High-dimensional two-sample test under spiked covariance

Rui Wang[a], Xingzhong Xu[a,b,*]

[a] *School of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100081,China*

[b] *Beijing Key Laboratory on MCAACI, Beijing Institute of Technology, Beijing 100081,China*

Abstract

This paper considers testing the means of two p-variate normal samples in high dimensional setting. The covariance matrix is assumed to be spiked, which often arises in practice. We derive the asymptotic distribution of Chen and Qin (2010)'s test statistic under a spiked covariance model. Also, a new test procedure is proposed through projection on the orthogonal complement of the principal subspace. The asymptotic normality of the new test statistic is proved and the power function of the test is given. Theoretical and simulation results show that the new test outperforms existing methods substantially when the covariance matrix is spiked.

Keywords: high dimension, mean test, principal subspace, spiked covariance model

1. Introduction

Suppose $X_{k,1}, \ldots, X_{k,n_k}$ are independent identically distributed (i.i.d.) p-dimensional normal random vectors with unknown mean vector μ_k and covariance matrix $\mathbf{\Sigma}$, $k = 1, 2$. We consider the hypothesis testing problem

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2. \tag{1}$$

*Corresponding author

Email address: `xuxz@bit.edu.cn` (Xingzhong Xu)

In this paper, the high dimensional setting is adopted, that is, the dimension p varies as n increases, where $n = n_1 + n_2$ is the total sample size. Testing hypotheses (1) is important in many fields, including biology, finance and economics.

A classical test statistic for hypotheses (1) is Hotelling's T^2 test statistic $(\bar{X}_1 - \bar{X}_2)^T \mathbf{S}^{-1} (\bar{X}_1 - \bar{X}_2)$ where $\bar{X}_1$ and $\bar{X}_2$ are the two sample means and $\mathbf{S} = (n-2)^{-1} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (X_{k,i} - \bar{X}_k)(X_{k,i} - \bar{X}_k)^T$ is the pooled sample covariance matrix. However, Hotelling's test statistic is not defined when $p \geq n-1$. Moreover, Bai and Saranadasa (1996) showed that even if $p < n-1$, Hotelling's test suffers from low power when p is comparable to n. Perhaps, the main reason for the low power of Hotelling's test is that $\mathbf{S}$ is a poor estimator of $\mathbf{\Sigma}$ is large compared with n. See Chen and Qin (2010) and the references therein. For testing hypotheses (1) in high dimensional settings, many test statistics are based on the estimation of $(\mu_1 - \mu_2)^T \mathbf{A} (\mu_1 - \mu_2)$ for a positive definite matrix $\mathbf{A}$. Bai and Saranadasa (1996) proposed a test based on

$$T_{BS} = \|\bar{X}_1 - \bar{X}_2\|^2 - (\frac{1}{n_1} + \frac{1}{n_2}) \operatorname{tr} \mathbf{S},$$

an unbiased estimator of $\|\mu_1 - \mu_2\|^2$. Chen and Qin (2010) modified T_{BS} by removing terms $\sum_{i=1}^{n_k} X_{ki}^T X_{ki}$, $k = 1, 2$, and proposed a test based on

$$\begin{aligned} T_{CQ} &= \frac{\sum_{i \neq j}^{n_1} X_{1i}^T X_{1j}}{n_1(n_1 - 1)} + \frac{\sum_{i \neq j}^{n_2} X_{2i}^T X_{2j}}{n_2(n_2 - 1)} - 2 \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} X_{1i}^T X_{2j}}{n_1 n_2} \\ &= \|\bar{X}_1 - \bar{X}_2\|^2 - \frac{1}{n_1} \operatorname{tr} \mathbf{S}_1 - \frac{1}{n_2} \operatorname{tr} \mathbf{S}_2, \end{aligned}$$

where $\mathbf{S}_k = (n_k - 1)^{-1} \sum_{i=1}^{n_k} (X_{k,i} - \bar{X}_k)(X_{k,i} - \bar{X}_k)^T$, $k = 1, 2$. As an estimator of $\|\mu_1 - \mu_2\|^2$, T_{CQ} is unbiased even if the covariances of the two populations are different. In contrast, T_{BS} is unbiased only when the covariances are the same or $n_1 = n_2$. Srivastava and Du (2008) proposed a test based on

$$T_{SD} = (\bar{X}_1 - \bar{X}_2)^T [\operatorname{diag}(\mathbf{S})]^{-1} (\bar{X}_1 - \bar{X}_2),$$

where $\operatorname{diag}(\mathbf{S})$ is a diagonal matrix with the same diagonal elements as $\mathbf{S}$'s.

In practice, it is often the case that the variables are strongly correlated. See, for example, Chen et al. (2011), Thulin (2014) and Ma et al. (2015). As noted by Ma et al. (2015), however, the tests of Bai and Saranadasa (1996), Srivastava and Du (2008) and Chen and Qin (2010) may not be valid when there are strong correlations between the variables. For example, the condition

$$\operatorname{tr}(\boldsymbol{\Sigma}^4) = o\big(\operatorname{tr}^2(\boldsymbol{\Sigma}^2)\big) \tag{2}$$

imposed by Chen and Qin (2010) is violated when $\boldsymbol{\Sigma}$ has a uniform correlation structure. More precisely, suppose $\boldsymbol{\Sigma} = (1-\rho)\mathbf{I}_p + \rho \mathbf{1}_p \mathbf{1}_p^T$ where $0 < \rho < 1$, $\mathbf{I}_p$ is the p dimensional identity matrix and $\mathbf{1}_p$ is the p dimensional vector with all the elements equal to one. In this case, $\boldsymbol{\Sigma}$ has eigenvalues $1 + \rho(p-1)$ and $1-\rho$ with multiplicities 1 and $p-1$ respectively. Then (2) is violated since as $p \to \infty$, we have

$$\frac{\operatorname{tr}(\boldsymbol{\Sigma}^4)}{\operatorname{tr}^2(\boldsymbol{\Sigma}^2)} = \frac{\big(1+\rho(p-1)\big)^4 + (1-\rho)^4(p-1)}{\big[\big(1+\rho(p-1)\big)^2 + (1-\rho)^2(p-1)\big]^2} \to 1.$$

Note that under the uniform correlation structure, the largest eigenvalue of $\boldsymbol{\Sigma}$ is significantly larger than the rest of eigenvalues. This is a special case of the spiked covariance model

$$\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T + \sigma^2 \mathbf{I}_p, \tag{3}$$

where $\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_r)$, $\boldsymbol{\lambda}_1 \geq \cdots \geq \boldsymbol{\lambda}_r > 0$, $r \geq 1$, $\mathbf{V}$ is a $p \times r$ orthonormal matrix and $\sigma^2 > 0$. The spiked covariance model (3) is adopted by many theoretical studies. See Cai et al. (2013), Birnbaum et al. (2013), Passemier et al. (2017) and the references therein. The spiked covariance arises when variables are strongly correlated and the correlations are determined by a small number of factors. In Section 2, we derive the asymptotic distribution of T_{CQ} under the spiked covariance model. Generally, T_{CQ} is asymptotically distributed as a weighted chi-squared random variable. In a special case, T_{CQ} is asymptotically distributed as the sum of a weighted chi-squared random variable and a normal random variable. We also correct the critical value of T_{CQ} under the spiked covariance model.

Recently, a class of test procedures are proposed through random projection. See, for example, Lopes et al. (2011), Thulin (2014) and Srivastava et al. (2016).

The idea is to project data onto some random lower-dimensional subspaces, and then perform the test using the projected data. It has been shown that these procedures have substantially higher power than competing tests when the variables are correlated. This suggests that projecting data onto certain subspaces may lead to an improvement of the test procedures. Instead of using randomly chosen subspaces, we would like to find the optimal subspace. In Section 3, we will see that under the spiked covariance model, the optimal subspace is the orthogonal complement of the principal subspace. Motivated by this, we propose a new test procedure through projection onto the (estimated) orthogonal complement of the principal subspace. The asymptotic null distribution of our test statistic is derived and the asymptotic power function is also given. Our analysis and simulations show that our test has very attractive power performance under the spiked covariance model.

The rest of the paper is organized as follows. In Section 2, we revisit Chen and Qin (2010)'s test. In Section 3, we propose a test procedure and exploit properties of the test. In Section 4, simulations are carried out and a real data example is given. Section 5 contains some discussion. The technical proofs are presented in Appendix.

2. Asymptotic properties of Chen and Qin (2010)'s test

Throughout the paper, we assume $p \to \infty$ as $n \to \infty$ and $n_1/n_2 \to c \in (0, +\infty)$, that is, we consider high dimensional and balanced data.

In Chen and Qin (2010), the asymptotic normality of T_{CQ} is derived under the condition (2). We shall show that under the null hypothesis, the condition (2) is essential for the asymptotic normality of T_{CQ}. We note that under the null hypothesis, T_{CQ} is a quadratic form of a standard normal random vector. To see this, let $Z_{k,i} = \mathbf{\Sigma}^{-1/2} X_{k,i}$, $i = 1, \ldots, n_k$, $k = 1, 2$. It can be seen that $Z_{k,i}$ is $N_p(0, \mathbf{I}_p)$ distributed under the null hypothesis. Write $Z = (Z_{1,1}^T, \ldots, Z_{1,n_1}^T, Z_{2,1}^T, \ldots, Z_{2,n_2}^T)^T$. Then $T_{CQ} = Z^T\big(\mathbf{B}_n \otimes \mathbf{\Sigma}\big)Z$, where $\otimes$ is the Kronecker product and

$$\mathbf{B}_n = \begin{pmatrix} \frac{1}{n_1(n_1-1)}(\mathbf{1}_{n_1}\mathbf{1}_{n_1}^T - \mathbf{I}_{n_1}) & -\frac{1}{n_1 n_2}\mathbf{1}_{n_1}\mathbf{1}_{n_2}^T \\ -\frac{1}{n_1 n_2}\mathbf{1}_{n_2}\mathbf{1}_{n_1}^T & \frac{1}{n_2(n_2-1)}(\mathbf{1}_{n_2}\mathbf{1}_{n_2}^T - \mathbf{I}_{n_2}) \end{pmatrix}.$$

Using characteristic function method, one can prove the following result which gives a necessary and sufficient condition for the asymptotic normality of the quadratic form of a standard normal random vector.

Lemma 1. *Suppose Y_n is a k_n dimensional standard normal random vector and $\mathbf{A}_n$ is a $k_n \times k_n$ symmetric matrix. Then as $n \to \infty$, a necessary and sufficient condition for*

$$\frac{Y_n^T \mathbf{A}_n Y_n - \mathrm{E}\, Y_n^T \mathbf{A}_n Y_n}{[\mathrm{Var}(Y_n^T \mathbf{A}_n Y_n)]^{1/2}} \xrightarrow{\mathcal{L}} N(0,1) \tag{4}$$

is that

$$\frac{\lambda_1(\mathbf{A}_n^2)}{\mathrm{tr}(\mathbf{A}_n^2)} \to 0, \tag{5}$$

where "$\xrightarrow{\mathcal{L}}$" means convergence of a sequence of random variables in law and $\lambda_i(\cdot)$ means the ith largest eigenvalue.

To apply Lemma 1 to T_{CQ}, one needs to calculate the eigenvalues of $\mathbf{B}_n \otimes \mathbf{\Sigma}$. Note that the eigenvalues of $\mathbf{B}_n$ are $-1/n_1(n_1-1)$, $-1/n_2(n_2-1)$, $(n_1+n_2)/n_1 n_2$ and 0 with multiplicities $n_1 - 1$, $n_2 - 1$, 1 and 1 respectively. Thus,

$$\mathrm{tr}\,(\mathbf{B}_n \otimes \mathbf{\Sigma})^2 = \mathrm{tr}(\mathbf{B}_n^2)\,\mathrm{tr}\,\mathbf{\Sigma}^2 = (\frac{1}{n_1(n_1-1)} + \frac{1}{n_2(n_2-1)} + \frac{2}{n_1 n_2})\,\mathrm{tr}\,\mathbf{\Sigma}^2,$$

and

$$\lambda_1\Big((\mathbf{B}_n \otimes \mathbf{\Sigma})^2\Big) = \lambda_1(\mathbf{B}_n^2)\lambda_1(\mathbf{\Sigma}^2) = \Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)^2 \lambda_1(\mathbf{\Sigma}^2).$$

Because $n_1/n_2 \to c$, the condition

$$\frac{\lambda_1\Big((\mathbf{B}_n \otimes \mathbf{\Sigma})^2\Big)}{\mathrm{tr}\,\Big(\mathbf{B}_n \otimes \mathbf{\Sigma}\Big)^2} \to 0$$

is equivalent to $\lambda_1(\mathbf{\Sigma}^2)/\,\mathrm{tr}\,\mathbf{\Sigma}^2 \to 0$. From

$$\frac{\lambda_1(\mathbf{\Sigma})^4}{(\sum_{i=1}^p \lambda_i(\mathbf{\Sigma})^2)^2} \le \frac{\sum_{i=1}^p \lambda_i(\mathbf{\Sigma})^4}{(\sum_{i=1}^p \lambda_i(\mathbf{\Sigma})^2)^2} \le \frac{\lambda_1(\mathbf{\Sigma})^2 \sum_{i=1}^p \lambda_i(\mathbf{\Sigma})^2}{(\sum_{i=1}^p \lambda_i(\mathbf{\Sigma})^2)^2} = \frac{\lambda_1(\mathbf{\Sigma})^2}{\sum_{i=1}^p \lambda_i(\mathbf{\Sigma})^2},$$

we can see that $\lambda_1^2(\mathbf{\Sigma})/\mathrm{tr}(\mathbf{\Sigma}^2) \to 0$ is equivalent to (2). Then Lemma 1 implies that under the null hypothesis, the condition (2) is a necessary and sufficient condition for

$$\frac{T_{CQ} - \mathrm{E}\, T_{CQ}}{\left[\mathrm{Var}(T_{CQ})\right]^{1/2}} \xrightarrow{\mathcal{L}} N(0,1).$$

The above result implies that Chen and Qin (2010)'s test procedure can be used only when the eigenvalues of $\mathbf{\Sigma}$ are concentrated around their average. In a class of applications, however, the correlations between variables are mainly driven by several common factors, and consequently, $\mathbf{\Sigma}$ has a few eigenvalues which are much larger than the others. See, for example, Jung and Marron (2009), Cai et al. (2013) and Wang and Fan (2017). To characterize such correlations between variables, we consider the spiked covariance structure (3). For $p \geq q$, let $\mathbb{O}_{p\times q}$ denote the collection of all $p \times q$ column orthogonal matrices. We make the following assumption for the covariance matrix $\mathbf{\Sigma}$.

