Estimating the Proportion of Nonzero Normal Means under Certain Strong Covariance Dependence

by

Xiongzhi Chen and R.W. Doerge
Department of Statistics, Purdue University

Technical Report #12-03

Department of Statistics
Purdue University

November 2012

Estimating the Proportion of Nonzero Normal Means under Certain Strong Covariance Dependence

Xiongzhi Chen and R.W. Doerge
Department of Statistics, Purdue University

Abstract

The proportion of certain type of hypotheses is a key component of adaptive false discovery procedures in multiple testing. To date, a good estimator of the proportion of false null hypotheses under dependence is lacking. For multiple testing normal means, we develop a (uniformly) consistent estimator of the proportion of nonzero normal means when the dependent test statistics follow a joint normal distribution with a known covariance matrix representing certain types of strong dependence. Theoretically and empirically, we demonstrate the better performance of our estimator than existing ones.

Contents

List of Figures

List of Tables

1 Introduction

Multiple testing has been widely conducted in differential gene expression studies (Efron, 2008), genome-wide association studies (Zhang and Liu, 2011), functional magnetic resonance imaging (Pacifico et al., 2004), and classification of interstellar objects (Liang et al., 2004). In typical multiple testing, there are m null hypotheses H_i with associated statistics T_i, $i = 1, ..., m$. The true status of each H_i is denoted by s_i such that $s_i = 0$ means H_i is true and that $s_i = 1$ means H_i is false, but it is unknown which among these m null hypotheses are true. Let $I_0^* = \{1 \leq i \leq m : s_i = 0\}$, $I_1^* = \{1, ..., m\} \setminus I_0^*$ with $m_l = |I_l^*|$ for $l = 0, 1$. The proportion of true nulls is defined as $\pi_0 = m_0/m_1$. A multiple testing procedure (MTP) $\mathcal{R}$ based on T_i measurably maps each $T = (T_1, ..., T_m)'$ to a unique $\hat{\mathbf{s}} = (\hat{s}_1, ..., \hat{s}_m)' \in \mathcal{S} = \{0, 1\}^m$ such that the inferred status of H_i is $\hat{s}_i$. The false discovery rate (FDR, Benjamini and Hochberg, 1995) of $\mathcal{R}$ is defined as $FDR(\mathcal{R}) = E[V(\mathcal{R})/R(\mathcal{R})| R(\mathcal{R}) > 0]$, where

$$V(\mathcal{R}) = |\{i \in I_0^* : \hat{s}_i = 1\}| \text{ and } R(\mathcal{R}) = |\{1 \leq i \leq m : \hat{s}_i = 1\}|. \quad (1)$$

When $\mathcal{R}$ depends on a threshold $t \in [0, \infty)$ such that $\hat{s}_i = 1_{\{|T_i| \leq t\}}$, $\mathcal{R}$ in (1) is replaced by t.

Due to its improved statistical power in multiple testing (Benjamini and Hochberg, 1995; Genovese and Wasserman, 2002) control of the FDR has become very popular in the aforementioned fields, and MTPs that incorporate a conservative estimate $\hat{\pi}_0$ of π_0, i.e.,

$$\pi_0 \leq \max\{\hat{\pi}_0, E[\hat{\pi}_0]\} < 1, \ P\text{-a.s.} \quad (2)$$

when $\pi_0 < 1$ and try to retain their FDRs under a prespecified level have been developed. Such MTPs (e.g., $\mathcal{R}_{E01}$ in Efron et al., 2001; $\mathcal{R}_{GW04}$ in Genovese and Wasserman, 2004; $\mathcal{R}_{BH06}$ in Benjamini et al., 2006; $\mathcal{R}_{BR09}$ in Blanchard and Roquain, 2009), commonly termed "adaptive FDR procedures", can be more powerful than their non-adaptive counterparts, and they show the importance of accurately estimating π_0 in the sense of (2). Many popular estimators of $\hat{\pi}_0$'s or of $\pi = 1 - \pi_0$ (whose meaning is to be specified later) have been proposed, and methods that derive them can be a roughly categorized into three classes:

1. The "slope" method to estimate π_0 (Storey et al., 2004; Benjamini et al., 2006; Blanchard and Roquain, 2009) whose origin can be traced back to the slope intuition in Schweder and Spjøtvoll (1982). Highly dependent on the (global or) local uniformity of the (empirical) density of the p-values, the method does not perform well under (strong) dependence (Blanchard and Roquain, 2009; Friguet and Causeur, 2011; Wang et al., 2011). The slope method yields the factor-slope hybrid (FSH) method in Friguet and Causeur (2011), which employs the same model in Friguet and Kloareg (2009) and uses $\hat{\pi}_0^S$ in Storey et al. (2004) to estimate π_0 based on the adjusted p-values. $\hat{\pi}_0^S$ is implemented by R library *qvalue* and the hybrid by R library *FAMT*.

2. Density estimation via mixture model to estimate π_0, which divides into three branches: (a) mode matching (Efron, 2008; Schwartzman, 2008). It is implemented by R package *locfdr,* robust to dependence but requires $\pi_0 \geq 0.9$; (b) nonparametric maximum likelihood estimator (NPMLE) in Langaas et al. (2005). It is implemented by the function *convest* in the R library *limma*, relies on the shape constraints on the marginal p-value density, and does not perform well under dependence; (c) Bayesian method (Ghosal and Roy, 2011), which depends on a good prior for the unknown parameters and is computationally demanding.

3. Fourier transform method (FTM) to estimate π (Jin and Cai, 2007; Jin, 2008), which works only for independent or strongly mixing test statistics each of whose density has the location-shift (and/or scale) property.

The above categories show that, consistency of estimation under strong dependence and adaptivity to smaller values of π_0 is not achieved by any of the existing estimators of π_0 (or π) simultaneously. Therefore, we aim to develop a better estimator of π that possesses these two properties under certain strong dependence. In particular, we consider the following setting. Suppose $\mathbf{Z} \sim N(\mu, \mathbf{\Sigma})$, i.e., $\mathbf{Z}$ is normally distributed with mean vector $\boldsymbol{\mu} = (\mu_1, ..., \mu_m)'$ and a known, deterministic correlation matrix $\mathbf{\Sigma} > 0$ representing certain types of strong dependency. We want to consistently estimate the proportion of nonzero normal means

$$\pi = m^{-1} \left|\{i : \mu_i \neq 0\}\right| \tag{3}$$

based on an observed vector $\mathbf{z}$ of $\mathbf{Z}$ (where a true null is equivalent to $\mu_i = 0$).

Our strategy to develop the consistent estimator of π has three steps. First, we utilize the principal factor approximation (PFA) developed in Fan et al. (2012) to decompose the jointly normally distributed random vector of test statistics into two independent random vectors, such that the major vector contributes the major part of the covariance dependence between the components of the original random vector and the minor vector consists of weakly dependent random variables. We then develop partial theory for non-concave partially penalized least squares estimate and apply the theory to consistently estimate the major vector. Finally, we extend a key bound needed for the Fourier transform method (FTM) to estimate π in Jin (2008) to the case of heterogeneous null distributions, and apply the extended Fourier transform method to the components of the mean-shifted, estimated minor vector to estimate π consistently.

2 Revisiting Principal Factor Approximation

We take the convention that each vector is a column vector, unless it is transposed. For a matrix $\mathbf{C} = (\mathbf{c}_1, ..., \mathbf{c}_m)$ and subset $A \subseteq \{1, ..., m\}$, we set $\mathbf{C}_A = (\mathbf{c}_i, i \in A)$ for which the order of the column indices of $\mathbf{C}_A$ is the same as that of the elements in A, and $\mathbf{C}^A = ((\mathbf{C}')_A)'$; in particular, we set $\mathbf{C}_{(k)} = (\mathbf{c}_1, ..., \mathbf{c}_k)$ and $\mathbf{C}_{(-k)} = (\mathbf{c}_{k+1}, ..., \mathbf{c}_m)$ for $1 \leq k < m$. When $\mathbf{C}$ is a column vector, this convection applies row-wise. We restate the PFA in Fan et al. (2012) as follows. Let $\mathbf{w} = (w_1, ..., w_m)' \sim N_m(\mathbf{0}, \mathbf{I})$. The spectral decomposition of $\mathbf{\Sigma}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$ and corresponding eigenvectors $\boldsymbol{\gamma}_i$, $i = 1, ..., m$ implies $Z_i = \mu_i + \eta_i + v_i$, where

$$\eta_i = \sum\nolimits_{j=1}^{k} \sqrt{\lambda_j}\gamma_{ij} w_j; \quad v_i = \sum\nolimits_{j=k+1}^{m} \sqrt{\lambda_j}\gamma_{ij} w_j.$$

Set $\mathbf{T} = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_m) = (\gamma_{ij})_{m \times m}$, $\mathbf{D} = diag\{\lambda_1, \lambda_2, ..., \lambda_m\}$, $\mathbf{G} = \mathbf{T}\sqrt{\mathbf{D}}$, $\boldsymbol{\eta} = (\eta_1, ..., \eta_m)' = \mathbf{G}_{(k)}\mathbf{w}_{(k)}$ and $\mathbf{v} = (v_1, ..., v_m)' = \mathbf{G}_{(-k)}\mathbf{w}_{(-k)}$. Then we have

$$\mathbf{Z} = \boldsymbol{\mu} + \boldsymbol{\eta} + \mathbf{v}. \tag{4}$$

Components of $\mathbf{w}_{(k)}$ are called the "principal factors" in Fan et al. (2012). We call $\boldsymbol{\eta}$ and $\mathbf{v}$ the "major" and "minor" vectors, respectively. In order to make the model (4) more flexible, we let

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

with $\mathbf{X}$ an $m \times p$ matrix and $\boldsymbol{\beta} \in \mathbb{R}^p$ a sparse vector such that $\|\boldsymbol{\beta}\|_0 = q_0 > 0$.

The decomposition (4) induces three appealing properties for asymptotic analysis later:

1. $\boldsymbol{\eta}$ is independent of $\mathbf{v}$.

2. There always exists a pair $\delta > 0$ and $0 \leq k \leq m$ such that

$$m^{-1}\sqrt{\lambda_{k+1}^2 + ... + \lambda_m^2} = O\left(m^{-\delta}\right). \tag{5}$$

3. Whenever (5) holds,

$$m^{-2}\sum_{1\leq i,j\leq m}\left|cov_{ij}^{\mathbf{v}}\right| \leq m^{-1}\left\|\mathbf{A}\right\|_F = m^{-1}\sqrt{\lambda_{k+1}^2 + ... + \lambda_m^2} = O\left(m^{-\delta}\right),$$

where $\mathbf{A} = cov\left(\mathbf{v}, \mathbf{v}\right) = \left(cov_{ij}^{\mathbf{v}}\right)_{m\times m}$ and $\left\|\cdot\right\|_F$ denotes the Froebinius norm of a matrix.

3 Non-concave Partially Penalized Least Squares

We introduce some conventions and notations. Let $\tilde{\mathbf{w}} = (\tilde{w}_1, ..., \tilde{w}_k)'$, $\tilde{\boldsymbol{\beta}} = \left(\tilde{\beta}_1, ..., \tilde{\beta}_p\right)'$ and take the convention that a scalar function applied to a vector results in a vector whose ith component is the function evaluated at the ith component of the vector. Further, we write $(\mathbf{v}_*', \mathbf{v}_{**}')'$ as $(\mathbf{v}_*, \mathbf{v}_{**})$ for two column vectors $\mathbf{v}_*$ and $\mathbf{v}_{**}$.

The accuracy of an estimator of π depends on that of $\mathbf{w}_{(k)}$ when PFA is used. The L_1-regression used in Fan et al. (2012) to estimate $\hat{\mathbf{w}}_{(k)}$ assumes $\pi \approx 0$ and omits all nonzero μ_i's in the optimization. To exploit the sparsity of $\boldsymbol{\mu}$ and allow for relative large $\pi \in (0, 1/2)$ when estimating $\mathbf{w}_{(k)}$, we consider the estimate, if it exists,

$$\left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{w}}_{(k)}\right) \in \operatorname*{arg\,min}_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^p, \tilde{\mathbf{w}}\in\mathbb{R}^k} L\left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{w}}; \lambda\right), \tag{6}$$

with

$$L\left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{w}}; \lambda\right) = \left\|\mathbf{Z} - \mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{G}_{(k)}\tilde{\mathbf{w}}\right\|^2 + \Lambda\left\|p_\lambda\left(\tilde{\boldsymbol{\beta}}\right)\right\|_1, \tag{7}$$

where $\Lambda > 0$ is a scale parameter, and $p_\lambda(\cdot) = \lambda\rho(\cdot)$ is the penalty function with $\rho(\cdot) = \rho(\cdot; \lambda)$. Specifically, we assume that ρ satisfies Condition 1 in Fan and Lv (2011):

Condition 1: $\rho(t)$ is increasing and concave in $t \in [0, \infty)$, and has a continuous derivative $\rho'(t)$ with $\rho'(0+) \in (0, \infty)$. If $\rho(t)$ is dependent on λ, $\rho'(t; \lambda)$ is increasing in $\lambda \in (0, \infty)$ and $\rho'(0+)$ is independent of λ.

