Supplementary Document for "MWPCR: Multiscale Weighted Principal Component Regression for High-dimensional Prediction"

1 Simulation Studies: Continuous Response

In this subsection, we apply MWPCR to the high-dimensional prediction problem. We are interested in predicting univariate continuous responses y_i by using $20 \times 20 \times 10$ image data $\mathbf{x}_i = \{\mathbf{x}_{i,\mathbf{g}} : \mathbf{g} \in \mathcal{G}\}$. We generated the 3D-images $\mathbf{x}_i$ as follows:

$$\mathbf{x}_{i,\mathbf{g}} = \beta_0(\mathbf{g}) l_i + \epsilon_i(\mathbf{g}) \quad \text{for} \quad i = 1, \cdots, n = 50, \tag{1}$$

where $l_i = 1 + 0.01(i-1)$ and $\boldsymbol{\beta}_0$ is the same as the image in the left panel of Figure (1). We set $\mathbf{y}_i = l_i$ for $i = 1, \ldots, 50$. In this case, we have $n = 50$ and $p = 4,000$.

We consider three types of noise $\epsilon_i(\mathbf{g})$ in (1). First, $\epsilon_i^{(1)}(\mathbf{g})$ were independently generated from a $N(0, 2^2)$ generator across all voxels. Second, $\epsilon_i^{(2)}(\mathbf{g}) = \sum_{\|\mathbf{g}'-\mathbf{g}\|_1 \leq 1} \epsilon_i^{(1)}(\mathbf{g}')/m_{\mathbf{g}}$ were generated from $\epsilon_i^{(1)}(\mathbf{g})$ by introducing the short range spatial correlation, where $\| \cdot \|_1$ is the L_1 norm of a vector and $m_{\mathbf{g}}$ is the number of locations in the set $\{\| \mathbf{g}' - \mathbf{g} \|_1 \leq 1\}$. Third, to introduce the long range spatial correlation, $\epsilon_i^{(3)}(\mathbf{g})$ were generated according to $\epsilon_i^{(3)}(\mathbf{g}) = 2\sin(\pi g_1/10)\xi_{i,1} + 2\cos(\pi g_2/10)\xi_{i,2} + 2\sin(\pi g_3/5)\xi_{i,3} + \epsilon_i^1(\mathbf{g})$, where $\mathbf{g} = (g_1, g_2, g_3)^T$ and $\xi_{i,k}$ for $k = 1, 2, 3$ were independently generated from a $N(0, 1)$ generator. Moreover, the noise variances in all voxels of the red cuboid region equal 4, 4/6, and $4\{\sin(\pi g_1/10)^2 + \cos(\pi g_2/10)^2 + \sin(\pi g_3/5)^2\} + 4$ for Type I, II, and III noises, respectively. Therefore, among the three types of noise, Type III noise has the smallest signal-to-noise ratio and Type II noise has the largest one.

We ran the three stages of MWPCR for the second set of simulations as follows. In Stage 1, we fitted the same linear model as (1) for $f(\mathbf{x}_{i,\mathbf{g}}|\mathbf{y}_i, \boldsymbol{\beta}(\mathbf{g}))$, in which $\mathbf{z}_i$ was dropped out. Then, W_I is calculated based on the p-value of Wald test associated with the correlation between $\mathbf{x}_{i,\mathbf{g}}$ and $\mathbf{y}_i$ at each voxel $\mathbf{g}$. Similar to the first set of simulations, MWPCR1, MWPCR2, and MWPCR3, respectively, correspond to the location kernel $K_1(.)$, the similarity kernel $K_2(.)$, and the combination of kernels $K_1(.)$ and $K_2(.)$. In Stage 2, we tried different numbers of principal components of PCA to reconstruct the low dimensional latent variables $\{\mathbf{u}_{k,i}\}_{k \leq K}$. The analysis results are very robust and we just report the results corresponding to 5, 7 and 10 principal components (PCs) in Figure

2. In Stage 3, we fitted a linear latent variable regression given by $\mathbf{y}_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k \mathbf{u}_{k,i} + \epsilon_i$ to do prediction.

We compared MWPCR and three other dimensional reduction methods including PCA, weighted PCA (WPCA) (Skocaj et al., 2007), and supervised PCA (SPCA) (Bair et al., 2006). We used the leave-one-out cross validation method to compute the prediction errors of all methods. Let $\hat{\mathbf{y}}_i$ be the fitted response value based on the linear latent variable regression, the prediction error is defined as $|\hat{\mathbf{y}}_i - \mathbf{y}_i|/|\mathbf{y}_i|$. Subsequently, we calculated the prediction error difference between MWPCR and all other three methods across different types of noise and different numbers of principal components. Figure 2 presents the box plots of the prediction error differences under different scenarios. These simulation results confirm that MWPCR outperforms all other methods for different types of noise and different numbers of PCs. Figure 3 reports some further results based on the variance thresholding. The green, red and blue curves in Figure 3, respectively, represent the first, second, and third quantiles of the error differences between MWPCR and the three other methods as the variance thresholding increases. These results further show that MWPCR outperforms all other methods.

We compared MWPCR with four other high-dimensional regression methods, including penalized regression (PR) (Tibshirani, 1996), sure independence screening (SIS) regression (Fan and Lv, 2008), support vector regression (SVR) (Basak et al., 2007), and SPLS (Chun and Keles, 2010). We used the software packages SLEP (Liu et al., 2009), SIS (Fan et al., 2010), LIBSVM (Chang and Lin, 2011), and spls (Chung et al., 2012) to run PR, SIS regression, SVR and SPLS, respectively. Figure 4 shows the boxplots of the prediction error difference between MWPCR and all the other regression methods under the three types of noise. The prediction error differences are almost always less than 0 (under the dashed line), indicating that MWPCR outperforms PR, SIS, SVR, and SPLS.

2 Theoretical Properties for Binary Classification

In this section, we theoretically compare MWPCR with standard and supervised PCAs, which do not incorporate the spatial and important score weights, in a high-dimensional binary classification problem. Our results shed some new insights on MWPCR in such problem.

2.1 Setup

We introduce several notation that will be used in the following context. Consider two sequences of constant values $\{a_n : n = 1, \ldots, \infty\}$ and $\{b_n : n = 1, \ldots, \infty\}$.

- Denote $a_n \gg b_n$ if $\lim_{n\to\infty} a_n/b_n = 0$.
- Denote $a_n \asymp b_n$ if $c_2 \leq \underline{\lim}_{n\to\infty} a_n/b_n \leq \overline{\lim}_{n\to\infty} a_n/b_n \leq c_1$ for two constants $c_1 \geq c_2 > 0$.

Consider a binary classification problem with $(\mathbf{x}_i, \mathbf{y}_i)$, where $\mathbf{y}_i = 0, 1$ is the class label. Without loss of generality, it is assumed that $\mathbf{y}_i = 0$ for $i = 1, \ldots, n_1$ and $\mathbf{y}_i = 1$ for $i = n_1 + 1, \ldots, n$ and $\mathbf{x}_i|\mathbf{y}_i \sim N\left(\boldsymbol{\mu}_{\mathbf{y}_i}, \Sigma\right)$. Let $\rho_{n,1} = n_1/n$ and $\overline{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{x}_i/n$. We have

$$\overline{\mathbf{x}} \sim N\left(\boldsymbol{\mu}, n^{-1}\Sigma\right), \qquad E(\overline{\mathbf{x}}) = \boldsymbol{\mu} = \rho_{n,1}\boldsymbol{\mu}_0 + (1 - \rho_{n,1})\boldsymbol{\mu}_1,$$
$$\mathbf{x}_i - \boldsymbol{\mu}|\mathbf{y}_i = 0 \sim N\left((1 - \rho_{n,1})(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1), \Sigma\right), \quad \mathbf{x}_i - \boldsymbol{\mu}|\mathbf{y}_i = 1 \sim N\left(\rho_{n,1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \Sigma\right).$$

2.2 Theory Under an Ideal Scenario

In this subsection, we assume that Σ, $\boldsymbol{\mu}_0$, and $\boldsymbol{\mu}_1$ are known and investigate the effect of applying the spatial weight and importance score weight matrices on PCA and its variants for classification. We consider the spectral decomposition of $\Sigma = VDV^T$, where D is the diagonal matrix of the population eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ and $V = [\mathbf{v}_1, \cdots, \mathbf{v}_p]$ is the matrix of corresponding population eigenvectors. Let $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ for any vector or matrix $\mathbf{a}$ and $||\cdot||_2$ be the Euclidean norm of a vector or matrix. In addition, we assume a multiple component spike model (Paul, 2007;

Jung and Marron, 2009) such that as $n \to \infty$, the eigenvalues satisfy

$$\infty > \lambda_1 > \lambda_2 > \ldots > \lambda_m \gg \lambda_{m+1} \to \cdots \to \lambda_p \to \sigma^2, \tag{2}$$

where m is a finite positive integer and σ is a positive constant. We define a signal set and a noise set, which are, respectively, denoted by $\mathbb{S}$ and $\mathbb{S}^{\perp}$, as follows:

$$\mathbb{S} = \text{span}\{\mathbf{v}_1, \cdots, \mathbf{v}_m\} \text{ and } \mathbb{S}^{\perp} = \text{span}\{\mathbf{v}_{m+1}, \cdots, \mathbf{v}_p\}. \tag{3}$$

We consider three different spatial weight matrices including a precision matrix, a kernel matrix, and a selection matrix. We state the following theorems.

The expectation of $n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T$ is equal to $\Sigma_O = \Sigma + \rho_{n,1}(1-\rho_{n,1})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^{\otimes 2}$. Similarly, for any weight matrix $Q^{(\ell)}$, the expectation of $Q^{(\ell)T} n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T Q^{(\ell)}$ is given by

$$\Sigma_E = Q^{(\ell)T}\Sigma Q^{(\ell)} + \rho_{n,1}(1-\rho_{n,1})Q^{(\ell)T}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^{\otimes 2}Q^{(\ell)}. \tag{4}$$

If we set $Q^{(\ell)} = \Sigma^{-1/2}$, then (4) reduces to

$$\Sigma_{E(1)} = I_p + \rho_{n,1}(1-\rho_{n,1})\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^{\otimes 2}\Sigma^{-1/2}. \tag{5}$$

We obtain the following results about the eigenvalue-eigenvector pairs of Σ_O and $\Sigma_{E(1)}$ and their impact on PCA as follows.

Theorem 1 *We have the following results:*

(a) Under Assumptions (2) and (3), if $\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \in \mathbb{S}^{\perp}$ *and* $||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1||_2^2 << \lambda_m$*, then the first* m *eigenvalue-eigenvector pairs of* Σ_O *are the same as those of* Σ*.*

(b) The eigenvalues and eigenvectors of $\Sigma_{E(1)}$ *are, respectively, given by*

$$\lambda_{E(1),1} = 1 + \rho_{n,1}(1-\rho_{n,1})||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2^2 > \lambda_{E(1),2} = \cdots = \lambda_{E(1),p} = 1, \tag{6}$$
$$\mathbf{v}_{E(1),1} = \Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)/||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2, \qquad \mathbf{v}_{E(1),l} \in \{\mathbf{v}_{E(1),1}\}^{\perp} \quad \textit{for} \;\; l \geq 2.$$

(c) Except for a constant, the extracted first principal component $(\mathbf{x}_i - \boldsymbol{\mu})^T\Sigma^{-1/2}\mathbf{v}_{E(1),1}$ *corresponds to the Bayesian optimal classifier and* $(\mathbf{x}_i - \boldsymbol{\mu})^T\Sigma^{-1/2}\mathbf{v}_{E(1),1}|\mathbf{y}_i \sim N((\lambda - 1 + \mathbf{y}_i)||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2, 1)$*. The misclassification rate of the Fisher discriminant based on the extracted first principal component is equal to* $1 - \Phi(||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2/2)$*, where* $\Phi(\cdot)$ *is the distribution function of the standard normal.*

Theorem 1 has several important implications on PCA and its variants. Theorem 1 (a) indicates that if the signal $||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1||_2^2$ is relatively weak and $\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ is orthogonal to $\mathbb{S}$, then the first m principal components of Σ_O do not contain useful information for better classification. Thus, PCA may produce misleading feature selection and inferior classification under the presence of strong correlations in $\mathbf{X}$. Similar comments are also correct, when one standardizes $\mathbf{X}$ before applying PCA. Theorem 1 (b) indicates that without noise, the normalized $\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ is always the most important direction selected by MWPCR when $Q^{(\ell)} = \Sigma^{-1/2}$. Furthermore, if the signal $||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2$ is very strong, we expect that the estimated largest eigenvector of $\Sigma_{E(1)}$ is approximately parallel to $\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$. Theorem 1 (c) indicates that the use of $Q^{(\ell)} = \Sigma^{-1/2}$ in MWPCR can substantially improve classification accuracy.

