

# LARGE SCALE MULTIPLE TESTING FOR HIGH-DIMENSIONAL NONPARANORMAL DATA

---

A Dissertation  
Submitted to  
the Temple University Graduate Board

---

In Partial Fulfillment of the  
Requirements for the Degree of  
Doctor of Philosophy

---

by  
Yanhui Xu  
Diploma Date May 2019

Examining Committee Members:

Sanat K. Sarkar, Advisory Chair, Statistics

Xu Han, Co-Advisor, Statistics

Kuang-Yao Lee, Statistics

Zhigen Zhao, Statistics

Li He, External Reader, Merck Research Laboratories

## ABSTRACT

False discovery control in high dimensional multiple testing has been frequently encountered in many scientific research. Under the multivariate normal distribution assumption, Fan et al. (2012) proposed an approximate expression for false discovery proportion (FDP) in large-scale multiple testing when a common threshold is used and provided a consistent estimate of realized FDP when the covariance matrix is known. They further extended their study when the covariance matrix is unknown (Fan & Han 2017). However, in reality, the multivariate normal assumption is often violated. In this paper, we relaxed the normal assumption by developing a testing procedure on nonparanormal distribution which extends the Gaussian family to a much larger population. The nonparanormal distribution is indeed a high dimensional Gaussian copula with nonparametric marginals. Estimating the underlying monotone functions is key to good FDP approximation. Our procedure achieved minimal mean error in approximating the FDP compared with other methods in simulation studies. We gave theoretical investigations regarding the performance of estimated covariance matrix and false rejections. In real dataset setting, our method was able to detect more differentiated genes while still maintaining the FDP under a small level. This thesis provides an important tool for approximating FDP in a given experiment where the normal assumption may not hold. We also developed a dependence-adjusted procedure which provides more power than fixed-threshold method. Our procedure also show robustness for heavy-tailed data under a variety of distributions in numeric studies.

## ACKNOWLEDGMENTS

All praise, honour and glory to my Lord Jesus Christ for His richest love and mercy for the accomplishment of this dissertation.

I would like to express my deepest gratitude towards all the people who have helped me through my journey of pursuing my Ph.D in Statistics.

First, I would like to thank my advisor, Dr. Sanat Sarkar, for his encouragement, support, guidance and insights for my research. Even during my early years in the Ph.D program, he has inspired and impressed me by his immense and extensive knowledge, meticulous attitude and genial personality. His guidance has continued through my Ph.D research. I am really thankful that he could be my advisor.

My sincere appreciation also goes to Dr. Xu Han for his detailed guidance, inspiring discussions and suggestions, for his patience and encouragement, which have helped me tremendously throughout all the stages of my research and thesis writing. I also would like to thank my dissertation committee members, Dr. Zhigen Zhao, Dr. Kuang-Yao Lee, and Dr. He for their helpful suggestions and editing remarks.

I am grateful to all the professors in the Department of Statistical Science, Temple University, and also to all the administrative staffs for their support.

I thank all my family members, especially my husband Matt Ford and my mom Yueqin Wang, for their endless support, encouragement and love throughout my study at Temple University!

## TABLE OF CONTENTS

	Page
ABSTRACT . . . . .	ii
ACKNOWLEDGMENTS . . . . .	iii
LIST OF TABLES . . . . .	vii
LIST OF FIGURES . . . . .	ix
 CHAPTER	
1 INTRODUCTION . . . . .	1
1.1 Motivation . . . . .	1
1.2 Background . . . . .	6
1.2.1 Nonparanormal . . . . .	6
1.2.2 False Discovery Proportion . . . . .	7
1.2.3 Estimating Marginal Transformation Functions . . . . .	9
1.3 Structure . . . . .	12
2 LITERATURE REVIEW . . . . .	15
2.1 Family Wise Error Rate . . . . .	15
2.2 False Discovery Rate . . . . .	16
2.2.1 The Benjamini and Hochberg Procedure . . . . .	17
2.2.2 The Benjamini and Liu Procedure . . . . .	17
2.2.3 The Benjamini and Yekutieli Procedure . . . . .	17
2.2.4 Storey's pFDR . . . . .	18
2.3 Adaptive Procedure . . . . .	19
2.3.1 Plug-in procedures . . . . .	19
2.3.2 Two-stage procedures . . . . .	20
2.4 Hierarchical Hypothesis Testing . . . . .	21

	Page
2.4.1	Benjamini and Bogomolov Procedure . . . . . 22
2.4.2	Hu's Procedure . . . . . 23
2.4.3	Benjamini and Heller Procedure . . . . . 23
2.5	Adjusting multiplicity under dependence . . . . . 24
2.5.1	p-value adjustment . . . . . 25
2.5.2	Factor based adjustment . . . . . 25
2.5.3	Robustness . . . . . 27
2.6	Nonparanormal Distribution . . . . . 28
2.6.1	Empirical Correlation Matrix . . . . . 29
2.6.2	Rank Correlations . . . . . 30
3	NPN-PFA PROCEDURE . . . . . 33
3.1	Estimating Marginal Transformation Functions . . . . . 33
3.2	Proposed Method . . . . . 36
3.3	Theoretical Investigations . . . . . 37
3.3.1	Correlation Matrix . . . . . 37
3.3.2	Number of False Rejections . . . . . 38
3.4	Dependence Adjusted Procedure . . . . . 39
4	SIMULATION AND REAL DATA ANALYSIS . . . . . 41
4.1	Simulation Result . . . . . 41
4.1.1	Known fixed number of factors . . . . . 42
4.1.2	Unknown number of factors . . . . . 46
4.1.3	Band-width selection . . . . . 50
4.1.4	Robustness . . . . . 52
4.1.5	Other Transformation Function . . . . . 55
4.1.6	Dependence Adjusted Testing procedure . . . . . 57
4.2	Real Data Analysis . . . . . 59
4.2.1	Prostate Cancer Data . . . . . 59
4.2.2	Breast Cancer Data . . . . . 62

	Page
5 CONCLUSION AND FUTURE DIRECTION . . . . .	65
5.1 Summary and Conclusion . . . . .	65
5.2 Ongoing and Future Work . . . . .	66
5.2.1 Three-way data hypothesis testing . . . . .	66
BIBLIOGRAPHY . . . . .	71
A THEORETICAL PROOF . . . . .	79

## LIST OF TABLES

Table	Page
1.1 Testing normality in the gene expression of prostate cancer data. . . . .	3
1.2 Hypothesis testing results classification. . . . .	8
1.3 Mean(SD) of $\widehat{FDP}(t) - FDP(t)$ , expressed in percentage, using <i>Winsorized</i> <sup>2</sup> estimator and rank-based correlation estimators. . . . .	12
4.1 Mean(SD) of $\widehat{FDP}(t) - FDP(t)$ , expressed in percentage, comparing <i>Winsorized</i> <sup>2</sup> vs NPN-PFA. Correlation matrix $\widehat{\Gamma}$ is estimated using mSCM. . . . .	44
4.2 Mean(SD) of $\widehat{FDP}(t) - FDP(t)$ , expressed in percentage, comparing <i>Winsorized</i> <sup>2</sup> vs NPN-PFA. Correlation matrix $\widehat{\Gamma}$ is estimated using POET. . . . .	45
4.3 Mean(S.D.) of $\widehat{FDP}(t) - FDP(t)$ , expressed in percentage, comparing different values of $c$ (using normal kernel and mSCM). . . . .	45
4.4 Mean(SD) of $\widehat{FDP}(t) - FDP(t)$ , expressed in percentage, comparing <i>Winsorized</i> <sup>2</sup> vs NPN-PFA with unknown number of factors. Correlation matrix $\widehat{\Gamma}$ is estimated using mSCM. . . . .	48
4.5 Mean (S.D.) of $\widehat{FDP}(t) - FDP(t)$ , expressed in percentage, comparing different bandwidth selection methods . . . . .	51
4.6 Mean (S.D.) of $\widehat{FDP}(t) - FDP(t)$ , expressed in percentage, using PFA directly vs Nonparanormal PFA. . . . .	52
4.7 Mean (S.D.) of $\widehat{FDP}(t) - FDP(t)$ , expressed in percentage, using PFA directly vs Nonparanormal PFA for $t$ -distribution with $\mu_1 = 0.5$ . . . . .	52
4.8 Mean (S.D.) of $\widehat{FDP}(t) - FDP(t)$ , expressed in percentage, using Box-Cox vs nonparanormal transformation . . . . .	56
4.9 Comparison of dependence-adjusted procedure with fixed threshold procedure under strict factor model and approximate factor model. The nonzero $\mu_i$ are simulated from $U(0.1, 0.5)$ and $p_1 = 200$ . . . . .	58
4.10 Top 10 genes selected from the prostate cancer data: comparing NPN-PFA vs Efron's method. . . . .	61
4.11 Testing normality of prostate cancer data after Nonparanormal transformation . . . . .	61

Table	Page
4.12 Testing sample independence in the prostate cancer data after Nonpara-normal transformation . . . . .	62
4.13 30 most differentially expressed genes that can discriminate breast cancers with BRCA1 mutations from those with BRCA2 mutations. . . . .	63
5.1 Mean (S.D.) of $\widehat{FDP}(t) - FDP(t)$ , expressed in percentage, using PFA in matrix normal data. . . . .	70

## LIST OF FIGURES

Figure	Page
1.1 Visualization of Nonparanormal data with density plot, contour plot and 3D plot (from left to right). . . . .	7
1.2 FDP estimation with known number of factors using <i>Winsorized</i> <sup>2</sup> Estimator and rank-based correlation estimation (n=50). Top panel: power transformation; Bottom panel: CDF transformation . . . . .	13
4.1 FDP estimation with known number of factors (n=50, p=1000, t=0.01). Top panel: CDF transformation; bottom panel: Power transformation. Correlation matrix is estimated using mSCM. . . . .	44
4.2 FDP estimation with known number of factors (n=50, p=1000, t=0.01). Top panel: CDF transformation; bottom panel: Power transformation. Correlation matrix is estimated using POET. . . . .	44
4.3 Comparison of $\widehat{FDP}(t)$ with realized FDP(t) for power transformation with unknown number of factors. From top to bottom panel corresponds to Model 1 to Model 3. n = 50. . . . .	49
4.4 Comparison of $\widehat{FDP}(t)$ with FDP(t) for CDF transformation. From top to bottom panel corresponds to Model 1 to Model 3. n = 50. . . . .	50
4.5 Comparing PFA alone vs NPN-PFA with $\mu_1 = 0.5$ . . . . .	53
4.6 Comparing PFA alone vs NPN-PFA with $\mu_1 = 0.5$ , with various degrees of freedom for <i>t</i> -distribution: from left to right: df=3, 4 (top); df=5, 6 (bottom). . . . .	54
4.7 Box-Cox vs nonparanormal transformation. Top left: CDF transformation; Top right: power transformation; Bottom left: gamma-distribution; Bottom right: log-normal distribution. . . . .	56
4.8 The estimated FDP and false discoveries as functions of the number of total discoveries for p=6033 genes, with the number of factors being 1, 3, 5 or 7. mSCM was used to estimate the correlation matrix. . . . .	60
4.9 The estimated false discovery proportion and the estimated number of false discoveries as functions of the number of total discoveries for $p = 3226$ genes, with the number of factors being 1, 2, 3, 4 or 5. Modified SCM was used to estimate the correlation matrix. . . . .	63

- 5.1 Comparison of  $\widehat{FDP}(t)$  with  $FDP(t)$  for matrix normal data. The order match the order of the table. Top Panel: Method 1; Bottom Panel: Method 2; First column:  $n=50$  with factors = (2,4); Second column:  $n=50$  with factors = (3,3); Third column:  $n=100$  with factors = (2,4); Forth column:  $n=100$  with factors = (3,3) . . . . . 70

# CHAPTER 1

## INTRODUCTION

### 1.1 Motivation

With modern technologies generating vast amounts of data at an unprecedented speed, statistical analyses of data arising in modern scientific investigations, such as those in genomics, astrophysics, brain imaging, and spatial epidemiology, involve simultaneous testing of thousands of hypotheses. Most of these applications produce high-dimensional data, in which the underlying variables are seen to be strongly dependent. Failing to account for such dependency in a multiple testing method can result in efficiency loss and introduction of bias in identifying true signals versus noise, see Efron (2007) and Schwartzman & Lin (2011). Strong correlation among the variables can also result in high variability among testing results, making the scientific findings hard to reproduce. Thus, taking into account of dependency among variables in the construction of a multiple testing method, especially when such dependency is very high, is an important undertaking; not doing it can negatively impact the results.

Controlling false discovery rate (FDR), the expected proportion of false discoveries, is a powerful approach to large-scale multiple testing compared to traditional methods of controlling the familywise error rate Benjamini & Hochberg (1995). However, the commonly used FDR controlling methods are applied directly to the marginal  $p$ -values that are often required to be independent. Efforts have been extended towards controlling FDR under dependence, but they are seen to be successful only in a qualitative sense, i.e., when the dependence structure meets certain conditions. For instance, Benjamini & Yekutieli (2001) showed that the Benjamini-Hochberg (BH) method controls the FDR if the  $p$ -values satisfy the condition of positive regression

dependency on the subset (PRDS) of null  $p$ -values; see also Sarkar (2002). Storey (2004) proposed a procedure that controls FDR asymptotically under weak dependence. Sun & Tony Cai (2009) assumed the hidden states of the hypotheses being tested follow a Markov model and considered controlling a measure closely related to the FDR based on a new test statistic called the local index of significance. Clarke & Hall (2009) assumed low dependence among the variables in showing that multiple testing procedures are robust when the null distributions are light-tailed, such as normal or Student's  $t$  distribution. Efron (2007, 2010a) took into account the variability among test statistics in formulating conditional false discovery rate by incorporating a dispersion factor in the empirical null distribution of the test statistics obtained by bootstrapping. Leek & Storey (2008), Friguet et al. (2009), and Desai & Storey (2012) considered weakening the dependence structure by assuming a strict factor model with independent idiosyncratic errors for the data and then subtracting the common factor.

Developing a FDR controlling method under dependency more effectively than simply making an assumption about its type by incorporating a measure of dependency directly into the method has been a challenging problem in multiple testing. Fan et al. (2012) overcome this challenge when the test statistics follow multivariate normal distribution with a known covariance matrix. More specifically, they considered thresholding an estimate of FDP. Controlling FDP by thresholding an estimate of it is often statistically more meaningful than controlling its expectation, the FDR, since the actual FDP can vary dramatically, especially under high dependency, even when the FDR can be controlled. Moreover, the underlying dependence can be captured more effectively by directly incorporating a quantitative measure of it into estimation of FDP. Fan et al. (2012) developed a procedure called Principal Factor Approximation (PFA) to approximate FDP in high-dimensional multiple testing. They showed theoretical consistency of PFA before proposing to use it for FDP control. Fan & Han (2017) extended this framework to one that includes situations with unknown covariance matrix, and evaluated the impact of estimating marginal

variances and eigenvalues/eigenvectors on the FDP approximation. Still, the assumption of multivariate normal distribution remains a restriction in both of these papers. It is often violated in reality, that high dimensional data that are often skewed and contaminated with outliers. For instance, we looked at the prostate cancer dataset analyzed in Efron (2007), which contains comparisons of 50 non-tumor subjects with 52 tumor patients for each of 6033 genes originally reported by Singh et al. (2002). Test statistics were calculated using two sample  $t$  test for each gene. Based on empirical null extrapolation of the test statistics, about 93% of the genes are seen to come from null population. We tested the marginal sample distribution of the 6033 genes as well as the test statistics of the prostate cancer data at the significance level of 0.05, with the results being shown in Table 1.1. It is clear that at most 8.2% of the 6033 genes would pass any of the four normality tests. Even with Holm’s correction, there are still more than 60% genes that fail to pass any normality test. So the normality assumption is clearly not valid in this case, potentially affecting hypothesis testing results.

Table 1.1. Testing normality in the gene expression of prostate cancer data.

Critical value (0.05)	Shapiro Wilk	Shapiro- Francia	Anderson- Darling	Pearson
Unadjusted	22(0.36%)	57(0.94%)	143(2.37%)	292(4.84%)
Adjusted with Holm’s correction	1552(25.7%)	2356(39.1%)	1833 (30.4%)	807(13.4%)

Data transformation methods, such as logarithmic, inverse, power, and the most famous Box-Cox transformation, have long been applied to achieve normality. However, these transformed data are hard to interpret when used in parametric modeling.

The purpose of this paper is to broaden the applicability of Fan & Han (2017) by relaxing the multivariate normal distribution assumption made therein. Relaxing multivariate normality using nonparanormal family of distributions has been found to be quite successful in modeling dependency in data from many fields. For instance,

Gaussian copula has been applied in (Ince et al. 2017) to estimate dependencies for two types of neuroimaging data, electroencephalogram (EEG) and magnetoencephalogram (MEG) data. In (Malevergne et al. 2003), the authors have studied three groups of financial assets, currencies, metals traded on the London Metal Exchange, and stocks chosen among the largest companies quoted on the New York Stocks Exchange, and investigated if the dependence between each pairs of assets within each group can be modeled by Gaussian copula. Use of nonparanormal family of distributions for modeling dependency has been seen, again in neuroscience ((Berkes et al. 2009)), and in hydrology and climate science (Renard & Lang (2007), AghaKouchak (2014)), genomics (Liu et al. (2012), Xue & Zou (2012)), and informatics ((Rey & Roth 2012)). Nonparanormal family of distributions was first introduced by Liu, Lafferty & L.Wasserman (2009) and further illustrated by Liu, Han & Yuan (2012) and Xue & Zou (2012). As a semi-parametric Gaussian copula (Liu, Lafferty & L.Wasserman (2009) and Xue & Zou (2012)), it relaxes many of the constraints of the multivariate normal family, and thus greatly enhances the scope of extending high-dimensional inferential methods to those that require much less restrictive assumptions than normality. We consider extending the work of Fan and Han Fan & Han (2017), including a thorough investigation of the approximation of FDP estimation, under this family of distributions with unknown correlations.

Estimating the marginal transformation functions, which is fundamental to modeling data using nonparanormal family of distributions, requires estimation of the cumulative distribution function (CDF) of the original data. Liu et al. (2009) proposed a truncated empirical distribution function, called *Winsorized*<sup>2</sup> estimator, which reduces the variance at tails under high-dimensional settings. However, a drawback of using this estimator in the present context is that the truncation of tail areas will affect the accuracy of multiple testings that are based on tails. So, instead of *Winsorized*<sup>2</sup> estimator, we will consider using kernel based estimator which has several desirable properties. Specifically, it has been shown to have a smaller mean integrated squared error than empirical CDF estimator. Moreover, Reiss (1981) proved that the relative

deficiency of the empirical CDF with respect to an appropriately chosen kernel CDF quickly deteriorates to infinity as the sample size increases under some conditions (Shirahata & Chu (1988) and Kendall (1990)). There has been extensive studies in the performance of kernel CDF estimation (see Parzen (1962), Azzalini (1981), Falk (1983) and more recently Cheng & Peng (2002)); however, it is worth noting that how such estimator affects multiple testing is not known yet.

Investigating the impact of unknown covariance matrix on the performance of our proposed multiple testing method under nonparanormal family of distributions is an important part of our research. There are estimators available in the literature for estimating covariance matrix. For instance, Fan et al. (2013) proposed a thresholding estimator under conditional sparsity assumption, called Principal Orthogonal complement Thresholding (POET) estimator, which achieves better convergence rate compared to other types of covariance matrix estimators. Also, there are rank based correlation estimators, such as Kendall's tau and Spearman's rho, which enjoy robustness Xue & Zou (2012) and can be modified to achieve consistency (Liu et al. 2012). However, we propose in this paper a modified sample correlation matrix estimator (mSCM). It not only simplifies arguments in the proofs of associated theoretical results, but also achieves the desired convergence. Moreover, extensive simulation studies comparing the performance of mSCM to those of POET and the rank-based estimator in terms of the error rate of FDP approximation indicate that while both mSCM and POET outperform the rank-based covariance matrix estimator and there is negligible difference between the performance of mSCM and POET. But mSCM is computationally much more convenient than POET.

The development of our proposed multiple testing procedure consists of the following steps: (i) Transformation of the original data from nonparanormal to multivariate normal using kernel quantile based estimators of the transformation functions; (ii) construction of test statistics using the transformed data; (iii) estimation of population correlation matrix using mSCM; and (iv) application of PFA to approximate FDP. Similar procedure using *Winsorized* estimator is also constructed to compare

its performance with ours. Simulation results indicate superior performance of our procedure over that based on *Winsorized* estimator in terms of FDP accuracy. We also provide theoretical support for our Kernel based approximation method.

## 1.2 Background

### 1.2.1 Nonparanormal

First, let us recall the definition of nonparanormal family of distributions.

**Definition 1.2.1** (*Nonparanormal*) A random vector  $\mathbf{X} = (X_1, \dots, X_p)^T$  is said to have a nonparanormal distribution; i.e.,  $\mathbf{X} \sim NPN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{g})$ , if there exist functions  $\{g_j\}_{j=1}^p$  such that  $\mathbf{Y} = \mathbf{g}(\mathbf{X}) \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$ , where  $\mathbf{g}(\mathbf{X}) = (g_1(X_1), \dots, g_p(X_p))$ . When the  $g_j$ s are monotone and differentiable, the joint probability density function of  $\mathbf{X}$  is given by

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \cdot \exp \left\{ -\frac{1}{2} (\mathbf{g}(\mathbf{x}) - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{g}(\mathbf{x}) - \boldsymbol{\mu}) \right\} \cdot \prod_{j=1}^p |g'_j(x_j)|, \quad (1.1)$$

where the product term is the Jacobian of transformation.

The density in (1.1) is generally not identifiable. To make it identifiable,  $g_j$  is chosen so that the marginal means and variances remains the same as the original data:

$$\mathbb{E}[g_j(X_j)] = \mathbb{E}(X_j) \text{ and } \text{Var}[g_j(X_j)] = \text{Var}(X_j).$$

These conditions depend only on  $\boldsymbol{\mu}$  and  $\text{diag}(\boldsymbol{\Sigma})$  (note that  $\text{diag}(\boldsymbol{\Sigma}) = \text{diag}(\boldsymbol{\Sigma}_0)$ ), but not the full covariance matrix  $\boldsymbol{\Sigma}$ .

To visualize the nonparanormal data, we plotted two nonparanomal data generated by power or CDF transformation in Figure 1.1, see Chapter 4 for details. The nonparanormal data covers a much wider variety of distributions.

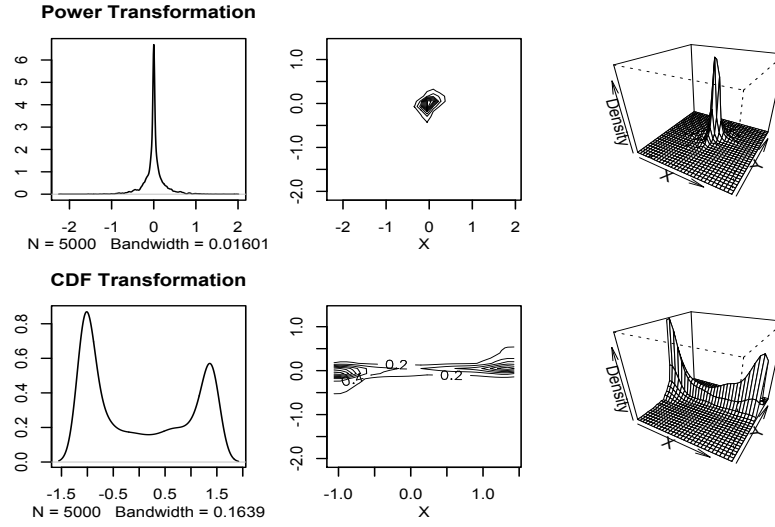


Figure 1.1. Visualization of Nonparanormal data with density plot, contour plot and 3D plot (from left to right).

### 1.2.2 False Discovery Proportion

Suppose that we have a random sample  $\mathbf{X}_1, \dots, \mathbf{X}_n$  from  $NPN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{g})$ , with  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)$ ,  $\boldsymbol{\Sigma}$  and  $\mathbf{g}$  being unknown, for the problem of testing  $H_{0j} : \mu_j = 0$  against  $H_j : \mu_j \neq 0$ , simultaneously for  $j = 1, \dots, p$ , subject to a control of FDR by thresholding an estimate of FDP at a suitably chosen constant.

Let  $p_j$  denote the  $p$ -value associated with testing the  $j$ th hypothesis,  $t$  be the thresholding constant, and  $\Theta$  be the null set. Then,

$$R(t) = \sum_{j=1}^p \mathbf{I}\{p_j \leq t\} \text{ and } V(t) = \sum_{j \in \Theta} \mathbf{I}\{p_j \leq t\}$$

are the number of discoveries and the number of false discoveries, respectively. Correspondingly, the FDP at  $t$  is defined as:

$$\text{FDP}(t) = V(t)/R(t),$$

with  $0/0$  being conventionally assumed to be zero in the paper. Note that  $R(t)$  is observable, but  $V(t)$  and FDP are both random variables and not observable. Given an experiment (observed sample data),  $FDP(t)$  is a realized but unknown quantity. Better approximation of  $FDP(t)$  will provide better control of the false discoveries in the experiment. The further classifications of the multiple hypothesis testing can be found in Table 1.2.

Table 1.2. Hypothesis testing results classification.

Hypothesis	Accept	Reject	Total
True Null	$U$	$V$	$p_0$
False Null	$T$	$S$	$p_1$
	$W$	$R$	$p$

Let us suppose for the time being that  $Diag(\boldsymbol{\Sigma})$  and the monotone transformation  $\mathbf{g}$  be known. We can simply obtain the multivariate normal vector as  $\mathbf{Y} = \mathbf{g}(\mathbf{X}) \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$ , keeping the mean vector and the diagonals of  $\boldsymbol{\Sigma}$  unchanged, and apply the PFA procedure in Fan, Han & Gu (2012) to estimate the FDP based on  $\mathbf{Y}_i = \mathbf{g}(\mathbf{X}_i)$ ,  $i = 1, \dots, n$ .