Assumption 1. *The covariance matrix $\mathbf{\Sigma}$ has structure $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T + \sigma^2\mathbf{I}_p$, where $\mathbf{V} \in \mathbb{O}_{p\times r}$, r is a known number and $\mathbf{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_r)$, $\boldsymbol{\lambda}_1 \geq \cdots \geq \boldsymbol{\lambda}_r > 0$. As n, p tend to infinity, the parameters r, σ^2 are fixed and $\mathbf{\Lambda}$ satisfies*

$$\kappa p^\beta \geq \boldsymbol{\lambda}_1 \geq \cdots \geq \boldsymbol{\lambda}_r \geq \kappa^{-1} p^\beta,$$

where $\kappa > 1$ and $\beta \geq 1/2$ are constants.

The covariance structure in Assumption 1 is commonly adopted in PCA study. See Cai et al. (2013), Birnbaum et al. (2013), Passemier et al. (2017) and the references therein. This covariance structure is also connected with the factor model. In fact, the model in Assumption 1 with $\beta = 1$ corresponds to the factor model in Ma et al. (2015) with homoscedastic noise.

In Assumption 1, the column space of $\mathbf{V}$ is the eigenspace of $\mathbf{\Sigma}$ associated with the r leading eigenvalues, and is therefore called principal subspace. Since $\mathbf{V}$ is a column orthogonal matrix, $\mathbf{V}\mathbf{V}^T$ is the orthogonal projection onto the principal subspace. Let $\tilde{\mathbf{V}}$ be a member of $\mathbb{O}_{p\times(p-r)}$ such that the columns of $\tilde{\mathbf{V}}$ are orthogonal to the columns of $\mathbf{V}$. Although such $\tilde{\mathbf{V}}$ is not unique, the orthogonal projection $\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T = \mathbf{I}_p - \mathbf{V}\mathbf{V}^T$ is unique and is equal to the orthogonal projection onto the orthogonal complement of the principal subspace.

For positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp b_n$ to denote $a_n = O(b_n)$ and $b_n = O(a_n)$ as $n \to \infty$. Under Assumption 1, we have

$$\frac{\mathrm{tr}(\mathbf{\Sigma}^4)}{\mathrm{tr}^2(\mathbf{\Sigma}^2)} = \frac{\sum_{i=1}^r (\boldsymbol{\lambda}_i + \sigma^2)^4 + (p-r)\sigma^8}{\left(\sum_{i=1}^r (\boldsymbol{\lambda}_i + \sigma^2)^2 + (p-r)\sigma^4\right)^2} \asymp \frac{p^{4\beta} + p}{(p^{2\beta} + p)^2}.$$

The right hand side tends to 0 if and only if $\beta < 1/2$. Our previous arguments assert that the asymptotic distribution of T_{CQ} won't be normal for $\beta \geq 1/2$. To derive the asymptotic distribution of T_{CQ} for $\beta \geq 1/2$, note that the variation of T_{CQ} is mainly due to $\|\bar{X}_1 - \bar{X}_2\|^2$. Let $\tau = 1/n_1 + 1/n_2$. Under the null hypothesis, we have

$$\mathrm{Var}(\|\bar{X}_1 - \bar{X}_2\|^2) = 2\tau^2\,\mathrm{tr}(\mathbf{\Sigma}^2) = 2\tau^2 \sum_{i=1}^r (\boldsymbol{\lambda}_i + \sigma^2)^2 + 2\tau^2(p-r)\sigma^4,$$

where the first term of the right hand side is of order $p^{2\beta}/n^2$ and the second term is of order p/n^2. If $\beta = 1/2$, the two terms are of the same order. If $\beta > 1/2$, however, the second term is dominated by the first term. This implies that the asymptotic distributions of T_{CQ} are different for $\beta = 1/2$ and $\beta > 1/2$. Since the variance of $(\tau p^\beta)^{-1}\|\bar{X}_1 - \bar{X}_2\|^2$ is bounded under the null hypothesis, we use τp^β to standardize T_{CQ}. The following two theorems give the asymptotic distributions of $(\tau p^\beta)^{-1} T_{CQ}$ when $\beta = 1/2$ and $\beta > 1/2$, respectively.

Theorem 1. *Under Assumption 1, suppose $\beta = 1/2$ and $\lambda_i/p^\beta \to \omega_i \in (0, +\infty)$, $i = 1, \ldots, r$. Let $Z_0, Z_1, \ldots, Z_r$ be i.i.d. $N(0,1)$ random variables, then the following results hold:*

(a) *If $\mu_1 = \mu_2$, then*

$$\frac{1}{\tau p^\beta} T_{CQ} \xrightarrow{w} \sqrt{2}\sigma^2 Z_0 + \sum_{i=1}^r \omega_i Z_i^2 - \sum_{i=1}^r \omega_i,$$

where "$\xrightarrow{w}$" denotes weak convergence.

(b) *If $(\tau p^\beta)^{-1/2}\big(\mathbf{V}^T(\mu_1 - \mu_2)\big)_i \to \zeta_i \in (-\infty, +\infty)$, $i = 1, \ldots, r$, and $\big(\tau p^\beta\big)^{-1}\|\tilde{\mathbf{V}}^T(\mu_1 - \mu_2)\|^2 \to \zeta^* \in [0, +\infty)$, then*

$$\frac{1}{\tau p^\beta} T_{CQ} \xrightarrow{w} \sqrt{2}\sigma^2 Z_0 + \sum_{i=1}^r (\sqrt{\omega_i} Z_i + \zeta_i)^2 + \zeta^* - \sum_{i=1}^r \omega_i.$$

Theorem 2. *Under Assumption 1, suppose $\beta > 1/2$ and $\lambda_i/p^\beta \to \omega_i \in (0,+\infty)$, $i = 1,\ldots,r$. Let $Z_1,\ldots,Z_r$ be i.i.d. $N(0,1)$ random variables, then the following results hold:*

(a) *If $\mu_1 = \mu_2$, then*

$$\frac{1}{\tau p^\beta} T_{CQ} \xrightarrow{w} \sum_{i=1}^r \omega_i Z_i^2 - \sum_{i=1}^r \omega_i.$$

(b) *If $(\tau p^\beta)^{-1/2}\big(\mathbf{V}^T(\mu_1-\mu_2)\big)_i \to \zeta_i \in (-\infty,+\infty)$, $i = 1,\ldots,r$, and $(\tau p^\beta)^{-1}\|\tilde{\mathbf{V}}^T(\mu_1-\mu_2)\|^2 \to \zeta^* \in [0,+\infty)$, then*

$$\frac{1}{\tau p^\beta} T_{CQ} \xrightarrow{w} \sum_{i=1}^r (\sqrt{\omega_i} Z_i + \zeta_i)^2 + \zeta^* - \sum_{i=1}^r \omega_i.$$

Remark 1. By the definitions of ζ_i and ζ^*, we have

$$\frac{1}{\tau p^\beta}\|\mu_1-\mu_2\|^2 = \frac{1}{\tau p^\beta}\|\mathbf{V}^T(\mu_1-\mu_2)\|^2 + \frac{1}{\tau p^\beta}\|\tilde{\mathbf{V}}^T(\mu_1-\mu_2)\|^2 \to \sum_{i=1}^r \zeta_i^2 + \zeta^*.$$

Thus, $\sum_{i=1}^r \zeta_i^2$ and ζ^* characterize the signal strength in the principal subspace and the complement of the principal subspace, respectively. Under the conditions of Theorem 1 or Theorem 2, the following statements are equivalent:

(1) $\zeta_1 = \cdots = \zeta_r = \zeta^* = 0$.

(2) $\|\mu_1 - \mu_2\|^2 = o(\tau p^\beta)$.

(3) The asymptotic distributions of $(\tau p^\beta)^{-1} T_{CQ}$ are the same under the null hypothesis and the alternative hypothesis.

(4) Any test procedure based on T_{CQ} has trivial power asymptotically.

It is implied by Theorem 1 and Theorem 2 that the original critical value of T_{CQ} can not be used when $\beta \geq 1/2$. Now we adjust the critical value of T_{CQ} such that the resulting test has correct level asymptotically. Consider the random variable $W = \sqrt{2p}\sigma^2 Z_0 + \sum_{i=1}^r \boldsymbol{\lambda}_i Z_i^2 - \sum_{i=1}^r \boldsymbol{\lambda}_i$, where $Z_0, Z_1, \ldots, Z_r$ are i.i.d. $N(0,1)$ random variables. Let $F(x; \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_r, \sigma^2)$ be the cumulative distribution function of W. Under the conditions of Theorem 1, we have

$$\frac{W}{p^\beta} \xrightarrow{w} \sqrt{2}\sigma^2 Z_0 + \sum_{i=1}^r \omega_i Z_i^2 - \sum_{i=1}^r \omega_i.$$

Under the conditions of Theorem 2, we have

$$\frac{W}{p^\beta} \xrightarrow{w} \sum_{i=1}^{r} \omega_i Z_i^2 - \sum_{i=1}^{r} \omega_i.$$

Hence in both case, we have

$$\sup_{x\in\mathbb{R}} \big| \Pr\big(\frac{1}{\tau} T_{CQ} \le x\big) - \Pr\big(W \le x\big)\big| = o(1).$$

Thus, if we reject the null hypothesis when

$$\frac{1}{\tau} T_{CQ} > F^{-1}(1-\alpha; \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_r, \sigma^2),$$

then the resulting test is asymptotically level α for $\beta \ge 1/2$. However, the distribution $F(x; \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_r, \sigma^2)$ involves some unknown parameters. In order to consistently estimate $F(x; \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_r, \sigma^2)$, we need to estimate $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_r$ and σ^2. In section 3, we will give their estimators $\hat{\boldsymbol{\lambda}}_1, \ldots, \hat{\boldsymbol{\lambda}}_r$ and $\hat{\sigma}_*^2$. Proposition 2 asserts that these estimators are ratio consistent. Now we propose a corrected T_{CQ} test procedure which reject the null hypothesis with α level of significance if

$$\tau^{-1} T_{CQ} > F^{-1}(1-\alpha; \hat{\boldsymbol{\lambda}}_1, \ldots, \hat{\boldsymbol{\lambda}}_r, \hat{\sigma}_*^2).$$

Then under the conditions of either Theorem 1 or Theorem 2, the corrected T_{CQ} test procedure is asymptotically level α.

As we have seen in Remark 1, for the corrected T_{CQ} test, the separation boundary between the testable and non-testable regions is $\|\mu_1 - \mu_2\|^2 \asymp p^\beta/n$. Then as β increases, the corrected T_{CQ} test procedure becomes less powerful. This implies that the power of the corrected T_{CQ} test procedure is negatively affected by the large eigenvalues of $\boldsymbol{\Sigma}$.

3. A projection test

In Section 2, we adjusted the critical value of T_{CQ} such that the corrected T_{CQ} test procedure is asymptotically level α under Assumption 1. However, the power of the corrected T_{CQ} test procedure is negatively affected by the large eigenvalues of $\mathbf{\Sigma}$. This motivates us to propose a new test for the hypotheses (1) under Assumption 1.

Recently, a class of test procedures have been proposed through random projection to lower dimensional subspace. See, for example, Lopes et al. (2011), Thulin (2014) and Srivastava et al. (2016). It is known that random projection based methods offer higher power when the variables are dependent. However, these test procedures are randomized, which is undesirable in practice. Then, is there an optimal projection which is nonrandomized?

For $\mathbf{O} \in \mathbb{O}_{p\times k}$ $(k \le p)$, define statistic

$$T(\mathbf{O}) = \|\mathbf{O}^T(\bar{X}_1 - \bar{X}_2)\|^2 - \frac{1}{n_1}\operatorname{tr}(\mathbf{O}^T\mathbf{S}_1\mathbf{O}) - \frac{1}{n_2}\operatorname{tr}(\mathbf{O}^T\mathbf{S}_2\mathbf{O}).$$

Then $T(\mathbf{O})$ is Chen and Qin (2010)'s statistic on the transformed data $\mathbf{O}^T X_{k,i}$. Denote by $\Phi(\cdot)$ the cumulative distribution function of the standard normal random variable. Under the condition (2), Chen and Qin (2010) proved that the asymptotic power of T_{CQ} under the local alternative is

$$\Phi\left(\Phi^{-1}(\alpha) + \frac{\|\mu_1 - \mu_2\|^2}{\sqrt{2\tau^2\operatorname{tr}(\mathbf{\Sigma}^2)}}\right).$$

Hence the power of T_{CQ} is largely impacted by $\|\mu_1 - \mu_2\|^2/\sqrt{2\tau^2\operatorname{tr}(\mathbf{\Sigma}^2)}$, which may be viewed as a signal to noise ratio (SNR). Consequently, $\|\mathbf{O}^T(\mu_1 - \mu_2)\|^2/\sqrt{2\tau^2\operatorname{tr}(\mathbf{O}^T\mathbf{\Sigma}^2\mathbf{O})}$ measures the power of $T(\mathbf{O})$. To consider an average-case scenario, like Lopes et al. (2011), we temporarily place a prior on $\mu_1 - \mu_2$ and assume that $\mu_1 - \mu_2$ is from the uniform distribution on the unit sphere. In this case, an average SNR can be defined as

$$\mathrm{E}\left(\frac{\|\mathbf{O}^T(\mu_1 - \mu_2)\|^2}{\sqrt{2\tau^2\operatorname{tr}(\mathbf{O}^T\mathbf{\Sigma}^2\mathbf{O})}}\right) = \frac{k/p}{\sqrt{2\tau^2\operatorname{tr}(\mathbf{O}^T\mathbf{\Sigma}^2\mathbf{O})}}. \tag{6}$$

It can be expected that the $T(\mathbf{O})$ maximizing the average SNR has the best average power behavior among $\{T(\mathbf{O}) : \mathbf{O} \in \mathbb{O}_{p\times k},\ k \le p\}$.