We consider the effect of applying a kernel matrix. Specifically, as discussed in Section 2, we may construct a spatial kernel matrix W_E by using either the local spatial weights or the weighted adjacency matrix in order to reduce noise in data (Buja et al., 1989). Therefore, for the binary classification problem, it is assumed that W_E satisfies

$$W_E\Gamma(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = \rho_0\Gamma(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \tag{7}$$

where ρ_0 is a known scalar and satisfies $0 < c_0 \leq |\rho_0| \leq 1$ and Γ is a pre-specified weight matrix, such as $\Sigma^{-1/2}$ considered above. Two trivial choices of W_E are I_p and $\Gamma(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^{\otimes 2}/||\Gamma(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2^2$. For the multiple component spike model, Σ can be approximated by $\tilde{V}\tilde{D}\tilde{V}^T + \sigma^2 I_p$, where $\tilde{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_m]$ is a $p \times m$ submatrix of Σ and $\tilde{D} = \text{diag}(\lambda_1 - \sigma^2, \cdots, \lambda_m - \sigma^2)$. In this case, we apply W_E to $\Gamma\Sigma_O\Gamma^T$ to get

$$\Sigma_{E(2)} = W_E\Gamma(\tilde{V}\tilde{D}\tilde{V}^T + \sigma^2 I_p)\Gamma^T W_E^T + \rho_{n,1}(1 - \rho_{n,1})\rho_0^2\{\Gamma(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\}^{\otimes 2}. \tag{8}$$

The eigenvalues of W_E usually satisfy three properties. First, all eigenvalues of W_E are smaller than or equal to one. For the locally spatial weight matrix, W_E can be regarded as the transition probability matrix of a Markov chain with p states, whose eigenvalues are not bigger than one. Moreover, all eigenvalues of $W_E = \exp(-0.5L/\gamma)$ based on the Laplace-Beltrami operator are not bigger than one, since $L = W_D - W$ is nonnegative definite. Second, when the kernels $K_1(t)$ and $K_2(t)$ are differentially continuous functions of t, the ordered eigenvalues of W_E usually decay in polynomial rates (Little and Reade, 1984; Reade, 1984). Moreover, since ρ_0 is usually close to one, $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ should be the eigenvector corresponding to one of the largest eigenvalues of W_E. Third, for $\Gamma = I_p$, the elements of $W_E W_E^T$ converge to zero for most smooth kernels (Buja et al., 1989), whereas those of $W_E\tilde{V}\tilde{D}\tilde{V}^T W_E^T$ do not change too much. This observation is also important as we set $\Gamma = \Sigma^{-1/2}$. In this case, $W_E\Gamma\Sigma\Gamma^T W_E^T$ reduces to $W_E W_E^T$, whose elements converge to zero.

Let $\{(\lambda_{w,k}^2, \mathbf{v}_{w,k}) : k = 1, \ldots, p\}$ be the eigenvalue-eigenvector pairs of $W_E W_E^T$. It is assumed that as $\min(n, p) \to \infty$, we have

$$1 \geq \lambda_{w,1}^2 \geq \cdots \geq \lambda_{w,m_w}^2 \gg \lambda_{w,m_w+1}^2 \to \cdots \to \lambda_{w,p}^2 \to 0, \tag{9}$$

where m_w is a positive integer. As discussed in Buja et al. (1989), most linear smoothers (e.g., spline) satisfy Assumption (9). For instance, $\lambda_{w,j}$s are either 0 or 1 only for many linear smoothers, such as bin smoother.

Let $\mathcal{F}$ be the linear space spanned by $W_E\mathbf{v}_1, \cdots, W_E\mathbf{v}_m$, $\mathbf{v}_{w,1}, \cdots, \mathbf{v}_{w,m_w}$, and $W_E(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$.

We obtain the following results for $\Sigma_{E(2)}$.

Theorem 2 *Under Assumptions* (2), (3), (7), *and* (9), *we have the following results:*

(a) For any $\mathbf{v} \perp \Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ *and* $\Gamma = \Sigma^{-1/2}$, *if* $\rho_{n,1}(1-\rho_{n,1})\rho_0^2||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2^2$ *is larger than* 1, *then we have*

$$\frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\Sigma^{-1/2}\Sigma_{E(2)}\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}{||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2^2} \geq \rho_{n,1}(1-\rho_{n,1})\rho_0^2||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2^2 > \lambda_{w,1}^2 \geq \frac{\mathbf{v}^T\Sigma_{E(2)}\mathbf{v}}{||\mathbf{v}||_2^2}.$$

If W_E *is symmetric and* $\lambda_{w,k_0}^2 = \rho_o^2$, *then the eigenvalues of* $\Sigma_{E(2)}$ *and their corresponding eigenvectors are, respectively, given by*

$$\lambda_{w,1}^2, \cdots, \lambda_{w,k_0-1}^2, \lambda_{w,k_0}^2 + \rho_{n,1}(1-\rho_{n,1})\rho_0^2||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2^2, \lambda_{w,k_0+1}^2, \cdots, \lambda_{w,p}^2$$
$$\mathbf{v}_{w,1}, \cdots, \mathbf{v}_{w,k_0-1}, \Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)/||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2, \mathbf{v}_{w,k_0+1}, \cdots, \mathbf{v}_{w,p}.$$

(b) For $\Gamma = I_p$, *if* $\rho_{n,1}(1-\rho_{n,1})\rho_0^2||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2^2$ *is larger than* $\sigma^2\lambda_{w,m_w+1}^2$, *then*

$$\frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\Sigma_{E(2)}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}{||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2^2} \geq \rho_{n,1}(1-\rho_{n,1})\rho_0^2||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2^2 > \sigma^2\lambda_{w,m_w+1}^2 \geq \frac{\mathbf{v}^T\Sigma_{E(2)}\mathbf{v}}{||\mathbf{v}||_2^2}.$$

holds for any $\mathbf{v} \perp \mathcal{F}$. *Moreover, if* $\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \perp W_E\mathbf{v}_j$ *holds for all* $j \leq m$ *and* W_E *is symmetric, then* $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)/||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1||_2$ *is the eigenvector of* $\Sigma_{E(2)}$ *corresponding to the eigenvalue* $\rho_0^2\sigma^2 + \rho_{n,1}(1-\rho_{n,1})\rho_0^2||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2^2$.

Theorem 2 has several important implications on PCA and its variants. Theorem 2 (a) reveals a key advantage of applying both W_E and $\Sigma^{-1/2}$. If $\rho_0 = 1$, then the optimal direction $\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)/||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2$ corresponds to the largest eigenvalue of $\Sigma_{E(2)}$. Compared with the eigenvalues in Theorem 1 (b), the ordered eigenvalues of $\Sigma_{E(2)}$ decay much faster to zero. Therefore, when the signal $||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2^2$ is relatively weak, applying both kernel matrix and covariance matrix can outperform solely applying covariance matrix in MWPCR. Theorem 2 (b) also reveals the key advantage of applying W_E. Specifically, the use of W_E can down-

weight all vectors that are orthogonal to $\mathcal{F}$ and $||W_E \mathbf{v}_j||_2 \leq 1$ holds for all j; thus, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ can easily appear in the space spanned by the leading eigenvectors of $\Sigma_{E(2)}$. Furthermore, if $\rho_{n,1}(1-\rho_{n,1})\rho_0^2||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2^2 >> \max(\lambda_1, 1)$, then $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)/||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1||_2$ is the first eigenvector of $\Sigma_{E(2)}$.

We examine the effect of applying the selection weight matrix $Q_1^{(\ell)}$ to either $\widetilde{\mathbf{X}}$ or $\widetilde{\mathbf{X}}\Sigma^{-1/2}$, where $\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n \mu^T$ is the centered data matrix. We consider two different sets, including a discriminative set $S_1 = \{\mathbf{g} \in \mathcal{G} : \{\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\}_{(\mathbf{g})} \neq 0\}$ and a signal set $S_2 = \{\mathbf{g} \in \mathcal{G} : \{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\}_{(\mathbf{g})} \neq 0\}$, where $\{\cdot\}_{(g)}$ denotes the $g-$th component of a vector. If Σ is diagonal, then we have $S_1 = S_2$. However, similar to the arguments in Mai et al. (2012), we can show that S_1 and S_2 can be very different for correlated features. Without loss of generality, for $k = 1, 2$, it is assumed that S_k contains the first p_{S_k} indices. We can partition $\Sigma^{1/2}$ and $\boldsymbol{\mu}_l$ according to S_1 and S_2 as follows:

$$\boldsymbol{\mu}_l = \begin{pmatrix} \boldsymbol{\mu}_{l,S_1} \\ \boldsymbol{\mu}_{l,S_1^c} \end{pmatrix}, \quad \boldsymbol{\mu}_l = \begin{pmatrix} \boldsymbol{\mu}_{l,S_2} \\ \boldsymbol{\mu}_{l,S_2^c} \end{pmatrix}, \quad \Sigma^{1/2} = \begin{pmatrix} \Sigma_{S_1 S_1} & \Sigma_{S_1 S_1^c} \\ \Sigma_{S_1^c S_1} & \Sigma_{S_1^c S_1^c} \end{pmatrix}, \quad \Sigma^{1/2} = \begin{pmatrix} \Sigma_{S_2 S_2} & \Sigma_{S_2 S_2^c} \\ \Sigma_{S_2^c S_2} & \Sigma_{S_2^c S_2^c} \end{pmatrix}$$

for $l = 0, 1$, where S_l^c is the complement of S_l for $l = 1, 2$. Let $\Sigma_{E(3)}$ be the expectation of $(I_{p_{S_k}}, \mathbf{0})\Gamma\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T\Gamma^T(I_{p_{S_k}}, \mathbf{0})^T$. We obtain the following results.

Theorem 3 *We have the following results:*

(a) $S_1 \subset S_2$ if and only if $\Sigma_{S_2^c S_2}\Sigma_{S_2 S_2}^{-1}(\boldsymbol{\mu}_{1,S_2} - \boldsymbol{\mu}_{0,S_2}) = \mathbf{0}$.

(b) $S_2 \subset S_1$ if and only if $\boldsymbol{\mu}_{1,S_1^c} = \boldsymbol{\mu}_{0,S_1^c}$ or $\Sigma_{S_1^c S_1}\Sigma_{S_1 S_1}^{-1}(\boldsymbol{\mu}_{1,S_1} - \boldsymbol{\mu}_{0,S_1}) = \mathbf{0}$.

(c) The eigenvalues and eigenvectors of $\Sigma_{E(3)}$ corresponding to $\Gamma = \Sigma^{-1/2}$ are given by

$$\lambda_{I(k),1} = 1 + \rho_{n,1}(1-\rho_{n,1})||(I_{p_{S_k}}, \mathbf{0})\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2^2 > \lambda_{I,2} = \cdots = \lambda_{I,p_{S_k}} = 1,$$
$$\mathbf{v}_{I(k),1} = (I_{p_{S_k}}, \mathbf{0})\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)/||(I_{p_{S_k}}, \mathbf{0})\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2, \quad \mathbf{v}_{I(k),l} \in \{\mathbf{v}_{I(k),1}\}^{\perp} \quad \textit{for} \quad l \geq 2.$$

(d) For $Q^{(\ell)} = \Sigma^{-1/2}(I_{p_{S_k}}, \mathbf{0})^T$, the misclassification rate of the Fisher discriminant based on the extracted first principal component is equal to $1 - \Phi\left(||(I_{p_{S_k}}, \mathbf{0})\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2/2\right)$. For the discriminative set, we can get the optimal Bayesian classifier.

