n$ is singular and T_n is thus not well-defined. There is a large literature accommodating the Hotelling T^2 statistic into the high-dimensional situation; see for example, Bai and Saranadasa [1996], Chen and Qin [2010], Srivastava et al. [2013], Chen et al. [2011] among others. Most of the existing work assumes that X_i is Gaussian or has a linear form and central limit theorems are derived for various quadratic forms of the sample. For example, by Bai and Saranadasa [1996], one can prove that the central limit theorem holds for $n\bar{X}_n^\top \bar{X}_n - \mathrm{tr}(\hat{\Sigma}_n)$, where X_i has a linear form, assuming that p/n tends to a finite constant and the largest eigenvalue of Σ is negligible relative to its Frobenius norm. The latter condition can be violated in cases such as factor models, as discussed in Katayama et al. [2013], who studied the asymptotic distribution of

$Z^\top Z - \mathrm{tr}(\Sigma)$ over different types of Σ with $Z \sim N(0, \Sigma)$.

A primary goal of the chapter concerns the asymptotic distribution of $|\bar{X}_n - \mu|^2 = (\bar{X}_n - \mu)^\top (\bar{X}_n - \mu)$, where for a vector $\mathbf{x} = (x_1, \ldots, x_m)^\top$, the Euclidean norm $|\mathbf{x}| = |\mathbf{x}|_2 = (\mathbf{x}^\top \mathbf{x})^{1/2} = (\sum_{i=1}^m x_i^2)^{1/2}$. The asymptotic theory for $|\bar{X}_n - \mu|^2$ can be applied to test the hypothesis (2.1). Unless otherwise specified, assume throughout the chapter that $\mu = 0$. In the classical setting with fixed dimension p, due to the Central Limit Theorem, we have the distributional convergence $\sqrt{n}\bar{X}_n \Rightarrow N(0, \Sigma)$. Hence, letting $Y \sim N(0, \Sigma)$, we have by Slutsky's Theorem that

$$\sup_{u \in \mathbb{R}} |\mathbb{P}(n\bar{X}_n^\top \bar{X}_n \leq u) - \mathbb{P}(Y^\top Y \leq u)| \to 0. \tag{2.3}$$

In this chapter we shall discuss the validity of (2.3) in situations in which p can be unbounded. In modern problems, the dimension p can be larger than the sample size n. In this case, the traditional methods may not work. For example, Portnoy [1986] showed that the CLT $\sqrt{n}\bar{X}_n \Rightarrow N(0, \Sigma)$ is generally not valid when p is large such that $\sqrt{n} = o(p)$. However, interestingly, (2.3) can still be valid, by using different methods.

In this chapter, we shall develop an asymptotic theory for $\bar{X}_n^\top \bar{X}_n$ for a generally distributed X, without requiring normality or linearity assumption. In particular, we shall apply the normal comparison method of Stein type and show that $\bar{X}_n^\top \bar{X}_n$ can be approximated by a mixture of independent χ_1^2 distributions. Differently from the classical CLT results (cf. Bai and Saranadasa [1996], Chen and Qin [2010], Srivastava et al. [2013], Chen et al. [2011]), the approximate distribution here may or may not be asymptotically normal. Specifically, we shall establish the following analogous form of (2.3):

$$\sup_{u \in \mathbb{R}} |\mathbb{P}(n\bar{X}_n^\top \bar{X}_n \leq u) - \mathbb{P}(n\bar{Y}_n^\top \bar{Y}_n \leq u)| \to 0, \tag{2.4}$$

where $Y_i, i \in \mathbb{Z}$, are i.i.d. $N(0, \Sigma)$ random vectors and $\bar{Y}_n = \sum_{i=1}^n Y_i/n$. We can view

(2.4) as a *quadratic form invariance principle* since the distributions of quadratic functionals of non-Gaussian random vectors can be approximated by those of Gaussian vectors with the same covariance structure. The invariance principle in the traditional sense refers to the Gaussian approximation of partial sum processes of non-Gaussian random variables; cf Berkes et al. [2014].

As an immediate application of (2.3) or (2.4), one can perform test for the hypothesis (2.1). Given the significance level $\alpha \in (0,1)$, let $u_{1-\alpha}$ be the $(1-\alpha)$th quantile of $Y^\top Y$. Namely

$$\mathbb{P}(Y^\top Y \le u_{1-\alpha}) = 1 - \alpha. \tag{2.5}$$

Then H_0 is rejected if $n\bar{X}_n^\top \bar{X}_n > u_{1-\alpha}$. By (2.3), the latter test has an asymptotic level α.

If Σ is known, the cutoff value $u_{1-\alpha}$ can be easily computed, either numerically or analytically, since the distribution of $Y^\top Y$ is completely known. In most applications, however, Σ is not known. As a natural solution, one uses a covariance matrix estimate $\tilde{\Sigma}$ (say), generates $\tilde{Y} \sim N(0, \tilde{\Sigma})$ given $\tilde{\Sigma}$, and then obtains the estimated cutoff values $\tilde{u}_{1-\alpha}$. To evaluate the validity of this plug-in procedure, we shall introduce a new matrix convergence criterion: the *normalized consistency*. It is closely related, but different from the widely used spectral norm convergence. From modern random matrix theory, it is now well-known that the sample covariance matrix $\hat{\Sigma}_n$ is not a (spectral norm) consistent estimator of Σ when p is large; see Marčenko and Pastur [1967], Bai and Silverstein [2010], to name a few. However, our results indicate that in certain situations the sample covariance matrix can be normalized consistent, and hence the corresponding estimated cutoff value is consistent. Details are given in Section 2.3.1.

As another contribution, we introduce a subsampling technique to estimate the distribution function $F(u) = \mathbb{P}(n\bar{X}_n^\top \bar{X}_n \le u)$, which avoids estimating Σ or its eigenvalues. The subsampling empirical distribution function is defined as

$$\check{F}(u) = \frac{1}{J} \sum_{j=1}^{J} \mathbf{1}_{m|\bar{X}_{A_j} - \bar{X}|_2^2 \le u(1-m/n)}, \tag{2.6}$$

where $A_j, j = 1 \ldots, J$, are i.i.d. samples from the set $\{1, \ldots, n\}$ with cardinality $|A_j| = m$ and $\mathbf{1}_B$ is the indicator function for event B. The sampling process of the subsets $(A_j)_{j\ge 1}$ is independent of $(X_i)_{i\ge 1}$; see Section 2.3.2 for details. The subsampling procedure avoids estimating Σ or its eigenvalues. We shall show that the subsampling approach leads to a consistent estimate of the distribution function $F(\cdot)$. The only assumption needed is a Lyapunov-type condition for the invariance principle.

Another type of approach for testing (2.1) is to use the maximum or L^∞ norm $|\bar{X}_n|_\infty = \max_{j\le p} |\bar{X}_{nj}|$ or the studentized version $\max_{j\le p} |\bar{X}_{nj}|/\hat{\sigma}_j$, where $\hat{\sigma}_j^2$ are estimates for the marginal variances $\sigma_j^2 = \mathrm{var}(X_{ij})$. Kosorok and Ma [2007] considered the uniform consistency problem. Fan et al. [2007] performed the L^∞ test via Bonferroni correction, thus completely ignoring dependencies among entries of X_i. In a recent work, Chernozhukov et al. [2013] derived a Gaussian approximation for $|\bar{X}_n|_\infty$ in the high-dimensional setting. In comparison with the marginal testing procedures, the procedure in Chernozhukov et al. [2013] is dependence-adjusted. Liu and Shao [2013] established a deep Cramér-type moderate deviation principle for Hotelling's T^2 statistic under mild moment condition. The L^2-based test can be more powerful if the alternative consists of many small but non-zero signals that are of similar magnitudes. Assuming sparseness, Fan et al. [2013b] proposed a power boosting test procedure based on a modified quadratic form. Their method requires an L^2-type test statistic that has an asymptotic correct size.

This chapter is organized as follows. In Section 2.2, we present the invariance principle result. Section 2.3 provides a plug-in Gaussian multiplier calibration procedure, a subsampling procedure and a power analysis. In evaluating the plug-in method, we introduce *normalized* *consistency*, a new matrix convergence criterion. Section 2.6 assesses finite sample performance via a simulation. Section 2.7 contains a real data application. Proofs are given in Sections 2.8.

We now introduce some notation. Let X be a random vector. Write $X \in \mathcal{L}^q$, $q > 0$, if $\|X\|_q := (\mathbb{E}|X|^q)^{1/q} < \infty$. For a matrix $A = (a_{jk})_{j,k}$, $\rho(A) = \max_{\mathbf{x}} |A\mathbf{x}|/|\mathbf{x}|$ (resp. $|A|_F = (\sum_{jk} a_{jk}^2)^{1/2}$) denotes its spectral (resp. Frobenius) norm. Denote by C a positive constant whose value may vary from place to place.

2.2 An Invariance Principle

Our testing procedures are based on the invariance principle (2.3). In this section we shall present a rigorous theoretical result. Consider i.i.d. random vectors $X_i \in \mathbb{R}^p$, $i \in \mathbb{Z}$, with $\mathbb{E}X_i = \mathbf{0}$ and covariance matrix $\mathrm{Cov}(X_i) = \Sigma$. Let $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ be eigenvalues of Σ and

$$\Sigma = Q\Lambda Q^\top, \text{ where } \Lambda = \mathrm{diag}\left(\lambda_1, \ldots, \lambda_p\right), \tag{2.7}$$

be the eigen-decomposition of Σ. Here Q is an orthonormal matrix with $Q^\top Q = \mathrm{Id}_p$, the $p \times p$ identity matrix. Recall that $Y, Y_1, \ldots$ are i.i.d. $N(0, \Sigma)$ and $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$.

Our main result is Theorem 2.2.2 which asserts that under suitable conditions the distributions of quadratic forms $\bar{X}_n^\top \bar{X}_n$ and $\bar{Y}_n^\top \bar{Y}_n$ are asymptotically close. In our asymptotic relation, we let $n \to \infty$ and allow the dimension $p = p_n \to \infty$ as $n \to \infty$. To state the theorem, we need to impose the following condition on X.

Assumption 2.2.1. *Let $\delta > 0$. Assume that*

$$K_\delta(X)^{2+\delta} := \mathbb{E}\left|\frac{|X_1|_2^2 - \operatorname{tr}(\Sigma)}{|\Sigma|_F}\right|^{2+\delta} < \infty; \tag{2.8}$$

$$D_\delta(X)^{2+\delta} := \mathbb{E}\left|\frac{X_1^\top X_2}{|\Sigma|_F}\right|^{2+\delta} < \infty. \tag{2.9}$$

In Condition 2.2.1, (2.8) and (2.9) are the Lyapunov-type conditions for the distribution of the sample. Here $K_\delta(X)$ and $D_\delta(X)$ depend on the distribution of X. In the sequel for notational convenience we abbreviate them as K_δ and D_δ, respectively. Note that $D_0 = 1$. Proposition 2.2.1 shows that for Gaussian vectors we can have explicit upper bounds.

Proposition 2.2.1. *Let Y_i be i.i.d. $N(0,\Sigma)$ and $\delta \geq 0$. Then*

$$\mathbb{E}\left|\frac{Y_1^\top Y_1 - \operatorname{tr}(\Sigma)}{|\Sigma|_F}\right|^{2+\delta} \leq c_\delta^{2+\delta}; \tag{2.10}$$

$$\mathbb{E}\left|\frac{Y_1^\top Y_2}{|\Sigma|_F}\right|^{2+\delta} \leq d_\delta^{2+\delta}, \tag{2.11}$$

where $c_\delta = (1+\delta)^{1/2}\|\xi^2 - 1\|_{2+\delta}$, $d_\delta = (1+\delta)^{1/2}\|\xi\|_{2+\delta}^2$ and $\xi \sim N(0,1)$.

Based on (2.8) and (2.9), we have the following asymptotic result, which provides a rate of convergence of the invariance principle (2.4).

Theorem 2.2.2. *Assume that (2.8) and (2.9) hold with $0 < \delta \leq 1$. Then*

$$\sup_t \left|\mathbb{P}\left(n\bar{X}_n^\top \bar{X}_n \leq t\right) - \mathbb{P}\left(Y^\top Y \leq t\right)\right| = O(\psi_n^{-1/2}), \tag{2.12}$$

where ψ_n is the solution to the equation $L_\delta(n,\psi) = \psi^{-1/2}$ with

$$L_\delta(n,\psi) = \psi^2\left(\frac{\tilde{K}_0^2}{n} + \frac{\tilde{K}_0}{n^{1/2}}\right) + \psi^q\left[\frac{\tilde{K}_\delta^q}{n^{q-1}} + \frac{\mathbb{E}(X_1^\top \Sigma X_1)^{q/2}}{n^{\delta/2}|\Sigma|_F^q} + \frac{\tilde{D}_\delta^q}{n^\delta}\right],$$

where $q = 2+\delta$, $\tilde{K}_\delta = K_\delta + c_\delta$, $\tilde{D}_\delta = D_\delta + d_\delta$, and c_δ and d_δ are given in Proposition 2.2.1. In particular, we have $\psi_n \to \infty$ if

$$\frac{\tilde{K}_0^2}{n} + \frac{\tilde{K}_\delta^q}{n^{q-1}} + \frac{\mathbb{E}(X_1^\top \Sigma X_1)^{q/2}}{n^{\delta/2}|\Sigma|_F^q} + \frac{\tilde{D}_\delta^q}{n^\delta} \to 0 \text{ as } n \to \infty. \tag{2.13}$$

Consequently the left hand side of (2.12) converges to 0.

Remark 1. Since $X_1^\top \Sigma X_1 = \mathbb{E}(X_1^\top X_2 X_2^\top X_1 | X_1)$, by Jensen's inequality,

$$\mathbb{E}(X_1^\top \Sigma X_1)^{q/2} \le \mathbb{E}(|X_1^\top X_2 X_2^\top X_1|^{q/2}) = \mathbb{E}(|X_1^\top X_2|^q) = D_\delta^q |\Sigma|_F^q. \tag{2.14}$$

So (2.13) follows from $\tilde{K}_0^2/n + \tilde{K}_\delta^q/n^{q-1} + \tilde{D}_\delta^q/n^{\delta/2} \to 0$ as $n \to \infty$. Namely if n is sufficiently large such that $\tilde{K}_0^2 + \tilde{K}_\delta^{q/(q-1)} + \tilde{D}_\delta^{2q/\delta} = o(n)$, then the left hand side of (2.12) holds with rate $\psi_n^{-1/2} \to 0$. □

Remark 2. If conditions (2.8) and (2.9) hold with K_δ and D_δ bounded, then we can choose $\psi_n \asymp n^{\delta/(5+2\delta)}$ and the corresponding convergence rate in (2.12) is $O(n^{-\delta/(10+4\delta)})$. □

2.2.1 Central and non-Central Limit Theorems

Theorem 2.2.2 clarifies an important issue in the literature. In early studies, the primary focus is on the central limit theorem

$$R_n := \frac{n|\bar{X}_n|_2^2 - \mathrm{tr}(\Sigma)}{|\Sigma|_F} = \frac{n\bar{X}_n^\top \bar{X}_n - \mathrm{tr}(\Sigma)}{|\Sigma|_F} \Rightarrow N(0,2) \tag{2.15}$$

or its modified version (cf. (2.18)); see for example Bai and Saranadasa [1996], Srivastava [2009], Chen and Qin [2010]. Note that $Y^\top Y$ has the same distribution as $\sum_{j=1}^p \lambda_j \eta_j$, where

η_j are i.i.d. χ^2_1. Since $\mathbb{E}(\eta_j - 1)^2 = 2$, by Lindeberg's Central Limit Theorem,

$$V := \frac{Y^\top Y - \mathrm{tr}(\Sigma)}{|\Sigma|_F} =_{\mathcal{D}} \sum_{j=1}^{p} |\Sigma|_F^{-1} \lambda_j (\eta_j - 1) \Rightarrow N(0, 2) \tag{2.16}$$

holds if and only if $\lambda_1/|\Sigma|_F = \rho(\Sigma)/|\Sigma|_F \to 0$. Let $\zeta \sim N(0,1)$. By the Berry–Esseen theorem, the accuracy of the Gaussian approximation

$$\sup_t |\mathbb{P}(V \le t) - \mathbb{P}(\sqrt{2}\zeta \le t)| \le C_0 \sum_{j=1}^{p} (|\Sigma|_F^{-1} \lambda_j)^3 = C_0 \frac{\mathrm{tr}(\Sigma^3)}{|\Sigma|_F^3}, \tag{2.17}$$

where C_0 is an absolute constant. The ratio $\mathrm{tr}(\Sigma^3)/|\Sigma|_F^3$ quantifies the goodness of the Gaussian approximation. As a rule of thumb, we regard the Gaussian approximation reasonable if $\mathrm{tr}(\Sigma^3)/|\Sigma|_F^3 \le 0.1$. Under $\lambda_1/|\Sigma|_F \to 0$, by Theorem 2.2.2 and (2.16), (2.15) holds. In general, if $\lambda_1/|\Sigma|_F$ does not converge to 0, then R_n does not have an asymptotic Gaussian distribution. When the dependence between entries of X is strong, the asymptotic distribution of R_n can be non-normal. For example, suppose $Y \sim N(0, \Sigma)$ and Σ is Toeplitz with diagonal 1 and $\sigma_{j,k} \sim |k - j|^{-D}$ for some $0 < D < 1/2$ as $|k - j| \to \infty$. Then $(Y^\top Y - \mathrm{tr}(\Sigma))/|\Sigma|_F \Rightarrow \sum_{j=1}^{\infty} c_j(\eta_j - 1)$, the Rosenblatt distribution, with $c_j \sim c j^{D-1}$ as $j \to \infty$, and c is a constant; see Veillette and Taqqu [2013]. In this latter non-Gaussian limit case, a test based on $n\bar{X}_n^\top \bar{X}_n$ with cutoff values obtained from CLT can have a wrong size and a wrong power.

An Invariance Principle without Condition (2.8)

The expression R_n in (2.15) involves the trace $\mathrm{tr}(\Sigma)$. A natural estimate for the latter is $\mathrm{tr}(\hat{\Sigma}_n)$, where $\hat{\Sigma}_n = n^{-1} \sum_{i=1}^{n} X_i X_i^\top$. Replacing $\mathrm{tr}(\Sigma)$ by $\mathrm{tr}(\hat{\Sigma}_n)$, R_n becomes

$$\tilde{R}_n = \frac{\sum_{i \ne j \le n} X_i^T X_j}{(n-1)|\Sigma|_F}. \tag{2.18}$$

If $\mu = \mathbb{E}X_i \neq 0$, then based on i.i.d. vectors $X_1, \ldots, X_n$, the statistic $(n(n-1))^{-1} \sum_{i \neq j} X_i^T X_j$ is an unbiased estimate of the quantity $|\mu|_2^2 = \mu^\top \mu$, while the natural plug-in estimator $\bar{X}_n^\top \bar{X}_n$ is biased; see also Chen and Qin [2010]. Using the arguments in the proof of Theorem 2.2.2, without essential extra difficulties, we have the invariance principle result:

Corollary 2.2.3. *Assume Condition (2.9) and $\mu = 0$. Further assume*

$$L_\delta^\dagger := \frac{\mathbb{E}(X_1^\top \Sigma X_1)^{q/2}}{n^{\delta/2} |\Sigma|_F^q} + \frac{\tilde{D}_\delta^q}{n^\delta} \to 0. \tag{2.19}$$

Then $\psi_n := (L_\delta^\dagger)^{-1/(q+1/2)} \to \infty$ and, recall $V = \sum_{j=1}^p |\Sigma|_F^{-1} \lambda_j (\eta_j - 1)$,

$$\sup_t |\mathbb{P}(\tilde{R}_n \leq t) - \mathbb{P}\,(V \leq t)\,| = O(\psi_n^{-1/2}) \to 0. \tag{2.20}$$

By (2.14), a simple sufficient condition for (2.19) is $D_\delta^q = o(n^{\delta/2})$. Then the rate in (2.20) becomes $D_\delta^{q/(5+2\delta)} n^{-\delta/(10+4\delta)}$. Notice that in Corollary 2.2.3 Condition (2.8) is not needed since $\tilde{R}_n$ does not involve the diagonal terms $X_i^T X_i$. Consequently the weaker moment condition $X_i \in \mathcal{L}^{2+\delta}$ suffices. In comparison, (2.8) necessarily requires the stronger moment condition $X_i \in \mathcal{L}^{4+2\delta}$.

Optimality of the Conditions for the Invariance Principle

Condition (2.19) requires that, to ensure the invariance principle

$$\sup_t |\mathbb{P}(\tilde{R}_n \leq t) - \mathbb{P}\,(V \leq t)\,| \to 0, \tag{2.21}$$

the sample size n needs to be sufficiently large such that $L_\delta^\dagger \to 0$, namely

$$\frac{\mathbb{E}(X_1^\top \Sigma X_1)^{q/2}}{|\Sigma|_F^q} = o(n^{\delta/2}) \text{ and } \tilde{D}_\delta^q = o(n^\delta). \tag{2.22}$$

In general the condition $L_\delta^\dagger \to 0$ in (2.19) is not relaxable. Let $\ell = p^\beta$, $\beta > 1/2$, and let $B_{ij}, i, j \in \mathbb{Z}$, be i.i.d. Bernoulli($\ell^{-1}$) random variables; let $X_{ij} = (\ell B_{ij} - 1)(\ell - 1)^{-1/2}$. Then $\mathbb{E}X_{ij} = 0$, $\mathbb{E}X_{ij}^2 = 1$, $\mathbb{E}|X_{ij}|^q \sim \ell^{q/2-1}$, $\Sigma = \mathrm{Id}_p$ and $|\Sigma|_F^2 = p$. By Burkholder's inequality, $\mathbb{E}|X_1^\top X_1|^{q/2} \le c_q \mathbb{E}|\sum_{j=1}^p X_{1j}|^q$. By Rosenthal's inequality (Rosenthal [1970]), $\mathbb{E}|\sum_{j=1}^p X_{1j}|^q \le c_q(p\mathbb{E}|X_{11}|^q + p^{q/2})$ and $\mathbb{E}|X_1^\top X_2|^q \le c_q(p\mathbb{E}|X_{11}X_{21}|^q + p^{q/2})$. Then (2.19) requires that

$$\ell = o(np^{1/2}), \text{ or } p^{\beta-1/2} = o(n). \tag{2.23}$$

We remark that Condition (2.23) is also necessary for (2.21). By (2.21),

$$\frac{(n-1)|\Sigma|_F \tilde{R}_n}{np^{1/2}} = \frac{\sum_{l=1}^p Q_l}{np^{1/2}} \Rightarrow N(0,2), \text{ where } Q_l = \sum_{i \ne j \le n} X_{il}X_{jl}. \tag{2.24}$$

By the Linderberg-Feller central limit theorem, (2.24) holds if and only if

$$p\mathbb{E}\{[Q_1/(np^{1/2})]^2 \mathbf{1}_{|Q_1| \ge \theta np^{1/2}}\} = \mathbb{E}\{n^{-2}Q_1^2 \mathbf{1}_{|Q_1| \ge \theta np^{1/2}}\} \to 0 \tag{2.25}$$

holds for every $\theta > 0$. Note that $W := \sum_{i=1}^n B_{i1}$ is binomial(n, ℓ^{-1}). If

$$np^{1/2} = O(\ell), \tag{2.26}$$

then for all large n, the event $\{|Q_1| < \theta np^{1/2}\}$ implies $\{W \le 1\}$, and

$$\mathbb{E}\{n^{-2}Q_1^2 \mathbf{1}_{|Q_1| < \theta np^{1/2}}\} \le \mathbb{E}\{n^{-2}Q_1^2 \mathbf{1}_{W \le 1}\} \le \frac{n^2}{\ell^2} + \frac{n}{\ell} \to 0, \tag{2.27}$$

by noting that $\mathbb{E}\{n^{-2}Q_1^2 \mathbf{1}_{W=0}\} \le n^2\ell^{-2}$ and $\mathbb{E}\{n^{-2}Q_1^2 \mathbf{1}_{W=1}\} \le n\ell^{-1}$. Clearly (2.27) violates (2.25) since $n^{-2}\mathbb{E}Q_1^2 \to 2$. $\square$

2.2.2 High Dimensional Cramér-von Mises Test

The classical Cramér-von Mises test is widely used for testing distributions based on empirical distribution functions. Consider the problem of testing whether the sample (i.i.d. random variables taking values in $[0,1]$) $U_1, \ldots, U_n$ follow the standard uniform$(0,1)$ distribution. Then one can use the Cramér-von Mises statistic

$$\Omega = \int_0^1 [\hat{F}(u) - u]^2 du, \text{ where } \hat{F}(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{U_i \le u}, \tag{2.28}$$

or the following version based on (2.18) with diagonal elements deleted:

$$\begin{aligned} \Omega_0 &= \frac{1}{n(n-1)} \sum_{i \ne j \le n} \int_0^1 (\mathbf{1}_{U_i \le u} - u)(\mathbf{1}_{U_j \le u} - u) du \\ &= \frac{1}{n(n-1)} \sum_{i \ne j \le n} [\frac{U_i^2 + U_j^2}{2} - \max(U_i, U_j) + \frac{1}{3}]. \end{aligned} \tag{2.29}$$

Under the null hypothesis that U_i are i.i.d. uniform$(0,1)$, we have

$$n\Omega_0 \to \sum_{h=1}^{\infty} h^{-2}\pi^{-2}(\eta_h - 1), \text{ where } \eta_h \text{ are i.i.d. } \chi_1^2. \tag{2.30}$$

The above test statistic can be easily generalized to high dimensional vectors. Let $W_i = (U_{i1}, \ldots, U_{id})^\top$, $i = 1, \ldots, n$, be i.i.d. random vectors, $d \in \mathbb{N}$. Consider for example testing the hypothesis that U_{il}, $1 \le l \le d$, follow uniform$(0,1)$. Similarly as (2.29), we define

$$\Omega_0^W = \frac{1}{n(n-1)} \sum_{l=1}^{d} \sum_{i \ne j \le n} \int_0^1 (\mathbf{1}_{U_{il} \le u} - u)(\mathbf{1}_{U_{jl} \le u} - u) du. \tag{2.31}$$

The above statistic concerns testing marginal distributions. Similar statistics can be defined for testing joint distributions. Define

$$\Psi = \sum_{l=1}^{d} \int_0^1 (\mathbf{1}_{U_{1l} \le s} - s)(\mathbf{1}_{U_{2l} \le s} - s) ds \tag{2.32}$$

and $\psi^2 = \mathbb{E}(\Psi^2)$ with

$$\psi^2 = \sum_{l,l'=1}^{d} \int_0^1 \int_0^1 [\mathbb{P}(U_{1l} \le s, U_{1l'} \le s') - ss']^2 ds ds'. \tag{2.33}$$

Assume that n is sufficiently large such that

$$\mathbb{E}|\Psi/\psi|^{2+\delta} = o(n^{\delta/2}). \tag{2.34}$$

To apply Corollary 2.2.3, let K be a large integer and $\mathcal{A}_K = \{(k/K, l) : k = 1, \ldots, K; l = 1, \ldots, d\}$. Let $\mathcal{A} = \{(u, l) : 0 \le u \le 1; l = 1, \ldots, d\}$. For $\alpha = (u, l) \in \mathcal{A}$, let $X_{i\alpha} = \mathbf{1}_{U_{il} \le u} - u$, $\gamma_{\alpha\beta} = \mathbb{E}(X_{i\alpha} X_{i\beta})$ and the covariance matrix $\Sigma_K = (\gamma_{\alpha\beta})_{\alpha,\beta \in \mathcal{A}_K}$. Clearly, as $K \to \infty$, we have $K^{-1} \sum_{\alpha \in \mathcal{A}_K} X_{1\alpha} X_{2\alpha} \to \Psi$ and $K^{-2} \sum_{\alpha,\beta \in \mathcal{A}_K} \gamma_{\alpha\beta}^2 \to \psi^2$. For $\alpha = (u, l) \in \mathcal{A}$ and $\beta \in \mathcal{A}$, define the linear operator $\mathcal{L}$ by $\mathcal{L} f(\beta) = \sum_{l=1}^{d} \int_0^1 \gamma_{\alpha\beta} f(\alpha) du$. By Corollary 2.2.3, under condition (2.34), we have

$$\sup_t |\mathbb{P}(n\Omega_0^W \le t) - \mathbb{P}(\sum_{h=1}^{\infty} \lambda_h (\eta_h - 1) \le t)| \to 0, \tag{2.35}$$

where $\lambda_h, h \ge 1$, are the eigenvalues of the linear operator $\mathcal{L}$. In the special case with $d = 1$, under the hypothesis that $U_i \sim$uniform$(0, 1)$, we have $|\Psi| \le 1$ and (2.34) trivially holds. The covariance function $\phi(u, v) = \min(u, v) - uv$ and the linear operator $\mathcal{L} f(v) = \int_0^1 \phi(u, v) f(u) du$ has eigenvalues $\lambda_h = h^{-2}\pi^{-2}, h = 1, 2, \ldots$ (cf. (2.30)) and eigenfunctions $f_h(v) = \sin(hv\pi)$. For details see Chapter 5 in Shorack and Wellner [1986].

