

LEAST FAVORABLE DIRECTION TEST FOR MULTIVARIATE ANALYSIS OF VARIANCE IN HIGH DIMENSION

Author(s)

Affiliation(s)

Abstract: This paper considers the problem of multivariate analysis of variance for normal samples. When the sample dimension is larger than the sample size, the classical likelihood ratio test is not defined since the likelihood function is unbounded. Based on the unboundedness of the likelihood function, we propose a new test. The asymptotic null distribution of the test statistic is derived and the local asymptotic power function of the test is also given. In our results, the covariance matrix may have a few large eigenvalues which can characterize the strong correlations between variables. The asymptotic power function and simulations show that the proposed test has particular high power when there are strong correlations between variables.

Key words and phrases: High dimensional data, least favorable direction test, multivariate analysis of variance, principal component analysis, spiked covariance.

1. Introduction

Suppose there are k ($k \geq 2$) groups of p dimensional data. Within the i th group ($1 \leq i \leq k$), we have observations $\{X_{ij}\}_{j=1}^{n_i}$ which are independent and identically distributed (i.i.d.) as $N_p(\xi_i, \mathbf{\Sigma})$, the p dimensional normal distribution with mean vector ξ_i and common variance matrix $\mathbf{\Sigma}$. We would like to test the hypotheses

$$H_0 : \xi_1 = \xi_2 = \cdots = \xi_k \quad \text{v.s.} \quad H_1 : \xi_i \neq \xi_j \text{ for some } i \neq j. \quad (1.1)$$

This testing problem is known as one-way multivariate analysis of variance (MANOVA) and has been well studied when p is small compared to n , where $n = \sum_{i=1}^k n_i$ is the total sample size.

Let $\mathbf{H} = \sum_{i=1}^k n_i(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})^T$ be the sum-of-squares between groups and $\mathbf{G} = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{\mathbf{X}}_i)(X_{ij} - \bar{\mathbf{X}}_i)^T$ be the sum-of-squares within groups, where $\bar{\mathbf{X}}_i = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}$ is the sample mean of group i and $\bar{\mathbf{X}} = n^{-1} \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}$ is the pooled sample mean. There are four classical test statistics for hypothesis (1.1), which are all based on the eigenvalues of $\mathbf{H}\mathbf{G}^{-1}$.

Wilks' Lambda:	$ \mathbf{G} + \mathbf{H} / \mathbf{G} $
Pillai trace:	$\text{tr}[\mathbf{H}(\mathbf{G} + \mathbf{H})^{-1}]$
Hotelling-Lawley trace:	$\text{tr}[\mathbf{H}\mathbf{G}^{-1}]$
Roy's maximum root:	$\lambda_{\max}(\mathbf{H}\mathbf{G}^{-1})$

In some modern scientific applications, people would like to test hypothesis (1.1) in high dimensional setting, i.e., p is greater than n . See, for example, Verstyne et al. (2005) and Tsai and Chen (2009). However, when $p \geq n$, the four classical test statistics are not defined. Researchers have done extensive work to study the testing problem (1.1) in high dimensional setting. So far, most tests are designed for two sample case, i.e., $k = 2$. See, for example, Bai and Saranadasa (1996), Srivastava (2007), Chen and Qin (2010), Cai et al. (2013) and Feng et al. (2014). For multiple sample case, Schott (2007) modified Hotelling-Lawley trace and proposed the test statistic

$$T_{SC} = \frac{1}{\sqrt{n-1}} \left(\frac{1}{k-1} \text{tr}(\mathbf{H}) - \frac{1}{n-k} \text{tr}(\mathbf{G}) \right).$$

Statistic T_{SC} is a representative of the so-called sum-of-squares type statistics as it is based on an estimation of squared Euclidean norm $\sum_{i=1}^k n_i \|\xi_i - \bar{\xi}\|^2$, where $\bar{\xi} = n^{-1} \sum_{i=1}^k n_i \xi_i$. See Srivastava and Kubokawa (2013), Yamada and Himeno (2015) and Bu Zhou (2017) for some other sum-of-squares type test statistics.

In another work, Cai and Xia (2014) proposed a test statistic

$$T_{CX} = \max_{1 \leq i \leq p} \sum_{1 \leq j < l \leq k} \frac{n_j n_l}{n_j + n_l} \frac{(\Omega(\bar{\mathbf{X}}_j - \bar{\mathbf{X}}_l))_i^2}{\omega_{ii}},$$

Where $\Omega = (\omega)_{ij} = \Sigma^{-1}$ is the precision matrix. When Ω is unknown, they

substitute it by an estimator $\hat{\Omega}$. Unlike T_{SC} , T_{CX} is an extreme value type statistic.

The likelihood ratio test (LRT) method has been very successful in leading to satisfactory procedures in many specific problems. However, the LRT statistic for hypotheses (1.1), i.e., Wilks' Lambda, is not defined for $p > n - k$. In high dimensional setting, both sum-of-squares type statistics and extreme value type statistics are not based on likelihood function. We try to construct a likelihood-based test in high dimensional setting. In a recent work, Zhao and Xu (2016) proposed a generalized likelihood ratio test in the context of one-sample mean vector test. They used a least favorable argument to construct a generalized likelihood ratio test statistic. Their simulation results showed that their test has good power performance, especially when the variables are correlated.

In this paper, we propose a generalized likelihood ratio test statistic for hypotheses (1.1), we call it least favorable direction (LFD) statistic. The asymptotic distributions of the test statistic are derived. In our results, the covariance matrix may have r significantly large eigenvalues. This covariance structure, known as spiked covariance, can characterize the strong correlations between variables. See, for example, Fan et al. (2008), Cai et al. (2013), Shen et al. (2013) and Ma et al. (2015). The asymptotic null dis-

tribution of the proposed test statistic involves some unknown parameters. We give their estimators and formulate a test with asymptotically correct level. The asymptotic local power function of LFD test is also given. The asymptotic power of LFD test doesn't depend on the large eigenvalues of the covariance matrix. For most existing tests, however, the asymptotic power decreases as the large eigenvalues of the covariance matrix increases. Hence LFD test is particularly powerful when there are strong correlations between variables. Further simulations show the good performance of LFD test.

The rest of the paper is organized as follows. In Section 2, we propose LFD test and give the asymptotic distributions of LFD test. Section 3 complements our study with numerical simulations. In Section 4, we give a short discussion. Finally, the proofs are gathered in the Appendix.

2. Least favorable direction test

We introduce some notations. Define the $p \times n$ pooled sample matrix \mathbf{X} as

$$\mathbf{X} = (X_{11}, X_{12}, \dots, X_{1n_1}, X_{21}, X_{22}, \dots, X_{2n_2}, \dots, X_{k1}, X_{k2}, \dots, X_{kn_k}).$$

The sum-of-squares within groups \mathbf{G} can be written as $\mathbf{G} = \mathbf{X}(\mathbf{I}_n - \mathbf{J}\mathbf{J}^T)\mathbf{X}^T$

where

$$\mathbf{J} = \begin{pmatrix} \frac{1}{\sqrt{n_1}}\mathbf{1}_{n_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\sqrt{n_2}}\mathbf{1}_{n_2} & \mathbf{0} \\ \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \frac{1}{\sqrt{n_k}}\mathbf{1}_{n_k} \end{pmatrix}$$

is an $n \times k$ matrix and $\mathbf{1}_{n_i}$ is an n_i -dimensional vector with all elements equal to 1, $i = 1, \dots, k$. Construct an $n \times (n - k)$ matrix $\tilde{\mathbf{J}}$ as

$$\tilde{\mathbf{J}} = \begin{pmatrix} \tilde{\mathbf{J}}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{J}}_2 & \mathbf{0} \\ \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \tilde{\mathbf{J}}_k \end{pmatrix},$$

where $\tilde{\mathbf{J}}_i$ is an $n_i \times (n_i - 1)$ matrix defined as

$$\tilde{\mathbf{J}}_i = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \cdots & \frac{1}{\sqrt{(n_i-2)(n_i-1)}} & \frac{1}{\sqrt{(n_i-1)n_i}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \cdots & \frac{1}{\sqrt{(n_i-2)(n_i-1)}} & \frac{1}{\sqrt{(n_i-1)n_i}} \\ 0 & -\frac{2}{\sqrt{6}} & \cdots & \vdots & \vdots \\ \vdots & \vdots & \cdots & -\frac{n_i-2}{\sqrt{(n_i-2)(n_i-1)}} & \frac{1}{\sqrt{(n_i-1)n_i}} \\ 0 & 0 & \cdots & 0 & -\frac{n_i-1}{\sqrt{(n_i-1)n_i}} \end{pmatrix}.$$

Then $\tilde{\mathbf{J}}$ is a column orthogonal matrix satisfying $\tilde{\mathbf{J}}^T\tilde{\mathbf{J}} = \mathbf{I}_{n-k}$ and $\tilde{\mathbf{J}}\tilde{\mathbf{J}}^T =$

$\mathbf{I}_n - \mathbf{J}\mathbf{J}^T$. Let $\mathbf{Y} = \mathbf{X}\tilde{\mathbf{J}}$ be a $p \times (n - k)$ random matrix. Then \mathbf{G} can be

written as

$$\mathbf{G} = \mathbf{Y}\mathbf{Y}^T.$$

The sum-of-squares between groups \mathbf{H} can be written as

$$\mathbf{H} = \mathbf{X}(\mathbf{J}\mathbf{J}^T - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{X}^T = \mathbf{X}\mathbf{J}(\mathbf{I}_k - \frac{1}{n}\mathbf{J}^T\mathbf{1}_n\mathbf{1}_n^T\mathbf{J})\mathbf{J}^T\mathbf{X}^T.$$

By some matrix algebra, $\mathbf{I}_k - \frac{1}{n}\mathbf{J}^T\mathbf{1}_n\mathbf{1}_n^T\mathbf{J} = \mathbf{C}\mathbf{C}^T$ where \mathbf{C} is a $k \times (k-1)$

matrix defined as $\mathbf{C} = \mathbf{C}_1\mathbf{C}_2$, and

$$\mathbf{C}_1 = \begin{pmatrix} \sqrt{n_1} & \sqrt{n_1} & \cdots & \sqrt{n_1} & \sqrt{n_1} \\ -\frac{n_1}{\sqrt{n_2}} & \sqrt{n_2} & \cdots & \sqrt{n_2} & \sqrt{n_2} \\ 0 & -\frac{n_1+n_2}{\sqrt{n_3}} & \cdots & \vdots & \vdots \\ \vdots & \vdots & \cdots & -\frac{\sum_{i=1}^{k-2} n_i}{\sqrt{n_{k-1}}} & \sqrt{n_{k-1}} \\ 0 & 0 & \cdots & 0 & -\frac{\sum_{i=1}^{k-1} n_i}{\sqrt{n_k}} \end{pmatrix},$$

$$\mathbf{C}_2 = \begin{pmatrix} \frac{n_1(n_1+n_2)}{n_2} & 0 & \cdots & 0 \\ 0 & \frac{(\sum_{i=1}^2 n_i)(\sum_{i=1}^3 n_i)}{n_3} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \frac{(\sum_{i=1}^{k-1} n_i)(\sum_{i=1}^k n_i)}{n_k} \end{pmatrix}^{-\frac{1}{2}}.$$

Then \mathbf{H} can be written as

$$\mathbf{H} = \mathbf{X}\mathbf{J}\mathbf{C}\mathbf{C}^T\mathbf{J}^T\mathbf{X}^T.$$

Define $\boldsymbol{\Theta} = (\sqrt{n_1}\xi_1, \dots, \sqrt{n_k}\xi_k)$ and the null hypothesis H_0 is equivalent to

$\boldsymbol{\Theta}\mathbf{C} = \mathbf{O}_{p \times (k-1)}$, where $\mathbf{O}_{p \times (k-1)}$ is a $p \times (k-1)$ matrix with all elements

equal to 0. Thus, the hypotheses in (1.1) are equivalent to

$$H_0 : \boldsymbol{\Theta}\mathbf{C} = \mathbf{O}_{p \times (k-1)} \quad \text{v.s.} \quad H_1 : \boldsymbol{\Theta}\mathbf{C} \neq \mathbf{O}_{p \times (k-1)}.$$

In low dimensional setting, the testing problem (1.1) is well studied. A classical test statistic is Roy's maximum root which is constructed by Roy (1953) using his well-known union intersection principle. The key idea is to decompose data \mathbf{X} into a set of univariate data $\{\mathbf{X}_a = a^T \mathbf{X} : a \in \mathbb{R}^p, a^T a = 1\}$. This induces a decomposition of the null hypothesis and the alternative hypothesis:

$$H_0 = \bigcap_{a \in \mathbb{R}^p, a^T a = 1} H_{0a} \quad \text{v.s.} \quad H_1 = \bigcup_{a \in \mathbb{R}^p, a^T a = 1} H_{1a},$$

where $H_{0a} : a^T \boldsymbol{\Theta}\mathbf{C} = \mathbf{O}_{1 \times (k-1)}$ and $H_{1a} : a^T \boldsymbol{\Theta}\mathbf{C} \neq \mathbf{O}_{1 \times (k-1)}$. Let $L_0(a)$ and $L_1(a)$ be the maximum likelihood of \mathbf{X}_a under H_{0a} and H_{1a} , respectively.