(e) For $\Gamma = I_p$, if $\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \in \mathbb{S}^{\perp}$ and Assumptions (2) and (3) hold, then the eigenvalues of $\Sigma_{E(3)}$ are given by $\lambda_1 c_1, \cdots, \lambda_m c_m, \sigma^2 + \rho_{n,1}(1-\rho_{n,1})||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2^2, \sigma^2, \cdots, \sigma^2$, where $c_1, \cdots, c_m$ are m nonnegative scalars less than or equal to 1 and will be introduced in the supplementary document. Moreover, $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)/||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1||_2$ is the eigenvector of $\Sigma_{E(3)}$ corresponding to the eigenvalue $\sigma^2 + \rho_{n,1}(1-\rho_{n,1})||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2^2$.

Theorem 3 has several important implications on PCA and its variants. Theorem 3 (a) and (b) are direct generalizations of Proposition 1 of Mai et al. (2012) for correlated features. Theorem 3 (c) indicates that the normalized $(I_{p_{S_k}}, \mathbf{0})\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ is always the most important direction selected by MWPCR when $Q^{(\ell)} = \Sigma^{-1/2}(I_{p_{S_k}}, \mathbf{0})^T$. Theorem 3 (d) quantifies the misclassification rate of the Fisher discriminant based on the extracted first principal component. In most high-dimensional problems, it is much easier to approximate the signal set compared with the discriminative set. Theorem 3 (e) indicates that the use of $(I_{p_{S_2}}, \mathbf{0})$ can dramatically improve the reconstruction of $(\boldsymbol{\mu}_{0,S_2} - \boldsymbol{\mu}_{1,S_2})$ in a very challenging scenario discussed in Theorem 1 (a). Specifically, λ_j is reduced to $c_j\lambda_j$ for all $j = 1, \ldots, m$, whereas the eigenvalue corresponding to $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)/||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1||_2$ does not change. Particularly, if $p_{S_2} << p$, then all c_m can be much smaller than 1. In this case, $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)/||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1||_2$ will easily show up in PCA.

2.3 Theory Under A Real Scenario

In this subsection, we focus the estimated Σ, $\boldsymbol{\mu}_0$, and $\boldsymbol{\mu}_1$ and investigates the effect of applying the spatial weight and score weight matrices on PCA and its variants for classification.

From Theorem 1 of Fan et al. (2008), we can find a sample based covariance matrix estimator $\hat{\Sigma}$ such that

$$\|\hat{\Sigma} - \Sigma\|_F = O_p(pn^{-\frac{1}{2}}) \tag{10}$$

We respectively replace Σ and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ in (5) by $\hat{\Sigma}$ in (10) and the sample mean difference $\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0$ to generate

$$\hat{\Sigma}_{E(1)} = I_p + \rho_{n,1}(1-\rho_{n,1})\hat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)^{\otimes 2}\hat{\Sigma}^{-1/2}, \tag{11}$$

which further yields $\hat{\Sigma}_{E(3)}$ in Theorem 3 (c)

$$\hat{\Sigma}_{E(3)} = (I_{p_{S_k}}, \mathbf{0})\hat{\Sigma}_{E(1)}(I_{p_{S_k}}, \mathbf{0})^T. \tag{12}$$

Theorem 4 *The first eigenvalue and eigenvector of* $\hat{\Sigma}_{E(1)}$ *in* (11) *and* $\hat{\Sigma}_{E(3)}$ *in* (12) *are, respectively, given by*

$$\begin{aligned}
\hat{\lambda}_{E(1),1} &= 1 + \rho_{n,1}(1-\rho_{n,1})||\hat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)||_2^2, \\
\hat{\mathbf{v}}_{E(1),1} &= \hat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)/||\hat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)||_2; \\
\hat{\lambda}_{E(3),1} &= 1 + \rho_{n,1}(1-\rho_{n,1})||(I_{p_{S_k}}, \mathbf{0})\hat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)||_2^2, \\
\hat{\mathbf{v}}_{E(3),1} &= (I_{p_{S_k}}, \mathbf{0})\hat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)/||(I_{p_{S_k}}, \mathbf{0})\hat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)||_2.
\end{aligned} \tag{13}$$

Under Assumptions (2) *and*(10), $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2 < \infty$, *and* $p^2 n^{-1/2} \to 0$, *the first eigenvalue and the corresponding eigenvector satisfy that as* $n \to \infty$,

$$\begin{aligned}
\hat{\lambda}_{E(1),1} \xrightarrow{p} \lambda_{E(1),1} \quad \text{and} \quad |<\hat{\mathbf{v}}_{E(1),1}, \mathbf{v}_{E(1),1}>| \xrightarrow{p} 1, \\
\hat{\lambda}_{E(3),1} \xrightarrow{p} \lambda_{E(3),1} \quad \text{and} \quad |<\hat{\mathbf{v}}_{E(3),1}, \mathbf{v}_{E(3),1}>| \xrightarrow{p} 1.
\end{aligned} \tag{14}$$

Theorem 4 further confirms the better performance of MWPCR. Depending on the sample data $\mathbf{X}$, MWPCR could generate the consistent estimator $\hat{\mathbf{v}}_{E(1),1}$ (or $\hat{\mathbf{v}}_{E(3),1}$) for $\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ (or $(I_{p_{S_k}}, \mathbf{0})\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$). Theorem 1 (c) (Theorem 3 (d)) has clearly shown that the extracted direction $\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ (or $(I_{p_{S_k}}, \mathbf{0})\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$) corresponds to the Bayesian optimal classifier. Thus, the sample based MWPCR could exact the optimal classification direction and improve classification accuracy even when dimension $p \to \infty$.

Now we consider the sample based estimators for $\Sigma_{E(2)}$ in Theorem 2 (a) and (b) as following:

$$\begin{aligned}
\hat{\Sigma}_{E(2)}^{(a)} &= W_E W_E^T + \rho_{n,1}(1-\rho_{n,1})\rho_0^2 \left\{\hat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)\right\}^{\otimes 2}, \\
\hat{\Sigma}_{E(2)}^{(b)} &= W_E \hat{\Sigma} W_E^T + \rho_{n,1}(1-\rho_{n,1})\rho_0^2 (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)^{\otimes 2}.
\end{aligned}$$

Theorem 5 *Assume that* $pn^{-1/2} \to 0$ *and* $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2 < \infty$.

(a) If $\infty > \lambda^{(a)}_{E(2),1} > \cdots > \lambda^{(a)}_{E(2),k_0}$, *where* $\lambda^{(a)}_{E(2),k}$ *is the* k*-th eigenvalue of* $\Sigma_{E(2)}$ *in Theorem 2 (a), then the* k_0*-th eigenvalue and eigenvector of* $\hat{\Sigma}^{(a)}_{E(2)}$ *satisfy*

$$\hat{\lambda}^{(a)}_{E(2),k_0} \xrightarrow{p} \lambda^2_{w,k_0} + \rho_{n,1}(1-\rho_{n,1})\rho_0^2||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2^2,$$
$$\left| < \hat{\mathbf{v}}^{(a)}_{E(2),k_0}, \Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)/||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2 > \right| \xrightarrow{p} 0.$$

(b) If $\infty > \lambda^{(b)}_{E(2),1} > \lambda^{(b)}_{E(2),2}$, *where* $\lambda^{(b)}_{E(2),k}$ *is the* k*-th eigenvalue of* $\Sigma_{E(2)}$ *in Theorem 2 (b), and* $\rho_{n,1}(1-\rho_{n,1})\rho_0^2||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2^2 >> \max(\lambda_1, 1)$, *then the first eigenvalue and eigenvector of* $\hat{\Sigma}^{(b)}_{E(2)}$ *satisfy*

$$\hat{\lambda}^{(b)}_{E(2),1} \xrightarrow{p} \rho_0^2\sigma^2 + \rho_{n,1}(1-\rho_{n,1})\rho_0^2||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2^2,$$
$$\left| < \hat{\mathbf{v}}^{(b)}_{E(2),1}, (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)/||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1||_2 > \right| \xrightarrow{p} 0.$$

Theorem 5 (a) reveals that the MWPCR estimators are consistent. This together with Theorem 2 shows that MWPCR with both kernel matrix and covariance matrix is better than MWPCR only with covariance matrix. Theorem 5 (b) shows that the sample based estimators $\hat{\lambda}^{(b)}_{E(2),1}$ and $\hat{\mathbf{v}}^{(b)}_{E(2),1}$ are respectively consistent with $\lambda^{(b)}_{E(2),1}$ and $\mathbf{v}^{(b)}_{E(2),1}$. This together with Theorem 2 indicates that the standard PCA could exact the useful information for classification when the signal of $\rho_{n,1}(1-\rho_{n,1})\rho_0^2||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2^2$ is much stronger than $\max(\lambda_1, 1)$.

2.4 Classification Accuracy

In this subsection, we consider a specific setting and show that the use of spatial kernel and importance score weight matrices in MWPCR can improve classification accuracy even when signals are weak (Fan and Fan, 2008). For notational simplicity, we assume that $\mathbf{x}_i|\mathbf{y}_i$ follows a single spike model as follows:

$$\mathbf{x}_i = \boldsymbol{\mu}_{\mathbf{y}_i} + \sqrt{\tilde{\lambda}_1}\xi_i\mathbf{e}_1 + \sigma\boldsymbol{\epsilon}_i, \tag{15}$$

where ξ_i and $\boldsymbol{\epsilon}_i$ are independent with $\xi_i \sim N(0,1)$ and $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, I_p)$, $\mathbf{e}_1$ is a $p \times 1$ unit vector, and $\tilde{\lambda}_1$ and σ are positive scalars. Without loss of generality, we further assume $\mathbf{e}_1 = (1, 0, \cdots, 0)^T$. In this case, Σ is equal to $\tilde{\lambda}_1 \mathbf{e}_1 \mathbf{e}_1^T + \sigma^2 I_p$.

Our MWPCR proceeds as follows. First, we calculate $\hat{\boldsymbol{\mu}}_0 = \sum_{i=1}^{n_1} \mathbf{x}_i / n_1$, $\hat{\boldsymbol{\mu}}_1 = \sum_{i=n_1+1}^{n} \mathbf{x}_i / n_2$, and $\hat{\boldsymbol{\mu}} = \rho_{n,1} \hat{\boldsymbol{\mu}}_0 + (1 - \rho_{n,1}) \hat{\boldsymbol{\mu}}_1$. Second, we perform PCA on the sample covariance matrix of $\{Q^T(\mathbf{x}_i - \hat{\boldsymbol{\mu}}) : i = 1, \cdots, n\}$, denoted as $S_{Q^T\mathbf{x}}$, as follows:

$$S_{Q^T\mathbf{x}} = Q^T \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{\otimes 2} Q = Q^T S_{\mathbf{x}} Q = \hat{V} \hat{D}^2 \hat{V}^T,$$

where $S_{\mathbf{x}} = \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_0)^{\otimes 2} / (n-1)$ and $\hat{V} = [\hat{\mathbf{v}}_1, \cdots, \hat{\mathbf{v}}_K]$. Without of loss of generality, we focus on the space spanned by $\hat{\mathbf{v}}_1$ and construct a projected linear discrimination function. Specifically, a new observation $\mathbf{x}$ is classified into Class 0 if

$$\delta(\mathbf{x}) = (\mathbf{x} - \hat{\boldsymbol{\mu}})^T Q \hat{\mathbf{v}}_1 \hat{\mathbf{v}}_1^T Q^T (\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1) > 0 \tag{16}$$

and its misclassification rate is $\mathbf{W}(\hat{\delta}) = \mathbf{P}\left\{\hat{\delta}(\mathbf{x}) \leq 0 \mid \mathbf{x}_i,\ i = 1, \cdots, n\right\} = 1 - \Phi(\psi_0)$, where ψ_0 is given by

$$\psi_0 = \frac{(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}})^T Q \hat{\mathbf{v}}_1 \hat{\mathbf{v}}_1^T Q^T (\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)}{\sqrt{(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)^T Q \hat{\mathbf{v}}_1 (\hat{\mathbf{v}}_1^T Q^T \Sigma Q \hat{\mathbf{v}}_1) \hat{\mathbf{v}}_1^T Q^T (\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)}}.$$

We compare our MWPCR with the independence classification rule (Fan and Fan, 2008). As shown in Fan and Fan (2008), the independence rule would be no better than the random guessing due to noise accumulation, when the signal level is relatively weak, that is $\sqrt{n/p}||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1||_2^2 \to 0$. Below, we will show that our MWPCR can improve classification accuracy under a key condition that $n||Q^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)||_2^2 / \mathrm{tr}(QQ^T)$ converges to ∞ as follows.