2.3 Estimating Distributions of Quadratic Forms

In Section 2.2 we shall present an asymptotic result for (2.3) (cf. Theorem 2.2.2). To apply (2.3) for testing the hypothesis $H_0 : \mu = 0$ (say) at level $\alpha \in (0, 1)$, we need to estimate the distribution of $Y^\top Y$ or compute $u_{1-\alpha}$, the $(1-\alpha)$th quantile of $Y^\top Y$. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ be the eigenvalues of the covariance matrix Σ. Then

$$Y^\top Y =_{\mathcal{D}} \sum_{j=1}^{p} \lambda_j \eta_j = V|\Sigma|_F + \operatorname{tr}(\Sigma), \tag{2.36}$$

where $\eta_1, \eta_2, \ldots,$ are i.i.d. χ_1^2 random variables, $=_{\mathcal{D}}$ denotes "equality in distribution", and

$$V = \sum_{j=1}^{p} |\Sigma|_F^{-1} \lambda_j (\eta_j - 1) \tag{2.37}$$

is the normalized version. Let $v_{1-\alpha}$ be the $(1-\alpha)$th quantile of V. Then $u_{1-\alpha} = v_{1-\alpha}|\Sigma|_F + \operatorname{tr}(\Sigma)$. In practice, however, Σ and hence λ_j are not known. To deal with this, we shall propose two re-sampling calibration procedures: a plug-in calibration and a subsampling approach, which are discussed in Sections 2.3.1 and 2.3.2, respectively, where asymptotic justification for the validity of those methods are also discussed.

2.3.1 A Plug-in Procedure

Let $\tilde{\Sigma}$ be an estimate of Σ based on the data $\mathbf{X}_n = (X_1, \ldots, X_n)$. Let $\tilde{\lambda}_1 \geq \ldots \geq \tilde{\lambda}_p \geq 0$ be the eigenvalues of $\tilde{\Sigma}$ and $|\tilde{\Sigma}|_F = (\sum_{j=1}^{p} \tilde{\lambda}_j^2)^{1/2}$. As (2.37) let $\tilde{V} = \sum_{j=1}^{p} |\tilde{\Sigma}|_F^{-1} \tilde{\lambda}_j (\eta_j^* - 1)$, where η_j^* are i.i.d. χ_1^2 random variables that are independent of $\mathbf{X}_n$. Let $\tilde{v}_{1-\alpha}$ be the $(1-\alpha)$th quantile of $\tilde{V}$, given $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_p$. Note that $\tilde{v}_{1-\alpha}$ can be obtained either numerically or by extensive simulation. Then we reject H_0 at level α if $\hat{R}_n := (n|\bar{X}_n|_2^2 - \widehat{\operatorname{tr}(\Sigma)})/\widehat{|\Sigma|_F} > \tilde{v}_{1-\alpha}$, where $\widehat{\operatorname{tr}(\Sigma)}$ and $\widehat{|\Sigma|_F}$ are estimates of $\operatorname{tr}(\Sigma)$ and $|\Sigma|_F$, respectively.

To ensure the validity of the above procedure, we need to impose suitable conditions so that the following requirements are met:

(i) the estimated quantile $\tilde{v}_{1-\alpha}$ is close to $v_{1-\alpha}$;

(ii) ratio consistency: $\widehat{|\Sigma|_F}/|\Sigma|_F \to 1$ in probability and $\widehat{\mathrm{tr}(\Sigma)} - \mathrm{tr}(\Sigma) = o_{\mathbb{P}}(|\Sigma|_F)$.

As will be discussed in Section 2.3.1, (i) requires that $\tilde{\Sigma}$ is a *normalized consistent* estimate of Σ in the sense that

$$\rho(\tilde{\Sigma}/|\tilde{\Sigma}|_F - \Sigma/|\Sigma|_F) = o_{\mathbb{P}}(1). \tag{2.38}$$

In Section 2.3.1 we shall also discuss its relation with the classical spectral norm convergence $\rho(\tilde{\Sigma} - \Sigma) = o_{\mathbb{P}}(1)$. Under structural assumptions such as bandedness or sparsity, various regularized procedures have been proposed so that the spectral norm consistency holds; see Wu and Pourahmadi [2003], Bickel and Levina [2008b,a], Fan et al. [2013a] among many others. We leave it as a future research problem on whether those regularized estimates are normalized consistent. In this chapter we shall consider (2.38) for sample covariance matrices.

For (ii), various estimators for $|\Sigma|_F$ exist in the literature. For example, Chen and Qin [2010] proved that $\widehat{|\Sigma|_F} = (\{n(n-1)\}^{-1}\mathrm{tr}[\sum_{j\neq k}^{n}(X_j - \bar{X}_{j,k})X_j^\top(X_k - \bar{X}_{j,k})X_k^\top])^{1/2}$ is ratio-consistent in the sense that $\widehat{|\Sigma|_F}/|\Sigma|_F - 1 = o_{\mathbb{P}}(1)$ for linear processes under suitable conditions, where $\bar{X}_{j,k} = (n-2)^{-1}\sum_{i\neq j,k} X_i$. See also Bai and Saranadasa [1996]. Note that $\mathrm{tr}(\Sigma)$ can be simply estimated by $\mathrm{tr}(\hat{\Sigma}_n)$, where $\hat{\Sigma}_n = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(X_i - \bar{X}_n)^\top$ is the sample covariance matrix. Assume $\mu = 0$ and write $\hat{\Sigma}_n = (n-1)^{-1}\sum_{i=1}^{n} X_i X_i^\top - (n-1)^{-1} n\bar{X}_n\bar{X}_n^\top$. If n is sufficiently large such that $\mathbb{E}(X_1^\top X_1 - \mathrm{tr}(\Sigma))^2 = o(n|\Sigma|_F^2)$, we have $\widehat{\mathrm{tr}(\Sigma)} - \mathrm{tr}(\Sigma) = o_{\mathbb{P}}(|\Sigma|_F)$.

Normalized Consistency

As a natural way to approximate the distribution of V, one can employ the plug-in procedure in Section 2.3.1. Let $\tilde{\lambda}_1 \geq \ldots \geq \tilde{\lambda}_p \geq 0$ be eigenvalues of the covariance matrix estimate $\tilde{\Sigma}$. By Lemma 2.3.1, if

$$\max_{j \leq p} ||\Sigma|_F^{-1} \lambda_j - |\tilde{\Sigma}|_F^{-1} \tilde{\lambda}_j| \to 0 \text{ in probability,} \tag{2.39}$$

then with probability converging to 1, we have

$$\sup_t |\mathbb{P}(V \leq t) - \mathbb{P}^*(\tilde{V} \leq t)| \to 0, \tag{2.40}$$

where $\mathbb{P}^*$ is the conditional probability given $\mathbf{X}_n$. With (2.40), the distribution of V can be approximated by that of $\tilde{V}$.

Lemma 2.3.1. *Let $a_{p,1} \geq a_{p,2} \geq \ldots \geq a_{p,p} \geq 0$ and $b_{p,1} \geq b_{p,2} \geq \ldots \geq b_{p,p} \geq 0$ be two sequences of real numbers satisfying $\sum_{j=1}^{p} a_{p,j}^2 = \sum_{j=1}^{p} b_{p,j}^2 = 1$. Assume $\max_{j \leq p} \left|a_{p,j} - b_{p,j}\right| \to 0$ as $p \to \infty$. Let η_j be i.i.d. χ_1^2 random variables and $\eta_j' = \eta_j - 1$. Let $V_a = \sum_{j=1}^{p} a_{p,j} \eta_j'$ and $V_b = \sum_{j=1}^{p} b_{p,j} \eta_j'$. Then as $p \to \infty$,*

$$\sup_x \left|\mathbb{P}\left(V_a \leq x\right) - \mathbb{P}\left(V_b \leq x\right)\right| \to 0. \tag{2.41}$$

Interestingly, there is a simple sufficient condition for (2.39). By Weyl's theorem (Golub and Van Loan [2013, Theorem 8.1.5]), (2.39) follows from

$$\rho(\tilde{\Sigma}/|\tilde{\Sigma}|_F - \Sigma/|\Sigma|_F) = o_{\mathbb{P}}(1). \tag{2.42}$$

We say that an estimate $\tilde{\Sigma}$ of Σ is *normalized consistent* if (2.42) holds. It is closely related to, but different from the classical definition of spectral norm consistency in the sense of

$$\rho(\tilde{\Sigma} - \Sigma) = o_{\mathbb{P}}(1). \tag{2.43}$$

Normalized consistency does not generally imply the spectral norm consistency (2.43). For example, let $n = p$ and X_i be i.i.d. standard $N(0, \mathrm{Id}_p)$ random vectors. By the random matrix theory, (2.43) does not hold for $\hat{\Sigma}_n$, which is not a consistent estimate of $\Sigma = \mathrm{Id}_p$; see Marčenko and Pastur [1967], Wachter [1978], Geman [1980]. Indeed, the largest eigenvalue of $\hat{\Sigma}_n$ converges to 4, while the smallest one converges to 0. However the normalized consistency (2.42) holds since both $\rho(\Sigma/|\Sigma|_F) = p^{-1/2} \to 0$ and $\rho(\hat{\Sigma}_n/|\hat{\Sigma}_n|_F) = O_{\mathbb{P}}(p^{-1/2}) \to 0$. Without further conditions, the spectral norm consistency (2.43) does not imply the normalized consistency either. Proposition 2.3.2 relates these two types of convergence.

Proposition 2.3.2. *For an estimate $\tilde{\Sigma}$ of Σ, assume that $|\tilde{\Sigma}|_F/|\Sigma|_F \to 1$ in probability. Then the normalized consistency (2.42) holds if and only if $\rho(\tilde{\Sigma} - \Sigma) = o_{\mathbb{P}}(f)$.*

We now apply Theorem 2.2.2 and approximate the $(1-\alpha)$th quantile of $Y^\top Y$ by that of $\sum_{j=1}^{p} \tilde{\lambda}_j \tilde{\eta}_j$ (cf. (2.36)). Assuming $|\tilde{\Sigma}|_F/|\Sigma|_F \to 1$ in probability, $\mathrm{tr}(\tilde{\Sigma}) - \mathrm{tr}(\Sigma) = o_{\mathbb{P}}(|\Sigma|_F)$ and normalized consistency of $\tilde{\Sigma}$, we have by Lemma 2.3.1 that

$$\sup_t |\mathbb{P}(Y^\top Y \le t) - \mathbb{P}^*(\tilde{\lambda}_1 \tilde{\eta}_1 + \ldots + \tilde{\lambda}_p \tilde{\eta}_p \le t)| \to 0 \tag{2.44}$$

holds with probability converging to 1. The latter shows the validity of the plug-in method.

Theorem 2.3.3 concerns the normalized consistency of the sample covariance matrix. Assume $\mu = 0$ and let $\hat{\Sigma}_n = n^{-1} \sum_{i=1}^{n} X_i X_i^\top$.

Theorem 2.3.3. *Assume $\mathbb{E}[(X_1^\top X_1)^2] = o\left(n|\Sigma|_F^2\right)$. Then $|\hat{\Sigma}_n|_F/|\Sigma|_F \to 1$ in probability,*

$$\mathbb{E}|\hat{\Sigma}/|\hat{\Sigma}_n|_F - \Sigma/|\Sigma|_F|_F^2 = o(1), \tag{2.45}$$

which further implies the normalized consistency (2.42).

Theorem 2.3.3 requires that n is big enough such that $\mathbb{E}[(X_1^\top X_1)^2]/|\Sigma|_F^2 = o(n)$. The latter condition trivially holds if the entries of X_1 are strongly dependent in the sense that $|\Sigma|_F^2 = \sum_{j,k\le p}\sigma_{j,k}^2 \asymp p^2$ and, for some constant $C<\infty$, $\max_{j\le p}\|X_{1j}\|_4 \le C$. In this case $\mathbb{E}[(X_1^\top X_1)^2] \le p^2C^4$ and the condition $\mathbb{E}[(X_1^\top X_1)^2]/|\Sigma|_F^2 = o(n)$ reduces to the natural one $n\to\infty$. As a simple example, let $X_{1j} = a_jZ+\xi_j$, where $Z,\xi_1,\ldots,\xi_p$ are i.i.d. $N(0,1)$ and a_j's are real coefficients. If $\sum_{j=1}^p a_j^2 \asymp p$, then $\mathbb{E}[(X_1^\top X_1)^2]/|\Sigma|_F^2 \asymp 1$ and the condition $n\to\infty$ suffices. In this case Σ has $p-1$ eigenvalues 1 and 1 eigenvalue $1+\sum_{j=1}^p a_j^2$, hence $V \Rightarrow \chi_1^2 - 1$. Let $a_j = p^{-\delta}$ with $0<\delta\le 1/4$. Then $|\Sigma|_F^2 = (p^{1-2\delta}+1)^2+p-1 \asymp p^{2-4\delta}$ and $\mathbb{E}[(X_1^\top X_1)^2] \asymp p^2$. If n is sufficiently large such that $p^{4\delta} = o(n)$, then $\mathbb{E}[(X_1^\top X_1)^2] = o\left(n|\Sigma|_F^2\right)$ and Theorem 2.3.3 is applicable. On the other hand, if $n = o(p^{4\delta})$, then elementary but tedious calculations show that $\lim_{p\to\infty}\rho(\Sigma/|\Sigma|_F)>0$, while $\operatorname{tr}(\hat\Sigma_n^4) = o_{\mathbb{P}}(|\hat\Sigma_n|_F^4)$. Hence (2.39) does not hold for the sample covariance matrix estimate, suggesting that the condition $\mathbb{E}[(X_1^\top X_1)^2] = o\left(n|\Sigma|_F^2\right)$ in Theorem 2.3.3 is sharp. We shall leave it as a future research problem whether regularized covariance matrix estimators (cf. Bickel and Levina [2008b,a], Fan et al. [2013a]) are normalized consistent in the setting where $\rho(\Sigma) = o(|\Sigma|_F)$ is violated.

2.3.2 A Subsampling Procedure

Estimation of high-dimensional covariance matrices or their eigenvalues is highly nontrivial. In this section we shall provide a subsampling approach which can avoid such estimation problems. For a subset $A\subset\{1,2,\ldots,n\}$, let $|A|$ be its cardinality and $\bar X_A = \sum_{a\in A}X_a/|A|$. Let $|A| = m$ and $r = m/n$. Note that $\sqrt{m}(\bar X_A - \bar X_n)$ has mean 0 and covariance matrix $(1-r)\Sigma$, while $\sqrt{n}(\bar X_n-\mu)$ has mean 0 and covariance matrix Σ. This motivates us to introduce the empirical cumulative distribution functions (2.46) and (2.47) and the following subsampling based procedure for estimating the distribution function $F(t) = \mathbb{P}(n|\bar X_n-\mu|^2\le t)$:

1. Let $m = m_n \in \mathbb{N}$ be such that $m \to \infty$ and $m = o(n)$; let $A_1, \ldots, A_J$ be i.i.d. uniformly sampled from the class $\mathcal{A} := \{A : A \subset \{1, \ldots, n\}, |A| = m\}$. Assume that the sampling process $(A_j)_{j \geq 1}$ is independent of $(X_i)_{i \geq 1}$. Define

$$\hat{F}_r(t) = \frac{1}{J} \sum_{j=1}^{J} \mathbf{1}_{m|\bar{X}_{A_j} - \bar{X}|_2^2 \leq t(1-r)}. \tag{2.46}$$

   Obtain the $(1-\alpha)$th empirical quantile $\hat{F}_r^{-1}(1-\alpha)$ from (2.46).

2. Reject $H_0 : \mu = 0$ at level α if $n\bar{X}_n\bar{X}_n > \hat{F}_r^{-1}(1-\alpha)$.

See Politis et al. [1999] for a detailed discussion of subsampling techniques in the low dimensional setting, where it can serve as an alternative to bootstrap when the latter is inconsistent. As a slightly different version, one can also employ the following procedure. Let the index set $B_j = \{l \in \mathbb{Z} : (j-1)m < l \leq jm\}$, $j = 1, \ldots, L$, where $L = \lfloor n/m \rfloor$ and $\lfloor u \rfloor = \max\{k \in \mathbb{Z} : k \leq u\}$. Define the subsampling empirical distribution function

$$\check{F}_r(t) = \frac{1}{L} \sum_{j=1}^{L} \mathbf{1}_{m|\bar{X}_{B_j} - \bar{X}|_2^2 \leq t(1-r)}. \tag{2.47}$$

Based on the invariance principle result (2.3), we can show that the subsampling approach is asymptotically valid; see Theorem 2.3.4 for details. The subsample proportion r is a tuning parameter. Theoretically we only require $m \to \infty$ and $r \to 0$.

Validity of the Subsampling Procedure

In this section we shall study the validity of the subsampling method introduced in Section 2.3.2. Theorem 2.3.4 shows that the empirical distribution functions defined in (2.46) and (2.47) both converge to the distribution function $F(t) = \mathbb{P}(n|\bar{X} - \mu|_2^2 \leq t)$.

Theorem 2.3.4. *Let $0 < \delta \le 1$. Assume (2.8), (2.9), $m \to \infty$, $m = o(n)$, and (2.13) holds with n therein replaced by m. Then (i)*

$$\sup_t |\check{F}_r(t) - \mathbb{P}(n|\bar{X} - \mu|_2^2 \le t)| \to 0 \text{ in probability.} \tag{2.48}$$

(ii) If $J \to \infty$, then the convergence (2.49) also holds for $\hat{F}_r(t)$.

Theorem 2.3.4 suggests that sample quantiles of $\hat{F}_r(\cdot)$ or $\check{F}_r(\cdot)$ can be used to approximate those of $F(t) = \mathbb{P}(n|\bar{X} - \mu|_2^2 \le t)$. Given a level $\alpha \in (0, 1)$, let $\check{u}_{1-\alpha}$ be the $(1-\alpha)$th quantile of $\check{F}_r(\cdot)$. Then at level α we can reject the null hypothesis $H_0 : \mu = 0$ if $n|\bar{X}|_2^2 \ge \check{u}_{1-\alpha}$.

2.3.3 A Half-Sampling Procedure

A particularly interesting special case of subsampling is when $m = \lfloor n/2 \rfloor$, which in this thesis is called half-sampling. Half-sampling can be dated back to as early as the 1940s, when Mahalanobis proposed the idea in the context of sample survey (Mahalanobis [1946]). Early contributors include McCarthy [1969], Hartigan [1969, 1975] among others. A brief review can be found in Hall [2003]. In the one dimensional case $p = 1$ with asymptotic Gaussian limit, Wu [1990] showed that, with sample size n, the error bound of the delete-d jackknife histogram estimate has the form $O(d^{-1/2} + (n - d)^{-1/2})$. The latter bound is minimized when $d = O(n)$ and $n - d = O(n)$. If we take $d = n/2$, the latter is equivalent to half-sampling. In the literature, half-sampling is shown to be capable in estimating sampling variance (Shao and et al. [1989]), constructing the confidence sets and performing hypothesis test (Wu [1990]); see also Shao and Tu [2012]. These conclusions, however, are only shown to hold when the underlying limiting distribution is normal. In the high-dimensional setting the asymptotic distribution can be non-normal or even non-existence.

In the rest of this subsection we denote $\hat{F}(\cdot) \equiv \hat{F}_{1/2}(\cdot)$, where $\hat{F}_r(\cdot)$ is defined in (2.46).

Theorem 2.3.5. *Let $0 < \delta \le 1$. Assume (2.8), (2.9), and (2.13) holds. Let $J \to \infty$. Then*

$$\sup_t |\hat{F}(t) - \mathbb{P}(n|\bar{X} - \mu|_2^2 \le t)| \to 0 \text{ in probability.} \tag{2.49}$$

The half-sampling scheme can be viewed as a balanced Rademacher multiplier bootstrap, namely, $\hat{F}(t)$ is the empirical distribution of $T_{BR}^\top T_{BR}$ conditional on $(X_i)_{i=1}^n$, where

$$T_{BR} = n^{-1} \sum_{i=1}^n \epsilon_i X_i, \tag{2.50}$$

where $\epsilon_i, i = 1, \dots, n$ are Rademacher random variables that are independent of the samples and satisfy the constraint $\sum_{i=1}^n \epsilon_i = 0$.

Below we briefly describe the intuition for the half-sampling procedure based on Hadamard matrices. A formal mathematical argument is postponed to Section 2.8. At the outset assume that the data is multivariate Gaussian and n is a multiple of 4, namely $n = 4\lfloor n/4 \rfloor$. For $j = 1, \dots, J$, denote

$$T_j = \sqrt{r(1-r)^{-1}}(\bar{X}_{A_j} - \bar{X}). \tag{2.51}$$

Define

$$\kappa_j(t) = \mathbb{I}\left\{|T_j|_2^2 \le t\right\} - F(t),$$

Fix A_j, the random vector T_j has mean 0 and covariance $n^{-1}\Sigma$ for all $1 \le j \le J$, thus having same distribution as T_0. Let $H = (h_{jk}) \in \mathbb{R}^{n \times n}$ be a *Hadamard matrix*, namely its entries are either 1 or -1, and $HH^\top = n\mathrm{Id}_n$. See Hedayat et al. [1978], Georgiou et al. [2003], Yarlagadda and Hershey [2012] for a detailed discussion of Hadamard matrices. We fix the first column to be all 1s. Let $J = n - 1$. Consider the situation where we force the half-samples to be balanced by letting the jth realization of $(\epsilon_1, \dots, \epsilon_n)$ in (2.50) both take values of h_j for $j = 1, 2, \ldots, J$, where h_j is the $(j-1)$th column of H. Then $h_j^\top h_k = 0$ if $j \neq k$ and $h_j^\top h_j = n$. Then vectors $T_j = n^{-1} X^\top h_j$, $j = 1, \ldots, J$ are i.i.d. $N(0, n^{-1}\Sigma)$, where $X = (X_1, \ldots, X_n)^\top$. Hence (2.63) easily follows from the classical Glivenko-Cantelli theorem on uniform consistency for empirical distribution functions.

In combinatorial design the problem of constructing Hadamard matrices is still open. It is unclear whether Hadamard matrices exist for any n which is a multiple of 4. As a simple way out, our half-sampling scheme can achieve approximate orthogonality in the sense that $\boldsymbol{\epsilon}_j^\top \boldsymbol{\epsilon}_k / n$ is close to the delta function $\delta_{jk} = \mathbb{I}\{j = k\}$, where $\boldsymbol{\epsilon}_j = (\epsilon_{1,j}, \ldots, \epsilon_{n,j})$, $j = 1, \ldots, J$ and $\epsilon_{i,j} = 2\mathbb{I}\left\{i \in A_j\right\} - 1$. The latter implies that the covariance between $\mathbb{I}\left\{T_j' T_j \leq t\right\}$ and $\mathbb{I}\left\{T_k' T_k \leq t\right\}$ is generally small for $j \neq k$. To see this, for given sets $A_1, A_2 \in \mathcal{A}$ with $m = n/2$, let

$$H_1 = A_1 \cap A_2,\ H_2 = A_1 \cap A_2^c,\ H_3 = A_1^c \cap A_2,\ H_4 = A_1^c \cap A_2^c.$$

Then $|H_1| = |H_4|$, $|H_2| = |H_3| = n/2 - |H_1|$. Let

$$\delta = (4|H_1| - n_1)/n_1.$$

In the appendix, we show that

$$\sup_t \left| \mathbb{P}\left(R_0^{(1)} \leq t, R_0^{(2)} \leq t\right) - \mathbb{P}^2\left(R_0^{(1)} \leq t\right) \right| \leq C\left(\|\delta_1\|_2^{2/5} + \|\delta_2\|_2^{2/5}\right), \tag{2.52}$$

where

$$R_0^{(j)} = \frac{n|\bar{\xi}_{A_j} - \bar{\xi}|_2^2 - f_1}{f}. \tag{2.53}$$

If δ is small, meaning approximate orthogonality that $|H_1| \approx n/4$, (2.52) suggests that for Gaussian data $\mathbb{I}\{T_j' T_j \leq t\}$ and $\mathbb{I}\left\{T_k' T_k \leq t\right\}$ are asymptotically independent. For half-sampling or Rademacher schemes, we indeed have

$$\delta = O_P(n^{-1/2}).$$

The above argument does not apply to non-Gaussian data directly. To this end, we present joint Gaussian approximation under Condition in Section 2.8. With the latter results, the validity of the half-sampling approach is guaranteed for non-Gaussian data.

2.4 Extension of Half-Sampling to Two-Sample Problem

The half-sampling procedure can be extended to the two-sample hypothesis testing problem of comparing the mean vectors of high-dimensional data. Let $X_i, i \in \mathbb{Z}$, (resp. $Y_i, i \in \mathbb{Z}$) be i.i.d. p-dimensional random vectors with mean $\boldsymbol{\mu}_X$ (resp. $\boldsymbol{\mu}_Y$) and covariance matrix Σ_X (resp. Σ_Y). Assume that $(X_i)_{i\in\mathbb{Z}}$ and $(Y_i)_{i\in\mathbb{Z}}$ are also independent. Given the observations $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$, we are interested in testing the hypothesis

$$H_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y \quad \text{v.s.} \quad H_A : \boldsymbol{\mu}_X \neq \boldsymbol{\mu}_Y. \tag{2.54}$$

The above testing problem appears frequently in many applied areas. In our setting the dimension p is allowed to be much larger than the total sample size $n = n_1 + n_2$. Let $\bar{X} = n_1^{-1} \sum_{i=1}^{n_1} X_i$ and $\bar{Y} = n_2^{-1} \sum_{i=1}^{n_2} Y_i$ be the sample mean estimates of $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$, respectively.

We consider the distribution of the test statistic

$$Q = T^\top T, \text{ where } T = \bar{X} - \bar{Y}, \tag{2.55}$$

and reject H_0 if Q exceeds some cutoff value. Let the cumulative distribution function

$$F(t) = \mathbb{P}(T_0^\top T_0 \le t), \text{ where } T_0 = \bar{X} - \mu_X - (\bar{Y} - \mu_Y). \tag{2.56}$$

For $\alpha \in (0,1)$ let $u_{1-\alpha}$ be the $(1-\alpha)$th quantile of F, namely $F(u_{1-\alpha}) = 1-\alpha$. Then we can reject H_0 at level α if $Q > u_{1-\alpha}$. The fact that Σ_X and Σ_Y may be unequal induces challenge in the hypothesis test. In this section we extend the half-sampling method into the two-sample scenario.