The basic idea of PFA is to apply spectral decomposition of  $\boldsymbol{\Sigma}_0$  and use principal factors to account for the dependence, so that the remaining dependence is weak. Let us consider the vector of test statistics  $\mathbf{Z} = \frac{1}{\sqrt{n}} \sum_{i=1}^n Diag(\boldsymbol{\Sigma}^{-1/2}) \mathbf{Y}_i$ , which is distributed as  $N_p(\boldsymbol{\mu}^*, \boldsymbol{\Gamma})$ , where  $\boldsymbol{\mu}^* = (\mu_1^*, \dots, \mu_p^*)^T$ , with  $\mu_j^* = \sqrt{n} \mu_j / \sigma_j$ ,  $j = 1, \dots, p$ ,  $\sigma_j$  is the marginal standard deviation, and  $\boldsymbol{\Gamma}$  is correlation matrix. The underlying problem is that of testing

$$H_{0j} : \mu_j^* = 0 \text{ vs } H_{1j} : \mu_j^* \neq 0 \text{ for } j = 1, \dots, p. \quad (1.2)$$

The vector  $\mathbf{Z}$  can be decomposed as

$$\mathbf{Z} = \boldsymbol{\mu}^* + \mathbf{B}\mathbf{W} + \mathbf{K} \quad (1.3)$$

where  $\mathbf{W} \sim N_k(0, \mathbf{I}_k)$  is the vector of  $k$  common factors, and  $\mathbf{K} \sim N(0, \mathbf{A})$  is the vector of idiosyncratic errors distributed independently of  $\mathbf{W}$ . The columns of the matrix  $\mathbf{B} = (\sqrt{\lambda_1}\gamma_1, \dots, \sqrt{\lambda_k}\gamma_k)$  are the non-normalized  $k$  principal components and  $\mathbf{A} = \sum_{j=k+1}^p \lambda_j \gamma_j \gamma_j^T$ , where  $\{\lambda_j\}_{j=1}^p$  are the eigenvalues and  $\{\gamma_j\}_{j=1}^p$  are the corresponding eigen vectors of  $\mathbf{\Gamma}$ . The oracle FDP( $t$ ) is defined as

$$\text{FDP}_{\text{oracle}}(t) = \sum_{j \in \Theta} [\Phi(a_j(z_{t/2} + \eta_j)) + \Phi(a_j(z_{t/2} - \eta_j))]/R(t),$$

where  $z_{t/2}$  is the  $(1 - t/2)$ th quantile of the standard normal with the cdf  $\Phi(\cdot)$ ,  $a_j = (1 - \|\mathbf{b}_j\|^2)^{-1/2}$ , and  $\eta_j = \mathbf{b}_j^T \mathbf{W}$  with  $\mathbf{b}_j^T$  being the  $j$ th row of  $\mathbf{B}$ . Fan et al. (2012) have proved that under sparsity assumption,  $\text{FDP}_{\text{oracle}}(t)$  can be approximated by

$$\text{FDP}_A(t) = \sum_{j=1}^p [\Phi(a_j(z_{t/2} + \eta_j)) + \Phi(a_j(z_{t/2} - \eta_j))]/R(t). \quad (1.4)$$

With known  $\mathbf{\Gamma}$  and number of factors, the only parameter we need to estimate from the data is  $\mathbf{W}$ . Thus  $\text{FDP}_A(t)$  can be estimated by

$$\widehat{\text{FDP}}_A(t) = \sum_{j=1}^p [\Phi(a_j(z_{t/2} + \hat{\eta}_j)) + \Phi(a_j(z_{t/2} - \hat{\eta}_j))]/R(t) \quad (1.5)$$

where  $\hat{\eta}_j = \mathbf{b}_j^T \widehat{\mathbf{W}}$  for some estimator  $\widehat{\mathbf{W}}$  of  $\mathbf{W}$ . Under mild conditions, Fan, Han & Gu (2012) showed that

$$\left| \widehat{\text{FDP}}_A(t) - \text{FDP}_A(t) \right| = O\left(\left\| \widehat{\mathbf{W}} - \mathbf{W} \right\|\right), \quad (1.6)$$

where  $\widehat{\mathbf{W}}$  can be estimated by minimizing the absolute deviation loss:  $\sum_{j=1}^p |Z_j - \mathbf{b}_j^T \widehat{\mathbf{W}}|$ .

### 1.2.3 Estimating Marginal Transformation Functions

Now, when  $\Sigma$  and the marginal transformation function  $\mathbf{g}$  are unknown under the assumed nonparanormal setting, the  $\mathbf{Y}_i$ 's cannot be directly observed and need to

be estimated from the  $\mathbf{X}_i$ 's. Correspondingly, all the parameters in the numerator of (1.4) have to be estimated properly.

Let  $F_j(x)$  denote the marginal CDF of  $X_j$ , the  $j$ th component of  $\mathbf{X}$ , and  $g_j$  be assumed to be increasing without loss of generality. Since  $g_j(X_j)$  is normally distributed, we have

$$F_j(x) = \Phi\left(\frac{g_j(x) - \mu_j}{\sigma_j}\right),$$

yielding

$$Y_j \equiv g_j(X_j) \equiv \mu_j + \sigma_j \Phi^{-1}(F_j(X_j)). \quad (1.7)$$

Thus  $Y_{ij}$  can be estimated by

$$\widehat{Y}_{ij} = \widehat{\mu}_j + \widehat{\sigma}_j \Phi^{-1}(\widehat{F}_j(X_{ij})), \quad i = 1, \dots, n, \quad j = 1, \dots, p, \quad (1.8)$$

where  $\widehat{F}_j$  is some estimate of the marginal cumulative distribution function  $F_j$  of the  $j$ th coordinate  $X_j$ , and  $\widehat{\sigma}_j$  is the sample standard deviation which can be calculated from the data directly.

For  $\widehat{F}_j(X_{ij})$ , Liu et al. (2012) proposed using *Winsorized*<sup>2</sup> estimator for graphical models. More specifically, let  $\widehat{F}'_j(x)$  be the empirical CDF of the  $X_j$ , which is

$$\widehat{F}'_j(x) \equiv \frac{1}{n} \sum_{i=1}^n I\{X_{ji} \leq x\}.$$

Then the *Winsorized*<sup>2</sup> estimator is defined as

$$\widehat{F}_j(x) = \begin{cases} \delta_n, & \text{if } \widehat{F}'_j(x) < \delta_n, \\ \widehat{F}'_j(x), & \text{if } \delta_n \leq \widehat{F}'_j(x) \leq 1 - \delta_n, \\ (1 - \delta_n), & \text{if } \widehat{F}'_j(x) > 1 - \delta_n, \end{cases} \quad (1.9)$$

where  $\delta_n$  is a truncation parameter set to be

$$\delta_n \equiv \frac{1}{4n^{1/4}\sqrt{\pi \log n}}.$$

We will show later in this section that the *Winsorized*<sup>2</sup> estimator above can lead to unsatisfactory FDP approximation. Correspondingly, we will propose a kernel based estimator which will be described with more details later.

From  $\hat{Y}_{ij}$ , the estimated test statistics  $\hat{\mathbf{Z}}$  can be obtained. The  $p$  value of the  $j^{\text{th}}$  hypothesis testing is  $\hat{p}_j = 2\Phi(-|\hat{Z}_j|)$ . Correspondingly we have

$$\hat{R}(t) = \sum_{j=1}^p \mathbf{I}\{\hat{p}_j \leq t\} \text{ the estimated total number of rejections;}$$

$$\hat{V}(t) = \sum_{j \in \Theta} \mathbf{I}\{\hat{p}_j \leq t\} \text{ the estimated number of false rejections;}$$

$$\widehat{\text{FDP}}(t) = \hat{V}(t)/\hat{R}(t).$$

Next, we will illustrate that the *Winsorized*<sup>2</sup> estimator can produce some bias in the FDP approximation. Let us first consider a toy example. Suppose  $X_j \sim N(1.5, 1)$ . Let  $n = 100$ , then the threshold  $\delta_n \approx 0.02$ . Suppose we have two sample points:  $X_{j1} = 4$  and  $X_{j2} = -0.6$ . Note that  $P(X_j \leq 4) \approx 0.9938$  and  $P(X_j \leq -0.6) \approx 0.0179$ . Therefore, it is fair to have  $\Phi^{-1}(\hat{F}_j(X_{j1})) + \Phi^{-1}(\hat{F}_j(X_{j2})) = \Phi^{-1}(1 - \delta_n) + \Phi^{-1}(\delta_n) = 0$  under the *Winsorized*<sup>2</sup> estimator. On the other hand,  $\Phi^{-1}(F_j(X_{j1})) + \Phi^{-1}(F_j(X_{j2})) \approx \Phi^{-1}(0.9938) + \Phi^{-1}(0.0179) = 0.4016$ . Now, it is clear that the *Winsorized*<sup>2</sup> estimator can produce some bias in the significance detection.

Suppose we generate some nonparanormal data according to the model in Chapter 4 based on the CDF transformation (Definition 4.1.1) or Power transformation (Definition 4.1.2). Let  $p = 1000$  and  $n = 50, 100$ . See further details of parameters in Chapter 4. We apply *Winsorized*<sup>2</sup> estimator for the estimate of the marginal cumulative distribution function of the nonparanormal data. To obtain a good FDP

approximation, we further need a good estimate of the correlation matrix  $\mathbf{\Gamma}$  Fan & Han (2017). Since the rank-based estimators: Kendall’s tau and Spearman’s rho, have been popularly favored for the nonparanormal data, we apply them here combined with the PFA method in Fan & Han (2017). As shown in Figure 1.2 and Table 1.3, the *Winsorized*<sup>2</sup> estimator based FDP approximation has greatly underestimated the true values of FDP in both simulation settings. Winsorization was applied in Liu et al. (2009) to avoid infinity value and to achieve better bias-variance trade-off. Unfortunately this will impact the tail probability and the estimated FDP. To avoid this drawback and preserve the structure and distribution of the tails, we explored the use of kernel distribution estimator for the transformation function estimation.

Table 1.3. Mean(SD) of  $\widehat{\text{FDP}}(t) - \text{FDP}(t)$ , expressed in percentage, using *Winsorized*<sup>2</sup> estimator and rank-based correlation estimators.

		Kendall’s tau	Spearman’s rho
CDF transformation	n=50	-4.4(8.12)	-4.17(8.05)
	n=100	-5.07(8.7)	-4.94(8.66)
Power transformation	n=50	-3.69(6.58)	-3.42(6.45)
	n=100	-3.23(6.73)	-3.08(6.66)

### 1.3 Structure

In Chapter 2 of this thesis, we will go over some existing false discovery controlling procedures for testing multiple hypotheses. We will also discuss some recent advances in multiple testing under high dimensional settings.

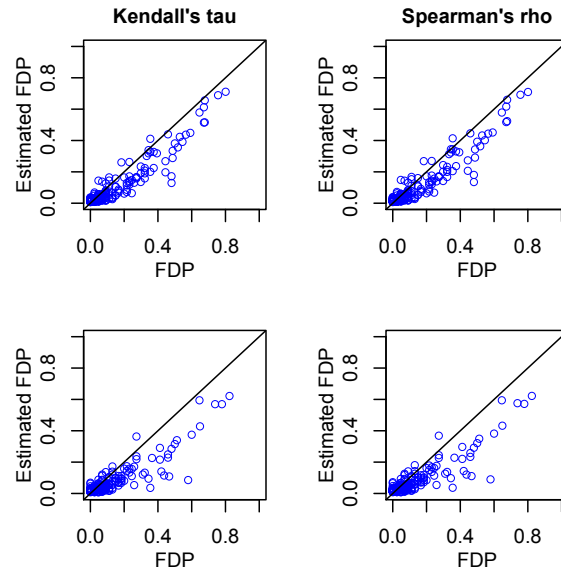


Figure 1.2. FDP estimation with known number of factors using *Winsorized<sup>2</sup>* Estimator and rank-based correlation estimation ( $n=50$ ). Top panel: power transformation; Bottom panel: CDF transformation

In Chapter 3, we will present the proposed methodology and investigate the theoretical properties of it in Chapter 3.

In Chapter 4, we will carry out extensive simulations and numerical studies to assess the performance of our method under known and unknown number of factors. We also investigate the power improvement under dependence adjusted procedure. We will also provide results in real data applications using the prostate cancer data and breast cancer data to discover the most differently differentiated genes between normal and prostate cancer tissues.

Finally, some summaries and future work are discussed in Chapter 5, and most of the proofs and technical details are given in Appendix.



## CHAPTER 2

### LITERATURE REVIEW

In this chapter, we will provide literature reviews for two topics: first is to review some commonly used multiple testing procedures in controlling family wise error rate (FWER), FDR, or estimating FDP; second is to provide an overview of the nonparametric extension of the Gaussian population through the use of Gaussian copula. Over the past two decades, the rapid advance in technology has made the data industry explode. Multiple testing has evolved from single step to hierarchical; from low dimension to high dimension; and from independent to highly correlated test statistics. The challenge is to adapt the current procedures and derive new ones to meet the demand of large scale inferences.

#### 2.1 Family Wise Error Rate

FWER is defined as the probability of making at least one false rejection when all null hypotheses are true. FWER control is the initial focus of multiplicity adjustment and it still frequently used today since it's easy to apply and interpret. The trade-off, however, is that when the number of hypotheses is large, they can be overly conservative, resulting in low-power tests. Many of these procedures involve ordering the raw  $p$ -values to compare with ordered critical values. Most FWER controlling methods are derived from the Bonferroni inequality or the Simes inequality.

Among many of the FWER controlling methods, Bonferroni procedure is the original and most widely used due to its simplicity, that is, a single step adjustment of all raw  $p$ -values. Considering testing  $m$  number of hypotheses  $H_{0(1)}, \dots, H_{0(m)}$ , let  $p_1, \dots, p_m$  be the corresponding  $p$ -values. Test all hypothesis with the critical value:  $c = \alpha/m$ . That is, hypothesis  $H_{0(i)}$  is rejected if  $p_i < \alpha/m$ . Bonferroni proce-

dures controls the FWER strongly at  $\pi_0\alpha$  under any arbitrary dependence structure, however it is too conservative with large  $m$ .

Sidak (1967) improved the power of Bonferroni method by using a single step critical value  $c = 1 - (1 - \alpha)^{1/m}$ . Sidak's procedure requires independence or positive dependence of the test statistics.

The Holm procedure (1979) is considered as a step-down version of the Bonferroni method. It is uniformly more powerful than Bonferroni procedure. The critical values are  $c_i = \alpha/(m - i + 1)$  for  $i = 1, 2, \dots, m$ . Compare the order p-values (ascending):  $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$  to the ordered critical values, if  $p_{(i)} \geq \alpha/(m - i + 1)$ , then accept  $H_{0(i)}, H_{0(i+1)}, \dots, H_{0(m)}$ ; otherwise, reject  $H_{0(i)}$  and increment to  $i + 1$ , while  $i < m$ . Holland and Copenhaver (1987) made some modifications to improve the power by using  $c_i = 1 - (1 - \alpha)^{1/(m-i+1)}$ ,  $i = 1, 2, \dots, m$  as critical values instead.

The Hochbergs Procedure (1988) is a step-up version of the Bonferroni method. By making some assumptions of the joint distribution of the test statistics (independent or positively dependent), it provides more power using the same critic value. The procedure rejects all  $p_{(i)} \leq c_k$ , where  $k = \max\{i : p_{(i)} < \alpha/(m-i+1)\}$ .

## 2.2 False Discovery Rate

FDR is an alternative way to maintain some principle bound on error while preserving the power of the tests. FDR is defined as the expected percentage or proportion of rejected hypotheses that have been wrongly rejected and has the following expression (Benjamini & Hochberg 1995) :  $\text{FDR} = \text{E}\left(\frac{V}{R \vee 1}\right) = \text{E}\left(\frac{V}{R} | R > 0\right) \text{Pr}(R > 0)$ , where  $R \vee 1 = \max(R, 1)$ . FDR is a better way to adjust for multiplicity especially when the number of hypothesis being tested is large. Instead of controlling the probability of a Type I error at a set level for each test, these methods control the overall FDR at level  $\alpha$ . When all null hypotheses are actually true, regardless of the number of rejections, the FDR is equivalent to the FWER. However, if the number of true nulls is smaller than the total number of hypothesis, the FDR is smaller than or

equal to the FWER (Benjamini & Hochberg 1995). Thus, controlling the FDR is less stringent than controlling the FWER, and consequently FDR procedures are more powerful.

### 2.2.1 The Benjamini and Hochberg Procedure

FDR was first introduced by Benjamini & Hochberg (1995), who developed a simple step-up procedure performed on the ordered  $p$ -values of the tests. Considering testing  $m$  number of hypotheses, suppose the number of true nulls is defined as  $m_0$  and  $0 \leq m_0 \leq m$ . Order all  $p$ -values such that:  $p_{(1)} < p_{(2)} < \dots < p_{(m)}$ . Let  $k = \max\{i : p_{(i)} < i\alpha/m\}$  and reject all hypotheses whose  $p$ -values are less than or equal to  $p_{(k)}$ . This procedure controls the FDR at level  $m_0\alpha/m$ , under the assumption of independence of the null  $p$ -values, or positive regression dependence, see Benjamini & Yekutieli (2001) and Sarkar (2002).

### 2.2.2 The Benjamini and Liu Procedure

Benjamini & Liu (1999) derived an alternative step-down procedure for controlling FDR which is called the BL procedure. To control the FDR at level  $\alpha$ , BL method starts with the smallest  $p$ -value to compare with the critical value,  $\delta_i = 1 - [1 - \min(1, \frac{m\alpha}{m-i+1})]^{1/(m-i+1)}$ . Let  $k = \min\{i : p_{(i)} > \delta_i\}$  and reject all hypotheses whose  $p$ -values are less than  $p_{(k)}$ . Their procedure neither dominates nor is dominated by the BH step-up procedure.

### 2.2.3 The Benjamini and Yekutieli Procedure

The Benjamini & Yekutieli (2001) procedure, abbreviated as BY procedure, extended the BH method so that it controls the FDR under any arbitrary dependence assumptions. The BY method is conducted as follows: order  $p$ -values, reject all  $p_i \leq p_{(k)}$  where  $k = \max\{i, p_{(i)} < \frac{i\alpha}{C_m m}\}$ ,  $C_m = \sum_i^m 1/i \approx \ln(m)$  when  $m$  is large.

Although this method sounds promising for its applicability of any dependence structure, it is typically very conservative, sometimes even more so than the Bonferroni method. A few other procedures based on BY method have been proposed to improve the power. Sarkar (2008b) discovered an alternative step-wise method that controls FDR under arbitrary dependence structure with improved power using the critical value  $\delta_i = i(i+1)\alpha/(2m^2)$ . Sarkar (2008a) further advanced the power of his algorithm by creating a pair-wise FDR control procedure using the critical value of  $\delta_i = i(i+1)\alpha/m(m-1)$ , for  $i = 2, \dots, m$  under the assumption that  $p$ -values are pair-wisely positively dependent.

#### 2.2.4 Storey's pFDR

Storey (2002, 2003) introduced a positive false discovery rate (pFDR) concept which is defined as:  $p\text{FDR} = E(\frac{V}{R}|R > 0)$ . In an asymptotic setting, pFDR and the FDR are equivalent, and consequently any asymptotic results about the FDR can essentially be directly translated into results for pFDR. Storey introduced the Bayesian interpretation of this estimator and the threshold  $q$  value: The conditional probability that a discovery is a false discovery given that its test statistic is in the rejection region. The rate is estimated as a posterior probability of a mixture model.

Suppose  $m$  identical hypotheses are being tested with the  $p$ -values  $p_1, \dots, p_m$  and a rejection region of  $t$ . Let  $H_i = 0$  or  $1$  represent the result of true or false null hypothesis, denote  $F_0$  and  $F_1$  be the cumulative distribution of  $p$ -values from the null and alternative hypothesis respectively, where  $F_0 \sim \text{Uniform}(0, 1)$ . Assume that  $(p_i, H_i)$  are i.i.d random variables.  $p_i|H_i \sim (1-H_i) \cdot F_0 + H_i \cdot F_1$ , and  $H_i \sim \text{Bernoulli}(1-\pi_0)$  for  $i = 1, \dots, m$ . Then  $p\text{FDR}(t) = Pr(H = 0|p < t) = \pi_0 t / (\pi_0 t + (1-\pi_0)F_1(t))$ , where  $\pi_0$  is the implicit prior probability used in the above posterior probability.

## 2.3 Adaptive Procedure

Adaptive control of FDR was first introduced by Benjamini & Hochberg (2000). The idea of the adaptation is to integrate an estimation of the unknown proportion of nulls  $\pi_0$  in the threshold of the procedures discussed above and that the FDR is still rigorously controlled at  $\alpha$ .

For instance, an adaptive multiple testing procedure would use  $c_i/\pi_0$  instead of  $c_i$  when comparing to the  $p$  values, as a way of compensating for the smaller number of true nulls other than assuming all hypotheses are nulls. By having a larger rejection region we gain in power. This is particularly important when the proportion of false nulls is large.

### 2.3.1 Plug-in procedures

Many adaptive FDR controlling methods belong to this family, where some initial estimator of  $\pi_0$  is directly plugged in as a shrinkage corrector to the usual procedures. The plug-in estimator used by Hochberg & Benjamini (1990) and Benjamini & Hochberg (2000) was constructed based on a modification of the graphical method of Schweder & Spjøtvoll (1982). Suppose there are  $m_0$  null hypothesis and the test statistics corresponding to them are independent, the  $p$ -values under null have uniform distribution over  $[0, 1]$ . The  $m_0$   $p$ -values can be considered as a realization of an ordered sample from the uniform distribution. In any given multiple testing situation, without knowing the value of  $m_0$ , we can safely assume that larger  $p$ -values are most likely from the true nulls. Let  $N_p$  be the number of  $p$ -values greater than a certain value  $p$ , we have  $E(N_p) \approx m_0(1 - p)$  when  $p$  is not too small. The plot of  $1 - p$  versus  $N_p$  (the quantile plot of the  $p$ -values) should exhibit linear relationship, along a line of slope  $S = m_0$ .

Storey (2002) developed his adaptive FDR controlling method using this plug-in estimator:  $\widehat{\text{FDR}}_\lambda(t) = m\widehat{\pi}_0(\lambda)t/\{R \vee 1\}$ , where  $\widehat{\pi}_0(\lambda)$  is defined as:  $\widehat{\pi}_0(\lambda) = \#(P_i >$

$\lambda)/(1 - \lambda)m$ .  $\lambda \in [0, 1)$  and is chosen such that the balance between the bias and the variance of  $\widehat{\pi}_0(\lambda)$  is reached, see details in Storey (2002).

Storey (2004) proved that his method is a conservative estimator for FDR over all significance regions simultaneously, and its asymptotic consistency is also established. For fixed  $\lambda$ ,  $E(\widehat{\text{FDR}}_\lambda(t)) \geq \text{FDR}(t)$ .  $\widehat{\text{FDR}}_\lambda(t)$  Storey's method become equivalent to the BH method when  $\pi_0 = 1$  but is more powerful than BH method under general circumstances.

### 2.3.2 Two-stage procedures

Two-stage approaches are also developed upon the basis of the BH method. A first round of multiple hypothesis testing is performed using some fixed algorithm such as step-up procedure, then the number of rejected hypothesis are used to estimate the number of true nulls. The threshold is updated using the estimated  $m_0$  for another round of testing. Two-stage procedures are proved to be more powerful than plug-in method.

Benjamini et al. (2006), often called the BKY method, estimated the proportion of null hypothesis using a modified threshold  $\frac{\alpha}{1+\alpha}$  before applying the BH method. The adaptive BKY method is implemented by the following steps:

1. Apply the BH step-up procedure at level  $\alpha' = \alpha/(1 + \alpha)$ . Let  $r_1$  be the number of rejected hypotheses. If  $r_1 = 0$  do not reject any hypothesis and stop; if  $r_1 = m$  reject all  $m$  hypotheses and stop; otherwise continue.
2. Let  $m_0 = (m - r_1)/(1 - \alpha') = (m - r_1)/(1 + \alpha)$ .
3. Use the BH step-up procedure again with  $\alpha^* = \alpha'/m_0$ .

The BKY adaptive BH method controls the FDR at  $\alpha$  under the assumption of independent  $p$ -values, improves the power over original BH procedure, but is less powerful than the Storey (2004) method. However, simulation studies show that the

BKY outperforms the Storey’s method when  $p$ -values are generated from a multivariate normal distribution with common positive correlations while still maintaining the FDR under  $\alpha$ .

Additional two-step or multiple-step adaptive procedures aiming at improving the BH method have been proposed. Among them are efforts for estimating the proportion of true nulls more accurately, see Sarkar (2008*b*) and Gavrilov et al. (2009); incorporating the dependence structure into test statistics before applying the BH method, see Efron (2007), Efron (2010*a*), and Sun & Tony Cai (2009); or generalizing the notion of FDR to  $k$ -FDR by relaxing control over at most  $k - 1$  false rejections Sarkar (2007), Sarkar & Guo (2009), and Sarkar & Guo (2010).

## 2.4 Hierarchical Hypothesis Testing

A special case of hypothesis testing arises when data come from heterogeneous groups or unknown intrinsic clusters. Hypotheses are divided into families and testing is carried out in multiple stages. For example, genome wide association study (GWAS), examination of a genome-wide set of genetic variants in different individuals is carried out to see if any variant is associated with a trait.

The data matrix has the rows indexed by the SNPs and columns by the traits. The first step is to select the SNPs that contains significant association with any traits. Then test individual traits within a selected SNP. Brain imaging and astronomical data usually contain a spatial structure where hypotheses are correlated locally. In other words, signals often come in clusters. Partitioning these hypotheses according to their spatial structure and perform hypothesis testing in a hierarchical order will improve the power and reduce the number of hypotheses tested. Multiple testing approaches that control the within group error rates do not guarantee the error rate of overall. A line of research has been focusing on adapting the FDR measures to maintain the error rate for each testing steps as well as overall experiment, see Benjamini & Heller (2007), Sun & Tony Cai (2009), Hu et al. (2010), Clements et al. (2011),

Clements et al. (2014), Benjamini & Bogomolov (2014), Heller et al. (2017). Most of the group hypothesis testing start from the group/family level then proceed to the individual hypothesis testing for the group(s) tested positive for signals, see Benjamini & Heller (2007), Yekutieli (2008), Benjamini & Bogomolov (2014), Clements et al. (2011), Clements et al. (2014), and Heller et al. (2017); or it can start with testing individual hypothesis within each group before moving on to identifying the significant group(s), see Liu et al. (2016). Liu et al. (2016) constructed a Bayesian method by decomposing a posterior measure of false discoveries across all hypotheses into within- and between-group components, allowing a portion of the overall FDR level to be used to maintain control over within-group false discoveries. Sun & Tony Cai (2009) developed a multiple-testing procedure that exploits the dependence structure among hypotheses assuming that the data were generated from a two-stage hidden Markov model. Here we will review some of the work listed above.

#### 2.4.1 Benjamini and Bogomolov Procedure

Recent work by Benjamini & Bogomolov (2014) addressed the testing of multiple families of hypotheses within the framework of selective inference. Testing each family separately while attending to some error rate control within each tested family has an obvious advantage that the control is achieved on the average across families. However, once only some families are selected, and inference is made on only the selected families, this average error rate across families deteriorates. They showed that applying a Bonferroni procedure in each selected row may result in a highly inflated conditional FWER when the selection is based on within group  $p$ -values. The authors presented a general framework to retains the control over expected average error over the selected families. They also explored the error control following selection under dependence of the hypothesis within the selected group. However, their procedure can only guarantee error rate on the average under positive regression dependent on the subset but not under general dependency.

### 2.4.2 Hu's Procedure

Hu et al. (2010) demonstrate the benefit of considering group structure by presenting a  $p$ -value weighting procedure which utilizes the relative importance of each group while controlling the overall false discovery rate under weak conditions. Their procedure, called group BH method, is as follows:

Assume that the  $N$  hypotheses can be divided into  $K$  disjoint groups with group sizes  $n_g$ ,  $g = 1, \dots, K$ . Let  $I_g$  be the index set of the  $g$ -th group,

1. For each  $p$ -value in group  $g$ , calculate the weighted  $p$ -values  $P_{wg}^{gi} = \frac{\pi_{g0}}{\pi_{g1}} P_{gi}$ . If  $\pi_{g0} = 1$ , let  $P_{wg}^{gi} = \infty$ ; If  $\pi_{g0} = 1$  for all  $g$ , accept all the hypotheses and stop. Otherwise go to the next step;
2. Pool all the weighted  $p$ -values together and let  $P_{(1)}^w, \dots, P_{(N)}^w$  be the corresponding ordered  $p$ -values.
3. Compute  $k = \max\{i, P_i^w \leq \frac{i\alpha^w}{N}\}$ , where  $\alpha^w = \frac{\alpha}{1-\pi_0}$ . If such a  $k$  exists, reject the  $k$  hypotheses associated with  $P_{(1)}^w, \dots, P_{(k)}^w$ ; otherwise do not reject any of the hypotheses.