In addition, we borrow from Lv and Fan (2009) the definition that for $\mathbf{b} = (b_1, ..., b_{q_0})' \in \mathbb{R}^{q_0}$ with $\left\|\mathbf{b}\right\|_0 = q_0$,

$$\kappa(\rho; \mathbf{b}) = \lim_{\varepsilon\to 0}\max_{1\leq j\leq q_0}\sup_{t_1, t_2\in(|b_j|-\varepsilon, |b_j|+\varepsilon), t_1<t_2} -\frac{\rho'(t_2) - \rho'(t_1)}{t_2 - t_1}.$$

Since the error distributions in (4) are weakly dependent, and in (6) only $\tilde{\mathbf{u}}$ but not $\tilde{\mathbf{w}}$ is penalized, we call (7) "non-concave partially penalized least squares (NCPPLS)" with dependent errors, for which no statistical theory on estimates in this setting seems to exist. Hereunder, we establish the existence of $\left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{w}}_{(k)}\right)$ in (6) and its weak oracle property (in the sense of Lv and Fan, 2009). Let $\hat{I}_1 = \left\{j : \hat{\beta}_j \neq 0\right\}$, $\hat{I}_0 = \left\{j : \hat{\beta}_j = 0\right\}$, $I_1 = \{1 \leq i \leq p : \beta_i \neq 0\}$, $I_0 = \{1, ..., p\} \backslash I_1$ and $\beta_* = \min_{i \in I_1} |\beta_i|$. Define $a_i = \left(1 - \sum_{j=1}^{k} \lambda_j \gamma_{ij}^2\right)^{-1/2}$, $a_{(1)} = \min_{i \in E_\infty^1, i \leq m} \{a_i\}$ and $a_{(m)} = \max_{i \in E_\infty^1, i \leq m} \{a_i\}$ when $E_\infty^1 \neq \varnothing$, where

$$E_\infty^1 = \{i \in \mathbb{N} : \cap_{k=1}^{\infty} \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} \{1 \leq i \leq n : \infty > a_i > k\}\}.$$

Define $\mathcal{N} \times \mathcal{N}^*$ with

$$\begin{cases} \mathcal{N} = \left\{\boldsymbol{\nu} \in \mathbb{R}^{q_0} : \|\boldsymbol{\nu}\|_0 = \left\|\boldsymbol{\beta}_{I_1}\right\|_0, \left\|\nu - \boldsymbol{\beta}_{I_1}\right\|_\infty \leq c_1\right\}, \\ \mathcal{N}^* = \left\{\boldsymbol{\varsigma} \in \mathbb{R}^k : \left\|\boldsymbol{\varsigma} - \mathbf{w}_{(k)}\right\|_\infty \leq c_2\right\}, \end{cases}$$

for some $c_1 = (1 - c_0)\beta_*$ for $c_0 \in (0, 1)$, $c_2 > 0$, and the event

$$\mathcal{E}_2 = \left\{\max_{1 \leq i \leq m} \{|v_i|\} \leq u\right\} = \{\|\mathbf{v}\|_\infty \leq u\} \text{ for } u > 0.$$

We now state the existence and weak oracle property of the NCPPLS estimates in (6):

Theorem 1 *Assume Conditions 2 to 6 in Section A.1 hold. Then, with probability at least*

$$1 - \left(a_{(1)} u\right)^{-1} \exp\left(-2^{-1} a_{(1)}^2 u^2\right)$$

on $\mathcal{E}_2$, setting $\lambda = \underline{\lambda}$ yields some $\left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{w}}_{(k)}\right) \in \arg\min_{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^p, \tilde{\mathbf{w}} \in \mathbb{R}^k} L\left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{w}}; \lambda\right)$ with $\left(\hat{\boldsymbol{\beta}}_{I_1}, \hat{\mathbf{w}}_{(k)}\right) \in \mathcal{N} \times \mathcal{N}^*$ and $\hat{\boldsymbol{\beta}}_{I_0} = \mathbf{0}$, whereby $c_2 = \lambda_k^{-1/2}\left(C_{7,k} c_1 + C_{8,k} u\right)$.

We remark that Theorem 1 generalize Theorems 4 of Lv and Fan (2009) to NCPPLS with weakly dependent errors.

4 Bound on Difference Between Phase Functions

Once we have (consistently) estimated $\mathbf{w}_{(k)}$ as $\hat{\mathbf{w}}_{(k)}$, the next step is to adapt the FTM to estimate π in Jin (2008) to our model to estimate π. The superior performance of the FTM to estimate π has been justified for statistical models with independent or strongly mixing normal random errors with homogeneous null distributions. However, due to the heterogeneity and dependence between v_i, $i=1,...,m$ in model (4), we need to extend the FTM to the case of heterogeneous null distributions.

Let $\phi_{\mu,\sigma}(x)=\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-2^{-1}\sigma^{-2}(x-\mu)^2\right)$ for $\sigma>0$. Define

$$\kappa_\sigma(t,x)=\int_{-1}^{1}\omega(\zeta)\,e^{(t\zeta\sigma)^2/2}\cos(t\zeta x)\,d\zeta, \tag{8}$$

which extends that in Jin (2008), and

$$\psi(t;\mu)=\int e^{it\mu\zeta}\omega(\zeta)\,d\zeta \tag{9}$$

for some $\omega(\zeta)$ that is bounded and symmetric around 0. We have $E[\kappa_\sigma(t,X)]=\psi(t;\mu)=\hat{\omega}(t\mu)$ when $X\sim N(\mu,\sigma^2)$, where $\hat{f}$ denotes the Fourier transform of a function $f\in L^1(\mathbb{R})$. Note that

$$E[\kappa_1(t,X)]=(\kappa_1(t,\cdot)*\phi_{0,1})(\mu)=\psi(t;\mu)=E[\kappa_\sigma(t,X)]=(\kappa_\sigma(t,\cdot)*\phi_{0,\sigma})(\mu),$$

Lemma 7.1 in Jin (2006) holds for the pair in (8) and (9), where $*$ denotes convolution of functions. Let $v_j^*=\mu_j+v_j$ and $\mathbf{v}^*=(v_1^*,...,v_m^*)'$. Then $v_j^*\sim N\left(\mu_j,a_i^{-2}\right)$. As done in Jin (2008), we define the underlying phase function

$$\varphi(t)=\varphi(t;\mu,m)=\frac{1}{m}\sum\nolimits_{j=1}^{m}[1-\psi(t;\mu_j)] \tag{10}$$

and empirical phase function

$$\varphi_m(t)=\varphi_m(t;\mathbf{v}^*)=\frac{1}{m}\sum\nolimits_{j=1}^{m}\left[1-\kappa_{a_j^{-1}}\left(t;v_j^*\right)\right]. \tag{11}$$

The performance of the FTM method to estimate π hinges upon how accurately φ_m approximates φ, the oracle that usually knows the true value of π. Therefore we will derive bounds on $|\varphi_m(t)-\varphi(t)|$ under the settings of the model (4).

Define

$$\Lambda_m(C_m)=\left\{(\boldsymbol{\mu},\boldsymbol{\sigma}):\max_{1\leq j\leq m}\{|\mu_j|+|\sigma_j|\}\leq C_m\right\}$$

with $\boldsymbol{\sigma}=(\sigma_1,...,\sigma_m)'$ where C_m depends only on m, and

$$\Theta_m(\gamma,C_m)=\left\{(\boldsymbol{\mu},\pi_m):\max_{i\in I_1^*}|\mu_i|\leq C_m;\mu_*=\min_{i\in I_1^*}|\mu_i|\geq\frac{\ln\ln m}{\sqrt{2\ln m}};\pi_m\geq m^{\gamma-1}\right\},$$

with $\gamma\in(0,1)$ where π in (3) is now written as π_m. We derive a bound on $\varphi_m(t)-\varphi(t)$ by fully exploiting the property of the minor vector.

Theorem 2 *Assume (5) and*

$$\limsup_{m\to\infty}\left|\rho_{ij}^{\mathbf{v}}\right|\leq 1-\varepsilon_0 \text{ for some } \varepsilon_0\in(0,1]. \tag{12}$$

Then for any $\tilde{\epsilon}>0$, *with probability at least* $1-M\tilde{\epsilon}^{-2}a_{(1)}^{-2}m^{-\delta}\ln m$,

$$\sup_{0\leq t\leq\sqrt{2\gamma\ln m}}\left|\varphi_m(t)-\varphi(t)\right|\leq M\vartheta_m(\gamma)\,\tilde{\epsilon}, \tag{13}$$

where

$$\vartheta_m(\gamma)\begin{cases}\leq e^M+o(1) & \text{if} \quad \limsup_{m\to\infty}a_{(1)}^{-2}\ln m=M,\\ =\dfrac{\exp\left(2^{-1}\gamma a_{(1)}^{-2}\ln m\right)}{\gamma a_{(1)}^{-2}\ln m}(1+o(1)) & \text{if} \quad \lim_{m\to\infty}a_{(1)}^{-2}\ln m=\infty.\end{cases}$$

Theorem 2 bounds $\delta_m(\varphi)=\sup_{0\leq t\leq\sqrt{2\gamma\ln m}}\left|\varphi_m(t)-\varphi(t)\right|$ when the random errors, i.e., v_j's, are heterogeneous and weakly dependent. Therefore, it generalizes the bound on $\delta_m(\varphi)$ in Jin and Cai (2007) and Jin (2008), where the random errors have the same null variances (since the null parameters need to be estimated there) and are strongly mixing.

5 Consistency of Plug-in Estimator

Set $\hat{\boldsymbol{\eta}}=\mathbf{G}_{(k)}\hat{\mathbf{w}}_{(k)}$. Since $\hat{\mathbf{w}}_{(k)}$ is a consistent estimator of $\mathbf{w}_{(k)}$, so is $\hat{\mathbf{v}}^*=(\hat{v}_1^*,...,\hat{v}_m^*)$ with $\hat{v}_j^*=Z_i-\hat{\eta}_i$ as that of $\mathbf{v}^*$. By Lipschitz property of φ_m, $\varphi_m(t;\hat{\mathbf{v}}^*)\to\varphi_m(t;\mathbf{v}^*)$ in probability. When the speed of convergence of $\hat{\mathbf{w}}_{(k)}$ to $\mathbf{w}_{(k)}$ is compatible with that of $\varphi_m(t;\hat{\mathbf{v}}^*)$ to $\varphi(t;\mathbf{v}^*)$, the plug in estimator $\varphi_m(t;\hat{\mathbf{v}}^*)$ in place of $\varphi(t;\mathbf{v}^*)$ to estimate π_m will be (uniformly) consistent.

Theorem 3 *Suppose Condition 7 in Section A.1, the conditions of Theorem 1 and of Theorem 2 hold. In addition, assume* $a_{(1)}\to\infty$ *and* $u\to 0$ *such that* $a_{(1)}u\to\infty$ *and that*

$$\limsup_{m\to\infty}a_{(1)}^{-2}\ln m=M \text{ and } \tilde{\epsilon}^{-2}m^{-\delta}=o(1).$$

Then, when $\tilde{\epsilon}\to 0$ *and*

$$\sqrt{2\gamma\ln m}\,(C_{8,k}c_1+C_{9,k}u)\,\tilde{\epsilon}=o(\pi_m),$$

the plug-in procedure $\varphi_m(t;\hat{\mathbf{v}}^*)$ *with* $\left(\hat{\boldsymbol{\beta}},\hat{\mathbf{w}}_{(k)}\right)\in\arg\min_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^p,\tilde{\mathbf{w}}\in\mathbb{R}^k}L\left(\tilde{\boldsymbol{\beta}},\tilde{\mathbf{w}};\lambda\right)$ *is consistent. When further*

$$\sqrt{2\gamma\ln m}\,(C_{8,k}c_1+C_{9,k}u)\,\tilde{\epsilon}=o\left(m^{1-\gamma}\right),$$

$\varphi_m(t;\hat{\mathbf{v}}^*)$ *is uniformly consistent on* $\Theta_m(\gamma,C_m)$, *i.e.,*

1. *For eligible*[1] ω, $\lim_{m\to\infty}\sup_{\Theta_m(\gamma,C_m)}\left|\frac{\varphi_m\left(\sqrt{2\gamma\ln m};\hat{\mathbf{v}}^*\right)}{\pi_m}-1\right|=0$,

2. *For good* ω, $\lim_{m\to\infty}\sup_{\Theta_m(\gamma,C_m)}\left|\frac{\sup_{0<t\leq\sqrt{2\gamma\ln m}}\varphi_m\left(t;\hat{\mathbf{v}}^*\right)}{\pi_m}-1\right|=0$.

6 Implementation of Plug-in Estimator

PFA is easily implemented by spectral decomposition and then choosing

$$k_\delta=\min\left\{k:m^{-1}\sqrt{\lambda_{k+1}^2+...+\lambda_m^2}\leq m^{-\delta}\right\}$$

with a preset $\delta\in(0,1)$, say, $\delta=0.5$. Then we implement the NCPPLS using the minimax concave penalty (MCP) $p_\lambda(\cdot)$ of Zhang (2010) such that

$$\lambda^{-1}p_\lambda(t)=\rho(t;\lambda)=\int_0^t(1-x/(\theta\lambda))_+\,dx \text{ for } t>0,$$

where $\theta=3.7$ is the parameter for the degree of concavity and λ the tuning parameter. A local optimum

$$\left(\hat{\boldsymbol{\beta}},\hat{\mathbf{w}}_{(k)},\hat{\lambda}\right)\in\underset{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^p,\tilde{\mathbf{w}}\in\mathbb{R}^k,\lambda\geq0}{\arg\min}L\left(\tilde{\boldsymbol{\beta}},\tilde{\mathbf{w}};\lambda\right)$$

is found via alternating optimization (e.g., Bezdek and Hathaway, 2002) through the following steps:

1. Set $j=0$ and initialize $\tilde{\mathbf{w}}_*^{(j)}=\mathbf{0}$.

2. Use the R package *sparsenet* of Mazumder et al. (2011) to compute the solution path of

$$\hat{\boldsymbol{\beta}}_\lambda^{(j)}\in\underset{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^m}{\arg\min}L\left(\tilde{\boldsymbol{\beta}},\tilde{\mathbf{w}}_*^{(j)};\lambda\right)$$

   for $\lambda\in S_\lambda=\{\lambda_{k^*},...,\lambda_0\}$, a grid of descendingly ordered λ values automatically set by *sparsenet*, where $\hat{\boldsymbol{\beta}}_\lambda^{(j)}$ has been indexed by $\lambda\in S_\lambda$. Pick $\hat{\boldsymbol{\beta}}_*^{(j)}$ for which

$$\left(\hat{\boldsymbol{\beta}}_*^{(j)},\lambda_*^{(j)}\right)=\underset{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^p,\lambda\in S_\lambda}{\arg\min}L\left(\tilde{\boldsymbol{\beta}},\tilde{\mathbf{w}}_*^{(j)};\lambda\right).$$

[1]From Jin (2008), a density function ω over $(-1,1)$ is eligible if it is continuous, symmetric (around 0), and bounded. It is good if additionally there exists some convex, super-additive function $g^\#$ such that $\omega(\zeta)=g^\#(1-\zeta)$ for all $0<\xi<1$.