For each a satisfying $a^T a = 1$, the LRT statistic

$$\frac{L_1(a)}{L_0(a)} = \left(\frac{a^T (\mathbf{G} + \mathbf{H})a}{a^T \mathbf{G}a} \right)^{n/2}$$

can be used to test H_{0a} v.s. H_{1a} . Using union intersection principle, Roy proposed the test statistic $\max_{a^T a = 1} L_1(a)/L_0(a) = \lambda_{\max}^{n/2}(\mathbf{H}\mathbf{G}^{-1})$, where $\lambda_{\max}(\cdot)$ means the largest eigenvalue. This statistic is an increasing function of Roy's maximum root.

From a likelihood point of view, log likelihood ratio is an estimator of the Kullback-Leibler divergence between the true distribution and the null

distribution. Hence the component LRT statistic $L_1(a)/L_0(a)$ characterizes the discrepancy between the true distribution and the null distribution along direction a . Hopefully, the direction $a^* = \arg \max_{a^T a=1} L_1(a)/L_0(a)$ achieves the largest discrepancy between the true distribution and the null distribution. Thus, H_{0a^*} is the component null hypothesis most likely to be not true. We shall call a^* the least favorable direction. Roy's maximum root is in fact the component LRT statistic along the least favorable direction.

Unfortunately, Roy's maximum root can only be defined when $n-k \geq p$, hence can not be used in the high dimensional setting. In what follows, we assume $p > n - k$. In this case, Roy's maximum root is not defined since

$$\mathcal{A} \stackrel{def}{=} \{a : L_1(a) = +\infty, a^T a = 1\} = \{a : a^T \mathbf{G} a = 0, a^T a = 1\}$$

is not empty. Note that

$$\{a : L_0(a) < +\infty, a^T a = 1\} = \{a : a^T (\mathbf{G} + \mathbf{H}) a \neq 0, a^T a = 1\}.$$

By the independence of \mathbf{G} and \mathbf{H} , with probability 1, we have $\mathcal{A} \cap \{a : L_0(a) < +\infty, a^T a = 1\} \neq \emptyset$. This suggests that the least favorable direction a^* , which hopefully maximizes the discrepancy between $L_1(a)$ and $L_0(a)$, should be defined as $a^* = \arg \min_{a \in \mathcal{A}} L_0(a)$. Equivalently,

$$a^* = \arg \min_{a \in \mathcal{A}} L_0(a) = \arg \max_{a^T a=1, a^T \mathbf{G} a=0} a^T \mathbf{H} a.$$

This motivates us to propose a new test statistic as

$$T(\mathbf{X}) = a^{*T} \mathbf{H} a^* = \max_{a^T \mathbf{a}=1, a^T \mathbf{G} a=0} a^T \mathbf{H} a.$$

We reject the null hypothesis when $T(\mathbf{X})$ is large enough. We shall call $T(\mathbf{X})$ the least favorable direction (LFD) test statistic. Since the least favorable direction a^* is obtained from likelihood function, the statistic $T(\mathbf{X})$ is also a generalized likelihood ratio test statistic.

Now we derive the explicit forms of LFD test statistic. Let $\mathbf{Y} = \mathbf{U}_Y \mathbf{D}_Y \mathbf{V}_Y^T$ be the singular value decomposition of \mathbf{Y} , where \mathbf{U}_Y and \mathbf{V}_Y are $p \times (n-k)$ and $(n-k) \times (n-k)$ both column orthogonal matrices, \mathbf{D}_Y is an $(n-k) \times (n-k)$ diagonal matrix. Let $\mathbf{P}_Y = \mathbf{U}_Y \mathbf{U}_Y^T$ be the projection matrix on the column space of \mathbf{Y} . Then Proposition 4 in Appendix implies that

$$T(\mathbf{X}) = \lambda_{\max}(\mathbf{C}^T \mathbf{J}^T \mathbf{X}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{X} \mathbf{J} \mathbf{C}). \quad (2.2)$$

Next, we derive another simple form of $T(\mathbf{X})$. By the relationship

$$\begin{pmatrix} \mathbf{J}^T \mathbf{X}^T \mathbf{X} \mathbf{J} & \mathbf{J}^T \mathbf{X}^T \mathbf{X} \tilde{\mathbf{J}} \\ \tilde{\mathbf{J}}^T \mathbf{X}^T \mathbf{X} \mathbf{J} & \tilde{\mathbf{J}}^T \mathbf{X}^T \mathbf{X} \tilde{\mathbf{J}} \end{pmatrix}^{-1} = \left(\begin{pmatrix} \mathbf{J}^T \\ \tilde{\mathbf{J}}^T \end{pmatrix} \mathbf{X}^T \mathbf{X} \begin{pmatrix} \mathbf{J} & \tilde{\mathbf{J}} \end{pmatrix} \right)^{-1} = \begin{pmatrix} \mathbf{J}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{J} & \mathbf{J}^T (\mathbf{X}^T \mathbf{X})^{-1} \tilde{\mathbf{J}} \\ \tilde{\mathbf{J}}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{J} & \tilde{\mathbf{J}}^T (\mathbf{X}^T \mathbf{X})^{-1} \tilde{\mathbf{J}} \end{pmatrix}$$

and matrix inverse formula, we have that

$$(\mathbf{J}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{J})^{-1} = \mathbf{J}^T \mathbf{X}^T \mathbf{X} \mathbf{J} - \mathbf{J}^T \mathbf{X}^T \mathbf{X} \tilde{\mathbf{J}} (\tilde{\mathbf{J}}^T \mathbf{X}^T \mathbf{X} \tilde{\mathbf{J}})^{-1} \tilde{\mathbf{J}}^T \mathbf{X}^T \mathbf{X} \mathbf{J} = \mathbf{J}^T \mathbf{X}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{X} \mathbf{J}.$$

Thus,

$$T(\mathbf{X}) = \lambda_{\max} \left(\mathbf{C}^T (\mathbf{J}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{J})^{-1} \mathbf{C} \right). \quad (2.3)$$

Compared with (2.2), (2.3) doesn't involve $\mathbf{P}_{\mathbf{Y}}$. Hence (2.3) is convenient for computation.

The case $k = 2$ and the least favorable direction.

Now we derive the asymptotic distribution of LFD test statistic. We are especially interested in the case when variables are correlated. For some real world problems, variables are heavily correlated with common factors, then the covariance matrix Σ is spiked in the sense that a few eigenvalues of Σ are significantly larger than the others (Fan et al., 2008; Cai et al., 2013; Shen et al., 2013; Ma et al., 2015). To characterize this correlation pattern, we make the following assumption for the eigenvalues of Σ .

Assumption 1. *Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ be the eigenvalues of Σ . Suppose the first r ($r \geq 0$) eigenvalues are significantly larger than the others. We assume $r = o(n)$. We assume $c_1 \geq \lambda_{r+1} \geq \dots \geq \lambda_p \geq c_2$ for some positive absolute constants c_1 and c_2 . If $r \neq 0$, we assume*

$$\frac{\lambda_r n}{p} \rightarrow \infty, \quad \frac{\lambda_1^2 p r^2}{\lambda_r^2 n^2} \rightarrow 0 \quad \frac{\log \lambda_r}{n} \rightarrow 0.$$

Remark 1. The spiked covariance model is commonly assumed in the study of PCA theory. Most existing works assumed r is fixed. Here we allow r to

vary as a smaller order of n . The condition $\boldsymbol{\lambda}_r n/p \rightarrow \infty$ requires $\boldsymbol{\lambda}_r$ to be much larger than p/n . This is satisfied, for example, for the factor model adopted by Ma et al. (2015). The harshest condition is $\boldsymbol{\lambda}_1^2 p r^2 / (\boldsymbol{\lambda}_r^2 n^2) \rightarrow 0$. If $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_r$ are of the same order and r is fixed, this condition is equivalent to $p/n^2 \rightarrow 0$. We require this condition since the PCA consistency results are not valid when p is too large. See, for example, (Cai et al., 2013). If $r > 0$, this condition is unavoidable and the asymptotic behavior of $T(\mathbf{X})$ is different if this condition is violated.

To establish the asymptotic distribution of $T(\mathbf{X})$ under Assumption 1, we need following notations. Let \mathbf{W}_{k-1} be a $(k-1) \times (k-1)$ symmetric random matrix whose entries above the main diagonal are i.i.d. $N(0, 1)$ and the entries on the diagonal are i.i.d. $N(0, 2)$. Let $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ denote the eigenvalue decomposition of $\boldsymbol{\Sigma}$, where $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_p)$. We denote $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)$ where \mathbf{U}_1 is $p \times r$ and \mathbf{U}_2 is $p \times (p-r)$. Denote $\boldsymbol{\Lambda}_1 = \text{diag}(\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_r)$ and $\boldsymbol{\Lambda}_2 = \text{diag}(\boldsymbol{\lambda}_{r+1}, \dots, \boldsymbol{\lambda}_p)$. Then $\boldsymbol{\Sigma} = \mathbf{U}_1\boldsymbol{\Lambda}_1\mathbf{U}_1^T + \mathbf{U}_2\boldsymbol{\Lambda}_2\mathbf{U}_2^T$.

The following theorem establishes the asymptotic distribution of LFD test statistic.

Theorem 1. *Under Assumption 1, suppose $p/n \rightarrow \infty$ and*

$$\text{tr} \left(\mathbf{\Lambda}_2 - \frac{1}{p-r} \text{tr}(\mathbf{\Lambda}_2) \mathbf{I}_{p-r} \right)^2 = o\left(\frac{p}{n}\right).$$

Then under local alternative hypothesis

$$\frac{1}{\sqrt{p}} \|\mathbf{\Theta C}\|_F^2 = O(1),$$

we have

$$\frac{T(\mathbf{X}) - \frac{p-r-n+k}{p-r} \text{tr}(\mathbf{\Lambda}_2)}{\sqrt{\text{tr}(\mathbf{\Lambda}_2^2)}} \sim \lambda_{\max} \left(\mathbf{W}_{k-1} + \frac{1}{\sqrt{\text{tr}(\mathbf{\Lambda}_2^2)}} \mathbf{C}^T \mathbf{\Theta}^T (\mathbf{I}_p - \mathbf{P}_{\mathbf{Y}}) \mathbf{\Theta C} \right) + o_P(1),$$

where \sim means having the same distribution.

To gain some insight into the asymptotic behavior of $T(\mathbf{X})$, suppose the null hypothesis holds and $k = 2$, then Theorem 1 implies that

$$\frac{T(\mathbf{X}) - \frac{p-r-n+k}{p-r} \text{tr}(\mathbf{\Lambda}_2)}{\sqrt{2 \text{tr}(\mathbf{\Lambda}_2^2)}} \xrightarrow{\mathcal{L}} N(0, 1).$$

The asymptotic variance of $T(\mathbf{X})$ is $2 \text{tr}(\mathbf{\Lambda}_2^2)$. If $r = 0$, this equals to $2 \text{tr}(\mathbf{\Sigma}^2)$ which is the asymptotic variance of Bai and Saranadasa (1996) and Chen and Qin (2010)'s statistic. While for $r > 0$, $2 \text{tr}(\mathbf{\Lambda}_2^2)$ is smaller than $2 \text{tr}(\mathbf{\Sigma}^2)$. In fact, if $\liminf_{n \rightarrow \infty} \lambda_1/p \in (0, +\infty]$, we have

$$\liminf_{n \rightarrow \infty} \frac{2 \text{tr}(\mathbf{\Sigma}^2)}{2 \text{tr}(\mathbf{\Lambda}_2^2)} \in (1, +\infty].$$

The reason for this is because the projection $\mathbf{I}_p - \mathbf{P}_{\mathbf{Y}}$ appeared in expression (2.2) can remove large variance terms of \mathbf{XJC} .

To formulate a test procedure with asymptotically correct level, the parameters r , $\text{tr}(\mathbf{\Lambda}_2)$ and $\text{tr}(\mathbf{\Lambda}_2^2)$ should be estimated. We use the following statistic to estimate r :

$$\hat{r} = \begin{cases} \arg \max_{1 \leq i \leq n-k-1} \frac{\lambda_i(\mathbf{Y}^T \mathbf{Y})}{\lambda_{i+1}(\mathbf{Y}^T \mathbf{Y})} \geq \gamma_n & \text{if } \max_{1 \leq i \leq n-k-1} \frac{\lambda_i(\mathbf{Y}^T \mathbf{Y})}{\lambda_{i+1}(\mathbf{Y}^T \mathbf{Y})} \geq \gamma_n \\ 0 & \text{otherwise} \end{cases}$$

where γ_n slowly tends to $+\infty$ as $n \rightarrow \infty$. The following proposition establishes the consistency of \hat{r} .

Proposition 1. *Suppose $p/n \rightarrow \infty$, $r = o(n)$, $\boldsymbol{\lambda}_r n/p \rightarrow \infty$ and $c_1 \geq \lambda_{r+1} \geq \dots \geq \lambda_p \geq c_2$. If $\gamma_n \rightarrow \infty$ and $\gamma_n = o(n\boldsymbol{\lambda}_r/p)$, then $\Pr(\hat{r} = r) \rightarrow 1$.*

Remark 2. For the factor model adopted by Ma et al. (2015), λ_r is of order p . Then we can take $\gamma_n = \sqrt{n}$.