$\pi_{g0}$  can be estimated when the number of null hypothesis is unknown for each group.

Hu et al.'s procedure was shown to be more powerful than the classical Benjamini-Hochberg procedure in both theoretical and simulation studies. By estimating the proportion of true null hypotheses, the data-driven procedure controls the false discovery rate asymptotically.

### 2.4.3 Benjamini and Heller Procedure

Benjamini & Heller (2007) provided a testing procedure for data containing spatial signals at varying locations. Signals usually appear in clusters for these spatial dataset. Traditionally, signals were tested individually. However the findings were reported as a group of nearby locations. If a prior knowledge of these natural clusters

are known or can be estimated, testing on the clusters levels will not only increases the strength of the signal but also reduces the number of hypotheses tests conducted. Let  $C_1, \dots, C_m$  be the  $m$  partition of  $N$  testing hypothesis denoted as the set  $D$ . Let  $D_0$  be the true null set and  $D_1$  be the false null set, both are unknown and need to be estimated from the data. There are two principals to classify data into clusters: first, the prior information is dependent on the given dataset to be analyzed; second, the quality of classification does affect the potential gain from using clusters of locations rather than individual locations. Benjamini & Heller (2007) developed both cluster FDR and weighted cluster FDR methods. The weighted procedure contains the following steps:

1. Order the cluster  $p$ -values  $p_{(1)} \leq \dots \leq p_{(j)} \leq \dots \leq p_{(m)}$ .
2. Let  $w_{(j)}$  be the weight associated with  $p_{(j)}$ , and  $w_{(j)} = m\lambda(C_i) / \sum_{i=1}^m \lambda(C_i)$ ,  $\lambda(C_i)$  is the number of rejections in cluster  $C_i$  under the threshold  $\lambda$ .
3. Let  $k = \max\{j : p_{(j)} \leq (\sum_{i=1}^j w_{(i)}/m)q\}$  and reject the clusters corresponding to the smallest  $k$   $p$ -values.

They also extended the above procedure to a two stage adaptive procedure where the sum of weights of null clusters are updated after stage 2. To further identify the signals within a selected cluster, they developed a hierarchical procedures, called the cluster testing and trimming (CTT) procedure.

They showed that the adaptive procedure controls the cluster FDR as well as size weighted cluster FDR under independence of test statistics. When this assumption is violated, the one stage method still maintains control of the WFDR if the test statistics satisfy PRDS.

## 2.5 Adjusting multiplicity under dependence

Many procedures discussed so far control FWER or FDR under a common assumption: the test statistics are independent. However, in many situations, this

assumption is not valid. For example, brain imaging data are often spatially dependent; gene expression data usually have a clustered dependence structure; and multiple hypothesis testings of the same endpoint at different stage of a clinical trial are certainly not independent. In reality, dependent test statistics are encountered more often than independent test statistics. Therefore, more investigation is needed for the multiple testing under dependence.

### 2.5.1 p-value adjustment

Benjamini & Yekutieli (2001) proved that the BH step-up procedure controls FDR when the test statistics are positively dependent under the null hypothesis. This positive dependence covers a variety of practical situations including comparisons of multiple treatment group with a single control group, multivariate normal, and multivariate  $t$  distribution, etc. The author also modified the BH method so that it can control FDR conservatively under other dependence structure. Sarkar (2002) extended Benjamini & Yekutieli (2001)'s result to a more general step-wise procedures and also proved theoretically that the BL step-down procedure controls the FDR when the test statistics are positively dependent. Sarkar (2008*b*) further explored the performance of general step-wise procedures under positive dependence. Storey (2004)'s point estimate of FDR provides a conservatively biased estimate of the FDR under weak dependence.

### 2.5.2 Factor based adjustment

Efron (2007) became the first to capture the dependent structure at the test statistics level instead of the  $p$ -value level. He used a dispersion factor to account for the variation in a given dataset. This algorithm performs well only when the dependence among the null hypothesis is low. It deviates from the true FDP when the dependence is high.

Leek & Storey (2008) and Desai & Storey (2012) assumed that the dependence structure of a high dimensional variation can be captured by a low dimensional latent vectors so that the remaining errors are independent. However, the assumption of strict factor model is stringent and lacks formal justification. Sun & Tony Cai (2009) incorporated the local dependence structure among hypotheses in their procedure assuming that the data come from a two-stage hidden Markov model. Clarke & Hall (2009) discovered under high dimensional settings, performance of FDR procedures developed under the assumption of independence will not suffer under weak dependence as long as the null distribution of the test statistics are lightly tailed.

Friguet et al. (2009) assumed the dependence structure of the responses conditioned by the predictors can be modeled by a strict factor model with independent idiosyncratic errors. Factors are estimated by EM algorithm and subtracted out from the data so that conventional FDR controlling method like BH method can be applied. They showed that this procedure is more powerful and results in more precise FDR estimates than the traditional BH algorithm when used on dependent tests. However, when the correlation presented in the data is small, the number of factors are underestimated using their method.

Instead of assuming a strict factor model, Fan et al. (2012) used an approximate factor model to capture the majority of the dependence among the test statistics so that the remaining dependence is weak. By removing restriction over the structure of the covariance matrix, the approximate factor model can be applied to any arbitrary dependence structure. The approximate common factors are estimated using  $L1$  regression instead of EM algorithm. Fan et al. (2012) also provide the asymptotic convergence rate of their estimator under certain conditions.

There are few restrictions in Fan et al.'s method: 1). The covariance matrix of the test statistics is known. In reality, the covariance matrix of a specific dataset is usually unknown. 2). The covariance matrix of the test statistics is assumed to be a correlation matrix. To overcome these limitations, Fan & Han (2017) evaluated the impact of marginal variances as well as eigenvalues/eigenvectors over the estimation

of FDP. The author outlined certain criteria for these estimate so that the convergence of FDP estimator can be achieved along with rigorous theoretical proof.

Fan et al. (2017) proposed a factor-Adjusted Robust Multiple Testing (FarmTest) procedure for large scale simultaneous inference with control of the false discovery proportion.

### 2.5.3 Robustness

Procedures that depend on adjusting  $p$ -values to control FDR are fast, easy to implement and robust. However, they suffer being overly conservative and are only able to control the FDR under weak dependence. All the factor based methods mentioned above assume joint normality of factors and noise. However, as we have stated in Chapter 1 that, in the real world, normality is rarely met especially for high dimensional data. As the dimension gets larger, more outliers are likely to appear, and this may lead to significant false discoveries. Procedures are needed to handle dependence and heavy-tailedness simultaneously.

Fan & Han (2017) provide some theoretical and numerical evidence that the PFA method could be extended to dependent  $t$ -distribution. They chose  $df = 6$  for the simulation of dependent  $t$ -distribution which is not a setting for heavy-tailed data. Recently, Fan et al. (2017) proposed a factor-adjusted robust multiple Testing (FarmTest) procedure which shows advantage in FDP control over a variety of factor-based methods for large-scale simultaneous inference when the data are generated from heavy-tailed distributions. The authors used an asymptotic estimation of the FDP by assuming the independence among the factor adjusted test statistics with the form  $FDP^A(z; \eta) = 2p\pi_0(\eta)\Phi(-z)/R(z)$ ;  $z > 0$ , instead of a direct estimation of the FDP in other methods such as Fan & Han (2017) and Fan et al. (2012). In this thesis, we proposed a procedure that can extend the existing FDP estimation using the exact form to a much broader distribution families including heavy-tailed distributions.

## 2.6 Nonparanormal Distribution

The nonparanormal family is a nonparametric extension of the normal family (Liu et al. 2009) and is equivalent to Gaussian copula model for continuous variables (Klaassen et al. (1997), Tsukahara (2005)). Copulas become of great interest to statisticians since they give a promising, flexible tool for understanding dependence among random variables and for modeling non-Gaussian multivariate data. In a truly nonparametric setting, one may be interested in estimating the associated copula  $C$  by the empirical copula, which is naturally defined using the joint and marginal empirical distributions. In semiparametric copula models, we assume that the copula  $C$  associated with  $F$  belongs to a parametric family  $(\theta)$ , where  $\theta$  is an  $p$ -dimensional parameter. Gaussian copulas allow any marginal distribution and any positive definite correlation matrix. Gaussian copulas consider only pairwise dependence between individual components of a random variable. The definition of copula was initially introduced by Sklar (1959).

**Theorem 2.6.1 (Sklar, 1959)** *Suppose  $X_1, \dots, X_p$  are random variables with continuous marginal distribution function  $F_{X_1}, \dots, F_{X_p}$  and joint distribution function  $F$ . Then there exists a  $p$ -copula  $C$  (distribution function on  $[0, 1]^d$  with uniform marginals) such that for all  $(x_1, \dots, x_p)^T \in \mathbb{R}$ ,*

$$F(x_1, \dots, x_p) = C(F_{X_1}, \dots, F_{X_p}) \quad (2.1)$$

The connection between nonparanormal distribution and Gaussian copula is made clear in Lemma 1 in Liu et al. (2009). For nonparanormal we have:

$$F(x_1, \dots, x_p) = \Phi_{\mu, \Sigma}(\Phi^{-1}(F_{X_1}), \dots, \Phi^{-1}(F_{X_p}))$$

$\Phi_{\mu, \Sigma}$  is the multivariate Gaussian CDF. Thus the corresponding copula becomes

$$C(u_1, \dots, u_p) = \Phi_{\mu, \Sigma}(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_p))$$

The nonparanormal family removes many constraints of the Gaussian family. Nonparanormal distribution can be multi-modal, heavy-tailed, can model nonlinear dependencies amongst variables, and can even handle discrete dataset.

Nonparanormal distributions have been used to model dependencies among high-dimensional data in many fields, such as discrete neural response data (Berkes et al. 2009), time series (Malevergne et al. (2003), Chen & Fan (2006)), text regression and prediction (Wang & Hua 2014), extreme value analysis in hydrology and climate data (Renard & Lang (2007), AghaKouchak (2014)), graphical analysis of gene expression data (Liu et al. (2012), Xue & Zou (2012)), data compression in communication and crime data (Rey & Roth 2012), and mutual information in neuroimaging data (Ince et al. 2017).

One of the most important step for nonparanormal modeling is to estimate the correlation matrix. We will review several approaches proposed by recent literatures.

### 2.6.1 Empirical Correlation Matrix

This method involves two steps. The first step is to estimate the transformation function  $\mathbf{g}$  defined in Definition 1.2.1 in Chapter 1. This process is called "Gaussianization" by Szabó et al. (2007). Liu et al. (2009) uses a winsorized empirical CDF estimator to estimate the marginal CDF. Then uses equation 1.8 to obtain the transformed data. Whereas we proposed a kernel CDF based estimation for transformed data which outperforms in FDP approximation compared to *Winsorized*<sup>2</sup> estimator.

Let  $\widehat{Y}_{ij}$  be the estimated transformed data. By the definition of nonparanormal, the transformed data is approximately Gaussian. The second step is to estimate empirical covariance matrix of the the transformed data, which is  $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \widehat{Y}_i \widehat{Y}_i^T$ .

Fan, Liao and Mincheva (2013) developed a method called POET to estimate the unknown  $\Sigma$  based on samples. They applied a thresholding method to the noise part of the correlation matrix under the assumption of sparsity.

In this Thesis, we proposed a modified sample correlation matrix (mSCM) estimator. We compared our results to POET in multiple simulations in Chapter 4. Our mSCM estimator performs well in FDP approximation and is computationally more efficient.

## 2.6.2 Rank Correlations

Rank correlations usually include Spearman's  $\rho$  and Kendall's  $\tau$ . For two random variables  $X$  and  $Y$  with CDFs  $F_X, F_Y: \mathbb{R} \in [0, 1]$ , Spearman's  $\rho$  and Kendall's  $\tau$  are defined by

$$\begin{aligned}\rho(X, Y) &= \text{Corr}(F_X(X), F_Y(Y)) \\ \tau(X, Y) &= \text{Corr}(\text{sign}(XX), \text{sign}(YY)).\end{aligned}$$

Liu et al. (2012) proposed the following rank correlation estimators. Let  $r_j^i$  be the rank of  $x_j^i$  among  $x_j^1, \dots, x_j^n$  and  $\bar{r}_j = \frac{1}{n} \sum_{i=1}^n r_j^i = \frac{n+1}{2}$ .

$$\begin{aligned}(\text{Spearman's rho}) \hat{\rho}_{jk} &= \frac{\sum_{i=1}^n (r_j^i - \bar{r}_j)(r_k^i - \bar{r}_k)}{\sqrt{\sum_{i=1}^n (r_j^i - \bar{r}_j)(r_k^i - \bar{r}_k)}}, \\ (\text{Kendall's tau}) \hat{\tau}_{jk} &= \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \text{sign}((x_j^i - x_j^{i'})(x_k^i - x_k^{i'})).\end{aligned}$$

Both  $\hat{\rho}_{jk}$  and  $\hat{\tau}_{jk}$  are nonparametric correlations between the empirical realizations of random variables  $X_j$  and  $X_k$ . Note that these statistics are invariant under monotone transformations. For nonparanormal distributions, the following lemma connects Spearman's rho and Kendall's tau to the underlying Pearson correlation coefficient  $\Sigma_{jk}^0$ .

**Lemma 2.6.2 ((Kendall 1948) (Kruskal 1958))** *Assuming  $X \sim NPN_d(\boldsymbol{\mu}, \mathbf{g}, \boldsymbol{\Sigma}_0)$ , without loss of generality, we assume  $\sigma_{jj} = 1$ , we have  $\Sigma_{0jk} = 2 \sin(\frac{\pi}{6} \rho_{jk}) = \sin(\frac{\pi}{6} \tau_{jk})$ .*

Based on this lemma, the following estimators  $\widehat{S}^\rho = [\widehat{S}_{jk}^\rho]$  and  $\widehat{S}^\tau = [\widehat{S}_{jk}^\tau]$  are unbiased estimators for the unknown correlation matrix  $\Sigma_0$ :

$$\widetilde{S}_{jk}^\rho = \begin{cases} 2 \sin(\frac{\pi}{6} \widehat{\rho}_{jk}) & j \neq k \\ 1 & j = k \end{cases} \quad \text{and} \quad \widetilde{S}_{jk}^\tau = \begin{cases} \sin(\frac{\pi}{2} \widehat{\tau}_{jk}) & j \neq k \\ 1 & j = k \end{cases}$$

Liu's main goal is to use Spearman's rho and Kendall's tau to directly estimate the unknown correlation matrix, without explicitly calculating the marginal transformations.

Liu has also proved that  $\widehat{S}^\rho$  and  $\widehat{S}^\tau$  have a fast exponential convergence rate to  $\Sigma_0$  in the  $\|\cdot\|_{\max}$  norm under the assumption  $d > n$ .

In Chapter 4, simulations were performed to compare Liu et al.'s rank based statistics vs our modified sample correlation matrix estimator.



## CHAPTER 3

### NPN-PFA PROCEDURE

Without further notice, all notations used in this section will be the same as in Chapter 1.

#### 3.1 Estimating Marginal Transformation Functions

Recall equation 1.8,  $\hat{Y}_{ij} = \hat{\mu}_j + \hat{\sigma}_j \Phi^{-1}(\hat{F}_j(X_{ij}))$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, p$ . To estimate the marginal transformation function, we need to estimate the marginal CDF  $F_j$ . Define the following kernel estimator of  $F_j$ :

$$\hat{F}_j(z) = \int_{-\infty}^z \hat{f}_j(t) dt = \frac{1}{n} \sum_{j=1}^n G\left(\frac{z - Z_j}{h}\right), \quad (3.1)$$

where  $G(z) = \int_{-\infty}^z g(y) dy$  is the CDF of a kernel function.

A range of commonly used kernel functions include uniform, triangular, Epanechnikov, normal. Normal kernel is often chosen for computational simplicity and for any given value of  $x$ , which uses all available sample points for smoothing. On the other hand, kernel functions like uniform, triangular, or Epanechnikov all have a fixed range so that points residing outside of the threshold window are not used for smoothing.

For normal kernel function, equation (3.1) becomes:

$$\hat{F}_j(x) = \frac{1}{n} \sum_{j=1}^n \Phi\left(\frac{x - X_j}{h}\right),$$

Any non-negative, symmetric function that integrates to one and has mean zero can be used as a kernel function. For the feasibility of theoretical proof, we use truncated normal kernel:

$$\widehat{F}_j(x) = \frac{1}{n} \sum_{j=1}^n G\left(\frac{x - X_j}{h}\right),$$

where

$$G(u) = \begin{cases} 1, & \text{if } u < -a, \\ \frac{\Phi(u) - \Phi(-a)}{\Phi(a) - \Phi(-a)}, & \text{if } -a \leq u \leq a, \\ 0, & \text{if } u > a. \end{cases} \quad (3.2)$$

By definition

$$\widehat{F}_j(X_i) = \frac{1}{n} \sum_{i=1}^n G\left(\frac{X_i - X_j}{h}\right) = \frac{1}{n} \sum_{\{j: X_j \leq X_i - ah\}} \mathbb{1} + \frac{1}{n} \sum_{\{j: X_i - ah \leq X_j \leq X_i + ah\}} G\left(\frac{X_i - X_j}{h}\right).$$

When  $n$  is sufficiently large,

$$\frac{1}{n} \sum_{\{j: X_j \leq X_i - ah\}} \mathbb{1} \xrightarrow{a.s.} P(X \leq X_i - ah),$$

and  $\widehat{F}_j(X_i)$  can be approximated by

$$\widetilde{F}_j(X_i) = P(X \leq X_i - ah) + \frac{1}{n} \sum_{\{j: X_i - ah \leq X_j \leq X_i + ah\}} G\left(\frac{X_i - X_j}{h}\right). \quad (3.3)$$

For estimation of sample mean  $\widehat{\mu}_j$ , we consider a thresholding estimator  $\widehat{\mu}_j$  for  $\mu_j$ . Let  $\mu'_j$  be the sample mean, then  $\mu'_j = \mu_j + O_p(n^{-1/2})$ . Since we know most of the  $\mu_j$ s are zero, we will consider

$$\widehat{\mu}_j = \begin{cases} 0, & \text{if } |\mu'_j| < \Delta_n, \\ \mu'_j, & \text{if } |\mu'_j| \geq \Delta_n \end{cases} \quad (3.4)$$

to improve the convergence rate, where the threshold  $\Delta_n = \frac{c}{\sqrt{2\pi \ln(n)}}$  for some pre-defined constant  $c$ . By thresholding,  $\hat{\mu}_j$  will converge to  $\mu_j$  at a faster rate. The detailed proof is provided in the Appendix.

Thus,

$$\tilde{y}_{ij} = \tilde{g}_j(x_{ij}) = \hat{\mu}_j + \hat{\sigma}_j \Phi^{-1}(\tilde{F}_j(x_{ij})), \quad i = 1, \dots, n, \quad j = 1, \dots, p. \quad (3.5)$$

Based on  $\tilde{y}_{ij}$ , we have the corresponding approximation of the test statistics,  $p$ -value, number of total rejections, false discoveries and FDP:

$$\tilde{Z}_j = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{Y}_{ij},$$

$$\tilde{p}_j = 2(1 - \Phi(|\tilde{Z}_j|)),$$

$$\tilde{R}(t) = \sum_{j=1}^p \mathbf{I}\{\tilde{p}_j \leq t\},$$

$$\tilde{V}(t) = \sum_{j=1}^p [\Phi(\hat{a}_j(z_{t/2} + \tilde{\eta}_j)) + \Phi(\hat{a}_j(z_{t/2} - \tilde{\eta}_j))], \quad (3.6)$$

$$\widehat{\text{FDP}}(t) = \sum_{j=1}^p [\Phi(\hat{a}_j(z_{t/2} + \tilde{\eta}_j)) + \Phi(\hat{a}_j(z_{t/2} - \tilde{\eta}_j))] / \tilde{R}(t) \quad (3.7)$$

where  $\hat{a}_j = (1 - \|\hat{\mathbf{b}}_j\|^2)^{-1/2}$ ,  $\tilde{\eta}_j = \hat{\mathbf{b}}_j^T \tilde{\mathbf{W}}$ .  $\hat{\mathbf{b}}_j$  is the  $j$ -th row of the estimated  $k$  principal components and  $\tilde{\mathbf{W}}$  are common factors estimated from  $\tilde{Z}_j$ .

For FDP approximation, like expression (3.7), we also need an estimate of the correlation matrix  $\Sigma$ , such that the eigenvalues and eigenvectors can be plugged in the formula. We propose the following estimation.

Let  $\Gamma_{jk}$  denote the  $jk^{\text{th}}$  element of the population correlation matrix, then

$$\begin{aligned} \Gamma_{jk} &= \mathbf{E} \left[ \frac{1}{\sigma_j \sigma_k} (y_j - \mu_j)(y_k - \mu_k) \right] \\ &= \mathbf{E} \left[ \Phi^{-1}(F_j(X_j)) \Phi^{-1}(F_k(X_k)) \right]. \end{aligned}$$

$\Gamma_{jk}$  can be estimated by

$$\widehat{\Gamma}_{jk} = \frac{1}{n} \sum_{i=1}^n \Phi^{-1}(\widehat{F}_j(X_{ij})) \Phi^{-1}(\widehat{F}_k(X_{ik})). \quad (3.8)$$

which we define as modified sample correlation matrix (mSCM) estimator.

### 3.2 Proposed Method

Step 1. Estimate marginal transformation function. Recall equation 1.8,

$$\widehat{Y}_{ij} = \widehat{\mu}_j + \widehat{\sigma}_j \Phi^{-1}(\widehat{F}_j(X_{ij})), \quad i = 1, \dots, n, \quad j = 1, \dots, p,$$

$\mu$  and  $\sigma$  and  $F_j(X_{ij})$  will be estimated using sample mean defined in 3.4, sample variance, and Kernel CDF estimator.

Step 2. Estimate correlation matrix using mSCM,

$$\widehat{S}_{jk} = \frac{1}{n} \sum_{i=1}^n \Phi^{-1}(\widehat{F}_j(X_{ij})) \Phi^{-1}(\widehat{F}_k(X_{ik})).$$

Step 3. Calculate test statistics based on  $\widehat{Y}_{ij}$  obtained from Step 1,

$$\widehat{Z}_j = \frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{Y}_{ij}$$

Step 4. Calculate the number of total rejections,

$$\widehat{R}(t) = \sum_{j=1}^p I(\widehat{p}_j < t), \quad \text{where } \widehat{p}_j = 2(1 - \Phi(|\widehat{Z}_j|))$$

Step 5. Estimate the factors and factor loading matrix, then the FDP,

$$\widehat{\text{FDP}}_A(t) = \sum_{j=1}^p [\Phi(\widehat{a}_j(z_{t/2} + \widehat{\eta}_j)) + \Phi(\widehat{a}_j(z_{t/2} - \widehat{\eta}_j))] / \widehat{R}(t).$$

### 3.3 Theoretical Investigations

#### 3.3.1 Correlation Matrix

As we have stated previously, a crucial step for a good FDP approximation is to develop a method for estimating the underlying correlation matrix of the multivariate normal distribution Fan & Han (2017). The estimated correlation matrix will be used to obtain the eigenvalues and eigenvectors in the FDP approximation. Under certain regularity conditions of the estimated correlation matrix, the subsequent FDP approximation can perform well. This motivates us to present the following Theorem 1. Here we consider the mSCM estimator defined in (3.8).

**Theorem 3.3.1** *Suppose a random vector  $\mathbf{X}$  has a nonparanormal distribution:  $\mathbf{X} \sim NPN(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$  and the sampled data  $\{X_{i,k}\}$  are bounded in such a way that satisfying  $0 < a_k \leq F_k(X_{i,k}) \leq b_k < 1$  for constants  $a_k, b_k \forall k, i$ . Let  $\boldsymbol{\Gamma}$  be the true correlation matrix of the underlying multivariate normal and  $\widehat{\boldsymbol{\Gamma}}$  be the estimated correlation matrix. We have the following*

$$\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\|_{op} = O_p(pn^{-\delta}) \text{ when } \ln p \leq \frac{1}{2}n^{1-2\delta} \text{ for some constant } 0 < \delta < 1/2.$$

Note that we do not need a consistent estimate of the correlation matrix for a good approximation of FDP as shown in Fan & Han (2017). As theorem 1 shows, we can handle a situation where  $p = O(e^{\frac{1}{2}n^{1-2\delta}})$  which is similar to the NP dimensionality defined in Fan & Lv (2008). The assumption of sample data bounded on the CDF ensures that the samples are free of extremely small or large values.  $\{X_{i,k}\}$  could still follow  $N(0, 1)$  or some distributions with compact support.

### 3.3.2 Number of False Rejections

For simplicity of presentation, we assume that each coordinate of  $\mathbf{X}$  follows the same distribution. The proof can be easily extended to situations where each coordinate of  $\mathbf{X}$  are not identically distributed. Let  $\mathbf{Y}$  be the underlying multivariate normal vector,  $\mathbf{f}$  be the density function of  $\mathbf{X}$ ,  $\mathbf{V}(t)$  be the false number of rejections defined in the numerator in equation (1.5),  $\tilde{\mathbf{Y}}$  be the vector described in (3.5) and  $\lambda_i$  be the  $i$ th largest eigenvalue of  $\mathbf{\Gamma}$ .

**Theorem 3.3.2** *With the conditions in Theorem 1, in addition, if the following conditions are satisfied,*

- (1) *Suppose the nonparanormal random vector  $\mathbf{X}$  has continuous density function  $\mathbf{f}$  and  $\mathbf{f}'(0)$  is bounded,*
- (2) *the density of  $X_{1i} - X_{1j}$  exists for  $i \neq j$  and is bounded at 0, and its second derivative is bounded in a neighborhood around 0,*
- (3)  *$\lambda_i - \lambda_{i+1} \geq d_p$  for a sequence  $d_p > 0$  for  $i = 1, \dots, k$ , where  $d_p = O(p^{1+\tau}n^{-\delta})$  for some constant  $\tau > 0$  and  $\delta > 0$ ,*
- (4)  *$\hat{a}_i \leq \tau_1$  and  $a_i \leq \tau_2 \quad \forall i = 1, \dots, p$  for some finite constants  $\tau_1$  and  $\tau_2$ .*

*when  $kn^{-\min\{1/6, \delta\}} = o(1)$ ,  $kp^{-\frac{1}{2}-\tau}\|\boldsymbol{\mu}^*\| = o(1)$ ,  $kp^{-\tau} = o(1)$ ,  $k \exp(-c_1 \frac{n}{2 \ln n}) = o(1)$  for some constant  $c_1 > 0$ , and  $p^{-1}\sqrt{\lambda_{k+1} + \dots + \lambda_p} = O(p^{-\epsilon})$  for some  $\epsilon > 0$ , we can conclude:*

$$p^{-1} \left| \tilde{V}(t) - V(t) \right| = o_p(1)$$

For a high dimensional factor model, it is usually satisfied that  $d_p = O(p)$  for the first few eigenvalues. Therefore, Condition 3 in Theorem 3.3.2 can be easily satisfied with such strong dependence structure (Fan, Liao & Mincheva 2013).