Note that for fixed k, (6) is maximized when the columns of $\mathbf{O}$ are equal to the last k eigenvectors of $\mathbf{\Sigma}$. Thus, it remains to maximize

$$\frac{k/p}{\sqrt{2\tau^2 \sum_{i=p-k+1}^{p} \lambda_i^2(\mathbf{\Sigma})}} \tag{7}$$

over k. If $k \leq p-r$, (7) is equal to $\sqrt{k}/(\sqrt{2}\sigma^2\tau p)$ which is an increasing function of k. If $k > p-r$, we have

$$\begin{aligned}
&\frac{k/p}{\sqrt{2\tau^2(\sum_{i=p-k+1}^{r} \lambda_i^2(\mathbf{\Sigma}) + k\sigma^4)}} \leq \frac{1}{\sqrt{2\tau^2(\kappa^{-2}p^{2\beta} + (p-r)\sigma^4)}} \\
=&\frac{p/(p-r)}{\sqrt{\kappa^{-2}p^{2\beta}/((p-r)\sigma^4) + 1}} \frac{(p-r)/p}{\sqrt{2\tau^2((p-r)\sigma^4)}}.
\end{aligned}$$

Hence for sufficiently large p, we have

$$\frac{k/p}{\sqrt{2\tau^2(\sum_{i=p-k+1}^{r} \lambda_i^2(\mathbf{\Sigma}) + k\sigma^4)}} < \frac{(p-r)/p}{\sqrt{2\tau^2((p-r)\sigma^4)}},$$

and (7) is maximized when $k = p - r$. Consequently, for sufficiently large p, (6) is maximized when $\mathbf{O} = \tilde{\mathbf{V}}$.

The above discussion motivates us to consider the variable

$$T_1 = \|\tilde{\mathbf{V}}^T(\bar{X}_1 - \bar{X}_2)\|^2 - \frac{1}{n_1}\operatorname{tr}(\tilde{\mathbf{V}}^T\mathbf{S}_1\tilde{\mathbf{V}}) - \frac{1}{n_2}\operatorname{tr}(\tilde{\mathbf{V}}^T\mathbf{S}_2\tilde{\mathbf{V}}).$$

Note that based on $\tilde{\mathbf{V}}^T X_{ki}$, the likelihood ratio test statistic for hypothesis (1) is $\|\tilde{\mathbf{V}}^T(\bar{X}_1 - \bar{X}_2)\|^2$. In this view, T_1 can be regarded as a restricted likelihood ratio statistic. It can be shown that T_1 is asymptotically normal.

Proposition 1. *Under Assumption 1, suppose $\|\mu_1 - \mu_2\|^2 = o(p/n)$, we have*

$$\frac{T_1 - \|\tilde{\mathbf{V}}^T(\mu_1 - \mu_2)\|^2}{\sigma^2\sqrt{2\tau^2 p}} \sim \frac{\chi^2_{p-r} - (p-r)}{\sqrt{2(p-r)}} + o_P(1).$$

Remark 2. Proposition 1 implies that

$$\frac{T_1 - \|\tilde{\mathbf{V}}^T(\mu_1 - \mu_2)\|^2}{\sigma^2\sqrt{2\tau^2 p}} \xrightarrow{\mathcal{L}} N(0,1).$$

However, compared with the standard normal distribution, the standardized Chi-squared distribution is a better approximation of the statistic. This fact is implied by the proof of Proposition 1.

Note that T_1 relies on the subspace $\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T$ which is typically unknown and thus needs to be estimated. Let $\hat{\mathbf{V}}$ and $\hat{\tilde{\mathbf{V}}}$ denote the first r and last $p-r$ eigenvectors of $\mathbf{S}$, respectively. Anderson (1963) proved that the maximum likelihood estimator (MLE) of $\mathbf{V}$ is $\hat{\mathbf{V}}$. This fact, together with the equalities $\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T = \mathbf{I}_p - \mathbf{V}\mathbf{V}^T$ and $\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T = \mathbf{I}_p - \hat{\mathbf{V}}\hat{\mathbf{V}}^T$, implies that the MLE of $\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T$ is $\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T$. Thus, as the main term of T_1, $\|\tilde{\mathbf{V}}^T(\bar{X}_1 - \bar{X}_2)\|^2$ can be estimated by $\|\hat{\tilde{\mathbf{V}}}^T(\bar{X}_1 - \bar{X}_2)\|^2$. In T_1, the centralization term $n_1^{-1}\operatorname{tr}(\tilde{\mathbf{V}}^T\mathbf{S}_1\tilde{\mathbf{V}}) + n_2^{-1}\operatorname{tr}(\tilde{\mathbf{V}}^T\mathbf{S}_2\tilde{\mathbf{V}})$ is an unbiased estimator of $\tau(p-r)\sigma^2$, which makes $\mathrm{E}\, T_1 = 0$ under the null hypothesis. However, the centralization of $\|\hat{\tilde{\mathbf{V}}}^T(\bar{X}_1 - \bar{X}_2)\|^2$ is more involved.

Under the null hypothesis,

$$
\begin{aligned}
&\mathrm{E}[\|\hat{\tilde{\mathbf{V}}}^T(\bar{X}_1 - \bar{X}_2)\|^2|\mathbf{S}] = \mathrm{E}[\|\hat{\tilde{\mathbf{V}}}^T(\bar{X}_1 - \bar{X}_2)\|^2|\mathbf{S}] = \tau\operatorname{tr}(\hat{\tilde{\mathbf{V}}}^T\boldsymbol{\Sigma}\hat{\tilde{\mathbf{V}}}) \\
=&\tau(p-r)\sigma^2 + \tau\operatorname{tr}(\hat{\tilde{\mathbf{V}}}^T\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T\hat{\tilde{\mathbf{V}}}).
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\operatorname{tr}(\hat{\tilde{\mathbf{V}}}^T\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T\hat{\tilde{\mathbf{V}}}) = \operatorname{tr}(\boldsymbol{\Lambda}^{1/2}\mathbf{V}^T\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T\mathbf{V}\boldsymbol{\Lambda}^{1/2}) = \operatorname{tr}(\boldsymbol{\Lambda}^{1/2}(\mathbf{I}_r - \mathbf{V}^T\hat{\mathbf{V}}\hat{\mathbf{V}}^T\mathbf{V})\boldsymbol{\Lambda}^{1/2}) \\
=&\sum_{i=1}^r(1 - \sum_{l=1}^r(\hat{v}_l^T v_i)^2)\boldsymbol{\lambda}_i,
\end{aligned}
$$

where $\hat{v}_i$ and v_i are the ith columns of $\hat{\mathbf{V}}$ and $\mathbf{V}$, respectively, $i = 1, \ldots, r$. Under $p = O(n\boldsymbol{\lambda}_r)$ and some other regular conditions, Wang and Fan (2017), Theorem 3.2 asserts that

$$
\hat{v}_j^T v_i = O_P(\epsilon_{j,n}), \quad 1 \le i \ne j \le r, \quad \text{and} \quad \hat{v}_i^T v_i = \frac{1}{\sqrt{1 + \frac{p}{n(\boldsymbol{\lambda}_i + \sigma^2)}\sigma^2}} + O_P(\epsilon_{i,n}), \quad 1 \le i \le r,
$$

where $\epsilon_{i,n}$ is higher order infinitesimal, $i = 1, \ldots, r$. This motivates us to approximate $\operatorname{tr}(\hat{\tilde{\mathbf{V}}}^T\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T\hat{\tilde{\mathbf{V}}})$ by

$$
\sum_{i=1}^r\Big(1 - \frac{1}{1 + \frac{p}{n(\boldsymbol{\lambda}_i + \sigma^2)}\sigma^2}\Big)\boldsymbol{\lambda}_i = \sum_{i=1}^r \frac{p\sigma^2}{n\boldsymbol{\lambda}_i + (n+p)\sigma^2}\boldsymbol{\lambda}_i.
$$

Hence

$$
\mathrm{E}[\|\hat{\tilde{\mathbf{V}}}^T(\bar{X}_1 - \bar{X}_2)\|^2|\mathbf{S}] \approx \tau(p-r)\sigma^2 + \tau\sum_{i=1}^r \frac{p\sigma^2}{n\boldsymbol{\lambda}_i + (n+p)\sigma^2}\boldsymbol{\lambda}_i.
$$

To centralize $\|\hat{\hat{\mathbf{V}}}^T(\bar{X}_1-\bar{X}_2)\|^2$, we need to estimate $\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_r$ and σ^2. Anderson (1963) proved that the MLE of σ^2 is $\hat{\sigma}^2=(p-r)^{-1}\sum_{i=r+1}^{p}\lambda_i(\mathbf{S})$ and the MLE of $\boldsymbol{\lambda}_i$ is $\lambda_i(\mathbf{S})-\hat{\sigma}^2$, $i=1,\ldots,r$. However, Lemma 5 in Appendix B implies that $\hat{\sigma}^2$ is downward biased and $\lambda_i(\mathbf{S})-\hat{\sigma}^2$ is upward biased. (See also Passemier et al. (2017) and Wang and Fan (2017).) Motivated by the results in Lemma 5, we propose the following bias-corrected estimators:

$$\hat{\sigma}_*^2=\Big(1-\frac{r}{n-2}\Big)^{-1}\hat{\sigma}^2,\quad \hat{\boldsymbol{\lambda}}_i=\lambda_i(\mathbf{S})-\frac{p+n-r-2}{n-2}\hat{\sigma}_*^2\quad i=1,\ldots,r.$$

The following proposition gives the convergence rate of these estimators.

Proposition 2. *Under Assumption 1, we have*

$$\hat{\sigma}_*^2=\sigma^2+O_P\Big(\max\big(\frac{1}{\sqrt{np}},\frac{1}{p}\big)\Big),\tag{8}$$

and

$$\frac{\hat{\boldsymbol{\lambda}}_i}{\boldsymbol{\lambda}_i}=1+O_P\Big(\max\big(\frac{1}{\sqrt{n}},\frac{1}{p^{\beta}}\big)\Big).\tag{9}$$

Remark 3. Recently, Passemier et al. (2017) proposed a bias-corrected estimator of σ^2:

$$\Big(1+\frac{1}{n-2}\big(r+\hat{\sigma}^2\sum_{i=1}^{r}\frac{1}{\boldsymbol{\lambda}_i}\big)\Big)\hat{\sigma}^2.$$

In their paper, $\boldsymbol{\lambda}_i$'s are fixed and known. This is different from our model where $\boldsymbol{\lambda}_i$'s are divergent and unknown. Under Assumption 1, we have

$$\Big(1+\frac{1}{n-2}\big(r+\hat{\sigma}^2\sum_{i=1}^{r}\frac{1}{\boldsymbol{\lambda}_i}\big)\Big)\hat{\sigma}^2=\Big(1+\frac{r}{n-2}\Big)\hat{\sigma}^2+O_P\big(\frac{1}{np^{\beta}}\big)=\hat{\sigma}_*^2+O_P\big(\frac{1}{n^2}\big)+O_P\big(\frac{1}{np^{\beta}}\big).$$

So the difference between Passemier et al. (2017)'s estimator and $\hat{\sigma}_*^2$ is minor in our model.

Remark 4. Recently, Wang and Fan (2017) proposed an estimator of $\lambda_i(\boldsymbol{\Sigma})=\boldsymbol{\lambda}_i+\sigma^2$, $i=1,\ldots,r$:

$$\max\Big(\lambda_i(\mathbf{S})-\frac{p}{n-2}\Big(1-\frac{p}{p-r}\frac{r}{n-2}\Big)^{-1}\hat{\sigma}^2,0\Big).$$

They showed that under $p>n-2$, $p=O(n\boldsymbol{\lambda}_r)$ and some other conditions,

$$\frac{1}{\boldsymbol{\lambda}_i+\sigma^2}\max\Big(\lambda_i(\mathbf{S})-\frac{p}{n-2}\Big(1-\frac{p}{p-r}\frac{r}{n-2}\Big)^{-1}\hat{\sigma}^2,0\Big)=1+O_P\Big(\frac{1}{\boldsymbol{\lambda}_i}\sqrt{\frac{p}{n}}+\frac{1}{\sqrt{n}}\Big).$$

Note that under Assumption 1 and $p > n-2$, we have

$$\frac{1}{\boldsymbol{\lambda}_i}\sqrt{\frac{p}{n}} + \frac{1}{\sqrt{n}} \asymp \frac{1}{\sqrt{n}} \asymp \max\big(\frac{1}{\sqrt{n}}, \frac{1}{p^{\beta}}\big).$$

In this case, Wang and Fan (2017)'s estimator and our estimator have the same convergence rate, although the estimands are slightly different. Compared with Wang and Fan (2017)'s result, Proposition 2 doesn't need the conditions $p > n-2$ and $p = O(n\boldsymbol{\lambda}_r)$.

Now we propose the following test statistic:

$$T_2 = \|\hat{\tilde{\mathbf{V}}}^T(\bar{X}_1 - \bar{X}_2)\|^2 - \tau(p-r)\hat{\sigma}_*^2 - \tau\sum_{i=1}^{r}\frac{p\hat{\sigma}_*^2}{n\hat{\boldsymbol{\lambda}}_i + (n+p)\hat{\sigma}_*^2}\hat{\boldsymbol{\lambda}}_i.$$

To standardize T_2, note that

$$\begin{aligned}
&\mathrm{Var}\big(\|\hat{\tilde{\mathbf{V}}}^T(\bar{X}_1 - \bar{X}_2)\|^2\big) = 2\tau^2\big(\operatorname{tr}(\hat{\tilde{\mathbf{V}}}^T\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T\hat{\tilde{\mathbf{V}}})^2 + 2\sigma^2\operatorname{tr}(\hat{\tilde{\mathbf{V}}}^T\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T\hat{\tilde{\mathbf{V}}}) + \sigma^4(p-r)\big)\\
=&2\tau^2\Big(\sum_{l=1}^{r}\big(1-\sum_{i=1}^{r}(\hat{v}_l^Tv_i)^2\big)^2\boldsymbol{\lambda}_i^2 + 2\sum_{1\le i<j\le r}\big(\sum_{l=1}^{r}(\hat{v}_l^Tv_i)(\hat{v}_l^Tv_j)\big)^2\boldsymbol{\lambda}_i\boldsymbol{\lambda}_j\\
&+2\sigma^2\sum_{l=1}^{r}\big(1-\sum_{i=1}^{r}(\hat{v}_l^Tv_i)^2\big)\boldsymbol{\lambda}_i + \sigma^4(p-r)\Big)\\
\approx&2\tau^2\Big(\sum_{i=1}^{r}\big(\frac{p\sigma^2}{n\boldsymbol{\lambda}_i + (n+p)\sigma^2}\boldsymbol{\lambda}_i\big)^2 + 2\sigma^2\sum_{i=1}^{r}\frac{p\sigma^2}{n\boldsymbol{\lambda}_i + (n+p)\sigma^2}\boldsymbol{\lambda}_i + \sigma^4(p-r)\Big).
\end{aligned}$$

$$Q = T_2 \Big/ \left(2\tau^2\Big(\sum_{i=1}^{r}\big(\frac{p\hat{\sigma}_*^2}{n\hat{\boldsymbol{\lambda}}_i + (n+p)\hat{\sigma}_*^2}\hat{\boldsymbol{\lambda}}_i\big)^2 + 2\sigma^2\sum_{i=1}^{r}\frac{p\hat{\sigma}_*^2}{n\hat{\boldsymbol{\lambda}}_i + (n+p)\hat{\sigma}_*^2}\hat{\boldsymbol{\lambda}}_i + \hat{\sigma}_*^4(p-r)\Big)\right)^{1/2}$$

In view of Proposition 1, we reject the null hypothesis if

$$Q > \frac{\chi^2_{\alpha,p-r} - (p-r)}{\sqrt{2(p-r)}},$$

where $\chi^2_{\alpha,p-r}$ is the upper α quantile of a χ^2_{p-r} random variable.