We first fix the notation before proceeding. Denote

$$\bar{\Sigma} = \frac{\Sigma_X}{n_1} + \frac{\Sigma_Y}{n_2},$$

which is the covariance matrix of $T = \bar{X} - \bar{Y}$. Let $\bar{f}_1 = \operatorname{tr}(\bar{\Sigma})$ and $\bar{f} = |\bar{\Sigma}|_F$. Let $\xi_i \sim N(\boldsymbol{\mu}_X, \Sigma_X)$, $\zeta_j \sim N(\boldsymbol{\mu}_Y, \Sigma_Y)$, and they are mutually independent for $i = 1, \cdots, n_1$, $j = 1, \cdots, n_2$. Let $\bar{\xi}$ and $\bar{\zeta}$ be the average of $(\xi_i)_{1\le i\le n_1}$ and $(\zeta_i)_{1\le i\le n_2}$ respectively.

For a positive integer n denote the set $[n] = \{1, 2, \ldots, n\}$. Let the classes $\mathcal{A}^\circ := \{A : A \subset [n_1], |A| = m_1\}$ and $\mathcal{B}^\circ := \{B : B \subset [n_2], |B| = m_2\}$. For $A \subset [n_1]$, let $S_A^X = \sum_{a\in A} X_a$ and $\bar{X}_A = S_A/|A|$. The quantities S_B^Y and $\bar{Y}_B$ are similarly defined. In the subsampling procedure, we take $m_1 = \lfloor rn_1 \rfloor, m_2 = \lfloor rn_2 \rfloor$ as the subsample sizes for the two samples respectively, where $r \in (0,1)$ is the subsample proportion. The subsampling estimator of $F(t) = \mathbb{P}(T_0^\top T_0 \le t)$ is

$$F_r(t) = \binom{n_1}{m_1}^{-1} \binom{n_2}{m_2}^{-1} \sum_{A\in\mathcal{A}^\circ, B\in\mathcal{B}^\circ} \mathbb{I}\left\{T(A,B)^\top T(A,B) \le t\right\}, \tag{2.57}$$

where $\mathbb{I}\{\cdot\}$ stands for the indicator function.

For practical implementation, since the denominator in (2.57) can be very large, we shall apply random sampling to approximate $F_r(\cdot)$. Let $A_1, \ldots, A_J$ and $B_1, \ldots, B_J$, independent of $(X_i)_{i\geq 1}$ and $(Y_i)_{i\geq 1}$, be i.i.d. uniformly sampled from $\mathcal{A}^\circ$ and $\mathcal{B}^\circ$, respectively, where J is a large positive integer. For each $1 \leq j \leq J$, let

$$T_{r,j} = T_r(A_j, B_j) := \sqrt{r(1-r)^{-1}}[\bar{X}_{A_j} - \bar{Y}_{B_j} - (\bar{X} - \bar{Y})]. \tag{2.58}$$

Then we estimate $F(t)$ by

$$\hat{F}_r(t) = \frac{1}{J}\sum_{j=1}^{J} \mathbb{I}\left\{T'_{r,j}T_{r,j} \leq t\right\}, \tag{2.59}$$

To test (2.54), we reject the null hypothesis H_0 at level $\alpha \in (0,1)$ if $T^\top T = (\bar{X}-\bar{Y})^\top(\bar{X}-\bar{Y})$ is larger than the $(1-\alpha)$th empirical quantile of $\hat{F}_r(t)$. In practice where n is moderately large, the rounding down procedure of $m_i = \lfloor n_i r \rfloor, i = 1, 2$ may induce additional error to the subsampling procedure such that $m_1/n_1 \neq m_2/n_2$. In this sense half-sampling is preferred. Again, we denote the half-sampling estimator of (2.56) as $\hat{F}(t) \equiv \hat{F}_{1/2}(t)$.

Now we present theoretical results for the half-sampling procedure without assuming $\Sigma_X = \Sigma_Y$. Let $\lambda_1^X \geq \cdots \geq \lambda_p^X$ and $\lambda_1^Y \geq \cdots \geq \lambda_p^Y$ be the eigenvalues of Σ_X and Σ_Y, respectively. Let $f_1^X = \sum_{j=1}^p \lambda_j^X = \mathrm{tr}(\Sigma_X)$ and $f^X = (\sum_{j=1}^p (\lambda_j^X)^2)^{1/2} = (\mathrm{tr}(\Sigma_X^2))^{1/2}$. Similarly we define f_1^Y and f^Y. In the rest of this section we assume that $\mu_X = \mu_Y = 0$.

Assumption 2.4.1. *Let $0 < \delta \leq 1$ be a constant and $q = 2 + \delta$. Assume that*

$$\begin{aligned}
K_{X,q}^q &:= \mathbb{E}\left|\frac{|X_1|_2^2 - f_1^X}{f^X}\right|^q < \infty; \\
K_{Y,q}^q &:= \mathbb{E}\left|\frac{|Y_1|_2^2 - f_1^Y}{f^Y}\right|^q < \infty; \\
D_{X,q}^q &:= \mathbb{E}\left|\frac{X_1^\top X_2}{f^X}\right|^q < \infty; \\
D_{Y,q}^q &:= \mathbb{E}\left|\frac{Y_1^\top Y_2}{f^Y}\right|^q < \infty; \\
D_{XY,q}^q &:= \mathbb{E}\left|\frac{X_1^\top Y_1}{\sqrt{f^X f^Y}}\right|^q < \infty.
\end{aligned} \tag{2.60}$$

For notational simplicity, we drop the subscript q. Let

$$\tilde{K}_\delta = \max(K_X,\, K_Y,\, c_q) \text{ and } \tilde{D}_\delta = \max(D_X,\, D_Y,\, D_{XY},\, d_q).$$

We use $\tilde{K}_0$ to denote $\tilde{K}_\delta$ with $\delta = 0$. Note again for the linear process model, there exists an upper bound for the latter quantities that does not depend on its loading matrices.

Theorem 2.4.1. *Let Condition 2.4.1 be satisfied and $\delta = q - 2$. Assume that*

$$\begin{aligned} \tilde{K}_0(n_1^{-1/2} + n_2^{-1/2}) &\to 0, \\ \tilde{K}_\delta^q(n_1^{-(1+\delta)} + n_2^{-(1+\delta)}) &\to 0, \\ \tilde{D}_\delta^q(n_1^{-\delta} + n_2^{-\delta}) &\to 0, \end{aligned} \tag{2.61}$$

and

$$n_1^{1-q}\mathbb{E}(X_1^\top \bar{\Sigma} X_1)^{q/2} + n_2^{1-q}\mathbb{E}(Y_1^\top \bar{\Sigma} Y_1)^{q/2} = o(\bar{f}^q). \tag{2.62}$$

Then we have that

$$\sup_t |\hat{F}(t) - F(t)| \to_{\mathbb{P}} 0. \tag{2.63}$$

2.5 Power Analysis

In this section we shall present an asymptotic expression for the power function of the test based on Theorem 2.2.2. Assume that $X_1, \ldots, X_n$ are i.i.d. with mean $\mu = \mathbb{E}X_i$ and covariance matrix Σ. Based on Theorem 2.2.2, we reject the null hypothesis $H_0 : \mu = 0$ at level $\alpha \in (0, 1)$ if $n\bar{X}_n^\top \bar{X}_n > u_{1-\alpha}$, where the cutoff value $u_{1-\alpha}$ is defined in (2.5). Let

$Y \sim N(0, \Sigma)$ and define

$$g_\alpha(\mu) = \mathbb{P}\big((Y + \sqrt{n}\mu)^\top (Y + \sqrt{n}\mu) \geq u_{1-\alpha}\big). \tag{2.64}$$

A similar argument as Theorem 2.2.2 reveals that under (2.13), we have

$$|\mathbb{P}\big(n\bar{X}_n^\top \bar{X}_n \geq u_{1-\alpha}\big) - g_\alpha(\mu)| = o(1). \tag{2.65}$$

Note that $v_{1-\alpha} = |\Sigma|_F^{-1}\big(u_{1-\alpha} - \mathrm{tr}(\Sigma)\big) = O(1)$ and $\mathbb{E}(\mu^\top Y)^2 = \mu^\top \Sigma \mu \leq \rho(\Sigma)\mu^\top \mu$. Elementary manipulations show that the asymptotic power function $g_\alpha(\mu) \to \alpha$ if $n\mu^\top \mu / |\Sigma|_F \to 0$, and $g_\alpha(\mu) \to 1$ if $n\mu^\top \mu / |\Sigma|_F \to \infty$.

If we were to apply the central limit theorem (2.15) or its variants (see for example Bai and Saranadasa [1996], Srivastava [2009], Chen and Qin [2010]) to perform the test, one may obtain an erroneous power. Specifically, with the CLT (2.15), instead of the cutoff value $u_{1-\alpha}$, one use $\mathrm{tr}(\Sigma) + \sqrt{2}|\Sigma|_F z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1-\alpha)$th quantile of the standard Gaussian distribution. Then one obtains the asymptotic power expression

$$g_\alpha^*(\mu) = \mathbb{P}\big((Y + \sqrt{n}\mu)^\top (Y + \sqrt{n}\mu) \geq \mathrm{tr}(\Sigma) + \sqrt{2}|\Sigma|_F z_{1-\alpha}\big). \tag{2.66}$$

As pointed out in the discussion following Theorem 2.2.2, the difference between $g_\alpha^*(\mu)$ and the asymptotically correct power $g_\alpha(\mu)$ can be substantial if entries in X_i are strongly dependent. As a numerical example, let $\alpha = 0.01$, $n = 100$, $p = 100$, $\Sigma = \mathrm{Id}_p + \mathbf{1}\mathbf{1}^\top$, where $\mathbf{1} = (1, 1, \cdots, 1)^\top$, and $\mu = a\mathbf{1}$. For $a = 0.2$, we have $g_\alpha^*(\mu) = 0.572$, while the correct power $g_\alpha(\mu) = 0.278$, and $a = 0.02$, we have $g_\alpha^*(\mu) = 0.074$, while $g_\alpha(\mu) = 0.011$.

2.6 A Simulation Study

In this section we shall provide a simulation study for the finite sample performances of the invariance principle Theorem 2.2.2 (Oracle hereafter), the plug-in and the subsampling procedures described in Sections 2.3.1 and 2.3.2, respectively. We also compare their performance with the methods proposed by Gregory et al. [2015] and Srivastava et al. [2013] (GCT and SK hereafter, respectively). We consider the following two data generating models.

Model 1 (Linear Process Model): Let $\xi_{i,k}, i, k \in \mathbb{Z}$ are i.i.d. Student t_5; let

$$X_{i,j} = \sum_{k=0}^{\infty} (k+1)^{-\beta} \xi_{i,j-k}, \text{ where } \beta > 1/2, 1 \le i \le n, 1 \le j \le p. \tag{2.67}$$

If $\beta < 1$, then the process $(X_{i,j})_j$ is long memory, thus having a strong cross-sectional dependence. In our simulations we choose $p = 200$ and $n = 20, 200$ and truncate the sum in (2.67) to $\sum_{k=0}^{2000}$. We consider two levels of β: $\beta = 2$ and $\beta = 0.6$, which correspond to short and long memory, respectively.

Model 2 (Nonlinear Factor Model): Let

$$X_{i,j} = (\xi_{ij} + aZ_i)^3,\ 1 \le i \le n, 1 \le j \le p, \tag{2.68}$$

where $\xi_{i,j}, Z_k \sim N(0,1)$ and they are mutually independent. We consider two cases: $a = 0.1$ and $a = 1$, which imply weak and strong factors, respectively. We also let $p = 200$ and $n = 100$ and 200.

Figure 2.1 and Figure 2.2 display the empirical cumulative distribution functions (CDF) of the p-values under the null, should be uniform if the distribution theory is accurate. We report the p-values of the tests obtained from 1,000 simulations. The number of resampling in the subsampling, half-sampling, plug-in calibration are also taken to be $J = 1000$. In the subsampling procedure, we choose $r = 1/\log n$. To obtain the oracle distribution and verify the invariance principle theorem, we plug-in the true covariance matrix into the Monte-Carlo simulation. The number of independent standard normal vectors is also $J = 1,000$ in each realization. In the plug-in approach, the scalars f_1 and f are simply estimated by the $\mathrm{tr}(\hat{\Sigma}_n)$ and $|\hat{\Sigma}_n|_F$, respectively.

From Figure 2.1 - Figure 2.2 we conclude that all the methods produce p-values that are close to uniform for data generated from the short-range dependent linear model and the weak nonlinear factor model. However, a clear non-uniform empirical CDF can be observed for the method GCT and SK under the factor model and the long-range dependence setups. In contrast, the p-values obtained by the half-sampling method are quite close to uniform distribution.

2.7 A Real Data Application

In this section we shall apply the subsampling method to a microarray dataset of pancreatic ductal adenocarcinoma (PDAC), one of the most lethal cancers. To better understand the complex nature of processes in PDAC on a genetic basis, Zhang et al. [2012, 2013] collected gene expression profiling of pancreatic tumor and adjacent non-tumor samples from 45 PDAC patients using Affymetrix GeneChip Human Gene 1.0 ST arrays. We use the preprocessed data from NCBI's Gene Expression Omnibus (Barrett et al. [2013]), accessible through GEO Series accession number GSE28735 (`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28735`). Our goal is to test whether the gene is differentially expressed between uninvolved and involved skin from psoriatic patients.

The dataset consists of $n = 45$ gene expression arrays, and each array has $p = 28,869$ genes. In Zhang et al. [2012], 36 genes were selected based on their expression level, survival status of the subjects and their potential association with cancer-related pathways. However, our analysis reveals statistical significance of many of the individually unidentified genes when they function as a group.

We first test the overall difference between the gene expression levels of tumor and non-tumor samples excluding the 36 genes detected in Zhang et al. [2012]. Due to the dependence of the gene expression differences, the normal approximation of R_n in (2.15) does not appear to be reasonable: the Lyapunov normalized third moment $\mathrm{tr}(\Sigma^3)/|\Sigma|_F^3$ is estimated as $\mathrm{tr}(\hat{\Sigma}_n^3)/|\hat{\Sigma}_n|_F^3 = 0.662$. Thus we shall resort to the invariance principle Theorem 2.2.2. We shall apply the subsampling method and take $m = \lfloor n/\log n \rfloor = 12$ and $J = 10,000$ and perform the test via the subsampling procedure. We obtain $|\bar{X}|_2^2 = 2461.30$. The cutoff value with respect to the significance level 0.001 is 563.54, and the p-value is ≈ 0, which suggests that the genes are differentially expressed as a whole on involved and noninvolved skin, even without the genes identified in Zhang et al. [2012].

We then investigate into the roles of genetic pathways (Kanehisa et al. [2014]) in PDAC. According to the KEGG database, the pathway "hsa05212" is relevant to pancreatic cancer. Among the $28,869$ genes, 66 are mapped to this pathway. The estimate $\mathrm{tr}(\hat{\Sigma}_n^3)/|\hat{\Sigma}_n|_F^3 = 0.807$, again suggesting that the normal approximation of R_n is not satisfactory. Using our subsampling technique, we perform the proposed test on this pathway and conclude that it is significant, where the test statistic is 11.399 and the cutoff value with respect to the 0.001 significance level is 3.072, and the p-value is ≈ 0.

On the other hand, there are totally 229 pathways known for the genes of interest. The largest one is mapped to $1,046$ genes in the dataset. Interestingly, pathways are found to be highly significantly associated with the disease even if they harbors a very small amount of the individually significant genes. For example, the pathway "Olfactory transduction" (ID "hsa04740") is mapped to 381 genes. We apply the FDR procedure on the p-values obtained from the 381 marginal t-tests with respect to a significance level 0.01, where only 16 (4.2%) of the genes were selected. We further performed the L_∞ type test (Chernozhukov et al. [2013]) via Gaussian multiplier calibration on the rest genes in this pathway and obtain a p-value of 0.07 (We persist on a significance level of 0.01 and conclude that there is no strong association). This suggests that non of the rest 365 genes can be further selected as individually significant. However, our L^2-based test leads to the opposite claim. For this pathway, the estimated $\mathrm{tr}(\hat{\Sigma}_n^3)/|\hat{\Sigma}_n|_F^3 = 0.757$, so we again perform a subsampling test ($m = 12$, $J = 10,000$), which indicates that the rest of the genes are still significantly associated with the disease (with a p-value of 0.003).

2.8 Proof of the Results

For ease of notation, we write

$$f_k := [\mathrm{tr}(\Sigma^k)]^{1/k} \text{ and } \hat{f}_k := [\mathrm{tr}(\hat{\Sigma}^k)]^{1/k}, \quad k = 1, 2, \ldots.$$

Then $f_k^k = \sum_{i=1}^p \lambda_i^k$ and $\hat{f}_k^k = \sum_{i=1}^p \hat{\lambda}_i^k$. For the Frobenius norm with $k = 2$, we simply write $f = f_2$ and $\hat{f} = \hat{f}_2$.

Proof of Proposition 2.2.1. Note that $\boldsymbol{\xi} = \Lambda^{-1/2} Q^\top Y_1 \sim N(\mathbf{0}, \mathrm{Id}_p)$. Then $Y_1^\top Y_1 = \sum_{j=1}^p \lambda_j \xi_j^2$, where ξ_j are entries of $\boldsymbol{\xi}$ and are i.i.d. $N(0,1)$. Let $q = 2 + \delta$. By Burkholder's inequality (Chow and Teicher [1997]),

$$\left\| Y^\top Y - f_1 \right\|_q^2 \leq (q-1) \sum_{j=1}^p \lambda_j^2 \|\xi_j^2 - 1\|_q^2.$$

Then (2.10) holds. Let $\boldsymbol{\zeta} = \Lambda^{-1/2} Q^\top Y_2$. Then $Y_1^\top Y_2 = \sum_{j=1}^p \lambda_j \xi_j \zeta_j$ and (2.11) similarly follows. $\square$

In Lemma 2.8.1 and in the proof of Theorem 2.2.2, we define

$$g_0(u) = (1 - \min(1, \max(u, 0))^4)^4. \tag{2.69}$$

Any non-increasing function $g_0(\cdot)$ with $g_0(\cdot) \in \mathbb{C}^3$, $g_0(u) = 1$ if $u \leq 0$, and $g_0(u) = 0$ if

$u \geq 1$, will meet our requirements. To make the calculations explicit, we can choose g_0 in the form of (2.69). Then

$$g_* = \max_u [|g_0'(u)| + |g_0''(u)| + |g_0'''(u)|] < \infty. \tag{2.70}$$

Proof of Theorem 2.2.2. Let $Y_i \in \mathbb{R}^p$ be i.i.d. $N(\mathbf{0}, \Sigma)$ random vectors and $\bar{Y} = \sum_{i=1}^n Y_i/n$. Then $\sum_{j=1}^p \lambda_j \eta_i$ and $n|\bar{Y}_n|_2^2$ are identically distributed. Note that $f_1 = \sum_{j=1}^p \lambda_j$. Hence, to show (2.12), since $L_\delta(n, \psi)$ is increasing in ψ, it suffices to prove the following relation holds for every ψ:

$$\sup_t |\mathbb{P}\left(R_n \leq t\right) - \mathbb{P}\left(R_n^\diamond \leq t\right)| = O(L_\delta(n, \psi) + \psi^{-1/2}), \tag{2.71}$$

where $R_n^\diamond$ is the Gaussian version of R_n in (2.15):

$$R_n^\diamond = \frac{n|\bar{Y}_n|_2^2 - f_1}{f}. \tag{2.72}$$

Recall (2.69) for g_0. We first approximate the indicator function $h(x) = \mathbb{I}\left\{x \leq t\right\}$ by the $\mathbb{C}^3$ function $g_{\psi,t}(x) = g_0(\psi(x - t))$ for t fixed. By (2.70),

$$\mathbb{I}\left\{x \leq t\right\} \leq g_{\psi,t}(x) \leq \mathbb{I}\left\{x \leq t + \psi^{-1}\right\},$$
$$\sup_{x,t} |g_{\psi,t}'(x)| \leq g_* \psi, \ \sup_{x,t} |g_{\psi,t}''(x)| \leq g_* \psi^2, \ \sup_{x,t} |g_{\psi,t}'''(x)| \leq g_* \psi^3.$$

Then $\mathbb{P}\left(R_n \leq t\right) \leq \mathbb{E} g_{\psi,t}\left(R_n\right)$. By Lemma 2.8.1,

$$\begin{aligned} \mathbb{E} g_{\psi,t}\left(R_n\right) &\leq \mathbb{E} g_{\psi,t}\left(R_n^\diamond\right) + C L_\delta(n, \psi) \\ &\leq \mathbb{P}\left(R_n^\diamond \leq t + \psi^{-1}\right) + C L_\delta(n, \psi). \end{aligned} \tag{2.73}$$

The reverse direction is similar: by applying Lemma 2.8.1 again, we have

$$\mathbb{P}\left(R_n \leq t\right) \geq \mathbb{P}\left(R_n^{\diamond} \leq t - \psi^{-1}\right) - C L_\delta(n, \psi). \tag{2.74}$$

By (2.73), (2.74) and (2.80) in Lemma 2.8.2, we have (2.71). $\square$

Lemma 2.8.1. *Assume (2.8) and (2.9). Let $\tilde{K}_\delta$ and $\tilde{D}_\delta$ be specified as in Theorem 2.2.2. Let $g_{\psi,t}(x) = g_0(\psi(x-t))$, where $g_0(\cdot)$ is given by (2.69). Recall (2.72) for R_n and $R_n^{\diamond}$. Then we have*

$$\sup_t \left|\mathbb{E} g_{\psi,t}\left(R_n\right) - \mathbb{E} g_{\psi,t}\left(R_n^{\diamond}\right)\right| = O[L_\delta(n, \psi)]. \tag{2.75}$$

Proof of Lemma 2.8.1. Let $H_i = \sum_{j=1}^{i-1} X_j + \sum_{j=i+1}^{n} Y_j$ and

$$\begin{aligned}
L_i &= \frac{H_i^\top H_i - (n-1) f_1}{n f}, \\
\Delta_i &= \frac{2 H_i^\top X_i + X_i^\top X_i - f_1}{n f}, \\
\Gamma_i &= \frac{2 H_i^\top Y_i + Y_i^\top Y_i - f_1}{n f}.
\end{aligned}$$

Note that H_i is independent of X_i and Y_i. Let

$$\begin{aligned}
\mathrm{I} &= g'_{\psi,t}(L_i)(\Delta_i - \Gamma_i), \\
\mathrm{II} &= \frac{1}{2} g''_{\psi,t}(L_i)(\Delta_i^2 - \Gamma_i^2), \\
\mathrm{III} &= \left[g_{\psi,t}\left(L_i + \Delta_i\right) - g_{\psi,t}\left(L_i + \Gamma_i\right)\right] - \mathrm{I} - \mathrm{II}.
\end{aligned}$$

Note that X_i and Y_i both have mean $\mathbf{0}$ and covariance matrix Σ. Then

$$\begin{aligned}\mathbb{E}\mathrm{I} &= \mathbb{E}\mathbb{E}\left[g'_{\psi,t}(L_i)(\Delta_i - \Gamma_i)\Big| X_i, Y_i\right] \\ &= \frac{1}{nf}\mathbb{E}\left[2(X_i^\top - Y_i^\top)\mathbb{E}(g'_{\psi,t}(L_i)H_i) + (X_i^\top X_i - Y_i^\top Y_i)\mathbb{E}g'_{\psi,t}(L_i)\right] = 0.\end{aligned}$$

For II, by (2.70), $|g''_{\psi,t}(u)| \le g_*\psi^2$. Then for $C_1 = g_*/2$,

$$\begin{aligned}|\mathbb{E}\mathrm{II}| &= \left|\frac{1}{2}\mathbb{E}[g''_{\psi,t}(L_i)(\Delta_i^2 - \Gamma_i^2)]\right| \\ &= \frac{1}{2}\left|\mathbb{E}\left[g''_{\psi,t}(L_i)\mathbb{E}\left(\Delta_i^2 - \Gamma_i^2\Big| H_i\right)\right]\right| \\ &\le C_1\psi^2\mathbb{E}\left|\mathbb{E}\left(\Delta_i^2 - \Gamma_i^2\Big| H_i\right)\right|.\end{aligned}$$

The term $n^2f^2\mathbb{E}\left(\Delta_i^2 - \Gamma_i^2\big| H_i\right)$ can be decomposed into

$$\begin{aligned}&4\mathbb{E}\left[H_i^\top X_iX_i^\top H_i - H_i^\top Y_iY_i^\top H_i\Big| H_i\right] + \mathbb{E}\left[(X_i^\top X_i - f_1)^2 - (Y_i^\top Y_i - f_1)^2\right] \\ &\quad + 4\mathbb{E}\left[H_i^\top X_i(X_i^\top X_i - f_1) - H_i^\top Y_i(Y_i^\top Y_i - f_1)\Big| H_i\right],\end{aligned}$$

where $\mathbb{E}\left(H_i^\top X_iX_i^\top H_i - H_i^\top Y_iY_i^\top H_i\Big| H_i\right) = 0$. By (2.8),

$$\mathbb{E}\left|(X_i^\top X_i - f_1)^2 - (Y_i^\top Y_i - f_1)^2\right| \le f^2(K_0^2 + c_0^2) \le f^2\tilde{K}_0^2.$$

Since Y_i is Gaussian, $\mathbb{E}\left[H_i^\top Y_i(Y_i^\top Y_i - f_1)\Big| H_i\right] = 0$. By the Cauchy-Schwarz inequality and (2.8), since $\|H_i^\top X_i\|^2 = (n-1)\mathrm{tr}(\Sigma^2) = (n-1)f^2$,

$$\begin{aligned}\mathbb{E}\left|\mathbb{E}\left(\Delta_i^2 - \Gamma_i^2\Big| H_i\right)\right| &\le \frac{\tilde{K}_0^2}{n^2} + \frac{\mathbb{E}|\mathbb{E}[H_i^\top X_i(X_i^\top X_i - f_1)\Big| H_i]|}{n^2f^2} \\ &\le \frac{\tilde{K}_0^2}{n^2} + \frac{\|H_i^\top X_i\|\|X_i^\top X_i - f_1\|}{n^2f^2} \\ &\le \frac{\tilde{K}_0^2}{n^2} + \frac{K_0}{n^{3/2}}.\end{aligned}$$

So

$$|\mathbb{E}\mathrm{II}| \le C\psi^2(n^{-2}\tilde{K}_0^2 + n^{-3/2}\tilde{K}_0). \tag{2.76}$$