We use the following statistic to estimate $\text{tr}(\mathbf{\Lambda}_2)$:

$$\widehat{\text{tr}(\mathbf{\Lambda}_2)} = \frac{1}{n-k} \sum_{i=\hat{r}+1}^{n-k} \lambda_i(\mathbf{Y}^T \mathbf{Y}).$$

Proposition 2. *Under the assumptions of Theorem 1, suppose $\gamma_n \rightarrow \infty$ and $\gamma_n = o(n\boldsymbol{\lambda}_r/p)$, then*

$$\widehat{\text{tr}(\mathbf{\Lambda}_2)} = \text{tr}(\mathbf{\Lambda}_2) + o_P(\sqrt{p}).$$

To estimate $\text{tr}(\mathbf{\Lambda}_2^2)$, we use the idea of leave-two-out. Let $\mathbf{Y}_{(i,j)}$ be the matrix obtained by deleting the i th and j th columns from \mathbf{Y} . Let

$\mathbf{Y}_{(i,j)} = \mathbf{U}_{\mathbf{Y};(i,j)} \mathbf{D}_{\mathbf{Y};(i,j)} \mathbf{V}_{\mathbf{Y};(i,j)}^T$ be the singular value decomposition of $\mathbf{Y}_{(i,j)}$.

Here $\mathbf{U}_{\mathbf{Y};(i,j)}$ is $p \times (n-k-2)$. Let $\tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}$ be a $p \times (p-n+k+2)$ orthogonal matrix satisfying $\tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T = \mathbf{I}_p - \mathbf{U}_{\mathbf{Y};(i,j)} \mathbf{U}_{\mathbf{Y};(i,j)}^T$.

Let w_{ij} be the (i, j) th element of $\mathbf{Y}^T \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T \mathbf{Y}$. Define

$$\widehat{\text{tr}(\Lambda_2^2)} = \frac{2}{(n-k)(n-k-1)} \sum_{1 \leq i < j \leq n-k} w_{ij}^2.$$

We use $\widehat{\text{tr}(\Lambda_2^2)}$ to estimate $\text{tr}(\Lambda_2^2)$. The following proposition shows that $\widehat{\text{tr}(\Lambda_2^2)}$ is a ratio consistent estimator.

Proposition 3. *Under the assumptions of Theorem 1, we have*

$$\frac{\widehat{\text{tr}(\Lambda_2^2)}}{\text{tr}(\Lambda_2^2)} \xrightarrow{P} 1.$$

Now we can construct LFD test procedure with asymptotic correct level α . Let

$$Q = \frac{T(\mathbf{X}) - \frac{p-\hat{r}-n+k}{p-\hat{r}} \widehat{\text{tr}(\Lambda_2^2)}}{\sqrt{\widehat{\text{tr}(\Lambda_2^2)}}}.$$

Let $F(x)$ be the cumulative distribution function of $\lambda_{\max}(\mathbf{W}_{k-1})$. LFD test reject the null hypothesis if $Q > F^{-1}(1 - \alpha)$.

Theorem 1, Proposition 2 and Proposition 3 implies that the resulting test procedure has asymptotic correct level under the assumptions of Theorem 1. And by Theorem 1, the asymptotic local power function of LFD

test procedure is

$$\Pr \left(\lambda_{\max}(\mathbf{W}_{k-1} + \frac{1}{\sqrt{\text{tr}(\mathbf{\Lambda}_2^2)}} \mathbf{C}^T \mathbf{\Theta}^T (\mathbf{I}_p - \mathbf{P}_{\mathbf{Y}}) \mathbf{\Theta} \mathbf{C}) \geq F_{\mathbf{W}}^{-1}(1 - \alpha) \right).$$

If $k = 2$, the asymptotic local power function of Bai and Saranadasa (1996) and Chen and Qin (2010)'s method can be written as

$$\Pr \left(\lambda_{\max}(\mathbf{W}_1 + \frac{1}{\sqrt{\text{tr}(\mathbf{\Sigma}^2)}} \mathbf{C}^T \mathbf{\Theta}^T \mathbf{\Theta} \mathbf{C}) \geq F_{\mathbf{W}}^{-1}(1 - \alpha) \right).$$

Hence the asymptotic relative efficiency between LFD test and Bai and Saranadasa (1996) and Chen and Qin (2010)'s method is

$$\sqrt{\frac{\text{tr}(\mathbf{\Sigma}^2)}{\text{tr}(\mathbf{\Lambda}_2^2)} \frac{\mathbf{C}^T \mathbf{\Theta}^T (\mathbf{I}_p - \mathbf{P}_{\mathbf{Y}}) \mathbf{\Theta} \mathbf{C}}{\mathbf{C}^T \mathbf{\Theta}^T \mathbf{\Theta} \mathbf{C}}}.$$

There's a random term $\mathbf{P}_{\mathbf{Y}}$ in the expression. To gain some insight into this, suppose, for example, $\sqrt{n_i} \xi_i$ is from prior distribution $N_p(0, \psi \mathbf{I}_p)$, $i = 1, \dots, k$. Then $\psi^{-1} \mathbf{C}^T \mathbf{\Theta}^T \mathbf{\Theta} \mathbf{C}$ is distributed as $\text{Wishart}_{k-1}(p, \mathbf{I}_{k-1})$, the $k - 1$ dimensional Wishart distribution with degree of freedom p and parameter \mathbf{I}_{k-1} . On the other hand, $\mathbf{C}^T \mathbf{\Theta}^T (\mathbf{I}_p - \mathbf{P}_{\mathbf{Y}}) \mathbf{\Theta} \mathbf{C}$ is distributed as $\text{Wishart}_{k-1}(p - n + k, \mathbf{I}_{k-1})$. In this case, we have

$$\frac{\mathbf{C}^T \mathbf{\Theta}^T (\mathbf{I}_p - \mathbf{P}_{\mathbf{Y}}) \mathbf{\Theta} \mathbf{C}}{\mathbf{C}^T \mathbf{\Theta}^T \mathbf{\Theta} \mathbf{C}} \xrightarrow{P} 1.$$

Thus, when

$$\liminf_{n \rightarrow \infty} \frac{\text{tr}(\mathbf{\Sigma}^2)}{\text{tr}(\mathbf{\Lambda}_2^2)} \in (1, +\infty],$$

LFD test tends to be more powerful than Chen and Qin (2010)'s test.

3. Numerical study

In this section, we compare the numerical performance of LFD test with some existing tests, including the MANOVA tests of Schott (2007) and Cai and Xia (2014) and the two sample tests of Srivastava (2007), Chen and Qin (2010), Cai et al. (2013) and Feng et al. (2014). Note that the critical values of these existing tests are not valid under spiked covariance model. Hence we use permutation method to determine the critical values during our simulations.

Permutation method is a powerful tool to determine the critical value of a test statistic. The test procedure resulting from permutation method is exact as long as the null distribution of observations are exchangeable (Romano, 1990). The major down-side to permutation method is that it can be computationally intensive. Fortunately, for the test statistic proposed in this paper, the permutation method can be computationally fast. By expression (2.3), a permuted statistic can be written as

$$T(\mathbf{X}\Gamma) = \lambda_{\max}\left(\mathbf{C}^T(\mathbf{J}^T\Gamma^T(\mathbf{X}^T\mathbf{X})^{-1}\Gamma\mathbf{J})^{-1}\mathbf{C}\right), \quad (3.4)$$

where Γ is an $n \times n$ permutation matrix. Note that $(\mathbf{X}^T\mathbf{X})^{-1}$, the most time-consuming component, can be calculated beforehand. The permutation procedure for LFD test statistic can be summarized as:

-
1. Calculate $T(\mathbf{X})$ according to (2.3), hold intermediate result $(\mathbf{X}^T \mathbf{X})^{-1}$.
 2. For a large M , independently generate M random permutation matrix $\Gamma_1, \dots, \Gamma_M$ and calculate $T(\mathbf{X}\Gamma_1), \dots, T(\mathbf{X}\Gamma_M)$ according to (3.4).
 3. Calculate the p -value by $\tilde{p} = (M + 1)^{-1} [1 + \sum_{i=1}^M I\{T(\mathbf{X}\Gamma_i) \geq T(\mathbf{X})\}]$.
Reject the null hypothesis if $\tilde{p} \leq \alpha$.

Here M is the permutation times. It can be seen that step 1 and step 2 cost $O(n^2p + n^3)$ and $O(n^2M)$ operations respectively. In large sample or high dimensional setting, step 2 has a negligible effect on total computational complexity.

Now we evaluate the empirical power performance of LFD test and competing tests. Define signal-to-noise ratio (SNR) as

$$\text{SNR} = \frac{\|\Theta \mathbf{C}\|_F^2}{\sqrt{\Lambda_2^2}}.$$

We use SNR to characterize the signal strength.

In our first simulation study, we take $k = 3$. For comparison, we also carry out simulations for the tests of Schott (2007) and Cai and Xia (2014). We denote these two tests by SC and CX, respectively. We take $r = 2$ and $\Sigma = \text{diag}(1.5p, p, 1, 1, \dots, 1)$. We consider two different structures of alternative hypotheses: the non-sparse alternative and the sparse alternative. In the non-sparse case, we set $\xi_1 = \kappa \mathbf{1}_p$, $\xi_2 = -\kappa \mathbf{1}_p$ and $\xi_3 = \mathbf{0}_p$, where κ

is selected to make the SNR equal to the given value. In the sparse case, we set $\xi_1 = \kappa(1_{p/5}^T, \mathbf{0}_{4p/5}^T)^T$, $\xi_2 = \kappa(\mathbf{0}_{p/5}^T, 1_{p/5}^T, \mathbf{0}_{3p/5}^T)^T$ and $\xi_3 = \mathbf{0}_p$. Again, κ is selected to make the SNR equal to the given value. The empirical power is computed based on 1000 simulations. The simulation results are summarized in Tables 1-4. It can be seen from the results that the proposed test outperforms the other two tests for both non-sparse and sparse alternatives. This verifies our theoretical results that LFD test performs well under spiked covariance.

In our second simulation study, we would like to investigate the effect of correlations between variables. We take $k = 2$ so that we can compare our test with some existing two sample tests. For comparison, we carry out simulations for the test of Srivastava (2007), Chen and Qin (2010), Cai et al. (2013) and Feng et al. (2014). We shall denote these tests by SR, CQ, CLX and FZWZ, respectively. Let the diagonal elements of Σ be 1 and the off-diagonal elements of Σ be ρ with $0 \leq \rho < 1$. The parameter ρ characterizes the correlations between variables. We set $\xi_1 = \kappa(\mathbf{1}_{p/2}^T, -\mathbf{1}_{p/2}^T)^T$ and $\xi_2 = \mathbf{0}_p$, where κ is selected such that SNR equals to 5. Figure 1 plots the empirical power versus ρ , where empirical power is computed based on 1000 simulations. We can see that the empirical power of LFD test holds nearly constant as ρ varies while the empirical powers of Chen and Qin

(2010) and Feng et al. (2014)’s tests decrease as ρ increases. When ρ is small, LFD test has reasonable performance. When ρ is larger than 0.1, LFD test outperforms all other tests.

Table 1: Empirical powers of tests under non-sparse alternative. $\alpha = 0.05$, $k = 3$, $n_1 = n_2 = n_3 = 10$.

SNR	$p = 50$			$p = 75$			$p = 100$		
	CX	SC	LFD	CX	SC	LFD	CX	SC	LFD
0	0.047	0.044	0.052	0.051	0.050	0.051	0.059	0.047	0.047
1	0.074	0.056	0.074	0.089	0.050	0.089	0.062	0.062	0.093
2	0.120	0.045	0.133	0.090	0.040	0.119	0.071	0.049	0.127
3	0.107	0.046	0.197	0.118	0.057	0.242	0.102	0.057	0.220
4	0.160	0.062	0.271	0.131	0.057	0.328	0.146	0.053	0.339
5	0.207	0.064	0.386	0.149	0.052	0.458	0.146	0.067	0.484
6	0.199	0.061	0.485	0.192	0.047	0.583	0.160	0.057	0.588
7	0.234	0.071	0.577	0.221	0.074	0.685	0.185	0.057	0.707
8	0.266	0.072	0.648	0.263	0.078	0.775	0.201	0.062	0.829
9	0.319	0.081	0.718	0.245	0.068	0.838	0.230	0.064	0.896
10	0.304	0.075	0.784	0.297	0.089	0.904	0.288	0.062	0.913

Table 2: Empirical powers of tests under non-sparse alternative. $\alpha = 0.05$,
 $k = 3$, $n_1 = n_2 = n_3 = 25$.

SNR	$p = 100$			$p = 150$			$p = 200$		
	CX	SC	LFD	CX	SC	LFD	CX	SC	LFD
0	0.045	0.041	0.054	0.052	0.046	0.043	0.048	0.043	0.049
1	0.074	0.061	0.099	0.054	0.056	0.082	0.057	0.061	0.107
2	0.092	0.066	0.128	0.086	0.050	0.146	0.079	0.065	0.174
3	0.097	0.070	0.207	0.094	0.058	0.258	0.087	0.053	0.307
4	0.117	0.050	0.249	0.116	0.053	0.375	0.127	0.061	0.412
5	0.147	0.057	0.334	0.139	0.058	0.535	0.122	0.034	0.570
6	0.204	0.057	0.444	0.169	0.070	0.666	0.139	0.055	0.738
7	0.215	0.065	0.523	0.190	0.054	0.774	0.165	0.061	0.847
8	0.247	0.074	0.618	0.200	0.064	0.851	0.181	0.055	0.915
9	0.274	0.073	0.650	0.229	0.059	0.915	0.212	0.052	0.943
10	0.291	0.069	0.729	0.245	0.064	0.930	0.225	0.051	0.977

Table 3: Empirical powers of tests under sparse alternative. $\alpha = 0.05$,
 $k = 3$, $n_1 = n_2 = n_3 = 10$.