The following theorem establishes the asymptotic normality of T_2.

Theorem 3. *Under Assumption 1, suppose $p/n^2 \to 0$ and $\|\mu_1 - \mu_2\|^2 = O(\sqrt{p}/n)$, we have*

$$\frac{T_2 - \|\tilde{\mathbf{V}}^T(\mu_1 - \mu_2)\|^2}{\sigma^2\sqrt{2\tau^2 p}} \xrightarrow{\mathcal{L}} N(0,1).$$

Remark 5. The asymptotic normality of T_2 is closely related to the convergence rate of $\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T$ to $\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T$. Lemma 6 in Appendix B and the equality $\|\hat{\mathbf{V}}\hat{\mathbf{V}}^T - \mathbf{V}\mathbf{V}^T\| = \|\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T - \tilde{\mathbf{V}}\tilde{\mathbf{V}}^T\|$ imply that $\|\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T - \tilde{\mathbf{V}}\tilde{\mathbf{V}}^T\| = O_P(p/(p^\beta n))$. Hence $\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T$ can consistently estimate $\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T$ only if $p/(p^\beta n) \to 0$. Moreover, Cai et al. (2013)'s Theorem 5 implies that no other estimator has faster convergence rate than $\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T$. The asymptotic normality of T_2 requires the condition

$$p^{-1}\big(p^\beta\|\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T - \tilde{\mathbf{V}}\tilde{\mathbf{V}}^T\|^2 + 1\big)^2 \xrightarrow{P} 0.$$

This is equivalent to $\|\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T - \tilde{\mathbf{V}}\tilde{\mathbf{V}}^T\|^2 = o_P(\sqrt{p}/p^\beta)$. Hence in Theorem 3 we assume $p/n^2 \to 0$. The proof of Theorem 3 implies that the asymptotic normality of T_2 is not valid if the condition $p/n^2 \to 0$ is violated.

By Proposition 2,

$$\frac{p\hat{\sigma}_*^2}{n\hat{\boldsymbol{\lambda}}_i + (n+p)\hat{\sigma}_*^2}\hat{\boldsymbol{\lambda}}_i \le \frac{p}{n}\hat{\sigma}_*^2 = o_P(p).$$

Then it follows from Proposition 2 and Theorem 3 that

$$Q = \frac{T_2}{\sigma^2\sqrt{2\tau^2(p-r)}}(1 + o_P(1)) \xrightarrow{\mathcal{L}} N(0,1).$$

Thus, our test is asymptotically level α.

The asymptotic power function of our test can be obtained immediately from Theorem 3.

Corollary 1. *Under the conditions of Theorem 3, the asymptotic power function of our test is*

$$\Phi\Big(-\Phi^{-1}(1-\alpha) + \frac{\|\tilde{\mathbf{V}}^T(\mu_1 - \mu_2)\|^2}{\sigma^2\sqrt{2\tau^2 p}}\Big).$$

From the expression of the asymptotic power function, we can see that the power of our test is not affected by the large eigenvalues of $\mathbf{\Sigma}$.

Having derived the asymptotic power function of our test, we are now in a position to provide a comparison with other tests. Again, to make an average-case comparison against other tests, we place a prior on $\mu_1 - \mu_2$. Suppose that the norm $\|\mu_1 - \mu_2\|$ is nonrandom while the orientation $\delta = (\mu_1 - \mu_2)/\|\mu_1 - \mu_2\|$

is from the uniform distribution on the unit sphere. Then $\|\tilde{\mathbf{V}}^T(\mu_1-\mu_2)\|^2 = \|\mu_1-\mu_2\|^2\|\tilde{\mathbf{V}}^T\delta\|^2$. Since $\mathrm{E}\,\|\tilde{\mathbf{V}}^T\delta\|^2 = (p-r)/p$ and $\mathrm{Var}\left(\|\tilde{\mathbf{V}}^T\delta\|^2\right) = 2r(p-r)/\big(p^2(p+2)\big)$, we have $\|\tilde{\mathbf{V}}^T\delta\|^2 = 1 + O_P(1/p)$. In this case, the asymptotic power function of our test is equal to

$$\Phi\Big(-\Phi^{-1}(1-\alpha) + \frac{\|\mu_1-\mu_2\|^2}{\sigma^2\sqrt{2\tau^2 p}}\Big).$$

So the separation boundary of our test is $\|\mu_1-\mu_2\|^2 \asymp \sqrt{p}/n$.

To compare our test with the corrected T_{CQ} test, recall that the separation boundary of the corrected T_{CQ} test procedure is $\|\mu_1-\mu_2\|^2 \asymp p^\beta/n$. Thus, when $\beta > 1/2$, the testable region of our test is larger than that of the corrected T_{CQ} test procedure.

We would also like to compare our test with T_{SD}. Srivastava and Du (2008) showed that the asymptotic power function of T_{SD} is

$$\Phi\Big(-\Phi^{-1}(1-\alpha) + \frac{\|\mathbf{\Sigma}_d^{-1/2}(\mu_1-\mu_2)\|^2}{\sqrt{2\tau^2\,\mathrm{tr}(\mathbf{R}^2)}}\Big),$$

where $\mathbf{\Sigma}_d = \mathrm{diag}(\mathbf{\Sigma})$ and $\mathbf{R} = \mathbf{\Sigma}_d^{-1/2}\mathbf{\Sigma}\mathbf{\Sigma}_d^{-1/2}$ is the population correlation matrix. It is known that the power of T_{SD} is highest when $\mathbf{\Sigma}$ is diagonal.

First we consider the uniform correlation structure $\mathbf{\Sigma} = (1-\rho)\mathbf{I}_p + \rho\mathbf{1}_p\mathbf{1}_p^T$ $(0<\rho<1)$ which is far away from a diagonal matrix. In this case, the diagonal entries of Σ are all 1. We have

$$\frac{\|\mathbf{\Sigma}_d^{-1/2}(\mu_1-\mu_2)\|^2}{\sqrt{2\tau^2\,\mathrm{tr}(\mathbf{R}^2)}} = \frac{1}{\sqrt{\rho^2 p + (1-\rho^2)}}\frac{\|\mu_1-\mu_2\|^2}{\sqrt{2\tau^2 p}}.$$

Hence the asymptotic relative efficiency of our test with respect to T_{SD} is $\sqrt{\rho^2 p + (1-\rho^2)}$. Our test has much higher power.

Now we consider the diagonal covariance matrix. In this case, $\mathbf{\Sigma} = \mathrm{diag}(\boldsymbol{\lambda}_1 + \sigma^2, \ldots, \boldsymbol{\lambda}_r + \sigma^2, \sigma^2, \ldots, \sigma^2)$ and $\mathbf{R} = \mathbf{I}_p$. Note that $\|\mathbf{\Sigma}_d^{-1/2}(\mu_1-\mu_2)\|^2 = \|\mu_1-\mu_2\|^2\|\mathbf{\Sigma}^{-1/2}\delta\|^2$. We have

$$\mathrm{E}\,\|\mathbf{\Sigma}^{-1/2}\delta\|^2 = \frac{1}{p}\,\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\right) = \sigma^{-2}(1+o_P(1))$$

and

$$\mathrm{Var}\left(\|\mathbf{\Sigma}^{-1/2}\delta\|^2\right) = \frac{2}{p+2}\Big(\frac{1}{p}\,\mathrm{tr}\left(\mathbf{\Sigma}^{-2}\right) - \big(\frac{1}{p}\,\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\right)\big)^2\Big) = o\big(\frac{1}{p}\big).$$

Hence $\|\mathbf{\Sigma}^{-1/2}\delta\|^2 = \sigma^{-2}(1 + o_P(1))$ and the asymptotic power function of T_{SD} equals to

$$\Phi\Big(- \Phi^{-1}(1-\alpha) + \frac{\|\mu_1 - \mu_2\|^2}{\sigma^2\sqrt{2\tau^2 p}}\Big).$$

Thus, under the diagonal covariance matrix, the asymptotic power function of our test is the same as that of T_{SD}.

4. Numerical studies

4.1. Simulation results

In this section, we consider the simulation performance of the proposed test and compare it with several other alternatives: (1) Chen and Qin (2010)'s test (CQ); (2) Ma et al. (2015)'s test (FAST); (3) the corrected T_{CQ} test procedure proposed in Section 2 (CCQ); (4) Srivastava and Du (2008)'s test (SD); (5) Lopes et al. (2011)'s test (LJW). The data generation mechanism is as follow. We randomly choose a $\mathbf{U} \in \mathbb{O}_{p\times p}$ from Haar invariant distribution. Let d_i equal to p^{β} plus a random error from $U(0,1)$ (Uniform distribution between 0 and 1), $i = 1, \ldots, r$. Construct $p \times p$ diagonal matrix $\mathbf{D} = \text{diag}(\sqrt{d_1}, \ldots, \sqrt{d_r}, 1, \ldots, 1)$. Then, we independently generate data by the formula

$$X_{k,i} = \mu_k + \mathbf{UD}Y_{k,i} \quad i = 1, \ldots, n_k \text{ and } k = 1, 2,$$

where $Y_{k,i}$ is a p dimensional random vector whose entries are i.i.d. random variables with common distribution F. We will consider three different distributions of F.

- Normal: $F \sim N(0, 1)$.
- Chi-squared: $F \sim (\chi^2_4 - 4)/\sqrt{8}$.
- Student's t: $F \sim t_4/\sqrt{2}$, where t_4 is a Student's t random variable with degree of freedom 4.

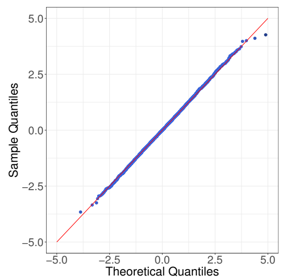
Throughout the simulations, we take nominal level $\alpha = 0.05$ and $r = 2$.

In Section 3, the critical value of our test is determined by the quantile of $(\chi^2_{p-r} - (p-r))/\sqrt{2(p-r)}$. We have proved that the distribution of Q is asymptotically equal to that of $(\chi^2_{p-r} - (p-r))/\sqrt{2(p-r)}$. Now we use Q-Q plot to compare these two distributions in the finite sample case. Figure 1 displays the Q-Q plots in different combinations of sample size and dimension. For $n_1 = n_2 = 50$, $p = 500$ or 800, the right tail of Q is a little heavier than the standardized Chi-squared distribution. Otherwise, the distribution of Q can be well approximated by the standardized Chi-squared distribution. Figure 2 displays the Q-Q plots in different combinations of F and β. It can be seen that the distribution of Q is very close to that of the standardized Chi-squared distribution, even under non-normal distributions.

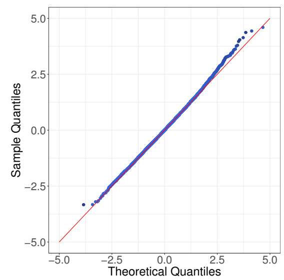
Next, we consider the simulation of the empirical level. Samples are repeatedly generated 2000 times to calculate empirical level. The result is listed in Table **??**. The empirical levels of the CQ test are larger than the nominal level in all cases, especially when β is large. The empirical levels of the SD test are close to the nominal level when $\beta = 1/2$, but tend to be smaller than the nominal level as β increases. The empirical levels of the FAST test are very close to the nominal level in most cases, but tend to be smaller than the nominal level when $n_1 = n_2 = 50$, $p = 800$ and $\beta = 1/2$. The empirical levels of the CCQ test are very close to the nominal level in all cases. Since the LJW test is exact, it's empirical levels are very close to the nominal level. The empirical level of our test is a little inflated for $n_1 = n_2 = 50$, but converges to the nominal level as the sample size increases.

Now we consider the simulation of the empirical power. In view of Corollary 1, we define SNR as $\text{SNR} = \|\mu_1 - \mu_2\|^2/(\sigma^2\sqrt{2\tau^2 p})$. We take $\mu_1 = \mathbf{0}_p$. The orientation of μ_2 is from the uniform distribution on the unit sphere. The norm of μ_2 is selected to make SNR equal to specific values. Samples are repeatedly generated 2000 times to calculate empirical power. The simulation results are illustrated in Figure **??**. From the results, we can find that when $\mathbf{\Sigma}$ is spiked, the new test outperforms T_{CQ} substantially; when $\mathbf{\Sigma}$ is not spiked, all three tests have similar performance.

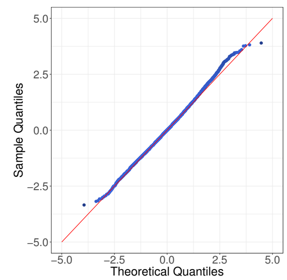
(a) $n_1 = n_2 = 50$, $p = 200$.



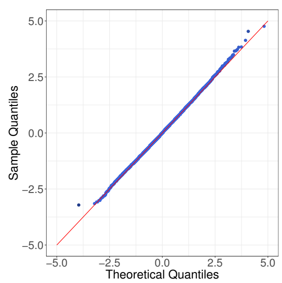
(b) $n_1 = n_2 = 50$, $p = 500$.



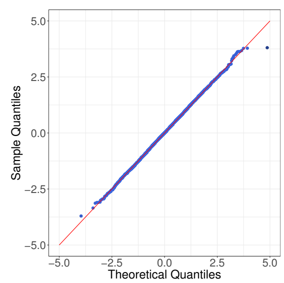
(c) $n_1 = n_2 = 50$, $p = 800$.