Since $0 \le g(t) \le 1$ for all t, and $|g'''_{\psi,t}(u)| \le g_*\psi^3$. We have that

$$\begin{aligned}
\mathbb{E}\,|\mathrm{III}| &\le \mathbb{E}\min\Big\{1 + |\mathrm{I}| + |\mathrm{II}|, g_*\psi^3(|\Delta_i|^3 + |\Gamma_i|^3)\Big\} \\
&\le C\mathbb{E}\min\Big\{1 + \psi(|\Delta_i| + |\Gamma_i|) + \psi^2(|\Delta_i|^2 + |\Gamma_i|^2), \psi^3(|\Delta_i|^3 + |\Gamma_i|^3)\Big\} \\
&\le C\psi^q(\mathbb{E}\,|\Delta_i|^q + \mathbb{E}\,|\Gamma_i|^q),
\end{aligned}$$

where $q = 2 + \delta$. Let $\mathbf{x} \in \mathbb{R}^p$ be a fixed vector. By Rosenthal's inequality,

$$\mathbb{E}\,|H_i\mathbf{x}|_q^q \le c_q[i\|X_1^\top\mathbf{x}\|_q^q + (n-i)\|Y_n^\top\mathbf{x}\|_q^q + n^{q/2}(\mathbf{x}^\top\Sigma\mathbf{x})^{q/2}], \tag{2.77}$$

where c_q and $c_{q,1}, \ldots$ hereafter are constants only depend on q and they may take different values at different appearances. Note that $Y_n^\top\mathbf{x} \sim N(0, \mathbf{x}^\top\Sigma\mathbf{x})$. Let $c_{q,1} = \|\xi_1\|_q^q$, $\xi_1 \sim N(0,1)$. Then $\mathbb{E}|Y_n^\top\mathbf{x}|^q = c_{q,1}(\mathbf{x}^\top\Sigma\mathbf{x})^{q/2}$ and

$$\|H_i^\top X_i\|_q^q \le c_q(n\left\|X_1^\top X_2\right\|_q^q + n^{q/2}\mathbb{E}(X_1^\top\Sigma X_1)^{q/2}) \tag{2.78}$$

Hence by (2.9) and (2.11), we have

$$\begin{aligned}
\mathbb{E}|\Delta_i|^q &\le C\frac{\mathbb{E}|H_i^\top X_i|^q + \mathbb{E}|X_i^\top X_i - f_1|^q}{n^q f^q} \\
&\le C\frac{n\tilde{D}_\delta^q f^q + n^{q/2}\mathbb{E}(X_1^\top\Sigma X_1)^{q/2} + K_\delta^q f^q}{n^q f^q}.
\end{aligned} \tag{2.79}$$

By (2.77), $\|H_i^\top Y_i\|_q^q \le c_q(n\mathbb{E}(X_1^\top \Sigma X_1)^{q/2} + n^{q/2} f^q)$, which implies that

$$\mathbb{E}\,|\Gamma_i|^q \le c_q(n\mathbb{E}(X_1^\top \Sigma X_1)^{q/2}/(nf)^q + n^{-q/2}).$$

Observe that $(|H_i + X_i|_2^2 - nf_1)/(nf) = L_i + \Delta_i$ and $(|H_i + Y_i|_2^2 - nf_1)/(nf) = L_i + \Gamma_i$. We write the telescope sum

$$g_{\psi,t}\left(R_n\right) - g_{\psi,t}\left(R_n^\diamond\right) = \sum_{i=1}^n \left[g_{\psi,t}(L_i + \Delta_i) - g_{\psi,t}(L_i + \Gamma_i)\right],$$

which entails (2.75) in view of (2.76), (2.79) and $\mathbb{E}\mathrm{I} = 0$. $\square$

Lemma 2.8.2. *Let $a_1 \ge \ldots \ge a_p \ge 0$ be such that $\sum_{i=1}^p a_i^2 = 1$; let η_i be i.i.d. χ_1^2 random variables. Then for all $h > 0$,*

$$\sup_t \mathbb{P}\left(t \le a_1\eta_1 + \ldots + a_p\eta_p \le t + h\right) \le h^{1/2}\sqrt{4/\pi}. \tag{2.80}$$

Proof of Lemma 2.8.2. Write $V = \sum_{i=1}^p a_i\eta_i$. Assume $a_1 \le 1/2$. Then its characteristic function $\phi_V(s) = \mathbb{E}\exp(\sqrt{-1}sV)$, $s \in \mathbb{R}$, satisfies

$$\begin{aligned} |\phi_V(s)| &= \left|\prod_{j=1}^p (1 - 2\sqrt{-1}a_j s)^{-1/2}\right| \\ &= \prod_{j=1}^p (1 + 4a_j^2 s^2)^{-1/4} \\ &\le (1 + 4s^2 + 8b_4 s^4 + 32/3 b_6 s^6)^{-1/4}, \end{aligned} \tag{2.81}$$

where $b_4 = \sum_{j\ne k} a_j^2 a_k^2 = 1 - \sum_{k=1}^p a_k^4 \ge 1 - a_1^2 \ge 3/4$ and

$$b_6 = \sum\nolimits_{j,k,l}^{\star} a_j^2 a_k^2 a_l^2 = 1 - 3\sum_{j\ne k} a_j^4 a_k^2 - \sum_{j=1}^p a_j^6$$

$$\geq \quad 1 - 3\sum_{j=1}^{p} a_j^4 \left(\sum_{k\neq j} a_k^2 + a_j^2\right) \geq 1 - 3a_1^2 \geq 1/4.$$

By the inversion formula and (2.81), the density function $f_V(\cdot)$ of V satisfies

$$f_V(v) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\sqrt{-1}vs}\phi_V(s)ds \leq \frac{1}{2\pi}\int_{-\infty}^{\infty} |\phi_V(s)|ds < 1$$

Now we shall deal with the case that $a_1 > 1/2$. Note that for all $w > 0$, $\sup_u \mathbb{P}(u \leq \eta_1 \leq u + w) \leq w^{1/2}\sqrt{2/\pi}$. Then $\sup_t \mathbb{P}(t \leq V \leq t + h) \leq (2h)^{1/2}\sqrt{2/\pi}$. Combining with the case $a_1 \leq 1/2$, we obtain the upper bound $\max(h^{1/2}\sqrt{4/\pi}, h)$. Note that (2.80) trivially holds if $h \geq 1$. □

Proof of Proposition 2.3.2. Note that $\rho(\Sigma/f) \leq |\Sigma/f|_F = 1$. Since $\tilde{f}/f - 1 = o_{\mathbb{P}}(1)$, $\rho(\Sigma/f)(f/\tilde{f} - 1) = o_{\mathbb{P}}(1)$. Hence for the "if" part,

$$\rho(\tilde{\Sigma}/\tilde{f} - \Sigma/f) \leq \rho(\tilde{\Sigma} - \Sigma)/\tilde{f} + \rho(\Sigma/f)|f/\tilde{f} - 1| = o_{\mathbb{P}}(1)$$

The "only if" part can be similarly proved. □

Proof of Lemma 2.3.1. Let $\rho_p = \max_j \left|a_{p,j} - b_{p,j}\right|$. Choose an integer sequence $K = K_p$ such that $K_p \to \infty$ and $K_p\rho_p \to 0$. Let $W = \sum_{j=1}^{K-1} a_{p,j}\eta'_j$, $W^\circ = \sum_{j=K}^{p} a_{p,j}\eta'_j$, $S = \sum_{j=1}^{K-1} b_{p,j}\eta'_j$, $S^\circ = \sum_{j=K}^{p} b_{p,j}\eta'_j$, $w = 2\sum_{j=K}^{p} a_{p,j}^2$ and $s = 2\sum_{j=K}^{p} b_{p,j}^2$. Let $u_K = a_{p,K}^{1/4}$. By the Gaussian approximation result in Sakhanenko [2006], on a richer probability space, we can construct a random variable $Z \sim N(0,1)$, independent of $(\eta_i)_{i=1}^{K-1}$, such that

$$\mathbb{P}(|W^\circ - w^{1/2}Z| \geq u_K) \leq \frac{c_4}{u_K^4}\sum_{j=K}^{p} a_{p,j}^4 \leq \frac{c_4}{u_K^4} a_{p,K}^2 = c_4 u_K^4, \tag{2.82}$$

where $c_4 > 0$ is an absolute constant. Since $u_K \to 0$, by Lemma 2.8.2,

$$\sup_x |\mathbb{P}(|W + W^\circ| \le x) - \mathbb{P}(|W + w^{1/2} Z| \le x)| \to 0. \tag{2.83}$$

Similarly, for $v_K = b_{p,K}^{1/4}$, we can also construct a probability space with a r.v. $Z^* \sim N(0,1)$ such that $\mathbb{P}(|S^\circ - w^{1/2} Z^*| \ge v_K) \le c_4 v_K^4$, and

$$\sup_x |\mathbb{P}(|S + S^\circ| \le x) - \mathbb{P}(|S + s^{1/2} Z^*| \le x)| \to 0. \tag{2.84}$$

Let $T = (W + w^{1/2} Z) - (S + s^{1/2} Z)$. Since $w - s = 2\sum_{j=1}^{K-1}(b_{p,j}^2 - a_{p,j}^2)$,

$$\begin{aligned} \mathbb{E}|T| &\le 2(K-1)\rho_p + |w^{1/2} - s^{1/2}| \\ &\le 2K\rho_p + |w - s|^{1/2} \le 2K\rho_p + (4K\rho_p)^{1/2} \to 0. \end{aligned} \tag{2.85}$$

Hence, by (2.83), (2.84) and Lemma 2.8.2, (2.41) follows. □

Proof of Theorem 2.3.3. Since X_i are i.i.d., we have

$$\begin{aligned} \mathbb{E}|\hat{\Sigma} - \Sigma|_F^2 &= \mathbb{E} \sum_{j,k=1}^{p} (\hat{\sigma}_{jk} - \sigma_{jk})^2 \\ &= \frac{1}{n} \sum_{j,k=1}^{p} \mathbb{E}\left(X_{1j}^2 X_{1k}^2 - \sigma_{jk}^2\right) \\ &= \frac{1}{n}\mathbb{E}\Big[\big(\sum_{j=1}^{p} X_{1j}^2\big)^2\Big] - \frac{1}{n} f^2 \\ &= \frac{1}{n}\mathbb{E}[(X_1^\top X_1)^2] - \frac{1}{n} f^2, \end{aligned}$$

which, by the assumption $\mathbb{E}[(X_1^\top X_1)^2] = o\left(nf^2\right)$, implies $\mathbb{E}|\hat{\Sigma} - \Sigma|_F^2 = o(f^2)$. Then $||\Sigma|_F -$

$|\hat{\Sigma}|_F\|_2 \leq \||\hat{\Sigma} - \Sigma|_F\|_2 = o(f)$, or $\|f - \hat{f}\|_2 = o(f)$, and

$$\|\hat{\Sigma}/\hat{f} - \Sigma/f\|_F \leq \|(\hat{\Sigma} - \Sigma)/f\|_F + \||\hat{\Sigma}/\hat{f}|_F|1 - \hat{f}/f|\| = o(1).$$

$\square$

Proof of Theorem 2.3.4. (i) Assume without loss of generality that $\mu = 0$. For a set $B \subset \{1, \dots, n\}$ define $W_B^\circ = (|B||\bar{X}_B|_2^2 - f_1)/f$ and $W_B = [|B||\bar{X}_B - \bar{X}|_2^2/(1 - |B|/n) - f_1]/f$. Using the identity $n\bar{X} = |B|\bar{X}_B + (n - |B|)\bar{X}_{B^c}$, where $B^c = \{1, \dots, n\} - B$, we have by elementary manipulations that

$$W_B = \frac{n - |B|}{n} W_B^\circ + \frac{|B|}{n} W_{B^c}^\circ - 2|B| \frac{n - |B|}{n} \frac{\bar{X}_B^\top \bar{X}_{B^c}}{f}. \tag{2.86}$$

Then for any $\theta > 0$, we have by the triangle inequality that

$$\mathbb{P}(W_{B_j}^\circ \leq \frac{t - \theta}{1 - m/n}) - \tau \leq \mathbb{P}(W_{B_j} \leq t) \leq \mathbb{P}(W_{B_j}^\circ \leq \frac{t + \theta}{1 - m/n}) + \tau, \tag{2.87}$$

where $\tau = \mathbb{P}(|R_j| \geq \theta)$, $R_j = (m/n)W_{B_j^c}^\circ - 2m(1 - m/n)^{-1} f^{-1} \bar{X}_{B_j}^\top \bar{X}_{B_j^c}$. Note that $\mathbb{E}|\bar{X}_{B_j}^\top \bar{X}_{B_j^c}|^2 = f^2/(m(n-m))$. Since $m = o(n)$, by Theorem 2.2.2, we have $R_j = o_{\mathbb{P}}(1)$ and $\tau \to 0$. Hence by Theorem 2.2.2, Lemma 2.8.2 and (2.87),

$$\mathbb{P}(W_{B_j} \leq t) - \mathbb{P}(W_{B_j}^\circ \leq t) \to 0. \tag{2.88}$$

A similar argument implies that, for $j \neq j'$, the joint probability

$$\mathbb{P}(W_{B_j} \leq t, W_{B_{j'}} \leq t) - \mathbb{P}(W_{B_j}^\circ \leq t, W_{B_{j'}}^\circ \leq t) \to 0. \tag{2.89}$$

Therefore, by Theorem 2.2.2, we have $\mathbb{E}|\check{F}_r(t) - \mathbb{P}(V \leq t)|^2 \to 0$, which implies the uniform version (2.49) via the standard Glivenko–Cantelli argument in view of the continuity result

Lemma 2.8.2.

We now prove (ii). Following the argument in (i), it suffices to show that

$$\mathbb{E}|\mathbb{P}(W^{\circ}_{A_j} \le t, W^{\circ}_{A_{j'}} \le t) - \mathbb{P}^2(V \le t)| \to 0. \tag{2.90}$$

For sets $A, A' \in \mathcal{A}$, let $A \cap A' = D_1$, $A - D_1 = D_2$ and $A' - D_1 = D_3$. Then

$$W^{\circ}_{A} = (1-k/m)W^{\circ}_{D_2} + (k/m)W^{\circ}_{D_1} + 2k(1-k/m)\frac{\bar{X}^{\top}_{D_1}\bar{X}_{D_2}}{f},$$

where $k = |D_1|$. A similar expression exists for $W^{\circ}_{A'}$. Choose a sequence $\rho_n \to 0$ with $m/n = o(\rho_n)$. If $k \le m\rho_n$, similarly as in part (i), we have $|\mathbb{P}(W^{\circ}_{A} \le t, W^{\circ}_{A'} \le t) - \mathbb{P}(W^{\circ}_{D_2} \le t, W^{\circ}_{D_3} \le t)| \to 0$ and $|\mathbb{P}(W^{\circ}_{D_2} \le t) - \mathbb{P}(V \le t)| \to 0$. Note that $\mathbb{E}|A_j \cap A_{j'}| \le m^2/n$. Then $\mathbb{P}(|A_j \cap A_{j'}| \ge m\rho_n) \le m/(n\rho_n) \to 0$. Then (2.90) follows by conditioning on $|A_j \cap A_{j'}| \le m\rho_n$. □

Proof of Theorem 2.3.5. We omit the proof of Theorem 2.3.5 and show Theorem 2.4.1 as the one sample result is a special case of the two sample result. □

The following notation is used in the proof of Theorem 2.4.1.

Denote

$$\begin{aligned}
\mathcal{T} &= \frac{S^X_{H_1} - S^X_{H_4}}{n_1} - \frac{S^Y_{M_1} - S^Y_{M_4}}{n_2}, \qquad & \mathcal{V} &= \frac{S^X_{H_2} - S^X_{H_3}}{n_1} - \frac{S^Y_{M_2} - S^Y_{M_3}}{n_2},\\
\tau &= \frac{S^\xi_{H_1} - S^\xi_{H_4}}{n_1} - \frac{S^\zeta_{M_1} - S^\zeta_{M_4}}{n_2}, \qquad & \nu &= \frac{S^\xi_{H_2} - S^\xi_{H_3}}{n_1} - \frac{S^\zeta_{M_2} - S^\zeta_{M_3}}{n_2}.
\end{aligned} \tag{2.91}$$

Note that given the sets A_1, B_1, A_2, B_2, the random variables $\mathcal{T}, \mathcal{V}, \tau, \nu$ are independent, and

$$R_0^{(1)} = \frac{(\tau+\nu)^\top(\tau+\nu) - \bar{f}_1}{\bar{f}}, \qquad R_0^{(2)} = \frac{(\tau-\nu)^\top(\tau-\nu) - \bar{f}_1}{\bar{f}}, \tag{2.92}$$

$$R_1^{(1)} = \frac{(\mathcal{T}+\mathcal{V})^\top(\mathcal{T}+\mathcal{V}) - \bar{f}_1}{\bar{f}}, \qquad R_1^{(2)} = \frac{(\mathcal{T}-\mathcal{V})^\top(\mathcal{T}-\mathcal{V}) - \bar{f}_1}{\bar{f}}. \tag{2.93}$$

Proof of Theorem 2.4.1. Given A_j, B_j, $\bar{X}_{A_j} - \bar{X} = n_1^{-1}(S^X_{A_j} - S^X_{A_j^c})$ and $\bar{Y}_{B_j} - \bar{Y} = n_2^{-1}(S^Y_{B_j} - S^Y_{B_j^c})$. By Lemma 2.8.3, we have that for all $A_j \in \mathcal{A}^\circ, B_j \in \mathcal{B}^\circ$,

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(R_1^{(j)} \le t | A_j, B_j) - \mathbb{P}(R_0 \le t)| = o(1),$$
$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\left(R_1 \le t | A_j, B_j\right) - \mathbb{P}(R_0 \le t)\right| = o(1).$$

Consequently,

$$\sup_t |\mathbb{E}\kappa_j(t)| \le \mathbb{E}[\sup_t |\mathbb{E}(\kappa_j(t)|A_j, B_j)|] = o(1). \tag{2.94}$$

On the other hand, Lemma 2.8.4 shows that for any $A_1, A_2 \in \mathcal{A}^\circ, B_1, B_2 \in \mathcal{B}^\circ$,

$$\sup_t \left|\mathbb{P}\left(R_1^{(1)} \le t, R_1^{(2)} \le t | A_1, A_2, B_1, B_2\right) - \mathbb{P}\left(R_0^{(1)} \le t, R_0^{(2)} \le t | A_1, A_2, B_1, B_2\right)\right| = o(1),$$

which implies that

$$\sup_t \left|\mathbb{P}\left(R_1^{(1)} \le t, R_1^{(2)} \le t\right) - \mathbb{P}\left(R_0^{(1)} \le t, R_0^{(2)} \le t\right)\right| = o(1).$$

By Lemma 2.8.6 and Lemma 2.8.3, we further have

$$\sup_t |\mathrm{Cov}\left(\kappa_1(t), \kappa_2(t)\right)| = \sup_t \left|\mathbb{P}\left(R_1^{(1)} \le t, R_1^{(2)} \le t\right) - \mathbb{P}^2\left(R_1^{(1)} \le t\right)\right| = o(1). \tag{2.95}$$

Therefore, we obtain from (2.94) and (2.95) that

$$\mathbb{E}\left(J^{-1}\sum_{j=1}^{J}\kappa_j(t)\right)^2 = o(1).$$

The uniform convergence (2.63) is obtained via the standard Glivenko-Cantelli argument. $\square$

Lemma 2.8.3. *Let Condition 2.4.1 be satisfied. Let $Z \sim N(0, \mathrm{I}_p)$. Then we have*

$$\sup_{t\in\mathbb{R}}\left|F(t) - \mathbb{P}\left(Z^\top \bar{\Sigma} Z \le t\right)\right| \le C[L_\delta(n_1, n_2, \psi) + \psi^{-1/2}], \tag{2.96}$$

for some constant C that does not grow with n or p,

$$\begin{aligned} L_\delta(n_1, n_2, \psi) =&\psi^2\left[\tilde{K}_0^2(n_1^{-1} + n_2^{-1}) + \tilde{K}_0(n_1^{-1/2} + n_2^{-1/2})\right] + \psi^q\Big\{\left(n_1^{-\delta} + n_2^{-\delta}\right)\tilde{D}_\delta^q \\ &+ \left(n_1^{1-q} + n_2^{1-q}\right)\tilde{K}_\delta^q + \bar{f}^{-q}\left(n_1^{1-q}\mathbb{E}(X_1^\top \bar{\Sigma} X_1)^{q/2} + n_2^{1-q}\mathbb{E}(Y_1^\top \bar{\Sigma} Y_1)^{q/2}\right)\Big\}. \end{aligned}$$

Furthermore, if (2.61) and (2.62) hold, then

$$\sup_{t\in\mathbb{R}}\left|F(t) - \mathbb{P}\left(Z^\top \bar{\Sigma} Z \le t\right)\right| = o(1).$$

Proof of Lemma 2.8.3. The lemma can be proved in a similar way as Theorem 2.2.2. Details are omitted. $\square$

Lemma 2.8.4. *For $j = 1, 2$, fix A_j, B_j. Recall (2.53) for $R_0^{(j)}$, and denote*

$$R_1^{(j)} = \frac{|\bar{X}_{A_j} - \bar{X} - \bar{Y}_{B_j} + \bar{Y}|_2^2 - \bar{f}_1}{\bar{f}}. \tag{2.97}$$

Assume that $(X_i)_{1\le i\le n_1}$, $(Y_j)_{1\le j\le n_2}$ are independent, and that the conditions in Theorem

2.4.1 are satisfied. We have that

$$\sup_t \left| \mathbb{P}\left(R_0^{(1)} \le t, R_0^{(2)} \le t\right) - \mathbb{P}\left(R_1^{(1)} \le t, R_1^{(2)} \le t\right) \right| = o(1). \tag{2.98}$$

Proof of Lemma 2.8.4. We approximate $|x|$ by the function

$$l(x) = \begin{cases} -16^{-1}\left(5x^8 - 21x^6 + 35x^4 - 35x^2\right) & , \text{ if } |x| \le 1; \\ |x|, \text{ otherwise} . \end{cases}$$

Then we have for some constant C_l, we have that

$$\sup_x |l'(x)| + \sup_x |l''(x)| + \sup_x |l'''(x)| < C_l.$$

Let $l_\psi(x) = \psi^{-1} l(\psi x)$, then

$$|x| - \psi^{-1} \le l_\psi(x) \le |x| \le l_\psi(x) + \psi^{-1}.$$

Recall $g_{\psi,t}(\cdot)$ in the proof of Theorem 2.2.2. Define

$$h_{\psi,t}(x,y) = g_{\psi,t}\left(x + 2l_\psi(y)\right),$$

It follows that

$$\mathbb{I}\left\{x + 2|y| \le t\right\} \le h_{\psi,t}(x,y) \le \mathbb{I}\left\{x + 2|y| \le t + 3\psi^{-1}\right\}. \tag{2.99}$$

Recall (2.91) for T, V, τ and ν. Let

$$\rho_0 = \frac{\tau^\top \tau + \nu^\top \nu - \bar{f}_1}{\bar{f}}, \qquad \varrho_0 = \frac{\tau^\top \nu}{\bar{f}}, \tag{2.100}$$

$$\rho_1 = \frac{\mathcal{T}^\top \mathcal{T} + \mathcal{V}^\top \mathcal{V} - \bar{f}_1}{\bar{f}}, \qquad \varrho_1 = \frac{\mathcal{T}^\top \mathcal{V}}{\bar{f}}. \tag{2.101}$$

Then we have

$$\mathbb{P}\left(R_1^{(1)} \leq t, R_1^{(2)} \leq t\right) = \mathbb{P}\left(\rho_1 + 2|\varrho_1| \leq t\right),$$
$$\mathbb{P}\left(R_0^{(1)} \leq t, R_0^{(2)} \leq t\right) = \mathbb{P}\left(\rho_0 + 2|\varrho_0| \leq t\right).$$

So by Lemma 2.8.5, (2.99) and Lemma 2.8.2, we get

$$\begin{aligned}
&\mathbb{P}\left(R_1^{(1)} \leq t, R_1^{(2)} \leq t\right) \\
&\leq \mathbb{P}\left(R_0^{(1)} \leq t + 3\psi^{-1}, R_0^{(2)} \leq t + 3\psi^{-1}\right) + CL_\delta(n_1, n_2, \psi) \\
&\leq \mathbb{P}\left(R_0^{(1)} \leq t, R_0^{(2)} \leq t\right) + C_1 L_\delta(n_1, n_2, \psi) + C_2 \psi^{-1/2}.
\end{aligned}$$

Similarly,

$$\mathbb{P}\left(R_1^{(1)} \leq t, R_2^{(1)} \leq t\right) \leq \mathbb{P}\left(R_0^{(1)} \leq t, R_0^{(2)} \leq t\right) - C_1 L_\delta(n_1, n_2, \psi) - C_2 \psi^{-1/2}.$$

The proof is completed. $\square$

Lemma 2.8.5. *Assume the conditions in Theorem 2.8.4 are satisfied. Recall (2.100) and (2.101). There exists a constant $C > 0$ such that*

$$\sup_t \left|\mathbb{E} h_{\psi,t}\left(\rho_1, \varrho_1\right) - \mathbb{E} h_{\psi,t}\left(\rho_0, \varrho_0\right)\right| \leq CL_\delta(n_1, n_2, \psi).$$

Proof of Lemma 2.8.5. Note that there exists a positive constant C_h such that

$$\sup_{x,y,t}\left|\frac{\partial h_{\psi,t}(x,y)}{\partial x}\right| \le C_h\psi, \qquad \sup_{x,y,t}\left|\frac{\partial h_{\psi,t}(x,y)}{\partial y}\right| \le C_h\psi,$$

$$\sup_{x,y,t}\left|\frac{\partial^2 h_{\psi,t}(x,y)}{\partial x^2}\right| \le C_h\psi^2, \qquad \sup_{x,y,t}\left|\frac{\partial^2 h_{\psi,t}(x,y)}{\partial x\partial y}\right| \le C_h\psi^2,$$

$$\sup_{x,y,t}\left|\frac{\partial^2 h_{\psi,t}(x,y)}{\partial y^2}\right| \le C_h\psi^2, \qquad \sup_{x,y,t}\left|\frac{\partial^3 h_{\psi,t}(x,y)}{\partial x^3}\right| \le C_h\psi^3,$$

$$\sup_{x,y,t}\left|\frac{\partial^3 h_{\psi,t}(x,y)}{\partial x^2\partial y}\right| \le C_h\psi^3, \qquad \sup_{x,y,t}\left|\frac{\partial^3 h_{\psi,t}(x,y)}{\partial x\partial y^2}\right| \le C_h\psi^3,$$

$$\sup_{x,y,t}\left|\frac{\partial^3 h_{\psi,t}(x,y)}{\partial y^3}\right| \le C_h\psi^3.$$

We can prove the lemma by expanding $h_{\psi,t}(\rho_1,\varrho_1) - h_{\psi,t}(\rho_0,\varrho_0)$ similarly as in Lemma 2.8.1. The details are omitted. □

Lemma 2.8.6. *Recall $R_0^{(j)}$ from (2.53) for $j = 1, 2$. There exists some constant $C > 0$ such that*

$$\sup_t\left|\mathbb{P}\left(R_0^{(1)} \le t, R_0^{(2)} \le t\right) - \mathbb{P}^2\left(R_0^{(1)} \le t\right)\right| \le C\left(n_1^{-1/5} + n_2^{-1/5}\right). \tag{2.102}$$

Proof of Lemma 2.8.6. Let $\delta_1 = (4|H_1| - n_1)/n_1$, $\delta_2 = (4|M_1| - n_2)/n_2$, and

$$\beta_1 = \bar{\xi}_{A_1} - \bar{\xi}, \quad \beta_2 = \bar{\xi}_{A_2} - \bar{\xi}, \quad \varpi_1 = \bar{\zeta}_{B_1} - \bar{\zeta}, \quad \varpi_2 = \bar{\zeta}_{B_2} - \bar{\zeta}.$$

Observe that $\tau + \nu = \beta_1 - \varpi_1$ and $\tau - \nu = \beta_2 - \varpi_2$. Let $\varpi = \beta_2 - \varpi_2 - \delta_1\beta_1 + \delta_2\varpi_1$, and a simple calculation leads to

$$\mathrm{cov}(\beta_1 - \varpi_1, \varpi | A_1, A_2, B_1, B_2) = 0.$$

We define

$$R_{\varpi} = \frac{|\varpi|_2^2 - f_1^{\varpi}}{\bar{f}},$$

where $f_1^{\varpi} = tr(\Sigma_{\varpi})$ and $\Sigma_{\varpi} = (1-\delta_1^2)\Sigma_X/n_1 + (1-\delta_2^2)\Sigma_Y/n_2$. Then for any $\iota > 0$, we have

$$\begin{aligned}
\mathbb{P}_{\mathcal{E}}(R_0^{(1)} \le t, R_0^{(2)} \le t) \le& \mathbb{P}_{\mathcal{E}}(R_0^{(1)} \le t, R_{\varpi} \le t+\iota) + \mathbb{P}_{\mathcal{E}}(|R_0^{(2)} - R_{\varpi}| > \iota) \\
=& \mathbb{P}_{\mathcal{E}}(R_0^{(1)} \le t)\mathbb{P}_{\mathcal{E}}(R_{\varpi} \le t+\iota) + \mathbb{P}_{\mathcal{E}}(|R_0^{(2)} - R_{\varpi}| > \iota) \\
\le& \mathbb{P}_{\mathcal{E}}(R_0^{(1)} \le t)P_{\mathcal{E}}(R_0^{(2)} \le t+2\iota) + 2\mathbb{P}_{\mathcal{E}}(|R_0^{(2)} - R_{\varpi}| > \iota).
\end{aligned}$$

Here the subscript $\mathbb{P}_{\mathcal{E}}$ and $\mathbb{E}_{\mathcal{E}}$ denotes conditional probability and expectation given A_1, A_2, B_1, B_2. For the first term on the right hand side of the inequality above, by Lemma 2.8.2, we have

$$\mathbb{P}_{\mathcal{E}}(R_0^{(1)} \le t)\mathbb{P}_{\mathcal{E}}(R_0^{(2)} \le t+2\iota) \le \mathbb{P}_{\mathcal{E}}(R_0^{(1)} \le t)\mathbb{P}_{\mathcal{E}}(R_0^{(2)} \le t) + \iota^{1/2}\sqrt{8/\pi}. \tag{2.103}$$

For the second term, we have

$$R_0^{(2)} - R_{\varpi} = \frac{|\delta_1\beta_1 - \delta_2\varpi_1|_2^2 - \bar{f}_1 + f_1^{\varpi}}{\bar{f}} + \frac{2\varpi^\top(\delta_1\beta_1 - \delta_2\varpi_1)}{\bar{f}}.$$

Observe that

$$\begin{aligned}
\mathbb{E}_{\mathcal{E}}\left|\frac{|\delta_1\beta_1 - \delta_2\varpi_1|_2^2 - \bar{f}_1 + f_1^{\varpi}}{\bar{f}}\right|_2^2 &= 2\frac{\mathrm{tr}((\delta_1^2\Sigma_X/n_1 + \delta_2^2\Sigma_Y/n_2)^2)}{\bar{f}^2} \le 4(\delta_1^4 + \delta_2^4), \\
\mathbb{E}_{\mathcal{E}}\left|\frac{2\varpi^\top(\delta_1\beta_1 - \delta_2\varpi_1)}{\bar{f}}\right|_2^2 &= 4\bar{f}^{-2}\mathrm{tr}\Big[\Sigma_{\varpi}\Big(\delta_1^2\frac{\Sigma_X}{n_1} + \delta_2^2\frac{\Sigma_Y}{n_2}\Big)\Big] \le 4(\delta_1^2 + \delta_2^2).
\end{aligned}$$

Then we have $\|R_0^{(2)} - R_{\varpi}\|_2^2 \le C(\delta_1^2 + \delta_2^2)$ for some constant C independent of A_1, A_2, B_1, B_2.