SNR	$p = 50$			$p = 75$			$p = 100$		
	CX	SC	LFD	CX	SC	LFD	CX	SC	LFD
0	0.038	0.043	0.037	0.046	0.058	0.059	0.049	0.044	0.047
1	0.064	0.054	0.076	0.067	0.061	0.088	0.066	0.053	0.084
2	0.101	0.052	0.097	0.085	0.048	0.114	0.111	0.058	0.114
3	0.144	0.060	0.169	0.132	0.050	0.188	0.112	0.049	0.166
4	0.181	0.060	0.220	0.161	0.052	0.239	0.157	0.063	0.249
5	0.236	0.063	0.295	0.194	0.061	0.313	0.216	0.057	0.311
6	0.285	0.070	0.333	0.253	0.065	0.419	0.243	0.060	0.398
7	0.344	0.081	0.425	0.299	0.061	0.506	0.291	0.066	0.543
8	0.401	0.082	0.513	0.363	0.077	0.620	0.299	0.065	0.611
9	0.455	0.079	0.600	0.407	0.067	0.667	0.392	0.060	0.709
10	0.522	0.076	0.641	0.467	0.086	0.784	0.417	0.071	0.766

Table 4: Empirical powers of tests under sparse alternative. $\alpha = 0.05$,
 $k = 3$, $n_1 = n_2 = n_3 = 25$.

SNR	$p = 100$			$p = 150$			$p = 200$		
	CX	SC	LFD	CX	SC	LFD	CX	SC	LFD
0	0.068	0.051	0.051	0.046	0.053	0.043	0.065	0.049	0.052
1	0.074	0.049	0.062	0.062	0.046	0.109	0.084	0.048	0.103
2	0.100	0.060	0.123	0.064	0.055	0.149	0.093	0.055	0.155
3	0.105	0.048	0.157	0.104	0.054	0.228	0.114	0.065	0.270
4	0.152	0.064	0.246	0.133	0.056	0.320	0.129	0.054	0.303
5	0.194	0.054	0.280	0.190	0.036	0.419	0.151	0.048	0.434
6	0.232	0.059	0.311	0.210	0.057	0.500	0.203	0.051	0.553
7	0.298	0.061	0.405	0.246	0.054	0.586	0.220	0.057	0.661
8	0.367	0.061	0.477	0.314	0.051	0.707	0.261	0.077	0.765
9	0.405	0.064	0.499	0.351	0.057	0.783	0.275	0.064	0.823
10	0.455	0.067	0.587	0.405	0.061	0.828	0.367	0.059	0.900

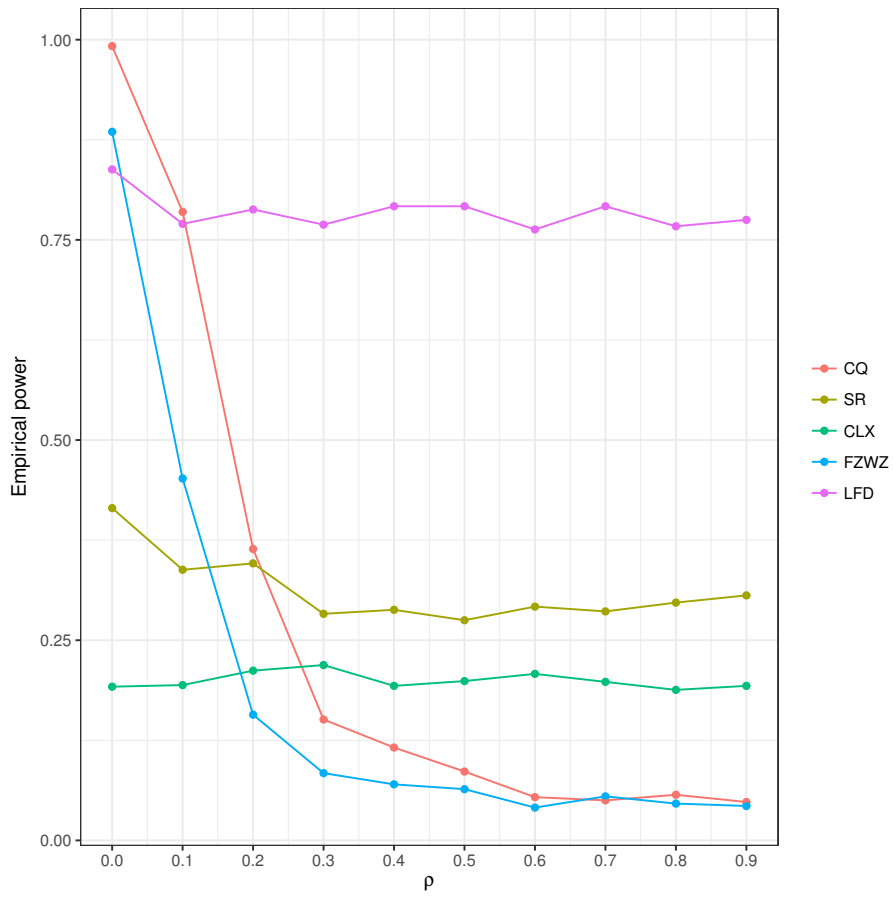


Figure 1: The empirical powers of tests. $\alpha = 0.05$, $k = 2$, $n_1 = n_2 = 20$, $p = 150$.

4. Concluding remarks

In this paper, using the idea of least favorable direction, we proposed LFD test for MANOVA in high dimensional setting. We derived the asymptotic distribution of LFD test statistic. We also gave the asymptotic local power function. Our theoretic work and simulation studies show that when the covariance matrix is spiked, LFD test tends to be more powerful than existing tests.

Our proof relies on the normality of the observations. It is interesting to investigate whether the theorems are still valid without normal assumption. Moreover, we assumed that p doesn't grow too fast. Without prior knowledge of Σ , this condition is unavoidable since when p is large, it's impossible to consistently estimate the principal space. See, for example, Cai et al. (2013). On the other hand, if we know some prior knowledge of Σ , for example, Σ is sparse, it's possible to construct a better test. We leave it for future research.

Appendix A Technical details

Proposition 4. *Suppose \mathbf{A} is a $p \times r$ matrix with rank r and \mathbf{B} is a $p \times p$ non-zero semi-definite matrix. Denote by $\mathbf{A} = \mathbf{U}_\mathbf{A} \mathbf{D}_\mathbf{A} \mathbf{V}_\mathbf{A}^T$ the singular value decomposition of \mathbf{A} , where $\mathbf{U}_\mathbf{A}$ and $\mathbf{V}_\mathbf{A}$ are $p \times r$ and $r \times r$ column*

orthogonal matrix, $\mathbf{D}_{\mathbf{A}}$ is a $r \times r$ diagonal matrix. Let $\mathbf{P}_{\mathbf{A}} = \mathbf{U}_{\mathbf{A}}\mathbf{U}_{\mathbf{A}}^T$ be the projection on the column space of \mathbf{A} . Then

$$\max_{a^T \mathbf{A} = 1, a^T \mathbf{A} \mathbf{A}^T a = 0} a^T \mathbf{B} a = \lambda_{\max}(\mathbf{B}(\mathbf{I}_p - \mathbf{P}_{\mathbf{A}})). \quad (\text{A.5})$$

Proof. Note that $a^T \mathbf{A} \mathbf{A}^T a = 0$ is equivalent to $\mathbf{P}_{\mathbf{A}} a = 0$ which in turn is equivalent to $a = (\mathbf{I}_p - \mathbf{P}_{\mathbf{A}})a$. Then

$$\max_{a^T \mathbf{A} = 1, a^T \mathbf{A} \mathbf{A}^T a = 0} a^T \mathbf{B} a = \max_{a^T \mathbf{A} = 1, \mathbf{P}_{\mathbf{A}} a = 0} a^T (\mathbf{I}_p - \mathbf{P}_{\mathbf{A}}) \mathbf{B} (\mathbf{I}_p - \mathbf{P}_{\mathbf{A}}) a, \quad (\text{A.6})$$

which is obviously no greater than $\lambda_{\max}((\mathbf{I} - \mathbf{P}_{\mathbf{A}}) \mathbf{B} (\mathbf{I} - \mathbf{P}_{\mathbf{A}}))$. To prove that they are equal, without loss of generality, we can assume $\lambda_{\max}((\mathbf{I} - \mathbf{P}_{\mathbf{A}}) \mathbf{B} (\mathbf{I} - \mathbf{P}_{\mathbf{A}})) > 0$. Let α_1 be one eigenvector corresponding to the largest eigenvalue of $(\mathbf{I} - \mathbf{P}_{\mathbf{A}}) \mathbf{B} (\mathbf{I} - \mathbf{P}_{\mathbf{A}})$. Since $(\mathbf{I} - \mathbf{P}_{\mathbf{A}}) \mathbf{B} (\mathbf{I} - \mathbf{P}_{\mathbf{A}}) \mathbf{P}_{\mathbf{A}} = (\mathbf{I} - \mathbf{P}_{\mathbf{A}}) \mathbf{B} (\mathbf{P}_{\mathbf{A}} - \mathbf{P}_{\mathbf{A}}) = \mathbf{O}_{p \times p}$ and $\mathbf{P}_{\mathbf{A}}$ is symmetric, the rows of $\mathbf{P}_{\mathbf{A}}$ are eigenvectors of $(\mathbf{I} - \mathbf{P}_{\mathbf{A}}) \mathbf{B} (\mathbf{I} - \mathbf{P}_{\mathbf{A}})$ corresponding to eigenvalue 0. It follows that $\mathbf{P}_{\mathbf{A}} \alpha_1 = 0$. Therefore, α_1 satisfies the constraint of (A.6) and (A.6) is no less than $\lambda_{\max}((\mathbf{I} - \mathbf{P}_{\mathbf{A}}) \mathbf{B} (\mathbf{I} - \mathbf{P}_{\mathbf{A}}))$. The conclusion now follows by noting that $\lambda_{\max}((\mathbf{I} - \mathbf{P}_{\mathbf{A}}) \mathbf{B} (\mathbf{I} - \mathbf{P}_{\mathbf{A}})) = \lambda_{\max}(\mathbf{B}(\mathbf{I} - \mathbf{P}_{\mathbf{A}}))$.

□

Lemma 1 (Davidson and Szarek (2001) Theorem II.7). *Let \mathbf{A} be $m \times n$*

with iid $N(0, 1)$ entries. If $m > n$, then for any $t > 0$,

$$\Pr(\sqrt{\lambda_1(\mathbf{A}\mathbf{A}^T)} > \sqrt{m} + \sqrt{n} + t) \leq \exp(-t^2/2),$$

$$\Pr(\sqrt{\lambda_n(\mathbf{A}\mathbf{A}^T)} < \sqrt{m} - \sqrt{n} - t) \leq \exp(-t^2/2).$$

Proves of the main results It can be seen that \mathbf{XJC} is independent of \mathbf{Y} . Since $\mathbf{E}\mathbf{Y} = \mathbf{O}_{p \times (n-k)}$, we can write $\mathbf{Y} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{G}_1$, where \mathbf{G}_1 is a $p \times (n-k)$ matrix with i.i.d. $N(0, 1)$ entries. We write $\mathbf{XJC} = \mathbf{\Theta}\mathbf{C} + \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{G}_2$, where \mathbf{G}_2 is a $p \times (k-1)$ matrix with i.i.d. $N(0, 1)$ entries.

Then

$$\begin{aligned} \mathbf{C}^T \mathbf{J}^T \mathbf{X}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{XJC} &= \mathbf{G}_2^T \mathbf{\Lambda}^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \mathbf{\Lambda}^{1/2} \mathbf{G}_2 + \mathbf{\Theta} \mathbf{C}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{\Theta} \mathbf{C} + \\ &\quad \mathbf{\Theta} \mathbf{C}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \mathbf{\Lambda}^{1/2} \mathbf{G}_2 + \mathbf{G}_2^T \mathbf{\Lambda}^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{\Theta} \mathbf{C}. \end{aligned} \quad (\text{A.7})$$

The first term of (A.7) can be represented as

$$\mathbf{G}_2^T \mathbf{\Lambda}^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \mathbf{\Lambda}^{1/2} \mathbf{G}_2 = \sum_{i=1}^p \lambda_i (\mathbf{\Lambda}^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \mathbf{\Lambda}^{1/2}) \eta_i \eta_i^T, \quad (\text{A.8})$$

where $\eta_i \stackrel{i.i.d.}{\sim} N(0, \mathbf{I}_{k-1})$.