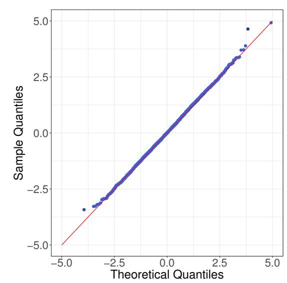
(d) $n_1 = n_2 = 100$, $p = 200$.



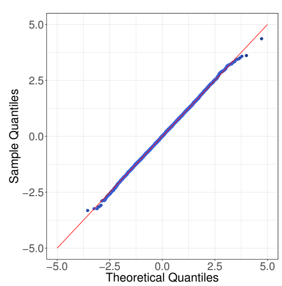
(e) $n_1 = n_2 = 100$, $p = 500$.



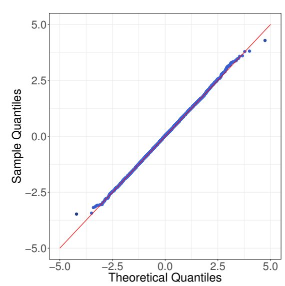
(f) $n_1 = n_2 = 100$, $p = 800$.



(g) $n_1 = n_2 = 150$, $p = 200$.



(h) $n_1 = n_2 = 150$, $p = 500$.



(i) $n_1 = n_2 = 150$, $p = 800$.

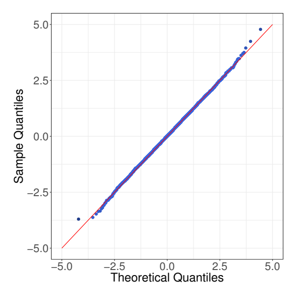
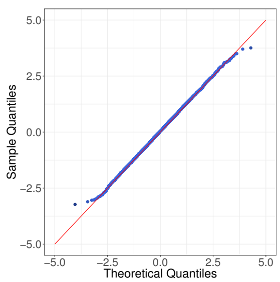
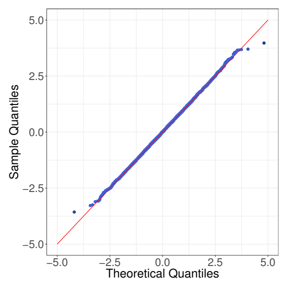


Figure 1: Q-Q plots of the empirical distribution of Q against that of $(\chi^2_{p-r} - (p - r))/\sqrt{2(p-r)}$ based on xxx independently generated Q. In all cases, F is normal and $\beta = 1$.

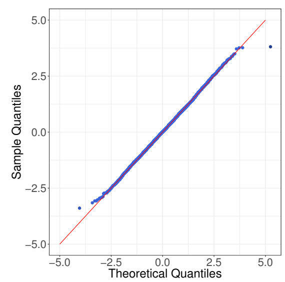
(a) Normal. $\beta = 0.5$.



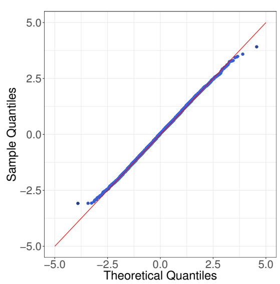
(b) Normal. $\beta = 1$.



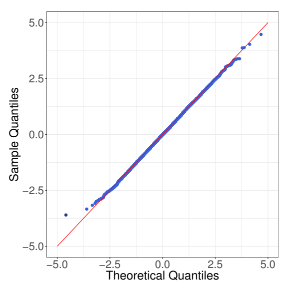
(c) Normal. $\beta = 2$.



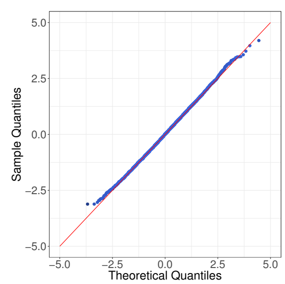
(d) Chi-squared. $\beta = 0.5$.



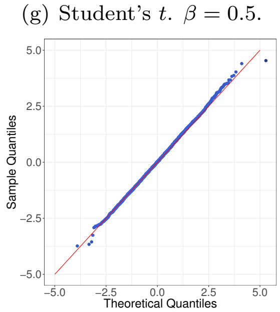
(e) Chi-squared. $\beta = 1$.



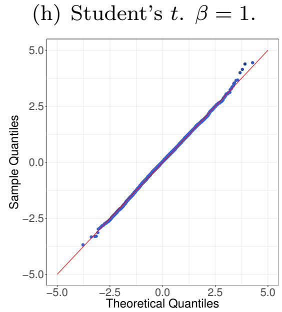
(f) Chi-squared. $\beta = 2$.



(g) Student's t. $\beta = 0.5$.



(h) Student's t. $\beta = 1$.



(i) Student's t. $\beta = 2$.

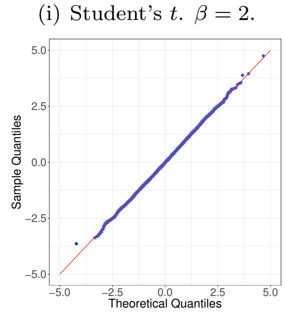
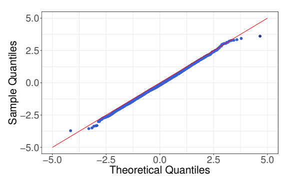
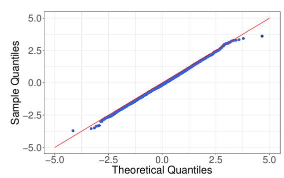


Figure 2: Q-Q plots of the empirical distribution of Q against that of $(\chi^2_{p-r} - (p - r))/\sqrt{2(p-r)}$ based on xxx independently generated Q. In all cases, $n_1 = n_2 = 100$ and $p = 500$.

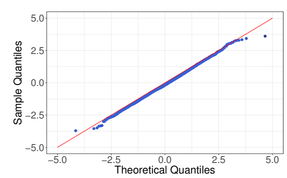
(a) $n_1 = n_2 = 50$, $p = 200$.



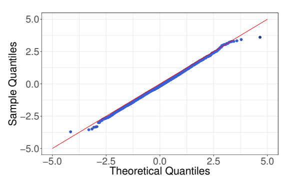
(b) $n_1 = n_2 = 50$, $p = 500$.



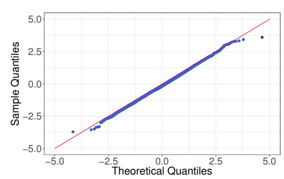
(c) $n_1 = n_2 = 50$, $p = 800$.



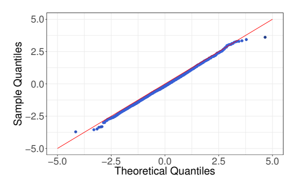
(d) $n_1 = n_2 = 100$, $p = 200$.



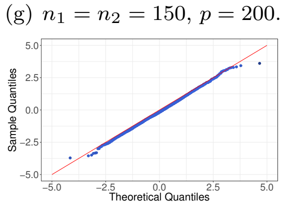
(e) $n_1 = n_2 = 100$, $p = 500$.



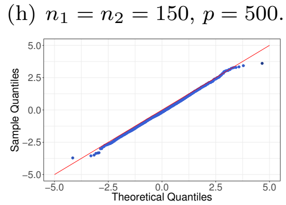
(f) $n_1 = n_2 = 100$, $p = 800$.



(g) $n_1 = n_2 = 150$, $p = 200$.



(h) $n_1 = n_2 = 150$, $p = 500$.



(i) $n_1 = n_2 = 150$, $p = 800$.

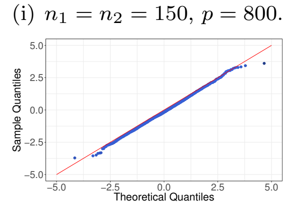
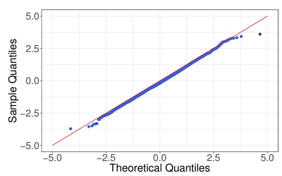
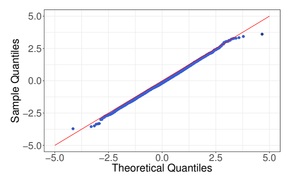


Figure 3: Empirical power simulation. In all cases, F is normal and $\beta = 1$.

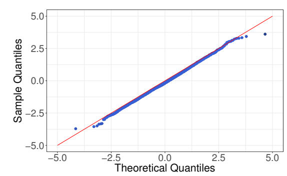
(a) Normal. $\beta = 0.5$.



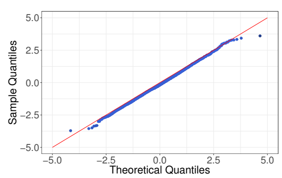
(b) Normal. $\beta = 1$.



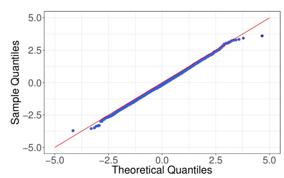
(c) Normal. $\beta = 2$.



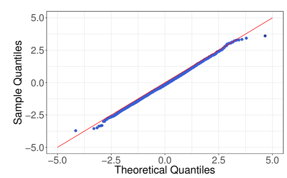
(d) Chi-squared. $\beta = 0.5$.



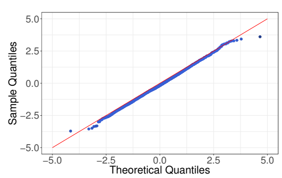
(e) Chi-squared. $\beta = 1$.



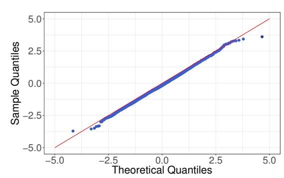
(f) Chi-squared. $\beta = 2$.



(g) Student's t. $\beta = 0.5$.



(h) Student's t. $\beta = 1$.



(i) Student's t. $\beta = 2$.

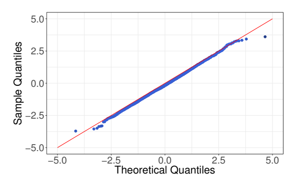


Figure 4: Empirical power simulation. In all cases, $n_1 = n_2 = 100$ and $p = 500$.

4.2. Real data analysis

In this section, we study the practical problem considered in Ma et al. (2015). The task is to test whether Monday stock returns are equal to those of other trading days on average. Define an observation be the log return of stocks in a day. Hence p is the total number of stocks. Let sample 1 and sample 2 be the observations on Monday and the other trading days, respectively. Then we would like to test $H_0 : \mu_1 = \mu_2$ v.s. $H_1 : \mu_1 \neq \mu_2$. We collected the data of $p = 710$ stocks of China from 01/04/2013 to 12/31/2014. There are total $n_1 = 95$ Monday and $n_2 = 388$ other trading days.

We assume $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. The first eigenvalue of $\mathbf{S}$ is 0.14, which is significantly larger than the others. In fact, the second eigenvalue is 0.02. Hence there's clearly a spiked eigenvalue. We take $r = 1$ and perform our new test. The p value is 0.149, which is obtained by permutation method with 1000 permutations. Hence, the null hypothesis can not be rejected for $\alpha = 0.05$. We draw the same conclusion as Ma et al. (2015).

5. Conclusion remark

This paper is concerned with the problem of testing the equality of means in the setting of high dimension and spiked covariance. We derived the asymptotic distribution of Chen and Qin (2010)'s test statistic. To reduce the variance of T_{CQ}, we dropped big variance terms and obtain a new test statistic. The asymptotic normality of the new statistic is proved and the asymptotic power is given.

In another paper, Zhao and Xu (2016) proved that their test statistic can be written in the form of projection. Their simulation results showed that their test performs well under strong correlations. Our work partially explains why their test performs well although the projections are slightly different.

Spiked covariance is an important correlation pattern and has been widely studied in terms of PCA. In PCA, authors focus on the principal subspace.

However, in some circumstances, as our work have shown, the complement of the principal subspace is more useful.

In our paper, we have assumed r is known. If r is an unknown positive number, a consistent estimator of r is

$$\hat{r} = \operatorname{argmax}_{l \leq R} \frac{\lambda_l(\mathbf{S})}{\lambda_{l+1}(\mathbf{S})}, \tag{10}$$

where R is a hyperparameter. See Ahn and Horenstein (2013) for detail.

The asymptotic normality of the new test statistic relies on the assumption $\sqrt{p}/n \to 0$. In the situation of small n or very large p, the critical value of the new test can be determined by permutation method. Our simulation shows that the new test still performs well. It remains a theoretical interest to study the asymptotic behavior of permutation based test in these situations.

Non normality

Appendix A Proofs of the results in Section 2

Proof of Lemma 1. By a standard orthogonal transformation, we can write

$$\frac{Y_n^T \mathbf{A}_n Y_n - \mathrm{E}\, Y_n^T \mathbf{A}_n Y_n}{[\operatorname{Var}(Y_n^T \mathbf{A}_n Y_n)]^{1/2}} = \sum_{i=1}^{k_n} \frac{\lambda_i(\mathbf{A}_n)}{\left[2\operatorname{tr}(\mathbf{A}_n^2)\right]^{1/2}} (Z_{ni}^2 - 1), \tag{11}$$

where $Z_{n1}, \ldots, Z_{nk_n}$ are independent standard normal random variables.

If 5 holds, then

$$\begin{aligned}
&\sum_{i=1}^{k_n} \mathrm{E}\left[\frac{\lambda_i^2(\mathbf{A}_n)}{2\mathrm{tr}(\mathbf{A}_n^2)} (Z_{ni}^2 - 1)^2 \left\{\frac{\lambda_i^2(\mathbf{A}_n)}{2\mathrm{tr}(\mathbf{A}_n^2)} (Z_{ni}^2 - 1)^2 \geq \epsilon\right\}\right] \\
\leq &\sum_{i=1}^{k_n} \frac{\lambda_i^2(\mathbf{A}_n)}{2\operatorname{tr}(\mathbf{A}_n^2)} \mathrm{E}\left[(Z_{n1}^2 - 1)^2 \left\{\frac{\lambda_1(\mathbf{A}_n^2)}{2\operatorname{tr}(\mathbf{A}_n^2)} (Z_{n1}^2 - 1)^2 \geq \epsilon\right\}\right] \\
= &\frac{1}{2} \mathrm{E}\left[(Z_{n1}^2 - 1)^2 \left\{\frac{\lambda_1(\mathbf{A}_n^2)}{2\operatorname{tr}(\mathbf{A}_n^2)} (Z_{n1}^2 - 1)^2 \geq \epsilon\right\}\right] \to 0.
\end{aligned}$$

Hence 4 follows by Lindeberg's central limit theorem.