Thus

$$\mathbb{P}_{\mathcal{E}}(|R_0^{(2)} - R_{\varpi}| > \iota) \leq C\frac{\delta_1^2 + \delta_2^2}{\iota^2}.$$

Balancing the above inequality with (2.103), we take $\iota = \delta_1^{4/5} + \delta_2^{4/5}$ and obtain

$$\mathbb{P}_{\mathcal{E}}(R_0^{(1)} \leq t, R_0^{(2)} \leq t) \leq \mathbb{P}_{\mathcal{E}}(R_0^{(1)} \leq t)\mathbb{P}_{\mathcal{E}}(R_0^{(2)} \leq t) + C(\delta_1^{2/5} + \delta_2^{2/5}).$$

Since

$$\mathbb{E}\left[\mathbb{P}_{\mathcal{E}}(R_0^{(1)} \leq t)\mathbb{P}_{\mathcal{E}}(R_0^{(2)} \leq t)\right] = \mathbb{E}\left[\mathbb{P}(R_0^{(1)} \leq t|A_1, B_1)\mathbb{P}(R_0^{(2)} \leq t|A_2, B_2)\right],$$

the lemma follow from the fact that $\|\delta_j\|_{2/5} \leq \|\delta_1\|_2 = O(n_1^{-1/2})$ for the hypergeometric random variable $\delta_j, j = 1, 2$. □

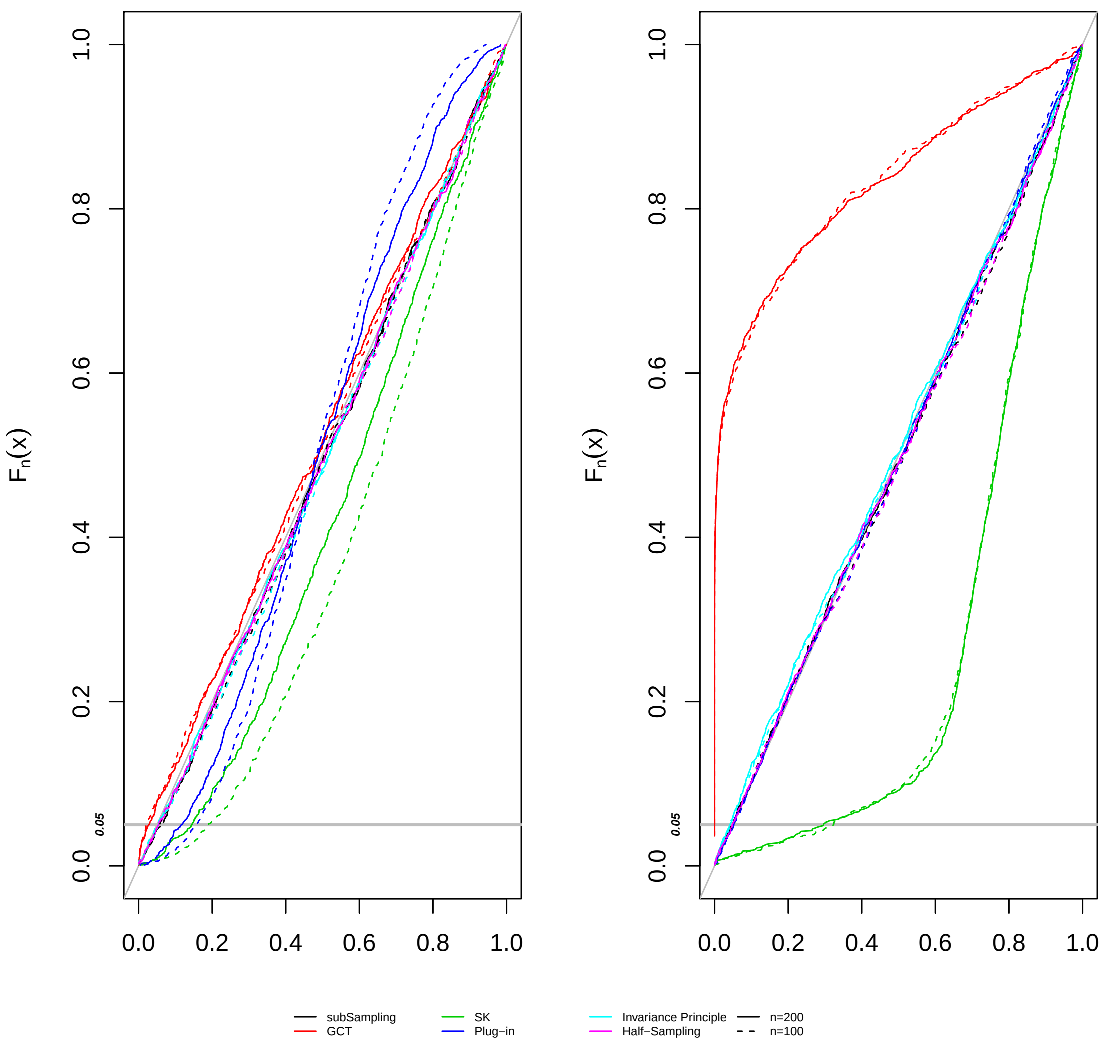


Figure 2.1: The empirical cumulative distribution function of the p-values based on 1000 simulations from Model 1. The data are generated with $\beta = 2$ and $\beta = 0.6$ in the left and right panels respectively.

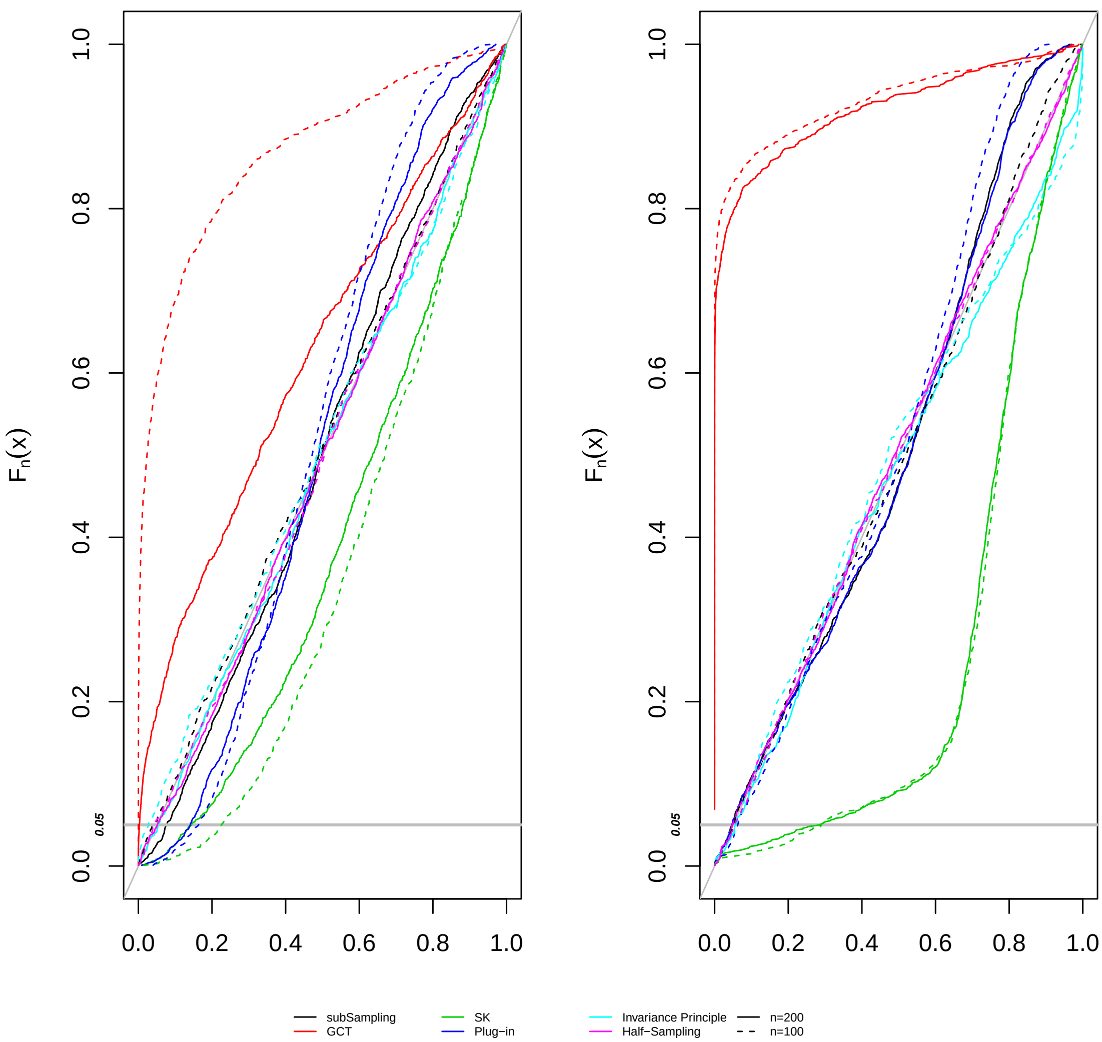


Figure 2.2: The empirical cumulative distribution function of the p-values based on 1000 simulations from Model 2. The data are generated with $a = 0.1$ and $a = 1$ in the left and right panels respectively.

CHAPTER 3

ESTIMATING HIGH-DIMENSIONAL DYNAMIC GRAPHS

3.1 Introduction

Networks are useful tools to visualize the relational information among a large number of variables. Edges in the partial correlation network have meaningful statistical interpretation corresponding to non-zero partial correlations, which corresponds to the non-zero entries in the inverse covariance matrix (precision matrix) (c.f. Lauritzen [1996], Peng et al. [2009]). There is a large volume of literature on estimating static network or precision matrix in the high-dimensional context $\min(n, p) \to \infty$, where n is the sample size and p is the number of variables; see Meinshausen and Bühlmann [2006], Friedman et al. [2008], Banerjee et al. [2008], Rothman et al. [2008], Yuan [2010], Yuan and Lin [2007], Ravikumar et al. [2008], Candès and Tao [2007], Cai et al. [2011], Cai and Liu [2011], Fan et al. [2009], Basu et al. [2015], Loh and Bühlmann [2014], Loh et al. [2013] among many others.

Most of the earlier work in graphical modeling assumes that the underlying network is time-invariant, which can be quite restrictive in practice. It rules out many real-world applications in systems such as gene networks, social networks, and stocking market, where the underlying data generating mechanisms are dynamic (c.f. Lèbre et al. [2010], Przytycka et al. [2010], Khandani and Lo [2011], Chi et al. [2010]), with potential slow-varying feature and abrupt changes. Despite the recent successful attempts towards the more flexible time-varying models (Zhou et al. [2010], Kolar and Xing [2011], Kolar et al. [2010], Kolar and Xing [2014], Qiu et al. [2015], Lu et al. [2015], Ahmed and Xing [2009], Tibshirani et al. [2005]), there are still a number of major challenges in the current high-dimensional literature. First, theoretical analysis were performed under the fundamental assumption that the observations are either *independent*, or the temporal dependence has very specific forms such as Gaussian processes or vector autoregression (e.g.Zhou et al. [2010], Kolar and Xing [2011], Basu et al. [2015], Cho and Fryzlewicz [2015], Roy et al. [2014], Qiu et al. [2015], Zhou [2014]). Second, few existing results consider distribution of the data heavier than sub-Gaussian. For continuous observations, sub-Gaussian tails were extensively assumed by Zhou et al. [2010], Kolar and Xing [2011], Cho and Fryzlewicz [2015]; for discrete observations, Markov random fields, a special case of sub-Gaussian distributions, were exclusively considered in Roy et al. [2014], Kolar and Xing [2014]. Both requirements are unduly demanding in view that a lot of time series encountered in real applications are nonlinear Tong [1993], Fan and Yao [2003]. Third, in addressing the change-point estimation problem in high-dimensional time series, piecewise constancy that is widely assumed (e.g. Roy et al. [2014], Cho and Fryzlewicz [2015], Fryzlewicz [2014], Kokoszka and Leipus [2000]) can be violated easily. For instance, financial data often appears to have time-dependent cross-volatilities with structural breaks (Aue et al. [2009]). For resting-state fMRI signals, correlation analysis reveals both slowly varying and abrupt changing characteristics corresponding to modularities in brain functional networks Chang and Glover [2010], Hutchison et al. [2013].

Explorations in high-dimensional (stationary) time series have been increasing recently; see Qiu et al. [2015], Wiesel et al. [2013], Qiu et al. [2014], Basu et al. [2015], Zhou [2014], Chen et al. [2013], Bhattacharjee and Bose [2014], Shu and Nan [2014] among many others. In Chen et al. [2013], Bhattacharjee and Bose [2014], Shu and Nan [2014], the authors considered the theoretical properties of regularized estimation of covariance and precision matrices, based on various dependence measure of high-dimensional time series. Lu et al. [2015] considered the non-paranormal graphs that evolves with a random variable. Qiu et al. [2015] discussed the joint estimation of Gaussian graphical models based on a stationary VAR(1) model with special coefficient matrices, which may also depend on certain covariates. The authors applied a CLIME estimator with a kernel estimator of covariance matrix and developed consistency in the graph recovery at a given time points. Basu et al. [2015] studied the recovery of granger causality across time and nodes assuming a stationary Gaussian vector autoregressive model with unknown order. They applied a bi-level threshold group lasso-type estimator and discussed support recovery assuming restricted eigenvalues and irrepresentative condition.

In this chapter, we focus on recovering the time-varying undirected graphs based on regularized estimation of the precision matrices. Our study allows two types of dynamics: unknown number of abrupt changes and gradual varies between the change points. In particular, we study high-dimensional *piecewise locally stationary processes* in a general nonlinear temporal dependency framework, where the observation are allowed to have only polynomial moments up to a finite order. The ultimate goal is to simultaneously recover the time-varying undirected graphs, where edges have meaningful statistical interpretation corresponding to non-zero partial correlations (c.f. Lauritzen [1996], Peng et al. [2009]). Our method has two steps. In the first step, the maximum norm of the local difference matrix is computed at each time point and the jumps in the covariance matrices are detected at the location where the maximum norms are above a certain level. In the second step, the precision matrices before and after the jump are estimated by a regularized kernel smoothing estimator. A boundary correction procedure is applied to reduce the bias near the change point.

We provide an asymptotic theory to justify the proposed method in high-dimensions: point-wise and uniform rates of convergence are established for the change-point estimation and graph recovery under mild and interpretable conditions. The rates are determined by subtle interplays of the sample size, dimensionality, temporal dependence, moment condition, and choice of bandwidth. They are significantly more involved than problems for sub-Gaussian tails and independent samples. Uniform consistent recovery of time-varying network structures is more challenging and difficult than point-wise consistency. For the multiple change point detection problem, we characterize the sharp threshold of the difference statistic that gives consistent selection of the number of change points.

We now introduce some notation. Positive constants, independent of n and p, are denoted by $C, C_1, C_2, \cdots$ and their values may differ from line to line. For the sequence of real numbers, a_n and b_n, we write $a_n = O(b_n)$ or $a_n \lesssim b_n$ if $\limsup_{n\to\infty}(a_n/b_n) \le C$ for some constant $C < \infty$ and $a_n = o(b_n)$ if $\lim_{n\to\infty}(a_n/b_n) = 0$. We say $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. For a sequence of random variables Y_n and a corresponding set of constants a_n, denote $Y_n = O_{\mathbb{P}}(a_n)$ if for any $\varepsilon > 0$ there is a constant $C > 0$ such that $\mathbb{P}(Y_n/a_n > C) < \varepsilon$ for all n. For a vector $\mathbf{x} \in \mathbb{R}^p$, we write $|\mathbf{x}| = (\sum_{j=1}^p x_j^2)^{1/2}$. For a matrix Σ, $|\Sigma|_1 = \sum_{j,k} |\sigma_{jk}|$, $|\Sigma|_\infty = \max_{j,k} |\sigma_{jk}|$, $|\Sigma|_{L_1} = \max_k \sum_j |\sigma_{jk}|$, $|\Sigma|_F = (\sum_{j,k} \sigma_{jk}^2)^{1/2}$ and $\rho(\Sigma) = \max\{|\Sigma\mathbf{x}| : |\mathbf{x}| = 1\}$. For a random vector $\mathbf{z} \in \mathbb{R}^p$, write $\mathbf{z} \in \mathcal{L}^a$, $a > 0$, if $\|\mathbf{z}\|_a =: [\mathbb{E}(|\mathbf{z}|^a)]^{1/a} < \infty$. Let $\|\mathbf{z}\| = \|\mathbf{z}\|_2$. Denote $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

3.2 Time Series Model and Assumptions

We first introduce a class of causal vector stochastic process. Let $\boldsymbol{\varepsilon}_i \in \mathbb{R}^p, i \in \mathbb{Z}$ be independent and identically distributed random vectors and $\mathcal{F}_i = (\ldots, \boldsymbol{\varepsilon}_{i-1}, \boldsymbol{\varepsilon}_i)$ be a shift process. Let $(\mathbf{X}_i^\circ(t))_{i\in\mathbb{Z}}$ be a p-dimensional stationary process, and $\mathbf{X}_i^\circ(t) = (X_{i1}^\circ(t), \cdots, X_{ip}^\circ(t))$ is generated as

$$\mathbf{X}_i^\circ(t) = \mathbf{H}(\mathcal{F}_i;\, t), \tag{3.1}$$

where $\mathbf{H}(\cdot;\cdot) = \big(H_1(\cdot;\cdot), \ldots, H_p(\cdot;\cdot)\big)$ is an $\mathbb{R}^p$-valued measurable function. We rescale the time index to $t_i = i/n$ and view $\mathbf{X}_i = \mathbf{X}_{i,n} = \mathbf{X}_i^\circ(t_i)$, $i = 1, 2, \cdots, n$. In other words,

$$\mathbf{X}_{i,n} = \mathbf{H}(\mathcal{F}_i;\, i/n). \tag{3.2}$$

We drop the subscription n in $\mathbf{X}_{i,n}$ in the rest of the chapter. Without loss of generality, the data is assumed to be preprocessed and has mean zero. Our focus is on the second-order properties.

Model (3.1) is introduced in Draghicescu et al. [2009]. The processes $\big(X_i^\circ(t)\big)_{i\in\mathbb{Z}, t\in[0,1)}$

constitute a triangular array, double indexed by i and t, and the observations $(X_i)_{i=1}^n$ are taken from the diagonal of the array. Fixing the time index t, the process $\big(X_i^\circ(t)\big)_{i\in\mathbb{Z}}$ is stationary. On the other hand, since $\mathbf{H}(\mathcal{F}_i; t_i)$ is allowed to vary with t_i, the form (3.2) is able to take into account of non-stationarity.

The process $(\mathbf{X}_i)_{i\in\mathbb{Z}}$ is causal or non-anticipative as $\mathbf{X}_i$ is an output of the past innovations $(\boldsymbol{\varepsilon}_j)_{j\le i}$ and does not depend on the future innovations. In fact, it covers a broad range of linear and nonlinear, stationary and non-stationary processes such as vector autoregressive moving average processes, locally stationary processes, Markov chains, nonlinear functional processes, *etc.* (Draghicescu et al. [2009], Zhou and Wu [2009, 2010], Chen et al. [2013]).

In application, non-stationary time series data can involve both abrupt breaks and smooth varies between the breaks. To model the non-stationarity, we assume the underlying processes are piecewise locally stationary with a finite number of structural breaks, defined as below.

Definition 3.2.1. $\mathrm{PLS}_\iota([0,1], L)$ is said to be the collection of zero-mean piecewise locally stationary processes on $[0,1]$, if for each $(X(t))_{0\le t\le 1} \in \mathrm{PLS}_\iota([0,1], L)$, there is a nonnegative integer ι such that $X(t)$ is piecewise stochastic Lipschitz continuous in t with Lipschitz constant L on the interval $[t^{(l)}, t^{(l+1)}), l = 0, \cdots, \iota$, where $0 = t^{(0)} < t^{(1)} \cdots < t^{(\iota)} < t^{(\iota+1)} = 1$. A vector process $(\mathbf{X}(t))_{0\le t\le 1} \in PLS_\iota([0,1], L)$ if all coordinates belong to $PLS_\iota([0,1], L)$. For the process $(X_0^\circ(t))_{0\le t\le 1}$ defined in (3.1), this means that there exists a non-negative integer ι and a constant $L > 0$, such that

$$\max_{1\le j\le p} \big\| H_j(\mathcal{F}_0; t) - H_j(\mathcal{F}_0; t') \big\| \le L|t - t'| \text{ for all } t^{(l)} \le t, t' < t^{(l+1)}, 0 \le l \le \iota.$$

If we assume $(\mathbf{X}_i^\circ(t))_{0\le t\le 1} \in \mathrm{PLS}_\iota([0,1], L), i \in \mathbb{Z}$, then it follows that for each $i' =$

$i-k,\ldots,i+k$, where $k/n \to 0$, and that $t^{(l)} \le i, i' < t^{(l+1)}$ for some $0 \le l \le \iota$, we have

$$\max_{1\le j\le p} \|H_j(\mathcal{F}_{i'}; i/n) - H_j(\mathcal{F}_{i'}; i'/n)\| \le Lk/n = o(1).$$

In other words, within a locally stationary time period, in a local window of i, $(X_{i'j})_{i-k\le i'\le i+k}$ can be approximated by the stationary process $(X^{\circ}_{i'j}(i/n))_{i-k\le i'\le i+k}$ for each $j = 1,\ldots,p$.

The covariance matrix function of the underlying process is $\Sigma(t) = \big(\sigma_{jk}(t)\big)_{1\le j,k\le p}$, $t \in [0,1]$, where $\sigma_{jk}(t) = \mathbb{E}\big(H_j(\mathcal{F}_0;t)H_k(\mathcal{F}_0;t)\big)$, and the precision matrix function is $\Omega(t) = \Sigma(t)^{-1} = \big(\omega_{jk}(t)\big)_{1\le j,k\le p}$. The graph at time t is denoted by $G(t) = (\mathcal{V}, \mathcal{E}(t))$, where $\mathcal{V}$ is the vertex set and $\mathcal{E}(t) = \{(j,k) : \omega_{jk}(t) \ne 0\}$. Note that $(\mathbf{X}^{\circ}_i(t))_t \in \mathrm{PLS}_\iota([0,1], L), i \in \mathbb{Z}$ implies piecewise Lipschitz continuity in $\Sigma(t)$ except at the breaks $t^{(1)},\cdots,t^{(\iota)}$. In particular, if $\sup_{0\le t\le 1}\max_{1\le j\le p}\big\|H_j(\mathcal{F}_0;t)\big\| \le C$ for some constant $C > 0$, we have that

$$|\Sigma(s) - \Sigma(t)|_\infty \;\le\; 2CL|s-t|, \qquad \forall s,t \in [t^{(l)}, t^{(l+1)}), l = 0,\ldots,\iota. \tag{3.3}$$

The other direction is not necessarily true, i.e., (3.3) does not indicate $(\mathbf{X}^{\circ}_i(t))_t \in \mathrm{PLS}_\iota([0,1], L)$, $i \in \mathbb{Z}$ in general. As a trivial example, let $\varepsilon_{ij} = 2^{-1/2}$ with probability $2/3$ and $\sqrt{2}$ with probability $1/3$ i.i.d for all i, j. At time $t_k = k/n$, let $X^{\circ}_{ij}(t_k) = (-1)^k\sqrt{t_k}\varepsilon_{ij}$. Then we find that for any k and k' such that $k + k'$ is odd, $|\Sigma(t_k) - \Sigma(t_{k'})|_\infty = |t_k - t_{k'}|$ but $\|X^{\circ}_{01}(t_k) - X^{\circ}_{01}(t_{k'})\|_2 = \sqrt{t_k} + \sqrt{t_{k'}}$.