Let $\rho_{Q,12} = \mathbf{e}_1^T QQ^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) / (||Q^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)||_2 ||Q^T \mathbf{e}_1||_2)$. We need the following assumptions.
(A.1) Let $1 \geq \lambda_{Q,1} \geq \cdots \geq \lambda_{Q,q} > \lambda_{Q,q+1} = \cdots = \lambda_{Q,p} = 0$ be the sorted eigenvalues of QQ^T such that $(\sum_{j=1}^{q} \lambda_{Q,j}) / (n||Q^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)||_2^2) \to^p 0$ as $\min(p, n) \to \infty$.
(A.2) $||Q^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)||_2^2 / \log(n) \to \infty$.

(A.3) $R_0 = \tilde{\lambda}_1 ||Q^T \mathbf{e}_1||_2^2 / ||Q^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)||_2^2 < \infty$ holds for $\rho_{Q,12} \neq 0$ and $R_0 < 1 - \delta_0$ for $\rho_{Q,12} = 0$, where $\delta_0 > 0$ is a sufficiently small constant.

Theorem 6 *Under Assumptions (A.1)-(A.3), the misclassification rate of (16) is given by*

$$\mathbf{W}(\hat{\delta}) \xrightarrow{p} 1 - \Phi\left(\alpha_0 ||Q^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)||_2\right), \tag{17}$$

where α_0 is a positive constant defined in the supplementary document.

Theorem 6 shows that the use of Q can reduce the effects of noise accumulation on MWPCR. Specifically, when the true signal levels are relatively weak, that is $||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1||_2^2 = o(\sqrt{p/n})$, our MWPCR outperforms the random guessing as $\text{tr}(QQ^T)/\sqrt{np}$ is relatively small. As discussed above, applying both smoothing and selection weight matrices can reduce $\text{tr}(QQ^T)$, so these weight matrices should be used.

3 Proofs

This sections shows the proofs of Theorems 1 to 6. The proof of Theorems 1 is straightforward and skipped here to save space. The proofs of Theorems 2 to 6 are respectively given out in Sections 3.2 to 3.6. Section 3.1 presents several lemmas that are used in the proofs of theorems.

3.1 Lemmas

This subsection shows several lemmas that are used in the proofs of theorems. This first one is the Wielandt's Inequality (Rao, 2002) that is used to prove Lemma 2.

Lemma 1 *(Wielandt's Inequality). If A, B are $p \times p$ real symmetric matrices, then for all $j =$*

$1, \ldots, p,$

$$\left\{ \begin{array}{lcl} \lambda_j(A) & + & \lambda_p(B) \\ \lambda_{j+1}(A) & + & \lambda_{p-1}(B) \\ & \vdots & \\ \lambda_p(A) & + & \lambda_j(B) \end{array} \right\} \leq \lambda_j(A+B) \leq \left\{ \begin{array}{lcl} \lambda_j(A) & + & \lambda_1(B) \\ \lambda_{j-1}(A) & + & \lambda_2(B) \\ & \vdots & \\ \lambda_1(A) & + & \lambda_j(B) \end{array} \right\}.$$

The second lemma shows the convergence of eigenvalues and eigenvectors for two sequences of matrices.

Lemma 2 *Assume that two $p \times p$ matrices sequences $\{A_k, k = 1, \cdots, \infty\}$ and $\{B_k, k = 1, \cdots, \infty\}$ satisfying that as $k \to \infty$, $\|A_k - B_k\|_F \to 0$, then we have*

$$\max_{1 \leq j \leq p} |\lambda_j(A_k) - \lambda_j(B_k)| \to 0, \tag{18}$$

where $\lambda_j(.)$ is the jth eigenvalue of matrix and p could go to infinite when $k \to \infty$. Furthermore, if the first m (m is finite) eigenvalues of B_k satisfy that as $k \to \infty$, $\infty > \lambda_1(B_k) > \cdots > \lambda_m(B_k) > 0$, then as $k \to \infty$,

$$| < \mathbf{v}_l(A_k), \mathbf{v}_l(B_k) > | \to 1, \quad l = 1, \cdots, m, \tag{19}$$

where $\mathbf{v}_l(.)$ is the lth eigenvector of the matrix.

Proof of Lemma 2. First, we show the proof of (18). According to Lemma 1, we have that

$$\lambda_j(B_k) + \lambda_p(A_k - B_k) \leq \lambda_j(A_k) \leq \lambda_j(B_k) + \lambda_1(A_k - B_k), \quad j = 1, \cdots, p,$$

which yields

$$\max_{1 \leq j \leq p} |\lambda_j(A_k) - \lambda_j(B_k)| \leq |\lambda_1(A_k - B_k)| + |\lambda_p(A_k - B_k)| \leq 2\|A_k - B_k\|_F. \tag{20}$$

Since $\|A_k - B_k\|_F \to 0$, then it follows from (20) that (18) is established.

Second, we show the proof of (19). Note that A_k and B_k have the following eigen-decomposition

$$A_k = \sum_{l=1}^{p} \lambda_l(A_k)\mathbf{v}_l(A_k)\mathbf{v}_l^T(A_k) \quad \text{and} \quad B_k = \sum_{l=1}^{p} \lambda_l(B_k)\mathbf{v}_l(B_k)\mathbf{v}_l^T(B_k),$$

which yields

$$\begin{aligned} \lambda_1(A_k) &= \mathbf{v}_1^T(A_k)B_k\mathbf{v}_1(A_k) + \mathbf{v}_1^T(A_k)(A_k - B_k)\mathbf{v}_1(A_k) \\ &\leq \lambda_1(B_k)(\mathbf{v}_1^T(A_k)\mathbf{v}_1(B_k))^2 + \lambda_2(B_k)[1 - (\mathbf{v}_1^T(A_k)\mathbf{v}_1(B_k))^2] + \|A_k - B_k\|_F. \end{aligned} \quad (21)$$

Since $\|A_k - B_k\|_F \to 0$ and $\lambda_1(A_k) \to \lambda_1(B_k) > \lambda_2(B_k)$, then it follows from (21) that

$$(\mathbf{v}_1^T(A_k)\mathbf{v}_1(B_k))^2 \to 1, \quad (22)$$

which yields (19) for $l = 1$. According to (22), we have that $|\mathbf{v}_2^T(A_k)\mathbf{v}_1(B_k)| \to 0$. Repeat the same procedure, we have (19) for $l = 2, \cdots, m$.

The third lemma is about the convergence of the sample based matrices.

Lemma 3 *Under the assumptions* (2) *and* (10) *and* $p^2 n^{-\frac{1}{2}} \to 0$*, we have following properties that as* $N \to \infty$*,*

$$\|\hat{\Sigma}^{-1/2} - \Sigma^{-1/2}\|_F = O_p(p^2 n^{-\frac{1}{2}}) \quad \text{and} \quad \|\hat{\Sigma}^{-1} - \Sigma^{-1}\|_F = O_p(p^2 n^{-\frac{1}{2}}). \quad (23)$$

Proof of Lemma 3. Since $\hat{\Sigma}^{-1/2} + \Sigma^{-1/2} - \lambda_p^{1/2} I_p$ is the non-negative matrix, then it follows from Proposition 2.1 in Van Hemmen and Ando (1980) that

$$\|\hat{\Sigma}^{-1/2} - \Sigma^{-1/2}\|_F \leq \lambda_p^{-1/2} \|\hat{\Sigma}^{-1} - \Sigma^{-1}\|_F. \quad (24)$$

According to (6.12) of Fan et al. (2008), we have that whenever $||\Sigma^{-1}||_F||\hat{\Sigma}-\Sigma||_F < 1$,

$$\|\hat{\Sigma}^{-1}-\Sigma^{-1}\|_F \leq \frac{||\Sigma^{-1}||_F^2||\hat{\Sigma}-\Sigma||_F}{1-||\Sigma^{-1}||_F||\hat{\Sigma}-\Sigma||_F} \leq \frac{p\lambda_p^{-2}||\hat{\Sigma}-\Sigma||_F}{1-p^{1/2}\lambda_p^{-1}||\hat{\Sigma}-\Sigma||_F},$$

which together with (2), (10) and (24), yields (23).

The fourth lemma is about the convergence of the sample mean difference.

Lemma 4 *The difference between* $\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_0$ *and* $\boldsymbol{\mu}_1-\boldsymbol{\mu}_0$ *satisfy*

$$\|(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_0)-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)\|_\infty = O_p\left(\log(p)^{\frac{1}{2}}n^{-\frac{1}{2}}\right)$$

Lemma 4 is from (A5) in the appendix of Mai et al. (2012).

The last lemmas show the uniform convergence of sample eigenvalues.

Lemma 5 *Under the assumptions* (2) *and* (10) *and* $pn^{-\frac{1}{2}} \to 0$*, the eigenvalues of* $\hat{\Sigma}$ *and* Σ *satisfy that as* $n \to \infty$*,*

$$\max_{1\leq j\leq p}|\hat{\lambda}_j-\lambda_j| \xrightarrow{p} 0.$$

Lemma 5 is directly from Lemma 2.

Lemma 6 *Consider a rank-2 matrix* $\Sigma_0 = \lambda_{w0,1}\mathbf{w}_1^{\otimes 2}+\lambda_{w0,2}\mathbf{w}_2^{\otimes 2}$*, where* $\mathbf{w}^{\otimes 2}=\mathbf{w}\mathbf{w}^T$ *for any vector* $\mathbf{w}$*,* $||\mathbf{w}_1||_2=||\mathbf{w}_2||_2=1$*, and* $\lambda_{w0,1}\geq\lambda_{w0,2}>0$*. Let* $\rho_{1,2}=<\mathbf{w}_1,\mathbf{w}_2>$*. The two non-zero eigenvalues of* Σ_0 *are given by*

$$\begin{aligned}\lambda_+ &= 0.5\{\lambda_{w0,1}+\lambda_{w0,2}+\sqrt{(\lambda_{w0,1}-\lambda_{w0,2})^2+4\lambda_{w0,1}\lambda_{w0,2}\rho_{1,2}^2}\},\\ \lambda_- &= 0.5\{\lambda_{w0,1}+\lambda_{w0,2}-\sqrt{(\lambda_{w0,1}-\lambda_{w0,2})^2+4\lambda_{w0,1}\lambda_{w0,2}\rho_{1,2}^2}\}.\end{aligned}$$

If $\rho_{1,2} \neq 0$, then the eigenvectors corresponding to λ_+ and λ_- are given by

$$\mathbf{v}_{0,1} = \frac{\mathbf{w}_1 + x_+(\mathbf{w}_2 - \rho_{1,2}\mathbf{w}_1)}{\sqrt{1 + x_+^2(1-\rho_{1,2}^2)}} \text{ and } \mathbf{v}_{0,2} = \frac{\mathbf{w}_1 + x_-(\mathbf{w}_2 - \rho_{1,2}\mathbf{w}_1)}{\sqrt{1 + x_-^2(1-\rho_{1,2}^2)}},$$

where x_+ and x_- are, respectively, given by

$$x_\pm = \frac{-\{\lambda_{w0,2}(2\rho_{1,2}^2 - 1) + \lambda_{w0,1}\} \pm \sqrt{(\lambda_{w0,1} - \lambda_{w0,2})^2 + 4\lambda_{w0,1}\lambda_{w0,2}\rho_{1,2}^2}}{2\lambda_{w0,2}\rho_{1,2}(1-\rho_{1,2}^2)}.$$