Condition 4 is satisfied when the variances of  $\mathbf{K}$  in (1.3) and their estimates do not converge to zero. For the terms with  $\|\boldsymbol{\mu}^*\|$ ,  $k\|\boldsymbol{\mu}^*\|p^{-1/2} \rightarrow 0$  as  $p \rightarrow \infty$ , when  $\{\mu_i^*\}_{i=1}^p$  are sparse. In false discovery control, the arguably most important part is the number of false discoveries. With further assumptions such as  $R(t) = O_p(p)$ , see Fan & Han (2017), the result in Theorem 3.3.2 can be connected to the false discovery proportion. The simulation results in Chapter 4 show that our proposed method performs well in terms of false discovery proportion approximation. Theorem 3.3.2 provides some theoretical support for this phenomena.

### 3.4 Dependence Adjusted Procedure

In multiple testing setting especially high dimensional testings, dependence among test statistics will affect both FDP and false non-discovery rate (FNP). Re-ordering test statistics and  $p$ -values by incorporating the dependent structure will improve the testing efficiency, see Fan, Han & Gu (2012) and Fan & Han (2017). Based on equation 1.3,  $a_j(Z_j - b_j^T W) \sim N(a_j \mu_j, 1)$ , whereas  $a_j = (1 - \|\mathbf{b}_j\|^2)^{-1/2}$ . Since  $a_j > 1$ , this increases the strength of signals and the test statistics value. The  $p$ -value based on this adjusted test statistic is now  $2\Phi(-|a_j(Z_j - b_j^T W)|)$ . The critical region is  $|a_j(Z_j - b_j^T W)| \leq |z_{t/2}|$ . In our setting, the covariance matrix  $\boldsymbol{\Sigma}$  is unknown, the underlying multivariate normal samples  $Z_{ij}$  is also unknown, we calculate the  $p$ -values as  $\hat{p}_j = 2\Phi(-|\hat{a}_j(\hat{Z}_j - \hat{b}_j^T \hat{W})|)$ , where  $\hat{a}_j$ ,  $\hat{b}_j$  and  $\hat{W}$  have been defined in previous section, and  $\hat{Z}_j$  is the test statistics based on estimated samples  $\hat{y}_{ij}$  in 1.8. Our simulation has confirmed the improved power of this dependence-adjusted procedure compared with the non-adjusted procedure.



## CHAPTER 4

### SIMULATION AND REAL DATA ANALYSIS

#### 4.1 Simulation Result

In this section, we simulate two different types of nonparanormal data to compare the *Winsorized*<sup>2</sup> estimator vs the kernel density estimator in terms of FDP approximation. We also evaluated the impact of FDP approximation by three different methods of estimating correlation matrix: evaluate modified sample correlation matrix (mSCM), rank-based estimators or POET.

To simplify the presentation, we set the transformations for all dimensions the same:  $f_1 = \dots = f_p = f$ . Two different transformations were utilized: Gaussian CDF transformation and power transformation, see Liu et al. (2009). Let  $g$  be the inverse transformation function from a normal distribution to a nonparanormal distribution:  $g \equiv f^{-1}$ .

**Definition 4.1.1** (*CDF Transformation*) Let  $g_0$  be a one-dimensional Gaussian cumulative distribution function with mean  $\mu_{g_0}$  and the standard deviation  $\sigma_{g_0}$ ,

$$g_0(t) \equiv \Phi\left(\frac{t - \mu_{g_0}}{\sigma_{g_0}}\right)$$

The inverse transformation function  $g_j = f_j^i$  for the  $j$  – th dimension as

$$g_j(z_j) \equiv \sigma_j \left( \frac{g_0(z_j) - \int g_0(t) \phi\left(\frac{t - \mu_j}{\sigma_j}\right) dt}{\sqrt{\int \left( g_0(y) - \int g_0(t) \phi\left(\frac{t - \mu_j}{\sigma_j}\right) dt \right)^2 \phi\left(\frac{y - \mu_j}{\sigma_j}\right) dy}} \right) + \mu_j$$

where  $\sigma_j = \Sigma_0(j, j)$ .

**Definition 4.1.2** (*Power Transformation*) Let  $g_0$  be the symmetric and odd transformation given by

$$g_0(t) = \text{sign}(t)|t|^\alpha$$

where  $\alpha$  is a nonnegative parameter. The power transformation for the  $j$ -th dimension is defined as

$$g_j(z_j) \equiv \sigma_j \left( \frac{g_0(z_j - \mu_j)}{\sqrt{\int g_0^2(t - \mu_j) \phi\left(\frac{t - \mu_j}{\sigma_j}\right) dt}} \right) + \mu_j$$

These transformations are constructed to preserve the marginal mean and standard deviation.

#### 4.1.1 Known fixed number of factors

We first generate mixed normal data from a strict 3-factor model. We set the dimensionality  $p = 1000$ , the sample size  $n = 50, 100$ , the number of false nulls (sparsity) to be  $p_1 = 40$ , the threshold value  $t = 0.01$ , and the number of simulation round to be 200. In this model:

$$z_j = \mu + Bf_j + u_j, \quad f_j \sim N_3(0, I_3), \quad u_j \sim N_p(0, \Sigma_u),$$

the population mean  $\mu_j = 1$  for  $j = 1, \dots, 40$  and 0 otherwise.  $B$  is the factor loading matrix and each entry is generated from the uniform distribution  $U(-1, 1)$ .  $f_j$  is the common factors and it's independent of the noise  $u_j$ .  $\Sigma_u$  is set to be a diagonal matrix with 0.5 on it diagonal. For simplicity and without loss of generality, the covariance matrix of  $\mathbf{Z}$  is normalized so that its variances are 1.

To generate the nonparanomal data, we set  $\mu_{g_0} = 0.05$ , and  $\sigma_{g_0} = 0.4$  for the cdf transformation. For the power transformation, we set  $\alpha = 0.5$ . Then *Winsorized*<sup>2</sup> or kernel method is used to estimate the cumulative density function of the nonparanor-

mal data. Threshold sample mean is based on 3.4 and  $\Delta$  is set to  $c = 0.25$ . Sample variance of each of the dimension are used along with estimated CDF to produce transformed multivariate normal data. To estimate FDP using PFA, correlation matrix of the transformed data is estimated using mSCM (Figure 4.1), POET (Figure 4.2) or rank-based estimators: Kendall's tau and Spearman's rho (Figure 1.2).

We used the following settings to compare the realized  $FDP(t)$  values with their estimators  $\widehat{FDP}(t)$ : (*Winsorized*<sup>2</sup> + mSCM), (*Winsorized*<sup>2</sup> estimation + Kendall's tau), and (*Winsorized*<sup>2</sup> estimation+Spearman's rho), (Normal kernel+mSCM), (Truncated normal kernel (3SD) + mSCM), (Truncated normal kernel (5SD) + mSCM), (*Winsorized*<sup>2</sup>+POET), (Normal kernel+POET), (Truncated normal kernel (3SD) estimation + POET), (Truncated normal kernel (5SD) + POET). Table 1.3, 4.1 and 4.2 displayed the mean error and its standard deviation for the above settings. Overall, Kernel density estimation performed better than *Winsorized*<sup>2</sup> estimation. For settings used mSCM (Table 4.1), normal and truncated kernel with any bandwidth works well for both CDF and power transformation. Regular normal kernel setting also produced a nearly unbiased result with mean error of 0.001 and 0.01 with  $n = 100$  for CDF transformation and power transformation respectively. Note that *Winsorized*<sup>2</sup> estimation always produce underestimated FDP which is undesirable for scientific studies. There is no significant difference between mSCM and POET, but mSCM is more computationally convenient than POET. On the other hand, both Kendall's tau and Spearman's rho performed poorly (Table 1.3).

Corresponding numerical values were plotted in Figure 1.2, 4.1 and 4.2 with  $n = 50$ . Figure 1.2 represents the performance of FDP approximation using *Winsorized*<sup>2</sup> estimation with mSCM or rank-based estimators with data generated using CDF transformation and power transformation functions. Settings using *Winsorized*<sup>2</sup> estimator deviate clearly from the 45 degree line, whereas kernel density estimations perform better for both CDF and power transformation. Figure 4.2 replaced mSCM with POET, and the results have a similar pattern as Figure 4.1.

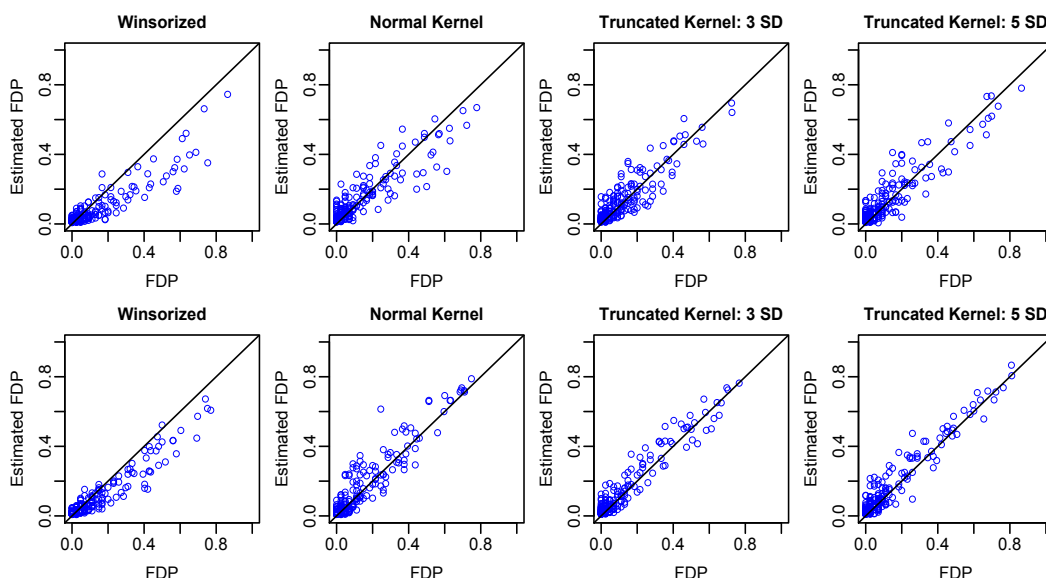


Figure 4.1. FDP estimation with known number of factors ( $n=50$ ,  $p=1000$ ,  $t=0.01$ ). Top panel: CDF transformation; bottom panel: Power transformation. Correlation matrix is estimated using mSCM.

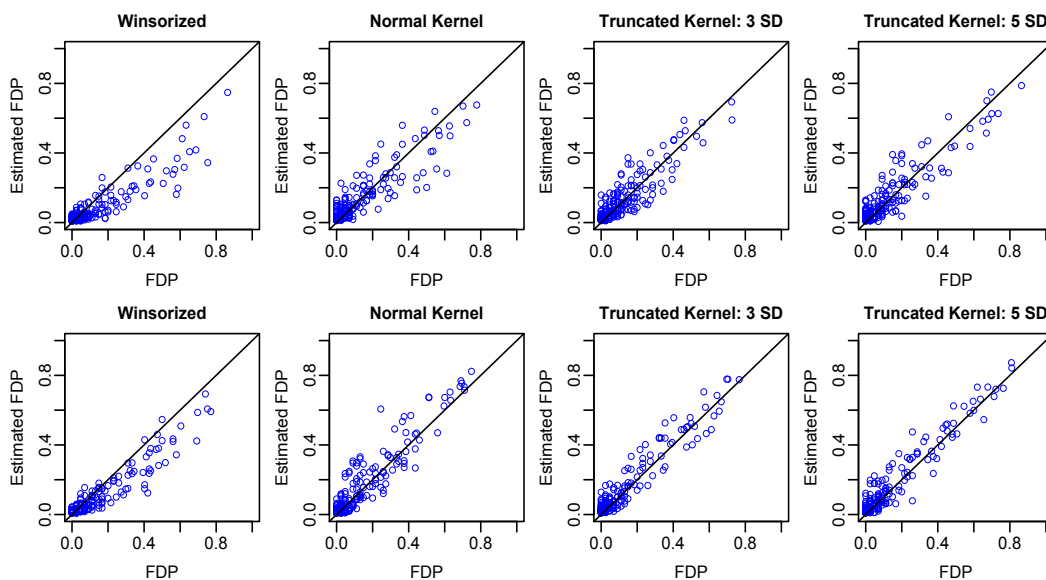


Figure 4.2. FDP estimation with known number of factors ( $n=50$ ,  $p=1000$ ,  $t=0.01$ ). Top panel: CDF transformation; bottom panel: Power transformation. Correlation matrix is estimated using POET.

Table 4.1. Mean(SD) of  $\widehat{\text{FDP}}(t) - \text{FDP}(t)$ , expressed in percentage, comparing *Winsorized*<sup>2</sup> vs NPN-PFA. Correlation matrix  $\widehat{\Gamma}$  is estimated using mSCM.

	<i>Winsorized</i> <sup>2</sup>	Normal	TN <sup>1</sup> 3SD	TN <sup>1</sup> 4SD	TN <sup>1</sup> 5SD
CDF Transformation					
n=50	-4.18(8.56)	0.81(7.66)	1.19(6.4)	0.86(7.11)	1.31 (6.58)
n=100	-3.67(7.15)	0.14(7.56)	0.6(5.89)	-0.51(6.88)	0.55(5.32)
Power Transformation					
n=50	-3.37(6.52)	1.91(7.08)	1.88(5.13)	2.32(6.39)	2.23(5.41)
n=100	-2.56(6.12)	1.08(5.42)	0.7(5.86)	0.7(5.39)	0.58(4.94)

<sup>1</sup>Truncated Normal

Table 4.2. Mean(SD) of  $\widehat{\text{FDP}}(t) - \text{FDP}(t)$ , expressed in percentage, comparing *Winsorized*<sup>2</sup> vs NPN-PFA. Correlation matrix  $\widehat{\mathbf{\Gamma}}$  is estimated using POET.

	<i>Winsorized</i> <sup>2</sup>	Normal	TN <sup>1</sup> 3SD	TN <sup>1</sup> 4SD	TN <sup>1</sup> 5SD
CDF Transformation					
n=50	-4.24(8.6)	0.74(7.65)	1.11(6.39)	0.78(7.11)	1.24(6.57)
n=100	-3.72(7.16)	0.07(7.56)	0.54(5.88)	-0.57(6.89)	0.49(5.33)
Power Transformation					
n=50	-3.46(6.56)	1.85(7.06)	1.83(5.11)	2.25(6.37)	2.17(5.4)
n=100	-2.62(6.13)	1.04(5.41)	0.65(5.86)	0.65(5.39)	0.52(4.93)

<sup>1</sup>Truncated Normal

We evaluated the choice of  $c$  on the estimation of FDP in simulations, see Table 4.3. Note that the impact of different  $c$  value over FDP approximation is minimum.

Table 4.3. Mean(S.D.) of  $\widehat{\text{FDP}}(t) - \text{FDP}(t)$ , expressed in percentage, comparing different values of  $c$  (using normal kernel and mSCM).

	c=0.05	c=0.15	c=0.25	c=0.35
CDF Transformation				
n=50	1.36 (7.5)	1.8 (6.4)	0.81(7.66)	1.0 (6.5)
n=100	1.4 (7.6)	1.4 (6.4)	0.14(7.56)	-0.9 (6.8)
Power Transformation				
n=50	2.3 (6.9)	2.3 (4.95)	2.4 (6.1)	1.76 (5.3)
n=100	1.9 (5.6)	1.2 (6.01)	1.1 (4.9)	0.57 (4.97)

#### 4.1.2 Unknown number of factors

In reality, the number of factors are unknown and we need to estimate it. If the covariance matrix  $\Sigma$  is known, we can choose the smallest  $k$  such that  $\frac{\sqrt{\lambda_{k+1}^2 + \dots + \lambda_p^2}}{\lambda_1 + \dots + \lambda_p}$  holds for a small  $\epsilon$ , say 0.01. However, in our setting the  $\Sigma$  is unknown. Fan & Han (2017) has elaborated an estimator defined as  $\hat{k}_{ER} = \operatorname{argmax}_{1 \leq k \leq k_{max}} (\tilde{\lambda}_k / \tilde{\lambda}_{k+1})$ , where  $\tilde{\lambda}_i$  is the  $i$ th largest eigenvalue of the sample correlation matrix and  $k_{max}$  is the maximum possible number of factors. We set the dimensionality  $p = 1000$ , the sample size  $n = 50, 100$ , the population mean for false nulls  $\mu_j = 1$  for  $j = 1, \dots, 40$ , the rejection region  $t = 0.01$ , the mean thresholding constant  $c = 0.25$  as defined in 3.4 and the number of simulation round 200. The maximum number of factors is set to be  $k_{max} = \lfloor 0.2n \rfloor$ .

To explore the applicability of our method under different dependent structures, we consider the following 3 settings which were originally defined by Fan & Han (2017). Data are generated from  $x_i \sim N_p(\mu, \Sigma)$ .

**Model 1: Strict Factor Model.** Recall the model described in Section 5.2.1:

$$z_j = \mu + Bf_j + u_j, \quad f_j \sim N_3(0, I_3), \quad \mu_j \sim N_p(0, \Sigma_u),$$

Data are generated based on a 3-factor model and then the number of factors are estimated before applying PFA method.

**Model 2: Approximate Factor Model.** The model set up is the same as Model 1, except that  $\Sigma_u$  is constructed as follows. First create a covariance matrix  $\Sigma_1$ , which was calibrated to the returns of *S&P500* constituent stocks. Then we construct a symmetric banded matrix  $\Sigma_2$ . For the  $(i, j)$ th element, if  $i \neq j$  and  $|i - j| \leq 25$ , set the element as 0.4 and zero otherwise. Next we construct a symmetric matrix  $\Sigma_3$  as the nearest positive definite matrix of  $\Sigma_1 + \Sigma_2$  by the algorithm of Higham (1988). Finally the covariance matrix  $\Sigma_u$  is set as  $0.5\Sigma_3$ .

**Model 3: Cluster Model.** Model 3 is designed against the eigengap condition in Theorem 3.3.2 and also to test the robustness of determining the number of factors. We first generate a  $p$ -dimensional vector  $\Lambda$ , where the first 4 elements are independent realizations from the uniform distribution  $U(160, 190)$ , the next 10 elements are independently from  $U(8, 12)$  and the rest are independently from  $U(0.1, 0.3)$ . Next we generate a  $p \times p$  matrix  $\mathbf{Q}$  in which each element is an independent realization from  $N(0, 1)$ . Let  $\mathbf{\Gamma}$  be the matrix, consisting of eigenvectors of  $\mathbf{Q}\Lambda\mathbf{Q}^T$ . Finally, let  $\Sigma = \mathbf{\Gamma}\Lambda\mathbf{\Gamma}^T$ .

Table 4.4 contains the empirical mean error (the mean difference between the true FDP and estimated FDP) and its SD comparing NPN-PFA method with normal kernel or truncated normal of 5SD and *Winsorized*<sup>2</sup> method. Both normal kernel and truncated normal kernel outperformed *Winsorized*<sup>2</sup> method under both power or CDF transformation for all 3 models. *Winsorized*<sup>2</sup> consistently produced larger mean error and underestimated the true FDP, which is not desirable for practical FDR control.

Figure 4.3 and 4.4 further demonstrates the performance of kernel based vs *Winsorized*<sup>2</sup> method under power or CDF transformation respectively. The sample size  $n = 50$ . Both normal kernel and truncated normal kernel method estimate the true FDP( $t$ ) very well for all 3 models and even capture the true value when FDP( $t$ ) is large.

Table 4.4. Mean(SD) of  $\widehat{\text{FDP}}(t) - \text{FDP}(t)$ , expressed in percentage, comparing *Winsorized*<sup>2</sup> vs NPN-PFA with unknown number of factors. Correlation matrix  $\hat{\mathbf{\Gamma}}$  is estimated using mSCM.

		<i>Winsorized</i> <sup>2</sup>	Normal	TN <sup>1</sup> 4SD
Model 1	CDF transformation			
	n=50	-3.78(6.88)	0.35(6.97)	-0.14(6.6)
	n=100	-3.05(7.07)	-0.24(6.34)	-0.27(6.62)
	Power transformation			
	n=50	-2.38(6.51)	0.87(5.13)	1.99(5.56)
	n=100	-2.11(6.28)	0.08(5.07)	0.26(6.65)
Model 2	CDF transformation			
	n=50	-3.55(7.83)	-0.68(6.88)	0.28(7.59)
	n=100	-2.92(7.05)	-0.63(6.82)	-0.32(6.87)
	Power transformation			
	n=50	-2.9(6.7)	0.46(6.29)	0.26(6.65)
	n=100	-2.38(6.35)	1.26(6.29)	0.01(5.95)
Model 3	CDF transformation			
	n=50	-4.07(8.8)	0.99(7.05)	1.89(8.15)
	n=100	-3.25(7.92)	0.72(6.58)	0.97(6.75)
	Power transformation			
	n=50	-3.34(7.87)	2.91(6.23)	3.16(7.06)
	n=100	-2.02(6.57)	2.74(6.48)	3.14(7.01)

<sup>1</sup> Truncated Normal

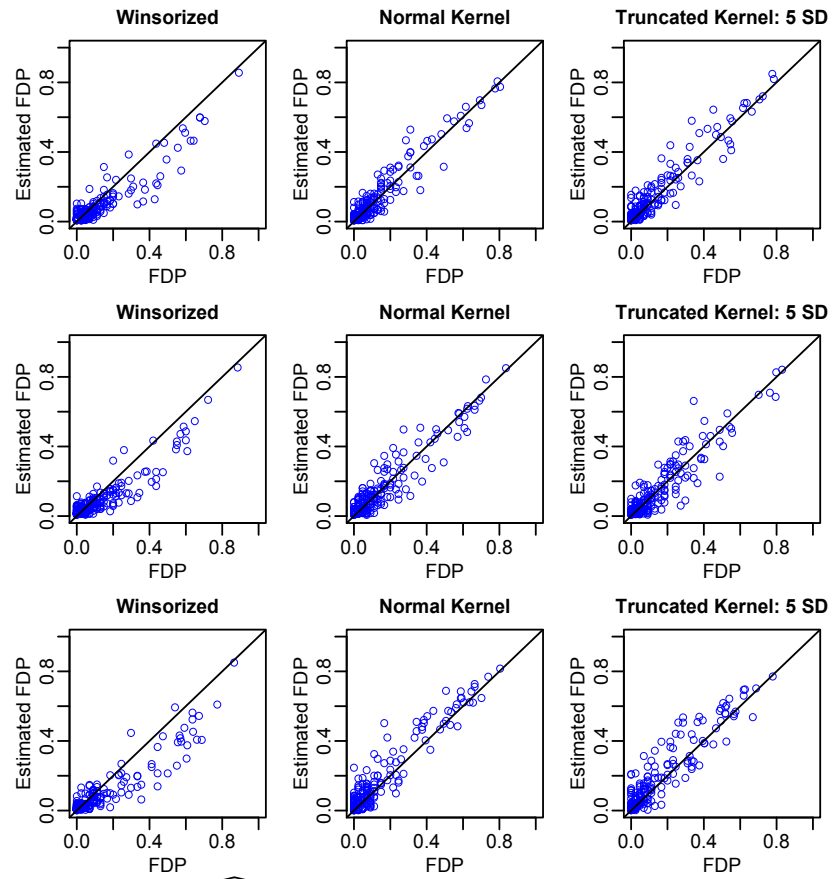


Figure 4.3. Comparison of  $\widehat{\text{FDP}}(t)$  with realized  $\text{FDP}(t)$  for power transformation with unknown number of factors. From top to bottom panel corresponds to Model 1 to Model 3.  $n = 50$ .

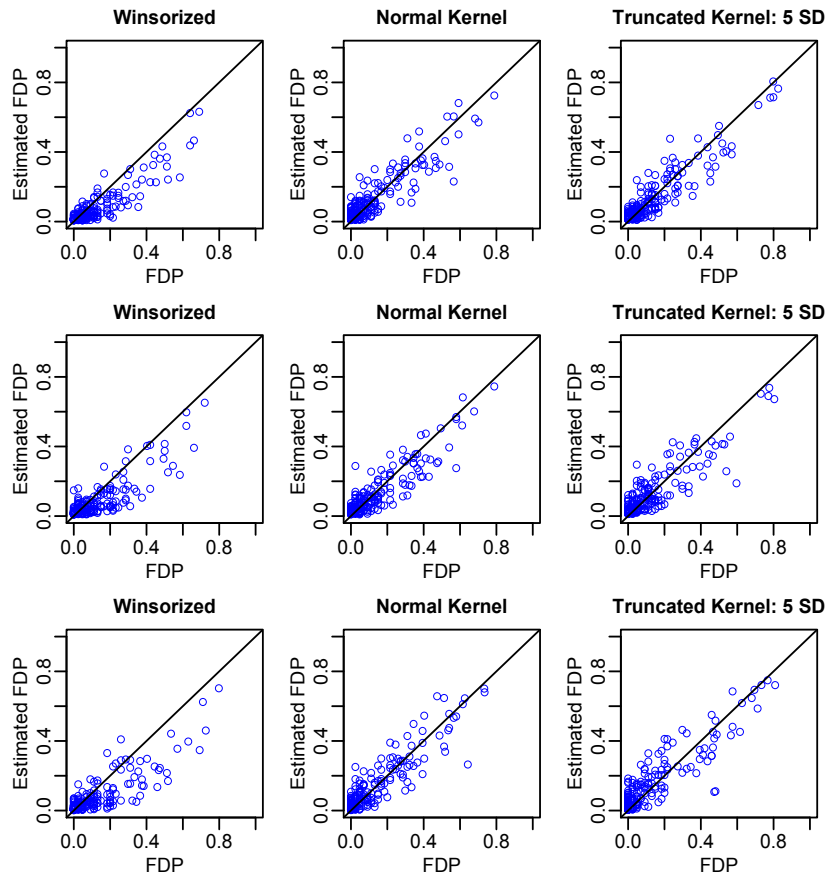


Figure 4.4. Comparison of  $\widehat{FDP}(t)$  with  $FDP(t)$  for CDF transformation. From top to bottom panel corresponds to Model 1 to Model 3.  $n = 50$ .

### 4.1.3 Band-width selection

It has been universally considered that the choice of bandwidth  $h$  is much more important than the choice of the type of kernel  $K$ . Many procedures aim to select the optimal bandwidth so that the mean integrated squared error (MISE) can be minimized. The MISE is defined as below.

$$MISE(h) = \mathbb{E}\left[\int \{\widehat{f}_h(x) - f(x)\}^2 dx\right]$$

For example, under the conditions in Sarda (1993), Altman & Léger (1995) showed

that

$$MISE(h) = V_1 n^{-1} - V_2 h n^{-1} - B_3 h^4 + Ch^2/n + \text{smaller order terms}$$

where  $V_1 = D_1(F)$ ,  $V_2 = 2A_1(K)D_2(F)$ ,  $B_3 = 0.25[A_2(K)]^2 D_3(F)$  and

$$A_1(K) = \int xk(x)K(x)dx,$$

$$A_2(K) = \int x^2k(x)dx,$$

$$D_1(F) = \int F(x)[1 - F(x)]f(x)W(x)dx,$$

$$D_2(F) = \int [f(x)]^2 W(x)dx,$$

$$D_3(F) = \int [f'(x)]^2 f(x)W(x)dx$$

The asymptotically optimal bandwidth is  $h_{opt} = (0.25V_2/B_3)^{1/3}n^{-1/3}$ .  $h_{opt}$  can be estimated either using cross-validation or using a plug-in estimator (Altman & Léger (1995)). In our simulations, we compared the FDP( $t$ ) estimation using multiple bandwidth selection methods in the R package sROC including unbiased cross-validation Scott (2015), Silverman's rule of thumb (Silverman 1986) (this is what's been used in simulations in Section 5.2.1 and 4.1.2), Sheather and Jones Sheather & Jones (1991). The differences are minimal.