Assumption 3.2.1. *(i) Assume $(\mathbf{X}^{\circ}_i(t))_{0\le t\le 1} \in \mathrm{PLS}_\iota([0,1], L)$ for each $i \in \mathbb{Z}$, where $L > 0$ and $\iota \ge 0$ are constants independent of n and p.*

(ii) For each $l = 0,\ldots,\iota$, and $1 \le j,k \le p$, we have $\sigma_{jk}(t) \in \mathcal{C}^2[t^{(l)}, t^{(l+1)})$.

Now we introduce the temporal dependence measure. We quantify the dependence of $\big(\mathbf{X}^{\circ}_i(t)\big)_{i\in\mathbb{Z}}$ by the dependence adjusted norm (DAN); c.f.Wu and Wu [2016]. Let $\boldsymbol{\varepsilon}'_i$ be an i.i.d. copy of $\boldsymbol{\varepsilon}_i$ and $\mathcal{F}_{i,\{m\}} = (\cdots, \boldsymbol{\varepsilon}_{i-m-1}, \boldsymbol{\varepsilon}'_{i-m}, \boldsymbol{\varepsilon}_{i-m+1}, \cdots, \boldsymbol{\varepsilon}_i)$. Denote $\mathbf{X}^{\circ}_{i,\{m\}}(t) = \big(X^{\circ}_{i1,\{m\}}(t),\ldots,X^{\circ}_{ip,\{m\}}(t)\big)$, where $X^{\circ}_{ij,\{m\}}(t) = H_j(\mathcal{F}_{i,\{m\}};t)$, $1 \le j \le p$. Here $\mathbf{X}^{\circ}_{i,\{m\}}(t)$

is a coupled version of $\mathbf{X}_i^\circ(t)$, with exactly the same generating mechanism and input, except $\boldsymbol{\varepsilon}_{i-m}$ being replaced by its iid copy $\boldsymbol{\varepsilon}'_{i-m}$.

Definition 3.2.2. Let constants $a \geq 1, A > 0$. Assume $\sup_{0\leq t\leq 1} \|X_{1j}^\circ(t)\|_a < \infty, j = 1,\ldots,p$. Define the uniform functional dependence measure for the sequences $(X_{ij}^\circ(t))_{i\in\mathbb{Z},t\in[0,1]}$ of form (3.1) as

$$\theta_{m,a,j} = \sup_{0\leq t\leq 1} \|X_{ij}^\circ(t) - X_{ij,\{m\}}^\circ(t)\|_a, \quad j = 1,\cdots,p,$$

and $\Theta_{m,a,j} = \sum_{i=m}^\infty \theta_{i,a,j}$. The dependence adjusted norm of $(X_{ij}^\circ(t))_{i\in\mathbb{Z},t\in[0,1]}$ is defined as

$$\left\|X_{\cdot,j}\right\|_{a,A} = \sup_{m\geq 0} (m+1)^A \Theta_{m,a,j},$$

whenever $\left\|X_{\cdot,j}\right\|_{a,A} < \infty$.

Intuitively, the physical dependence measure quantifies the adjusted stochastic difference between the random variable and its coupled version by replacing past innovations. Indeed, $\theta_{m,a,j}$ measures the impact on $X_{ij}^\circ(t)$ uniform over t by replacing $\boldsymbol{\varepsilon}_{i-m}$ while freezing all the other inputs, while $\Theta_{m,a,j}$ quantifies the cumulative influence of replacing $\boldsymbol{\varepsilon}_{-m}$ on $(X_{ij}^\circ(t))_{i\geq 0}$ uniform over t. DAN $\left\|X_{\cdot,j}\right\|_{a,A}$ controls the uniform polynomial decay in the lag of the cumulative physical dependence, where a depends on the the tail of marginal distributions of $X_{1,j}^\circ(t)$ and A quantifies the polynomial decay power, and hence the temporal dependence strength. It is apparent that $\left\|X_{\cdot,j}\right\|_{a,A}$ is a semi-norm, i.e., it has absolute scalability and satisfies the triangular inequality.

Assumption 3.2.2. *Let $\mathbf{X}_i^\circ(t)$ be defined in (3.1) and $\mathbf{X}_i$ in (3.2). There exist $q > 2$ and $A > 0$ such that*

$$\nu_{2q} := \sup_{t\in[0,1]} \max_{1\leq j\leq p} \mathbb{E}|X_j^\circ(t)|^{2q} < \infty \qquad \textit{and} \qquad N_{X,2q} := \max_{1\leq j\leq p} \left\|X_{\cdot,j}\right\|_{2q,A} < \infty. \quad (3.4)$$

We let $M_{X,q} := \left(\sum_{1\le j\le p} \|X_{\cdot,j}\|_{2q,A}^q\right)^{1/q}$ and write $N_X = N_{X,4}$, $M_X = M_{X,2}$. The quantity $M_{X,q}$ and $N_{X,2q}$ measures the L^q aggregated effect and the largest effect of the element-wise DANs respectively. Both quantities play a role in the convergence rates of our estimator.

Apparently we have that $\|X_{ij} - X_{ij,\{m\}}\|_a \le \theta_{m,a,j}$ and that $\max_{1\le j\le p} \mathbb{E}|X_{ij}|^{2q} \le \nu_{2q}$ for all $1 \le i \le n$. In contrast to other works in high-dimensional covariance matrix and network estimation, where sub-Gaussian tails and independence are the keys to ensure consistent estimation, Assumption 3.2.2 only requires that the time series have finite polynomial moment and allows linear and nonlinear processes with short memory in the time domain.

Example 3.2.1. Consider the following linear process model

$$\mathbf{H}(\mathcal{F}_i; t) = \sum_{m=0}^{\infty} A_m(t)\boldsymbol{\varepsilon}_{i-m},$$

where $\boldsymbol{\varepsilon}_i = (\varepsilon_1, \ldots, \varepsilon_p)$ and ε_{ij} are iid with mean 0 and variance 1, and $\|\varepsilon_{ij}\|_q \le C_q$ for each $i \in \mathbb{Z}$ and $1 \le j \le p$ with some constant $q > 2$ and $C_q > 0$. The linear process is commonly seen in literature and application. It includes the vector autoregressive model as a special example. Assume the coefficient matrices $A_m(t) = (a_{m,jk}(t))_{1\le j,k\le p}, m = 0, 1, \ldots$ satisfy the following condition.

(A1) For each $1 \le j, k \le p$, $a_{m,jk}(t) \in \mathcal{C}^2[0, 1]$.

(A2) For each $1 \le j \le p$, there is a constant $C_{A,j} > 0$ such that for each $t \in [0, 1]$, $\sum_{k=1}^p a_{m,jk}(t)^2 \le C_{A,j}(m + 1)^{-2(A+1)}$ for all $m \ge 0$.

(A3) For any $t, t' \in [0, 1]$, $\sum_{m=0}^{\infty} \sum_{k=1}^p [a_{m,jk}(t) - a_{m,jk}(t')]^2 \le L^2|t - t'|^2$ for each $j = 1, \ldots, p$.

We have that

$$\sigma_{jk}(t) = \sum_{m\geq 0} A_{m,j\cdot}^{\top}(t) A_{m,k\cdot}(t),$$

$$\Theta_{m,q,j} \leq 2C_q\sqrt{q-1}\sum_{m=0}^{\infty}(A_{m,j\cdot}^{\top}A_{m,j\cdot})^{1/2},$$

$$\|X_{ij}^{\circ}(t) - X_{ij}^{\circ}(t')\|^2 = \sum_{m=0}^{\infty} A_{m,j\cdot}\sum_{k=1}^{p}[a_{m,jk}(t) - a_{m,jk}(t')]^2,$$

where $A_{m,j\cdot}(t)$ is the jth row of $A_m(t)$.

Under condition (A1)-(A3), one can easily verify that for each $1 \leq j,k \leq p$, the process has (1) $\sigma_{jk}(t) \in \mathcal{C}^2[0,1]$; (2) $\|X_{\cdot,j}\|_{q,A} \leq \sqrt{(q-1)C_{A,j}}C_q$ by the Burkholder inequality (Chow and Teicher [1997]); and (3) $\|H_j(\mathcal{F}_0;t) - H_j(\mathcal{F}_0;t')\| \leq L|t-t'|$.

Conditions (A1)-(A3) implicitly assume smoothness in each entry of the coefficient matrices, sparseness in each column of the entry and evolution, and polynomial decay rate in the lag m of each entry and its derivative.

For $1 \leq l \leq \iota$, let $\delta_{jk}(t^{(l)}) := \sigma_{jk}(t^{(l)}) - \sigma_{jk}(t^{(l)}-)$ and $\Delta(t^{(l)}) = \big(\delta_{jk}(t^{(l)})\big)_{1\leq j,k\leq p}$, where $\sigma_{jk}(t^{(l)}-) = \lim_{t\to t^{(l)}-}\sigma_{jk}(t)$ exists as a consequence of (3.3). We assume that the change points are distinguishable and separated.

Assumption 3.2.3. *There exist positive constants* $c_1 \in (0,1)$ *and* $c_2 > 0$ *independent of* n *and* p *such that* $\max_{0\leq l\leq\iota}(t^{(l+1)} - t^{(l)}) \geq c_1$ *and* $\delta(t_l) := |\Delta(t_l)|_\infty \geq c_2$.

In the high-dimensional context, we assume that the inverse covariance matrices are sparse in the sense of their L_1 norms.

Assumption 3.2.4. *The precision matrix* $|\Omega(t)|_{L^1} \leq \kappa_p$ *for each* $t \in [0,1]$, *where* κ_p *is allowed to grow with* p.

If we further assume that the eigenvalues of the covariance matrices are bounded from below and above, i.e., there exists a constant $0 < c < 1$ such that $c \leq \inf_{t\in[0,1]}|\Sigma(t)|_2 \leq$

$\sup_{t\in[0,1]} |\Sigma(t)|_2 \le c^{-1}$, then the covariance matrices and precision matrices are well-conditioned. In particular, as $|\Omega(t) - \Omega(t')| \le c^{-2}|\Sigma(t) - \Sigma(t')|$, a small perturbation in the covariance matrix would cause no big jump in the precision matrix (in the sense of the spectral norm).

3.3 Method: Change Point Detection and Support Recovery

In Gaussian graphical models, the network structure is a second-order feature of the data. Specifically, the appearance of edges coincides with the support of the inverse covariance matrix. Same thing happens in the context of partial correlation graph. Detection of the potential change points is necessary in view of $|\Omega(t)-\Omega(t-)|_\infty \ge |\Sigma(t)|_{L^1}^{-1}|\Sigma(t-)|_{L^1}^{-1}|\Delta(t)|_\infty$, i.e., a non-negligible abrupt change in the covariance matrix will result in a substantial change of the graph structure for sparse covariance matrices. More importantly, recovering the graph structure using time series data that has abrupt jumps in the covariance matrix will induce remarkable error, as the estimates of the covariance components are themselves inaccurate. In other words, local smoothness in the covariance matrix is required for the inference of the graphical structure via constraint l_1 minimization estimator.

Proposition 3.4.1 and 3.4.2 in Section 3.4 show that under appropriate conditions, if each element of the covariance matrix varies smoothly in time, one can obtain accurate snapshot estimation of the precision matrices as well as the time-varying graphs with high probability via a kernel smoothed constrained l_1 minimization approach. As a consequence, our graph recovery method consists of two steps: change point detection and support recovery.

Let $h \equiv h_n > 0$ be the bandwidth such that $h \to 0$ and $n^{-1} \ll h$ as $n \to \infty$, and $\mathcal{D}_h(0) = \{h, h+1/n, \cdots, 1-h\}$ be the search grid. Define

$$D(s) = \frac{1}{n}\left(\sum_{i=0}^{hn-1} \mathbf{X}_{ns-i}\mathbf{X}_{ns-i}^\top - \sum_{i=1}^{hn} \mathbf{X}_{ns+i}\mathbf{X}_{ns+i}^\top\right), \qquad s \in \mathcal{D}_h(0). \tag{3.5}$$

To estimate the change points, compute

$$\hat{s}_1 = \text{argmax}_{s \in \mathcal{D}_h(0)} |D(s)|_\infty. \tag{3.6}$$

The following steps are performed recursively. For $l = 1, 2, \ldots$, let

$$\mathcal{D}_h(l) = \mathcal{D}_h(l-1) \cap \{\hat{s}_l - 2h, \cdots, \hat{s}_l + 2h\}^c, \tag{3.7}$$

$$\hat{s}_{l+1} = \arg\max_{s \in \mathcal{D}_h(l)} |D(s)|_\infty, \tag{3.8}$$

until the following criterion is attained:

$$\max_{s \in \mathcal{D}_h(l)} |D(s)|_\infty < \nu, \tag{3.9}$$

where ν is the early stopping threshold. Here ν is taken according to the error level, which as described in Section 3.4, depends on the dimension and sample size, as well as the serial dependence level, tail condition and local smoothness. Since our method only utilizes data in the localized neighborhood, ranked multiple change points can be estimated in a single pass, which offers the computational advantage than the binary segmentation algorithm Cho and Fryzlewicz [2015], Fryzlewicz [2014].

Once the change points are claimed, in the second step, we apply a time-varying (tv-) CLIME estimator for the covariance matrix functions of the multiple pieces of locally stationary processes before and after the structural breaks (Cai et al. [2011]). Let

$$\hat{\Sigma}(t) = \sum_{i=1}^{n} w(t, t_i) \mathbf{X}_i \mathbf{X}_i^\top, \tag{3.10}$$

where

$$w(t,i) = \frac{K_b(t_i,t)}{\sum_{i=1}^{n} K_b(t_i,t)} \tag{3.11}$$

and $K_b(u,v) = K(|u-v|/b)/b$. The bandwidth b satisfies that $b \to 0$ and $n^{-1} \ll b$ as $n \to \infty$. Also denote $B_n = nb$. The kernel function $K(\cdot)$ is chosen to have properties as follows.

Assumption 3.3.1. *The kernel function $K(\cdot)$ is non-negative, symmetric, and Lipschitz continuous with bounded support in $[-1,1]$, and that $\int_{-1}^{1} K(u)du = 1$.*

Assumption 3.3.1 is a common requirement on the kernel functions and can be fulfilled by a range of kernel functions such as the uniform kernel, triangular kernel, Epanechnikov kernel, *etc.*

The tv-CLIME estimator of the precision matrix $\Omega(t)$ is defined by $\tilde{\Omega}(t) = \big(\tilde{\omega}_{jk}(t)\big)_{1\le j,k\le p}$, where $\tilde{\omega}_{jk}(t) = \min(\hat{\omega}_{jk}(t), \hat{\omega}_{kj}(t))$, and $\hat{\Omega}(t) \equiv \hat{\Omega}_\lambda(t) = (\hat{\omega}_{jk}(t))_{1\le j,k\le p}$,

$$\hat{\Omega}_\lambda(t) = \arg\min_{\Omega\in\mathbb{R}^{p\times p}} |\Omega|_1 \quad \text{s.t. } |\hat{\Sigma}(t)\Omega - \mathrm{Id}_p|_\infty \le \lambda. \tag{3.12}$$

Then, the network is estimated by the "effective support" defined as follows.

$$\hat{G}(t;u) = (\hat{g}_{jk}(t;u))_{1\le j,k\le p}, \quad \text{where } \hat{g}_{jk}(t;u) = \mathbb{I}\left\{|\tilde{\omega}_{jk}(t)| \ge u\right\}. \tag{3.13}$$

It should be note that the kernel smoothing estimator of the covariance matrix does not reflect the abrupt jump in the covariance matrix at each change point (c.f. Assumption 3.2.3), as it element-wisely smoothly spreads the point mass around the investigated time point. In addition, in the neighborhood of the change points, non-negligible bias will be induced in estimating $\Sigma(t)$ by $\hat{\Sigma}(t)$ due to the jump. As a simple remedy, we apply the following reflection method for boundary correction. Suppose $t \in \hat{\mathcal{T}}_{b+h^2}(j)$ for $1 \le j \le \iota$, Denote $\hat{\mathcal{T}}_d(j) := [\hat{s}_j - d, \hat{s}_j + d)$ for $d \in (0,1)$. We replace (3.10) by

$$\hat{\Sigma}(t) = \sum_{i=1}^{n} w(t, t_i)\breve{\mathbf{x}}_i\breve{\mathbf{x}}_i^\top,$$

and then apply the rest of the tv-CLIME approach. Here

$$\breve{\mathbf{x}}_i = \begin{cases} \mathbf{x}_i & \text{if } (i - \hat{s}_j n)(t - \hat{s}_j n) \geq 0; \\ \mathbf{x}_{2\hat{s}_j n - i} & \text{otherwise.} \end{cases} \tag{3.14}$$

3.4 Theoretical Results

Our first result is that under appropriate conditions, element-wise local smoothness of the covariance matrix is sufficient for the consistency of support recovery via tv-CLIME. The latter observation motivates us to first identify the abrupt changes in covariance matrices in the sense of the max norm.

Define $J_{q,A}(n,p) = M_{X,q}(p\varpi_{q,A}(n))^{\frac{1}{q}}$, where $\varpi_{q,A}(n) = n, n(\log n)^{1+2q}, n^{q/2-Aq}$ if $A > 1/2 - 1/q$, $A = 1/2 - 1/q$, and $0 < A < 1/2 - 1/q$, respectively.

Proposition 3.4.1. *Suppose Assumptions 3.2.2, 3.2.4 and 3.3.1 hold with $\iota = 0$. Let $B_n = bn$ for $n^{-1} \ll b \ll 1$. Choose the parameter $\lambda^\circ \geq C\kappa_p(b^2 + B_n^{-1}J_{q,A}(B_n,p) + N_X(\log p/B_n)^{1/2})$ in the tv-CLIME estimator $\hat{\Omega}_{\lambda^\circ}(t)$ in (3.12), where C is a sufficiently large constant independent of n and p. Then for any $t \in [b, 1-b]$, we have*

$$|\hat{\Omega}_{\lambda^\circ}(t) - \Omega(t)|_\infty = O_{\mathbb{P}}(\kappa_p\lambda^\circ). \tag{3.15}$$

Furthermore, if we choose $\lambda^\diamond \geq C\kappa_p\left(b^2 + B_n^{-1}J_{q,A}(n,p) + N_X B_n^{-1}(n\log(p))^{1/2}\right)$, then

$$\sup_{t\in[b,1-b]} |\hat{\Omega}_{\lambda^\diamond}(t) - \Omega(t)|_\infty = O_{\mathbb{P}}(\kappa_p\lambda^\diamond). \tag{3.16}$$

The optimal order of the bandwidth parameter $b = b_\sharp$ is the solution to the following equation

$$b^2 \;=\; B_n^{-1} \max(J_{q,A}(n,p),\, N_X(n \log(p^2))^{1/2}).$$

In fact, closed-form expression for $b_\sharp$ is available. For constants C_1 and C_2 that are independent of n and p, take

$$b_\sharp = C_1\big(n^{-1}J_{q,A}(n,p)\big)^{1/3} + C_2 N_X^{1/3} n^{-1/6}\log(p)^{1/6}.$$

Given a finite sample, to distinguish the low entries in the precision matrix from the noise is challenging. In addition, a low magnitude of a certain element of the precision matrix implies weak connection of the edge in the graphical model. As a result, we only consider the estimation of *significant* edges in the graphs. Define the set of *significant* edges at level u as $\mathcal{E}^*(t;u) = \Big\{(j,k) : g^*_{jk}(t;u) \neq 0\Big\}$, where

$$g^*_{jk}(t;u) = \mathbb{I}\left\{|\omega_{jk}(t)| > u\right\}.$$

Then, as an immediate consequences of (3.16), we have the following support recovery consistency result.

Proposition 3.4.2. *Choose u as $u_\sharp = C_0\kappa_p^2 b_\sharp^2$, where C_0 is taken as a sufficiently large constant independent of n and p. Suppose that $u_\sharp = o(1)$ as $n,p \to \infty$. Then under conditions of Proposition 3.4.1, we have that as $n,p \to \infty$,*

$$\mathbb{P}\Big(\sup_{t\in[b,1-b]} \sum_{(j,k)\in\mathcal{E}^c(t)} \mathbb{I}\left\{\hat{g}_{jk}(t;u_\sharp) \neq 0\right\} \neq 0\Big) \to 0, \tag{3.17}$$

$$\mathbb{P}\Big(\sup_{t\in[b,1-b]} \sum_{(j,k)\in\mathcal{E}^*(t;2u_\sharp)} \mathbb{I}\left\{\hat{g}_{jk}(t;u_\sharp) = 0\right\} \neq 0\Big) \to 0. \tag{3.18}$$

Proposition 3.4.2 states that the pattern of significant edges in the time-varying true graphs $G(t), t \in [b, 1-b]$, can be correctly recovered with high probability. However, it is still an open question to what extent the edges with magnitude below u can be consistently estimated. Hypothesis test for the partial correlation graph in the non-stationary high-dimensional time series is rather challenging. We leave it as a future problem.

Propositions 3.4.1 and 3.4.2 yield that consistent estimation of the precision matrices and the graphs can be achieved before and after the change points. Now we provide theoretical result of the change point detection. Theorem 3.4.3 shows that if the change points are distinguishable and separated, then we can consistently identify them via the single pass segmentation approach under regular conditions. Denote

$$h_\diamond = C_1\big(n^{-1}J_{q,A}(n,p)\big)^{1/3} + C_2 N_X^{1/3} n^{-1/6}\log(p)^{1/6},$$

where C_1 and C_2 are constants independent of n and p.

Theorem 3.4.3. *Assume $\mathbf{X}_i \in \mathbb{R}^p$ admits the form (3.2). Suppose that Assumptions 3.2.2 to 3.2.3 are satisfied. Choose the bandwidth $h = h_\diamond$, and $\nu = (1+L)h_\diamond^2$ in (3.5) and (3.9) respectively. Assume that $h_\diamond = o(1)$ as $n, p \to \infty$. We have that there exist constants C_1, C_2, C_3 independent of n and p such that*

$$\mathbb{P}(|\hat{\iota} - \iota| > 0) \le C_1\big(\frac{p\varpi_{q,A}(n)M_{X,q}^q \nu_{2q}^q}{n^q c_2^q}\big)^{1/3} + C_2 p^2 \exp\big(-C_3(\frac{n\log^2(p)}{N_X^2})^{1/3}\big). \tag{3.19}$$

Furthermore, on the event $\{\iota = \hat{\iota}\}$, the ordered change-point estimator $(\hat{s}_{(1)} < \hat{s}_{(2)} < \cdots < \hat{s}_{(\hat{\iota})})$ defined in (3.7) satisfies

$$\max\nolimits_{1\le j\le \iota} |\hat{s}_{(j)} - t^{(j)}| = O_{\mathbb{P}}(h_\diamond^2). \tag{3.20}$$

Proposition 3.4.2 and Theorem 3.4.3 together indicate the consistency in the snapshot estimation of the time-varying graphs before and after the change points. Around the change points, we have the following result for the recovery of the time-varying network. Denote $\mathcal{S} := \big[b_\sharp, 1-b_\sharp\big] \cap \big(\cup_{1\le j\le \hat{\imath}} \hat{\mathcal{T}}^c_{h_\diamond^2+b_\sharp}(j)\big)$ as the time intervals between the estimated change points, and $\mathcal{N} := [0, b_\sharp) \cup \big(\cup_{1\le j\le \hat{\imath}} (\hat{\mathcal{T}}_{h_\diamond^2+b_\sharp} \cap \hat{\mathcal{T}}^c_{h_\diamond^2})\big) \cup (1-b_\sharp, 1]$ as the recoverable neighborhood of the jump.

Theorem 3.4.4. *Let Assumptions 3.2.2 to 3.3.1 be satisfied. We have the following results as $n, p \to \infty$.*

(i) For $t \in \mathcal{S}$, take $b = b_\sharp$ and $u = u_\sharp$, where $b_\sharp$ and $u_\sharp$ are defined in Proposition 3.4.2. Suppose $u_\sharp = o(1)$ as $n, p \to \infty$. we have

$$\sup_{t\in\mathcal{S}} \max_{j,k} |\hat{\sigma}_{j,k}(t) - \sigma_{j,k}(t)| = O_{\mathbb{P}}(b_\sharp^2). \tag{3.21}$$

Choose the penalty parameter as $\lambda_\sharp := C_1 \kappa_p b_\sharp^2$, where C_1 is a constant independent of n and p. Then

$$\sup_{t\in\mathcal{S}} \|\hat{\Omega}_{\lambda_\sharp}(t) - \Omega(t)\|_\infty = O_{\mathbb{P}}(\kappa_p^2 b_\sharp^2). \tag{3.22}$$

Moreover,

$$\mathbb{P}\Big(\sup_{t\in\mathcal{S}} \sum_{(j,k)\in\mathcal{E}^c(t)} \mathbb{I}\left\{\hat{g}_{j,k}(t; u_\sharp) \neq 0\right\} = 0\Big) \to 1, \tag{3.23}$$

$$\mathbb{P}\Big(\sup_{t\in\mathcal{S}} \sum_{(j,k)\in\mathcal{E}^*(t;2u_\sharp)} \mathbb{I}\left\{\hat{g}_{jk}(t; u_\sharp) = 0\right\} = 0\Big) \to 1. \tag{3.24}$$

(ii) For $s \in \mathcal{N}$, take $b = b_\star := C_1\big(n^{-1} J_{q,A}(n,p)\big)^{1/2} + C_2 N_X^{1/2} n^{-1/4} \log(p)^{1/4}$, and $u = u_\star := C_0 \kappa_p^2 b_\star$, where C_0, C_1 and C_2 are constants independent of n and p. Suppose

$u_\star = o(1)$ *as* $n, p \to \infty$. *We have*

$$\sup_{t\in\mathcal{N}} \max_{j,k} |\hat{\sigma}_{j,k}(t) - \sigma_{j,k}(t)| = O_{\mathbb{P}}(b_\star). \tag{3.25}$$

Choose the penalty parameter as $\lambda_\star := C_1 \kappa_p b_\star$, *where* C_1 *is a constant independent of* n *and* p. *Then*

$$\sup_{t\in\mathcal{N}} \|\hat{\Omega}_{\lambda_\star}(t) - \Omega(t)\|_\infty = O_{\mathbb{P}}(\kappa_p^2 b_\star). \tag{3.26}$$

Moreover,

$$\mathbb{P}\Big(\sup_{t\in\mathcal{N}} \sum_{(j,k)\in\mathcal{E}^c(t)} \mathbb{I}\left\{\hat{g}_{j,k}(t;u_\star) \neq 0\right\} = 0\Big) \to 1, \tag{3.27}$$

$$\mathbb{P}\Big(\sup_{t\in\mathcal{N}} \sum_{(j,k)\in\mathcal{E}^*(t;2u_\star)} \mathbb{I}\left\{\hat{g}_{j,k}(t;u_\star) = 0\right\} = 0\Big) \to 1. \tag{3.28}$$

Note around the jump points, the convergence rates for the covariance matrix entries and precision matrix entries in case (i) are slower than that of the between jump intervals in case (ii). This is because on the boundary due to the reflection, the smooth condition may no longer be true. Indeed, we only take advantage of the Lipschitz continuous property of the covariance matrix. Thus the bias term b^2 in the convergence rate of the between-jump area becomes b around the jumps. We note that around the smaller neighborhood of the jump $\mathcal{J} := \cup_{1\le j\le \hat{\iota}} \hat{\mathcal{T}}_{h_\diamond^2}$, due to the error in the change point detection, consistent recovery the graphs is not achievable.