3. Compute
$$\tilde{\mathbf{w}}_*^{(j+1)} = \mathbf{D}_{(k)}^{-1/2}\mathbf{T}_{(k)}' \left(\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}_*^{(j)}\right)$$
where $\mathbf{z}$ is an observation from $\mathbf{Z}$.

4. Set j to be $j+1$. Repeat steps 2 and 3 until
$$\left\|\hat{\boldsymbol{\beta}}_*^{(j+1)} - \hat{\boldsymbol{\beta}}_*^{(j)}\right\|_2 < \tau_* \left\|\hat{\boldsymbol{\beta}}_*^{(j)}\right\|_\infty \text{ with } \tau_* = 10^{-5}.$$

Suppose a local minimizer $\left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{w}}_{(k)}\right)$ has been found, we compute
$$\hat{\mathbf{v}}^* = \mathbf{z} - \mathbf{G}_{(k)}\hat{\mathbf{w}}_{(k)}.$$

To implement the extended FTM to estimate π, we have adapted the R codes of Jin (2008) to accommodate (10) and (11) since a_j^{-1} are not equal to each other. The adapted codes, available from the authors, are then applied to components of $\hat{\mathbf{v}}_*$ to estimate π, where the triangular density $\omega(x) = (1 - |x|)\, 1_{\{|x|\leq 1\}}$ is used.

7 Numerical Studies

7.1 Settings

Even though our theory has been developed for $\mathbf{X}$ not necessarily being the identity matrix, we set $\mathbf{X} = \mathbf{I}$ in our simulations, which means $\boldsymbol{\beta} = \boldsymbol{\mu}$ and $\pi = \pi_{\boldsymbol{\mu}} = \pi_{\boldsymbol{\beta}} = \#\{j : \beta_j \neq 0\}/m$. For $m = 2000$, $p = 1.5m$, $\pi_{\boldsymbol{\mu}} = 0.2$ and 0.4,

1. generate vector $\boldsymbol{\beta} \in \mathbb{R}^p$, where $\beta_j = 0$ for $1 \leq j \leq p(1-\pi_\beta)$ and $\beta_j \neq 0$ for $p(1-\pi_\beta) + 1 \leq j \leq p$. For $\beta_* = 0.5$ and 3, nonzero $|\beta_j|$'s are generated from the uniform distribution on $[\beta_*, \beta_* + 1]$; for $\beta_* = 0.01$, they are generated from the uniform distribution on $[\beta_*, \beta_* + 0.03]$. The signs of $\beta_j \neq 0$ are generated from $p(1-\pi_\beta)$ independent Bernoulli random variables each taking values ± 1 with equal probability.

2. generate $\mathbf{w} = (w_1, ..., w_m)' \sim N(\mathbf{0}, \mathbf{I})$ and set $\tilde{\mathbf{Z}} = \left(\tilde{Z}_1, ..., \tilde{Z}_m\right)$.

3. generate two covariance structures:

   - equicorrelation: set $\boldsymbol{\Sigma} = 0.5\mathbf{I} + 0.5\mathbf{1}\mathbf{1}'$ and generate $\tilde{\mathbf{Z}} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$; set $\mathbf{Z} = \boldsymbol{\beta} + \tilde{\mathbf{Z}}$.
   - two-component long-range dependence: set $\tilde{Z}_1 = w_1$, $\tilde{Z}_2 = (-w_1 + w_2)/\sqrt{2}$, ..., $\tilde{Z}_{m-1} = (-w_1 + w_{m-1})/\sqrt{2}$ and $\tilde{Z}_m = (-w_1 + w_m)/\sqrt{2}$; set $\mathbf{Z} = \boldsymbol{\beta} + \tilde{\mathbf{Z}}$.

4. Repeat Steps 2. and 3. 100 times.

7.2 Partial Simulation Results

To maintain concise presentation, we set $\boldsymbol{\zeta} = (\zeta_1, ..., \zeta_m)'$ and $\hat{\boldsymbol{\zeta}} = \left(\hat{\zeta}_1, ..., \hat{\zeta}_m\right)$, where $\zeta_j = \mu_j + \eta_j$, $\hat{\zeta}_j = \hat{\mu}_j + \hat{\eta}_j = \hat{\mu}_j + \left(\mathbf{G}_{(k)}\hat{\mathbf{w}}_{(k)}\right)_j$. Note that when $m = 2000$, the theoretical threshold for the minimal conditional mean $\zeta_* = \min\{|\zeta_j| : |\zeta_j| > 0\}$ is

$$\kappa_* = \frac{\ln \ln m}{\sqrt{2 \ln m}} \approx 0.52.$$

Denote the plug-in estimator $\varphi_m(t; \hat{\mathbf{v}}^*)$ by $\tilde{\pi}$ and its variant $\varphi_m(t; \hat{\mathbf{v}}^+)$ obtained using the method in Jin (2008) by $\tilde{\pi}^*$, where $\hat{\mathbf{v}}^+ = \hat{\mathbf{v}}^* \circ \mathbf{a}$ with $\mathbf{a} = (a_1, .., a_m)'$ is the standardized version of $\hat{\mathbf{v}}^*$. By the duality between π and $\pi_0 = 1 - \pi$, we will transform an estimator $\hat{\pi}_0$ of π_0 into $\hat{\pi} = 1 - \hat{\pi}_0$, and compare $\tilde{\pi}$ and $\tilde{\pi}^*$ with $\hat{\pi}^{BH} = 1 - \hat{\pi}_0^{BH}$ where

$$\hat{\pi}_0^{BH} = \frac{m - [m/2] + 1}{m\left(1 - p_{([m/2])}\right)}$$

in Benjamini et al. (2006), $\hat{\pi}^L = 1 - \hat{\pi}_0^L$ with $\hat{\pi}_0^L$ in Langaas et al. (2005), and $\hat{\pi}^{SLIM} = 1 - \hat{\pi}_0^{SLIM}$ with $\hat{\pi}_0^{SLIM}$ in Wang et al. (2011). It should be noted that $\hat{\pi}^{SLIM}$ was claimed by its developers to be robust to dependence. We did not compare $1 - \tilde{\pi}$ or $1 - \tilde{\pi}^*$ (as estimators of π_0) with that in Friguet and Causeur (2011) since the latter needs multiple observations for the response vector, or with those in Storey et al. (2004) and Efron (2010) since they directly break down with error messages when applied to our simulations.

We will not discuss the performance of the NCPPLS estimator since our focus is on estimators of π. Due to the usage of PFA, the conditional means ζ_j and the dependence between components $\hat{v}_j$ of the estimate $\hat{\mathbf{v}}_* = \mathbf{z} - \mathbf{G}_{(k)}\hat{\mathbf{w}}_{(k)}$ are the essential factors that effect the performance of $\tilde{\pi}$ and $\tilde{\pi}^*$. We remark that for the marginal means μ_j even though μ_* is relatively large, ζ_* obtained from the conditional means ζ_j can be very small due to the randomness of $\mathbf{w}_{(k)}$ and $\mathbf{G}_{(k)}$. Further, as inputs to the extended FTM to estimate π, the minimum $\hat{v}^*_{\min} = \min\left\{\left|\hat{v}^*_j\right| : \hat{v}_j \neq 0\right\}$ can be much smaller than μ_*, which creates additional difficulty in estimating π. Such differences are illustrated by Table 1 and Table 2.

We also remark that for our simulations the conditions of Theorem 3 are violated. For $\beta_* = \mu_* = 3$, $\pi = 0.2$ and $|\mu_j| \sim U(3, 4)$ when $\mu_j \neq 0$, Figure 1 and Figure 2 show the performances of five competing estimators of π and that of the NCPPLS estimator respectively for the equicorrelation and two-component long-range covariance dependence. We choose to display the results for $\pi = 0.2$ to better illustrate the advantage of the new estimator $\tilde{\pi}$, in that the smaller π is, the more the noise there is and the harder it is to estimate π. As shown by these two figures, for both types of covariance dependence the new estimators $\tilde{\pi}$ is the most accurate and most stable. Besides, it is almost always less than π, which is necessary for an adaptive FDR procedure that uses $\tilde{\pi}$ to maintain conservative control of the prespecified FDR level. In contrast, $\hat{\pi}^{SLIM}$, even though claimed to be robust to dependence, is very unstable and often larger than π; the quantile-based estimator $\hat{\pi}^{BH}$ is always larger than π and unstable; the NPMLE $\hat{\pi}^{L}$, depending on the shape of the (empirical) distribution of p-values, is always very close to 1, i.e., erroneous.

Additional details on Figure 1 and Figure 2 for the case $\pi = 0.4$ are provided in Table 1. We point out that, for the two types of covariance dependence considered, ζ_{**}, the minimum of all ζ_*'s for the 100 replications across the two settings for π are far less than the theoretically required minimal signal strength κ_*. Specifically, the maximum, $\zeta_{**}^{\max}$, of ζ_{**} for these four settings (two covariance dependences and two values of π) is 0.04311, implying

$$\frac{\zeta_{**}^{\max}}{\kappa_*} = \frac{0.04311}{0.52} = 0.08290.$$

Despite the existence of such weak signals (i.e., nonzero conditional means of small magnitudes), the new estimator $\tilde{\pi}$ still maintains high accuracy as its sample mean is less than π but is within 0.01 allowance from π, and high stability since it has the smallest sample standard deviation (being no larger than 0.0045) among all the competing estimators of π. It is interesting to note that the variant $\tilde{\pi}^*$ also has excellent performance since it is almost as accurate as $\tilde{\pi}$ (but is less stable because of the amplification induced by standardization of $\hat{\mathbf{v}}_*$). These observations demonstrate the superior performance of the new estimator $\tilde{\pi}$ and the slight improvement of $\tilde{\pi}$ in stability over the method in Jin (2008) for heterogeneous nulls when the minimal signal strength is not too small. We remark that this batch of simulations supplements Experiment (a) in Jin (2008) where in our notations $m = 10^5$, $\mu_* = 0.5$, 0.75, 1, 1.25 and $\kappa_* = 0.5092$. That is, in this experiment of Jin, $\mu_* \approx$ or $> \kappa_*$ holds, but in our simulations $\zeta_{**}^{\max} \ll \min\{0.52, 0.5092\}$. Our simulation results add to the evidence that the (extended) FTM to estimate π is able to detect weak signals (where we have called a nonzero (conditional) normal mean a "signal").

For the above two types of strong covariance dependence, we conducted another batch of simulations where $|\mu_j| \sim U(0.5, 1.5)$ with and $U(0.01, 0.04)$ when $\mu_j \neq 0$ for $\pi_\mu = 0.4$ and 0.2. The results are reported in Table 2. In terms of accuracy, stability and being no larger than the true proportion π, the variant $\tilde{\pi}^*$ is the winner when $\pi = 0.2$ while all other competing estimators give erroneously estimates much larger or smaller than π, with $\hat{\pi}^{SLIM}$ being the second worst. When $\pi = 0.4$ and $\mu_* = 0.5$, $\hat{\pi}^{BH}$ is the winner, $\hat{\pi}^{SLIM}$ the second best, and the variant $\hat{\pi}^*$ the third, with $\hat{\pi}^{L}$ being very close to 1; when $\pi = 0.4$ and $\mu_* = 0.01$, $\hat{\pi}^{BH}$ is the winner, the variant $\tilde{\pi}^*$ the second best, while $\hat{\pi}^{SLIM}$ and $\hat{\pi}^{L}$ are the worst. In all settings, $\tilde{\pi}$ and $\tilde{\pi}^*$ has the smallest sample standard deviations, and is the most stable. The better performance of $\tilde{\pi}^*$ than $\tilde{\pi}$ in such cases is due to the fact that $\tilde{\pi}^*$ uses $\hat{\mathbf{v}}^+$, each component of which has larger magnitude than that of $\hat{\mathbf{v}}_*$.

The simulation results seem to suggest preferences to different estimators of π under different sparsity settings. When the signal strength is large, i.e., μ_* is large, the new estimator $\tilde{\pi}$ is recommended, since in this case better performance of the NCPPLS estimator can be expected and relatively larger signals are available for the extended FTM to estimate

π. On the other hand, when the signal is very sparse, the variant $\tilde{\pi}^*$ of the new estimator $\tilde{\pi}$ is recommended, since $\tilde{\pi}^*$ uses amplified signals and the extended FTM method is able to detect the existence of fairly weak signals. Finally, when the signal strength is small but the number of signals is relatively large, $\hat{\pi}^{BH}$ is recommended and preferred to $\hat{\pi}^{SLIM}$, since in this case in an overall fashion more information on the distribution of nonzero normal means is available to the quantile-based estimator $\hat{\pi}^{BH}$ even though every bit of the available information is weak.