Table 4.5. Mean (S.D.) of  $\widehat{FDP}(t) - FDP(t)$ , expressed in percentage, comparing different bandwidth selection methods

	Silverman	UCV	Sheather& Jones	BCV
CDF Transformation				
n=50	0.8 (7.7)	-2.7 (8.7)	-0.7(6.9)	1.7 (7.4)
n=100	0.1 (7.6)	-1.0 (6.6)	-2.7 (6.5)	1.0 (6.2)
Power Transformation				
n=50	1.9 (6.9)	-0.2 (6.6)	-0.1 (6.6)	-0.1 (6.6)
n=100	1.1 (5.4)	-0.27 (4.88)	-0.1 (4.8)	-0.1 (4.8)

#### 4.1.4 Robustness

To explore the robustness of NPN-PFA when the underlying distributions deviate from nonparanormal. We investigate the robustness of our method vs direct application of PFA under the simulation settings similar to what Fan et al. (2017) has used except that both factors and the idiosyncratic errors are heavy-tailed.

We chose  $p = 500$ ,  $p_1 = 25$ , and  $n = 100$ , with both  $\mathbf{f}$  and the idiosyncratic errors coming from *iid*  $t$  distribution with degrees of freedom = 3, 4, 5, 6; *iid* gamma with shape = 7.5 and rate = 1; *iid* log-normal with  $a(\exp(0.5 + 0.5Z) - b)$ , where  $Z \sim N(0, 1)$  with  $a$  and  $b$  chosen such that the distribution has mean 0 and variance 1.

As you can see from Tables 4.6 - 4.7 and Figures 4.5 - 4.6, applying principal factor approximation (PFA) directly to these heavy tailed data produced dramatic underestimation of the true FDP value. The results are more desirable when nonparanormal transformation was applied on these data before the use of PFA.

Table 4.6. Mean (S.D.) of  $\widehat{FDP}(t) - FDP(t)$ , expressed in percentage, using PFA directly vs Nonparanormal PFA.

Critical value (0.05)	PFA	NPN-PFA
$t$ -distribution ( $\mu_1 = 0.5$ )	-4.7(12.5)	-0.11(9.65)
$t$ -distribution ( $\mu_1 = 1$ )	-2.3(10.2)	-0.02(8.7)
gamma-distribution ( $\mu_1 = 0.5$ )	-2.9(10.6)	-0.07(8.3)
gamma-distribution ( $\mu_1 = 1$ )	-0.55(6.7)	1.4(5.8)
lognormal-distribution ( $\mu_1 = 0.5$ )	-6.4(9.1)	1.5(9)
lognormal-distribution ( $\mu_1 = 1$ )	-2.7(9.7)	3.7(9.6)

Table 4.7. Mean (S.D.) of  $\widehat{FDP}(t) - FDP(t)$ , expressed in percentage, using PFA directly vs Nonparanormal PFA for  $t$ -distribution with  $\mu_1 = 0.5$ .

Critical value (0.05)	PFA	NPN-PFA
$t$ -distribution (df=3)	-4.7(12.5)	-0.11(9.65)
$t$ -distribution (df=4)	-5.2(10.2)	1.6(9.1)
$t$ -distribution (df=5)	-4(10.5)	3.1(9.6)
$t$ -distribution (df=6)	-3(8.4)	3.9(8.8)

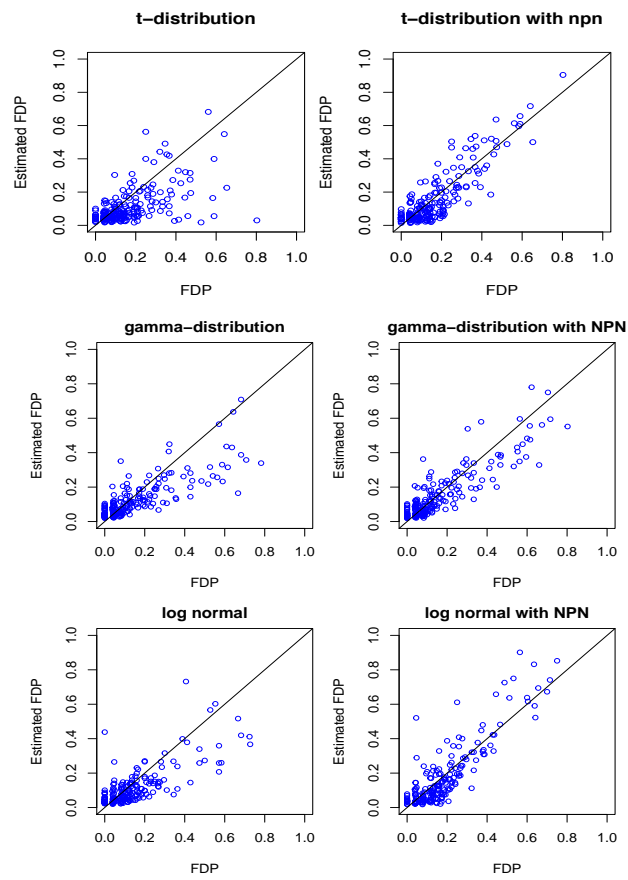


Figure 4.5. Comparing PFA alone vs NPN-PFA with  $\mu_1 = 0.5$ .

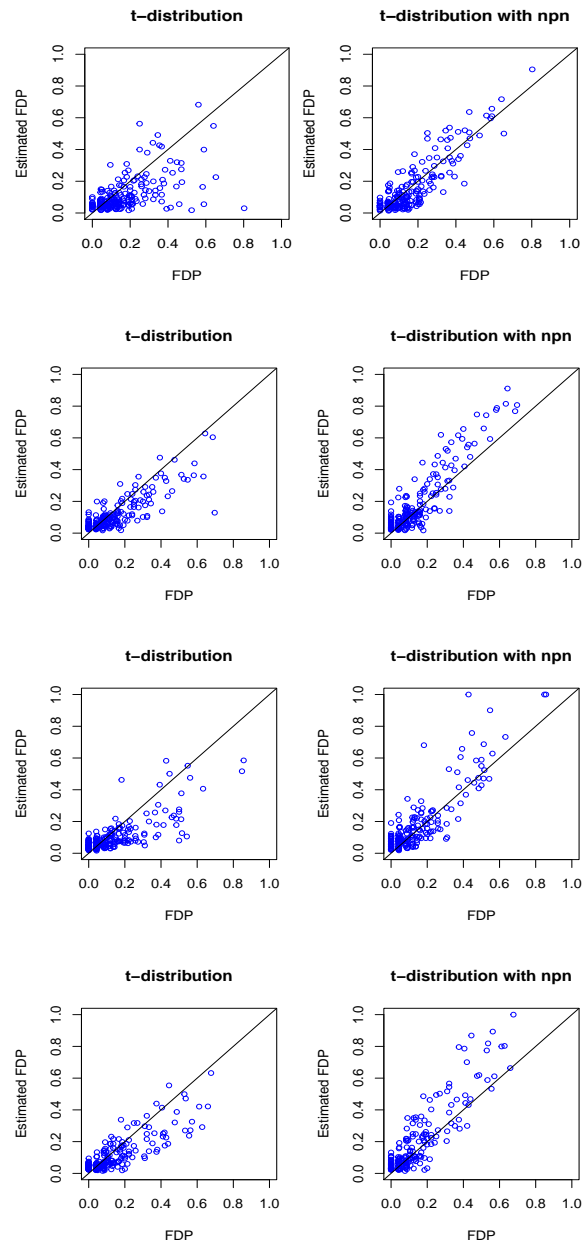


Figure 4.6. Comparing PFA alone vs NPN-PFA with  $\mu_1 = 0.5$ , with various degrees of freedom for  $t$ -distribution: from left to right:  $df=3, 4$  (top);  $df=5, 6$  (bottom).

#### 4.1.5 Other Transformation Function

Box-Cox is a power transformation that has been widely used to transform non-normal data to normal distribution. There is a restriction in using the Box-Cox transformation in the present context, which is that it requires the data to be positive. In order to make this transformation work, we have added to each of the observations an appropriately chosen constant determined from the range of all observations before applying the Box-Cox transformation. Then the data is adjusted through location-shift/re-scaling using the original sample mean and standard deviation.

We use the following two sets of data to see how the Box-Cox transformation performs compared to NPN-transformation in estimating  $FDP(t)$ , for some fixed rejection threshold  $t$ :

- 1) We generated NPN data using the 3-factor model as in Section 5.2.1, with  $p = 1000$  and  $p_1 = 40$ ,  $\mu_j = 1$  for  $j \in (1, 40)$  and  $\mu_j = 0$  otherwise, and the rejection threshold  $t = 0.01$ . For parameters used in CDF and power transformations to generate the nonparanormal data, please refer to Section 4 in our manuscript.
- 2) Since data with heavy tails are commonly seen in many scientific fields, we have also generated a similar 3 factor model as in 1), but with the factors and idiosyncratic errors being generated from gamma or log-normal distribution, with  $p = 500$  and  $p_1 = 25$ ,  $\mu_j = 0.5$  for  $j \in (1, 25)$  and  $\mu_j = 0$  otherwise, and the rejection threshold  $t = 0.01$ .

Simulation results are shown in Table 4.8 and Figure 4.7. As you can see,  $FDP(t)$  is universally underestimated with Box-Cox transformation whereas the estimated  $FDP(t)$  using kernel smoothing is reasonably close to the actual  $FDP(t)$ .

Table 4.8. Mean (S.D.) of  $\widehat{\text{FDP}}(t) - \text{FDP}(t)$ , expressed in percentage, using Box-Cox vs nonparanormal transformation

	CDF Transformation	Power Transformation	Gamma Distribution	Log-normal Distribution
Box-Cox	-2.1(8.7)	-2.3(7.1)	-3.6(8.8)	-4.4(7.8)
Nonparanormal	0.6(7.7)	2.2(6.7)	-0.07(8.3)	1.5(9.0)

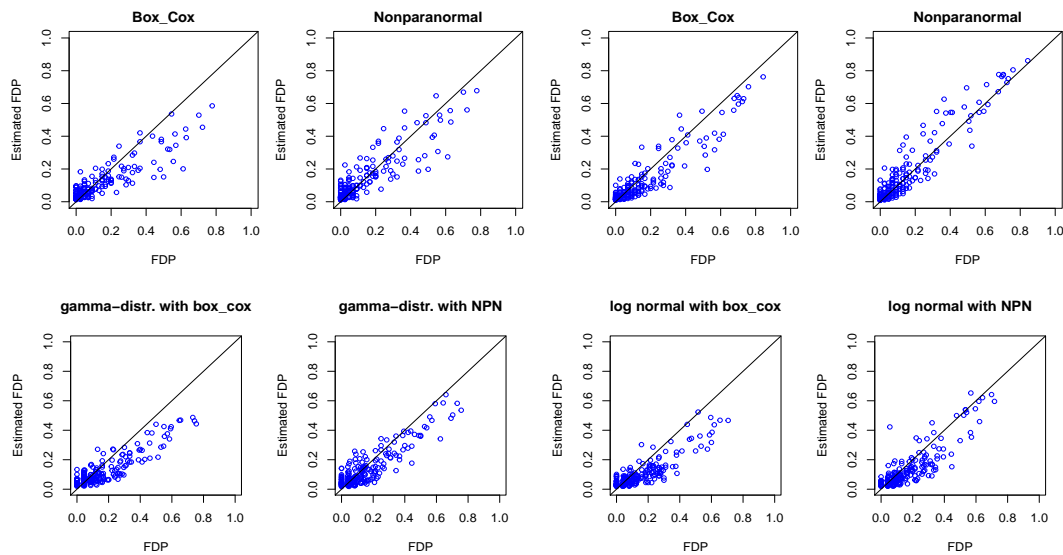


Figure 4.7. Box-Cox vs nonparanormal transformation. Top left: CDF transformation; Top right: power transformation; Bottom left: gamma-distribution; Bottom right: log-normal distribution.

#### 4.1.6 Dependence Adjusted Testing procedure

Recall Table 1.2, false negative rate is defined as  $\text{FNR} = E[T/(p - R)]$  where  $T$  is the number of falsely accepted null hypotheses. We compare the dependence adjusted procedure described in 3.4 with the fixed threshold procedure that uses non-adjusted  $p$ -values. With dependence-adjusted  $p$ -values, we fix rejection region  $t = 0.001$  and reject the hypotheses when the  $p$ -value is smaller than 0.001. Then we find the corresponding threshold value for the fixed threshold procedure such that the FDR in the two testing procedures are approximately the same. We performed simulation using model 1 & 2 in 4.1.2 for these are clear factor-model structure. Our simulation results (Table 4.9) suggested that the FNR for the dependence-adjusted procedure is consistently smaller than that of the fixed threshold procedure, which suggests that dependence-adjusted procedure is more powerful. Note that the population of signals was increased to  $p_1 = 200$  from  $p_1 = 40$ , suggesting that the dependence-adjusted procedure performs well even under non-sparse situation. When the covariance matrix is known, this dependence adjusted procedure has even more robust effect (Fan et al. 2012).

Table 4.9. Comparison of dependence-adjusted procedure with fixed threshold procedure under strict factor model and approximate factor model. The nonzero  $\mu_i$  are simulated from  $U(0.1, 0.5)$  and  $p_1 = 200$ .

	Fixed Threshold Procedure			Dependence Adjusted Procedure		
	FDR	FNR	Threshold	FDR	FNR	Threshold
Model 1						
CDF Transformation						
n=50	10.07%	14.39%	0.004	10.03%	9.89%	0.001
n=100	4.93%	9.16%	0.0052	4.92%	5.38%	0.001
Power Transformation						
n=50	8.82%	12.09%	0.002	8.96%	6.54%	0.001
n=100	4.75%	9.80%	0.0018	4.78%	4.59%	0.001
Model 2						
CDF Transformation						
n=50	10.79%	14.90%	0.0012	10.70%	10.20%	0.001
n=100	5.01%	9.82%	0.0014	4.99%	5.46%	0.001
Power Transformation						
n=50	9.54%	12.49%	0.0016	9.38%	7.11%	0.001
n=100	4.64%	9.50%	0.0022	4.68%	5.07%	0.001

## 4.2 Real Data Analysis

### 4.2.1 Prostate Cancer Data

We apply our multiple testing procedure to the prostate cancer data that was tested for normality in the beginning of this paper. This dataset has been analyzed extensively in previous publications using various multiple testing procedures Efron (2007), Efron (2010b), and Efron (2016). 52 prostate cancer patients of various stages and 50 healthy controls were included in the microarray experiment. To compare the genetic profile between these two groups of subjects, 6033 genes expression levels were measured. Let  $\mathbf{X}_1, \dots, \mathbf{X}_p$  denote the microarray of the healthy group with sample size  $n = 50$ , and  $\mathbf{Y}_1, \dots, \mathbf{Y}_p$  denote that of the prostate cancer group with sample size  $m = 52$ . Both  $\mathbf{X}_i$  and  $\mathbf{Y}_i$  are  $p$ -dimensional column vectors. The goal is to identify differentially expressed genes between two groups of subjects. This will enable scientists to identify cases of prostate cancer on the basis of gene-expression profiles.

Assuming  $X_j$  and  $Y_j$  are from two nonparanormal distribution with different mean vector but the same covariance matrix:  $X_j \sim NPN_d(\mu_{xj}, g_x, \Sigma)$  and  $Y_j \sim NPN_d(\mu_{yj}, g_y, \Sigma)$ . Multiple hypothesis testing is constructed to test:

$$H_{0j} : \mu_{xj} = \mu_{yj} \quad \text{vs} \quad H_j : \mu_{xj} \neq \mu_{yj} \quad \text{for } j = 1, \dots, p.$$

Suppose  $V_j \sim N(\mu_{xj}, \Sigma_0)$  and  $W_j \sim N(\mu_{yj}, \Sigma_0)$  are the underlying multivariate normal distribution of  $X_j$  and  $Y_j$  respectively. Consider the test statistics  $\mathbf{Z} = \sqrt{\frac{nm}{n+m}}(\bar{\mathbf{V}} - \bar{\mathbf{W}})/\sigma$ , and  $Z_j \sim N(\mu_{zj}, \Sigma)/\sigma$  with  $\mu_{zj} = \sqrt{\frac{nm}{n+m}}(\mu_{xj} - \mu_{yj})/\sigma$ . The original multiple hypothesis testing is now transformed to test

$$H_{0j} : \mu_{zj} = 0 \quad \text{vs} \quad H_j : \mu_{zj} \neq 0 \quad \text{for } j = 1, \dots, p.$$

We applied our kernel-mSCM method to this dataset using a grid of threshold value  $t$ . The estimated number of factors is 7 based on eigen-ratio method. With the current smaller sample size, this estimation can deviate from the actual number of factors. We also included results for  $k = 1, 3,$  and  $5$ . Estimated FDP and estimated number of false rejections were plotted against total number of rejections in Figure 4.8. We can see that the estimated FDP is close to zero when the number of rejections is below 125, which is 2% of the total 6033 genes. When the number of rejections increases, the estimated FDP also increases. Compared to the estimated results in Efron (2010b), where the data was considered low powered, and only 49 genes were considered differentiated, our method is able to identify more differentiated genes. These results suggest that the accuracy of true discoveries among the rejected hypothesis is high using our method for the rejection range studied. In addition, smaller estimated FDP values are obtained for higher number of factors for the kernel-mSCM method. We also compared the top 10 most significantly differentiated genes with the results in Efron (2010b) in Table 4.10. Only 2 of the 10 genes are overlapping. The signals were strengthened using NPN-PFA method as seen the larger absolute value of z-scores compared to that of Efron’s method.

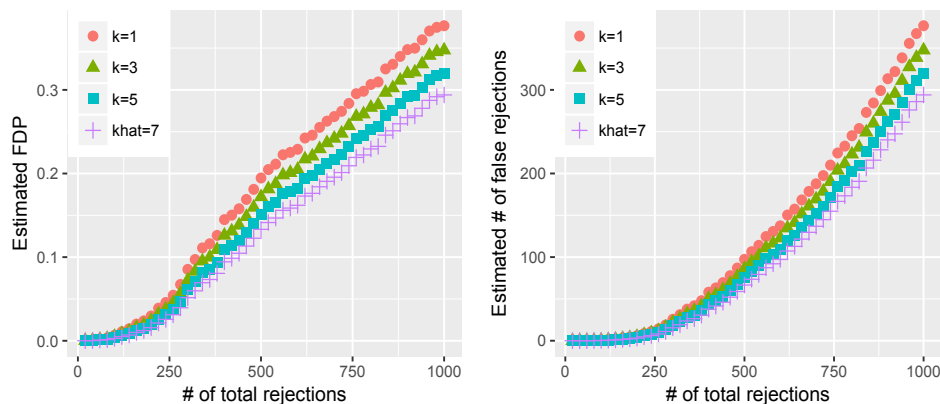


Figure 4.8. The estimated FDP and false discoveries as functions of the number of total discoveries for  $p=6033$  genes, with the number of factors being 1, 3, 5 or 7. mSCM was used to estimate the correlation matrix.

Table 4.10. Top 10 genes selected from the prostate cancer data: comparing NPN-PFA vs Efron’s method.

Efron(2010)		NPN-PFA	
Gene ID	z-score	Gene ID	z-score
610	5.29	735	6.40
1720	4.83	739	6.21
332	4.47	694	5.98
364	-4.42	698	5.94
914	4.4	452	-5.78
3940	-4.33	292	-5.71
4546	-4.29	298	5.63
1068	4.25	721	-5.50
579	4.19	610	-5.46
4331	-4.14	3940	5.38

To confirm that the non-paranormal transformation achieves desired normality without impacting the independence across observations, we re-run the normality test as in Table 1.1 and also calculated the Pearson correlation and Spearman’s rank correlation coefficient across observations for all 6033 genes in the prostate cancer data after transformation. Results are shown in Table 4.11 and Table 4.12.

Table 4.11. Testing normality of prostate cancer data after Nonparanormal transformation

Critical value (0.05)	Shapiro Wilk	Shapiro-Francia	Anderson-Darling	Pearson
Unadjusted	774(12.8%)	1164(19.3%)	1521(25.2%)	1505(25%)
With Holm’s correction	5024(83.3%)	5393(89.4%)	4589(76.1%)	2905(48.2%)

Recall that in Table 1 of our manuscript, the percentage of genes that didn’t have enough evidence of non-normal distribution is less than 5% before adjustment and between 13% and 39% after Holm’s correction. After transformation, these percentages increased significantly to 12%-25% before correction and 48% to 89.4% after Holm’s correction.

The small mean and median of both Pearson correlation coefficient and Spearman’s correlation coefficient indicated that the samples are near-independent.

Table 4.12. Testing sample independence in the prostate cancer data after Nonparanormal transformation

Critical value (0.05)	Mean	SD	Median
Pearson correlation	0.093	0.13	0.074
Spearman's rank	0.086	0.13	0.067

#### 4.2.2 Breast Cancer Data

We also apply our multiple testing procedure to a well-known breast cancer study used in Hedenfalk & et al (2001) and Efron (2007). 15 women were included in the study with 7 of them carrying a BRCA1 mutation, 8 of them carrying BRCA2 mutation, both known as the best-known genes linked to breast cancer risk. To compare the genetic profile between these two groups of women, 3226 genes expression levels were measured and the sample size is 7 and 8 for two groups, respectively.

Using the same notation as in Section 4.2.1, we applied our kernel-mSCM method to this dataset using a grid of threshold value  $t$ . The estimated number of factors from the dataset is 1. Results for  $k = 2, 3, 4$ , and 5 are also included for comparison. Estimated FDP and estimated number of false rejections were plotted against total number of rejections in Figure 4.9. We can see that the estimated FDP is close to zero when the number of rejections is below 500, which is 15% of the total 3226 genes. When the number of rejections increases, the estimated FDP also increases. These results suggest that the accuracy of true discoveries among the rejected hypothesis is high using our method for the rejection range studied. In addition, smaller estimated FDP values are obtained for higher number of factors for the NPN-PFA method.

30 most differentially expressed genes are listed in Table 4.13 using the kernel-POET method with 1 factor.

Table 4.13. 30 most differentially expressed genes that can discriminate breast cancers with BRCA1 mutations from those with BRCA2 mutations.

Clone ID	UniGene Title
3	phosphofructokinase, platelet
21	ARP1 (actin-related protein 1, yeast) homolog A (centractin alpha)
273	cold shock domain protein A
438	phosphofructokinase, platelet
557	SELENOPHOSPHATE SYNTHETASE ; Human selenium donor protein
600	D123 gene product
608	Human mRNA for ornithine decarboxylase antizyme, ORF 1 and ORF 2
816	zinc finger protein, subfamily 1A, 1 (Ikaros)
852	nuclear matrix protein p84
872	nuclease sensitive element binding protein 1
879	lysophospholipase II
892	EphB4
971	chromobox homolog 3 (Drosophila HP1 gamma)
1315	glutathione peroxidase 4 (phospholipid hydroperoxidase)
1404	nucleoporin 155kD
1426	ESTs, Weakly similar to putative [C.elegans]
1447	ESTs, Highly similar to CGI-26 protein [H.sapiens]
1489	nuclear receptor co-repressor 1
1591	ESTs
1641	ESTs, Moderately similar to atypical PKC specific binding protein
1921	GDP dissociation inhibitor 2
1942	polymyositis/scleroderma autoantigen 1 (75kD)
1984	ESTs
2133	UDP-galactose transporter related
2286	v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1
2319	O-linked N-acetylglucosamine (GlcNAc) transferase
2335	gamma-aminobutyric acid (GABA) A receptor, pi
2352	general transcription factor II, i, pseudogene 1
2353	tumor protein p53-binding protein, 2
2380	phytanoyl-CoA hydroxylase (Refsum disease)

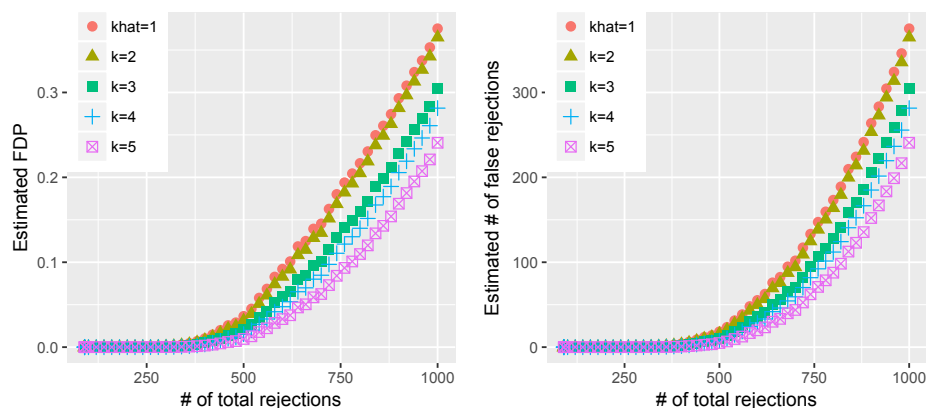


Figure 4.9. The estimated false discovery proportion and the estimated number of false discoveries as functions of the number of total discoveries for  $p = 3226$  genes, with the number of factors being 1, 2, 3, 4 or 5. Modified SCM was used to estimate the correlation matrix.



## CHAPTER 5

### CONCLUSION AND FUTURE DIRECTION

#### 5.1 Summary and Conclusion

In this thesis, we have proposed a new algorithm for high dimensional hypothesis testing under dependence which extends a factor based testing procedure to non-Gaussian distributions, specifically, nonparanormal population. Our motivation of this study comes from the common limitation of the recently developed factor based multiple testing methods: the assumption of normality which is often violated in reality. The normality test of one of the frequently used cancer microarray dataset proved our reasoning. The selection of nonparanormal population was based on its unique connection with Gaussian population through monotone marginal transformation functions. By keeping the marginal mean and variance of the NPN unchanged, the strength of signals being tested remains the same as well.

In Chapter three, we gave definitions and notations for nonparanormal distribution, algorithm to estimate the marginal transformation function and correlation matrix, expressions of FDP based on estimated test statistics, and theoretical investigation, simulations and real dataset. We proposed to use kernel distribution estimation for estimating the marginal transformation function. Simulation revealed that NPN-PFA method outperforms the *Winsorized*<sup>2</sup> estimator which is based on an empirical CDF estimator. We also proposed a simplified sample correlation estimator which outperforms the rank-based estimators and is computationally more efficient than thresholding estimator POET. We provide theoretical evidence that our estimator provides consistent estimations for both correlation matrix and false discoveries under some conditions. Furthermore, NPN+PFA is robust when the data deviates from nonparanormal distribution.

NPN+PFA was also applied to some real data. For the prostate cancer data, with total of 6033 genes, we were able to reject about 250 genes while maintaining the estimated FDP under 5%. Comparing with the previous results in Efron (2010*b*), our method is more powerful.

## 5.2 Ongoing and Future Work

### 5.2.1 Three-way data hypothesis testing

Among the recent data explosion in the past decade, more complex structure are seen and majority of them are three-way data. Observations are collected from various units in different settings or locations. For example, longitudinal data of multiple variables at a series of time points or spatio-temporal data existed in a format of layers of matrices. The challenge is to capture the dependence which are coming from random matrices instead of conventional random univariate or multivariate variables. For example, in genomics, gene expression data of  $p$  genes collected over  $q$  different tissues from the same subject are often correlated. Many studies have shown that matrix normal distribution is a feasible tool to model matrix variate random variables (Finn (1974), Timm (1980), Teng & Huang (2009), Efron (2009)).