3.2 Proof of Theorem 2

First, we prove Theorem 2 (a). For $\Gamma = \Sigma^{-1/2}$, it is easy to show that for $\mathbf{v} \perp \Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$, we have

$$\Sigma_{E(2)} = W_E W_E^T + \rho_{n,1}(1-\rho_{n,1})\rho_0^2\{\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\}^{\otimes 2}, \tag{25}$$

$$\frac{\mathbf{v}^T\Sigma_{E(2)}\mathbf{v}}{||\mathbf{v}||_2^2} = \frac{\mathbf{v}^T W_E W_E^T \mathbf{v}}{||\mathbf{v}||_2^2} \leq \lambda_{w,1}. \tag{26}$$

If W_E is symmetric, $\lambda_{w,k_0}^2 = \rho_o^2$ and $\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ is an eigenvector of W_E, then $W_E W_E^T$ has the eigen-decomposition

$$W_E W_E^T = \lambda_{w,k_0}^2 \left\{\frac{\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}{||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2}\right\}\left\{\frac{\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}{||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2}\right\}^T + \sum_{k\neq k_0, 1\leq k\leq p} \lambda_{w,k}^2 \mathbf{v}_{w,k}\mathbf{v}_{w,k}^T.$$

Then it follows from (25) that $\Sigma_{E(2)}$ has the eigen-decomposition

$$\Sigma_{E(2)} = \left\{\lambda_{w,k_0}^2 + \rho_{n,1}(1-\rho_{n,1})\rho_0^2||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2^2\right\}\left\{\frac{\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}{||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2}\right\}\left\{\frac{\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}{||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2}\right\}^T + \sum_{k\neq k_0, 1\leq k\leq p} \lambda_{w,k}^2 \mathbf{v}_{w,k}\mathbf{v}_{w,k}^T. \tag{27}$$

Until now, we have showed the eigenvalues and eigenvectors of $\Sigma_{E(2)}$ in Theorem 2 (a).

According to (27) and (26), we have that

$$\frac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\Sigma^{-1/2}\Sigma_{E(2)}\Sigma^{-1/2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}{||\Sigma^{-1/2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)||_2^2}=\lambda_{w,k_0}^2+\rho_{n,1}(1-\rho_{n,1})\rho_0^2||\Sigma^{-1/2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)||_2^2$$
$$\geq \rho_{n,1}(1-\rho_{n,1})\rho_0^2||\Sigma^{-1/2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)||_2^2>1\geq\lambda_{w,1}^2\geq\frac{\mathbf{v}^T\Sigma_{E(2)}\mathbf{v}}{||\mathbf{v}||_2^2}.$$

Until now, we finished the proof of Theorem 2 (a).

Finally, we show the proof of Theorem 2 (b). For $\Gamma=I_p$, it follows from (8) that

$$\Sigma_{E(2)}=W_E(\tilde{V}\tilde{D}\tilde{V}^T+\sigma^2I_p)W_E^T+\rho_{n,1}(1-\rho_{n,1})\rho_0^2\left\{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)\right\}^{\otimes 2}. \tag{28}$$

Since $\mathbf{v}\perp\mathcal{F}$ and (28), then

$$\frac{\mathbf{v}^T\Sigma_{E(2)}\mathbf{v}}{||\mathbf{v}||_2^2}=\sigma^2\frac{\mathbf{v}^TW_EW_E^T\mathbf{v}}{||\mathbf{v}||_2^2}\leq\sigma^2\lambda_{w,m_w+1}^2. \tag{29}$$

According to (28), we have

$$\frac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\Sigma_{E(2)}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}{||\boldsymbol{\mu}_1-\boldsymbol{\mu}_0||_2^2}\geq\rho_{n,1}(1-\rho_{n,1})\rho_0^2||\boldsymbol{\mu}_1-\boldsymbol{\mu}_0||_2^2. \tag{30}$$

Since $\rho_{n,1}(1-\rho_{n,1})\rho_0^2||\boldsymbol{\mu}_1-\boldsymbol{\mu}_0||_2^2$ is larger than $\sigma^2\lambda_{w,m_w+1}^2$, then it follows from (29) and (30)

$$\frac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\Sigma_{E(2)}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}{||\boldsymbol{\mu}_1-\boldsymbol{\mu}_0||_2^2}\geq\rho_{n,1}(1-\rho_{n,1})\rho_0^2||\boldsymbol{\mu}_1-\boldsymbol{\mu}_0||_2^2>\sigma^2\lambda_{w,m_w+1}^2\geq\frac{\mathbf{v}^T\Sigma_{E(2)}\mathbf{v}}{||\mathbf{v}||_2^2}.$$

In addition, if $\boldsymbol{\mu}_0-\boldsymbol{\mu}_1\perp W_E\mathbf{v}_j$ holds for all $j\leq m$ and W_E is symmetric, according to (28), it is easy to check that $(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)/||\boldsymbol{\mu}_0-\boldsymbol{\mu}_1||_2$ is the eigenvector of $\Sigma_{E(2)}$ corresponding to the eigenvalue $\rho_0^2\sigma^2+\rho_{n,1}(1-\rho_{n,1})\rho_0^2||\boldsymbol{\mu}_1-\boldsymbol{\mu}_0||_2^2$. This finishes the proof of Theorem 2 (b).

3.3 Proof of Theorem 3

We only prove Theorem 3 (e). We have

$$\Sigma_{E(3)} \quad = \quad (I_{p_{S_2}}, \mathbf{0})(\tilde{V}\tilde{D}\tilde{V}^T + \sigma^2 I_p)(I_{p_{S_2}}, \mathbf{0})^T + \rho_{n,1}(1-\rho_{n,1})\rho_0^2\{(\boldsymbol{\mu}_{1,S_2} - \boldsymbol{\mu}_{0,S_2})\}^{\otimes 2}.$$

Furthermore, we have

$$(I_{p_{S_2}}, \mathbf{0})\tilde{V}\tilde{D}\tilde{V}^T(I_{p_{S_2}}, \mathbf{0})^T = \sum_{j=1}^{m} \lambda_j \mathbf{v}_{S_2,j}\mathbf{v}_{S_2,j}^T = \sum_{j=1}^{m} \lambda_j ||\mathbf{v}_{S_2,j}||_2^2 \tilde{\mathbf{v}}_{S_2,j}\tilde{\mathbf{v}}_{S_2,j}^T,$$

where $\mathbf{v}_{S_2,j} = (I_{p_{S_2}}, \mathbf{0})\mathbf{v}_j$ for $j = 1, \ldots, m$ and $\tilde{\mathbf{v}}_{S_2,j} = \mathbf{v}_{S_2,j}/||\mathbf{v}_{S_2,j}||_2$, in which $||\mathbf{v}_{S_2,j}||_2 \leq 1$. It follows from the Cauchy (eigenvalue) interlacing theorem that the j-th eigenvalue of $(I_{p_{S_2}}, \mathbf{0})\tilde{V}\tilde{D}\tilde{V}^T(I_{p_{S_2}}, \mathbf{0})^T$, denoted as $\lambda_j c_j$, is smaller than λ_j for $j = 1, \ldots, m$. Thus, we have $c_j \leq 1$.

3.4 Proof of Theorem 4

Note that $\hat{\Sigma}_{E(3)}$ is a sub matrix of the $\hat{\Sigma}_{E(1)}$, and the proof of the consistency of the first eigenvalue and eigenvector of $\hat{\Sigma}_{E(1)}$ and $\hat{\Sigma}_{E(2)}$ are essentially same. Thus we only give out the proof of the properties of $\hat{\Sigma}_{E(1)}$ here. Proof of (13) is straightforward and skipped here to save space. We now show the detailed proof of (14).

First, we show the asymptotic property $\hat{\lambda}_{E(1),1} \xrightarrow{p} \lambda_{E(1),1}$. According to (6) and (13), we just need to show

$$||\hat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)||_2 \xrightarrow{p} ||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2, \tag{31}$$

which is equivalent with

$$||\hat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - \Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2 \xrightarrow{p} 0. \tag{32}$$

We rewrite $\hat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - \Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ as following:

$$\hat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - \Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = \mathbf{T}_1 + \mathbf{T}_2, \tag{33}$$

where

$$\mathbf{T}_1 = (\hat{\Sigma}^{-1/2} - \Sigma^{-1/2})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \quad \text{and} \quad \mathbf{T}_2 = \hat{\Sigma}^{-1/2}\left[(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right].$$

In order to show (32), it follows from (33) that we just need to show

$$||\mathbf{T}_i||_2 \xrightarrow{p} 0, \quad i = 1, 2. \tag{34}$$

Note that

$$||\mathbf{T}_1||_2 \leq ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2 ||\hat{\Sigma}^{-1/2} - \Sigma^{-1/2}||_F$$

which together with Lemma 3 and the assumptions $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2 < \infty$ and $p^2 n^{-1/2} \to 0$, yields (34) for $i = 1$. In addition, note that

$$||\mathbf{T}_2||_2^2 \leq p\hat{\lambda}_p^{-1} \|(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\|_\infty^2,$$

which together with Lemmas 4 and 5, and the assumption $p^2 n^{-1/2} \to 0$, yields (34) for $i = 2$. Then it follows that (32) is established.

Second, we shows the asymptotic property of the eigenvector $\hat{\mathbf{v}}_{E(1),1}$ such that

$$| < \hat{\mathbf{v}}_{E(1),1}, \mathbf{v}_{E(1),1} > | \xrightarrow{p} 1. \tag{35}$$

Since

$$| < \hat{\mathbf{v}}_{E(1),1}, \mathbf{v}_{E(1),1} > | = \frac{|(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)\hat{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)|}{||\hat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)||_2 ||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2},$$

then in order to show (35), it follows from (31) that we just need to show

$$(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)\hat{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \xrightarrow{p} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0). \tag{36}$$

Note that

$$(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)\hat{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = \mathbf{T}_1^* + \mathbf{T}_2^*, \tag{37}$$

where

$$\mathbf{T}_1^* = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T(\hat{\Sigma}^{-1} - \Sigma^{-1})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \quad \text{and} \quad \mathbf{T}_2^* = \{(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\}^T\hat{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0).$$

Thus, in order to prove (35), it follows from (36) and (37) that we just need to show

$$|\mathbf{T}_i^*| \xrightarrow{p} 0, \quad i = 1, 2. \tag{38}$$

Note that

$$|\mathbf{T}_1^*| \leq ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2^2||\hat{\Sigma}^{-1} - \Sigma^{-1}||_F$$

which together with Lemma 3 and the assumptions $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2 < \infty$ and $p^2n^{-1/2} \to 0$, yields (38) for $i = 1$. In addition, it follows from Cauchy-Schwarz inequality that

$$\begin{aligned} \mathbf{T}_2^{*2} &\leq ||\hat{\Sigma}^{-1/2}\left[(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right]||_2^2||\hat{\Sigma}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2^2 \\ &\leq p\hat{\lambda}_p^{-2}\|(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\|_\infty^2||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2^2, \end{aligned}$$

which together with Lemmas 4 and 5, and the assumptions $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2 < \infty$ and $p^2n^{-1/2} \to 0$, yields (38) for $i = 2$. It follows from that (35) is established.