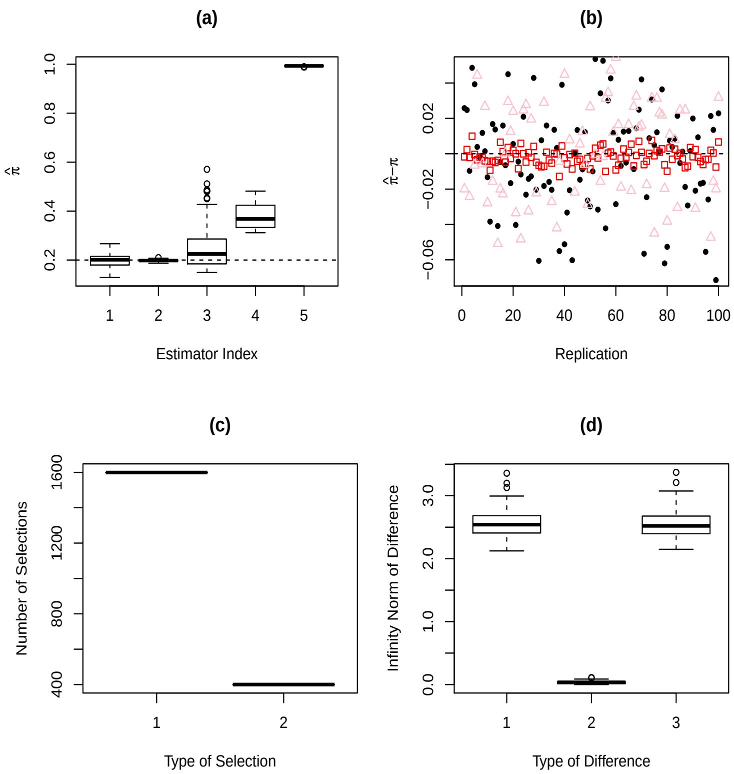

Figure 1: Equicorrelation with $\pi_{\boldsymbol{\mu}} = 0.2$. (a) Boxplots of five estimators of $\pi_{\boldsymbol{\mu}}$: 1 for $\tilde{\pi}^* = \hat{\pi}^1$; 2 for $\tilde{\pi} = \hat{\pi}^2$; 3 for $\hat{\pi}^{SLIM} = \hat{\pi}^3$; 4 for $\hat{\pi}^{BH} = \hat{\pi}^4$; 5 for $\hat{\pi}^L = \hat{\pi}^5$. (b) (truncated) difference between each estimator and the truth: $\circ$ for $\hat{\pi}^1 - \pi_{\boldsymbol{\mu}}$; $\square$ for $\hat{\pi}^2 - \pi_{\boldsymbol{\mu}}$; $\triangle$ for $\hat{\pi}^3 - \pi_{\boldsymbol{\mu}}$; $+$ for $\hat{\pi}^4 - \pi_{\boldsymbol{\mu}}$; $*$ for $\hat{\pi}^5 - \pi_{\mu}$. An undisplayed symbol indicates the corresponding difference is out of the specified range for the differences. (c) Boxplots of false and true selections: 1 for the number of incorrectly selected β_j's, (i.e., false selection); 2 for that of correctly selected nonzero β_j's, (i.e., true selections). (d) Boxplots of infinity norms of penalized estimates: 1 for $\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_\infty$; 2 for $\left\|\boldsymbol{w}_{(k)} - \hat{\boldsymbol{w}}_{(k)}\right\|_\infty$; 3 for $\left\|\boldsymbol{\zeta} - \hat{\boldsymbol{\zeta}}\right\|_\infty$.

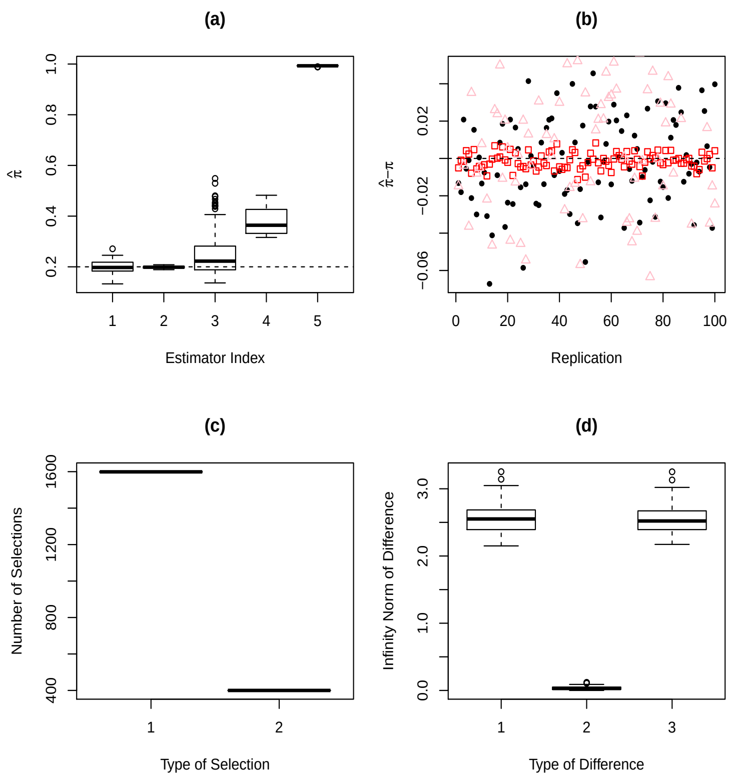

Figure 2: Two-component long-range dependence with $\pi_{\boldsymbol{\mu}} = 0.2$. (a) Boxplots of five estimators of $\pi_{\boldsymbol{\mu}}$: 1 for $\tilde{\pi}^* = \hat{\pi}^1$; 2 for $\tilde{\pi} = \hat{\pi}^2$; 3 for $\hat{\pi}^{SLIM} = \hat{\pi}^3$; 4 for $\hat{\pi}^{BH} = \hat{\pi}^4$; 5 for $\hat{\pi}^L = \hat{\pi}^5$. (b) (truncated) difference between each estimator and the truth: $\circ$ for $\hat{\pi}^1 - \pi_{\boldsymbol{\mu}}$; $\square$ for $\hat{\pi}^2 - \pi_{\boldsymbol{\mu}}$; $\triangle$ for $\hat{\pi}^3 - \pi_{\boldsymbol{\mu}}$; $+$ for $\hat{\pi}^4 - \pi_{\boldsymbol{\mu}}$; $*$ for $\hat{\pi}^5 - \pi_{\boldsymbol{\mu}}$. An undisplayed symbol indicates the corresponding difference is out of the specified range for the differences. (c) Boxplots of false and true selections: 1 for the number of incorrectly selected β_j's, (i.e., false selection); 2 for that of correctly selected nonzero β_j's, (i.e., true selections). (d) Boxplots of infinity norms of penalized estimates: 1 for $\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_\infty$; 2 for $\left\|\boldsymbol{w}_{(k)} - \hat{\boldsymbol{w}}_{(k)}\right\|_\infty$; 3 for $\left\|\boldsymbol{\zeta} - \hat{\boldsymbol{\zeta}}\right\|_\infty$.

8 Conclusion and Discussion

To induce better adaptive FDR procedures for multiple testing normal means under covariance dependence, we have developed a (uniformly) consistent estimator of the proportion of nonzero normal means when the strongly dependent test statistics follow a joint normal

	Two-comp Long-range		Equal Correlation	
$\pi_{\boldsymbol{\mu}}$	0.4	0.2	0.4	0.2
μ_*	3	3	3	3
$\zeta_{*,med}$	0.54499	0.50795	0.54499	0.50795
ζ_{**}	0.00509	0.04311	0.00509	0.04311
$a_{(m)}$	44.76606	44.76606	1.41457	1.41457
κ_*	0.52021	0.52021	0.52021	0.52021
$\widehat{v}^*_{\min *} \times 10^3$	0.00847	0.00577	0.00918	0.00922
$\widehat{v}^*_{\max *}$	9.75182	9.54028	9.26243	9.47015
$\widehat{\pi}_1$	0.39772 (0.02419)	0.19866 (0.02482)	0.39929 (0.02347)	0.19790 (0.02908)
$\widehat{\pi}_2$	**0.39631** **(0.00403)**	**0.19864** **(0.00424)**	**0.39614** **(0.00407)**	**0.19824** **(0.00450)**
$\widehat{\pi}_3$	0.41388 (0.04638)	0.25245 (0.09478)	0.41597 (0.04643)	0.25420 (0.09480)
$\widehat{\pi}_4$	0.43688 (0.02737)	0.38007 (0.05338)	0.43728 (0.02713)	0.38065 (0.05290)
$\widehat{\pi}_5$	0.98853 (0.00113)	0.99303 (0.00164)	0.98854 (0.00111)	0.99303 (0.00162)
FSP	0.99935 (0.00073)	0.99936 (0.00062)	0.99938 (0.00077)	0.99928 (0.00059)
TSP	1 (0)	1 (0)	1 (0)	1 (0)
$\left\Vert \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right\Vert_\infty$	2.54087 (0.22432)	2.55756 (0.20322)	2.54298 (0.17811)	2.56988 (0.23434)
$\left\Vert \mathbf{w}_{(k)} - \hat{\mathbf{w}}_{(k)} \right\Vert_\infty$	0.03566 (0.02403)	0.03601 (0.02713)	0.03553 (0.02542)	0.03671 (0.02654)
$\left\Vert \boldsymbol{\zeta} - \hat{\boldsymbol{\zeta}} \right\Vert_\infty$	2.53536 (0.22542)	2.54690 (0.20398)	2.54243 (0.17894)	2.56347 (0.23175)

Table 1: Estimators of π with $\mu_* = 3$: μ_* is the minimal marginal mean. $\zeta_{*,med}$ is the median of all ζ_*'s for all 100 replications; ζ_{**}, $\hat{v}^*_{\min *}$ and $\hat{v}^*_{\max *}$ are respectively the minimum of ζ_*'s, the minimum of $\hat{v}^*_{\min}$'s and the maximum of $\hat{v}^*_{\max}$'s for the 100 replications. From the 10th row on, in each cell the top and bottom numbers are respectively the sample mean and sample standard deviation of the corresponding estimator in the 1st column. FSP denotes the false selection proportion and TSP the true selection proportion.

	Two-component Long-range				Equal Correlation			
$\pi_{\boldsymbol{\mu}}$	0.4		0.2		0.4		0.2	
μ_*	0.5	0.01	0.5	0.01	0.5	0.01	0.5	0.01
$\zeta_{*,med}$	0.00392	0.50511	0.02397	0.46822	0.00413	0.50511	0.01776	0.46822
$\zeta_{**} \times 10^3$	0.03018	0.02483	0.37225	3.59922	0.00407	0.00468	0.00945	3.16668
$a_{(m)}$	44.76606	44.76606	44.76606	44.76606	1.41457	1.41457	1.41457	1.41457
κ_*	0.52021	0.52021	0.52021	0.52021	0.52021	0.52021	0.52021	0.52021
$\widehat{v}^*_{\min *} \times 10^3$	0.03844	0.00769	0.00769	0.00221	0.00153	0.00302	0.00843	0.00760
$\widehat{v}^*_{\max *}$	6.18574	5.73176	6.50909	5.72607	5.69584	4.44044	5.92522	4.76299
$\widehat{\pi}_1$	**0.28941** (**0.03010**)	**0.01353** (**0.01757**)	**0.14705** (**0.02704**)	**0.01114** (**0.01713**)	**0.29276** (**0.03304**)	**0.01242** (**0.01689**)	**0.14678** (**0.03274**)	**0.01329** (**0.01794**)
$\widehat{\pi}_2$	0.11543 (0.00561)	0.00231 (0.00298)	0.05954 (0.00480)	0.00218 (0.00312)	0.11594 (0.00694)	0.00243 (0.00313)	0.05941 (0.00596)	0.00232 (0.00320)
$\widehat{\pi}_3$	0.33673 (0.32966)	0.83585 (0.27250)	0.52275 (0.38207)	0.79240 (0.28800)	0.29229 (0.30431)	0.82889 (0.27903)	0.51618 (0.37526)	0.80611 (0.28072)
$\widehat{\pi}_4$	**0.36381** (0.04048)	**0.33129** (0.05922)	0.35380 (0.05710)	0.33858 (0.06673)	**0.36400** (0.04041)	**0.33166** (0.05948)	0.35395 (0.05736)	0.33904 (0.06668)
$\widehat{\pi}_5$	0.99928 (0.00190)	0.99977 (0.00111)	0.99889 (0.00256)	0.99928 (0.00186)	0.99924 (0.00194)	0.99976 (0.00109)	0.99893 (0.00250)	0.99929 (0.00184)

Table 2: Estimators of π: μ_* is the minimal marginal mean. $\zeta_{*,med}$ is the median of all ζ_*'s for all 100 replications; ζ_{**}, $\hat{v}^*_{\min *}$ and $\hat{v}^*_{\max *}$ are respectively the minimum of ζ_*'s, the minimum of $\hat{v}^*_{\min}$'s and the maximum of $\hat{v}^*_{\max}$'s for the 100 replications. From the 10th row on, in each cell the top and bottom numbers are respectively the sample mean and sample standard deviation of the corresponding estimator in the 1st column.

distribution with a known but non-arbitrary covariance matrix. As by-products of developing this estimator, we have established the existence and weak oracle property of the solution of non-concave partially penalized least squares with weakly dependent normal random errors, and extended the Fourier transform method to estimate this proportion in Jin (2008) to the case of weakly dependent, heterogeneous null distributions. It is by far the only consistent estimator of the proportion of nonzero normal means under certain strong covariance dependencies, and it can be plugged into the latest FDR procedures under dependence (e.g., Fan et al., 2012) to make them adaptive to lower levels of sparsity of the normal means.

The regularity Condition 4 in Section A.1 to establish Theorem 1 is closely related to the irrepresentable condition of Zhao and Yu (2006) when $\mathbf{w}_{(k)} = \mathbf{0}$ in model (4), and it can be violated in practice. Besides Condition 4, another restriction on the applicability of the consistent estimator is the assumption of a known covariance matrix for the jointly normally distributed test statistics since usually such a covariance matrix needs to be estimated. Unfortunately, existing theory can only accurately estimate a few types of large-scale variance-covariance matrices, while in practice the complex dependence structure between the test statistics may render such theory inapplicable.