Conversely, if 4 holds, we will prove that there is a subsequence of $\{n\}$ along which 5 holds. Then 5 follows by a standard contradiction argument.

Denote $c_{ni} = \lambda_i(\mathbf{A}_n)/\big[2\,\mathrm{tr}(\mathbf{A}_n^2)\big]^{1/2}$, $i = 1, \ldots, k_n$. Since 4 holds, the characteristic function of $\sum_{i=1}^{k_n} c_{ni}(Z_{ni}^2 - 1)$ converges to $\exp(-t^2/2)$ for every t. Denote by $\log z$ ($z \in \mathbb{C}$) the principal branch of the complex logarithm. For $t \in (-1/2, 1/2)$, we have

$$\begin{aligned}
&\mathrm{E}\big[\exp\big(it\sum_{j=1}^{k_n} c_{nj}(Z_{nj}^2 - 1)\big)\big] = \exp\Big(-i(\sum_{j=1}^{k_n} c_{nj})t - \frac{1}{2}\sum_{j=1}^{k_n}\log(1 - i2c_{nj}t)\Big)\\
&= \exp\Big(-i(\sum_{j=1}^{k_n} c_{nj})t + \frac{1}{2}\sum_{j=1}^{k_n}\sum_{l=1}^{+\infty}\frac{1}{l}(i2c_{nj}t)^l\Big) = \exp\Big(-i(\sum_{j=1}^{k_n} c_{nj})t + \frac{1}{2}\sum_{l=1}^{+\infty}\Big[\sum_{j=1}^{k_n}(c_{nj})^l\Big]\frac{1}{l}(i2t)^l\Big)\\
&= \exp\Big(-\frac{1}{2}t^2 + \frac{1}{2}\sum_{l=3}^{+\infty}\Big[\sum_{j=1}^{k_n}(c_{nj})^l\Big]\frac{1}{l}(i2t)^l\Big),
\end{aligned}$$

where the second equality holds since $0 \le c_{ni} \le \sqrt{2}/2$ by definition. Let $b_{nl} = \sum_{j=1}^{k_n}(c_{nj})^l$, $n = 1, 2, \cdots$ and $l = 3, 4, \cdots$. Note that for $l \ge 3$, we have

$$|b_{nl}| = \big|\sum_{j=1}^{k_n}(c_{nj})^l\big| \le \big|\sum_{j=1}^{k_n}(c_{nj})^2\big| = 1/2.$$

By Helly's selection theorem, there's a subsequence of $\{n\}$ along which $\lim_{n\to\infty} b_{nl} = b_l$ exists for every l. For this subsequence, applying dominated convergence theorem yields

$$\mathrm{E}\big[\exp\big(it\sum_{j=1}^{k_n} c_{nj}(Z_{nj}^2 - 1)\big)\big] \to \exp\Big(-\frac{1}{2}t^2 + \frac{1}{2}\sum_{l=3}^{+\infty} b_l\frac{1}{l}(i2t)^l\Big), \quad t \in \big(-\frac{1}{2}, \frac{1}{2}\big).$$

But the left hand side converges to $\exp(-t^2/2)$. It follows that

$$-\frac{1}{2}t^2 + \frac{1}{2}\sum_{l=3}^{+\infty} b_l\frac{1}{l}(i2t)^l = -\frac{1}{2}t^2 + 2\pi mi, \quad t \in \big(-\frac{1}{2}, \frac{1}{2}\big),$$

for some integer m. By the uniqueness of power series, we mulst have $m = 0$ and $b_l = 0$ for $l \ge 3$. Then 5 follows by noting that $b_{n4} \ge \max_j (c_{nj})^4$. $\square$

Proofs of Theorems 1 and 2. In both Theorems, (a) is a corrolary of (b). We shall prove (b) of Theorems 1 and 2 simultaneously.

Note that $(n_k - 1)\mathbf{S}_k \sim \mathrm{Wishart}_p(n_k - 1, \mathbf{\Sigma})$, $k = 1, 2$, where $\mathrm{Wishart}_p(m, \Psi)$ is the p dimensional Wishart distribution with parameter Ψ and m degrees of freedom. We have

$$\mathrm{E}\Big(\frac{1}{n_1}\operatorname{tr}\mathbf{S}_1+\frac{1}{n_2}\operatorname{tr}\mathbf{S}_2\Big)=\tau\operatorname{tr}\mathbf{\Sigma},$$

and

$$\begin{aligned}&\operatorname{Var}\Big(\frac{1}{n_1}\operatorname{tr}\mathbf{S}_1+\frac{1}{n_2}\operatorname{tr}\mathbf{S}_2\Big)=\Big(\frac{2}{n_1^2(n_1-1)}+\frac{2}{n_2^2(n_2-1)}\Big)\operatorname{tr}\mathbf{\Sigma}^2\\=&O\Big(\frac{1}{n^3}(p^{2\beta}+p)\Big)=O\Big(\frac{p^{2\beta}}{n^3}\Big).\end{aligned}$$

It follows that

$$\begin{aligned}&\frac{1}{n_1}\operatorname{tr}\mathbf{S}_1+\frac{1}{n_2}\operatorname{tr}\mathbf{S}_2=\tau\operatorname{tr}\mathbf{\Sigma}+O_P\Big(\frac{1}{n\sqrt{n}}p^{\beta}\Big)\\=&\tau\sum_{i=1}^{r}(\boldsymbol{\lambda}_i+\sigma^2)+\tau(p-r)\sigma^2+O_P\Big(\frac{1}{n\sqrt{n}}p^{\beta}\Big)\\=&\tau p^{\beta}\sum_{i=1}^{r}\omega_i+\tau(p-r)\sigma^2+o_P\Big(\frac{1}{n}p^{\beta}\Big).\end{aligned}$$

Thus,

$$\frac{1}{\tau p^{\beta}}\big(\frac{1}{n_1}\operatorname{tr}\mathbf{S}_1+\frac{1}{n_2}\operatorname{tr}\mathbf{S}_2\big)=\sum_{i=1}^{r}\omega_i+p^{1-\beta}\sigma^2+o_P(1). \tag{12}$$

Next we deal with $\|\bar{X}_1-\bar{X}_2\|^2$. Note that we have

$$\|\bar{X}_1-\bar{X}_2\|^2=\|\mathbf{V}^T(\bar{X}_1-\bar{X}_2)\|^2+\|\tilde{\mathbf{V}}^T(\bar{X}_1-\bar{X}_2)\|^2.$$

These two terms are independent. For the first term, note that $\mathbf{V}^T(\bar{X}_1-\bar{X}_2)\sim N_r\big(\mathbf{V}^T(\mu_1-\mu_2),\tau(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)\big)$, we have

$$\begin{aligned}\|\mathbf{V}^T(\bar{X}_1-\bar{X}_2)\|^2&\sim\sum_{i=1}^{r}\Big(\sqrt{\tau(\lambda_i+\sigma^2)}Z_i+\big(\mathbf{V}^T(\mu_1-\mu_2)\big)_i\Big)^2\\&=\tau p^{\beta}\sum_{i=1}^{r}\Big(\sqrt{p^{-\beta}(\lambda_i+\sigma^2)}Z_i+\frac{1}{\sqrt{\tau p^{\beta}}}\big(\mathbf{V}^T(\mu_1-\mu_2)\big)_i\Big)^2.\end{aligned}$$

By the assumptions of the theorem, we have that

$$\frac{1}{\tau p^{\beta}}\|\mathbf{V}^T(\bar{X}_1-\bar{X}_2)\|^2\xrightarrow{w}\sum_{i=1}^{r}(\sqrt{\omega_i}Z_i+\zeta_i)^2. \tag{13}$$

As for $\|\tilde{\mathbf{V}}^T(\bar{X}_1-\bar{X}_2)\|^2$, we have that

$$\begin{aligned}&\|\tilde{\mathbf{V}}^T(\bar{X}_1-\bar{X}_2)\|^2=\big\|\tilde{\mathbf{V}}^T(\mu_1-\mu_2)+\tilde{\mathbf{V}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\big\|^2\\=&\|\tilde{\mathbf{V}}^T(\mu_1-\mu_2)\|^2+\big\|\tilde{\mathbf{V}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\big\|^2+2(\mu_1-\mu_2)^T\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big).\end{aligned}$$

Since $\tilde{\mathbf{V}}^T(\bar{X}_1 - \bar{X}_2) \sim N_{p-r}(\tilde{\mathbf{V}}^T(\mu_1 - \mu_2), \sigma^2\tau\mathbf{I}_{p-r})$, by central limit theorem, we have

$$\frac{\left\|\tilde{\mathbf{V}}^T\big((\bar{X}_1 - \mu_1) - (\bar{X}_2 - \mu_2)\big)\right\|^2 - \sigma^2\tau(p-r)}{\sigma^2\tau\sqrt{2(p-r)}} \xrightarrow{\mathcal{L}} N(0,1).$$

For the intersection term, we have

$$\begin{aligned}
&2(\mu_1 - \mu_2)^T\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T\big((\bar{X}_1 - \mu_1) - (\bar{X}_2 - \mu_2)\big) \sim N(0, 4\sigma^2\tau\|\tilde{\mathbf{V}}^T(\mu_1 - \mu_2)\|^2)\\
=&O_P(\sqrt{\tau}\|\tilde{\mathbf{V}}^T(\mu_1 - \mu_2)\|) = o_P(\tau p^\beta).
\end{aligned}$$

It follows that

$$\frac{1}{\tau p^\beta}\big(\left\|\tilde{\mathbf{V}}^T(\bar{X}_1 - \bar{X}_2)\right\|^2 - \sigma^2\tau(p-r) - \left\|\tilde{\mathbf{V}}^T(\mu_1 - \mu_2)\right\|^2\big) \xrightarrow{\mathcal{L}} \sqrt{2}\sigma^2\delta_{\{\frac{1}{2}\}}(\beta)Z_0, \quad (14)$$

where $\delta_{\frac{1}{2}}(\beta)$ equals 1 if $\beta = 1/2$ and equals 0 otherwise.

Combining (12) (13) and (14) leads to

$$\begin{aligned}
&\frac{1}{\tau p^\beta}T_{CQ} = \frac{1}{\tau p^\beta}\big(\|\bar{X}_1 - \bar{X}_2\|^2 - \frac{1}{n_1}\operatorname{tr}\mathbf{S}_1 - \frac{1}{n_2}\operatorname{tr}\mathbf{S}_2\big)\\
=&\frac{1}{\tau p^\beta}\|\mathbf{V}^T(\bar{X}_1 - \bar{X}_2)\|^2 + \frac{1}{\tau p^\beta}\big(\|\tilde{\mathbf{V}}^T(\bar{X}_1 - \bar{X}_2)\|^2 - \sigma^2\tau(p-r) - \|\tilde{\mathbf{V}}^T(\mu_1 - \mu_2)\|^2\big)\\
&- \frac{1}{\tau p^\beta}\Big(\frac{1}{n_1}\operatorname{tr}\mathbf{S}_1 + \frac{1}{n_2}\operatorname{tr}\mathbf{S}_2\Big) + \frac{\sigma^2(p-r)}{p^\beta} + \frac{1}{\tau p^\beta}\|\tilde{\mathbf{V}}^T(\mu_1 - \mu_2)\|^2\\
=&\sum_{i=1}^r(\sqrt{\omega_i}Z_i + \zeta_i)^2 + \sqrt{2}\sigma^2\delta_{\{\frac{1}{2}\}}(\beta)Z_0 - (\sum_{i=1}^r\omega_i + p^{1-\beta}\sigma^2) + \frac{\sigma^2(p-r)}{p^\beta} + \zeta^* + o_P(1)\\
\xrightarrow{\mathcal{L}}&\sum_{i=1}^r(\sqrt{\omega_i}Z_i + \zeta_i)^2 + \zeta^* + \sqrt{2}\sigma^2\delta_{\{\frac{1}{2}\}}(\beta)Z_0 - \sum_{i=1}^r\omega_i.
\end{aligned}$$

This implies the conclusions of Theorem 1 and Theorem 2.

□

Appendix B Proofs of the results in Section 3

Lemma 2 (Weyl's inequality)**.** *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two symmetric* $n \times n$ *matrices and* $\mathbf{C} = \mathbf{A} + \mathbf{B}$*. If* $r + s - 1 \le i \le j + k - n$*, we have*

$$\lambda_j(\mathbf{A}) + \lambda_k(\mathbf{B}) \le \lambda_i(\mathbf{C}) \le \lambda_r(\mathbf{A}) + \lambda_s(\mathbf{B}).$$

See, for example, Horn and Johnson (2012), Theorem 4.3.1.

Lemma 3 (Cai et al. (2015), Proposition 1)**.** *Let* $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *be* $p \times p$ *symmetric matrices. Let* $r < p$ *be arbitrary and let* $\mathbf{V}_1, \mathbf{V}_2 \in \mathbb{O}_{p,r}$ *be formed by the* r *leading singular vectors of* $\mathbf{A}_1$ *and* $\mathbf{A}_2$*, respectively. Then*

$$\|\mathbf{A}_1 - \mathbf{A}_2\| \geq \frac{1}{2}(\lambda_r(\mathbf{A}_2) - \lambda_{r+1}(\mathbf{A}_2))\|\mathbf{V}_1\mathbf{V}_1^T - \mathbf{V}_2\mathbf{V}_2^T\|.$$

Lemma 4 (Davidson and Szarek (2001), Theorem II.7)**.** *Let* $\mathbf{Z}$ *be a* $p \times n$ *random matrix with i.i.d.* $N(0,1)$ *entries. Then for any* $t > 0$,

$$\Pr(\sqrt{\lambda_1(\mathbf{Z}\mathbf{Z}^T)} > \sqrt{n} + \sqrt{p} + t) \leq e^{-t^2/2},$$
$$\Pr(\sqrt{\lambda_{\min(n,p)}(\mathbf{Z}\mathbf{Z}^T)} < \sqrt{n} - \sqrt{p} - t) \leq e^{-t^2/2}.$$

We give two useful corollaries of Lemma 4.