Let $\mathbf{G}_1 = (\mathbf{G}_{1A}^T, \mathbf{G}_{1B}^T)^T$, where \mathbf{G}_{1A} is the first r rows of \mathbf{G}_1 and \mathbf{G}_{1B} is the last $p-r$ rows of \mathbf{G}_1 . The following lemma gives the asymptotic property of $\lambda_i(\mathbf{Y}^T \mathbf{Y})$, $i = 1, \dots, r$.

Lemma 2. Suppose $p/n \rightarrow \infty$, $r = o(n)$, $\boldsymbol{\lambda}_r n/p \rightarrow \infty$ and $c_1 \geq \boldsymbol{\lambda}_{r+1} \geq \dots \geq \boldsymbol{\lambda}_p \geq c_2$. Then

$$\sup_{1 \leq i \leq r} \left| \frac{\lambda_i(\mathbf{Y}^T \mathbf{Y})}{n \boldsymbol{\lambda}_i} - 1 \right| \rightarrow 0, \quad (\text{A.9})$$

$$\limsup_{n \rightarrow +\infty} \frac{\lambda_{r+1}(\mathbf{Y}^T \mathbf{Y})}{p} \leq c_1, \quad (\text{A.10})$$

$$\liminf_{n \rightarrow +\infty} \frac{\lambda_{n-k}(\mathbf{Y}^T \mathbf{Y})}{p} \geq c_2, \quad (\text{A.11})$$

almost surely.

Proof. Note that

$$\mathbf{Y}^T \mathbf{Y} = \mathbf{G}_1^T \boldsymbol{\Lambda} \mathbf{G}_1 = \mathbf{G}_{1A}^T \boldsymbol{\Lambda}_1 \mathbf{G}_{1A} + \mathbf{G}_{1B}^T \boldsymbol{\Lambda}_2 \mathbf{G}_{1B}.$$

For $1 \leq i \leq r$, we have

$$\lambda_i(\mathbf{G}_{1A}^T \boldsymbol{\Lambda}_1 \mathbf{G}_{1A}) \leq \lambda_i(\mathbf{Y}^T \mathbf{Y}) \leq \lambda_i(\mathbf{G}_{1A}^T \boldsymbol{\Lambda}_1 \mathbf{G}_{1A}) + c_1 \lambda_1(\mathbf{G}_{1B}^T \mathbf{G}_{1B}). \quad (\text{A.12})$$

Using Weyl's inequality, we can derive a lower bound for $\lambda_i(\mathbf{G}_{1A}^T \boldsymbol{\Lambda}_1 \mathbf{G}_{1A})$,

$i = 1, \dots, r$.

$$\begin{aligned} \lambda_i(\mathbf{G}_{1A}^T \boldsymbol{\Lambda}_1 \mathbf{G}_{1A}) &\geq \lambda_i(\mathbf{G}_{1A}^T \text{diag}(\boldsymbol{\lambda}_i \mathbf{I}_i, \mathbf{O}_{(r-i) \times (r-i)}) \mathbf{G}_{1A}) \\ &= \lambda_i \left(\boldsymbol{\lambda}_i \mathbf{G}_{1A}^T \mathbf{G}_{1A} - \boldsymbol{\lambda}_i \mathbf{G}_{1A}^T \text{diag}(\mathbf{O}_{i \times i}, \mathbf{I}_{r-i}) \mathbf{G}_{1A} \right) \\ &\geq \lambda_r \left(\boldsymbol{\lambda}_i \mathbf{G}_{1A}^T \mathbf{G}_{1A} \right) + \lambda_{p+i-r} \left(- \boldsymbol{\lambda}_i \mathbf{G}_{1A}^T \text{diag}(\mathbf{O}_{i \times i}, \mathbf{I}_{r-i}) \mathbf{G}_{1A} \right) \\ &= \boldsymbol{\lambda}_i \lambda_r (\mathbf{G}_{1A} \mathbf{G}_{1A}^T). \end{aligned} \quad (\text{A.13})$$

Similarly, we can obtain the upper bound.

$$\begin{aligned}
& \lambda_i(\mathbf{G}_{1A}^T \mathbf{\Lambda}_1 \mathbf{G}_{1A}) \\
&= \lambda_i \left(\mathbf{G}_{1A}^T \left(\text{diag}(\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_{i-1}, \mathbf{O}_{(r-i+1) \times (r-i+1)}) + \text{diag}(\mathbf{O}_{(i-1) \times (i-1)}, \boldsymbol{\lambda}_i, \dots, \boldsymbol{\lambda}_r) \right) \mathbf{G}_{1A} \right) \\
&\leq \lambda_1(\mathbf{G}_{1A}^T \text{diag}(\mathbf{O}_{(i-1) \times (i-1)}, \boldsymbol{\lambda}_i \mathbf{I}_{r-i+1}) \mathbf{G}_{1A}) \leq \boldsymbol{\lambda}_i \lambda_1(\mathbf{G}_{1A} \mathbf{G}_{1A}^T).
\end{aligned} \tag{A.14}$$

The inequality (A.12), (A.13) and (A.14) implies that

$$\sup_{1 \leq i \leq r} \left| \frac{\lambda_i(\mathbf{Y}^T \mathbf{Y})}{n \boldsymbol{\lambda}_i} - 1 \right| \leq \max \left(\left| \frac{\lambda_1(\mathbf{G}_{1A} \mathbf{G}_{1A}^T)}{n} - 1 \right|, \left| \frac{\lambda_r(\mathbf{G}_{1A} \mathbf{G}_{1A}^T)}{n} - 1 \right| \right) + \frac{c_1}{n \boldsymbol{\lambda}_r} \lambda_1(\mathbf{G}_{1B}^T \mathbf{G}_{1B}).$$

We only need to prove the right hand side converges to 0 almost surely.

By Lemma 1, for every $t > 0$, we have

$$\begin{aligned}
& \Pr \left(\sqrt{1 - \frac{k}{n}} - \sqrt{\frac{r}{n}} - \frac{t}{\sqrt{n}} \leq \sqrt{\frac{\lambda_r(\mathbf{G}_{1A} \mathbf{G}_{1A}^T)}{n}} \leq \sqrt{\frac{\lambda_1(\mathbf{G}_{1A} \mathbf{G}_{1A}^T)}{n}} \leq \sqrt{1 - \frac{k}{n}} + \sqrt{\frac{r}{n}} + \frac{t}{\sqrt{n}} \right) \\
& \geq 1 - 2 \exp\left(-\frac{t^2}{2}\right).
\end{aligned} \tag{A.15}$$

Let $t = n^{1/4}$. Since $r = o(n)$, we have

$$\sqrt{1 - \frac{k}{n}} - \sqrt{\frac{r}{n}} - \frac{t}{\sqrt{n}} \rightarrow 1 \quad \text{and} \quad \sqrt{1 - \frac{k}{n}} + \sqrt{\frac{r}{n}} + \frac{t}{\sqrt{n}} \rightarrow 1.$$

This, together with Borel-Cantelli lemma, yields

$$\frac{\lambda_r(\mathbf{G}_{1A} \mathbf{G}_{1A}^T)}{n} \rightarrow 1 \quad \frac{\lambda_1(\mathbf{G}_{1A} \mathbf{G}_{1A}^T)}{n} \rightarrow 1,$$

almost surely. As for $\lambda_1(\mathbf{G}_{1B}^T \mathbf{G}_{1B})$, by Lemma 1, we have

$$\Pr \left(\frac{c_1}{n \boldsymbol{\lambda}_r} \lambda_1(\mathbf{G}_{1B}^T \mathbf{G}_{1B}) \leq \frac{c_1}{n \boldsymbol{\lambda}_r} (\sqrt{n-k} + \sqrt{p-r} + t)^2 \right) \geq 1 - \exp\left(-\frac{t^2}{2}\right). \tag{A.16}$$

Let $t = n^{1/2}$, since we have assumed $\boldsymbol{\lambda}_r n/p \rightarrow \infty$, we have

$$\frac{c_1}{n\boldsymbol{\lambda}_r} \lambda_1(\mathbf{G}_{1B} \mathbf{G}_{1B}^T) \rightarrow 0$$

almost surely. Then (A.9) follows.

Inequality (A.10) and (A.11) follows from the fact

$$\lambda_{r+1}(\mathbf{Y}^T \mathbf{Y}) \leq \lambda_1(\mathbf{G}_{1B}^T \boldsymbol{\Lambda}_2 \mathbf{G}_{1B}) \leq c_1 \lambda_1(\mathbf{G}_{1B}^T \mathbf{G}_{1B}),$$

$$\lambda_{n-k}(\mathbf{Y}^T \mathbf{Y}) \geq \lambda_{n-k}(\mathbf{G}_{1B}^T \boldsymbol{\Lambda}_2 \mathbf{G}_{1B}) \geq c_2 \lambda_{n-k}(\mathbf{G}_{1B}^T \mathbf{G}_{1B}),$$

and Lemma 1. □

Let $\mathbf{U}_{\mathbf{Y}} = (\mathbf{U}_{\mathbf{Y},1}, \mathbf{U}_{\mathbf{Y},2})$, where $\mathbf{U}_{\mathbf{Y},1}$ and $\mathbf{U}_{\mathbf{Y},2}$ are the first r and last $p - r$ columns of $\mathbf{U}_{\mathbf{Y}}$ respectively.

Lemma 3. *Under the assumptions of Lemma 2, we have*

$$\lambda_{\max}(\mathbf{I}_r - \mathbf{U}_1^T \mathbf{U}_{\mathbf{Y},1} \mathbf{U}_{\mathbf{Y},1}^T \mathbf{U}_1) = O_P\left(\frac{p}{\boldsymbol{\lambda}_r n}\right).$$

If in addition, we assume

$$\frac{\log \boldsymbol{\lambda}_r}{n} \rightarrow 0, \tag{A.17}$$

then

$$\mathbb{E} \lambda_{\max}(\mathbf{I}_r - \mathbf{U}_1^T \mathbf{U}_{\mathbf{Y},1} \mathbf{U}_{\mathbf{Y},1}^T \mathbf{U}_1) = O\left(\frac{p}{\boldsymbol{\lambda}_r n}\right).$$

Proof. From $\mathbf{U} \boldsymbol{\Lambda}^{1/2} \mathbf{G}_1 \mathbf{G}_1^T \boldsymbol{\Lambda}^{1/2} \mathbf{U}^T = \mathbf{U}_{\mathbf{Y}} \mathbf{D}_{\mathbf{Y}}^2 \mathbf{U}_{\mathbf{Y}}^T$, we have

$$\begin{pmatrix} \boldsymbol{\Lambda}_1^{\frac{1}{2}} \mathbf{G}_{1A} \mathbf{G}_{1A}^T \boldsymbol{\Lambda}_1^{\frac{1}{2}} & \boldsymbol{\Lambda}_1^{\frac{1}{2}} \mathbf{G}_{1A} \mathbf{G}_{1B}^T \boldsymbol{\Lambda}_2^{\frac{1}{2}} \\ \boldsymbol{\Lambda}_2^{\frac{1}{2}} \mathbf{G}_{1B} \mathbf{G}_{1A}^T \boldsymbol{\Lambda}_1^{\frac{1}{2}} & \boldsymbol{\Lambda}_2^{\frac{1}{2}} \mathbf{G}_{1B} \mathbf{G}_{1B}^T \boldsymbol{\Lambda}_2^{\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} \mathbf{U}_1^T \mathbf{U}_{\mathbf{Y}} \mathbf{D}_{\mathbf{Y}}^2 \mathbf{U}_{\mathbf{Y}}^T \mathbf{U}_1 & \mathbf{U}_1^T \mathbf{U}_{\mathbf{Y}} \mathbf{D}_{\mathbf{Y}}^2 \mathbf{U}_{\mathbf{Y}}^T \mathbf{U}_2 \\ \mathbf{U}_2^T \mathbf{U}_{\mathbf{Y}} \mathbf{D}_{\mathbf{Y}}^2 \mathbf{U}_{\mathbf{Y}}^T \mathbf{U}_1 & \mathbf{U}_2^T \mathbf{U}_{\mathbf{Y}} \mathbf{D}_{\mathbf{Y}}^2 \mathbf{U}_{\mathbf{Y}}^T \mathbf{U}_2 \end{pmatrix}$$

It follows that

$$\mathbf{\Lambda}_2^{\frac{1}{2}} \mathbf{G}_{1B} \mathbf{G}_{1B}^T \mathbf{\Lambda}_2^{\frac{1}{2}} \geq \lambda_r(\mathbf{Y}^T \mathbf{Y}) \mathbf{U}_2^T \mathbf{U}_{\mathbf{Y},1} \mathbf{U}_{\mathbf{Y},1}^T \mathbf{U}_2.$$

Hence

$$\lambda_1(\mathbf{U}_2^T \mathbf{U}_{\mathbf{Y},1} \mathbf{U}_{\mathbf{Y},1}^T \mathbf{U}_2) \leq \frac{c_1}{\lambda_r(\mathbf{Y}^T \mathbf{Y})} \lambda_1(\mathbf{G}_{1B} \mathbf{G}_{1B}^T). \quad (\text{A.18})$$

By Lemma 1, for every $t > 0$, we have

$$\Pr \left(\frac{1}{p} (\sqrt{p-r} - \sqrt{n-k} - t)^2 \leq \frac{1}{p} \lambda_1(\mathbf{G}_{1B} \mathbf{G}_{1B}^T) \leq \frac{1}{p} (\sqrt{p-r} + \sqrt{n-k} + t)^2 \right) \geq 1 - 2 \exp\left(-\frac{t^2}{2}\right). \quad (\text{A.19})$$