Despite the implausibility of the assumptions made in Theorem 3, our strategy to develop the new consistent estimator depends only on the dependency between the components of the mean-shifted estimated minor vector and on the minimal magnitude of the nonzero conditional normal means. This implies that, even when certain assumptions of Theorem 3 do not hold, the proportion of nonzero normal means π can still be consistently estimated as long as

1. the major dependency among the components of the test statistics $\mathbf{Z}$ can be extracted, resulting in weak dependency between the components of the mean-shifted estimated minor vector $\hat{\mathbf{v}}^*$,

2. the estimated conditional mean vector $\hat{\boldsymbol{\zeta}}$ retains the same proportion of nonzero components as that of the marginal mean vector $\boldsymbol{\mu}$,

3. ζ_* is no less than the theoretical minimal signal strength κ_*.

The above intuition is supported by our empirical finding that the new estimator performs well in our simulation studies where some conditions of Theorem 3 are not fully satisfied. We will leave the development of a consistent estimator of the proportion of nonzero normal means under strong covariance dependence without using the consistent estimators of the principal factors to future research.

Acknowledgements

The authors would like to thank Professor Jiashun Jin for providing the technical report Jin (2006). This research was funded in part by a National Science Foundation Plant Genome Research Program grant (no. IOS-1025976) to RWD.

A Technical Proofs

A.1 Regularity Conditions

Condition 2: $\mathbf{X}$ and $\mathbf{G}_{(k)}$ satisfy $\mathbf{X}'_{I_1}\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}'_{(k)}\right)\mathbf{X}_{I_1}>0$ and $\liminf_{m\to\infty}\lambda_k>0$.

Condition 3: $\mathbf{X}$ and $\mathbf{G}_{(k)}$ satisfy

$$\left\|\left[\mathbf{X}'_{I_1}\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}'_{(k)}\right)\mathbf{X}_{I_1}\right]^{-1}\right\|_\infty<C_{3,k},\ \left\|\mathbf{X}'_{I_1}\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}'_{(k)}\right)\right\|_\infty<C_{4,k},$$

$$\left\|\mathbf{X}'_{I_0}\mathbf{T}_{(k)}\mathbf{T}'_{(k)}\mathbf{X}_{I_1}\right\|_\infty\le C_{5,k}\ ,\ \left\|\mathbf{X}'_{I_0}\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}'_{(k)}\right)\right\|_\infty\le C_{6,k},$$

$$\left\|\mathbf{T}'_{(k)}\mathbf{X}_{I_1}\right\|_\infty\le C_{7,k}\ ,\ \left\|\mathbf{T}'_{(k)}\right\|_\infty\le C_{8,k}.$$

Condition 4: ρ, $\mathbf{X}$ and $\mathbf{G}_{(k)}$ satisfy $\dfrac{\rho'(c_0\beta_*)}{\rho'(0+)}<\dfrac{1}{C_{3,k}C_{5,k}}$.

Condition 5: u, ρ, $\mathbf{X}$ and $\mathbf{G}_{(k)}$ satisfy

$$u<\max\left\{C_k^*\left(\rho,\mathbf{X},\mathbf{G}_{(k)}\right),C_k^+\left(\rho,\mathbf{X},\mathbf{G}_{(k)}\right),\frac{c_1}{C_{3,k}C_{4,k}}\right\},$$

where

$$C_k^*\left(\rho,\mathbf{X},\mathbf{G}_{(k)}\right)=c_1\left(\frac{1}{\rho'(c_0\beta_*)\,C_{3,k}}-\frac{C_{5,k}}{\rho'(0+)}\right)\left(\frac{C_{6,k}}{\rho'(0+)}+\frac{C_{4,k}}{\rho'(c_0\beta_*)}\right)^{-1}$$

and

$$C_k^+\left(\rho,\mathbf{X},\mathbf{G}_{(k)}\right)=C_{6,k}^{-1}\left(\kappa_0^{-1}\rho'(0+)\,\lambda_{\min}\left(\mathbf{X}'_{I_1}\mathbf{X}_{I_1}\right)-C_{5,k}c_1\right)$$

with $\kappa_0=\sup\{\kappa(\rho;\nu):\nu\in\mathcal{N}\}$.

Condition 6: λ satisfies $\underline{\lambda}\le\lambda\le\bar{\lambda}$, where

$$\underline{\lambda}=\frac{C_{5,k}c_1+C_{6,k}u}{2^{-1}\Lambda\rho'(0+)}\text{ and }\bar{\lambda}=\frac{1}{2^{-1}\Lambda\rho'(c_0\beta_*)}\left(\frac{c_1}{C_{3,k}}-C_{4,k}u\right).$$

Condition 7: $\left\|\mathbf{T}_{(k)}\mathbf{T}'_{(k)}\mathbf{X}_{I_1}\right\|_\infty \le C_{8,k}$ and $\left\|\mathbf{T}_{(k)}\mathbf{T}'_{(k)}\right\|_\infty \le C_{9,k}$.

It should be noted the similarity between the condition

$$\left\|\mathbf{X}'_{I_0}\mathbf{T}_{(k)}\mathbf{T}'_{(k)}\mathbf{X}_{I_1}\left[\mathbf{X}'_{I_1}\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}'_{(k)}\right)\mathbf{X}_{I_1}\right]^{-1}\right\|_\infty \le C_{3,k}C_{5,6} < \frac{\rho'(0+)}{\rho'(c_0\beta_*)}$$

and the irrepresentable condition of Zhao and Yu (2006) when $\mathbf{w}_{(k)} = \mathbf{0}$ in model (4).

A.2 Proof of Weak Oracle Property of Estimators (Theorem 1)

Write the objective function in (6) as

$$L = L\left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{w}}_{(k)}; \lambda\right) = \sum_{j=1}^m \left(Z_j - \tilde{\mathbf{x}}_j\tilde{\boldsymbol{\beta}} - \sum_{l=1}^k \sqrt{\lambda_l}\gamma_{j,l}\tilde{w}_l\right)^2 + \Lambda\lambda\sum_{j=1}^m \rho\left(\left|\tilde{\beta}_j\right|\right)$$

where $\tilde{\mathbf{x}}_j$ is the jth row of $\mathbf{X} = (x_{ij})$. Then the partial derivatives are

$$\begin{cases} \dfrac{\partial L}{\partial \tilde{w}_l} = -2\sum_{j=1}^m \left(\tilde{Y}_j - \tilde{\mathbf{x}}_j\tilde{\boldsymbol{\beta}}\right)\sqrt{\lambda_l}\gamma_{jl} & \text{for} \quad l = 1, .., k, \\ \dfrac{\partial L}{\partial \tilde{\beta}_j} = -2\sum_{i=1}^m \left(\tilde{Y}_i - \tilde{\mathbf{x}}_i\tilde{\boldsymbol{\beta}}\right)x_{ij} + \Lambda\lambda\,\mathrm{sgn}\left(\tilde{\beta}_j\right)\rho'\left(\left|\tilde{\beta}_j\right|\right) & \text{for} \quad \tilde{\beta}_j \ne 0, \end{cases}$$

where $\tilde{Y}_j = Z_j - \sum_{l=1}^k \sqrt{\lambda_l}\gamma_{j,l}\tilde{w}_l$. Let $\mathbf{G}_{(k)} = \mathbf{T}_{(k)}\sqrt{\mathbf{D}_{(k)}}$. Set

$$\begin{cases} \dfrac{\partial L\left(\tilde{\mu}, \tilde{\mathbf{w}}_{(k)}; \lambda\right)}{\partial \tilde{\boldsymbol{\beta}}} = -2\mathbf{X}'\left(\mathbf{Z} - \mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{G}_{(k)}\tilde{\mathbf{w}}_{(k)}\right) + \Lambda\lambda\tau, \\ \dfrac{\partial L\left(\tilde{\mu}, \tilde{\mathbf{w}}_{(k)}; \lambda\right)}{\partial \tilde{\mathbf{w}}_{(k)}} = -2\mathbf{G}'_{(k)}\left(\mathbf{Z} - \mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{G}_{(k)}\tilde{\mathbf{w}}_{(k)}\right), \end{cases}$$

where $\tau = (\tau_1, ..., \tau_m)'$ with $\tau_j = \tau\left(\tilde{\beta}_j\right)$, and $\tau(x) = \mathrm{sgn}(x)\rho'(|x|)$ for $x \ne 0$ and $\tau(x) \in [-\rho'(0+), \rho'(0+)]$ when $x = 0$.

Necessary conditions. When $\left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{w}}_{(k)}\right) \in \arg\min_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^p, \tilde{\mathbf{w}}_{(k)}\in\mathbb{R}^k} L\left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{w}}_{(k)}; \lambda\right)$, then necessarily $\dfrac{\partial L}{\partial \hat{w}_l} = 0$ for $1 \le l \le k$,

$$\frac{\partial L}{\partial \hat{\beta}_j} = -2\sum_{i=1}^m \left(\hat{Y}_i - \tilde{\mathbf{x}}_i\hat{\boldsymbol{\beta}}\right)x_{ij} + \Lambda\lambda\,\mathrm{sgn}\left(\hat{\beta}_j\right)\rho'\left(\left|\hat{\beta}_j\right|\right) = 0$$

for $j \in \hat{I}_1$, and

$$-\frac{\lambda\Lambda}{2}\rho'(0+) \leq \sum_{i=1}^{m}\left(\tilde{Y}_i - (\tilde{\mathbf{x}}_i)_{\hat{I}_1}\hat{\boldsymbol{\beta}}_{\hat{I}_1}\right)x_{ij} \leq \frac{\lambda\Lambda}{2}\rho'(0+) \text{ for } j \in \hat{I}_0,$$

where $\hat{I}_0 = \left\{i : \hat{\beta}_j = 0\right\}$ and $\hat{I}_1 = \left\{i : \hat{\beta}_j \neq 0\right\}$. In compact form, these conditions are:

$$2\mathbf{G}'_{(k)}\left(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{G}_{(k)}\hat{\mathbf{w}}_{(k)} - \mathbf{Z}\right) = \mathbf{0}, \tag{14}$$

$$\left(\mathbf{X}'\right)^{\hat{I}_1}\left(\mathbf{Z} - \mathbf{G}_{(k)}\hat{\mathbf{w}}_{(k)} - \mathbf{X}_{\hat{I}_1}\hat{\boldsymbol{\beta}}_{\hat{I}_1}\right) = \frac{\lambda\Lambda}{2}\operatorname{sgn}\left(\hat{\boldsymbol{\beta}}_{\hat{I}_1}\right) \circ \rho'\left(\hat{\boldsymbol{\beta}}_{\hat{I}_1}\right), \tag{15}$$

and

$$-\frac{\lambda\Lambda}{2}\rho'(\mathbf{0}+) \leq \left(\mathbf{X}'\right)^{\hat{I}_0}\left(\mathbf{Z} - \mathbf{G}_{(k)}\hat{\mathbf{w}}_{(k)} - \mathbf{X}_{\hat{I}_1}\hat{\boldsymbol{\beta}}_{\hat{I}_1}\right) \leq \frac{\lambda\Lambda}{2}\rho'(\mathbf{0}+). \tag{16}$$

Let $\left|\hat{I}_1\right| = q$ and $\tilde{\mathbb{R}}^q = \left\{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^p : \tilde{\boldsymbol{\beta}}_{\hat{I}_0} = \mathbf{0}\right\}$. By definition of $\left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{w}}_{(k)}\right)$, there exists a ball $\mathcal{B}$ in $\tilde{\mathbb{R}}^q \otimes \mathbb{R}^k$ of small radius centered at $\left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{w}}_{(k)}\right)$ such that $L\left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{w}}_{(k)}; \lambda\right) \geq L\left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{w}}_{(k)}; \lambda\right)$ for any $\left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{w}}_{(k)}\right) \in \mathcal{B}$, where $\otimes$ denotes the Cartesian product with relative topology inherited from $\mathbb{R}^q \otimes \mathbb{R}^k$. Further, $2\lambda_{\min}\left(\mathbf{X}'_{\hat{I}_1}\mathbf{X}_{\hat{I}_1}\right) \geq \lambda\Lambda\kappa\left(\rho; \hat{\boldsymbol{\beta}}_{\hat{I}_1}\right)$, where $\lambda_{\min}$ denotes the smallest eigenvalue of a matrix.