3.5 A Real Data Application

In this section, we apply our method to a real financial dataset from Yahoo! Finance (`finance.yahoo.com`). The data matrix contains daily closing prices of 452 stocks that are always in the S&P 500 index between January 1, 2003 through January 1, 2008. In total, there are $n = 1258$ time points. We select 50 stocks with the largest volatility and consider their log-returns; that is, for $j = 1, \cdots, 50$,

$$X_{ij} = \log\left(p_{i+1,j}/p_{ij}\right),$$

where p_{ij} is the daily closing price of the stock j at time point i. For each stock, the log-returns X_{ij} are then standardized to zero mean and unit variance. We first compute the statistic (3.5) and (3.6) for the change point detection. We look at the top three statistics for different bandwidths. For bandwidth $n_k = 15$, we rank the test statistic and find that the top three locations for the change points are: May 26, 2006 ($n_{\hat{s}_1} = 858$), May 10, 2007 ($n_{\hat{s}_2} = 1097$), and June 12, 2003 ($n_{\hat{s}_3} = 113$). The detected change points are quite robust to a variety of choices of bandwidth, except that as the bandwidth becomes larger, the test statistic is more clustered together. Our result is partially consistent with the change point detection method in Aue et al. [2009]. In particular, the two breaks in 2006 and 2007 were also found in Aue et al. [2009] and it is conjected that the 2007 break may be associated to the U.S. house market collapse. Meanwhile, it is interesting to observe the increased volatility before the 2008 financial crisis.

Next, we estimate the time-varying networks before and after the change point at May 26, 2006 with the largest jump size. Specifically, we look at four time points at: 813, 828, 888, and 903, corresponding to March 23, 2006 April 13, 2006, July 11, 2006, and August 1, 2006. We use tv-CLIME (3.12) with Epanechnikov kernel with the same bandwidth as in the change point detection to estimate the networks at the four points. Optimal tuning parameter λ is automatically selected according to the stability approach Liu et al. [2010]. The following matrix shows the number of different edges at those four time points. It is observed that time the first two time points (813 and 828) and the last two (888 and 903) have higher similarity than across the change point at time 858. The estimated networks

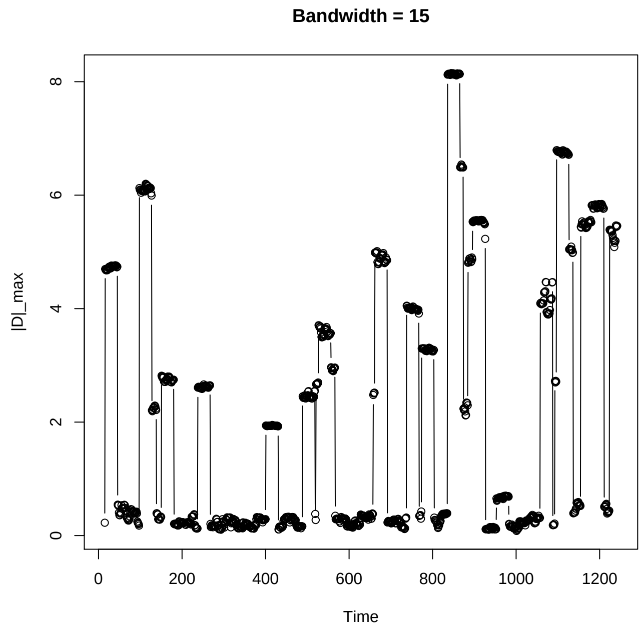
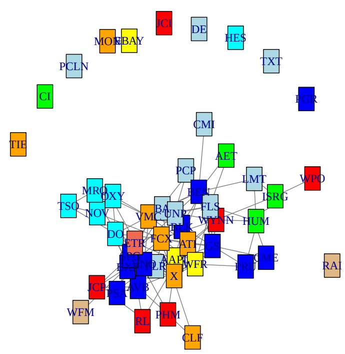


Figure 3.1: Break size $|D_s|_\infty$. From January 1, 2003 to January 1, 2008.

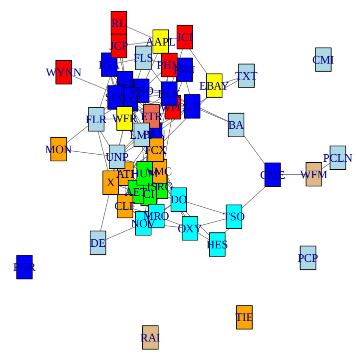
are shown in Figure 3.2. It is observed that at each time point the companies in the same section tend to be clustered together such as companies in the `Energy` section: OXY, NOV, TSO, MRO and DO (highlighted in cyan).

$$\text{Distance matrix of estimated networks: } \begin{pmatrix} 0 & 332 & 350 & 396 \\ 332 & 0 & 394 & 428 \\ 350 & 394 & 0 & 234 \\ 396 & 428 & 234 & 0 \end{pmatrix}$$



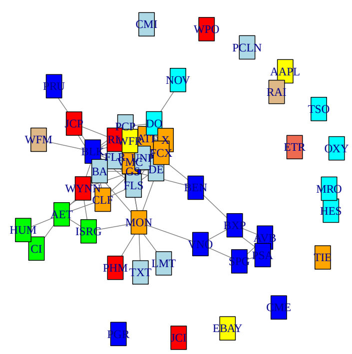
813.pdf

(a) Time 813.



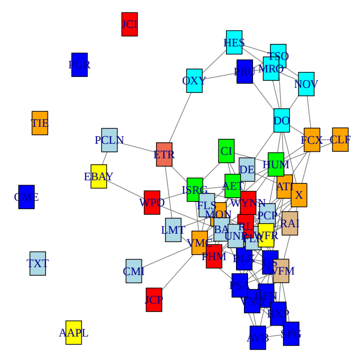
828.pdf

(b) Time 828.



888.pdf

(c) Time 888.



903.pdf

(d) Time 903.

Figure 3.2: Estimated networks at time points 813, 828, 888 and 903, corresponding to March 23, 2006, April 13, 2006, July 11, 2006, and August 1, 2006. Networks in the first and second row are estimated before and after the estimated change point at time 858, respectively. Different colors correspond to the nine sections in the S&P dataset.

3.6 Proof of the Results

3.6.1 Preliminary Lemmas

Lemma 3.6.1. *Let $(Y_i)_{i\in Z}$ be a sequence that admits (3.2). Assume $Y_i \in \mathcal{L}^q$ for $i = 1, 2, \cdots$, and the dependence adjusted norm (DAN) of the corresponding underlying array $(Y_i^\circ(t))$ satisfies $\|Y.\|_{q,A} < \infty$ for $q > 2$ and $A > 0$. Let $(\omega(t, t_i))_{i=1}^n$ be defined in (3.11) and suppose that the kernel function $K(\cdot)$ satisfies Assumption 3.3.1. Denote $\varpi_{q,A}(n) = n, n(\log n)^{1+2q}, n^{q/2-Aq}$ if $A > 1/2 - 1/q$, $A = 1/2 - 1/q$, and $0 < A < 1/2 - 1/q$, respectively. Then there exist constants C_1, C_2 and C_3 independent of n, such that for all $x > 0$,*

$$\sup_{t\in(0,1)} \mathbb{P}\left(\left|\sum_{i=1}^n w(t,t_i)\big(Y_i - \mathbb{E}(Y_i)\big)\right| > x\right) \le C_1 \frac{\varpi_{q,A}(B_n)\,\|Y.\|_{q,A}^q}{B_n^q x^q} + C_2 \exp\left(\frac{-C_3 B_n x^2}{\|Y.\|_{2,A}^2}\right). \tag{3.29}$$

$$\mathbb{P}\left(\sup_{t\in(0,1)}\left|\sum_{i=1}^n w(t,t_i)\big(Y_i - \mathbb{E}(Y_i)\big)\right| > x\right) \le C_1 \frac{\varpi_{q,A}(n)\,\|Y.\|_{q,A}^q}{B_n^q x^q} + C_2 \exp\left(\frac{-C_3 B_n^2 x^2}{n\,\|Y.\|_{2,A}^2}\right). \tag{3.30}$$

Proof. Let $S_i = \sum_{j=1}^i \big(Y_i - \mathbb{E}(Y_i)\big)$. Note that

$$\begin{aligned}
\sup_{t\in(0,1)}\left|\sum_{i=1}^n w(t,t_i)Y_i\right| &= \sup_{t\in(0,1)}\left|\sum_{i=1}^n w(t,t_i)(S_i - S_{i-1})\right| \\
&\le \sup_t \left|\sum_{i=1}^{n-1}\left[\big(w(t,t_i) - w(t,t_{i+1})\big)S_i\right]\right| + \sup_t |w(t,1)S_n| \\
&\lesssim B_n^{-1} \max_{1\le i\le n} |S_i|,
\end{aligned}$$

where the last inequality follows from the fact that $\sup_t \sum_{i=1}^n |w(t,t_i) - w(t - t_{i+1})| \asymp B_n^{-1}$

due to Assumption 3.3.1.

To see (3.30), it suffices to show

$$\mathbb{P}\left(\max_{1\le i\le n}|S_i|>x\right)\le C_1\frac{\varpi_{q,A}(n)\left\|Y_\cdot\right\|_{q,A}^q}{x^q}+C_2\exp\left(\frac{-C_3x^2}{n\left\|Y_\cdot\right\|_{2,A}^2}\right). \tag{3.31}$$

Now we develop a probability deviation inequality for $\max_{1\le i\le n}|\sum_{j=1}^i\alpha_jY_j|$, where $\alpha_j\ge 0$, $1\le j\le n$ are constants such that $\sum_{1\le j\le n}\alpha_j=1$. Denote $\mathcal{P}_0(Y_i)=\mathbb{E}(Y_i|\varepsilon_i)-\mathbb{E}(Y_i)$ and

$$\mathcal{P}_k(Y_i)=\mathbb{E}(Y_i|\varepsilon_{i-k},\ldots,\varepsilon_i)-\mathbb{E}(Y_i|\varepsilon_{i-k+1},\ldots,\varepsilon_i).$$

Then we can write

$$\begin{aligned}\max_{1\le i\le n}|\sum_{j=1}^i\alpha_jY_j|\le\max_{1\le i\le n}|\sum_{j=1}^i\alpha_j\mathcal{P}_0(Y_j)|+\max_{1\le i\le n}|\sum_{k=1}^n\sum_{j=1}^i\alpha_j\mathcal{P}_k(Y_j)|\\+\max_{1\le i\le n}|\sum_{k=n+1}^\infty\sum_{j=1}^i\alpha_j\mathcal{P}_k(Y_j)|.\end{aligned}\tag{3.32}$$

Note that $(\mathcal{P}_0(Y_j))_{j\in\mathbb{Z}}$ is an independent sequence. By Nagaev's inequality and Ottaviani's inequality, we have that

$$\begin{aligned}\mathbb{P}(\max_{1\le i\le n}|\sum_{j=1}^i\alpha_j\mathcal{P}_0(Y_j)|\ge x)&\lesssim\frac{\sum_{j=1}^n\alpha_j^q\left\|\mathcal{P}_0(Y_j)\right\|_q^q}{x^q}+\exp\big(-\frac{C_3x^2}{\sum_{j=1}^n\alpha_j^2\|\mathcal{P}_0(Y_j)\|_2^2}\big)\\&\lesssim\frac{\sum_{j=1}^n\alpha_j^q}{x^q\|Y_j\|_q}+\exp\big(-C_3\frac{x^2}{\sum_{j=1}^n\alpha_j^2}\big),\end{aligned}\tag{3.33}$$

where the last inequality holds because $\|\mathcal{P}_0(Y_j)\|_q\le 2\|Y_j\|_q$ by Jensen's inequality. Since $\sum_{j=i+1}^\infty\alpha_j\mathcal{P}_k(Y_j)$ is a martingale difference sequence with respect to $\sigma(\varepsilon_{i+1-k},\varepsilon_{i+2-k},\ldots)$, we have that $|\sum_{k=1+n}^\infty\sum_{j=i+1}^n\alpha_j\mathcal{P}_k(Y_j)|$ is a non-negative sub-martingale. Then by Doob's inequality and Burkholder's inequality, we have

$$\begin{aligned}
&\mathbb{P}\big(\max_{1\le i\le n}|\sum_{k=n+1}^{\infty}\sum_{j=1}^{i}\alpha_j\mathcal{P}_k(Y_j)|\ge x\big)\\
&\le \mathbb{P}\big(|\sum_{k=n+1}^{\infty}\sum_{j=1}^{n}\alpha_j\mathcal{P}_k(Y_j)|\ge \frac{x}{2}\big)+\mathbb{P}\big(\max_{1\le i\le n}|\sum_{k=n+1}^{\infty}\sum_{j=1+i}^{n}\alpha_j\mathcal{P}_k(Y_j)|\ge \frac{x}{2}\big)\\
&\lesssim \frac{\left\|\sum_{k=1+n}^{\infty}\sum_{j=1}^{n}\alpha_j\mathcal{P}_k(Y_j)\right\|_q^q}{x^q}\\
&\lesssim \frac{(\sum_{j=1}^{n}\alpha_j^2)^{q/2}\Theta_{n,q}^q}{x^q}\le \frac{\Theta_{n,q}^q n^{q/2-1}\sum_{j=1}^{n}\alpha_j^q}{x^q}. \qquad (3.34)
\end{aligned}$$

Now we deal with the term $\max_{1\le i\le n}|\sum_{k=1}^{n}\sum_{j=1}^{i}\alpha_j\mathcal{P}_k(Y_j)|$. Define $a_m=\min(2^m,n)$ and $M_n=\lceil \log n/\log 2\rceil$. Then

$$\max_{1\le i\le n}\Big|\sum_{k=1}^{n}\sum_{j=1}^{i}\alpha_j\mathcal{P}_k(Y_j)\Big|\le \sum_{m=1}^{M_n}\max_{1\le i\le n}\Big|\sum_{l=1}^{\lceil i/a_m\rceil}\sum_{j=1+(l-1)a_m}^{\min(la_m,i)}\sum_{k=1+a_{m-1}}^{a_m}\alpha_j\mathcal{P}_k(Y_j)\Big|. \qquad (3.35)$$

Let $\mathcal{A}_{odd}=\{1\le l\le\lceil i/a_m\rceil, l \text{ is odd}\}$ and $\mathcal{A}_{even}=\{1\le l\le\lceil i/a_m\rceil, l \text{ is even}\}$. We have

$$\mathbb{P}\big(\max_{1\le i\le n}\Big|\sum_{l=1}^{\lceil i/a_m\rceil}Z_{l,m,i}\Big|\ge x\big)\le \mathbb{P}\big(\max_{1\le i\le n}\Big|\sum_{\mathcal{A}_{odd}}Z_{l,m,i}\Big|\ge x/2\big)+\mathbb{P}\big(\max_{1\le i\le n}\Big|\sum_{\mathcal{A}_{even}}Z_{l,m,i}\Big|\ge x/2\big),$$

where we have that $Z_{l,m,i}:=\sum_{j=1+(l-1)a_m}^{\min(la_m,i)}\alpha_j\mathcal{P}_{a_{m-1}}^{a_m}(Y_j)$ is independent of $Z_{l+2,m,i}$ for $1\le l\le\lceil i/a_m\rceil, 1\le m\le M_n, 1\le i\le n$, as $\mathcal{P}_{a_{m-1}}^{a_m}(Y_j):=\sum_{k=1+a_{m-1}}^{a_m}\mathcal{P}_k(Y_j)$ is a_m-dependent. Therefore, we can apply Ottaviani's inequality and Nagaev's inequality for independent variables. As a consequence,

$$\mathbb{P}\big(\max_{1\le i\le n}\Big|\sum_{l=1}^{\lceil i/a_m\rceil}Z_{l,m,i}\Big|\ge x\big)\lesssim \frac{\sum_{1\le l\le\lceil n/a_m\rceil}\|Z_{l,m,n}\|_q^q}{x^q}+\exp\big(-\frac{C_3x^2}{\sum_{1\le l\le\lceil n/a_m\rceil}\|Z_{l,m,n}\|_2^2}\big).$$

Again, by Burkholder's inequality, we have that for $q \geq 2$,

$$\begin{aligned}\|Z_{l,m,n}\|_q &\leq \sum_{k=1+a_{m-1}}^{a_m} \| \sum_{j=1+(l-1)a_m}^{\min(la_m,n)} \alpha_j \mathcal{P}_k(Y_j)\|_q \\ &\lesssim (\sum_{j=1+(l-1)a_m}^{\min(la_m,n)} \alpha_j^2)^{1/2}(\Theta_{a_{m-1}} - \Theta_{a_m}).\end{aligned}$$

Note $\sum_{j=1+(l-1)a_m}^{\min(la_m,n)} \alpha_j^2 \leq a_m^{(q-2)/q}(\sum_{j=1+(l-1)a_m}^{\min(la_m,n)} \alpha_j^q)^{2/q}$. Let $\tau_m = m^{-2}/\sum_{m=1}^{M_n} m^{-2}$, and we have $\tau_m \asymp m^{-2}$ as $1 \leq \sum_{m=1}^{M_n} m^{-2} \leq \pi^2/6$. In respect to (3.35), we have that

$$\begin{aligned}\mathbb{P}\big(\max_{1\leq i\leq n} \big|\sum_{k=1}^{n}\sum_{j=1}^{i} \mathcal{P}_k(Y_j)\big| \geq x\big) &\leq \sum_{m=1}^{M_n} \mathbb{P}\big(\max_{1\leq i\leq n} \big|\sum_{l=1}^{\lceil i/a_m\rceil} Z_{l,m,i}\big| \geq \tau_m x\big) \\ &\lesssim \frac{\sum_{i=1}^{n} \alpha_j^q}{x^q}\|Y_\cdot\|_{q,A}^q \sum_{m=1}^{M_n} \tau_m^{-q} a_m^{(1/2-A)q-1} + \sum_{m=1}^{M_n} \exp\big(-\frac{C_3 x^2 \tau_m^2 a_m^{2A}}{\sum_{j=1}^{n} \alpha_j^2 \|Y_\cdot\|_{2,A}^2}\big).\end{aligned} \tag{3.36}$$

Note $\sum_{m=1}^{M_n} \tau_m^{-q} a_m^{(1/2-A)q-1} \asymp n^{-1}\varpi_{q,A}(n)$, and

$$\sum_{m=1}^{M_n} \exp\big(-\frac{C_3 x^2 \tau_m^2 a_m^{2A}}{\sum_{j=1}^{n} \alpha_j^2 \|Y_\cdot\|_{2,A}^2}\big) \lesssim \exp\big(-\frac{C_3 x^2}{\sum_{j=1}^{n} \alpha_j^2 \|Y_\cdot\|_{2,A}^2}\big).$$

Combining (3.32), (3.33), (3.34) and (3.36), we obtain

$$\begin{aligned}&\mathbb{P}\big(\max_{1\leq i\leq n} \big|\sum_{j=1}^{i} \alpha_j\big(Y_j - \mathbb{E}(Y_j)\big)\big| > x\big) \\ &\quad \leq C_1 \frac{\varpi_{q,A}(n)\sum_{j=1}^{n} \alpha_j^q \|Y_\cdot\|_{q,A}^q}{nx^q} + C_2 \exp\big(\frac{-C_3 x^2}{\sum_{j=1}^{n} \alpha_j^2 \|Y\|_{2,A}^2}\big).\end{aligned} \tag{3.37}$$

Now we have (3.31) by taking $\alpha_j = n^{-1}$ for $j = 1, \ldots, n$. Note that since $K(\cdot)$ has bounded support, for any given $t \in [b, 1-b]$, we have

$$\mathbb{P}\big(\big|\sum_{i=1}^{n} w(t,t_i)(Y_i - \mathbb{E}Y_i)\big| > x\big) \le \mathbb{P}\big(\big|\sum_{i=-B_n}^{B_n} w(t,t_{tn+i})(Y_{tn+i} - \mathbb{E}Y_{tn+i})\big| > x\big)$$
$$\le C_1 \frac{\varpi_{q,A}(B_n)\sum_{i=-B_n}^{B_n} w(t,t_{tn+i})^q \|Y_\cdot\|_{q,A}^q}{B_n x^q} + C_2 \exp\big(\frac{-C_3 x^2}{\sum_{i=-B_n}^{B_n} w(t,t_{tn+i})^2 \|Y_\cdot\|_{2,A}^2}\big).$$

Therefore (3.29) follows from (3.37) by taking $\alpha_j = w(t, tn+j)$ and note that Note that for any $t \in [b, 1-b]$, $\sum_{i=-B_n}^{B_n} w(t,t_{tn+i})^\beta \asymp B_n^{1-\beta}$ for a constant $\beta \ge 2$. $\square$

Lemma 3.6.2. *Suppose* $(X_{ij})_{i\in\mathbb{Z},1\le j\le p}$ *satisfy Assumption 3.2.2. Also let Assumption 3.3.1 hold. Let* $\varpi_{q,A}(n)$ *be defined as in Lemma 3.6.1. Then there exist constants* C_1, C_2 *and* C_3 *independent of* n *and* p*, such that for all* $x > 0$*, we have*

$$\sup_{t\in(0,1)} \mathbb{P}\big(\big|\sum_{i=1}^{n} \omega(t,t_i)\big(\mathbf{X}_i\mathbf{X}_i^\top - \mathbb{E}(\mathbf{X}_i\mathbf{X}_i^\top)\big)\big|_\infty \ge x\big)$$
$$\le C_1 \nu_{2q}^q \frac{p\varpi_{q,A}(B_n) M_{X,q}^q}{B_n^q x^q} + C_2 p^2 \exp\left(-C_3 \frac{B_n x^2}{\nu_4^2 N_X^2}\right), \tag{3.38}$$

and

$$\mathbb{P}\big(\sup_{t\in(0,1)} \big|\sum_{i=1}^{n} w(t,t_i)\big(\mathbf{X}_i\mathbf{X}_i^\top - \mathbb{E}(\mathbf{X}_i\mathbf{X}_i^\top)\big)\big|_\infty \ge x\big)$$
$$\le C_1 \nu_{2q}^q \frac{p\varpi_{q,A}(n) M_{X,q}^q}{B_n^q x^q} + C_2 p^2 \exp\left(-C_3 \frac{B_n^2 x^2}{n\nu_4^2 N_X^2}\right). \tag{3.39}$$

Proof. For $1 \le j,k \le p$, let $Y_{i,jk} = X_{ij}X_{ik}$. We now check the conditions in Lemma 3.6.1 for $(Y_{i,jk})_{1\le i\le n}$. Denote $Y_{i,jk,\{m\}} = X_{ij,\{m\}}X_{ik,\{m\}}$. Then the uniform functional dependence measure of $(Y_{i,jk})_i$ is

$$\begin{aligned}
\theta^Y_{m,q,jk} &= \sup_i \|Y_{i,jk} - Y_{i,jk,\{m\}}\|_q \\
&= \sup_i \|X_{ij}X_{ik} - X_{ij,\{m\}}X_{ik,\{m\}}\|_q \\
&\leq \sup_i \|X_{ij}(X_{ik} - X_{ik,\{m\}})\|_q + \sup_i \|X_{ik,\{m\}}(X_{ij} - X_{ij,\{m\}})\|_q.
\end{aligned}$$

Thus the DAN of the process $Y_{\cdot,jk}$ satisfies that

$$\|Y_{\cdot,jk}\|_{q,A} \leq \sup_i \|X_{ij}\|_{2q}\ \|X_{\cdot,k}\|_{2q,A} + \sup_i \|X_{ik}\|_{2q}\ \|X_{\cdot,j}\|_{2q,A} \leq \nu_q(\|X_{\cdot,k}\|_{2q,A} + \|X_{\cdot,j}\|_{2q,A}).$$

The result follows immediately from Lemma 3.6.1 and the Bonferroni inequality. □

Lemma 3.6.3. *We adopt the notation in Lemma 3.6.2. Suppose Assumptions 3.2.2, 3.2.1 and 3.3.1 hold with $\iota = 0$. Recall $B_n = nb$, where $b \to 0$ and $B_n/\sqrt{n} \to \infty$ as $n \to \infty$. Then there exists a constant C independent of n and p such that $\hat{\Sigma}(t)$ in (3.10) satisfies that for any $t \in [c, 1-c]$,*

$$|\hat{\Sigma}(t) - \Sigma(t)|_\infty = O_{\mathbb{P}}\left(b^2 + M_{X,q}\nu_{2q}B_n^{-1}(p\varpi_{q,A}(B_n))^{1/q} + \nu_4 N_X(\log p/B_n)^{1/2}\right). \quad (3.40)$$

Furthermore,

$$\sup_{t\in[c,1-c]} |\hat{\Sigma}(t) - \Sigma(t)|_\infty = O_{\mathbb{P}}\left(b^2 + M_{X,q}\nu_{2q}B_n^{-1}(p\varpi_{q,A}(n))^{1/q} + \nu_4 N_X B_n^{-1}[n\log p]^{1/2}\right). \quad (3.41)$$

Proof. First we have

$$\mathbb{E}\hat{\sigma}_{jk}(t) - \sigma_{jk}(t) = \sum_{i=1}^n w(t,t_i)[\sigma_{jk}(t_i) - \sigma_{jk}(t)].$$

Approximating the discrete summation with integral, we obtain for all $1 \le j, k \le p$,

$$\sup_{t\in[b,1-b]} \left| \mathbb{E}\hat{\sigma}_{jk}(t) - \sigma_{jk}(t) - \int_{-1}^{1} K(u)[\sigma_{jk}(ub+t) - \sigma_{jk}(t)]du \right| = O\left(B_n^{-1}\right).$$

By Assumption 3.2.1, we have

$$\sigma_{jk}(ub+t) - \sigma_{jk}(t) = ub\sigma'_{jk}(t) + \frac{1}{2}u^2b^2\sigma''_{jk}(t) + o(b^2u^2).$$

Thus we have $\sup_{t\in[c,1-c]} |\mathbb{E}\hat{\sigma}(t) - \sigma(t)|_\infty = O\left(B_n^{-1} + b^2\right)$, in view of Assumption 3.3.1. By Lemma 3.6.2, we have

$$\sup_{t\in(0,1)} \mathbb{P}\left(\left|\hat{\Sigma}(t) - \mathbb{E}\hat{\Sigma}(t)\right|_\infty \ge x\right) \le C_1 p \nu_q^q \frac{M_{X,q}^q \varpi_{q,A}(B_n)}{B_n^q x^q} + C_2 p^2 \exp\left(-C_3 \frac{B_n x^2}{N_X^2}\right).$$

Denote $u = C_4\big(M_{X,q}\nu_{2q}B_n^{-1}(p\varpi_{q,A}(B_n))^{1/q} + \nu_4 N_X(\log p/B_n)^{1/2}\big)$ for a large enough constant C_4, then for any $t \in (0,1)$,

$$\left|\hat{\Sigma}(t) - \mathbb{E}\hat{\Sigma}(t)\right|_\infty = O_{\mathbb{P}}(u).$$

Thus (3.40) is proved. The result (3.41) can be obtained similarly. □

3.6.2 Proof of the Main Results

Proof of Proposition 3.4.1. Given (3.40) and (3.41), the proof of (3.15) is standard. (See, e.g. [Cai et al., 2011, Theorem 6]). For λ° and λ^* given in Proposition 3.4.1, by Lemma 3.6.3, we have that respectively,

$$\lambda^\circ \ge \sup_t \mathbb{E}\big(\kappa_p|\hat{\Sigma}(t) - \Sigma(t)|_\infty\big), \tag{3.42}$$

$$\lambda^\diamond \ge \mathbb{E}\big(\kappa_p \sup_t |\hat{\Sigma}(t) - \Sigma(t)|_\infty\big). \tag{3.43}$$

Then note that for any $t \in [0,1]$, for any $\lambda > 0$,

$$
\begin{aligned}
|\hat{\Omega}_\lambda(t) - \Omega(t)|_\infty &\le |\Omega(t)|_{L_1} |\Sigma(t)\hat{\Omega}_\lambda(t) - \mathrm{Id}_p|_\infty \\
&\le |\Omega(t)|_{L_1} \big[|\hat{\Sigma}(t)\hat{\Omega}_\lambda(t) - \mathrm{Id}_p|_\infty + |(\Sigma(t) - \hat{\Sigma}(t))\Omega(t)|_\infty + |\hat{\Omega}_\lambda(t) - \Omega(t)|_{L^1} |\hat{\Sigma}(t) - \Sigma(t)|_\infty \big]
\end{aligned}
$$

where by construction, we have $|\hat{\Sigma}(t)\hat{\Omega}_\lambda(t) - \mathrm{Id}_p|_\infty \le \lambda$ and $|\hat{\Omega}_\lambda(t) - \Omega(t)|_{L^1} \le 2\kappa_p$. Consequently,

$$
|\hat{\Omega}_\lambda(t) - \Omega(t)|_\infty \le \kappa_p \big(\lambda + 3\kappa_p |\hat{\Sigma}(t) - \Sigma(t)|_\infty \big). \tag{3.44}
$$

Then (3.15) and (3.16) follow from (3.42) to (3.44). □

Proof of Proposition 3.4.2. Theorem 3.4.2 is an immediate result of (3.16). □

Proof of Theorem 3.4.3. Denote $r_j, 1 \le j \le \iota$ as the time point(s) of the time of jump ordered decreasingly in the sense of the infinite norm of covariance matrices, i.e., $|\Delta(r_1)|_\infty \ge |\Delta(r_2)|_\infty \ge \ldots \ge |\Delta(r_\iota)|_\infty \ge |\Delta(s)|_\infty$ for $s \in (0,1) \cap \{r_1, \ldots, r_\iota\}^c$. (Temporal order is applied if there is a tie.) Let $\mathcal{T}_h(j) = [r_j - h, r_j + h)$. For $h = o(1)$, as a result of Assumption 3.2.3, $\mathcal{T}_h(j) \cap \mathcal{T}_h(i) = \emptyset$ if $i \ne j$ for n sufficiently large. That is to say, each time point $s \in (0,1)$ is in the neighborhood of at most one change point.