Let $t = n^{1/2}$, then Borel-Cantelli lemma implies that

$$\frac{1}{p} \lambda_1(\mathbf{G}_{1B} \mathbf{G}_{1B}^T) \rightarrow 1 \quad (\text{A.20})$$

almost surely. Then (A.20), (A.18) and Lemma 2 implies that

$$\lambda_1(\mathbf{U}_2^T \mathbf{U}_{\mathbf{Y},1} \mathbf{U}_{\mathbf{Y},1}^T \mathbf{U}_2) = O_P\left(\frac{p}{\lambda_r n}\right).$$

The first conclusion then follows by the following simple relationship

$$\begin{aligned} \lambda_{\max}(\mathbf{U}_2^T \mathbf{U}_{\mathbf{Y},1} \mathbf{U}_{\mathbf{Y},1}^T \mathbf{U}_2) &= \lambda_{\max}(\mathbf{U}_{\mathbf{Y},1}^T \mathbf{U}_2 \mathbf{U}_2^T \mathbf{U}_{\mathbf{Y},1}) \\ &= \lambda_{\max}(\mathbf{U}_{\mathbf{Y},1}^T (\mathbf{I}_p - \mathbf{U}_1 \mathbf{U}_1^T) \mathbf{U}_{\mathbf{Y},1}) = \lambda_{\max}(\mathbf{I}_r - \mathbf{U}_{\mathbf{Y},1}^T \mathbf{U}_1 \mathbf{U}_1^T \mathbf{U}_{\mathbf{Y},1}) \\ &= 1 - \lambda_{\min}(\mathbf{U}_{\mathbf{Y},1}^T \mathbf{U}_1 \mathbf{U}_1^T \mathbf{U}_{\mathbf{Y},1}) = 1 - \lambda_{\min}(\mathbf{U}_1^T \mathbf{U}_{\mathbf{Y},1} \mathbf{U}_{\mathbf{Y},1}^T \mathbf{U}_1) \\ &= \lambda_{\max}(\mathbf{I}_r - \mathbf{U}_1^T \mathbf{U}_{\mathbf{Y},1} \mathbf{U}_{\mathbf{Y},1}^T \mathbf{U}_1). \end{aligned}$$

Next we prove the second conclusion of the lemma. In (A.15), (A.16) and (A.19), we take $t = \sqrt{2\log(\boldsymbol{\lambda}_r n/p)}$. Then these inequalities, (A.18) and condition (A.17) implies that

$$\lambda_1(\mathbf{U}_2^T \mathbf{U}_{\mathbf{Y},1} \mathbf{U}_{\mathbf{Y},1}^T \mathbf{U}_2) \leq \frac{c^* p}{\boldsymbol{\lambda}_r n}$$

with probability at least $1 - 3p/\boldsymbol{\lambda}_r n$ for some constant c^* . Since $\lambda_1(\mathbf{U}_2^T \mathbf{U}_{\mathbf{Y},1} \mathbf{U}_{\mathbf{Y},1}^T \mathbf{U}_2) \leq 1$, we have

$$\mathbb{E} \lambda_1(\mathbf{U}_2^T \mathbf{U}_{\mathbf{Y},1} \mathbf{U}_{\mathbf{Y},1}^T \mathbf{U}_2) \leq \frac{c^* p}{\boldsymbol{\lambda}_r n} (1 - \frac{3p}{\boldsymbol{\lambda}_r n}) + \frac{3p}{\boldsymbol{\lambda}_r n} = O(\frac{p}{\boldsymbol{\lambda}_r n}).$$

This completes the proof. \square

Lemma 4. *Under the assumptions of Lemma 2, we have the following upper and lower bound for $\boldsymbol{\Lambda}^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_{\mathbf{Y}}) \mathbf{U} \boldsymbol{\Lambda}^{1/2}$.*

$$\lambda_i(\boldsymbol{\Lambda}^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_{\mathbf{Y}}) \mathbf{U} \boldsymbol{\Lambda}^{1/2}) \geq \boldsymbol{\lambda}_{i+n-k}, \quad i = 1, \dots, p - n + k, \quad (\text{A.21})$$

$$\lambda_i(\boldsymbol{\Lambda}^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_{\mathbf{Y}}) \mathbf{U} \boldsymbol{\Lambda}^{1/2}) = O_P(\frac{\boldsymbol{\lambda}_1 p}{\boldsymbol{\lambda}_r n}) + c_1, \quad i = 1, \dots, r, \quad (\text{A.22})$$

$$\lambda_i(\boldsymbol{\Lambda}^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_{\mathbf{Y}}) \mathbf{U} \boldsymbol{\Lambda}^{1/2}) \leq \boldsymbol{\lambda}_i, \quad i = r + 1, \dots, p. \quad (\text{A.23})$$

Proof. The inequality (A.23) follows from the fact $\mathbf{I}_p - \mathbf{P}_{\mathbf{Y}} \leq \mathbf{I}_p$. The inequality (A.21) follows from the fact that $\text{Rank}(\mathbf{P}_{\mathbf{Y}}) \leq n - k$ and Weyl's inequality. As for inequality (A.22), note that the positive eigenvalues of $\boldsymbol{\Lambda}^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_{\mathbf{Y}}) \mathbf{U} \boldsymbol{\Lambda}^{1/2}$ equal to the positive eigenvalues of $(\mathbf{I}_p - \mathbf{P}_{\mathbf{Y}}) \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^T (\mathbf{I}_p -$

\mathbf{P}_Y). We write $(\mathbf{I}_p - \mathbf{P}_Y)\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T(\mathbf{I}_p - \mathbf{P}_Y)$ as the sum of two terms

$$\begin{aligned} & (\mathbf{I}_p - \mathbf{P}_Y)\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T(\mathbf{I}_p - \mathbf{P}_Y) \\ &= (\mathbf{I}_p - \mathbf{P}_Y)\mathbf{U}_1\mathbf{\Lambda}_1\mathbf{U}_1^T(\mathbf{I}_p - \mathbf{P}_Y) + (\mathbf{I}_p - \mathbf{P}_Y)\mathbf{U}_2\mathbf{\Lambda}_2\mathbf{U}_2^T(\mathbf{I}_p - \mathbf{P}_Y) \stackrel{def}{=} \mathbf{R}_1 + \mathbf{R}_2. \end{aligned}$$

Lemma 3 can be applied to control the largest eigenvalue of \mathbf{R}_1 :

$$\begin{aligned} \lambda_1(\mathbf{R}_1) &= \lambda_1(\mathbf{\Lambda}_1^{1/2}\mathbf{U}_1^T(\mathbf{I}_p - \mathbf{P}_Y)\mathbf{U}_1\mathbf{\Lambda}_1^{1/2}) \leq \lambda_1(\mathbf{\Lambda}_1^{1/2}\mathbf{U}_1^T(\mathbf{I}_p - \mathbf{U}_{Y,1}\mathbf{U}_{Y,1}^T)\mathbf{U}_1\mathbf{\Lambda}_1^{1/2}) \\ &\leq \lambda_1\lambda_1(\mathbf{U}_1^T(\mathbf{I}_p - \mathbf{U}_{Y,1}\mathbf{U}_{Y,1}^T)\mathbf{U}_1) = \lambda_1\lambda_1(\mathbf{I}_r - \mathbf{U}_1^T\mathbf{U}_{Y,1}\mathbf{U}_{Y,1}^T\mathbf{U}_1) = O_P\left(\frac{\lambda_1 p}{\lambda_r n}\right). \end{aligned}$$

Thus, for $i = 1, \dots, r$, we have

$$\lambda_i(\mathbf{\Lambda}^{1/2}\mathbf{U}^T(\mathbf{I}_p - \mathbf{P}_Y)\mathbf{U}\mathbf{\Lambda}^{1/2}) \leq \lambda_1(\mathbf{R}_1) + \lambda_1(\mathbf{R}_2) = O_P\left(\frac{\lambda_1 p}{\lambda_r n}\right) + c_1.$$

□

Lemma 5. *Under the assumptions of Theorem 1, we have*

$$\text{tr}(\mathbf{\Lambda}^{1/2}\mathbf{U}^T(\mathbf{I}_p - \mathbf{P}_Y)\mathbf{U}\mathbf{\Lambda}^{1/2}) = \frac{p - r - n + k}{p - r} \text{tr}(\mathbf{\Lambda}_2) + o_P(\sqrt{p}), \quad (\text{A.24})$$

and

$$\text{tr}(\mathbf{\Lambda}^{1/2}\mathbf{U}^T(\mathbf{I}_p - \mathbf{P}_Y)\mathbf{U}\mathbf{\Lambda}^{1/2})^2 = (1 + o_P(1)) \text{tr}(\mathbf{\Lambda}_2^2). \quad (\text{A.25})$$

Proof. By Lemma 4, we have

$$\sum_{i=n-k+1}^p \lambda_i^2 \leq \text{tr}(\mathbf{\Lambda}^{1/2}\mathbf{U}^T(\mathbf{I}_p - \mathbf{P}_Y)\mathbf{U}\mathbf{\Lambda}^{1/2})^2 \leq r(O_P\left(\frac{\lambda_1 p}{\lambda_r n}\right) + c_1)^2 + \sum_{i=r+1}^p \lambda_i^2.$$

Hence

$$\begin{aligned}
& \left| \text{tr} \left(\Lambda^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \Lambda^{1/2} \right)^2 - \sum_{i=r+1}^p \lambda_i^2 \right| \\
& \leq \max \left(\sum_{i=r+1}^{n-k} \lambda_i^2, r \left(O_P \left(\frac{\lambda_1 p}{\lambda_r n} \right) + c_1 \right)^2 \right) \\
& \leq r \left(O_P \left(\frac{\lambda_1 p}{\lambda_r n} \right) + c_1 \right)^2 + O(n) = o_P(p).
\end{aligned}$$

Then (A.25) holds.

Now we prove (A.24). Note that

$$\text{tr} \left(\Lambda^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \Lambda^{1/2} \right) = \text{tr} \left(\Lambda_1^{1/2} \mathbf{U}_1^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U}_1 \Lambda_1^{1/2} \right) + \text{tr} \left(\Lambda_2^{1/2} \mathbf{U}_2^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U}_2 \Lambda_2^{1/2} \right).$$

By Lemma 4, we have

$$\text{tr} \left(\Lambda_1^{1/2} \mathbf{U}_1^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U}_1 \Lambda_1^{1/2} \right) = O_P \left(\frac{\lambda_1 p r}{\lambda_r n} \right) = o_P(\sqrt{p}).$$

The second term can be written as $\text{tr} \left(\Lambda_2^{1/2} \mathbf{U}_2^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U}_2 \Lambda_2^{1/2} \right) = \text{tr}(\Lambda_2) -$

$\text{tr}(\mathbf{P}_Y \mathbf{U}_2 \Lambda_2 \mathbf{U}_2^T)$. For $\text{tr}(\mathbf{P}_Y \mathbf{U}_2 \Lambda_2 \mathbf{U}_2^T)$, we have

$$\begin{aligned}
& \left| \text{tr}(\mathbf{P}_Y \mathbf{U}_2 \Lambda_2 \mathbf{U}_2^T) - \frac{n-k}{p-r} \text{tr}(\Lambda_2) \right| = \left| \text{tr} \left(\mathbf{P}_Y \mathbf{U} \left(\Lambda_2 - \frac{1}{p-r} (\text{tr} \Lambda_2) \mathbf{I}_{p-r} \right) \mathbf{U}^T \right) \right| \\
& \leq \sqrt{\text{tr}(\mathbf{P}_Y^2)} \sqrt{\text{tr} \left(\Lambda_2 - \frac{1}{p-r} (\text{tr} \Lambda_2) \mathbf{I}_{p-r} \right)^2} = \sqrt{(n-k) \text{tr} \left(\Lambda_2 - \frac{1}{p-r} (\text{tr} \Lambda_2) \mathbf{I}_{p-r} \right)^2} = o(\sqrt{p}).
\end{aligned}$$

Hence

$$\text{tr} \left(\Lambda_2^{1/2} \mathbf{U}_2^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U}_2 \Lambda_2^{1/2} \right) = \frac{p-r-n+k}{p-r} \text{tr}(\Lambda_2) + o(\sqrt{p}).$$

Then (A.24) holds. \square

Proof of Theorem 1. Lemma (4) implies that the first term of (A.7) satisfies

the Lyapunov condition

$$\frac{\lambda_1 \left((\Lambda^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \Lambda^{1/2})^2 \right)}{\text{tr} \left((\Lambda^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \Lambda^{1/2})^2 \right)} = \frac{(O_P(\frac{\lambda_1 p}{\lambda_r n}) + c_1)^2}{(1 + o_P(1)) \text{tr}(\Lambda_2)} \xrightarrow{P} 0.$$

Apply Lyapunov central limit theorem conditioning on \mathbf{P}_Y , we have

$$\left(\text{tr} \left((\Lambda^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \Lambda^{1/2})^2 \right) \right)^{-1/2} (\mathbf{G}_2^T \Lambda^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \Lambda^{1/2} \mathbf{G}_2 - \text{tr}(\Lambda^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \Lambda^{1/2}) \mathbf{I}_{k-1}) \xrightarrow{\mathcal{L}} \mathbf{W}_{k-1}.$$