Consider a genomics study comparing gene expression between treatment and control group with  $p$  genes collected over  $q$  different tissues from the same subject. Let  $\mathbf{Y}_1, \dots, \mathbf{Y}_n$  i.i.d. from matrix normal distribution  $\mathcal{MN}(\boldsymbol{\mu}_y, \mathbf{U}, \mathbf{V})$  be the sample data for the treatment group, where  $\mathbf{U} \in \mathbb{R}^{p \times p}$  and  $\mathbf{V} \in \mathbb{R}^{q \times q}$ , and  $\mathbf{Z}_1, \dots, \mathbf{Z}_m$  i.i.d. from matrix normal distribution  $\mathcal{MN}(\boldsymbol{\mu}_z, \mathbf{U}, \mathbf{V})$  be the sample data for the control group. Let  $\bar{\mathbf{Y}}_{i,j} = n^{-1} \sum_{l=1}^n \mathbf{Y}_l^{(i,j)}$  and  $\bar{\mathbf{Z}}_{i,j} = m^{-1} \sum_{l=1}^m \mathbf{Z}_l^{(i,j)}$ , where  $\mathbf{Y}_l^{(i,j)}$  and  $\mathbf{Z}_l^{(i,j)}$  are the  $(i, j)$ th element of  $\mathbf{Y}_l$  and  $\mathbf{Z}_l$  respectively. Suppose standard deviation of the  $(i, j)$ th element in  $\mathbf{Y}_l$  or  $\mathbf{Z}_l$  is  $\sigma_{i,j}$ , then we estimate  $\sigma_{i,j}^2$  by  $\hat{\sigma}_{i,j}^2 = (n + m - 2)^{-1} \{ \sum_{l=1}^n (\mathbf{Y}_l^{(i,j)} - \bar{\mathbf{Y}}_{i,j})^2 + \sum_{k=1}^m (\mathbf{Z}_k^{(i,j)} - \bar{\mathbf{Z}}_{i,j})^2 \}$ . Let  $\boldsymbol{\Sigma} = (\sigma_{i,j}^{-1})$  and  $\hat{\boldsymbol{\Sigma}} = (\hat{\sigma}_{i,j}^{-1})$ . Let  $\tilde{\mathbf{Y}}_l = \mathbf{Y}_l \circ \hat{\boldsymbol{\Sigma}}$  for  $l = 1, \dots, n$  and  $\tilde{\mathbf{Z}}_k = \mathbf{Z}_k \circ \hat{\boldsymbol{\Sigma}}$  for  $k = 1, \dots, m$  where  $\circ$  is the Hadamard product. Let  $\tilde{\mathbf{Y}}^* = n^{-1} \sum_{l=1}^n \tilde{\mathbf{Y}}_l$  and  $\tilde{\mathbf{Z}}^* = m^{-1} \sum_{k=1}^m \tilde{\mathbf{Z}}_k$ . Consider

$\mathbf{X} = \sqrt{\frac{nm}{n+m}}(\tilde{\mathbf{Y}}^* - \tilde{\mathbf{Z}}^*)$ , then approximately  $\mathbf{X} \sim \mathcal{MN}(\sqrt{\frac{nm}{n+m}}(\boldsymbol{\mu}_y - \boldsymbol{\mu}_z) \circ \boldsymbol{\Sigma}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ , where  $\boldsymbol{\Sigma}_1$  and  $\boldsymbol{\Sigma}_2$  are the correlation matrices of  $\mathbf{U}$  and  $\mathbf{V}$  respectively. Here we propose two methods to model the dependence and estimate the FDP.

### Method 1

If we vectorize the matrix data, then we have  $\text{vec}(\mathbf{X}) \sim \mathcal{N}(\sqrt{\frac{nm}{n+m}}\text{vec}((\boldsymbol{\mu}_y - \boldsymbol{\mu}_z) \circ \boldsymbol{\Sigma}), \boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1)$ , where  $\otimes$  denotes the Kronecker product. Let  $\lambda_1, \dots, \lambda_p$  be the non-increasing eigenvalues of  $\boldsymbol{\Sigma}_1$ , and  $\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_p$  be the corresponding eigenvectors. Let  $\xi_1, \dots, \xi_q$  be the non-increasing eigenvalues of  $\boldsymbol{\Sigma}_2$ , and  $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_q$  be the corresponding eigenvectors. Then the eigenvalues of  $\boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1$  are  $\xi_j \times \lambda_i$  for  $1 \leq i \leq p, 1 \leq j \leq q$  and the corresponding eigenvectors are  $\boldsymbol{\gamma}_j \otimes \boldsymbol{\nu}_i$ . We will estimate  $\boldsymbol{\Sigma}_1$  by  $\hat{\boldsymbol{\Sigma}}_1 = (n+m-2)^{-1}q^{-1}\{\sum_{l=1}^n(\tilde{\mathbf{Y}}_l - \tilde{\mathbf{Y}}^*)(\tilde{\mathbf{Y}}_l - \tilde{\mathbf{Y}}^*)^T + \sum_{k=1}^m(\tilde{\mathbf{Z}}_k - \tilde{\mathbf{Z}}^*)(\tilde{\mathbf{Z}}_k - \tilde{\mathbf{Z}}^*)^T\}$  and estimate  $\boldsymbol{\Sigma}_2$  by  $\hat{\boldsymbol{\Sigma}}_2 = (n+m-2)^{-1}p^{-1}\{\sum_{l=1}^n(\tilde{\mathbf{Y}}_l - \tilde{\mathbf{Y}}^*)^T(\tilde{\mathbf{Y}}_l - \tilde{\mathbf{Y}}^*) + \sum_{k=1}^m(\tilde{\mathbf{Z}}_k - \tilde{\mathbf{Z}}^*)^T(\tilde{\mathbf{Z}}_k - \tilde{\mathbf{Z}}^*)\}$ . Let  $\hat{\lambda}_1, \dots, \hat{\lambda}_p$  be the eigenvalues of  $\hat{\boldsymbol{\Sigma}}_1$ , and  $\hat{\boldsymbol{\nu}}_1, \dots, \hat{\boldsymbol{\nu}}_p$  be the corresponding eigenvectors. Let  $\hat{\xi}_1, \dots, \hat{\xi}_q$  be the eigenvalues of  $\hat{\boldsymbol{\Sigma}}_2$ , and  $\hat{\boldsymbol{\gamma}}_1, \dots, \hat{\boldsymbol{\gamma}}_q$  be the corresponding eigenvectors. Therefore, we can consider  $\mathbf{X}$  as the test statistics and apply PFA in Fan & Han (2017) based on  $\{\hat{\lambda}_i\}$ ,  $\{\hat{\boldsymbol{\nu}}_i\}$ ,  $\{\hat{\xi}_i\}$  and  $\{\hat{\boldsymbol{\gamma}}_i\}$ .

Note that we want to test  $H_{0,i,j} : ((\boldsymbol{\mu}_y - \boldsymbol{\mu}_z) \circ \boldsymbol{\Sigma})_{i,j} = 0$  vs  $H_{1,i,j} : ((\boldsymbol{\mu}_y - \boldsymbol{\mu}_z) \circ \boldsymbol{\Sigma})_{i,j} \neq 0$  for  $i = 1, \dots, p$  and  $j = 1, \dots, q$ . Suppose all  $\sigma_{i,j}$  are known, then the test statistics should be  $\tilde{\mathbf{X}} = \sqrt{\frac{nm}{n+m}}(n^{-1}\sum_{l=1}^n \mathbf{Y}_l \circ \boldsymbol{\Sigma} - m^{-1}\sum_{k=1}^m \mathbf{Z}_k \circ \boldsymbol{\Sigma})$ . Correspondingly, the P-value for the  $(i, j)$ th hypothesis is  $2\Phi(-|\tilde{\mathbf{X}}_{i,j}|)$ , where  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal distribution. Define  $R(t) = \#\{(i, j) : P_{i,j} \leq t\}$ ,  $V(t) = \#\{\text{true null} : P_{i,j} \leq t\}$ , we are interested in  $\text{FDP}(t) = V(t)/R(t)$  where the convention  $0/0 = 0$  is always used. We can approximate  $\text{FDP}(t)$  by the following procedure.

Approximate the  $P_{i,j}$  by  $\hat{P}_{i,j} = 2\Phi(-|\mathbf{X}_{i,j}|)$  and calculate  $\hat{R}(t) = \#\{(i, j) : \hat{P}_{i,j} \leq t\}$ . For the eigenvalues  $\{\hat{\lambda}_i\}$  and  $\{\hat{\xi}_j\}$ , we calculate the product of each possible pair and arrange these values in a non increasing order, written as  $\{\theta_l\}$ ,

correspondingly the eigenvectors are  $\{\boldsymbol{\rho}_l\}$ . For a given integer value  $h$ , define  $\mathbf{B} = (\sqrt{\theta_1}\boldsymbol{\rho}_1, \dots, \sqrt{\theta_h}\boldsymbol{\rho}_h)$ . Then we can approximate FDP( $t$ ) by

$$\widehat{\text{FDP}}(t) = \sum_{l=1}^{p \times q} [\Phi(a_l(z_{t/2} + \eta_l)) + \Phi(a_l(z_{t/2} - \eta_l))] / \widehat{R}(t)$$

where  $a_l = (1 - \|\mathbf{b}_l\|^2)^{-1/2}$ ,  $\eta_l = \mathbf{b}_l^T (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{X}$  and  $\mathbf{b}_l^T$  is the  $l$ th row of  $\mathbf{B}$ . To determine  $h$ , we can use the eigenvalue ratio estimator in Ahn & Horenstein (2013). The estimator is  $\widehat{h} = \operatorname{argmax}_{1 \leq l \leq l_{\max}} (\theta_l / \theta_{l+1})$ , where  $l_{\max}$  is the maximum possible number of factors.

## Method 2

If  $\mathbf{X} \sim \mathcal{MN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ , then we can rewrite  $\mathbf{X}$  as

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{C}\mathbf{W}\mathbf{D} + \boldsymbol{\epsilon} \quad (5.1)$$

where  $\mathbf{C} = (\sqrt{\lambda_1}\boldsymbol{\nu}_1, \dots, \sqrt{\lambda_{k_1}}\boldsymbol{\nu}_{k_1})$ ,  $\mathbf{D} = (\sqrt{\xi_1}\boldsymbol{\gamma}_1, \dots, \sqrt{\xi_{k_2}}\boldsymbol{\gamma}_{k_2})^T$ ,  $\mathbf{W} \sim \mathcal{MN}(\mathbf{0}, \mathbf{I}_{k_1}, \mathbf{I}_{k_2})$ , and  $\boldsymbol{\epsilon} \sim \mathcal{MN}(\mathbf{0}, \sum_{i=k_1+1}^p \lambda_i \boldsymbol{\nu}_i \boldsymbol{\nu}_i^T, \sum_{j=k_2+1}^q \xi_j \boldsymbol{\gamma}_j \boldsymbol{\gamma}_j^T)$ .

Note that by the properties of Kronecker product, the expression (5.1) is equivalent to

$$\operatorname{vec}(\mathbf{X}) = \operatorname{vec}(\boldsymbol{\mu}) + (\mathbf{D}^T \otimes \mathbf{C}) \operatorname{vec}(\mathbf{W}) + \operatorname{vec}(\boldsymbol{\epsilon}) \quad (5.2)$$

Based on model (5.2), we consider a least squares estimator for  $\mathbf{W}$ :

$$\begin{aligned} \operatorname{vec}(\widehat{\mathbf{W}}) &= [(\mathbf{D}^T \otimes \mathbf{C})^T (\mathbf{D}^T \otimes \mathbf{C})]^{-1} (\mathbf{D}^T \otimes \mathbf{C})^T \operatorname{vec}(\mathbf{X}) \\ &= [(\mathbf{D}\mathbf{D}^T) \otimes (\mathbf{C}^T \mathbf{C})]^{-1} (\mathbf{D} \otimes \mathbf{C}^T) \operatorname{vec}(\mathbf{X}) \\ &= [\operatorname{diag}(\xi_1^{-1}, \dots, \xi_{k_2}^{-1}) \otimes \operatorname{diag}(\lambda_1^{-1}, \dots, \lambda_{k_1}^{-1})] (\mathbf{D}^T \otimes \mathbf{C})^T \operatorname{vec}(\mathbf{X}). \end{aligned}$$

Correspondingly,

$$(\mathbf{D}^T \otimes \mathbf{C}) \operatorname{vec}(\widehat{\mathbf{W}}) = [(\xi_1^{-1/2} \boldsymbol{\gamma}_1, \dots, \xi_{k_2}^{-1/2} \boldsymbol{\gamma}_{k_2}) \otimes (\lambda_1^{-1/2} \boldsymbol{\nu}_1, \dots, \lambda_{k_1}^{-1/2} \boldsymbol{\nu}_{k_1})]$$

$$\begin{aligned}
& \times [(\xi_1^{1/2} \boldsymbol{\gamma}_1, \dots, \xi_{k_2}^{1/2} \boldsymbol{\gamma}_{k_2})^T \otimes (\lambda_1^{1/2} \boldsymbol{\nu}_1, \dots, \lambda_{k_1}^{1/2} \boldsymbol{\nu}_{k_1})^T] \text{vec}(\mathbf{X}) \\
& = [(\sum_{i=1}^{k_2} \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^T) \otimes (\sum_{j=1}^{k_1} \boldsymbol{\nu}_j \boldsymbol{\nu}_j^T)] \text{vec}(\mathbf{X})
\end{aligned}$$

We will consider a FDP approximation formula

$$\widehat{\text{FDP}}(t) = \sum_{l=1}^{p \times q} [\Phi(a_l(z_{t/2} + \eta_l)) + \Phi(a_l(z_{t/2} - \eta_l))] / \widehat{R}(t)$$

where  $a_l = (1 - \|\mathbf{b}_l\|^2)^{-1/2}$ ,  $\mathbf{b}_l$  is the  $l$ th row of  $\widehat{\mathbf{D}}^T \otimes \widehat{\mathbf{C}}$ , and  $\eta_l$  is the  $l$ th element of  $[(\sum_{i=1}^{k_2} \widehat{\boldsymbol{\gamma}}_i \widehat{\boldsymbol{\gamma}}_i^T) \otimes (\sum_{j=1}^{k_1} \widehat{\boldsymbol{\nu}}_j \widehat{\boldsymbol{\nu}}_j^T)] \text{vec}(\mathbf{X})$ .

To determine  $k_1$  and  $k_2$ , we consider the eigenvalue ratio estimator:

$$\widehat{k}_1 = \operatorname{argmax}_{1 \leq l \leq l_{\max}} (\widehat{\lambda}_l / \widehat{\lambda}_{l+1}), \quad \widehat{k}_2 = \operatorname{argmax}_{1 \leq l \leq l_{\max}} (\widehat{\xi}_l / \widehat{\xi}_{l+1})$$

**Simulation Results** We perform simulations according to the following settings. To simulate matrix normal data, we created  $\Sigma_1$  and  $\Sigma_2$  according to the method in , setting the number of factors to be (3 and 3) or (2 and 4) for  $\Sigma_1$  and  $\Sigma_2$  respectively. For simplicity, the diagonal elements are set to be 1. We set sample size  $n = m = 50$  or 100 for both groups, dimension of  $\Sigma_1$  and  $\Sigma_2$  to be  $p = 40$  and  $q = 100$  respectively for both groups, the first 8 rows and first 25 columns for treatment group to contain signals with  $\mu_1 = 1$  and the remaining elements to have  $\mu_1 = 0$ , for control group  $\mu_1 = 0$ . The rejection region was set to be  $t = 0.01$ . The number of independent simulations is 200 for each setting. Table 5.1 contains the mean and SD of  $\widehat{\text{FDP}}(t) - \text{FDP}(t)$ , expressed in percentage. All the mean errors are within  $\pm 1\%$ . Figure 5.1 compares the estimated FDP vs true FDP. PFA performed consistently in all settings in estimating the false discovery proportions.

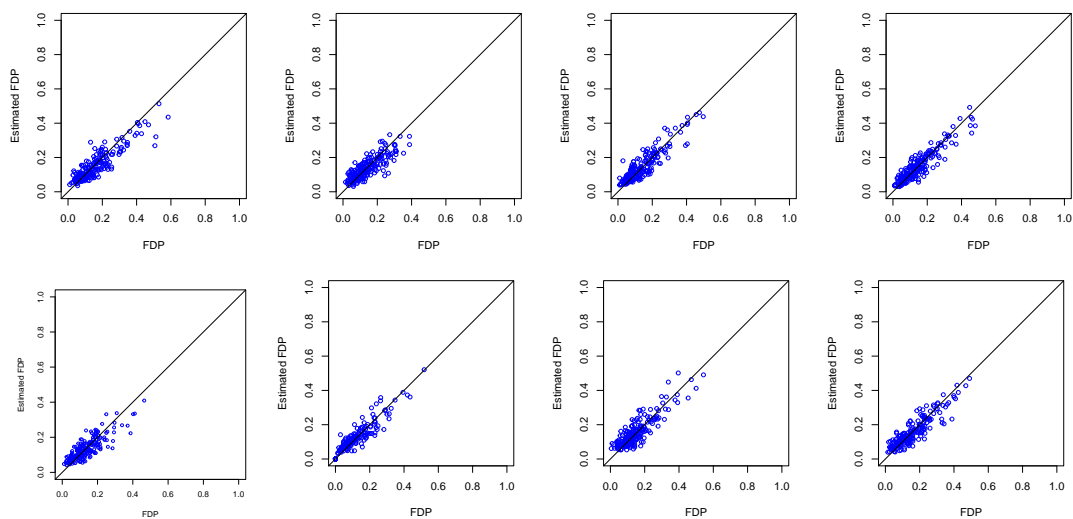


Figure 5.1. Comparison of  $\widehat{FDP}(t)$  with  $FDP(t)$  for matrix normal data. The order match the order of the table. Top Panel: Method 1; Bottom Panel: Method 2; First column:  $n=50$  with factors = (2,4); Second column:  $n=50$  with factors = (3,3); Third column:  $n=100$  with factors = (2,4); Forth column:  $n=100$  with factors = (3,3)

Table 5.1. Mean (S.D.) of  $\widehat{FDP}(t) - FDP(t)$ , expressed in percentage, using PFA in matrix normal data.

Settings	n=50		n=100	
	factors=2,4	factors=3,3	factors=2,4	factors=3,3
Method-1 two groups	0.9(4.9)	0.3(4.6)	-0.66(4.1)	-0.19(3.8)
Method-2 two groups	0.18(4.4)	-0.51(4.5)	-0.6(4.5)	0.15(4.1)

## Bibliography

- AghaKouchak, A. (2014), ‘Entropy–copula in hydrology and climatology’, *Journal of Hydrometeorology* **15**(6), 2176–2189.
- Ahn, S. C. & Horenstein, A. R. (2013), ‘Eigenvalue ratio test for the number of factors’, *Econometrica* **81**(3), 1203–1227.
- Altman, N. & Léger, C. (1995), ‘Bandwidth selection for kernel distribution function estimation’, *Journal of Statistical Planning and Inference* **46**(2), 195–214.
- Azzalini, A. (1981), ‘A note on the estimation of a distribution function and quantiles by a kernel method’, *Biometrika* pp. 326–328.
- Benjamini, Y. & Bogomolov, M. (2014), ‘Selective inference on multiple families of hypotheses’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(1), 297–318.
- Benjamini, Y. & Heller, R. (2007), ‘False discovery rates for spatial signals’, *Journal of the American Statistical Association* **102**(480), 1272–1281.
- Benjamini, Y. & Hochberg, Y. (1995), ‘Controlling the false discovery rate: a practical and powerful approach to multiple testing’, *Journal of the royal statistical society. Series B (Methodological)* pp. 289–300.
- Benjamini, Y. & Hochberg, Y. (2000), ‘On the adaptive control of the false discovery rate in multiple testing with independent statistics’, *Journal of educational and Behavioral Statistics* **25**(1), 60–83.
- Benjamini, Y., Krieger, A. M. & Yekutieli, D. (2006), ‘Adaptive linear step-up procedures that control the false discovery rate’, *Biometrika* pp. 491–507.

- Benjamini, Y. & Liu, W. (1999), ‘A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence’, *Journal of Statistical Planning and Inference* **82**(1), 163–170.
- Benjamini, Y. & Yekutieli, D. (2001), ‘The control of the false discovery rate in multiple testing under dependency’, *The Annals of Statistics* **29**, 1165–1188.
- Berkes, P., Wood, F. & Pillow, J. W. (2009), Characterizing neural dependencies with copula models, *in* ‘Advances in neural information processing systems’, pp. 129–136.
- Chacon, J. & Rodriguez-Casal, A. (2010), ‘A note on the universal consistency of the kernel distribution function estimator’, *Statistics and Probability Letters* **17**, 1414–1419.
- Chen, X. & Fan, Y. (2006), ‘Estimation of copula-based semiparametric time series models’, *Journal of Econometrics* **130**(2), 307–335.
- Cheng, M. & Peng, L. (2002), ‘Regression modeling for nonparametric estimation of distribution and quantile functions’, *Statist. Sinica* **12**, 10431060.
- Clarke, S. & Hall, P. (2009), ‘Robustness of multiple testing procedures against dependence’, *Ann. Statist.* **37**, 332358.
- Clements, N., Sarkar, S. K. & Guo, W. (2011), ‘Astronomical transient detection using grouped p-values and controlling the false discovery rate’, *Unpublished manuscript*.
- Clements, N., Sarkar, S. K., Zhao, Z. & Kim, D.-Y. (2014), ‘Applying multiple testing procedures to detect change in east african vegetation’, *The Annals of Applied Statistics* **8**(1), 286–308.
- Davis, C. & Kahan, W. M. (1970), ‘The rotation of eigenvectors by a perturbation. iii’, *SIAM Journal on Numerical Analysis* **7**(1), 1–46.

- Desai, K. H. & Storey, J. D. (2012), ‘Cross-dimensional inference of dependent high-dimensional data’, *Journal of the American Statistical Association* **107**(497), 135–151.
- Efron, B. (2007), ‘Correlation and large-scale simultaneous significance testing’, *Journal of the American Statistical Association* **102**(477), 93–103.
- Efron, B. (2009), ‘Are a set of microarrays independent of each other?’, *The annals of applied statistics* **3**(3), 922.
- Efron, B. (2010a), ‘Correlated z-values and the accuracy of large-scale statistical estimates’, *Journal of the American Statistical Association* **105**(491), 1042–1055.
- Efron, B. (2010b), ‘The future of indirect evidence’, *Statistical science: a review journal of the Institute of Mathematical Statistics* **25**(2), 145.
- Efron, B. (2016), ‘Empirical bayes deconvolution estimates’, *Biometrika* **103**(1), 1–20.
- Falk, M. (1983), ‘Relative efficiency and deficiency of kernel type estimators of smooth distribution functions’, *Journal of the Royal Statistical Society Series B* **37**, 7383.
- Fan, J. & Han, X. (2017), ‘Estimation of the false discovery proportion with unknown dependence’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* .
- Fan, J., Han, X. & Gu, W. (2012), ‘Estimating false discovery proportion under arbitrary covariance dependence’, *Journal of the American Statistical Association* **107**(499), 1019–1035.
- Fan, J., Ke, Y., Sun, Q. & Zhou, W. (2017), ‘Farm-test: Factor-adjusted robust multiple testing with false discovery control’, *arXiv preprint arXiv:1711.05386* .
- Fan, J., Liao, Y. & Mincheva, M. (2013), ‘Large covariance estimation by thresholding principal orthogonal complements’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(4), 603–680.

- Fan, J. & Lv, J. (2008), ‘Sure independence screening for ultrahigh dimensional feature space’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5), 849–911.
- Finn, J. D. (1974), *A general model for multivariate analysis.*, Holt, Rinehart & Winston.
- Friguet, C., Kloareg, M. & Causeur, D. (2009), ‘A factor model approach to multiple testing under dependence’, *Journal of the American Statistical Association* **104**(488), 1406–1415.
- Gavrilov, Y., Benjamini, Y. & Sarkar, S. K. (2009), ‘An adaptive step-down procedure with proven fdr control under independence’, *The Annals of Statistics* pp. 619–629.
- Hedenfalk, I. & et al, D. C. (2001), ‘Gene-expression profiles in hereditary breast cancer’, *New England Journal of Medicine* **344**, 539–548.
- Heller, R., Chatterjee, N., Krieger, A. & Shi, J. (2017), ‘Post-selection inference following aggregate level hypothesis testing in large scale genomic data’, *bioRxiv* p. 058404.
- Higham, N. (1988), ‘Computing a nearest symmetric positive semidefinite matrix’, *Journal of the Royal Statistical Society Series B* **103**, 103–118.
- Hochberg, Y. & Benjamini, Y. (1990), ‘More powerful procedures for multiple significance testing’, *Statistics in medicine* **9**(7), 811–818.
- Horn & Johnson (1990), ‘Matrix analysis’, *Cambridge University Press* .
- Hu, J. X., Zhao, H. & Zhou, H. H. (2010), ‘False discovery rate control with groups’, *Journal of the American Statistical Association* **105**(491), 1215–1227.
- Ince, R. A., Giordano, B. L., Kayser, C., Rousselet, G. A., Gross, J. & Schyns, P. G. (2017), ‘A statistical framework for neuroimaging data analysis based on mutual

- information estimated via a gaussian copula', *Human brain mapping* **38**(3), 1541–1573.
- Kendall, M. (1948), 'Rank correlation methods', *Griffin* pp. 139–140.
- Kendall, M. (1990), 'The performance of kernel density functions in kernel distribution function estimation', *Statistics & Probability Letters* **9**, 129132.
- Klaassen, C. A., Wellner, J. A. et al. (1997), 'Efficient estimation in the bivariate normal copula model: normal margins are least favourable', *Bernoulli* **3**(1), 55–77.
- Kruskal, W. (1958), 'Ordinal measures of association.', *Journal of the American Statistical Association* **53**, 814–861.
- Leek, J. & Storey, J. (2008), 'A general framework for multiple testing dependence', *Proceedings of the National Academy of Sciences of the United States of America* **105**, 18718–18723.
- Liu, H., Han, F. & Yuan, M. (2012), 'High-dimensional semiparametric gaussian copula graphical models', *Ann Stat* **40**, 2293–326.
- Liu, H., Lafferty, J. & L.Wasserman (2009), 'The nonparanormal: Semiparametric estimation of high dimensional undirected graphs', *Journal of Machine Learning Research* **10**(Oct), 2295–2328.
- Liu, Y., Sarkar, S. K. & Zhao, Z. (2016), 'A new approach to multiple testing of grouped hypotheses', *Journal of Statistical Planning and Inference* **179**, 1–14.
- Malevergne, Y., Sornette, D. et al. (2003), 'Testing the gaussian copula hypothesis for financial assets dependences', *Quantitative Finance* **3**(4), 231–250.
- Parzen, E. (1962), 'On estimation of a probability density function and mode', *The Annals of Mathematical Statistics* **33**(3), 1065–1076.
- Reiss, R. (1981), 'Nonparametric estimation of smooth distribution functions', *Scandinavian Journal of Statistics* **8**, 116119.