Form of solution. (14) is equivalent to $\mathbf{G}'_{(k)}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{D}_{(k)}\hat{\mathbf{w}}_{(k)} = \mathbf{G}'_{(k)}\mathbf{Z}$, and further to

$$\mathbf{T}'_{(k)}\mathbf{X}_{\hat{I}_1}\hat{\boldsymbol{\beta}}_{\hat{I}_1} + \mathbf{D}^{1/2}_{(k)}\hat{\mathbf{w}}_{(k)} = \mathbf{T}'_{(k)}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{D}^{1/2}_{(k)}\hat{\mathbf{w}}_{(k)} = \mathbf{T}'_{(k)}\mathbf{Z}$$

when $\hat{\boldsymbol{\beta}}_{\hat{I}_0} = \mathbf{0}$, i.e.,

$$\hat{\mathbf{w}}_{(k)} = \mathbf{D}^{-1/2}_{(k)}\mathbf{T}'_{(k)}\left(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}\right). \tag{17}$$

Plugging (17) into the RHS of (15) gives

$$\begin{aligned} & \mathbf{Z} - \mathbf{G}_{(k)}\hat{\mathbf{w}}_{(k)} - \mathbf{X}_{\hat{I}_1}\hat{\boldsymbol{\beta}}_{\hat{I}_1} \\ = \ & \mathbf{Z} - \mathbf{T}_{(k)}\mathbf{T}'_{(k)}\left(\mathbf{Z} - \mathbf{X}_{\hat{I}_1}\hat{\boldsymbol{\beta}}_{\hat{I}_1}\right) - \mathbf{X}_{\hat{I}_1}\hat{\boldsymbol{\beta}}_{\hat{I}_1} = \left(\mathbf{I} - \mathbf{T}_{(k)}\mathbf{T}'_{(k)}\right)\left(\mathbf{Z} - \mathbf{X}_{\hat{I}_1}\hat{\boldsymbol{\beta}}_{\hat{I}_1}\right). \end{aligned}$$

Therefore, (15) becomes

$$\left(\mathbf{X}'\right)^{\hat{I}_1}\left(\mathbf{I} - \mathbf{T}_{(k)}\mathbf{T}'_{(k)}\right)\left(\mathbf{Z} - \mathbf{X}_{\hat{I}_1}\hat{\boldsymbol{\beta}}_{\hat{I}_1}\right) = \frac{\lambda\Lambda}{2}\tau\left(\hat{\boldsymbol{\beta}}_{\hat{I}_1}\right)$$

and (16) changes into

$$2\left(\lambda\Lambda\right)^{-1}\left\|\left(\mathbf{X}'\right)^{\hat{I}_0}\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\right)\left(\mathbf{Z}-\mathbf{X}_{\hat{I}_1}\hat{\boldsymbol{\beta}}_{\hat{I}_1}\right)\right\|_\infty\leq\rho'\left(0+\right).$$

Setting

$$\hat{\tau}=2\left(\lambda\Lambda\right)^{-1}\mathbf{X}'\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\right)\left(\mathbf{Z}-\mathbf{X}_{\hat{I}_1}\hat{\boldsymbol{\beta}}_{\hat{I}_1}\right).$$

Then $\hat{\tau}_{\hat{I}_1}$ is equivalent to (15) and $\left\|\tau_{\hat{I}_0}\right\|_\infty\leq\rho'\left(0+\right)$ is just (16). We can also write

$$\hat{\boldsymbol{\beta}}_{\hat{I}_1}=\mathbf{Q}_{\hat{I}_1}^{-}\left[\left(\mathbf{X}'\right)^{\hat{I}_1}\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\right)\mathbf{Z}-\frac{\lambda\Lambda}{2}\tau\left(\hat{\boldsymbol{\beta}}_{\hat{I}_1}\right)\right],$$

where

$$\mathbf{Q}_{\hat{I}_1}=\mathbf{X}_{\hat{I}_1}'\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\right)\mathbf{X}_{\hat{I}_1}$$

and $^-$ for a matrix denotes its Moore-Penrose inverse.

Existence of a solution with weak oracle property. We want to find a local minimizer of L such that $\left\|\hat{\boldsymbol{\beta}}_{I_1}-\boldsymbol{\beta}_{I_1}\right\|_\infty\leq c_1$ and $\left\|\hat{\mathbf{w}}_{(k)}-\mathbf{w}_{(k)}\right\|_\infty\leq c_2$. For the moment, let us assume $I_1=\hat{I}_1$. Then from (17), (15), we would have,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{I_1} &= \left(\mathbf{X}_{I_1}'\mathbf{X}_{I_1}\right)^{-1}\left[\mathbf{X}_{I_1}'\left(\mathbf{Z}-\mathbf{G}_{(k)}\hat{\mathbf{w}}_{(k)}\right)-2^{-1}\lambda\Lambda\tau\left(\hat{\boldsymbol{\beta}}_{I_1}\right)\right]\\ &= \left(\mathbf{X}_{I_1}'\mathbf{X}_{I_1}\right)^{-1}\mathbf{X}_{I_1}'\mathbf{X}_{I_1}\boldsymbol{\beta}_{I_1}+\left(\mathbf{X}_{I_1}'\mathbf{X}_{I_1}\right)^{-1}\mathbf{X}_{I_1}'\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\mathbf{X}_{I_1}\left(\hat{\boldsymbol{\beta}}_{I_1}-\boldsymbol{\beta}_{I_1}\right)\\ &\quad+\left(\mathbf{X}_{I_1}'\mathbf{X}_{I_1}\right)^{-1}\left[\mathbf{X}_{I_1}'\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\right)\mathbf{v}-2^{-1}\lambda\Lambda\tau\left(\hat{\boldsymbol{\beta}}_{I_1}\right)\right],\end{aligned}$$

using

$$\begin{aligned}&\mathbf{Z}-\mathbf{G}_{(k)}\hat{\mathbf{w}}_{(k)}\\ =\;&\mathbf{X}_{I_1}\boldsymbol{\beta}_{I_1}+\mathbf{G}_{(k)}\mathbf{w}_{(k)}+\mathbf{v}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\left(\mathbf{Z}-\mathbf{X}_{I_1}\hat{\boldsymbol{\beta}}_{I_1}\right)\\ =\;&\mathbf{X}_{I_1}\boldsymbol{\beta}_{I_1}+\mathbf{G}_{(k)}\mathbf{w}_{(k)}+\mathbf{v}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\mathbf{X}_{I_1}\left(\boldsymbol{\beta}_{I_1}-\hat{\boldsymbol{\beta}}_{I_1}\right)-\mathbf{G}_{(k)}\mathbf{w}_{(k)}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\mathbf{v}\\ =\;&\mathbf{X}_{I_1}\boldsymbol{\beta}_{I_1}+\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\mathbf{X}_{I_1}\left(\hat{\boldsymbol{\beta}}_{I_1}-\boldsymbol{\beta}_{I_1}\right)+\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\right)\mathbf{v}\end{aligned}$$

and

$$\begin{aligned}\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\left(\mathbf{Z}-\mathbf{X}_{I_1}\hat{\boldsymbol{\beta}}_{I_1}\right) &= \mathbf{T}_{(k)}\mathbf{T}_{(k)}'\left(\mathbf{X}_{I_1}\boldsymbol{\beta}_{I_1}+\mathbf{G}_{(k)}\mathbf{w}_{(k)}+\mathbf{v}-\mathbf{X}_{I_1}\hat{\boldsymbol{\beta}}_{I_1}\right)\\ &= \mathbf{T}_{(k)}\mathbf{T}_{(k)}'\mathbf{X}_{I_1}\left(\boldsymbol{\beta}_{I_1}-\hat{\boldsymbol{\beta}}_{I_1}\right)+\mathbf{G}_{(k)}\mathbf{w}_{(k)}+\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\mathbf{v}.\end{aligned}$$

Let

$$\mathbf{P}=\mathbf{I}-\left(\mathbf{X}_{I_1}'\mathbf{X}_{I_1}\right)^{-1}\mathbf{X}_{I_1}'\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\mathbf{X}_{I_1}=\left(\mathbf{X}_{I_1}'\mathbf{X}_{I_1}\right)^{-1}\mathbf{X}_{I_1}'\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\right)\mathbf{X}_{I_1}$$

and

$$\boldsymbol{\Upsilon}\left(\hat{\beta}_{I_1},\mathbf{v}\right)=\mathbf{X}_{I_1}'\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\right)\mathbf{v}-2^{-1}\lambda\Lambda\tau\left(\hat{\boldsymbol{\beta}}_{I_1}\right)$$

Then

$$\mathbf{P}\left(\hat{\boldsymbol{\beta}}_{\hat{I}_1}-\boldsymbol{\beta}_{I_1}\right)=\left(\mathbf{X}_{I_1}'\mathbf{X}_{I_1}\right)^{-1}\boldsymbol{\Upsilon}\left(\hat{\boldsymbol{\beta}}_{I_1},\mathbf{v}\right).$$

Further, from (17) and (16), we would have $\left(\mathbf{X}'\right)^{I_0}\left(\mathbf{Z}-\mathbf{G}_{(k)}\hat{\mathbf{w}}_{(k)}-\mathbf{X}_{I_1}\hat{\boldsymbol{\beta}}_{I_1}\right)=\mathbf{X}_{I_0}'\boldsymbol{\Upsilon}_1\left(\hat{\boldsymbol{\beta}}_{I_1},\mathbf{v}\right)$ and

$$-\frac{\lambda\Lambda}{2}\rho'\left(\mathbf{0}+\right)\leq\mathbf{X}_{I_0}'\boldsymbol{\Upsilon}_1\left(\hat{\boldsymbol{\beta}}_{I_1},\mathbf{v}\right)\leq\frac{\lambda\Lambda}{2}\rho'\left(\mathbf{0}+\right),$$

where

$$\boldsymbol{\Upsilon}_1\left(\hat{\boldsymbol{\beta}}_{I_1},\mathbf{v}\right)=\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\mathbf{X}_{I_1}\left(\hat{\boldsymbol{\beta}}_{I_1}-\boldsymbol{\beta}_{I_1}\right)+\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\right)\mathbf{v}.$$

The previous arguments essentially suggest that we only need to show that there is a solution $(\hat{\boldsymbol{\nu}},\hat{\boldsymbol{\varsigma}})\in\mathcal{N}\times\mathcal{N}^*$ such that the following

$$\mathbf{P}\left(\hat{\boldsymbol{\nu}}-\boldsymbol{\beta}_{I_1}\right)=\left(\mathbf{X}_{I_1}'\mathbf{X}_{I_1}\right)^{-1}\boldsymbol{\Upsilon}\left(\hat{\boldsymbol{\nu}},\mathbf{v}\right)\tag{18}$$

$$-\frac{\lambda\Lambda}{2}\rho'\left(\mathbf{0}+\right)\leq\mathbf{X}_{I_0}'\boldsymbol{\Upsilon}_1\left(\hat{\boldsymbol{\nu}},\mathbf{v}\right)\leq\frac{\lambda\Lambda}{2}\rho'\left(\mathbf{0}+\right)\tag{19}$$

$$\hat{\boldsymbol{\varsigma}}=\mathbf{D}_{(k)}^{-1/2}\mathbf{T}_{(k)}'\left(\mathbf{Z}-\mathbf{X}_{I_1}\hat{\boldsymbol{\nu}}\right)\tag{20}$$

$$2\lambda_{\min}\left(\mathbf{X}_{I_1}'\mathbf{X}_{I_1}\right)\geq\Lambda\lambda\kappa\left(\rho;\hat{\boldsymbol{\nu}}\right)\tag{21}$$

hold simultaneously.

Define

$$\Xi^*\left(\boldsymbol{\nu}\right)=\left(\boldsymbol{\beta}_{I_1}-\boldsymbol{\nu}\right)+\delta_{\boldsymbol{\beta},I_1}$$

with component functions $\Psi_i\left(\boldsymbol{\nu}\right)$, $i=1,..,q$ and $\boldsymbol{\nu}=\left(\nu_1,...,\nu_{q_0}\right)'$, where

$$\delta_{\boldsymbol{\mu},I_1}=\mathbf{P}^{-1}\left(\mathbf{X}_{I_1}'\mathbf{X}_{I_1}\right)^{-1}\boldsymbol{\Upsilon}\left(\hat{\boldsymbol{\nu}},\mathbf{v}\right)=\left[\mathbf{X}_{I_1}'\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\right)\mathbf{X}_{I_1}\right]^{-1}\boldsymbol{\Upsilon}\left(\hat{\boldsymbol{\nu}},\mathbf{v}\right).$$

when $\mathbf{M}=\mathbf{X}_{I_1}'\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\right)\mathbf{X}_{I_1}>0$. By Miranda existence theorem (e.g., Vrahatis, 1989), Ξ^* has a zero in $\mathcal{N}$ when $\Xi^*\left(\mathbf{e}\right)\neq\mathbf{0}$ for $\mathbf{e}=\left(e_1,...,e_{q_0}\right)'\in\partial\mathcal{N}$ and that

$$\begin{cases}\Psi_i\left(e_1,...,e_{i-1},e_i,e_{i+1},...,e_{q_0}\right)\leq 0 \text{ for } 1\leq i\leq q_0\\ \Psi_i\left(e_1,...,e_{i-1},\tilde{e}_i,e_{i+1},...,e_{q_0}\right)\geq 0 \text{ for } 1\leq i\leq q_0\end{cases}$$

for e_i such that $e_i - \boldsymbol{\beta}_{I_1,i} = -c_1$ and $\tilde{e}_i - \boldsymbol{\beta}_{I_1,i} = c_1$ when $i = 1, ..., q_0$, respectively.

First, we claim that Ξ^* has a zero in $\mathcal{N}$ on the event

$$\mathcal{E}_2 = \left\{ \max_{1 \le i \le m} \{|v_i|\} \le u \right\} = \{\|\mathbf{v}\|_\infty \le u\}$$

when m is large enough. Clearly,

$$P(\mathcal{E}_2) \le \sum_{i=1}^{m} P(\{a_i |v_i| > a_i u\}) \le \sum_{i=1}^{m} P\left(\left\{a_i |v_i| > a_{(1)} u\right\}\right) \le \frac{m \exp\left(-2^{-1} a_{(1)}^2 u^2\right)}{a_{(1)} u}$$

and $P(\Omega \backslash \mathcal{E}_2) = 1 + o(1)$ when $\left(a_{(1)} u\right)^{-1} m \exp\left(-2^{-1} a_{(1)}^2 u^2\right) \to 0$ as $m \to \infty$. Let $\beta_* = \min_{j \in I_1} |\beta_j|$ and $\boldsymbol{\beta}_{I_1} = \left(\beta_{j_1}, ..., \beta_{j_{q_0}}\right)'$ for $j_l \in I_1$. When $c_1 = (1 - c_0)\beta_*$ for $c_0 \in (0, 1)$ and $\boldsymbol{\nu} = \left(\nu_{j_1}, ..., \nu_{j_{q_0}}\right)' \in \mathcal{N}$,

$$\left\|\boldsymbol{\nu} - \boldsymbol{\beta}_{I_1}\right\|_\infty = \max_{j_l \in I_1} \{|\nu_{j_l} - \beta_{j_l}|\} \ge \max_{1 \le l \le q_0} \{||\nu_{j_l}| - |\beta_{j_l}||\},$$

and $|\nu_{j_l}| \ge -c_1 + |\beta_{j_l}| \ge -c_1 + \beta_* = c_0 \beta_*$. So $\|\mathrm{sgn}(\boldsymbol{\nu}) \circ \rho'(\boldsymbol{\nu})\|_\infty \le \rho'(c_0 \beta_*)$. When

$$\|\mathbf{M}\|_\infty < C_{3,k} \text{ and } \left\|\mathbf{X}_{I_1}' \left(\mathbf{I} - \mathbf{T}_{(k)} \mathbf{T}_{(k)}'\right)\right\|_\infty < C_{4,k},$$

it follows that

$$\begin{aligned} \|\delta_{\boldsymbol{\beta}, I_1}\|_\infty &= \|\mathbf{M}\|_\infty \left[\left\|\mathbf{X}_{I_1}' \left(\mathbf{I} - \mathbf{T}_{(k)} \mathbf{T}_{(k)}'\right)\right\|_\infty \|\mathbf{v}\|_\infty + 2^{-1} \lambda \Lambda \rho'(c_0 \beta_*)\right] \\ &\le C_{3,k} \left(C_{4,k} u + \frac{\lambda \Lambda}{2} \rho'(c_0 \beta_*)\right). \end{aligned}$$

Let $\mathbf{e}_i = (e_1, ..., e_{i-1}, e_i, e_{i+1}, ..., e_{q_0})'$ and $\tilde{\mathbf{e}}_i = (e_1, ..., e_{i-1}, \tilde{e}_i, e_{i+1}, ..., e_{q_0})'$. When $u < \dfrac{c_1}{C_{3,k} C_{4,k}}$ and

$$\lambda \le \bar{\lambda} = \frac{1}{2^{-1} \Lambda \rho'(c_0 \mu_*)} \left(\frac{c_1}{C_{3,k}} - C_{4,k} u\right),$$

we have $\|\delta_{\beta, I_1}\|_\infty < c_1 = (1 - c_0)\mu_*$ and

$$\begin{cases} \Psi_i(\mathbf{e}_i) \le c_1 - \|\delta_{\boldsymbol{\beta}, I_1}\|_\infty < 0, \\ \Psi_i(\tilde{\mathbf{e}}_i) \ge -c_1 + \|\delta_{\boldsymbol{\beta}, I_1}\|_\infty > 0 \end{cases} \tag{22}$$

for $1 \le i \le q_0$. Hence, Ξ^* has a zero in $\mathcal{N}$.