Corollary 2. *Suppose that* $\mathbf{W}_n$ *is a* $p \times p$ *random matrix which is distributed as* $\mathrm{Wishart}_p(n, \mathbf{I}_p)$*. Then as* $n, p \to \infty$,

$$\lambda_1(\mathbf{W}_n) = O_P(\max(n,p)).$$

Proof. The result follows from the inequality

$$\Pr\Big(\frac{\lambda_1(\mathbf{W}_n)}{\max(n,p)} > 16\Big) \leq \Pr\Big(\lambda_1(\mathbf{W}_n) > 8(n+p)\Big) \leq \Pr\Big(\lambda_1(\mathbf{W}_n) > 4(\sqrt{n}+\sqrt{p})^2\Big)$$
$$= \Pr\Big(\sqrt{\lambda_1(\mathbf{W}_n)} > 2(\sqrt{n}+\sqrt{p})\Big) \leq \Pr\Big(\sqrt{\lambda_1(\mathbf{W}_n)} > 2\sqrt{n}+\sqrt{p}\Big) \leq e^{-n/2},$$

where the last inequality follows from Lemma 4 with $t = \sqrt{n}$. □

Corollary 3. *Suppose that* $\mathbf{W}_n$ *is a* $p \times p$ *random matrix which is distributed as* $\mathrm{Wishart}_p(n, \mathbf{I}_p)$*. Then as* $n, p \to \infty$,

$$\|\frac{1}{n}\mathbf{W}_n - \mathbf{I}_p\| = O_P\Big(\max\big(\sqrt{\frac{p}{n}}, \frac{p}{n}\big)\Big).$$

Proof. Since the eigenvalues of $n^{-1}\mathbf{W}_n - \mathbf{I}_p$ are $n^{-1}\lambda_1(\mathbf{W}_n) - 1, \ldots, n^{-1}\lambda_p(\mathbf{W}_n) - 1$, we have

$$\|\frac{1}{n}\mathbf{W}_n - \mathbf{I}_p\| = \max\Big(\frac{1}{n}\lambda_1(\mathbf{W}_n) - 1, 1 - \frac{1}{n}\lambda_p(\mathbf{W}_n)\Big).$$

This, together with the union bound, yields

$$\Pr\Big(\|\frac{1}{n}\mathbf{W}_n - \mathbf{I}_p\| > 4\big(\sqrt{\frac{p}{n}} + \frac{p}{n}\big)\Big) \leq \Pr\Big(\lambda_1(\mathbf{W}_n) > \big(\sqrt{n}+2\sqrt{p}\big)^2\Big) + \Pr\Big(\lambda_p(\mathbf{W}_n) < n - 4\sqrt{np} - 4p\Big).$$

For the first term, we have

$$\Pr\Big(\lambda_1(\mathbf{W}_n) > \big(\sqrt{n}+2\sqrt{p}\big)^2\Big) = \Pr\Big(\sqrt{\lambda_1(\mathbf{W}_n)} > \sqrt{n}+2\sqrt{p}\Big) \le e^{-p/2},$$

where the last inequality follows from Lemma 4 with $t=\sqrt{p}$.

We now show that the second term is also bounded by $e^{-p/2}$. To see this, note that if $p > n/4$, then $n - 4\sqrt{np} - 4p \le n - 4p < 0$. In this case, $\Pr\Big(\lambda_p(\mathbf{W}_n) < n - 4\sqrt{np} - 4p\Big) = 0$. If $p \le n/4$, we have

$$\begin{aligned} &\Pr\Big(\lambda_p(\mathbf{W}_n) < n - 4\sqrt{np} - 4p\Big) \le \Pr\Big(\lambda_p(\mathbf{W}_n) < n - 4\sqrt{np} + 4p\Big) \\ =&\Pr\Big(\sqrt{\lambda_p(\mathbf{W}_n)} < \sqrt{n} - \sqrt{2p}\Big) \le e^{-p/2}, \end{aligned}$$

where the last inequality follows from Lemma 4 with $t=\sqrt{p}$.

Now we conclude that

$$\Pr\Big(\|\frac{1}{n}\mathbf{W}_n - \mathbf{I}_p\| > 4\big(\sqrt{\frac{p}{n}}+\frac{p}{n}\big)\Big) \le 2e^{-p/2}.$$

Consequently,

$$\|\frac{1}{n}\mathbf{W}_n - \mathbf{I}_p\| = O_P\Big(\sqrt{\frac{p}{n}}+\frac{p}{n}\Big) = O_P\Big(\max\big(\sqrt{\frac{p}{n}},\frac{p}{n}\big)\Big).$$

$\square$

Lemma 5. *Under Assumption 1, we have*

$$\lambda_i(\mathbf{S}) = \boldsymbol{\lambda}_i + \frac{p+n-r-2}{n-2}\sigma^2 + O_P\Big(\max\big(\frac{p^\beta}{\sqrt{n}},1\big)\Big), \quad i=1,\dots,r, \tag{15}$$

and

$$\hat{\sigma}^2 = (1-\frac{r}{n-2})\sigma^2 + O_P\Big(\max\big(\frac{1}{\sqrt{np}},\frac{1}{p}\big)\Big). \tag{16}$$

Proof. Let $\mathbf{\Sigma} = \mathbf{U}\mathbf{E}\mathbf{U}^T$ denote the spectral decomposition of $\mathbf{\Sigma}$, where $\mathbf{U} = (\mathbf{V},\tilde{\mathbf{V}})$ and $\mathbf{E} = \mathrm{diag}(\boldsymbol{\lambda}_1+\sigma^2,\dots,\boldsymbol{\lambda}_r+\sigma^2,\sigma^2,\dots,\sigma^2)$. Let $\mathbf{Z}$ be a $p\times(n-2)$ random matrix with i.i.d. $N(0,1)$ entries. Denote $\mathbf{Z} = (\mathbf{Z}_{(1)}^T,\mathbf{Z}_{(2)}^T)^T$, where $\mathbf{Z}_{(1)}$ and $\mathbf{Z}_{(2)}$ are the first r rows and last $p-r$ rows of $\mathbf{Z}$. Then the sample covariance matrix $\mathbf{S}$ has the same distribution as the random matrix $(n-2)^{-1}\mathbf{U}\mathbf{E}^{1/2}\mathbf{Z}\mathbf{Z}^T\mathbf{E}^{1/2}\mathbf{U}^T$. So we have $\lambda_i(\mathbf{S}) \sim (n-2)^{-1}\lambda_i(\mathbf{Z}^T\mathbf{E}\mathbf{Z})$, $i=1,\dots,r$ and $\mathrm{tr}(\mathbf{S}) \sim (n-2)^{-1}\,\mathrm{tr}(\mathbf{Z}^T\mathbf{E}\mathbf{Z})$.

To prove (15), we only need to consider $(n-2)^{-1}\lambda_i(\mathbf{Z}^T\mathbf{E}\mathbf{Z})$. Note that $\mathbf{Z}^T\mathbf{E}\mathbf{Z} = \mathbf{Z}_{(1)}^T(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)\mathbf{Z}_{(1)} + \sigma^2\mathbf{Z}_{(2)}^T\mathbf{Z}_{(2)}$. From this and Weyl's inequality, for $i=1,\ldots,r$, we have

$$
\begin{aligned}
&\Big|\frac{1}{n-2}\lambda_i(\mathbf{Z}^T\mathbf{E}\mathbf{Z}) - \frac{1}{n-2}\lambda_i\big(\mathbf{Z}_{(1)}^T(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)\mathbf{Z}_{(1)}\big) - \frac{p-r}{n-2}\sigma^2\Big| \\
=&\Big|\lambda_i\big(\frac{1}{n-2}\mathbf{Z}^T\mathbf{E}\mathbf{Z}\big) - \lambda_i\big(\frac{1}{n-2}\mathbf{Z}_{(1)}^T(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)\mathbf{Z}_{(1)} + \frac{p-r}{n-2}\sigma^2\mathbf{I}_{n-2}\big)\Big| \\
\le&\Big\|\frac{1}{n-2}\mathbf{Z}^T\mathbf{E}\mathbf{Z} - \frac{1}{n-2}\mathbf{Z}_{(1)}^T(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)\mathbf{Z}_{(1)} - \frac{p-r}{n-2}\sigma^2\mathbf{I}_{n-2}\Big\| \\
=&\Big\|\frac{1}{n-2}\sigma^2\mathbf{Z}_{(2)}^T\mathbf{Z}_{(2)} - \frac{p-r}{n-2}\sigma^2\mathbf{I}_{n-2}\Big\| \\
=&\frac{p-r}{n-2}\sigma^2\Big\|\frac{1}{p-r}\mathbf{Z}_{(2)}^T\mathbf{Z}_{(2)} - \mathbf{I}_{n-2}\Big\|.
\end{aligned}
$$

But Corollary 3 implies that

$$
\Big\|\frac{1}{p-r}\mathbf{Z}_{(2)}^T\mathbf{Z}_{(2)} - \mathbf{I}_{n-2}\Big\| = O_P\Big(\max\big(\sqrt{\frac{n-2}{p-r}}, \frac{n-2}{p-r}\big)\Big).
$$

Thus, for $i=1,\ldots,r$,

$$
\frac{1}{n-2}\lambda_i(\mathbf{Z}^T\mathbf{E}\mathbf{Z}) = \frac{1}{n-2}\lambda_i\big(\mathbf{Z}_{(1)}^T(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)\mathbf{Z}_{(1)}\big) + \frac{p-r}{n-2}\sigma^2 + O_P\Big(\max\big(\sqrt{\frac{p}{n}}, 1\big)\Big). \tag{17}
$$

Next we deal with $(n-2)^{-1}\lambda_i\big(\mathbf{Z}_{(1)}^T(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)\mathbf{Z}_{(1)}\big)$. For $i=1,\ldots,r$, we have

$$
\begin{aligned}
&\Big|\frac{1}{n-2}\lambda_i\big(\mathbf{Z}_{(1)}^T(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)\mathbf{Z}_{(1)}\big) - (\boldsymbol{\lambda}_i+\sigma^2)\Big| \\
=&\Big|\frac{1}{n-2}\lambda_i\big((\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)^{1/2}\mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^T(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)^{1/2}\big) - \lambda_i\big(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r\big)\Big| \\
\le&\Big\|\frac{1}{n-2}(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)^{1/2}\mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^T(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)^{1/2} - (\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)\Big\| \\
=&\Big\|(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)^{1/2}\big(\frac{1}{n-2}\mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^T - \mathbf{I}_r\big)(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)^{1/2}\Big\| \\
\le&(\boldsymbol{\lambda}_1+\sigma^2)\Big\|\frac{1}{n-2}\mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^T - \mathbf{I}_r\Big\| \\
=&O_P\Big(\frac{\boldsymbol{\lambda}_1}{\sqrt{n}}\Big),
\end{aligned}
$$

where the first inequality follows from Weyl's inequality and the last equality follows from Corollary 3. This, together with (17), leads to

$$
\frac{1}{n-2}\lambda_i(\mathbf{Z}^T\mathbf{E}\mathbf{Z}) = \boldsymbol{\lambda}_i + \frac{p+n-r-2}{n-2}\sigma^2 + O_P\Big(\max\big(\sqrt{\frac{p}{n}}, \frac{\boldsymbol{\lambda}_1}{\sqrt{n}}, 1\big)\Big), \quad i=1,\ldots,r.
$$

Then (15) follows from $\boldsymbol{\lambda}_1 \asymp p^\beta$ and $\beta \geq 1/2$.

Next we prove (16). We note that

$$\hat{\sigma}^2 = \frac{1}{p-r}\Big(\operatorname{tr}(\mathbf{S}) - \sum_{i=1}^r \lambda_i(\mathbf{S})\Big) \sim \frac{1}{(n-2)(p-r)}\Big(\operatorname{tr}(\mathbf{Z}^T\mathbf{E}\mathbf{Z}) - \sum_{i=1}^r \lambda_i(\mathbf{Z}^T\mathbf{E}\mathbf{Z})\Big).$$

For $\operatorname{tr}(\mathbf{Z}^T\mathbf{E}\mathbf{Z})$, we have $\operatorname{tr}\big(\mathbf{Z}^T\mathbf{E}\mathbf{Z}\big) = \operatorname{tr}\big(\mathbf{Z}_{(1)}^T(\boldsymbol{\Lambda}+\sigma^2\mathbf{I}_r)\mathbf{Z}_{(1)}\big) + \sigma^2\operatorname{tr}\big(\mathbf{Z}_{(2)}^T\mathbf{Z}_{(2)}\big)$. Since $\operatorname{tr}\big(\mathbf{Z}_{(2)}^T\mathbf{Z}_{(2)}\big) \sim \chi^2_{(n-2)(p-r)}$, by central limit theorem, we have

$$\operatorname{tr}\big(\mathbf{Z}_{(2)}^T\mathbf{Z}_{(2)}\big) = (n-2)(p-r) + O_P(\sqrt{np}).$$

It follows that

$$\operatorname{tr}\big(\mathbf{Z}^T\mathbf{E}\mathbf{Z}\big) = \operatorname{tr}\big(\mathbf{Z}_{(1)}^T(\boldsymbol{\Lambda}+\sigma^2\mathbf{I}_r)\mathbf{Z}_{(1)}\big) + (n-2)(p-r)\sigma^2 + O_P(\sqrt{np}). \tag{18}$$

In view of (17), we have

$$\sum_{i=1}^r \lambda_i(\mathbf{Z}^T\mathbf{E}\mathbf{Z}) = \operatorname{tr}\big(\mathbf{Z}_{(1)}^T(\boldsymbol{\Lambda}+\sigma^2\mathbf{I}_r)\mathbf{Z}_{(1)}\big) + r(p-r)\sigma^2 + O_P\Big(\max\big(\sqrt{np}, n\big)\Big).$$

This, together with (18), yields

$$\frac{1}{(n-2)(p-r)}\Big(\operatorname{tr}(\mathbf{Z}^T\mathbf{E}\mathbf{Z}) - \sum_{i=1}^r \lambda_i(\mathbf{Z}^T\mathbf{E}\mathbf{Z})\Big) = (1-\frac{r}{n-2})\sigma^2 + O_P\Big(\max\big(\frac{1}{\sqrt{np}}, \frac{1}{p}\big)\Big).$$

□

Lemma 6. *Under Assumption 1, we have*

$$\|\hat{\mathbf{V}}\hat{\mathbf{V}}^T - \mathbf{V}\mathbf{V}^T\|^2 = O_P(\frac{p}{p^\beta n}).$$

Remark 6. it has been proved in Cai et al. (2013), Theorem 5 that under certain conditions,

$$\operatorname{E}\|\hat{\mathbf{V}}\hat{\mathbf{V}}^T - \mathbf{V}\mathbf{V}^T\|_F^2 = O(\frac{p}{p^\beta n}),$$

where $\|\cdot\|_F$ is the Frobenius norm. Moreover, they proved that the convergence rate $p/(p^\beta n)$ is in fact minimax optimal. However, Cai et al. (2013), Theorem 5 needs the condition $\log p = O(n)$ which is unwanted. This condition is used to control the expectation and hence can be dropped in Lemma 6.