For any $s \in [t^{(j)}, t^{(j+1)})$, $j = 0, 1, \ldots, \iota$, denote $\mathbb{D}(s) = \mathbb{E}[D(s)]$ and

$$
\mathbb{D}^\diamond(s) = \begin{cases} (h - s + t^{(j)})\Delta(t^{(j)}), & t^{(j)} \le s < t^{(j)} + h \\ 0, & t^{(j)} + h \le s < t^{(j+1)} - h \\ (h + s - r)\Delta(t^{(j+1)}), & t^{(j+1)} - h \le s \le t^{(j+1)}. \end{cases} \tag{3.45}
$$

Then, for $s \in \cup_{1 \le j \le \iota} [t^{(j)} + h, < t^{(j+1)} - h)$, by (3.3), we have

$$
|\Sigma(s+t) - \Sigma(s)|_\infty \le Lt, \qquad \forall |t| \le h,
$$

we can easily verify that

$$\sup_{s\in[0,1]} |\mathbb{D}(s) - \mathbb{D}^{\diamond}(s)|_\infty \le Lh^2. \tag{3.46}$$

Note that $|\mathbb{D}^{\diamond}(s)|_\infty$ is maximized at $s = r_1$ and $|\mathbb{D}^{\diamond}(r_1)|_\infty = h|\Delta(r_1)|_\infty$. By the triangle inequalities, we have that for some positive constant C, for any $s \in [0,1]$,

$$\begin{aligned} |\mathbb{D}(r_1)|_\infty - |\mathbb{D}(s)|_\infty &\ge hc_2 - |\mathbb{D}(r_1) - \mathbb{D}^{\diamond}(r_1)|_\infty - |\mathbb{D}^{\diamond}(s)|_\infty - |\mathbb{D}(s) - \mathbb{D}^{\diamond}(s)|_\infty \\ &\ge hc_2 - |\mathbb{D}^{\diamond}(s)|_\infty - 2Lh^2 \\ &\ge c_2(|s - r_1| \wedge h) - 2Lh^2. \end{aligned} \tag{3.47}$$

On the other hand, since $|D(r_1)|_\infty \le |D(\hat{s}_1)|_\infty$, we have

$$\begin{aligned} |\mathbb{D}(r_1)|_\infty - |\mathbb{D}(\hat{s}_1)|_\infty &\le |D(r_1)|_\infty - |D(\hat{s}_1)|_\infty + |\mathbb{D}(r_1) - D(r_1)|_\infty + |\mathbb{D}(\hat{s}_1) - D(\hat{s}_1)|_\infty \\ &\le |\mathbb{D}(r_1) - D(r_1)|_\infty + |\mathbb{D}(\hat{s}_1) - D(\hat{s}_1)|_\infty. \end{aligned} \tag{3.48}$$

Denote the event $\mathcal{A} := \{\sup_{s\in[h,1-h]} |D(s) - \mathbb{D}(s)|_\infty \le h_\diamond^2\}$ and let $\mathbf{Y_i} = (Y_{i,jk})_{1\le j,k\le p}$, $Y_{i,jk} = X_{ij}X_{ik} - \sigma_{i,jk}$. Note that

$$|D_{jk}(s) - \mathbb{D}_{jk}(s)| = \frac{1}{n}\left|\sum_{i=1}^{hn} Y_{n_s+1-i,jk} - \sum_{i=1}^{hn} Y_{n_s+i,jk}\right|. \tag{3.49}$$

By Lemma 3.6.2, we have for any $x > 0$,

$$\mathbb{P}\left(\sup_{s\in[h,1-h]} |D(s) - \mathbb{D}(s)|_\infty \ge x\right) \le C_1 \frac{p\varpi_{q,A}(n) M_{X,q}^q \nu_{2q}^q}{n^q x^q} + C_2 p^2 \exp\left(-C_3 \frac{nx^2}{N_X^2}\right). \tag{3.50}$$

It follows that

$$|\mathbb{D}(r_1)|_\infty - |\mathbb{D}(\hat{s}_1)|_\infty = O_{\mathbb{P}}\big(h^{-1} J_{q,A}(n,p) + N_X h^{-1}(n^{-1}\log(p))^{1/2}\big).$$

Taking $h = h_\diamond$, we have

$$|\hat{s}_1 - r_1| = O_{\mathbb{P}}(h_\diamond^2).$$

Furthermore we have

$$\mathbb{P}(\mathcal{A}) \geq 1 - C_1\big(\frac{p\varpi_{q,A}(n)M_{X,q}^q\nu_{2q}^q}{n^q c_2^q}\big)^{1/3} - C_2p^2\exp\big(-C_3(\frac{n\log^2(p)}{N_X^2})^{1/3}\big).$$

Let $\mathcal{A}_k := \{\max_{1\leq j\leq k}|\hat{s}_j - r_j| \leq c_2^{-1}2(L+1)h_\diamond^2\}$ for some $1 \leq k \leq \iota$. Assume $\mathcal{A}_k \subset \mathcal{A}$. Under $\mathcal{A}_k$ we have that $[r_j - h_\diamond, r_j + h_\diamond) \subset \hat{\mathcal{T}}_{2h_\diamond}(j) =: [\hat{s}_j - 2h_\diamond, \hat{s}_j + 2h_\diamond)$ for $1 \leq j \leq k$ and $r_{k+1} \notin \cup_{1\leq j\leq k}\hat{\mathcal{T}}_{2h_\diamond}(j)$ as a consequence of Assumption 3.2.3. According to (3.47) and (3.48), we have if $\mathcal{A}$ is true, $|\hat{s}_{k+1} - r_{k+1}| \leq c_2^{-1}2(L+1)h_\diamond^2$, which implies $\mathcal{A}_{k+1} \subset \mathcal{A}$. The result (3.20) follows from deduction.

Suppose $\mathcal{A}$ holds. By the choice of ν, as a consequence of (3.46) and (3.50), and that $\nu \ll h_\diamond$, we have that

$$\sup_{s\in[0,1]}|D(s) - \mathbb{D}^\diamond(s)|_\infty \leq \nu.$$

As a result,

$$\min_{1\leq j\leq \iota}|D(r_j)|_\infty \geq c_2h_\diamond - \nu \geq \nu,$$

i.e., $\hat{\iota} \geq \iota$. On the other hand, since $\cup_{1\leq j\leq \iota}\hat{\mathcal{T}}_{2h_\diamond}(j)$ is excluded from the searching region for $s_{\iota+1}$, we have

$$\sup_{s\in\left(\cup_{1\leq j\leq \iota}\hat{\mathcal{T}}_{2h_\diamond}(j)\right)^c}|D(s)|_\infty \leq \nu.$$

In other words, $\{\hat{\iota} = \iota\} \subset \mathcal{A}$. Thus (3.19) is proved.

$\square$

Proof of Theorem 3.4.4. We adopt the notations in the proof of Theorem 3.4.3 and assume that $\mathcal{E}$ holds. Similar as in Lemma 3.6.3, we have that by Lemma 3.6.2, for any $t \in (0,1)$,

$$\left|\hat{\Sigma}(t) - \mathbb{E}\hat{\Sigma}(t)\right|_{\infty} = O_{\mathbb{P}}(u),$$

where $u = C_4\big(M_{X,q}\nu_{2q}B_n^{-1}(p\varpi_{q,A}(B_n))^{1/q} + \nu_4 N_X(\log p/B_n)^{1/2}\big)$ for a large enough constant C_4.

Since under $\mathcal{E}$, $\mathcal{T}_b(j) \subset \hat{\mathcal{T}}_{b+h_\diamond^2}(j)$. For $t \in \big(\cup_{1\le j\le \iota}\hat{\mathcal{T}}_{b+h_\diamond^2}(j)\big)^c \cap [b, 1-b]$, we have that for all $1 \le j,k \le p$,

$$\begin{aligned}
\left|\mathbb{E}\hat{\sigma}_{jk}(t) - \sigma_{jk}(t)\right| &= \int_{-1}^{1} K(u)[\sigma_{jk}(ub+t) - \sigma_{jk}(t)]du + O\left(B_n^{-1}\right) \\
&= b\sigma'_{jk}(t)\int_{-1}^{1} uK(u)du + \big(\frac{1}{2}b^2\sigma''_{jk}(t) + o(b^2)\big)\int_{-1}^{1} u^2K(u)du + O\left(B_n^{-1}\right) \\
&= O(b^2 + B_n^{-1}).
\end{aligned}$$

On the other hand, for $t \in \cup_{1\le j\le\iota}\big(\hat{\mathcal{T}}_{b+h_\diamond^2}(j) \cap \mathcal{T}^c_{h_\diamond^2}(j)\big) \cup [0,b] \cup [1-b,1]$, due to reflection, we no longer have that differentiability. As a result of the Lipschitz continuity, we get

$$\left|\mathbb{E}\hat{\sigma}_{jk}(t) - \sigma_{jk}(t)\right| = \int_{-1}^{1} K(u)[\sigma_{jk}(ub+t) - \sigma_{jk}(t)]du + O\left(B_n^{-1}\right) = O(b + B_n^{-1}).$$

The result (3.25) follows by the choices of b. The rest of the proof are similar as in that of Proposition 3.4.1 and Theorem 3.4.2. □

REFERENCES

Amr Ahmed and Eric P. Xing. Recovering time-varying networks of dependencies recovering time-varying networks of dependencies in social and biological studies. *PNAS*, 106(29): 11878–11883, 2009.

Alexander Aue, Siegfried Hörmann, Lajos Horváth, and Matthew Reimherr. Break detection in the covariance structure of multivariate time series models. *Ann. Statist.*, 37(6B):4046–4087, 2009. ISSN 0090-5364. doi: 10.1214/09-AOS707. URL `http://dx.doi.org/10.1214/09-AOS707`.

Zhidong Bai and Hewa Saranadasa. Effect of high dimension: by an example of a two sample problem. *Statist. Sinica*, 6(2):311–329, 1996. ISSN 1017-0405.

Zhidong Bai and Jack W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, second edition, 2010. ISBN 978-1-4419-0660-1. doi: 10.1007/978-1-4419-0661-8. URL `http://dx.doi.org/10.1007/978-1-4419-0661-8`.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516, 2008. ISSN 1532-4435.

Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data setsupdate. *Nucleic acids research*, 41(D1):D991–D995, 2013.

Sumanta Basu, Ali Shojaie, and George Michailidis. Network Granger causality with inherent grouping structure. *J. Mach. Learn. Res.*, 16:417–453, 2015. ISSN 1532-4435.

István Berkes, Weidong Liu, and Wei Biao Wu. Komlós-Major-Tusnády approximation under dependence. *Ann. Probab.*, 42(2):794–817, 2014. ISSN 0091-1798. doi: 10.1214/13-AOP850. URL `http://dx.doi.org/10.1214/13-AOP850`.

Monika Bhattacharjee and Arup Bose. Consistency of large dimensional sample covariance matrix under weak dependence. *Stat. Methodol.*, 20:11–26, 2014. ISSN 1572-3127. doi: 10.1016/j.stamet.2013.08.005. URL `http://dx.doi.org/10.1016/j.stamet.2013.08.005`.

Peter J. Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227, 2008a. ISSN 0090-5364. doi: 10.1214/009053607000000758. URL `http://dx.doi.org/10.1214/009053607000000758`.

Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *Ann. Statist.*, 36(6):2577–2604, 2008b. ISSN 0090-5364. doi: 10.1214/08-AOS600. URL `http://dx.doi.org/10.1214/08-AOS600`.

Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.*, 106(494):672–684, 2011. ISSN 0162-1459. doi: 10.1198/jasa.2011.tm10560. URL `http://dx.doi.org/10.1198/jasa.2011.tm10560`.

Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106(494):594–607, 2011. ISSN 0162-1459. doi: 10.1198/jasa.2011.tm10155. URL `http://dx.doi.org/10.1198/jasa.2011.tm10155`.

Emmanuel Candès and Terence Tao. Rejoinder: "The Dantzig selector: statistical estimation when p is much larger than n" [Ann. Statist. **35** (2007), no. 6, 2313–2351; mr2382644]. *Ann. Statist.*, 35(6):2392–2404, 2007. ISSN 0090-5364. doi: 10.1214/009053607000000532. URL `http://dx.doi.org/10.1214/009053607000000532`.

Catie Chang and Gary H. Glover. Time-frequency dynamics of resting-state brain connectivity measured with fMRI. *NeuroImage*, 50:81–98, 2010.

Lin S. Chen, Debashis Paul, Ross L. Prentice, and Pei Wang. A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. *J. Amer. Statist. Assoc.*, 106(496):1345–1360, 2011. ISSN 0162-1459. doi: 10.1198/jasa.2011.ap10599. URL `http://dx.doi.org/10.1198/jasa.2011.ap10599`.

Song Xi Chen and Ying-Li Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.*, 38(2):808–835, 2010. ISSN 0090-5364. doi: 10.1214/09-AOS716. URL `http://dx.doi.org/10.1214/09-AOS716`.

Xiaohui Chen, Mengyu Xu, Wei Biao Wu, et al. Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics*, 41(6):2994–3021, 2013.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819, 2013. ISSN 0090-5364. doi: 10.1214/13-AOS1161. URL `http://dx.doi.org/10.1214/13-AOS1161`.

K Tse Chi, Jing Liu, and Francis CM Lau. A network perspective of the stock market. *Journal of Empirical Finance*, 17(4):659–667, 2010.

Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507, 2015.

Yuan Shih Chow and Henry Teicher. *Probability theory.* Springer Texts in Statistics. Springer-Verlag, New York, third edition, 1997. ISBN 0-387-98228-0. doi: 10.1007/978-1-4612-1950-7. URL `http://dx.doi.org/10.1007/978-1-4612-1950-7`. Independence, interchangeability, martingales.

Dana Draghicescu, Serge Guillas, and Wei Biao Wu. Quantile curve estimation and visualization for nonstationary time series. *Journal of Computational and Graphical Statistics*, 18(1):1–20, 2009.

Jianqing Fan and Qiwei Yao. *Nonlinear Time Series: Nonparmatric and Parametric Methods*. Springer-Verlag, 2003.

Jianqing Fan, Peter Hall, and Qiwei Yao. To how many simultaneous hypothesis tests can normal, Student's t or bootstrap calibration be applied? *J. Amer. Statist. Assoc.*, 102(480):1282–1288, 2007. ISSN 0162-1459. doi: 10.1198/016214507000000969. URL http://dx.doi.org/10.1198/016214507000000969.

Jianqing Fan, Yang Feng, and Yichao Wu. Network Exploration via the Adaptive Lasso and SCAD penalties. *The Annals of Applied Statistics*, 3(2):521–541, 2009.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 75(4): 603–680, 2013a. ISSN 1369-7412. doi: 10.1111/rssb.12016. URL http://dx.doi.org/10.1111/rssb.12016. With 33 discussions by 57 authors and a reply by Fan, Liao and Mincheva.

Jianqing Fan, Yuan Liao, and Jiawei Yao. Power enhancement in high dimensional cross-sectional tests. *Available at SSRN 2340313*, 2013b.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, 9(3):432–441, 2008.

Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

Stuart Geman. A limit theorem for the norm of random matrices. *Ann. Probab.*, 8(2): 252–261, 1980. ISSN 0091-1798.

Stelios Georgiou, Christos Koukouvinos, and Jennifer Seberry. Hadamard matrices, orthogonal designs and construction algorithms. In *Designs 2002*, pages 133–205. Springer, 2003.

Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013. ISBN 978-1-4214-0794-4; 1-4214-0794-9; 978-1-4214-0859-0.

Karl Bruce Gregory, Raymond J. Carroll, Veerabhadran Baladandayuthapani, and Soumendra N. Lahiri. A two-sample test for equality of means in high dimension. *J. Amer. Statist. Assoc.*, 110(510):837–849, 2015. ISSN 0162-1459. doi: 10.1080/01621459.2014.934826. URL `http://dx.doi.org/10.1080/01621459.2014.934826`.

Peter Hall. A short prehistory of the bootstrap. *Statistical Science*, pages 158–167, 2003.

J. A. Hartigan. Using subsample values as typical values. *J. Amer. Statist. Assoc.*, 64: 1303–1317, 1969. ISSN 0162-1459.

J. A. Hartigan. Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values. *Ann. Statist.*, 3:573–580, 1975. ISSN 0090-5364.

A Hedayat, WD Wallis, et al. Hadamard matrices and their applications. *The Annals of Statistics*, 6(6):1184–1238, 1978.

M. Hutchison, T. Womelsdorf, J. Gati, S. Everling, and R. Menon. Resting-state networks show dynamic functional connectivity in awake humans and anesthetized macaques. *Human Brain Mapping*, 34:2154–2177, 2013.

Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic acids research*, 42(D1):D199–D205, 2014.

Shota Katayama, Yutaka Kano, and Muni S. Srivastava. Asymptotic distributions of some test criteria for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.*, 116:410–421, 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2013.01.008.

Amir E. Khandani and Andrew W. Lo. What happened to the quants in august 2007? evidence from factors and transactions data. *Journal of Financial Markets*, 14(1):1–46, 2011.

Piotr Kokoszka and Remigijus Leipus. Change-point estimation in arch models. *Bernoulli*, 6(3):513–539, 2000.

Mladen Kolar and Eric Xing. On time varying undirected graphs. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS) 2011*, (JMLR) 15:407–415, 2011.

Mladen Kolar and Eric Xing. Sparsistent estimation of time-varying markov sparsistent estimation of time-varying markov random fields. *Preprint*, 2014.

Mladen Kolar, Le Song, and Eric Xing. Estimating time-varying networks. *Annals of Applied Statistics*, 4(1):94–123, 2010.

Michael R. Kosorok and Shuangge Ma. Marginal asymptotics for the "large p, small n" paradigm: with applications to microarray data. *Ann. Statist.*, 35(4):1456–1486, 2007. ISSN 0090-5364. doi: 10.1214/009053606000001433. URL `http://dx.doi.org/10.1214/009053606000001433`.

S.L. Lauritzen. *Graphical Models*. Clarendon Press, 1996. ISBN 9780191591228. URL `http://books.google.com/books?id=mGQWkx4guhAC`.

Sophie Lèbre, Jennifer Becq, Frèdèric Devaux, Michael PH Stumpf, and Gaëlle Lelandais. Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology*, 4(130):1–16, 2010.

Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (stars) for high-dim graphical models. *Neural Information Processing Systems (NIPS)*, 23, 2010.

Weidong Liu and Qi-Man Shao. A Carmér moderate deviation theorem for Hotelling's T^2-statistic with applications to global tests. *Ann. Statist.*, 41(1):296–322, 2013. ISSN 0090-5364. doi: 10.1214/12-AOS1082. URL `http://dx.doi.org/10.1214/12-AOS1082`.

Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

Po-Ling Loh, Martin J Wainwright, et al. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics*, 41(6):3022–3049, 2013.

Junwei Lu, Mladen Kolar, and Han Liu. Post-regularization inference for dynamic nonparanormal graphical models. *arXiv preprint arXiv:1512.08298*, 2015.

PC Mahalanobis. Sample surveys of crop yields in india. *Sankhyā: The Indian Journal of Statistics*, pages 269–280, 1946.

V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72 (114):507–536, 1967.

Philip J McCarthy. Pseudo-replication: Half samples. *Revue de l'Institut International de Statistique*, pages 239–264, 1969.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006. ISSN 0090-5364. doi: 10.1214/009053606000000281. URL `http://dx.doi.org/10.1214/009053606000000281`.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial Correlation Estimation by Joint Sparse Regression Models. *Journal of American Statistical Association*, 104(486):735–746, 2009.

D.N. Politis, J.P. Romano, and M. Wolf. *Subsampling*. Springer Series in Statistics. Springer New York, 1999. ISBN 9780387988542. URL `https://books.google.com/books?id=nGu6rqjE6JoC`.

Stephen Portnoy. On the central limit theorem in $\mathbf{R}^p$ when $p \to \infty$. *Probab. Theory Related Fields*, 73(4):571–583, 1986. ISSN 0178-8051. doi: 10.1007/BF00324853. URL `http://dx.doi.org/10.1007/BF00324853`.

Teresa M. Przytycka, Mona Singh, and Donna K. Slonim. Toward the dynamic interactome: it's toward the dynamic interactome: it's about time. *Briefings in Bioinformatics*, 11(1): 15–29, 2010.

Huitong Qiu, Fang Han, Han Liu, and Brian Caffo. Robust portfolio optimization under high dimensional heavy-tailed time series. Technical report, Technical report, Johns Hopkins University, 2014.

Huitong Qiu, Fang Han, Han Liu, and Brian Caffo. Joint estimation of multiple graphical models from high dimensional time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2015.

Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 2008.

Haskell P. Rosenthal. On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.*, 8:273–303, 1970. ISSN 0021-2172.

Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse Permutation Invariant Covariance Estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Sandipan Roy, Yves Atchadè, and George Michailidis. Change-point estimation in high-dimensional markov random field models. *preprint*, 2014.

A. I. Sakhanenko. Estimates in the invariance principle in terms of truncated power moments. *Sibirsk. Mat. Zh.*, 47(6):1355–1371, 2006. ISSN 0037-4474. doi: 10.1007/s11202-006-0119-1. URL `http://dx.doi.org/10.1007/s11202-006-0119-1`.

Jun Shao and et al. A general theory for jackknife variance estimation. *Ann. Statist.*, 17(3):1176–1197, 1989. ISSN 0090-5364. doi: 10.1214/aos/1176347263. URL `http://dx.doi.org/10.1214/aos/1176347263`.

Jun Shao and Dongsheng Tu. *The jackknife and bootstrap*. Springer Science & Business Media, 2012.

Galen R. Shorack and Jon A. Wellner. *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986. ISBN 0-471-86725-X.

Hai Shu and Bin Nan. Estimation of large covariance and precision matrices from temporally dependent observations. *arXiv preprint arXiv:1412.5059*, 2014.

Muni S. Srivastava. A test for the mean vector with fewer observations than the dimension under non-normality. *J. Multivariate Anal.*, 100(3):518–532, 2009. ISSN 0047-259X. doi: 10.1016/j.jmva.2008.06.006. URL `http://dx.doi.org/10.1016/j.jmva.2008.06.006`.

Muni S. Srivastava, Shota Katayama, and Yutaka Kano. A two sample test in high dimensional data. *J. Multivariate Anal.*, 114:349–358, 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2012.08.014. URL `http://dx.doi.org/10.1016/j.jmva.2012.08.014`.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67 (1):91–108, 2005.

Howell Tong. *Non-Linear Time Series: A Dynamical System Approach.* Oxford Statistical Science Series, 1993.

Mark S. Veillette and Murad S. Taqqu. Properties and numerical evaluation of the Rosenblatt distribution. *Bernoulli*, 19(3):982–1005, 2013. ISSN 1350-7265. doi: 10.3150/12-BEJ421. URL `http://dx.doi.org/10.3150/12-BEJ421`.

Kenneth W. Wachter. The strong limits of random matrix spectra for sample matrices of independent elements. *Ann. Probability*, 6(1):1–18, 1978.

Ami Wiesel, Ofir Bibi, and Amir Globerson. Time varying autoregressive moving average models for covariance estimation. *Signal Processing, IEEE Transactions on*, 61(11):2791–2801, 2013.

C.-F. J. Wu. On the asymptotic properties of the jackknife histogram. *Ann. Statist.*, 18(3): 1438–1452, 1990. ISSN 0090-5364. doi: 10.1214/aos/1176347759. URL `http://dx.doi.org/10.1214/aos/1176347759`.

Wei Biao Wu and Mohsen Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003. ISSN 0006-3444. doi: 10.1093/biomet/90.4.831. URL `http://dx.doi.org/10.1093/biomet/90.4.831`.

Wei-Biao Wu and Ying Nian Wu. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Statist.*, 10(1):352–379, 2016. doi: 10.1214/16-EJS1108. URL `http://dx.doi.org/10.1214/16-EJS1108`.

Rao K Yarlagadda and John E Hershey. *Hadamard matrix analysis and synthesis: with* *applications to communications and signal/image processing*, volume 383. Springer Science & Business Media, 2012.

Ming Yuan. High Dimensional Inverse Covariance Matrix Estimation via Linear Programming. *Journal of Machine Learning Research*, 11:2261–2286, 2010.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Geng Zhang, Aaron Schetter, Peijun He, Naotake Funamizu, Jochen Gaedcke, B Michael Ghadimi, Thomas Ried, Raffit Hassan, Harris G Yfantis, Dong H Lee, et al. Dpep1 inhibits tumor cell invasiveness, enhances chemosensitivity and predicts clinical outcome in pancreatic ductal adenocarcinoma. *PloS one*, 7(2):e31507, 2012.

Geng Zhang, Peijun He, Hanson Tan, Anuradha Budhu, Jochen Gaedcke, B Michael Ghadimi, Thomas Ried, Harris G Yfantis, Dong H Lee, Anirban Maitra, et al. Integration of metabolomics and transcriptomics revealed a fatty acid network exerting growth inhibitory effects in human pancreatic cancer. *Clinical cancer research*, 19(18):4983–4993, 2013.

Shuheng Zhou. Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, 42(2):532–562, 2014.

Shuheng Zhou, John Lafferty, and Larry Wasserman. Time Varying Undirected Graphs. *Machine Learning*, 80:295–319, 2010.

Zhou Zhou and Wei Biao Wu. Local linear quantile estimation for nonstationary time series. *The Annals of Statistics*, pages 2696–2729, 2009.

Zhou Zhou and Wei Biao Wu. Simultaneous inference of linear models with time varying coefficients. *Journal of the Royal Statistical Society*, 72(4):513–531, 2010.