Secondly, we verify that $\hat{\nu}$ satisfies (19). When

$$\left\|\mathbf{X}_{I_0}'\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\mathbf{X}_{I_1}\right\|_\infty \leq C_{5,k} \text{ and } \left\|\mathbf{X}_{I_0}'\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\right)\right\|_\infty \leq C_{6,k},$$

it holds that

$$\begin{aligned}
& \left\|\mathbf{X}_{I_0}'\boldsymbol{\Upsilon}_1\left(\hat{\boldsymbol{\nu}},\mathbf{v}\right)\right\|_\infty \\
\leq & \left\|\mathbf{X}_{I_0}'\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\mathbf{X}_{I_1}\right\|_\infty c_1+\left\|\mathbf{X}_{I_0}'\left(\mathbf{I}-\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\right)\right\|_\infty\|\mathbf{v}\|_\infty \\
\leq & C_{5,k}c_1+C_{6,k}u.
\end{aligned}$$

When

$$\lambda \geq \underline{\lambda}=\frac{C_{5,k}c_1+C_{6,k}u}{2^{-1}\Lambda\rho'\left(0+\right)},$$

$\|\delta_{\boldsymbol{\beta},I_0}\|_\infty \leq \frac{\lambda\Lambda}{2}\rho'\left(0+\right)$ and (19) holds. Therefore, $\lambda\in\left[\underline{\lambda},\bar{\lambda}\right]$ can be chosen. When

$$\frac{\rho'\left(c_0\beta_*\right)}{\rho'\left(0+\right)}<\frac{1}{C_{3,k}C_{5,k}}$$

and $u<C_k^*\left(\rho,\mathbf{X},\mathbf{G}_{(k)}\right)$ where

$$C_k^*\left(\rho,\mathbf{X},\mathbf{G}_{(k)}\right)=c_1\left(\frac{1}{\rho'\left(c_0\beta_*\right)C_{3,k}}-\frac{C_{5,k}}{\rho'\left(0+\right)}\right)\left(\frac{C_{6,k}}{\rho'\left(0+\right)}+\frac{C_{4,k}}{\rho'\left(c_0\beta_*\right)}\right)^{-1},$$

the interval $\left[\underline{\lambda},\bar{\lambda}\right]$ is non-empty with $\underline{\lambda}\geq 0$. Hence $\lambda=\underline{\lambda}$ is well defined.

Thirdly, we verify (21) when $\lambda=\underline{\lambda}$ is set. Since

$$u<C_k^+\left(\rho,\mathbf{X},\mathbf{G}_{(k)}\right):=C_{6,k}^{-1}\left(\frac{\rho'\left(0+\right)\lambda_{\min}\left(\mathbf{X}_{I_1}'\mathbf{X}_{I_1}\right)}{\kappa_0}-C_{5,k}c_1\right)$$

and $\kappa_0=\sup\left\{\kappa\left(\rho;\boldsymbol{\nu}\right):\boldsymbol{\nu}\in\mathcal{N}\right\}$, it follows that

$$\lambda\leq\frac{2\lambda_{\min}\left(\mathbf{X}_{I_1}'\mathbf{X}_{I_1}\right)}{\Lambda\kappa_0}\leq\frac{2\lambda_{\min}\left(\mathbf{X}_{I_1}'\mathbf{X}_{I_1}\right)}{\Lambda\kappa\left(\rho;\hat{\boldsymbol{\nu}}\right)},$$

i.e., (21) holds.

Finally, we bound the difference between $\hat{\boldsymbol{\varsigma}}$ and $\mathbf{w}_{(k)}$. From (20), we see

$$\begin{aligned}
\hat{\boldsymbol{\varsigma}}-\mathbf{w}_{(k)} &= \mathbf{D}_{(k)}^{-1/2}\mathbf{T}_{(k)}'\left(\mathbf{Z}-\mathbf{X}_{I_1}\hat{\boldsymbol{\nu}}\right)-\mathbf{w}_{(k)} \\
&= \mathbf{D}_{(k)}^{-1/2}\mathbf{T}_{(k)}'\mathbf{X}_{I_1}\left(\boldsymbol{\beta}_{I_1}-\hat{\boldsymbol{\nu}}\right)+\mathbf{D}_{(k)}^{-1/2}\mathbf{T}_{(k)}'\mathbf{v}.
\end{aligned}$$

Therefore

$$\begin{aligned}\left\|\hat{\boldsymbol{\varsigma}}-\mathbf{w}_{(k)}\right\|_\infty &\le \left\|\mathbf{D}_{(k)}^{-1/2}\mathbf{T}'_{(k)}\mathbf{X}_{I_1}\left(\boldsymbol{\beta}_{I_1}-\hat{\boldsymbol{\nu}}\right)\right\|_\infty+\left\|\mathbf{D}_{(k)}^{-1/2}\mathbf{T}'_{(k)}\mathbf{v}\right\|_\infty \\ &\le \lambda_k^{-1/2}C_{7,k}c_1+\lambda_k^{-1/2}\left\|\mathbf{T}'_{(k)}\right\|_\infty\|\mathbf{v}\|_\infty \\ &\le \lambda_k^{-1/2}\left(C_{7,k}c_1+C_{8,k}u\right)=c_2\end{aligned}$$

since

$$\left\|\mathbf{T}'_{(k)}\mathbf{X}_{I_1}\right\|_\infty \le C_{7,k} \text{ and } \left\|\mathbf{T}'_{(k)}\right\|_\infty \le C_{8,k}.$$

This completes this proof.

A.3 Proof of Bound on $|\varphi_m(t)-\varphi(t)|$ (Theorem 2)

We directly bound

$$m^{-1}\left|\sum_{j=1}^m\left(\cos\left(t\zeta v_j^*\right)-E\left[\cos\left(t\zeta v_j^*\right)\right]\right)\right|$$

using Markov inequality. Let $\tilde{s}_m(t)=m^{-1}\sum_{j=1}^m\cos\left(t\zeta v_j^*\right)$ and $\tilde{s}(t)=E\left[\tilde{s}_m(t)\right]$. The trick is to transform $var\left(\tilde{s}_m(t)-\tilde{s}(t)\right)$ into a sum of quadratic functions of $\left|\rho_{ij}^{\mathbf{v}}\right|$, and use (5), (12) to bound $var\left(\tilde{s}_m(t)-\tilde{s}(t)\right)$. When

$$P\left(\left\{\left|\tilde{s}_m(t)-\tilde{s}(t)\right|>\tilde{\epsilon}\right\}\right)\le\frac{var\left(\tilde{s}_m(t)-\tilde{s}(t)\right)}{\tilde{\epsilon}^2}$$

and $var\left(\tilde{s}_m(t)-\tilde{s}(t)\right)$ dominates $\tilde{\epsilon}^2$ with certain order as they converge to zero, the assertion will be justified.

Given $\tilde{\epsilon}>0$,

$$\begin{aligned}&var\left(\tilde{s}_m(t)-\tilde{s}(t)\right)\\ =\;& m^{-2}\sum_{i=1}^m var\left(\cos\left(t\zeta v_j^*\right)\right)+2m^{-2}\sum_{1\le i<j\le m}cov\left(\cos\left(t\zeta v_i^*\right),\cos\left(t\zeta v_j^*\right)\right).\end{aligned}$$

From

$$\cos\left(t\zeta v_i^*\right)=\cos\left(t\zeta\mu_i\right)-t\zeta\sin\left(\zeta\tilde{v}_i^*\right)\left(v_i^*-\mu_i\right)$$

where $\tilde{v}_i^*$ is strictly between μ_i and v_i^*, we get

$$\begin{aligned}
& cov\left(\cos\left(t\zeta v_i^*\right), \cos\left(t\zeta v_j^*\right)\right) \\
= \; & cov\left(\cos\left(t\zeta\mu_i\right) - t\zeta\sin\left(\zeta\tilde{v}_i^*\right)\left(v_i^* - \mu_i\right), \cos\left(t\zeta\mu_j\right) - t\zeta\sin\left(\zeta\tilde{v}_j^*\right)\left(v_j^* - \mu_j\right)\right) \\
= \; & t^2\zeta^2 cov\left(\sin\left(\zeta\tilde{v}_i^*\right)\left(v_i^* - \mu_i\right), \sin\left(\zeta\tilde{v}_j^*\right)\left(v_j^* - \mu_j\right)\right) \\
\leq \; & t^2\zeta^2 cov\left(\left|v_i^* - \mu_i\right|, \left|v_j^* - \mu_j\right|\right).
\end{aligned}$$

By Wellner and Smythe (2002), $E\left[a_i^{-1}\left|v_i^* - \mu_i\right| a_j^{-1}\left|v_j^* - \mu_j\right|\right] = \frac{2}{\pi}\tilde{\kappa}\left(\rho_{ij}^{\mathbf{v}}\right)$ where

$$\tilde{\kappa}\left(\rho_{ij}^{\mathbf{v}}\right) = \rho_{ij}^{\mathbf{v}} \arcsin\rho_{ij}^{\mathbf{v}} + \sqrt{1 - \left(\rho_{ij}^{\mathbf{v}}\right)^2}$$

and $E\left[\left|v_i^* - \mu_i\right|\right] = a_i^{-1}\sqrt{\frac{2}{\pi}}$. Noticing additionally that

$$\left|\sqrt{1 - \left(\rho_{ij}^{\mathbf{v}}\right)^2} - 1\right| = 2^{-1}\left(\rho_{ij}^{\mathbf{v}}\right)^2\left(1 - \tilde{\theta}\right)^{-1/2}$$

for some $0 < \tilde{\theta} < \left(\rho_{ij}^{\mathbf{v}}\right)^2$ and $\rho_{ij}^{\mathbf{v}} \leq 1 - \varepsilon_0$ for all $i, j \in S_{\varepsilon_0}^C$, we have

$$\left|\tilde{\kappa}\left(\rho_{ij}^{\mathbf{v}}\right) - 1\right| \leq \frac{\pi}{2}\left|\rho_{ij}^{\mathbf{v}}\right| + \frac{\left(\rho_{ij}^{\mathbf{v}}\right)^2}{2\sqrt{\varepsilon_0\left(2 - \varepsilon_0\right)}}$$

and

$$\begin{aligned}
& \left|cov\left(\cos\left(t\zeta v_i^*\right), \cos\left(t\zeta v_j^*\right)\right)\right| \leq t^2\zeta^2 cov\left(\left|v_i^* - \mu_i\right|, \left|v_j^* - \mu_j\right|\right) \\
= \; & t^2\zeta^2\left(E\left[\left|v_i^* - \mu_i\right|\left|v_j^* - \mu_j\right|\right] - E\left[\left|v_i^* - \mu_i\right|\right]E\left[\left|v_j^* - \mu_j\right|\right]\right) \\
= \; & \frac{2}{\pi}t^2\zeta^2 a_j^{-1}a_i^{-1}\left|\tilde{\kappa}\left(\rho_{ij}^{\mathbf{v}}\right) - 1\right| \leq Mt^2 a_{(1)}^{-2}\left(\left|\rho_{ij}^{\mathbf{v}}\right| + \left|\rho_{ij}^{\mathbf{v}}\right|^2\right).
\end{aligned}$$

Therefore

$$\begin{aligned}
& 2m^{-2}\sum_{1\leq i<j\leq m} cov\left(\cos\left(t\zeta v_i^*\right), \cos\left(t\zeta v_j^*\right)\right) \\
\leq \; & 2M\frac{m\left(m-1\right)}{m^2}t^2\zeta^2 a_{(1)}^{-2}\sum_{1\leq i<j\leq m}\left(\left|\rho_{ij}^{\mathbf{v}}\right| + \left|\rho_{ij}^{\mathbf{v}}\right|^2\right) \leq Mt^2 a_{(1)}^{-2} m^{-\delta}.
\end{aligned}$$