- Renard, B. & Lang, M. (2007), 'Use of a gaussian copula for multivariate extreme value analysis: some case studies in hydrology', *Advances in Water Resources* **30**(4), 897–912.
- Rey, M. & Roth, V. (2012), Meta-gaussian information bottleneck, in 'Advances in Neural Information Processing Systems', pp. 1916–1924.
- Sarda, P. (1993), 'Smoothing parameter selection for smooth distribution functions', *Journal of Statistical Planning and Inference* **35**(1), 65–75.
- Sarkar, S. (2002), 'Some results on false discovery rate in stepwise multiple testing procedures', *The Annals of Statistics* **30**, 239–257.
- Sarkar, S. K. (2007), 'Stepup procedures controlling generalized fwer and generalized fdr', *The Annals of Statistics* pp. 2405–2420.
- Sarkar, S. K. (2008a), 'On methods controlling the false discovery rate', *Sankhyā: The Indian Journal of Statistics, Series A (2008-)* pp. 135–168.
- Sarkar, S. K. (2008b), 'Two-stage stepup procedures controlling fdr', *Journal of Statistical Planning and Inference* **138**(4), 1072–1084.
- Sarkar, S. K. & Guo, W. (2009), 'On a generalized false discovery rate', *The Annals of Statistics* pp. 1545–1565.
- Sarkar, S. K. & Guo, W. (2010), 'Procedures controlling the k-fdr using bivariate distributions of the null p-values', *Statistica Sinica* pp. 1227–1238.
- Schwartzman, A. & Lin, X. (2011), 'The effect of correlation in false discovery rate estimation', *Biometrika* **98**(1), 199–214.
- Schweder, T. & Spjøtvoll, E. (1982), 'Plots of p-values to evaluate many tests simultaneously', *Biometrika* **69**(3), 493–502.
- Scott, D. W. (2015), *Multivariate density estimation: theory, practice, and visualization*, John Wiley & Sons.

- Sheather, S. J. & Jones, M. C. (1991), 'A reliable data-based bandwidth selection method for kernel density estimation', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 683–690.
- Shirahata, S. & Chu, I.-S. (1988), 'Mean integrated squared error properties and optimal kernels when estimating a distribution function', *Communications in Statistics Theory and Methods* **17**, 37853799.
- Silverman, B. (1986), 'Density estimation london', *UK: Chapman and Hall* .
- Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P. et al. (2002), 'Gene expression correlates of clinical prostate cancer behavior', *Cancer cell* **1**(2), 203–209.
- Sklar, M. (1959), 'Fonctions de repartition an dimensions et leurs marges', *Publ. inst. statist. univ. Paris* **8**, 229–231.
- Storey, J. D. (2002), 'A direct approach to false discovery rates', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(3), 479–498.
- Storey, J. D. (2003), 'The positive false discovery rate: a bayesian interpretation and the q-value', *Annals of statistics* pp. 2013–2035.
- Storey, J. D. (2004), 'Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach', *The Journal of the Royal Statistical Society: Series B* **66**, 187205.
- Sun, W. & Tony Cai, T. (2009), 'Large-scale multiple testing under dependence', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(2), 393–424.
- Szabó, Z., Póczos, B., Szirtes, G. & Lőrincz, A. (2007), 'Post nonlinear independent subspace analysis', *Artificial Neural Networks–ICANN 2007* pp. 677–686.

- Teng, S. L. & Huang, H. (2009), ‘A statistical framework to infer functional gene relationships from biologically interrelated microarray experiments’, *Journal of the American Statistical Association* **104**(486), 465–473.
- Timm, N. H. (1980), ‘2 multivariate analysis of variance of repeated measurements’, *Handbook of statistics* **1**, 41–87.
- Tsukahara, H. (2005), ‘Semiparametric estimation in copula models’, *Canadian Journal of Statistics* **33**(3), 357–375.
- Wang, W. Y. & Hua, Z. (2014), A semiparametric gaussian copula regression model for predicting financial risks from earnings calls., *in* ‘ACL (1)’, pp. 1155–1165.
- Xue, L. & Zou, H. (2012), ‘Regularized rank-based estimation of high-dimensional nonparanormal graphical models’, *The Annals of Statistics* **40**(5), 2541–2571.
- Yekutieli, D. (2008), ‘Hierarchical false discovery rate–controlling methodology’, *Journal of the American Statistical Association* **103**(481), 309–316.

## APPENDIX

## APPENDIX A

### THEORETICAL PROOF

#### Estimating Correlation Matrices

Liu et al. (2012) has exploit Spearman's rho and Kendall's tau statistics to directly estimate the correlation matrices  $\Sigma_0$ , without calculating the marginal transformation functions  $\{f_j\}_{j=1}^d$ . The details of these two statistics are as follow. Let  $r_j^i$  be the rank of  $x_j^i$  among  $x_j^1, \dots, x_j^n$  and  $\bar{r}_j = \frac{1}{n} \sum_{i=1}^n r_j^i = \frac{n+1}{2}$ .

$$\begin{aligned} \text{(Spearman's rho)} \quad \hat{\rho}_{jk} &= \frac{\sum_{i=1}^n (r_j^i - \bar{r}_j)(r_k^i - \bar{r}_k)}{\sqrt{\sum_{i=1}^n (r_j^i - \bar{r}_j)(r_k^i - \bar{r}_k)}}, \\ \text{(Kendall's tau)} \quad \hat{\tau}_{jk} &= \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \text{sign}((x_j^i - x_j^{i'})(x_k^i - x_k^{i'})). \end{aligned}$$

Both  $\hat{\rho}_{jk}$  and  $\hat{\tau}_{jk}$  are nonparametric correlations between the empirical realizations of random variables  $X_j$  and  $X_k$ . For nonparanormal distributions, the following lemma connects Spearman's rho and Kendall's tau to the underlying Pearson correlation coefficient  $\Sigma_{jk}^0$ .

**Lemma A.0.1** *Let  $\{\lambda_i\}_{i=1}^p$  be eigenvalues of  $\Sigma$  in non-increasing order and  $\{\gamma_i\}_{i=1}^p$  be their corresponding eigenvectors. Let  $\{\hat{\lambda}_i\}_{i=1}^p$  be eigenvalues of  $\hat{\Sigma}$  in non-increasing order and  $\{\hat{\gamma}_i\}_{i=1}^p$  be their corresponding eigenvectors.*

1. (sin  $\theta$  Theorem, Davis & Kahan (1970))  $|\hat{\lambda}_i - \lambda_i| \leq \|\hat{\Sigma} - \Sigma\|,$
2. (Weyl's Theorem, Horn & Johnson (1990))
 
$$\|\hat{\gamma}_i - \gamma_i\| \leq \frac{\sqrt{2}\|\hat{\Sigma} - \Sigma\|}{\min(|\hat{\lambda}_{i-1} - \lambda_i|, |\lambda_i - \hat{\lambda}_{i+1}|)}.$$

### Proof of Theorem 1

To prove the convergence of the estimate of correlation matrix, suppose  $S_{jk}$  is the  $jk$ th element of the estimated correlation matrix,

$$\widehat{S}_{jk} = \frac{1}{n} \sum_{i=1}^n \Phi^{-1}(\widehat{F}_j(X_{ij})) \Phi^{-1}(\widehat{F}_k(X_{ik})).$$

We first use the following term:

$$\begin{aligned} & |\widehat{S}_{jk} - S_{jk}| \\ &= \left| \frac{1}{n} \sum_{i=1}^n \Phi^{-1}(\widehat{F}_j(X_{ij})) \Phi^{-1}(\widehat{F}_k(X_{ik})) - \frac{1}{n} \sum_{i=1}^n \Phi^{-1}(F_j(X_{ij})) \Phi^{-1}(F_k(X_{ik})) \right| \\ &= \frac{1}{n} \left| \sum_{i=1}^n [\Phi^{-1}(\widehat{F}_j(X_{ij})) [\Phi^{-1}(\widehat{F}_k(X_{ik})) - \Phi^{-1}(F_k(X_{ik}))] \right. \\ &\quad \left. + \Phi^{-1}(F_k(X_{ik})) [\Phi^{-1}(\widehat{F}_j(X_{ij})) - \Phi^{-1}(F_j(X_{ij}))]] \right| \\ &\leq \frac{1}{n} \left| \sum_{i=1}^n [\Phi^{-1}(\widehat{F}_j(X_{ij})) [\Phi^{-1}(\widehat{F}_k(X_{ik})) - \Phi^{-1}(F_k(X_{ik}))]] \right| \\ &\quad + \frac{1}{n} \left| \sum_{i=1}^n [\Phi^{-1}(F_k(X_{ik})) [\Phi^{-1}(\widehat{F}_j(X_{ij})) - \Phi^{-1}(F_j(X_{ij}))]] \right| \\ &= \frac{1}{n} \left| \sum_{i=1}^n [\Phi^{-1}(\widehat{F}_j(X_{ij})) (\Phi^{-1}(\widehat{F}_k^*(X_{ik})))' [\widehat{F}_k(X_{ik}) - F_k(X_{ik})]] \right| \\ &\quad + \frac{1}{n} \left| \sum_{i=1}^n [\Phi^{-1}(F_k(X_{ik})) (\Phi^{-1}(\widehat{F}_j^*(X_{ij})))' [\widehat{F}_j(X_{ij}) - F_j(X_{ij})]] \right| \\ &= L_1 + L_2. \end{aligned}$$

The second to last equality is due to mean value theory.

Previously, Chacon & Rodriguez-Casal (2010) have shown that kernel CDF has the following properties when  $h = h_0$  (optimal bandwidth), for every  $F \in \mathcal{F}$  and every  $\varepsilon > 0$  there exists  $n_0 = n_0(\varepsilon, F)$  such that

$$P(\|F_{nhn} - F\| > \varepsilon) \leq 2e^{(-n\varepsilon^2/2)}, \forall n \geq n_0,$$

where  $\|F_{nh} - F\| = \sup_{x \in \mathcal{R}} |F_{nh}(x) - F(x)|$  is the uniform distance.

Since the sample data are from a bounded domain,  $\Phi^{-1}(\widehat{F}_j(X_{ij}))$  and  $(\Phi^{-1}(\widehat{F}_k^*(X_{ik})))'$  are bounded by a constant  $d$ . Hence, we can conclude that for sufficient large  $n$ ,

$$\begin{aligned}
P(L_1 > \varepsilon) &= P\left(\frac{1}{n} \left| \sum_{i=1}^n [\Phi^{-1}(\widehat{F}_j(X_{ij}))(\Phi^{-1}(\widehat{F}_k^*(X_{ik})))]' [\widehat{F}_k(X_{ik}) - F_k(X_{ik})] \right| > \varepsilon\right) \\
&\leq P\left(\frac{1}{n} \sum_{i=1}^n |[\Phi^{-1}(\widehat{F}_j(X_{ij}))(\Phi^{-1}(\widehat{F}_k^*(X_{ik})))]'| |\widehat{F}_k(X_{ik}) - F_k(X_{ik})| > \varepsilon\right) \\
&\leq P\left(\frac{1}{n} \sum_{i=1}^n |\widehat{F}_k(X_{ik}) - F_k(X_{ik})| > \frac{\varepsilon}{d}\right) \\
&= P(\|\widehat{F}_k - F_k\| > \frac{\varepsilon}{d}) \leq 2e^{(-n\varepsilon^2/2d^2)}.
\end{aligned}$$

Convergence of  $L_2$  is similar as above. On the other hand,

$$\begin{aligned}
\|\widehat{\Sigma} - \Sigma\|_{op} &\leq \|\widehat{\Sigma} - \Sigma\|_1 = \max_k \sum_{j=1}^p |\widehat{S}_{jk} - \Sigma_{jk}| \\
&\leq \max_k \sum_{j=1}^p |\widehat{S}_{jk} - S_{jk}| + \max_k \sum_{j=1}^p |S_{jk} - \Sigma_{jk}|.
\end{aligned}$$

For the first term we have

$$\begin{aligned}
&P\left(\max_k \sum_{j=1}^p |\widehat{S}_{jk} - S_{jk}| > \varepsilon\right) \\
&\leq P\left(dp \max_k \|\widehat{F}_k - F_k\| > \frac{\varepsilon}{2}\right) + P\left(d \max_k \sum_{j=1}^p \|\widehat{F}_j - F_j\| > \frac{\varepsilon}{2}\right) \\
&\leq pP\left(dp \|\widehat{F}_k - F_k\| > \frac{\varepsilon}{2}\right) + P\left(d \sum_{j=1}^p \|\widehat{F}_j - F_j\| > \frac{\varepsilon}{2}\right) \\
&\leq pP\left(\|\widehat{F}_k - F_k\| > \frac{\varepsilon}{2dp}\right) + pP\left(\|\widehat{F}_j - F_j\| > \frac{\varepsilon}{2dp}\right) \\
&\leq 4p \exp\left(-\frac{n\varepsilon^2}{2d^2p^2}\right).
\end{aligned}$$

For the second term we have

$$\begin{aligned}
& P\left(\max_k \sum_{j=1}^p |S_{jk} - \Sigma_{jk}| > \varepsilon\right) \\
&= P\left(\max_k \sum_{j=1}^p \left| \frac{1}{n} \sum_{i=1}^n \Phi^{-1}(F_j(X_{ij}))\Phi^{-1}(F_k(X_{ik})) - E[\Phi^{-1}(F_j(X_{ij}))\Phi^{-1}(F_k(X_{ik}))]] \right| > \varepsilon\right) \\
&\leq pP\left(\sum_{j=1}^p \left| \frac{1}{n} \sum_{i=1}^n \Phi^{-1}(F_j(X_{ij}))\Phi^{-1}(F_k(X_{ik})) - E[\Phi^{-1}(F_j(X_{ij}))\Phi^{-1}(F_k(X_{ik}))]] \right| > \varepsilon\right) \\
&\leq p^2P\left(\left| \frac{1}{n} \sum_{i=1}^n \Phi^{-1}(F_j(X_{ij}))\Phi^{-1}(F_k(X_{ik})) - E[\Phi^{-1}(F_j(X_{ij}))\Phi^{-1}(F_k(X_{ik}))]] \right| > \varepsilon/p\right) \\
&\leq p^2 \exp\left(-\frac{2n\varepsilon^2}{p^2}\right).
\end{aligned}$$

In the last step, we have used the Hoeffding inequality. We have also used the fact that the sample data are on the bounded domain.

Thus,  $\left\| \widehat{\Sigma} - \Sigma \right\|_{op} = O_p(pn^{-\delta})$  when  $\log p \leq \frac{1}{2}n^{1-2\delta}$ . The proof is now complete.

## Proof of Theorem 2

Let  $\widetilde{V}(t) = \sum_{j=1}^p [\Phi(\widehat{a}_j(z_{t/2} + \widetilde{\eta}_j)) + \Phi(\widehat{a}_j(z_{t/2} - \widetilde{\eta}_j))]$  where  $\widetilde{\eta}_j = \widehat{\mathbf{b}}_j^T \widetilde{\mathbf{W}}$  based on estimated samples in 3.5. Then we have

$$\widetilde{V}(t) - V(t) = \left(\widetilde{V}(t) - \widehat{V}_A(t)\right) + \left(\widehat{V}_A(t) - V_A(t)\right) + \left(V_A(t) - V(t)\right). \quad (\text{A.1})$$

Note that:

$$\begin{aligned}
\widehat{V}_A(t) &= \sum_{i=1}^p [\Phi(a_i(z_{t/2} + \mathbf{b}_i^T \widehat{\mathbf{W}})) + \Phi(a_i(z_{t/2} - \mathbf{b}_i^T \widehat{\mathbf{W}}))], \\
V_A(t) &= \sum_{i=1}^p [\Phi(a_i(z_{t/2} + \mathbf{b}_i^T \mathbf{W})) + \Phi(a_i(z_{t/2} - \mathbf{b}_i^T \mathbf{W}))]
\end{aligned}$$

Let

$$\begin{aligned}\Delta_1 &= \sum_{i=1}^p [\Phi(\widehat{a}_i(z_{t/2} + \widehat{\mathbf{b}}_i^T \widetilde{\mathbf{W}})) - \Phi(a_i(z_{t/2} + \mathbf{b}_i^T \widehat{\mathbf{W}}))], \\ \Delta_2 &= \sum_{i=1}^p [\Phi(\widehat{a}_i(z_{t/2} - \widehat{\mathbf{b}}_i^T \widetilde{\mathbf{W}})) - \Phi(a_i(z_{t/2} - \mathbf{b}_i^T \widehat{\mathbf{W}}))].\end{aligned}$$

Thus,  $\widetilde{V}(t) - \widehat{V}_A(t) = \Delta_1 + \Delta_2$ .

We first deal with the term  $\Delta_1$ . The term  $\Delta_2$  can be dealt with analogously and hence omitted. By the proof of Theorem 1 in Fan & Han (2017), for some constant  $c_1$  and  $c_2$  we have

$$|\Delta_1| \leq c_1 \left( \sum_{h=1}^k |\widetilde{\lambda}_h - \lambda_h| + \lambda_h \|\widetilde{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h\| \right) + c_2 \sum_{i=1}^p |\widehat{\mathbf{b}}_i^T \widetilde{\mathbf{W}} - \mathbf{b}_i^T \widehat{\mathbf{W}}|. \quad (\text{A.2})$$

By Condition 3 in Theorem 2, Lemma 1 in Appendix, and proof of Theorem 3 in Fan & Han (2017), it's not hard to see that  $\|\gamma_i - \widehat{\gamma}_i\| = O_p(p^{-\tau})$  and  $\sum_{h=1}^k |\widetilde{\lambda}_h - \lambda_h| = O_p(kpn^{-\delta})$  for some constant  $\tau > 0$  and  $0 < \delta < 1/2$ . This proves the convergence of the first item in (A.2).

For the second term in (A.2), we have

$$\begin{aligned}\sum_{i=1}^p |\widehat{\mathbf{b}}_i^T \widetilde{\mathbf{W}} - \mathbf{b}_i^T \widehat{\mathbf{W}}| &= \left[ 1 \dots 1 \right]_{1 \times p} |\widehat{\mathbf{b}}_i^T \widetilde{\mathbf{W}} - \mathbf{b}_i^T \widehat{\mathbf{W}}| \\ &= \left[ 1 \dots 1 \right]_{1 \times p} \left| \sum_{h=1}^k [\widehat{\boldsymbol{\gamma}}_h \widehat{\boldsymbol{\gamma}}_h^T \widetilde{\mathbf{Z}} - \boldsymbol{\gamma}_h \boldsymbol{\gamma}_h^T \mathbf{Z}] \right| \\ &= \left[ 1 \dots 1 \right]_{1 \times p} \left| \sum_{h=1}^k \widehat{\boldsymbol{\gamma}}_h \widehat{\boldsymbol{\gamma}}_h^T (\widetilde{\mathbf{Z}} - \mathbf{Z}) + \sum_{h=1}^k (\widehat{\boldsymbol{\gamma}}_h \widehat{\boldsymbol{\gamma}}_h^T - \boldsymbol{\gamma}_h \boldsymbol{\gamma}_h^T) \mathbf{Z} \right| \\ &\leq \left[ 1 \dots 1 \right]_{1 \times p} \left| \sum_{h=1}^k \widehat{\boldsymbol{\gamma}}_h \widehat{\boldsymbol{\gamma}}_h^T (\widetilde{\mathbf{Z}} - \mathbf{Z}) \right| + \left[ 1 \dots 1 \right]_{1 \times p} \left| \sum_{h=1}^k (\widehat{\boldsymbol{\gamma}}_h \widehat{\boldsymbol{\gamma}}_h^T - \boldsymbol{\gamma}_h \boldsymbol{\gamma}_h^T) \mathbf{Z} \right| \\ &\leq \sqrt{p} \left\| \sum_{h=1}^k \widehat{\boldsymbol{\gamma}}_h \widehat{\boldsymbol{\gamma}}_h^T \right\| \|\widetilde{\mathbf{Z}} - \mathbf{Z}\| + \sqrt{p} \left\| \sum_{h=1}^k [\widehat{\boldsymbol{\gamma}}_h \widehat{\boldsymbol{\gamma}}_h^T - \boldsymbol{\gamma}_h \boldsymbol{\gamma}_h^T] \right\| \|\mathbf{Z}\|\end{aligned}$$

$$\leq k\sqrt{p}\|\tilde{\mathbf{Z}} - \mathbf{Z}\| + \sqrt{p}\left\|\sum_{h=1}^k[\hat{\gamma}_h\hat{\gamma}_h^T - \gamma_h\gamma_h^T]\right\|\|\mathbf{Z}\|. \quad (\text{A.3})$$

The first factor in the second term in (A.3) can be shown as

$$\begin{aligned} \left\|\sum_{h=1}^k[\hat{\gamma}_h\hat{\gamma}_h^T - \gamma_h\gamma_h^T]\right\| &\leq \sum_{h=1}^k\|\hat{\gamma}_h\hat{\gamma}_h^T - \gamma_h\gamma_h^T\| \\ &= \sum_{h=1}^k\|\hat{\gamma}_h(\hat{\gamma}_h - \gamma_h)^T + (\hat{\gamma}_h - \gamma_h)\gamma_h^T\| \\ &\leq \sum_{h=1}^k\|\hat{\gamma}_h(\hat{\gamma}_h - \gamma_h)^T\| + \sum_{h=1}^k\|(\hat{\gamma}_h - \gamma_h)\gamma_h^T\| \\ &\leq 2\sum_{h=1}^k\|\hat{\gamma}_h - \gamma_h\| \\ &= O_p(kp^{-\tau}). \end{aligned}$$

For the right part of second term in (A.3), we also have,

$$\|\mathbf{Z}\|^2 \leq 2\|\boldsymbol{\mu}^*\|^2 + 2\sum_{i=1}^p\lambda_i\varepsilon_i^2 = O_p(\|\boldsymbol{\mu}^*\|^2 + p).$$

Thus

$$\sqrt{p}\left\|\sum_{h=1}^k[\hat{\gamma}_h\hat{\gamma}_h^T - \gamma_h\gamma_h^T]\right\|\|\mathbf{Z}\| = O_p(kp^{1/2-\tau}\|\boldsymbol{\mu}^*\| + kp^{1-\tau}). \quad (\text{A.4})$$

For the first factor in expression (A.3),

$$\begin{aligned} \|\tilde{\mathbf{Z}} - \mathbf{Z}\|^2 &= \sum_{j=1}^p\left[\frac{1}{\sqrt{n}\hat{\sigma}_j}\sum_{i=1}^n\tilde{z}_{ij} - \frac{1}{\sqrt{n}\sigma_j}\sum_{i=1}^nz_{ij}\right]^2 \\ &= \sum_{j=1}^p\left\{\frac{1}{\sqrt{n}\hat{\sigma}_j}\sum_{i=1}^n[\tilde{\mu}_j + \tilde{\sigma}_j\Phi^{-1}(\tilde{F}_j(X_{ij}))] - \frac{1}{\sqrt{n}\sigma_j}\sum_{i=1}^n[\mu_j + \sigma_j\Phi^{-1}(F_j(X_{ij}))]\right\}^2 \\ &= \sum_{j=1}^p\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{\tilde{\mu}_j}{\tilde{\sigma}_j} - \frac{\mu_j}{\sigma_j}\right) + \frac{1}{\sqrt{n}}\sum_{i=1}^n[\Phi^{-1}(\tilde{F}_j(X_{ij})) - \Phi^{-1}(F_j(X_{ij}))]\right\}^2 \end{aligned}$$

$$\leq 2 \sum_{j=1}^p n \left( \frac{\tilde{\mu}_j}{\tilde{\sigma}_j} - \frac{\mu_j}{\sigma_j} \right)^2 + 2 \sum_{j=1}^p \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n [\Phi^{-1}(\tilde{F}_j(X_{ij})) - \Phi^{-1}(F_j(X_{ij}))] \right)^2. \quad (\text{A.5})$$

We will work on the second term from above first by assessing the order of  $\frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \Phi^{-1}(\tilde{F}_j(X_{ij})) - \Phi^{-1}(F_j(X_{ij})) \right\}$ . To simplify the notation, without loss of generality, we ignore the index of  $j$ . Define the following indicator random variable:

$$U_{ik} = \mathbf{I}\left\{ \left| \frac{X_i - X_k}{h} \right| \geq 6 \right\}.$$

$$\begin{aligned} & \text{Var} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i)) \right\} \right] \\ &= \frac{1}{n} \sum_{i=1}^n \text{Var} \left[ \Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i)) \right] \\ &+ \frac{1}{n} \sum_{i=1}^n \sum_{\substack{k=1 \\ \{i \neq k\}}}^n \text{Cov} \left\{ [\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))], [\Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k))] \right\} \end{aligned} \quad (\text{A.6})$$

To evaluate the second term in the expression above, we can evaluate  $\text{Cov} \left\{ [\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))], [\Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k))] \right\}$  for  $i \neq k$ . By the law of total covariance, if  $X$ ,  $Y$ , and  $Z$  are random variables on the same probability space,

$$\text{Cov}(X, Y) = \mathbf{E} \left[ \text{Cov}(X, Y | Z) \right] + \text{Cov} \left[ \mathbf{E}(X | Z), \mathbf{E}(Y | Z) \right].$$

Thus,

$$\begin{aligned} & \text{Cov} \left\{ [\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))], [\Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k))] \right\} \\ &= \mathbf{E} \left[ \text{Cov} \left( [\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))], [\Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k))] | U_{ik} \right) \right] \\ &+ \text{Cov} \left[ \mathbf{E} \left( [\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))] | U_{ik} \right), \mathbf{E} \left( [\Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k))] | U_{ik} \right) \right] \end{aligned}$$

$$\begin{aligned}
&= \text{Cov} \left( [\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))], [\Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k))] | U_{ik} = 1 \right) P(U_{ik} = 1) \\
&+ \text{Cov} \left( [\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))], [\Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k))] | U_{ik} = 0 \right) P(U_{ik} = 0) \\
&+ \text{Cov} \left[ E \left( [\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))] | U_{ik} \right), E \left( [\Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k))] | U_{ik} \right) \right] \\
&= E \left( \text{Cov} \left[ [\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))], \right. \right. \\
&\quad \left. \left. [\Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k))] | X_i, X_k, U_{ik} = 1 \right] | U_{ik} = 1 \right) P(U_{ik} = 1) \\
&+ \text{Cov} \left( E \left[ [\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))] | X_i, X_k, U_{ik} = 1 \right], \right. \\
&\quad \left. E \left[ [\Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k))] | X_i, X_k, U_{ik} = 1 \right] \right) P(U_{ik} = 1) \\
&+ \text{Cov} \left( [\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))], [\Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k))] | U_{ik} = 0 \right) P(U_{ik} = 0) \\
&+ \text{Cov} \left[ E \left( [\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))] | U_{ik} \right), E \left( [\Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k))] | U_{ik} \right) \right] \\
&= I_1 + I_2 + I_3 + I_4
\end{aligned}$$

As defined in equation (3.2), we choose  $a = 3$  for the threshold for truncated kernel for simplicity. This proof can be easily extended to a general  $a$ . We have

$$\begin{aligned}
I_1 &= E \left[ \text{Cov} \left( [\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))], \right. \right. \\
&\quad \left. \left. [\Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k))] | X_i, X_k, U_{ik} = 1 \right) | U_{ik} = 1 \right] P(U_{ik} = 1) \\
&= 0 \tag{A.7}
\end{aligned}$$