3.5 Proof of Theorem 5

Let $\Sigma^{(a)}_{E(2)}$ be $\Sigma_{E(2)}$ in Theorem 2 (a). In order to prove Theorem 5 (a), it follows from Lemma 2 that we just need to show

$$||\hat{\Sigma}^{(a)}_{E(2)} - \Sigma^{(a)}_{E(2)}||_F \xrightarrow{p} 0,$$

which can be realized by proving

$$||\hat{\Sigma}^{(a)}_{E(2)} - \Sigma^{*}_{E(2)}||_F \xrightarrow{p} 0, \tag{39}$$

$$||\Sigma^{*}_{E(2)} - \Sigma^{(a)}_{E(2)}||_F \xrightarrow{p} 0, \tag{40}$$

where $\Sigma^{*}_{E(2)} = W_E W_E^T + \rho_{n,1}(1-\rho_{n,1})\rho_0^2 \left\{\hat{\Sigma}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right\}^{\otimes 2}$.

First, we give out the proof of (39). The $\hat{\Sigma}^{(a)}_{E(2)} - \Sigma^{*}_{E(2)}$ can be rewritten as:

$$\begin{aligned}
\frac{\hat{\Sigma}^{(a)}_{E(2)} - \Sigma^{*}_{E(2)}}{\rho_{n,1}(1-\rho_{n,1})\rho_0^2} &= \left\{\hat{\Sigma}^{-1/2}\left[(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right]\right\}\left\{\hat{\Sigma}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right\}^T \\
&+ \left\{\hat{\Sigma}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right\}\left\{\hat{\Sigma}^{-1/2}\left[(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right]\right\}^T \\
&+ \left\{\hat{\Sigma}^{-1/2}\left[(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right]\right\}^{\otimes 2},
\end{aligned} \tag{41}$$

where the right three terms are defined as $\mathbb{I}^*_i$ for $i = 1, 2, 3$. In order to prove (39), it follows from (41) that we just need to show

$$||\mathbb{I}^*_i||_F \xrightarrow{p} 0, \quad i = 1, 2, 3. \tag{42}$$

Since for $i = 1, 2$,

$$\begin{aligned}
||\mathbb{I}^*_i||_F &= ||\hat{\Sigma}^{-1/2}\left[(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right]||_2 ||\hat{\Sigma}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2 \\
&\leq \hat{\lambda}_p^{-1}||(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2 ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2 \\
&\leq p^{1/2}\hat{\lambda}_p^{-1}||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2 ||(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_\infty,
\end{aligned}$$

then it follows from $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2 < \infty$, $p^2 n^{-1/2} \to 0$, and Lemmas 4 and 5 that (42) is established for $i = 1, 2$. Similarly, we have

$$||\mathbb{I}_3^*||_F \leq p\hat{\lambda}_p^{-1}||(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_\infty^2,$$

which together with Lemmas 4 and 5 and $p^2 n^{-1/2} \to 0$, yields (42) for $i = 3$. Then it follows from (41) and (42) that (39) is established.

Second, we give out the proof of (40). The $\Sigma_{E(2)}^* - \Sigma_{E(2)}^{(a)}$ is rewritten as following

$$\begin{aligned}
\frac{\Sigma_{E(2)}^* - \Sigma_{E(2)}^{(a)}}{\rho_{n,1}(1-\rho_{n,1})\rho_0^2} &= \left\{(\hat{\Sigma}^{-1/2} - \Sigma^{-1/2})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\}\{\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right\}^T \\
&+ \left\{\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right\}\left\{(\hat{\Sigma}^{-1/2} - \Sigma^{-1/2})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right\}^T \\
&+ \left\{(\hat{\Sigma}^{-1/2} - \Sigma^{-1/2})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right\}\left\{(\hat{\Sigma}^{-1/2} - \Sigma^{-1/2})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right\}^T, \qquad (43)
\end{aligned}$$

where the right three terms are defined as $\mathbb{I}_i^{**}$, $i = 1, 2, 3$. In order to prove (40), it follows (43) that we just need to show

$$||\mathbb{I}_i^{**}||_F \xrightarrow{p} 0, \quad i = 1, 2, 3. \qquad (44)$$

Since for $i = 1, 2$,

$$\begin{aligned}
||\mathbb{I}_i^{**}||_F &= ||(\hat{\Sigma}^{-1/2} - \Sigma^{-1/2})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2||\Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_2 \\
&\leq \lambda_p^{-1/2}||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2^2||\hat{\Sigma}^{-1/2} - \Sigma^{-1/2}||_F,
\end{aligned}$$

then it follows from Lemma 3 and the assumptions $||\mu_1 - \mu_0||_2 < \infty$ and $p^2 n^{-1/2} \to 0$ that (44) is established for $i = 1, 2$. Similarly, we have

$$||\mathbb{I}_3^{**}||_F \leq ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2^2||\hat{\Sigma}^{-1/2} - \Sigma^{-1/2}||_F^2,$$

together with Lemma 3 and $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2 < \infty$ and $p^2 n^{-1/2} \to 0$, yields (44) for $i = 3$. It follows

that (40) is established.

Let $\Sigma^{(b)}_{E(2)}$ be $\Sigma_{E(2)}$ in Theorem 2 (b). In order to prove Theorem 5 (b), it follows from Lemma 2 that we just need to show

$$||\hat{\Sigma}^{(b)}_{E(2)} - \Sigma^{(b)}_{E(2)}||_F \xrightarrow{p} 0. \tag{45}$$

The $\hat{\Sigma}^{(b)}_{E(2)} - \Sigma^{(b)}_{E(2)}$ can be rewritten as following:

$$\begin{aligned}
\hat{\Sigma}^{(b)}_{E(2)} - \Sigma^{(b)}_{E(2)} &= W_E(\hat{\Sigma} - \Sigma)W_E^T \\
&+ \rho_{n,1}(1-\rho_{n,1})\rho_0^2 \left[(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right](\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \\
&+ \rho_{n,1}(1-\rho_{n,1})\rho_0^2 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\left[(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right]^T \\
&+ \rho_{n,1}(1-\rho_{n,1})\rho_0^2 \left[(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right]^{\otimes 2},
\end{aligned}$$

where the four right terms are defined as $\mathbb{I}_i$ for $i = 1, 2, 3, 4$. In order to prove (45), we just need to show

$$||\mathbb{I}_i||_F \xrightarrow{p} 0, \quad i = 1, 2, 3, 4. \tag{46}$$

Note that

$$||\mathbb{I}_1||_F \leq \lambda^2_{w,1}||\hat{\Sigma} - \Sigma||_F \leq ||\hat{\Sigma} - \Sigma||_F.$$

In addition, since $\lambda_{w,1} \leq 1$, $pn^{-1/2} \to 0$ and (10), then (46) is established for $i = 1$. Since

$$||\mathbb{I}_i||_F \leq \rho_{n,1}(1-\rho_{n,1})\rho_0^2 p^{1/2}||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2||(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_\infty, \quad i = 2, 3,$$

then it follows from Lemma 4, $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_2 < \infty$ and $pn^{-1/2} \to 0$ that (46) is established for $i = 2, 3$. Finally, since

$$||\mathbb{I}_4||_F \leq \rho_{n,1}(1-\rho_{n,1})\rho_0^2 p||(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||^2_\infty,$$

then according to Lemma 4 and $pn^{-1/2} \to 0$, we have (46) for $i = 4$.

3.6 Proof of Theorem 6

Recall that ψ_0 is given by

$$\psi_0 = \frac{(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}})^T Q \hat{\mathbf{v}}_1 \hat{\mathbf{v}}_1^T Q^T (\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)}{\sqrt{(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)^T Q \hat{\mathbf{v}}_1 \hat{\mathbf{v}}_1^T Q^T \Sigma Q \hat{\mathbf{v}}_1 \hat{\mathbf{v}}_1^T Q^T (\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)}} \xrightarrow{p} \alpha_0 ||Q^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)||_2. \tag{47}$$

The proof of (47) consists of four key steps:

- Step (I) is to derive an approximation to $\hat{\mathbf{v}}_1 Q^T(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}})$.
- Step (II) is to derive a bound for $||Q^T n^{-1}\{\sum_{i=1}^n \boldsymbol{\epsilon}_i^{\otimes 2}\}Q||_{op}$, where $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, I_p)$.
- Step (III) is to derive an approximation to $||Q^T(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)||_2$.
- Step (IV) is to derive an approximation to $\hat{\mathbf{v}}_1$.
- Step (V) is to derive an approximation to ψ_0.

For Step (I), we proceed as follows. Since $\hat{\boldsymbol{\mu}} = \rho_{n,1}\hat{\boldsymbol{\mu}}_0 + (1-\rho_{n,1})\hat{\boldsymbol{\mu}}_1$ and $||\hat{\mathbf{v}}_1||_2 = 1$, we have

$$\hat{\mathbf{v}}_1^T Q^T(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}) = (1-\rho_{n,1})\hat{\mathbf{v}}_1^T Q^T(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1) + \hat{\mathbf{v}}_1^T Q^T(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_0).$$

It can be shown that

$$\hat{\mathbf{v}}_1^T Q^T(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_0) = \hat{\mathbf{v}}_1^T Q^T \mathbf{e}_1 \sqrt{\tilde{\lambda}} \sum_{i=1}^{n_1} \xi_i / n_1 + \sigma \hat{\mathbf{v}}_1^T Q^T \sum_{i=1}^{n_1} \boldsymbol{\epsilon}_i / n_1 = O_p(n_1^{-1/2})(\sqrt{\tilde{\lambda}}||Q^T\mathbf{e}_1||_2 + \sigma||Q\hat{\mathbf{v}}_1||_2).$$

For Step (II), we proceed as follows. Let $Q = U_Q^T \Lambda_Q V_Q$ be the singular decomposition of Q, where $\Lambda_Q = \mathrm{diag}(\lambda_{Q,1}^{1/2}, \cdots, \lambda_{Q,q}^{1/2})$ and $U_Q U_Q^T = V_Q V_Q^T = I_q$. We have

$$Q^T n^{-1}\{\sum_{i=1}^n \boldsymbol{\epsilon}_i^{\otimes 2}\}Q = V_Q^T \Lambda_Q n^{-1}\{\sum_{i=1}^n (U_Q \boldsymbol{\epsilon}_i)^{\otimes 2}\}\Lambda_Q V_Q.$$

Note that the eigenvalues of AB are the same as those of BA, where the column of B is the same as the row of A. Therefore, the eigenvalues of $Q^T n^{-1}\{\sum_{i=1}^n \boldsymbol{\epsilon}_i^{\otimes 2}\}Q$ are the same as those

of $n^{-1}\sum_{j=1}^{q}\lambda_{Q,j}\tilde{\boldsymbol{\epsilon}}_i^{\otimes 2}$, where $\tilde{\boldsymbol{\epsilon}}_i$ is an $n\times 1$ Gaussian random vector $N(\mathbf{0}, I_n)$. It follows from the matrix Bernstein theorem (Tropp, 2012) that with a high probability, we have

$$||n^{-1}\sum_{j=1}^{q}\lambda_{Q,j}\tilde{\boldsymbol{\epsilon}}_i^{\otimes 2}||_{op} \leq n^{-1}\sum_{j=1}^{q}\lambda_{Q,j} + C_1\log(n)\{1+[n^{-1}\sum_{j=1}^{q}\lambda_{Q,j}^2/\log(n)]^{1/2}\}. \tag{48}$$

For Step (III), we process as follows. It is easy to show that $||Q^T(\hat{\boldsymbol{\mu}}_0-\hat{\boldsymbol{\mu}}_1)||_2^2$ is equal to

$$||Q^T(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)||^2 + ||\Delta\hat{\boldsymbol{\mu}}_{0,1}||_2^2 + 2(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^TQ\Delta\hat{\boldsymbol{\mu}}_{0,1}, \tag{49}$$

where $\Delta\hat{\boldsymbol{\mu}}_{0,1} = Q^T(\hat{\boldsymbol{\mu}}_0-\boldsymbol{\mu}_0-\hat{\boldsymbol{\mu}}_1+\boldsymbol{\mu}_1)$, which follows $N(\mathbf{0}, (n_0^{-1}+n_1^{-1})(\tilde{\lambda}Q^T\mathbf{e}_1\mathbf{e}_1^TQ+\sigma^2Q^TQ))$. Therefore, it is easy to show that with a high probability, $|(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^TQ\Delta\hat{\boldsymbol{\mu}}_{0,1}|$ is bounded by

$$\frac{\log(n)}{\sqrt{n}}||Q^T(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)||_2\sqrt{\tilde{\lambda}||Q^T\mathbf{e}_1||_2^2+\sigma^2}$$