This, in addition with

$$m^{-2}\sum_{i=1}^{m} var\left(\cos\left(t\zeta v_j^*\right)\right) \leq Mm^{-1}t^2 a_{(1)}^{-2},$$

implies

$$P\left(\left\{\left|\tilde{s}_m\left(t\right)-\tilde{s}\left(t\right)\right|>\tilde{\epsilon}\right\}\right) \leq \frac{Mt^2 a_{(1)}^{-2} m^{-\delta}}{\tilde{\epsilon}^2}+\frac{Mm^{-1}t^2 a_{(1)}^{-2}}{\tilde{\epsilon}^2}$$

and

$$\sup_{0\leq t\leq\sqrt{2\gamma\ln m}} P\left(\left\{\left|\tilde{s}_m\left(t\right)-\tilde{s}\left(t\right)\right|>\tilde{\epsilon}\right\}\right) \leq \frac{Ma_{(1)}^{-2}m^{-\delta}\ln m}{\tilde{\epsilon}^2}.$$

Hence, with

$$\left|\varphi_m\left(t\right)-\varphi\left(t\right)\right| \leq \int_0^1 e^{\left(t\zeta a_{(1)}^{-1}\right)^2/2}\omega\left(\zeta\right)m^{-1}\left|\sum_{j=1}^{m}\left[\cos\left(t\zeta v_j^*\right)-E\left[\cos\left(t\zeta v_j^*\right)\right]\right]\right|d\zeta,$$

we have

$$\begin{aligned} & \sup_{0\leq t\leq\sqrt{2\gamma\ln m}}\left|\varphi_m\left(t\right)-\varphi\left(t\right)\right| \\ \leq & \sup_{0\leq t\leq\sqrt{2\gamma\ln m}} 2\tilde{\epsilon}\int_0^1 e^{\left(t\zeta a_{(1)}^{-1}\right)^2/2}\omega\left(\zeta\right)d\zeta \leq M\tilde{\epsilon}\int_0^1 e^{\gamma\zeta^2 a_{(1)}^{-2}\ln m}d\zeta \end{aligned}$$

with probability at least $1-M\tilde{\epsilon}^{-2}a_{(1)}^{-2}m^{-\delta}\ln m$. Noticing that

$$\begin{aligned} \vartheta_m\left(\gamma\right) &= \int_0^1 e^{(\gamma\ln m)\zeta^2 a_{(1)}^{-2}/2}d\zeta \\ & \begin{cases} \leq e^M+o\left(1\right) & \text{if} \quad \limsup_{m\to\infty} a_{(1)}^{-2}\ln m = M, \\ = \dfrac{\exp\left(2^{-1}\gamma a_{(1)}^{-2}\ln m\right)}{\gamma a_{(1)}^{-2}\ln m}\left(1+o\left(1\right)\right) & \text{if} \quad \lim_{m\to\infty} a_{(1)}^{-2}\ln m=\infty, \end{cases} \end{aligned}$$

by properties of the Dawson integral (e.g., Abramowitz and Stegun, 1972), the assertion is proved.

A.4 Proof of (Uniform) Consistency of Plug-in Estimator (Theorem 3)

Let $\varphi_m\left(t;\hat{\mathbf{v}}^*\right)=\hat{\varphi}_m\left(t;\mathbf{v}^*\right)=\hat{\varphi}_m\left(t\right)$. With

$$\hat{\varphi}_m\left(t\right)-\varphi\left(t\right)=\varphi_m\left(t;\hat{\mathbf{v}}^*\right)-\varphi_m\left(t;\mathbf{v}^*\right)+\varphi_m\left(t;\mathbf{v}^*\right)-\varphi\left(t;\mu,m\right)$$

and

$$\frac{\varphi_m(t;\hat{\mathbf{v}}^*)}{\pi_m}-1=\frac{\varphi_m(t;\hat{\mathbf{v}}^*)-\varphi_m(t;\mathbf{v}^*)}{\pi_m}+\frac{\varphi_m(t;\mathbf{v}^*)}{\pi_m}-1,$$

it suffices to bound appropriately $|\varphi_m(t;\hat{\mathbf{v}}^*)-\varphi_m(t;\mathbf{v}^*)|$ and show $\dfrac{\varphi_m(t;\hat{\mathbf{v}}^*)-\varphi_m(t;\mathbf{v}^*)}{\pi_m}=$ $o_P(1)$ since $\limsup_{m\to\infty} a_{(1)}^{-2}\ln m=M$ already implies

$$\sup_{0\leq t\leq\sqrt{2\gamma\ln m}}|\varphi_m(t)-\varphi(t)|\leq M\tilde{\epsilon}(1+o(1))\to 0$$

except on an event with probability at most $M\tilde{\epsilon}^{-2}a_{(1)}^{-2}m^{-\delta}\ln m\to 0$ for $\tilde{\epsilon}\to 0+$.

Even though

$$\hat{v}_j^*=Z_j-\sum\nolimits_{j=1}^k\sqrt{\lambda_j}\gamma_{ij}\hat{w}_j$$

may not still be normal because $\hat{w}_j$ is a penalized estimator, it is normal with variance a_j^{-2} conditional on $\hat{w}_j$ and $\hat{\mu}_j$. So we apply a conditional argument to $\varphi_m(t;\hat{\mathbf{v}}^*)-\varphi_m(t;\mathbf{v}^*)$. Once the claim is justified conditional on each $\hat{w}_j$ and $\hat{\mu}_j$, it also holds unconditionally since the sigma-algebras involved are countably generated. Conditional on $\hat{w}_j$ and $\hat{\mu}_j$,

$$\begin{aligned}\varphi_m(t;\hat{\mathbf{v}}^*)-\varphi_m(t;\mathbf{v}^*) &= \frac{1}{m}\sum_{j=1}^m\left(\kappa_{a_j^{-1}}\left(t;v_j^*\right)-\kappa_{a_j^{-1}}\left(t;\hat{v}_j^*\right)\right)\\ &= m^{-1}\sum_{j=1}^m\int_{-1}^1\omega(\zeta)\,e^{\left(t\zeta a_j^{-1}\right)^2/2}\left[\cos\left(t\zeta v_j^*\right)-\cos\left(t\zeta\hat{v}_j^*\right)\right]d\zeta\\ &= m^{-1}\sum_{j=1}^m\int_{-1}^1\omega(\zeta)\,e^{\left(t\zeta a_{(1)}^{-1}\right)^2/2}t\zeta\sin\left(t\zeta\tilde{v}_j^*\right)\left(v_j^*-\hat{v}_j^*\right)d\zeta\end{aligned}$$

for some $\tilde{v}_j^*$ strictly between v_j^* and $\hat{v}_j^*$. By Theorem 1, there is some

$$\left(\hat{\boldsymbol{\beta}},\hat{\mathbf{w}}_{(k)}\right)\in\underset{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^p,\tilde{\mathbf{w}}\in\mathbb{R}^k}{\arg\min}\;L\left(\tilde{\boldsymbol{\beta}},\tilde{\mathbf{w}};\lambda\right)$$

such that $\left\|\hat{\mathbf{w}}_{(k)}-\mathbf{w}_{(k)}\right\|_\infty\leq c_2$ except on an event with probability at most

$$m\left(a_{(1)}u\right)^{-1}\exp\left(-2^{-1}a_{(1)}^2u^2\right)\to 0.$$

Therefore,

$$\|\mathbf{v}^*-\hat{\mathbf{v}}^*\|_\infty\leq\left\|\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\mathbf{X}_{I_1}\right\|_\infty c_1+\left\|\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\right\|u\leq C_{8,k}c_1+C_{9,k}u$$

and

$$
\begin{aligned}
&\sup_{0\leq t\leq\sqrt{2\gamma\ln m}}\left|\varphi_m\left(t;\hat{\mathbf{v}}^*\right)-\varphi_m\left(t;\mathbf{v}^*\right)\right| \\
\leq\quad & \sup_{0\leq t\leq\sqrt{2\gamma\ln m}} m^{-1}\sum_{j=1}^{m}\int_{-1}^{1}\omega\left(\zeta\right)e^{\left(t\zeta a_{(1)}^{-1}\right)^2/2}t\left|v_j^*-\hat{v}_j^*\right|d\zeta \\
\leq\quad & M\sqrt{2\gamma\ln m}\left\|\mathbf{v}^*-\hat{\mathbf{v}}^*\right\|_\infty\leq M\sqrt{2\gamma\ln m}\left(C_{8,k}c_1+C_{9,k}u\right).
\end{aligned}
$$

Let $\theta_m^+=\sqrt{2\gamma\ln m}\left(C_{8,k}c_1+C_{9,k}u\right)$. When $\theta_m^+=o\left(\pi_m\right)$,

$$
\frac{\sup_{0\leq t\leq\sqrt{2\gamma\ln m}}\left|\varphi_m\left(t;\hat{\mathbf{v}}^*\right)-\varphi_m\left(t;\mathbf{v}^*\right)\right|}{\pi_m}=o_P\left(1\right)
$$

and

$$
\sup_{0\leq t\leq\sqrt{2\gamma\ln m}}\left|\frac{\varphi_m\left(t;\hat{\mathbf{v}}^*\right)}{\pi_m}-1\right|=o_P\left(1\right).
$$

When $\theta_m^+=o\left(m^{1-\gamma}\right)$, we have

$$
\frac{\sup_{\Theta_m(\gamma,C_m)}\sup_{0\leq t\leq\sqrt{2\gamma\ln m}}\left|\varphi_m\left(t;\hat{\mathbf{v}}^*\right)-\varphi_m\left(t;\mathbf{v}^*\right)\right|}{\pi_m}=o_P\left(1\right)
$$

and

$$
\sup_{\Theta_m(\gamma,C_m)}\sup_{0\leq t\leq\sqrt{2\gamma\ln m}}\left|\frac{\varphi_m\left(t;\hat{\mathbf{v}}^*\right)}{\pi_m}-1\right|=o_P\left(1\right).
$$

This completes the proof.

References

Abramowitz, M. and I. A. Stegun (1972). *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables* (9 ed.). Dover books on mathematics. New York: Dover.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B 57*(1), 289–300.

Benjamini, Y., A. M. Krieger, and D. Yekutieli (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika 93*(3), 491–507.

Bezdek, J. C. and R. J. Hathaway (2002). Some notes on alternating optimization. *Lecture Notes in Computer Science*.

Blanchard, G. and E. Roquain (2009). Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research 10*, 2837–2871.

Efron, B. (2008). Microarrays, empirical bayes, and the two-groups model. *Statistical Science 23*, 1–22.

Efron, B. (2010). Correlated z -values and the accuracy of large-scale statistical estimates. *J. Amer. Statist. Assoc. 105*(491), 1042–1055.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc. 96*(456), 1151–1160.

Fan, J., X. Han, and W. Gu (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc. To appear.*

Fan, J. and J. Lv (2011). Non-concave penalized likelihood with np-dimensionality. *IEEE Trans Inf Theory 57*(8), 5467–5484.

Friguet, C. and D. Causeur (2011). Estimation of the proportion of true null hypotheses in high-dimensional data under dependence. *Computational Statistics and Data Analysis 55*(9), 2665–2676.

Friguet, C. and D. Kloareg, M.and Causeur (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc. 104*(488), 1406–1415.

Genovese, C. and L. Wasserman (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Statist. Soc. B 64*(3), 499–517.

Genovese, C. and L. Wasserman (2004). A stochastic process approach to false discovery control. *Ann. Statist. 32*(3), 1035–1061.

Ghosal, S. and A. Roy (2011). Predicting false discovery proportion under dependence. *J. Amer. Statist. Assoc. 106*(495), 1208–1218.

Jin, J. (2006). Propotions of nonzero normal means: universal oracle equivalence and uniformly consistent estimations. *Technical report. Department of statistics, Purdue University, West Lafayette*.

Jin, J. (2008). Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *J. R. Statist. Soc. B 70*(3), 461–493.

Jin, J. and T. T. Cai (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc. 102*(478), 495–506.

Langaas, M., L. B. Henery, and E. Ferkingstad (2005). Estimating the proportion of true null hypotheses, with application to dna microarray data. *J. R. Statist. Soc. B 67*(4), 555–572.

Liang, C.-L., J. A. Rice, I. d. Pater, C. Alcock, C. Alcock, A. Wang, and S. Marshall (2004). Statistical methods for detecting stellar occultations by kuiper belt objects: The taiwanese-american occultation survey. *Statist. Sci. 19*(2), 265–274.

Lv, J. and Y. Fan (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist. 37*(6A), 3498–3528.

Mazumder, R., J. H. Friedman, and T. Hastie (2011). parsenet: Coordinate descent with nonconvex penalties. *J. Amer. Statist. Assoc. 106*(495), 1125–1138.

Pacifico, M. P., C. Genovese, I. Verdinelli, and L. Wasserman (2004). False discovery control for random fields. *J. Amer. Statist. Assoc. 99*(468), 1002–1014.

Schwartzman, A. (2008). Empirical null and false discovery rate inference for exponential families. *Ann. Appl. Stat 2*(4), 1332–1359.

Schweder, T. and E. Spjøtvoll (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika 69*(3), 493–502.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation in simultaneous conservative consistency of false discover rates: a unified approach. *J. R. Statist. Soc. B 66*(1), 187–205.

Vrahatis, M. N. (1989). A short proof and a generalization of mirandas existence theorem. *Proc. Amer. Math. Soc. 107*(3), 701–703.

Wang, H.-Q., L. K. Tuominen, and C.-J. Tsai (2011). Slim: A sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics 27*, 225–231.

Wellner, J. A. and R. T. Smythe (2002). Computing the covariance of two brownian area integrals. *Statistica Neerlandica 56*(1), 101–109.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist. 38*(2), 894–942.

Zhang, Y. and J. S. Liu (2011). Fast and accurate approximation to significance tests in genome-wide association studies. *J. Amer. Statist. Assoc. 106*(495), 846–857.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *J. Mach. Learn. Res. 7*(11), 2541–2563.