This, combined with Lemma 5 and Slutsky's theorem, yields

$$\frac{1}{\sqrt{\text{tr}(\Lambda_2^2)}} (\mathbf{G}_2^T \Lambda^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \Lambda^{1/2} \mathbf{G}_2 - \frac{p-r-n+k}{p-r} \text{tr}(\Lambda_2) \mathbf{I}_{k-1}) \xrightarrow{\mathcal{L}} \mathbf{W}_{k-1}.$$

Next we show that the cross term of (A.7) is negligible. Note that

$$\begin{aligned} & \mathbb{E}[\|\mathbf{C}^T \Theta^T (\mathbf{I}_p - \mathbf{P}_{Z\bar{J}}) \mathbf{U} \Lambda^{1/2} \mathbf{G}_2\|_F^2 | \mathbf{Y}] \\ &= (k-1) \text{tr}(\mathbf{C}^T \Theta^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \Lambda \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_Y) \Theta \mathbf{C}) \\ &\leq (k-1) \lambda_1((\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \Lambda \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_Y)) \|\Theta \mathbf{C}\|_F^2 \\ &\leq (k-1) \lambda_1(\Lambda^{1/2} \mathbf{U}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \Lambda^{1/2}) \|\Theta \mathbf{C}\|_F^2 \\ &= (k-1) O_P(\frac{\lambda_1 p}{\lambda_r n} + c_1) \|\Theta \mathbf{C}\|_F^2 \\ &= (k-1) O_P(\frac{\lambda_1 \sqrt{p}}{\lambda_r n} + \frac{c_1}{\sqrt{p}}) \sqrt{p} \|\Theta \mathbf{C}\|_F^2 = o_P(p), \end{aligned}$$

where the last equality holds since we have assumed $\frac{1}{\sqrt{p}} \|\Theta \mathbf{C}\|_F^2 = O(1)$.

Hence $\|\mathbf{C}^T \Theta^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{U} \Lambda^{1/2} \mathbf{G}_2\|_F^2 = o_P(p)$. Now,

$$\frac{1}{\sqrt{\text{tr}(\Lambda_2^2)}} (\mathbf{C}^T \mathbf{Y}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{Y} \mathbf{C} - \frac{p-r-n+k}{p-r} \text{tr}(\Lambda_2) \mathbf{I}_{k-1} - \mathbf{C}^T \Theta^T (\mathbf{I}_p - \mathbf{P}_Y) \Theta \mathbf{C}) \xrightarrow{\mathcal{L}} \mathbf{W}_{k-1}.$$

Equivalently,

$$\begin{aligned} & \frac{1}{\sqrt{\text{tr}(\mathbf{\Lambda}_2^2)}} \left(\mathbf{C}^T \mathbf{Y}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{Y} \mathbf{C} - \frac{p - r - n + k}{p - r} \text{tr}(\mathbf{\Lambda}_2) \mathbf{I}_{k-1} \right) \\ & \sim \frac{1}{\sqrt{\text{tr}(\mathbf{\Lambda}_2^2)}} \mathbf{C}^T \mathbf{\Theta}^T (\mathbf{I}_p - \mathbf{P}_Y) \mathbf{\Theta} \mathbf{C} + \mathbf{W}_{k-1} + o_P(1). \end{aligned}$$

The conclusion follows by taking the maximum eigenvalue. \square

Proof of Proposition 1. First we consider the case of $r > 0$. By the construction of \hat{r} ,

$$\{\hat{r} = r\} \supseteq \left\{ \frac{\lambda_r(\mathbf{Y}^T \mathbf{Y})}{\lambda_{r+1}(\mathbf{Y}^T \mathbf{Y})} \geq \gamma_n \right\} \cap \left\{ \frac{\lambda_{r+1}(\mathbf{Y}^T \mathbf{Y})}{\lambda_{n-k}(\mathbf{Y}^T \mathbf{Y})} \leq \gamma_n \right\}.$$

Suppose $0 < \epsilon < 1$ is a fixed number. By assumption, there exists an n_0^* ,

for $n \geq n_0^*$, $\gamma_n \leq (1 - \epsilon)n\lambda_r/(c_1 p)$ and $\gamma_n \geq (1 + \epsilon)c_1/c_2$. Thus

$$\{\hat{r} = r\} \supseteq \left\{ \frac{\lambda_r(\mathbf{Y}^T \mathbf{Y})}{\lambda_{r+1}(\mathbf{Y}^T \mathbf{Y})} \geq (1 - \epsilon) \frac{n\lambda_r}{c_1 p} \right\} \cap \left\{ \frac{\lambda_{r+1}(\mathbf{Y}^T \mathbf{Y})}{\lambda_{n-k}(\mathbf{Y}^T \mathbf{Y})} \leq (1 + \epsilon) \frac{c_1}{c_2} \right\}.$$

Lemma 2 implies that almost surely, there exists an n_0 , for $n \geq n_0$, we have

$$\frac{\lambda_r(\mathbf{Y}^T \mathbf{Y})}{\lambda_{r+1}(\mathbf{Y}^T \mathbf{Y})} \geq (1 - \epsilon) \frac{n\lambda_r}{c_1 p}, \quad \frac{\lambda_{r+1}(\mathbf{Y}^T \mathbf{Y})}{\lambda_{n-k}(\mathbf{Y}^T \mathbf{Y})} \leq (1 + \epsilon) \frac{c_1}{c_2}.$$

This yields $\Pr(\hat{r} = r) \rightarrow 1$ for $r > 0$. The case of $r = 0$ can be similarly

proved by noting that

$$\{\hat{r} = r\} \supseteq \left\{ \frac{\lambda_1(\mathbf{Y}^T \mathbf{Y})}{\lambda_{n-k}(\mathbf{Y}^T \mathbf{Y})} \leq \gamma_n \right\}.$$

\square

Proof of Proposition 2. Since \hat{r} is a consistent estimator of r , we only need to prove

$$\frac{1}{n-k} \sum_{i=r+1}^{n-k} \lambda_i(\mathbf{Y}^T \mathbf{Y}) = \text{tr}(\mathbf{\Lambda}_2) + O_P(\sqrt{p}).$$

Note that

$$\mathbf{Y}^T \mathbf{Y} = \mathbf{G}_1^T \mathbf{\Lambda} \mathbf{G}_1 = \mathbf{G}_{1A}^T \mathbf{\Lambda}_1 \mathbf{G}_{1A} + \mathbf{G}_{1B}^T \mathbf{\Lambda}_2 \mathbf{G}_{1B}.$$

By Weyl's inequality, for $i = r+1, \dots, n-k$, we have

$$\lambda_i(\mathbf{G}_{1B}^T \mathbf{\Lambda}_2 \mathbf{G}_{1B}) \leq \lambda_i(\mathbf{Y}^T \mathbf{Y}) \leq \lambda_{i-r}(\mathbf{G}_{1B}^T \mathbf{\Lambda}_2 \mathbf{G}_{1B}).$$

It follows that

$$\sum_{i=r+1}^{n-k} \lambda_i(\mathbf{G}_{1B}^T \mathbf{\Lambda}_2 \mathbf{G}_{1B}) \leq \sum_{i=r+1}^{n-k} \lambda_i(\mathbf{Y}^T \mathbf{Y}) \leq \sum_{i=1}^{n-k-r} \lambda_i(\mathbf{G}_{1B}^T \mathbf{\Lambda}_2 \mathbf{G}_{1B}).$$

Hence

$$\left| \sum_{i=r+1}^{n-k} \lambda_i(\mathbf{Y}^T \mathbf{Y}) - \text{tr}(\mathbf{G}_{1B}^T \mathbf{\Lambda}_2 \mathbf{G}_{1B}) \right| \leq r \lambda_1(\mathbf{G}_{1B}^T \mathbf{\Lambda}_2 \mathbf{G}_{1B}) = O_P(rp).$$

But central limit theorem implies that

$$\text{tr}(\mathbf{G}_{1B}^T \mathbf{\Lambda}_2 \mathbf{G}_{1B}) - (n-k) \text{tr}(\mathbf{\Lambda}_2) = O_P(\sqrt{np}).$$

Thus

$$\begin{aligned} \frac{1}{n-k} \sum_{i=r+1}^{n-k} \lambda_i(\mathbf{Y}^T \mathbf{Y}) &= \text{tr}(\mathbf{\Lambda}_2) + \frac{1}{n-k} \left| \sum_{i=r+1}^{n-k} \lambda_i(\mathbf{Y}^T \mathbf{Y}) - (n-k) \text{tr}(\mathbf{\Lambda}_2) \right| \\ &= \text{tr}(\mathbf{\Lambda}_2) + O_P\left(\frac{rp}{n}\right) = \text{tr}(\mathbf{\Lambda}_2) + O_P(\sqrt{p}), \end{aligned}$$

where the last equality follows from Assumption 1. \square

Proof of Proposition 3. Let $\mathbf{U}_{\mathbf{Y},1;(i,j)}$ be the first r columns of $\mathbf{U}_{\mathbf{Y};(i,j)}$. Let $\mathbf{U}_{\mathbf{Y},2;(i,j)}$ be a $p \times (p - r)$ orthogonal matrix satisfying $\mathbf{U}_{\mathbf{Y},2;(i,j)} \mathbf{U}_{\mathbf{Y},2;(i,j)}^T = \mathbf{I}_p - \mathbf{U}_{\mathbf{Y},1;(i,j)} \mathbf{U}_{\mathbf{Y},1;(i,j)}^T$. Then by Lemma 3, we have

$$\lambda_1(\mathbf{U}_1^T \mathbf{U}_{\mathbf{Y},2;(i,j)} \mathbf{U}_{\mathbf{Y},2;(i,j)}^T \mathbf{U}_1) = \lambda_1(\mathbf{I}_r - \mathbf{U}_1^T \mathbf{U}_{\mathbf{Y},1;(i,j)} \mathbf{U}_{\mathbf{Y},1;(i,j)}^T \mathbf{U}_1) = O_P\left(\frac{p}{\lambda_r n}\right).$$

First, we prove that w_{ij}^2 is an approximation of $Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j$. For $1 \leq i < j \leq n - k$, define $\epsilon_{ij} = w_{ij} - Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j$, then we have

$$\begin{aligned} \epsilon_{ij} &= Y_i^T (\tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T - \mathbf{U}_2 \mathbf{U}_2^T) Y_j \\ &= Y_i^T (\mathbf{U}_1 \mathbf{U}_1^T + \mathbf{U}_2 \mathbf{U}_2^T) (\tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T - \mathbf{U}_2 \mathbf{U}_2^T) (\mathbf{U}_1 \mathbf{U}_1^T + \mathbf{U}_2 \mathbf{U}_2^T) Y_j \\ &= Y_i^T \mathbf{U}_2 \mathbf{U}_2^T (\tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T - \mathbf{U}_2 \mathbf{U}_2^T) \mathbf{U}_2 \mathbf{U}_2^T Y_j \\ &\quad + Y_i^T \mathbf{U}_1 \mathbf{U}_1^T \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T \mathbf{U}_2 \mathbf{U}_2^T Y_j \\ &\quad + Y_i^T \mathbf{U}_2 \mathbf{U}_2^T \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T \mathbf{U}_1 \mathbf{U}_1^T Y_j \\ &\quad + Y_i^T \mathbf{U}_1 \mathbf{U}_1^T \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T \mathbf{U}_1 \mathbf{U}_1^T Y_j \\ &= Y_i^T \mathbf{U}_2 \mathbf{U}_2^T (\tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T - \mathbf{U}_{\mathbf{Y},2;(i,j)} \mathbf{U}_{\mathbf{Y},2;(i,j)}^T) \mathbf{U}_2 \mathbf{U}_2^T Y_j \\ &\quad + Y_i^T \mathbf{U}_2 \mathbf{U}_2^T (\mathbf{U}_{\mathbf{Y},2;(i,j)} \mathbf{U}_{\mathbf{Y},2;(i,j)}^T - \mathbf{U}_2 \mathbf{U}_2^T) \mathbf{U}_2 \mathbf{U}_2^T Y_j \\ &\quad + Y_i^T \mathbf{U}_1 \mathbf{U}_1^T \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T \mathbf{U}_2 \mathbf{U}_2^T Y_j \\ &\quad + Y_i^T \mathbf{U}_2 \mathbf{U}_2^T \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T \mathbf{U}_1 \mathbf{U}_1^T Y_j \\ &\quad + Y_i^T \mathbf{U}_1 \mathbf{U}_1^T \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T \mathbf{U}_1 \mathbf{U}_1^T Y_j \\ &\stackrel{def}{=} \epsilon_{ij}^{(1)} + \epsilon_{ij}^{(2)} + \epsilon_{ij}^{(3)} + \epsilon_{ij}^{(4)} + \epsilon_{ij}^{(5)}. \end{aligned}$$

(A.26)