Because

$$\Phi^{-1} \left[ P(X \leq X_i - 3h) + \frac{1}{n} \sum_{\{j: X_i - ah \leq X_j \leq X_i + ah\}}^n G\left(\frac{X_i - X_j}{h}\right) \right] - \Phi^{-1}(F(X_i))$$

is independent of

$$\Phi^{-1} \left[ P(X \leq X_k - 3h) + \frac{1}{n} \sum_{\{j: X_k - ah \leq X_j \leq X_k + ah\}} G\left(\frac{X_k - X_j}{h}\right) \right] - \Phi^{-1}(F(X_k))$$

when conditioned on  $X_i$ ,  $X_k$ , and  $U_{ik} = 1$ .

$$\begin{aligned} I_2 &= \text{Cov} \left( \mathbb{E} \left[ \Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i)) \mid X_i, X_k, U_{ik} = 1 \right], \right. \\ &\quad \left. \mathbb{E} \left[ \Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k)) \mid X_i, X_k, U_{ik} = 1 \right] \right) P(U_{ik} = 1) \\ &= \text{Cov} \left( \mathbb{E} \left[ \Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i)) \mid X_i \right], \right. \\ &\quad \left. \mathbb{E} \left[ \Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k)) \mid X_k \right] \right) P(U_{ik} = 1) \\ &= 0 \end{aligned} \tag{A.8}$$

This is due to the fact that  $\Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k))$  is independent of  $X_i$  and  $\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))$  is independent of  $X_k$  when  $U_{ik} = 1$ .

$$\begin{aligned} I_3 &= \text{Cov} \left( \left[ \Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i)) \right], \left[ \Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k)) \right] \mid U_{ik} = 0 \right) P(U_{ik} = 0) \\ &= \mathbb{E} \left( \left[ \Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i)) \right] \left[ \Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k)) \right] \mid U_{ik} = 0 \right) P(U_{ik} = 0) \\ &\quad - \mathbb{E} \left( \Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i)) \mid U_{ik} = 0 \right) \mathbb{E} \left( \Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k)) \mid U_{ik} = 0 \right) P(U_{ik} = 0) \\ &\leq \mathbb{E} \left( \left[ \Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i)) \right] \left[ \Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k)) \right] \mid U_{ik} = 0 \right) P(U_{ik} = 0) \\ &= \mathbb{E} \left[ \mathbb{E} \left( \left[ \Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i)) \right] \left[ \Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k)) \right] \mid X_i, X_k, U_{ik} = 0 \right) \right. \\ &\quad \left. \mid U_{ik} = 0 \right] P(U_{ik} = 0) \end{aligned}$$

By mean value theory, there exists a  $F^*(X_i) \in \left( \tilde{F}(X_i), F(X_i) \right)$  and a  $F^c(X_k) \in \left( \tilde{F}(X_k), F(X_k) \right)$  such that, the last term from above equals

$$\begin{aligned}
& \mathbb{E} \left[ \mathbb{E} \left( [\Phi^{-1}(\tilde{F}^*(X_i))]' [\tilde{F}(X_i) - F(X_i)] [\Phi^{-1}(\tilde{F}^c(X_k))]' [\tilde{F}(X_k) - F(X_k)] \mid X_i, X_k, U_{ik} = 0 \right) \right. \\
& \qquad \qquad \qquad \left. \mid U_{ik} = 0 \right] P(U_{ik} = 0) \\
& \leq \mathbb{E} \left[ [\Phi^{-1}(\tilde{F}^*(X_i))]' [\Phi^{-1}(\tilde{F}^c(X_k))]' \mathbb{E}([\tilde{F}(X_i) - F(X_i)][\tilde{F}(X_k) - F(X_k)] \mid X_i, X_k, U_{ik} = 0) \right. \\
& \qquad \qquad \qquad \left. \mid U_{ik} = 0 \right] P(U_{ik} = 0) \\
& \leq c \mathbb{E} \left[ \mathbb{E} \left( \left| [\tilde{F}(X_i) - F(X_i)][\tilde{F}(X_k) - F(X_k)] \right| \mid X_i, X_k, U_{ik} = 0 \right) \mid U_{ik} = 0 \right] P(U_{ik} = 0) \\
& \leq c \mathbb{E} \left[ \left\{ \mathbb{E}([\tilde{F}(X_i) - F(X_i)]^2 \mid X_i, X_k, U_{ik} = 0) \mathbb{E}([\tilde{F}(X_k) - F(X_k)]^2 \mid X_i, X_k, U_{ik} = 0) \right\}^{1/2} \mid U_{ik} = 0 \right] \times \\
& \qquad \qquad \qquad P(U_{ik} = 0) \\
& = c \mathbb{E} \left[ \mathbb{E}([\tilde{F}(X_i) - F(X_i)]^2 \mid X_i, X_k, U_{ik} = 0) \mid U_{ik} = 0 \right] P(U_{ik} = 0).
\end{aligned}$$

The first inequality is due to the fact that  $[\Phi^{-1}(\tilde{F}^*(X_i))]'$  and  $[\Phi^{-1}(\tilde{F}^c(X_k))]'$  are both positive and upper bounded since sample values  $X_i$  and  $X_k$  are bounded. We assume the upper bound of their product is  $c$ . The third inequality is based on Cauchy Schwartz inequality.

Since  $X_i$  and  $X_k$  are iid, let  $V = X_i - X_k$  so  $f_V$  is symmetrical around mean 0 and the derivative of  $f_V$  is 0 at  $v = 0$ . By Condition 2 in Theorem 2, and with the optimal  $h = c_0 n^{-1/3}$ , we have the following:

$$\begin{aligned}
P(U_{ik} = 0) &= P\left(\left|\frac{X_i - X_k}{h}\right| \geq 6\right) \\
&= P(-6h \leq V \leq 6h) \\
&= \int_{-6h}^{6h} f_V(v) dv
\end{aligned}$$

By Taylor expansion residual theory, we expand  $f_V(v)$  around the point 0, there exists a value  $\epsilon \in (0, v)$  such that the remainder term is  $R_n(v) = \frac{1}{2}f''(\epsilon)(v - 0)^2$ . Thus,

$$\begin{aligned} \int_{-6h}^{6h} f_V(v)dv &= \int_{-6h}^{6h} [f(0) + f'(0)(v - 0) + \frac{1}{2}f''(\epsilon)(v - 0)^2]dv \\ &= \int_{-6h}^{6h} f(0)dv + \frac{1}{6} \int_{-6h}^{6h} f''(\epsilon)v^2dv \\ &= O_p(n^{-1/3}). \end{aligned} \tag{A.9}$$

With Condition (1) in Theorem 2, we have,

$$\begin{aligned} &E\left[\tilde{F}(X_i)|X_i, X_k, U_{ik} = 0\right] \\ &= E\left[P(X \leq X_i - 3h) + \frac{1}{n} \sum_{\{j:j \neq k\}} G\left(\frac{X_i - X_j}{h}\right)\mathbf{I}(X_i - 3h \leq X_j \leq X_i + 3h) \right. \\ &\quad \left. + \frac{1}{n}G\left(\frac{X_i - X_k}{h}\right)|X_i, X_k, U_{ik} = 0\right] \\ &= F(X_i - 3h) + \frac{n-1}{n} \int_{X_i-3h}^{X_i+3h} G\left(\frac{X_i - Z}{h}\right)f(z)dz + \frac{1}{n}G\left(\frac{X_i - X_k}{h}\right) \quad \text{let } v = \frac{X_i - Z}{h} \\ &= F(X_i - 3h) + \frac{n-1}{n}h \int_{-3}^3 G(v)f(X_i - hv)dv + \frac{1}{n}G\left(\frac{X_i - X_k}{h}\right) \\ &= F(X_i - 3h) - \frac{n-1}{n}G(3)F(X_i - 3h) + \frac{n-1}{n}G(-3)F(X_i + 3h) \\ &\quad + \frac{n-1}{n} \int_{-3}^3 g(v)F(X_i - hv)dv + \frac{1}{n}G\left(\frac{X_i - X_k}{h}\right). \end{aligned}$$

By Taylor expansion residual theory, we will expand  $F(X_i - hv)$  around  $X_i$ . There exists a number  $c \in (X_i, X_i - hv)$  such that the remainder term  $R_n(X_i - hv) = \frac{f^{n+1}(c)}{(n+1)!}(-hv)^{n+1}$ , the above expression can be simplified as

$$\begin{aligned} &F(X_i - 3h) - \frac{n-1}{n}G(3)F(X_i - 3h) + \frac{n-1}{n}G(-3)F(X_i + 3h) + \frac{n-1}{n}G(3)F(X_i) \\ &= \frac{n-1}{n}F(X_i) + \frac{1}{n}F(X_i - 3h) + G(-3)\left(F(X_i + 3h) + F(X_i - 3h) - 2F(X_i)\right) \end{aligned}$$

$$\begin{aligned}
& + \frac{n-1}{n} \frac{h^2}{2} k_2 f'(c) + \frac{1}{n} G\left(\frac{X_i - X_k}{h}\right) \\
= & F(X_i) - \frac{3h}{n} f(X_i) + \frac{9h^2}{2n} f'(c) + \frac{n-1}{n} G(-3) 9h^2 f'(c^*) \\
& + \frac{n-1}{n} \frac{h^2}{2} k_2 f'(c) + \frac{1}{n} G\left(\frac{X_i - X_k}{h}\right). \tag{A.10}
\end{aligned}$$

Similarly,

$$\begin{aligned}
& \mathbb{E}\left[G^2\left(\frac{X_i - X_j}{h}\right) | X_i\right] \\
& = \int_{-3}^3 G^2\left(\frac{X_i - z}{h}\right) f(z) dz \quad \text{let } v = \frac{X_i - Z}{h} \\
& = h \int_{-3}^3 G^2(v) f(X_i - hv) dv \\
& = 3G^2(3)hf(X_i) - \frac{9}{2}G^2(3)h^2f'(c^*) \\
& \quad + 3G^2(-3)hf(X_i) + \frac{9}{2}G^2(-3)h^2f'(c^*) - 2hk_1f(X_i) + h^2k_3f'(c^*)
\end{aligned}$$

where  $k_1 = \int_{-\infty}^{\infty} G(v)g(v)v dv$ ,  $k_2 = \int_{-\infty}^{\infty} g(v)v^2 dv$ , and  $k_3 = \int_{-\infty}^{\infty} G(v)g(v)v^2 dv$ .

Thus, we have

$$\begin{aligned}
& \text{Var}\left[\tilde{F}(X_i) | X_i, X_k, U_{ik} = 0\right] \\
& = \text{Var}\left[P(X \leq X_i - 3h) + \frac{1}{n} \sum_{\{j:j \neq k\}} G\left(\frac{X_i - X_j}{h}\right) \mathbf{I}(X_i - 3h \leq X_j \leq X_i + 3h) \right. \\
& \quad \left. + \frac{1}{n} G\left(\frac{X_i - X_k}{h}\right) | X_i, X_k, U_{ik} = 0\right] \\
& = \frac{n-1}{n^2} \mathbb{E}\left[G^2\left(\frac{X_i - X_j}{h}\right) | X_i\right] - \frac{n-1}{n^2} \left(\mathbb{E}\left[G\left(\frac{X_i - X_j}{h}\right) | X_i\right]\right)^2 \\
& \leq \frac{1}{n} \mathbb{E}\left[G^2\left(\frac{X_i - X_j}{h}\right) | X_i\right]. \tag{A.11}
\end{aligned}$$

Combining results from (A.10) and (A.11), we have

$$\begin{aligned}
\mathbb{E} \left[ \left( \tilde{F}(X_i) - F(X_i) \right)^2 | X_i \right] &= \left[ \mathbb{E}[\tilde{F}(X_i) | X_i] - F(X_i) \right]^2 + \text{Var} \left[ \tilde{F}(X_i) | X_i \right] \\
&\leq \left[ \mathbb{E}[\tilde{F}(X_i) | X_i] - F(X_i) \right]^2 + \frac{1}{n} \mathbb{E} \left[ G^2 \left( \frac{X_i - X_j}{h} \right) | X_i \right] \\
&\leq \left\{ 9G(-3)h^2 f'(X_i) + \frac{h^2}{2} k_2 f'(c) + \frac{1}{n} G \left( \frac{X_i - X_k}{h} \right) - \frac{3h}{n} f(X_i) + \frac{9h^2}{2n} f'(c) \right\}^2 \\
&+ \frac{h}{n} \left[ 3 \left( G^2(3) + G^2(-3) \right) f(X_i) - 2k_1 f(X_i) - \frac{9}{2} G^2(3) h f'(c^*) + \frac{9}{2} G^2(-3) h f'(c^*) + h k_3 f'(c) \right].
\end{aligned}$$

With the Condition 1 in Theorem 2, and  $G(\frac{X_i - X_k}{h})$  is bounded on  $U_{ik} = 0$ , we have

$$\mathbb{E} \left[ \mathbb{E} \left( (\tilde{F}(X_i) - F(X_i))^2 | X_i, X_k, U_{ik} = 0 \right) | U_{ik} = 0 \right] = O(n^{-4/3}) \quad (\text{A.12})$$

when  $h = h_0 = cn^{-1/3}$ . With the result from (A.9), we have

$$I_3 = c \times \mathbb{E} \left[ \mathbb{E} \left( (\tilde{F}(X_i) - F(X_i))^2 | X_i, X_k, U_{ik} = 0 \right) \right] P(U_{ik} = 0) = O(n^{-5/3}). \quad (\text{A.13})$$

Next, we consider about  $I_4$ . By Taylor series,  $\Phi^{-1}(\tilde{F}(X_i))$  can be expanded towards the point

$$a = P(X \leq X_i - 3h) + \frac{1}{n} \sum_{\substack{j: X_i - ah \leq X_j \leq X_i + ah \\ j \neq k}}^n G \left( \frac{X_i - X_j}{h} \right).$$

Hence,

$$\begin{aligned}
&\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i)) \\
&= \Phi^{-1}(a) + \frac{1}{n} [\Phi^{-1}(a)]' G \left( \frac{X_i - X_k}{h} \right) - \Phi^{-1}(F(X_i)) \\
&= \Phi^{-1} \left[ P(X \leq X_i - 3h) + \frac{1}{n} \sum_{\substack{j: X_i - ah \leq X_j \leq X_i + ah \\ j \neq k}}^n G \left( \frac{X_i - X_j}{h} \right) \right] - \Phi^{-1}(F(X_i))
\end{aligned}$$

$$\begin{aligned}
& + \frac{1}{n} [\Phi^{-1}(a)]' G\left(\frac{X_i - X_k}{h}\right) \\
\equiv S_1 + S_2.
\end{aligned}$$

In the same way, expand  $\Phi^{-1}(\tilde{F}(X_i))$  towards the point

$$b = P(X \leq X_i - 3h) + \frac{1}{n} \sum_{\substack{j: X_i - ah \leq X_j \leq X_i + ah \\ j \neq i}}^n G\left(\frac{X_i - X_j}{h}\right),$$

we have,

$$\begin{aligned}
& \Phi^{-1}(\tilde{F}(X_k)) - \Phi^{-1}(F(X_k)) \\
= & \Phi^{-1}(b) + \frac{1}{n} [\Phi^{-1}(b)]' G\left(\frac{X_k - X_i}{h}\right) - \Phi^{-1}(F(X_i)) \\
= & \Phi^{-1}\left[P(X \leq X_k - 3h) + \frac{1}{n} \sum_{\substack{j: X_i - ah \leq X_j \leq X_i + ah \\ j \neq i}}^n G\left(\frac{X_k - X_j}{h}\right)\right] - \Phi^{-1}(F(X_k)) \\
& + \frac{1}{n} [\Phi^{-1}(b)]' G\left(\frac{X_k - X_i}{h}\right) \\
\equiv & S_3 + S_4.
\end{aligned}$$

Therefore,

$$\begin{aligned}
I4 & = \text{Cov}\left[\mathbb{E}(S_1 + S_2|U_{ik}), \mathbb{E}(S_3 + S_4|U_{ik})\right] \\
& = \text{Cov}\left[\mathbb{E}(S_1|U_{ik}) + \mathbb{E}(S_2|U_{ik}), \mathbb{E}(S_3|U_{ik}) + \mathbb{E}(S_4|U_{ik})\right] \\
& = \text{Cov}\left[\mathbb{E}(S_1|U_{ik}), \mathbb{E}(S_3|U_{ik})\right] + \text{Cov}\left[\mathbb{E}(S_1|U_{ik}), \mathbb{E}(S_4|U_{ik})\right] \\
& \quad + \text{Cov}\left[\mathbb{E}(S_2|U_{ik}), \mathbb{E}(S_3|U_{ik})\right] + \text{Cov}\left[\mathbb{E}(S_2|U_{ik}), \mathbb{E}(S_4|U_{ik})\right]
\end{aligned}$$

For  $\mathbb{E}(S_1|U_{ik})$ , since  $X_k$  does not contribute to  $S_1$ ,  $\mathbb{E}(S_1|U_{ik})$  is a constant. Similarly,

$\mathbb{E}(S_3|U_{ik})$  is also a constant. Therefore,  $\text{Cov}\left[\mathbb{E}(S_1|U_{ik}), \mathbb{E}(S_3|U_{ik})\right] = \text{Cov}\left[\mathbb{E}(S_1|U_{ik}), \mathbb{E}(S_4|U_{ik})\right] = \text{Cov}\left[\mathbb{E}(S_2|U_{ik}), \mathbb{E}(S_3|U_{ik})\right] = 0$ , and we only need to evaluate  $\text{Cov}(\mathbb{E}(S_2|U_{ik}), \mathbb{E}(S_4|U_{ik}))$ .

Note that,

$$\begin{aligned}
I4 &= \text{Cov}\left[\mathbb{E}(S_2|U_{ik}), \mathbb{E}(S_4|U_{ik})\right] \\
&= E(E(S_2|U_{ik})E(S_4|U_{ik}) - E(S_2)E(S_4)) \\
&\leq E(E(S_2|U_{ik})E(S_4|U_{ik})) \\
&= O(n^{-2}). \tag{A.14}
\end{aligned}$$

The second step is because both  $E(S_2)$  and  $E(S_4)$  are nonnegative. The last step is because  $G$  is bounded by 1 and due to the assumption of sample data on bounded domain, the expectation of  $((\Phi^{-1}\tilde{F}(X_i))')$  and  $((\Phi^{-1}\tilde{F}(X_k))')$  are bounded.

The first term in (A.6) can be written as:

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n \text{Var}\left[\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))\right] \\
&= \mathbb{E}\left[\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))\right]^2 - \left(E[\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))]\right)^2 \\
&\leq \mathbb{E}\left[\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))\right]^2 \quad \text{By mean value theory, } F^*(X_i) \in [\tilde{F}(X_i), F(X_i)] \\
&= \mathbb{E}\left(\frac{1}{\phi[\Phi^{-1}(F^*(X_i))]}[\tilde{F}(X_i) - F(X_i)]\right)^2 \\
&\leq c'\mathbb{E}[\tilde{F}(X_i) - F(X_i)]^2 \\
&= O(n^{-4/3}). \tag{A.15}
\end{aligned}$$

The fourth step is due to sample data on bounded domain. The last step is using the results from equation (A.12). We also need to evaluate the expectation of the second term in (A.5). By Hölder's Inequality,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n [\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))]\right] &= \sqrt{n}\mathbb{E}[\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))] \\
&\leq \sqrt{n}\left[\mathbb{E}[\Phi^{-1}(\tilde{F}(X_i)) - \Phi^{-1}(F(X_i))]^2\right]^{\frac{1}{2}}
\end{aligned}$$

$$\begin{aligned}
&= \sqrt{n}O_p(n^{-\frac{2}{3}}) \\
&= O_p(n^{-\frac{1}{6}})
\end{aligned}$$

To evaluate the convergence of the first term in (A.5), we consider two scenarios. When  $\mu_j = 0$  (true null) and when  $\mu_j \neq 0$  (false null). First, we evaluate the situation where  $\mu_j = 0$ . We have

$$\frac{\hat{\mu}_j}{\hat{\sigma}_j} - \frac{\mu_j}{\sigma_j} = \frac{\hat{\mu}_j}{\hat{\sigma}_j}$$

To estimate  $\hat{\mu}_j$ , we use the estimator defined in (3.4). Define  $p^* = p(\bar{X}_j \geq \frac{c}{\sqrt{2\pi \ln n}})$ . Then we have

$$\frac{\hat{\mu}_j}{\hat{\sigma}_j} = \begin{cases} \frac{\bar{X}_j}{\hat{\sigma}_j}, & \text{with probability } p^*, \\ 0, & \text{with probability } 1 - p^* \end{cases}$$

By Hoeffding's inequality and assuming sample data on a bounded domain,

$$p^* = p(\bar{X}_j \geq \frac{c}{\sqrt{2\pi \ln n}}) \leq \exp(-c_1 \frac{n}{\ln n}) \text{ for some constant } c_1 > 0$$

Under true null,

$$\text{Var}\left(\frac{\hat{\mu}_j}{\hat{\sigma}_j} - \frac{\mu_j}{\sigma_j}\right) = \text{Var}\left(\frac{\hat{\mu}_j}{\hat{\sigma}_j}\right) \leq \text{E}\left(\frac{\hat{\mu}_j}{\hat{\sigma}_j}\right)^2 = p^* \text{E}\left(\frac{\hat{\mu}_j}{\hat{\sigma}_j}\right)^2 = O\left(n^{-1} \exp(-c_1 \frac{n}{\ln n})\right)$$

The last step is based on the fact that  $\hat{\sigma}_j$  is lower bounded by a constant and the following:

$$\text{E}(\bar{X}_j)^2 = \text{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right)^2 = \frac{1}{n^2} \sum_{i=1}^n \text{E}(X_i)^2 = O\left(\frac{1}{n}\right)$$

Under false null,

$$\begin{aligned}
\frac{\hat{\mu}_j}{\hat{\sigma}_j} - \frac{\mu_j}{\sigma_j} &= \frac{\hat{\mu}_j \sigma_j - \hat{\sigma}_j \mu_j}{\hat{\sigma}_j \sigma_j} \\
&= \frac{\hat{\mu}_j \sigma_j - \sigma_j \mu_j + \sigma_j \mu_j - \hat{\sigma}_j \mu_j}{\hat{\sigma}_j \sigma_j}
\end{aligned}$$

$$\begin{aligned}
&= \frac{\widehat{\mu}_j - \mu_j}{\widehat{\sigma}_j} + \frac{\mu_j(\sigma_j - \widehat{\sigma}_j)}{\widehat{\sigma}_j\sigma_j} \\
&= O_p(n^{-1/2}) + O_p(\mu_j n^{-1/2})
\end{aligned}$$

The last step comes from the fact about  $\widehat{\sigma}_j$  that

$$\begin{aligned}
\widehat{\sigma}_j^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - (\bar{X} - \mu)^2 \\
&= \sigma_j^2 + O_p(n^{-1})
\end{aligned}$$

Therefore,  $\widehat{\sigma}_j = \sigma_j + O_p(n^{-1/2})$

Let  $p_0$  and  $p_1$  be the number of true nulls and false nulls respectively. Due to sparsity,  $p_0 \approx p$ . Thus,

$$\begin{aligned}
\sum_j^p \left( \frac{\widehat{\mu}_j}{\widehat{\sigma}_j} - \frac{\mu_j}{\sigma_j} \right)^2 &= \sum_{j \in \{\text{true null}\}} \left( \frac{\widehat{\mu}_j}{\widehat{\sigma}_j} - \frac{\mu_j}{\sigma_j} \right)^2 + \sum_{j \in \{\text{false null}\}} \left( \frac{\widehat{\mu}_j}{\widehat{\sigma}_j} - \frac{\mu_j}{\sigma_j} \right)^2 \\
&= O_p(pn^{-1} \exp(-c_1 \frac{n}{\ln n})) \tag{A.16}
\end{aligned}$$

Plugging all results from (A.7), (A.8), (A.13), (A.14), (A.15), (A.16), and plug them into (A.5), we have  $\|\widetilde{\mathbf{Z}} - \mathbf{Z}\| = O_p(p^{1/2}n^{-1/6} + p^{1/2} \exp(-c_1 \frac{n}{2 \ln n}))$ . Thus, along with (A.4), we have

$$\begin{aligned}
\sum_{i=1}^p |\widehat{\mathbf{b}}_i^T \widetilde{\mathbf{W}} - \mathbf{b}_i^T \widetilde{\mathbf{W}}| &= k\sqrt{p}\|\widetilde{\mathbf{Z}} - \mathbf{Z}\| + \sqrt{p} \left\| \sum_{h=1}^k [\widehat{\gamma}_h \widehat{\gamma}_h^T - \gamma_h \gamma_h^T] \right\| \|\mathbf{Z}\| \\
&= O_p\left(kpn^{-1/6} + kp \exp(-c_1 \frac{n}{2 \ln n}) + kp^{\frac{1}{2}-\tau} \|\boldsymbol{\mu}^*\| + kp^{1-\tau}\right)
\end{aligned}$$

Thus,

$$\begin{aligned}
\Delta_1 &= O_p\left(kpn^{-1/6} + kp \exp(-c_1 \frac{n}{2 \ln n}) + kp^{\frac{1}{2}-\tau} \|\boldsymbol{\mu}^*\| + kp^{1-\tau} + kpn^{-\delta} + p^{-\tau}\right) \\
&= O_p\left(kpn^{-\min\{1/6, \delta\}} + kp \exp(-c_1 \frac{n}{2 \ln n}) + kp^{1-\tau} + kp^{\frac{1}{2}-\tau} \|\boldsymbol{\mu}^*\|\right)
\end{aligned}$$

$\Delta_2$  can be handled in the same way with similar results.

Let  $\widehat{V}_A(t)$  be the numerator in (1.5). The least-squares estimator of the corresponding  $\widehat{\mathbf{W}}$  is

$$\widehat{\mathbf{W}} = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{Z} = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \boldsymbol{\mu}^* + \mathbf{W} + (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{K} = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \boldsymbol{\mu}^* + \mathbf{W}$$

which is due to the fact that  $\mathbf{B}$  and  $\mathbf{K}$  are orthogonal. Using similar proof as we did for  $|\widetilde{V}(t) - \widehat{V}_A(t)|$ , we can show that

$$\left| \widehat{V}_A(t) - V_A(t) \right| = O\left( \mathbf{1}^T \left| \mathbf{B}(\widehat{\mathbf{W}} - \mathbf{W}) \right| \right)$$

and

$$\mathbf{1}^T \left| \mathbf{B}(\widehat{\mathbf{W}} - \mathbf{W}) \right| = \mathbf{1}^T \left| \sum_{h=1}^k \gamma_h \gamma_h^T \boldsymbol{\mu}^* \right| \leq kp^{1/2} \|\boldsymbol{\mu}^*\|$$

Thus,

$$\frac{1}{p} \left| \widehat{V}_A(t) - V_A(t) \right| = O_p(kp^{-1/2} \|\boldsymbol{\mu}^*\|)$$

By proposition 1 in Fan & Han (2017), if  $p^{-1} \sqrt{\lambda_{k+1} + \dots + \lambda_p} = O(p^{-\epsilon})$  for some  $\epsilon > 0$ , then  $p^{-1} |V_A(t) - V(t)| = o_p(1)$ . Thus,

$$\begin{aligned} & \frac{1}{p} [\widetilde{V}(t) - V(t)] \\ &= O_p\left( kn^{-\min\{1/6, \delta\}} + k \exp(-c_1 \frac{n}{2 \ln n}) + kp^{-\tau} + kp^{-1/2} \|\boldsymbol{\mu}^*\| \right) \\ &= o_p(1) \end{aligned}$$

The proof of Theorem 2 is completed.