Moreover, $||\Delta\hat{\boldsymbol{\mu}}_{0,1}||_2^2$ can be represented as

$$(n_0^{-1}+n_1^{-1})\tilde{\lambda}_1(\boldsymbol{\eta}^TQ^T\mathbf{e}_1)^2 + (n_0^{-1}+n_1^{-1})\sigma^2\boldsymbol{\eta}^TQ^TQ\boldsymbol{\eta},$$

where $\boldsymbol{\eta}\sim N(\mathbf{0}, I_p)$. Similar to (48), it is easy to show that with a large probability, $\boldsymbol{\eta}^TQ^TQ\boldsymbol{\eta}$ is smaller than $\sum_{j=1}^{q}\lambda_{Q,j} + \sqrt{\sum_{j=1}\lambda_{Q,j}^2}C\log(n)$, where C is a generic constant. Moreover, since $\boldsymbol{\eta}^TQ^T\mathbf{e}_1\sim N(0, ||Q^T\mathbf{e}_1||_2^2)$, with a large probability, we have

$$||\Delta\hat{\boldsymbol{\mu}}_{0,1}||_2^2 \leq Cn^{-1}\log(n)\tilde{\lambda}_1||Q^T\mathbf{e}_1||_2^2 + Cn^{-1}\{\sum_{j=1}^{q}\lambda_{Q,j} + \sqrt{\sum_{j=1}^{2}\lambda_{Q,j}^2}\log(n)\}.$$

By combining above results, we can derive an approximation to $||Q^T(\hat{\boldsymbol{\mu}}_0-\hat{\boldsymbol{\mu}}_1)||_2^2$ as follows:

$$C||Q^T(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)||_2^2\{1+\frac{\log(n)}{\sqrt{n}} + n^{-1}||Q^T(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)||_2^2\sum_{j=1}^{q}\lambda_{Q,j}\},$$

where C is a generic constant.

For Step (IV), we proceed as follows. Recall that $\hat{\mathbf{v}}_1$ is the eigenvector of $S_{Q^T\mathbf{x}}$ corresponding to its largest eigenvalue. We construct an approximation of $S_{Q^T\mathbf{x}}$, denoted as $\tilde{S}_{Q^T\mathbf{x}}$, given by

$$\tilde{S}_{Q^T\mathbf{x}} = K_{n,1}\{\frac{Q^T(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)}{||Q^T(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)||_2}\}^{\otimes 2} + K_{n,2}(\frac{Q^T\mathbf{e}_1}{||Q^T\mathbf{e}_1||_2})^{\otimes 2}, \tag{50}$$

where $K_{n,1} = \rho_{n,1}(1-\rho_{n,1})n||Q^T(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)||_2^2/(n-1)$ and $K_{n,2} = \tilde{\lambda}||Q^T\mathbf{e}_1||_2^2\sum_{i=1}^{n-2}\xi_i^2/(n-1)$. With some calculations, we have

$$\frac{n-1}{n-2}(S_{Q^T\mathbf{x}} - \tilde{S}_{Q^T\mathbf{x}}) = \sigma^2 Q^T\overline{\boldsymbol{\epsilon}^{\otimes 2}}Q + \sigma\sqrt{\tilde{\lambda}}Q^T\{\mathbf{e}_1(\overline{\xi\boldsymbol{\epsilon}})^T + (\overline{\xi\boldsymbol{\epsilon}})\mathbf{e}_1^T\}Q, \tag{51}$$

where $\overline{\boldsymbol{\epsilon}^{\otimes 2}} = \sum_{i=1}^{n-2}\boldsymbol{\epsilon}_i^{\otimes 2}/(n-2)$ and $\overline{\xi\boldsymbol{\epsilon}} = \sum_{i=1}^{n-2}\xi_i\boldsymbol{\epsilon}_i/(n-2)$. It follows from the random matrix theory that with a high probability, we have

$$\begin{aligned}
&\sigma^2||Q^T\overline{\boldsymbol{\epsilon}^{\otimes 2}}Q||_{op} \leq \sigma^2(n^{-1}\sum_{j=1}^{q}\lambda_{Q,j} + C_1\log(n)\{1 + [n^{-1}\sum_{j=1}^{q}\lambda_{Q,j}^2/\log(n)]^{1/2}\}), \\
&\sigma||\sqrt{\tilde{\lambda}}Q^T\{\mathbf{e}_1(\overline{\xi\boldsymbol{\epsilon}})^T + (\overline{\xi\boldsymbol{\epsilon}})\mathbf{e}_1^T\}Q||_{op} = C_3\sigma n^{-1/2}\sqrt{\tilde{\lambda}}||Q^T\mathbf{e}_1||_2,
\end{aligned} \tag{52}$$

where $||\cdot||_{op}$ is the spectral norm of a matrix and C_3 is a generic constant.

We consider an approximation of $\hat{\mathbf{v}}_1$ by using the eigenvector of $\tilde{S}_{Q^T\mathbf{x}}$, denoted as $\tilde{\mathbf{v}}_1$, corresponding to its largest eigenvalue. It follows from Lemma 6 that $\tilde{\mathbf{v}}_1$ is given by

$$\frac{(1-\hat{\rho}_{Q,12}x_{Q,+})}{\sqrt{1+x_{Q,+}^2(1-\hat{\rho}_{Q,12}^2)}}\frac{Q^T(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)}{||Q^T(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)||_2} + \frac{x_{Q,+}}{\sqrt{1+x_{Q,+}^2(1-\hat{\rho}_{Q,12}^2)}}\frac{Q^T\mathbf{e}_1}{||Q^T\mathbf{e}_1||_2}, \tag{53}$$

where $\hat{\rho}_{Q,12} = \mathbf{e}_1^TQQ^T(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)||Q^T(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)||_2^{-1}||Q^T\mathbf{e}_1||_2^{-1}$ and $x_{Q,+}$ is given by

$$x_{Q,+} = \frac{-\{K_{n,2}(2\hat{\rho}_{Q,12}^2 - 1) + K_{n,1}\} \pm \sqrt{(K_{n,1}-K_{n,2})^2 + 4K_{n,1}K_{n,2}\hat{\rho}_{Q,12}^2}}{2K_{n,2}\hat{\rho}_{Q,12}(1-\hat{\rho}_{Q,12}^2)}.$$

It follows from Theorem 2 of Yu et al. (2015) that if $\tilde{\mathbf{v}}_1^T\hat{\mathbf{v}}_1 \geq 0$, then with a large probability, we

have

$$||\tilde{\mathbf{v}}_1 - \hat{\mathbf{v}}_1||_2 \leq \frac{C||\tilde{S}_{Q^T\mathbf{x}} - S_{Q^T\mathbf{x}}||_{op}}{\sqrt{(K_{n,1} - K_{n,2})^2 + 4K_{n,1}K_{n,2}\hat{\rho}^2_{Q,12}}} = C(\frac{\sum_{j=1}^q \lambda_{Q,j}}{nK_{n,1}} + \frac{\log(n)}{K_{n,1}} + (nK_{n,1})^{-1/2}). \quad (54)$$

For Step (V), we proceed as follows. We first derive an approximation to $\hat{\mathbf{v}}_1^T Q^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$ and then approximate $\hat{\mathbf{v}}_1^T Q^T \Sigma Q \hat{\mathbf{v}}_1$. It follows from (53) that we have

$$\begin{aligned} \hat{\mathbf{v}}_1^T Q^T(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1) &= \tilde{\mathbf{v}}_1^T Q^T(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1) + (\hat{\mathbf{v}}_1 - \tilde{\mathbf{v}}_1)^T Q^T(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1) \\ &= [\frac{1}{\sqrt{1 + x^2_{Q,+}(1 - \hat{\rho}^2_{Q,12})}} + C||\hat{\mathbf{v}}_1 - \tilde{\mathbf{v}}_1||_2]||Q^T(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)||_2. \end{aligned}$$

Similarly, since $\hat{\mathbf{v}}_1^T Q^T \Sigma Q \hat{\mathbf{v}}_1 = \tilde{\lambda}(\hat{\mathbf{v}}_1^T Q^T \mathbf{e}_1)^2 + \sigma^2 ||Q\hat{\mathbf{v}}_1||_2^2$, we have

$$\hat{\mathbf{v}}_1^T Q^T \mathbf{e}_1 = \tilde{\mathbf{v}}_1^T Q^T \mathbf{e}_1 + (\hat{\mathbf{v}}_1 - \tilde{\mathbf{v}}_1)^T Q^T \mathbf{e}_1 = ||Q^T \mathbf{e}_1||_2 \{\frac{(1 - \hat{\rho}_{Q,12} x_{Q,+})\hat{\rho}_{Q,12} + x_{Q,+}}{\sqrt{1 + x^2_{Q,+}(1 - \hat{\rho}^2_{Q,12})}} + C||\hat{\mathbf{v}}_1 - \tilde{\mathbf{v}}_1||_2\}.$$

Therefore, with a high probability, we have

$$\hat{\mathbf{v}}_1^T Q^T \Sigma Q \hat{\mathbf{v}}_1 = \left(\tilde{\lambda}_1 ||Q^T \mathbf{e}_1||_2^2 \{\frac{(1 - \hat{\rho}_{Q,12} x_{Q,+})\hat{\rho}_{Q,12} + x_{Q,+}}{\sqrt{1 + x^2_{Q,+}(1 - \hat{\rho}^2_{Q,12})}}\}^2 + \sigma^2 ||Q\tilde{\mathbf{v}}_1||_2^2 \right) [1 + o(1)].$$

Moreover, if $\rho_{Q,12} \neq 0$, then it can be shown $\hat{\rho}_{Q,12} = \rho_{Q,12} + O(\frac{\log(n)}{\sqrt{n}})$ and

$$x_{Q,+} \to^p \frac{-\{R_0(2\rho^2_{Q,12} - 1) + 1\} \pm \sqrt{(1 - R_0)^2 + 4R_0\rho^2_{Q,12}}}{2R_0\rho_{Q,12}(1 - \rho^2_{Q,12})}.$$

By combining the results in Steps (I)-(V), we can finish the proof.

References

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006), "Prediction by supervised principal components." *J. Amer. Statist. Assoc.*, 101, 119–137.

Basak, D., Pal, S., and Patranabis, D. C. (2007), "Support vector regression," *Neural Information Processing-Letters and Reviews*, 11, 203–224.

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear smoothers and additive models," *The Annals of Statistics*, 453–510.

Chang, C.-C. and Lin, C.-J. (2011), "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2, 27.

Chun, H. and Keles, S. (2010), "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," *J. Roy. Statist. Soc. Ser. B*, 72,, 3–25.

Chung, D., Chun, H., and Keles, S. (2012), "An Introduction to the 'spls' Package, Version 1.0," .

Fan, J. and Fan, Y. (2008), "High-dimensional classification using features annealed independence rules," *Ann. Statist.*, 36, 2605–2637.

Fan, J., Fan, Y., and Lv, J. (2008), "High dimensional covariance matrix estimation using a factor model," *Journal of Econometrics*, 147, 186–197.

Fan, J., Feng, Y., Samworth, R., and Wu, Y. (2010), "SIS: Sure Independence Screening. R package version 0.6," .

Fan, J. and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 849–911.

Jung, S. and Marron, J. S. (2009), "PCA Consistency in High Dimension, Low Sample Size Context," *The Annals of Statistics*, 37, 4104–4130.

Little, G. and Reade, J. B. (1984), "Eigenvalues of analytic kernels," *SIAM J. Math. Anal.*, 15, 133–136.

Liu, J., Ji, S., and Ye, J. (2009), "SLEP: Sparse learning with efficient projections," *Arizona State University*, 6.

Mai, Q., Zou, H., and Yuan, M. (2012), "A direct approach to sparse discriminant analysis in ultra-high dimensions," *Biometrika*, 99, 29–42.