We deal with the five terms separately. First we deal with $\epsilon_{ij}^{(1)}$. It can be seen that Y_i , Y_j and $\tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}$ are mutually independent and $\tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T - \mathbf{U}_{\mathbf{Y},2;(i,j)} \mathbf{U}_{\mathbf{Y},2;(i,j)}^T$ is a projection matrix whose rank is not larger than $n - k - 2 - r$. Then

$$\begin{aligned}
& \mathbb{E}(\epsilon_{ij}^{(1)})^2 \\
&= \mathbb{E} \operatorname{tr}(\Lambda_2^{1/2} \mathbf{U}_2^T (\tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T - \mathbf{U}_{\mathbf{Y},2;(i,j)} \mathbf{U}_{\mathbf{Y},2;(i,j)}^T) \mathbf{U}_2 \Lambda_2^{1/2})^2 \\
&\leq c_1^2 (n - k - 2 - r) = o(p).
\end{aligned}$$

Next we deal with $\epsilon_{ij}^{(2)}$. we have

$$\begin{aligned}
& \mathbb{E}(\epsilon_{ij}^{(2)})^2 \\
&= \mathbb{E} \operatorname{tr}(\Lambda_2^{1/2} \mathbf{U}_2^T (\mathbf{U}_{\mathbf{Y},2;(i,j)} \mathbf{U}_{\mathbf{Y},2;(i,j)}^T - \mathbf{U}_2 \mathbf{U}_2^T) \mathbf{U}_2 \Lambda_2^{1/2})^2 \\
&\leq c_1^2 \mathbb{E} \operatorname{tr}(\mathbf{U}_2^T (\mathbf{U}_{\mathbf{Y},2;(i,j)} \mathbf{U}_{\mathbf{Y},2;(i,j)}^T - \mathbf{U}_2 \mathbf{U}_2^T) \mathbf{U}_2)^2 \\
&= c_1^2 \mathbb{E} \operatorname{tr}(\mathbf{I}_{p-r} - \mathbf{U}_2^T \mathbf{U}_{\mathbf{Y},2;(i,j)} \mathbf{U}_{\mathbf{Y},2;(i,j)}^T \mathbf{U}_2)^2 \\
&= c_1^2 \mathbb{E} \operatorname{tr}(\mathbf{I}_r - \mathbf{U}_1^T \mathbf{U}_{\mathbf{Y},1;(i,j)} \mathbf{U}_{\mathbf{Y},1;(i,j)}^T \mathbf{U}_1)^2 \\
&\leq c_1^2 r \mathbb{E} \lambda_1^2 (\mathbf{I}_r - \mathbf{U}_1^T \mathbf{U}_{\mathbf{Y},1;(i,j)} \mathbf{U}_{\mathbf{Y},1;(i,j)}^T \mathbf{U}_1) \\
&= O(r (\frac{p}{\lambda_r n})^2) = o(p).
\end{aligned}$$

Note that $\epsilon_{i,j}^{(3)}$ and $\epsilon_{i,j}^{(4)}$ have the same distribution, we have

$$\begin{aligned}
& \mathbb{E}(\epsilon_{ij}^{(3)})^2 = \mathbb{E}(\epsilon_{ij}^{(4)})^2 \\
&= \mathbb{E} \operatorname{tr}(\Lambda_1^{1/2} \mathbf{U}_1^T \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T \mathbf{U}_2 \Lambda_2 \mathbf{U}_2^T \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T \mathbf{U}_1 \Lambda_1^{1/2}) \\
&\leq c_1 \lambda_1 \mathbb{E} \operatorname{tr}(\mathbf{U}_1^T \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T \mathbf{U}_2 \mathbf{U}_2^T \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T \mathbf{U}_1) \\
&\leq c_1 \lambda_1 \mathbb{E} \operatorname{tr}(\mathbf{U}_1^T \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T \mathbf{U}_1) \\
&\leq c_1 \lambda_1 \mathbb{E} \operatorname{tr}(\mathbf{U}_1^T \mathbf{U}_{\mathbf{Y},2;(i,j)} \mathbf{U}_{\mathbf{Y},2;(i,j)}^T \mathbf{U}_1) \\
&\leq c_1 \lambda_1 r \frac{p}{\lambda_r n} = o(p).
\end{aligned}$$

As for $\epsilon_{i,j}^{(5)}$, we have

$$\begin{aligned}
& \mathbb{E}(\epsilon_{ij}^{(5)})^2 \\
&= \mathbb{E} \operatorname{tr}(\Lambda_1^{1/2} \mathbf{U}_1^T \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T \mathbf{U}_1 \Lambda_1^{1/2})^2 \\
&\leq \lambda_1^2 \mathbb{E} \operatorname{tr}(\mathbf{U}_1^T \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)} \tilde{\mathbf{U}}_{\mathbf{Y};(i,j)}^T \mathbf{U}_1)^2 \\
&\leq \lambda_1^2 \mathbb{E} \operatorname{tr}(\mathbf{U}_1^T \mathbf{U}_{\mathbf{Y},2;(i,j)} \mathbf{U}_{\mathbf{Y},2;(i,j)}^T \mathbf{U}_1)^2 \\
&\leq \lambda_1^2 r \left(\frac{p}{\lambda_r n}\right)^2 = o(p).
\end{aligned}$$

Note that

$$\begin{aligned}
& \widehat{\operatorname{tr}(\Lambda_2^2)} \\
&= \frac{2}{(n-k)(n-k-1)} \sum_{1 \leq i < j \leq n-k} (Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j)^2 \\
&+ \frac{2}{(n-k)(n-k-1)} \sum_{1 \leq i < j \leq n-k} \left(\left(\sum_{l=1}^5 \epsilon_{ij}^{(l)} \right) (Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j) + \left(\sum_{l=1}^5 \epsilon_{ij}^{(l)} \right)^2 \right).
\end{aligned}$$

We have

$$\begin{aligned}
& \mathbb{E} \left| \frac{2}{(n-k)(n-k-1)} \sum_{1 \leq i < j \leq n-k} \left(\left(\sum_{l=1}^5 \epsilon_{ij}^{(l)} \right) (Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j) + \left(\sum_{l=1}^5 \epsilon_{ij}^{(l)} \right)^2 \right) \right| \\
& \leq \mathbb{E} \left| \left(\sum_{l=1}^5 \epsilon_{12}^{(l)} \right) (Y_1^T \mathbf{U}_2 \mathbf{U}_2^T Y_2) + \left(\sum_{l=1}^5 \epsilon_{12}^{(l)} \right)^2 \right| \\
& \leq \sqrt{\mathbb{E} \left(\sum_{l=1}^5 \epsilon_{12}^{(l)} \right)^2 \mathbb{E} (Y_1^T \mathbf{U}_2 \mathbf{U}_2^T Y_2)^2 + \mathbb{E} \left(\sum_{l=1}^5 \epsilon_{12}^{(l)} \right)^2} = o(p).
\end{aligned}$$

It follows that

$$\widehat{\text{tr}(\mathbf{\Lambda}_2^2)} = \frac{2}{(n-k)(n-k-1)} \sum_{1 \leq i < j \leq n-k} (Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j)^2 + o_P(p).$$

Now we only need to prove that

$$\frac{2}{(n-k)(n-k-1)} \sum_{1 \leq i < j \leq n-k} (Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j)^2$$

is ratio consistent. Since $\mathbb{E}(Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j)^2 = \text{tr}(\mathbf{\Lambda}_2^2)$ for $i < j$, we have

$$\mathbb{E} \frac{2}{(n-k)(n-k-1)} \sum_{1 \leq i < j \leq n-k} (Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j)^2 = \text{tr}(\mathbf{\Lambda}_2^2).$$

To prove the proposition, we only need to show that

$$\text{Var} \left(\frac{2}{(n-k)(n-k-1)} \sum_{1 \leq i < j \leq n-k} (Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j)^2 \right) = o(\text{tr}^2(\mathbf{\Lambda}_2^2)).$$

Note that

$$\begin{aligned}
 & \mathbb{E} \left(\frac{2}{(n-k)(n-k-1)} \sum_{1 \leq i < j \leq n-k} (Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j)^2 \right)^2 \\
 &= \frac{4}{(n-k)^2(n-k-1)^2} \left(\sum_{1 \leq i < j \leq n-k} (Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j)^2 \right)^2 \\
 &= \frac{4}{(n-k)^2(n-k-1)^2} \mathbb{E} \left(\sum_{i < j} (Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j)^4 \right. \\
 &\quad + \sum_{i < j, k < l: \{i,j\} \cap \{k,l\} = \emptyset} (Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j)^2 (Y_k^T \mathbf{U}_2 \mathbf{U}_2^T Y_l)^2 \\
 &\quad + 2 \sum_{i < j < k} ((Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j)^2 (Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_k)^2 \\
 &\quad + (Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j)^2 (Y_j^T \mathbf{U}_2 \mathbf{U}_2^T Y_k)^2 + (Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_k)^2 (Y_j^T \mathbf{U}_2 \mathbf{U}_2^T Y_k)^2) \Big) \\
 &= \frac{4}{(n-k)^2(n-k-1)^2} \left(\frac{(n-k)(n-k-1)}{2} (6 \operatorname{tr}(\mathbf{\Lambda}_2^4) + 3 \operatorname{tr}^2(\mathbf{\Lambda}_2^2)) \right. \\
 &\quad + \frac{(n-k)(n-k-1)(n-k-2)(n-k-3)}{4} \operatorname{tr}^2(\mathbf{\Lambda}_2^2) \\
 &\quad \left. + (n-k)(n-k-1)(n-k-2) (2 \operatorname{tr}(\mathbf{\Lambda}_2^4) + \operatorname{tr}^2(\mathbf{\Lambda}_2^2)) \right) \\
 &= \operatorname{tr}^2(\mathbf{\Lambda}_2^2) (1 + o(1)).
 \end{aligned}$$

It follows that

$$\operatorname{Var} \left(\frac{2}{(n-k)(n-k-1)} \sum_{1 \leq i < j \leq n-k} (Y_i^T \mathbf{U}_2 \mathbf{U}_2^T Y_j)^2 \right) = o(\operatorname{tr}^2(\mathbf{\Lambda}_2^2)).$$

This completes the proof. \square

Acknowledgements

Write the acknowledgements here.

References

- Bai, Z. and H. Saranadasa (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* 6(2), 311–329.
- Bu Zhou, J. G. a. J.-T. Z. (2017, sep). High-dimensional general linear hypothesis testing under heteroscedasticity. *Journal of Statistical Planning and Inference* 188, 67–81.
- Cai, T. T., W. Liu, and Y. Xia (2013, aug). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(2), 349–372.
- Cai, T. T., Z. Ma, and Y. Wu (2013). Sparse pca: Optimal rates and adaptive estimation. *Annals of Statistics* 41(6), 3074–3110.
- Cai, T. T. and Y. Xia (2014). High-dimensional sparse manova. *Journal of Multivariate Analysis* 131(4), 174–196.
- Chen, S. X. and Y. L. Qin (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Annals of Statistics* 38(2), 808–835.
- Davidson, K. R. and S. J. Szarek (2001). *Handbook of the Geometry of Banach Spaces*, Volume 1. Amsterdam: North-Holland. Handbook of the Geometry of Banach Spaces.
- Fan, J., Y. Fan, and J. Lv (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* 147(1), 186–197.
- Feng, L., C. Zou, Z. Wang, and L. Zhu (2014). Two-sample behrens-fisher problem for high-

REFERENCES

- dimensional data. *Statistica Sinica*.
- Lopes, M. E., L. J. Jacob, and M. J. Wainwright (2015). A more powerful two-sample test in high dimensions using random projection. *Statistics*, 1206–1214.
- Ma, Y., W. Lan, and H. Wang (2015). A high dimensional two-sample test under a low dimensional factor structure. *Journal of Multivariate Analysis* 140, 162–170.
- Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association* 85(411), 686–692.
- Roy, S. N. (1953, jun). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics* 24(2), 220–238.
- Schott, J. R. (2007). Some high-dimensional tests for a one-way manova. *Journal of Multivariate Analysis* 98(9), 1825–1839.
- Shen, D., H. Shen, and J. S. Marron (2013). Consistency of sparse pca in high dimension, low sample size contexts. *Journal of Multivariate Analysis* 115(1), 317–333.
- Srivastava, M. S. (2007). Multivariate theory for analyzing high dimensional data. *Journal of the Japan Statistical Society* 37(1), 53–86.
- Srivastava, M. S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis* 100(3), 518–532As the access to this document is restricted, you may want to look for a different version under "Related research" (further below) orfor a different version of it.

REFERENCES

- Srivastava, M. S. and T. Kubokawa (2013). Tests for multivariate analysis of variance in high dimension under non-normality. *Journal of Multivariate Analysis* 115(1), 204–216.
- Tsai, C.-A. and J. J. Chen (2009). Multivariate analysis of variance test for gene set analysis. *Bioinformatics* 25(7), 897.
- Verstynen, T., J. Diedrichsen, N. Albert, P. Aparicio, and R. B. Ivry (2005). Ipsilateral motor cortex activity during unimanual hand movements relates to task complexity. *Journal of Neurophysiology* 93(3), 1209–1222.
- Yamada, T. and T. Himeno (2015, jul). Testing homogeneity of mean vectors under heteroscedasticity in high-dimension. *Journal of Multivariate Analysis* 139, 7–27.
- Zhao, J. and X. Xu (2016). A generalized likelihood ratio test for normal mean when p is greater than n . *Computational Statistics & Data Analysis*.

first author affiliation

E-mail: (first author email)

second author affiliation

E-mail: (second author email)