Paul, D. (2007), "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," *Statistica Sinica*, 17, 1617–1642.

Rao, C. R. (2002), *Linear statistical inference and its applications*, John Willey & Sons, New York.

Reade, J. (1984), "Eigenvalues of positive definite kernels II." *SIAM Journal on Mathematical Analysis*, 15, 137–142.

Skocaj, D., Leonardis, A., and Bischof, H. (2007), "Weighted and robust learning of subspace representations," *Pattern Recogn.*, 40, 1556–1569.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, 58, 267–288.

Tropp, J. A. (2012), "User-friendly tail bounds for sums of random matrices," *Found Computational Mathematics*, 12, 389–434.

Van Hemmen, J. and Ando, T. (1980), "An inequality for trace ideals," *Communications in Mathematical Physics*, 76, 143–148.

Yu, Y., Wang, T., and Samworth, R. (2015), "A useful variant of the Davis–Kahan theorem for statisticians," *Biometrika*, 102, 315–323.

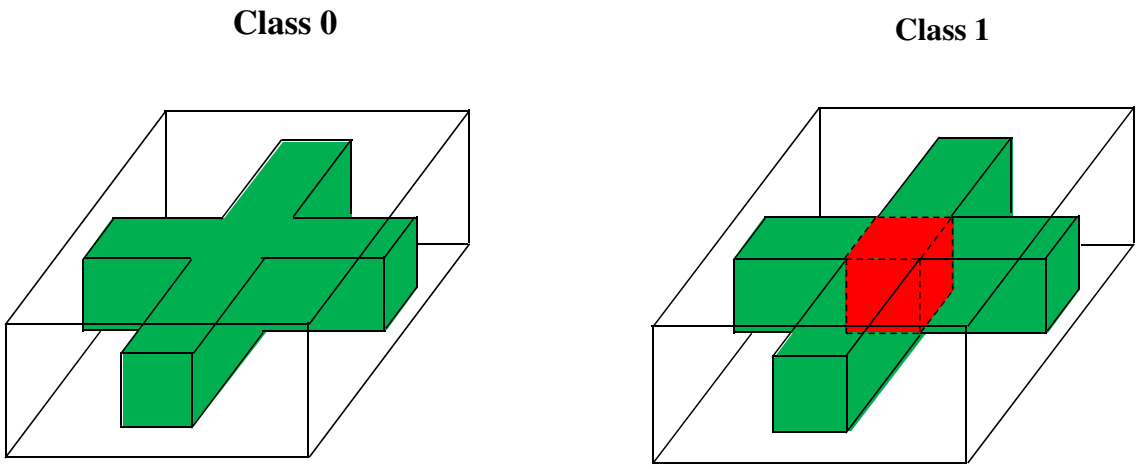

Figure 1: True mean images for the first set of simulations: Class 0 in the left panel and Class 1 in the right panel. The white, green, and red colors, respectively, correspond to 0, 1, and 2.

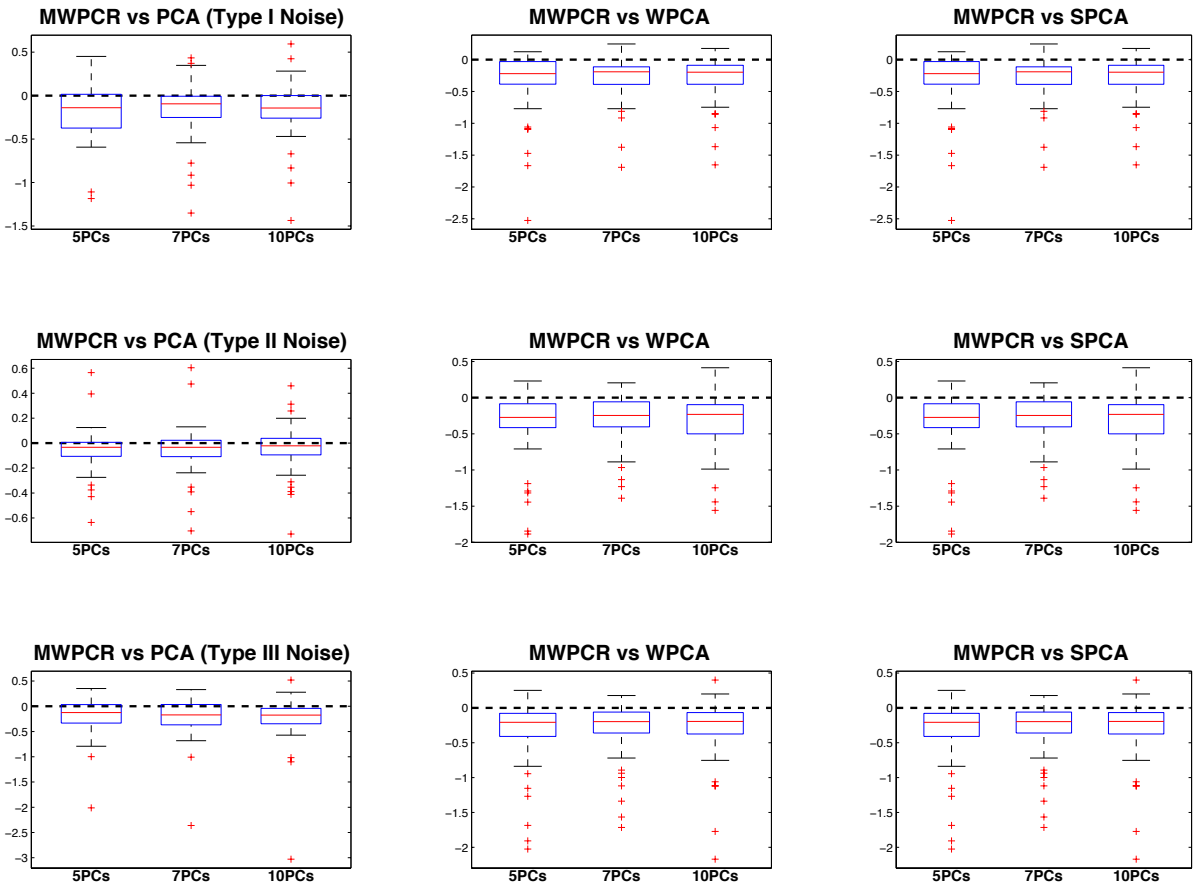

Figure 2: Simulation results for the second set of simulations: comparison of MWPCR and the three dimension reduction methods including principal component analysis (PCA), weighted PCA (WPCA), and supervised PCA (SPCA) for the three types of noise. All panels show the box plots of the prediction error differences between MWPCR and PCA (or WPCA, or SPCA) for three different numbers of principal components. The error differences are almost less than 0 (below the dished line) and confirm that MWPCR outperforms PCA, WPCA, and SPCA.

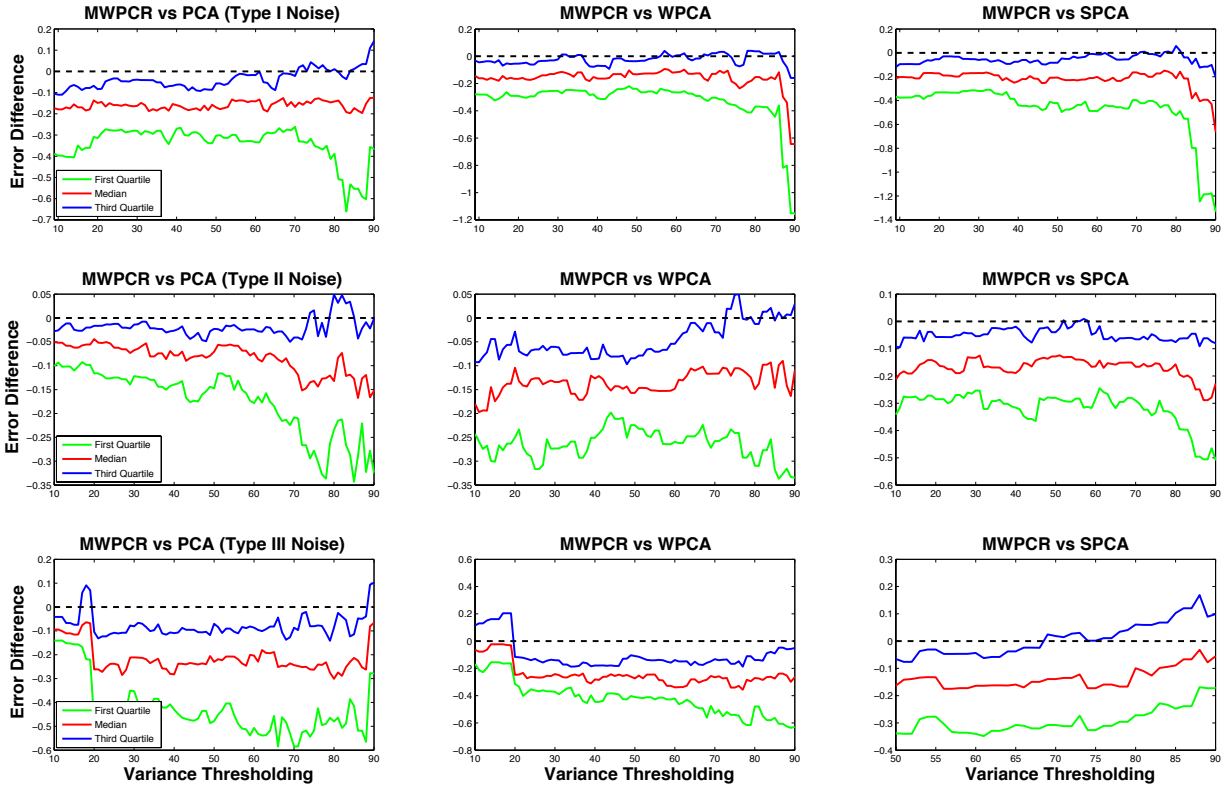

Figure 3: Simulation results for the second set of simulations: comparison of MWPCR and PCA, WPCA, and SPCA based on the variance thresholding for the three types of noise. In all panels, the green, red and blue curves are respectively the first, the second, and the third quantiles of error differences between MWPCR and PCA (or WPCA, or SPCA) for different variance thresholding. These results further confirm that MWPCR outperforms PCA, WPCA, and SPCA.

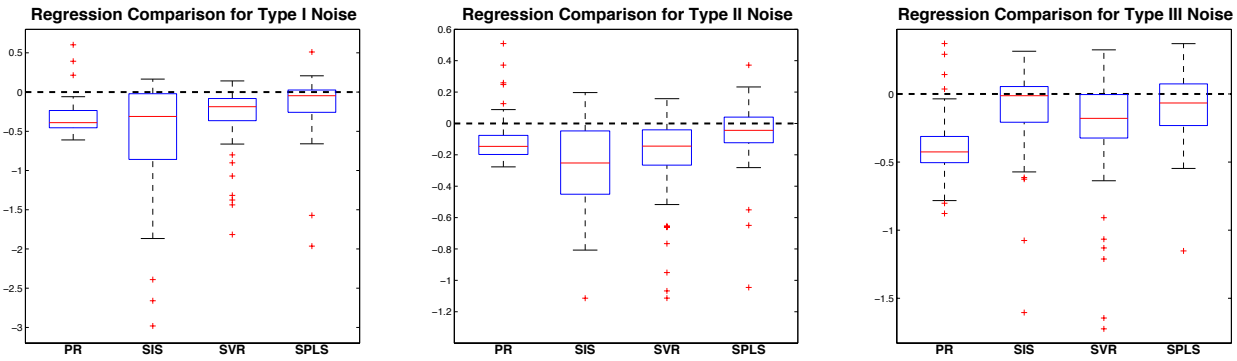

Figure 4: Simulation results for the second set of simulations: the box plots of the prediction error differences between MWPCR and other high-dimensional regression models including Lasso, SIS, SVR, and SPLS. The error differences are almost less than 0 (below the dished line) and confirm that MWPCR outperforms Lasso, SIS, SVR, and SPLS.