Proof. Since $\|\hat{\mathbf{V}}\hat{\mathbf{V}}^T - \mathbf{V}\mathbf{V}^T\|^2 \leq 1$, the conclusion is trivial when $p/(p^\beta n)$ is unbounded. So without loss of generality, we assume $p/(p^\beta n) = O(1)$. Define $\mathbf{U}$, $\mathbf{E}$, $\mathbf{Z}$, $\mathbf{Z}_{(1)}$ and $\mathbf{Z}_{(2)}$ as in the proof of Lemma 5. Without loss of generality, we assume that $\mathbf{S} = (n-2)^{-1}\mathbf{U}\mathbf{E}^{1/2}\mathbf{Z}\mathbf{Z}^T\mathbf{E}^{1/2}\mathbf{U}^T$. Similar to the proof of Cai et al. (2013), Theorem 5, we define

$$\mathbf{S}_0 = \frac{1}{n-2}\mathbf{V}(\mathbf{\Lambda} + \sigma^2\mathbf{I}_r)^{1/2}\mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^T(\mathbf{\Lambda} + \sigma^2\mathbf{I}_r)^{1/2}\mathbf{V}^T + \sigma^2\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T.$$

It can be seen that the set of eigenvalues of $\mathbf{S}_0$ is the union of the nonzero eigenvalues of the matrix $(n-2)^{-1}\mathbf{V}(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)^{1/2}\mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^T(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)^{1/2}\mathbf{V}^T$ and $\sigma^2\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T$. All the nonzero eigenvalues of $\sigma^2\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T$ are σ^2. We note that with probability 1, the first matrix is of rank r. Define the event

$$A = \Big\{\frac{1}{n-2}\lambda_r\big(\mathbf{V}(\mathbf{\Lambda} + \sigma^2\mathbf{I}_r)^{1/2}\mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^T(\mathbf{\Lambda} + \sigma^2\mathbf{I}_r)^{1/2}\mathbf{V}^T\big) > \sigma^2\Big\}.$$

On the event A, the eigenspace of $\mathbf{S}_0$ associated with the r leading eigenvalues is exactly $\mathbf{V}\mathbf{V}^T$, although the individual columns of $\mathbf{V}$ need not be the leading eigenvectors of $\mathbf{S}_0$. Applying Lemma 3 to $\mathbf{S}$ and $\mathbf{S}_0$ yields

$$\|\hat{\mathbf{V}}\hat{\mathbf{V}}^T - \mathbf{V}\mathbf{V}^T\|\mathbf{1}_{\{A\}} \leq \frac{2}{\lambda_r(\mathbf{S}_0) - \lambda_{r+1}(\mathbf{S}_0)}\|\mathbf{S} - \mathbf{S}_0\|\mathbf{1}_{\{A\}}. \tag{19}$$

Note that

$$\lambda_r\Big(\frac{1}{n-2}\mathbf{V}(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)^{1/2}\mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^T(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)^{1/2}\mathbf{V}^T\Big) = \frac{1}{n-2}\lambda_r\big((\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)^{1/2}\mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^T(\mathbf{\Lambda}+\sigma^2\mathbf{I}_r)^{1/2}\big).$$

By Weyl's inequality,

$$\begin{aligned}
&\Big|\frac{1}{n-2}\lambda_r((\mathbf{\Lambda} + \sigma^2\mathbf{I}_r)^{1/2}\mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^T(\mathbf{\Lambda} + \sigma^2\mathbf{I}_r)^{1/2}) - (\boldsymbol{\lambda}_r + \sigma^2)\Big| \\
=&\Big|\frac{1}{n-2}\lambda_r((\mathbf{\Lambda} + \sigma^2\mathbf{I}_r)^{1/2}\mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^T(\mathbf{\Lambda} + \sigma^2\mathbf{I}_r)^{1/2}) - \lambda_r(\mathbf{\Lambda} + \sigma^2\mathbf{I}_r)\Big| \\
\leq&\Big\|(\mathbf{\Lambda} + \sigma^2\mathbf{I}_r)^{1/2}\big(\frac{1}{n-2}\mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^T - \mathbf{I}\big)(\mathbf{\Lambda} + \sigma^2\mathbf{I}_r)^{1/2}\Big\| \\
\leq&(\boldsymbol{\lambda}_1 + \sigma^2)\Big\|\frac{1}{n-2}\mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^T - \mathbf{I}\Big\| \\
=&O_P\Big(\frac{\boldsymbol{\lambda}_1 + \sigma^2}{\sqrt{n}}\Big),
\end{aligned}$$

where the last equality follows from Corollary 3. Hence

$$\frac{1}{n-2}\lambda_r(\mathbf{V}(\mathbf{\Lambda} + \sigma^2\mathbf{I}_r)^{1/2}\mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^T(\mathbf{\Lambda} + \sigma^2\mathbf{I}_r)^{1/2}\mathbf{V}^T) = (\boldsymbol{\lambda}_1 + \sigma^2)\Big(1 + O_P\big(\frac{1}{\sqrt{n}}\big)\Big).$$

It follows that $P(A) \to 1$. In view of (19), it is sufficient to show that

$$\frac{2}{\lambda_r(\mathbf{S}_0) - \lambda_{r+1}(\mathbf{S}_0)} \|\mathbf{S} - \mathbf{S}_0\| \mathbf{1}_{\{A\}} = O_P(\sqrt{\frac{p}{p^\beta n}}).$$

Note that on event A, $\lambda_{r+1}(\mathbf{S}_0) = \sigma^2$ and

$$\lambda_r(\mathbf{S}_0) = \lambda_r((n-2)^{-1} \lambda_r(\mathbf{V}(\boldsymbol{\Lambda} + \sigma^2 \mathbf{I}_r)^{1/2} \mathbf{Z}_{(1)} \mathbf{Z}_{(1)}^T (\boldsymbol{\Lambda} + \sigma^2 \mathbf{I}_r)^{1/2} \mathbf{V}^T)).$$

Hence

$$\begin{aligned}
&\frac{2}{\lambda_r(\mathbf{S}_0) - \lambda_{r+1}(\mathbf{S}_0)} \mathbf{1}_{\{A\}} = \frac{2}{\lambda_r(\frac{1}{n-2} \lambda_r(\mathbf{V}(\boldsymbol{\Lambda} + \sigma^2 \mathbf{I}_r)^{1/2} \mathbf{Z}_{(1)} \mathbf{Z}_{(1)}^T (\boldsymbol{\Lambda} + \sigma^2 \mathbf{I}_r)^{1/2} \mathbf{V}^T)) - \sigma^2} \mathbf{1}_{\{A\}} \\
=& \frac{2}{\boldsymbol{\lambda}_r \Big(1 + O_P\big(\frac{1}{\sqrt{n}}\big)\Big)} \mathbf{1}_{\{A\}} = \frac{2}{\boldsymbol{\lambda}_r} (1 + o_P(1)).
\end{aligned}$$

Next we bound $\|\mathbf{S} - \mathbf{S}_0\|$. We have

$$\begin{aligned}
&\|\mathbf{S} - \mathbf{S}_0\| = \|(\mathbf{V}\mathbf{V}^T + \tilde{\mathbf{V}}\tilde{\mathbf{V}}^T)(\mathbf{S} - \mathbf{S}_0)(\mathbf{V}\mathbf{V}^T + \tilde{\mathbf{V}}\tilde{\mathbf{V}}^T)\| \\
\leq& \|\mathbf{V}\mathbf{V}^T(\mathbf{S} - \mathbf{S}_0)\mathbf{V}\mathbf{V}^T\| + 2\|\mathbf{V}\mathbf{V}^T(\mathbf{S} - \mathbf{S}_0)\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T\| + \|\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T(\mathbf{S} - \mathbf{S}_0)\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T\| \\
\leq& \|\mathbf{V}^T(\mathbf{S} - \mathbf{S}_0)\mathbf{V}\| + 2\|\mathbf{V}^T(\mathbf{S} - \mathbf{S}_0)\tilde{\mathbf{V}}\| + \|\tilde{\mathbf{V}}^T(\mathbf{S} - \mathbf{S}_0)\tilde{\mathbf{V}}\| \\
=& 2\Big\|\frac{\sigma}{n-2}(\boldsymbol{\Lambda} + \sigma^2 \mathbf{I}_r)^{1/2} \mathbf{Z}_{(1)} \mathbf{Z}_{(2)}^T\Big\| + \sigma^2 \Big\|\frac{1}{n-2} \mathbf{Z}_{(2)} \mathbf{Z}_{(2)}^T - \mathbf{I}_{p-r}\Big\| \\
\leq& \frac{2\sqrt{(\lambda_1 + \sigma^2)\sigma^2}}{n-2} \|\mathbf{Z}_{(1)} \mathbf{Z}_{(2)}^T\| + \sigma^2 \Big\|\frac{1}{n-2} \mathbf{Z}_{(2)} \mathbf{Z}_{(2)}^T - \mathbf{I}_{p-r}\Big\|.
\end{aligned}$$

By Corollary 3, we have $\left\|(n-2)^{-1} \mathbf{Z}_{(2)} \mathbf{Z}_{(2)}^T - \mathbf{I}_{p-r}\right\| = O_p(\max(\sqrt{p/n}, p/n))$. By the independence of $\mathbf{Z}_{(1)}$ and $\mathbf{Z}_{(2)}$, we have

$$\begin{aligned}
&\mathrm{E}\,\|\mathbf{Z}_{(1)} \mathbf{Z}_{(2)}^T\|^2 \leq \mathrm{E}\,\|\mathbf{Z}_{(1)} \mathbf{Z}_{(2)}^T\|_F^2 = \mathrm{E}\left[\,\mathrm{tr}\left(\mathbf{Z}_{(1)} \mathbf{Z}_{(2)}^T \mathbf{Z}_{(2)} \mathbf{Z}_{(1)}^T\right)\right] \\
=& \mathrm{E}\,\mathrm{E}\left[\,\mathrm{tr}\left(\mathbf{Z}_{(1)} \mathbf{Z}_{(2)}^T \mathbf{Z}_{(2)} \mathbf{Z}_{(1)}^T\right) \big| \mathbf{Z}_{(1)}\right] = (p - r)\,\mathrm{E}\left[\,\mathrm{tr}\left(\mathbf{Z}_{(1)} \mathbf{Z}_{(1)}^T\right)\right] = r(n-2)(p-r).
\end{aligned}$$

Hence $\|\mathbf{Z}_{(1)} \mathbf{Z}_{(2)}^T\| = O_P(\sqrt{np})$. Combining these bounds leads to

$$\|\mathbf{S} - \mathbf{S}_0\| = O_P(\sqrt{\frac{\lambda_1 p}{n}}) + O_P\Big(\max\big(\sqrt{\frac{p}{n}}, \frac{p}{n}\big)\Big) = O_P(\sqrt{\frac{p^\beta p}{n}}) + O_P(\frac{p}{n}).$$

Thus,

$$\frac{2}{\lambda_r(\mathbf{S}_0) - \lambda_{r+1}(\mathbf{S}_0)} \|\mathbf{S} - \mathbf{S}_0\| = O_P(\sqrt{\frac{p}{p^\beta n}}) + O_P(\frac{p}{p^\beta n}) = O_P(\sqrt{\frac{p}{p^\beta n}}),$$

where the last equality holds since we have assumed $p/(p^\beta n) = O(1)$. This completes the proof.

$\square$

Proof of Proposition 1. Note that

$$\begin{aligned}
&\|\tilde{\mathbf{V}}^T(\bar{X}_1-\bar{X}_2)\|^2 = \left\|\tilde{\mathbf{V}}^T(\mu_1-\mu_2)+\tilde{\mathbf{V}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\right\|^2\\
=&\|\tilde{\mathbf{V}}^T(\mu_1-\mu_2)\|^2+\left\|\tilde{\mathbf{V}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\right\|^2+2(\mu_1-\mu_2)^T\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\\
=&\|\tilde{\mathbf{V}}^T(\mu_1-\mu_2)\|^2+\left\|\tilde{\mathbf{V}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\right\|^2+o_P(\frac{\sqrt{p}}{n}).
\end{aligned} \tag{20}$$

The last equality holds since

$$\begin{aligned}
&2(\mu_1-\mu_2)^T\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\sim N(0,4\sigma^2\tau\|\tilde{\mathbf{V}}^T(\mu_1-\mu_2)\|^2)\\
=&O_P(\sqrt{\tau}\|\tilde{\mathbf{V}}^T(\mu_1-\mu_2)\|)=o_P(\frac{\sqrt{p}}{n}).
\end{aligned}$$

For $k=1,2$, we have

$$\frac{1}{n_k}\operatorname{tr}(\tilde{\mathbf{V}}^T\mathbf{S}_k\tilde{\mathbf{V}})\sim\frac{\sigma^2}{n_k(n_k-1)}\chi^2_{(p-r)(n_k-1)}=\sigma^2\frac{p-r}{n_k}\Big(1+O_P(\frac{1}{\sqrt{(p-r)(n_k-1)}})\Big),$$

where the last equality comes from central limit theorem. Thus,

$$\frac{1}{n_1}\operatorname{tr}(\tilde{\mathbf{V}}^T\mathbf{S}_1\tilde{\mathbf{V}})+\frac{1}{n_2}\operatorname{tr}(\tilde{\mathbf{V}}^T\mathbf{S}_2\tilde{\mathbf{V}})=\sigma^2\tau(p-r)+o_P(\frac{\sqrt{p}}{n}). \tag{21}$$

It follows from (20) and (21) that

$$\frac{T_1-\|\tilde{\mathbf{V}}^T(\mu_1-\mu_2)\|^2}{\sigma^2\sqrt{2\tau^2p}}=\frac{\|\tilde{\mathbf{V}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\|^2-\sigma^2\tau(p-r)}{\sigma^2\sqrt{2\tau^2p}}+o_P(1).$$

Then the proposition follows from $\|\tilde{\mathbf{V}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\|^2\sim\sigma^2\tau\chi^2_{p-r}$. □

Proof of Proposition 2. The conclusion (8) is a direct corollary of (15) in Lemma 5. By (16) and (8), for $i=1,\ldots,r$, we have

$$\begin{aligned}
\hat{\boldsymbol{\lambda}}_i=&\lambda_i(\mathbf{S})-\frac{p+n-r-2}{n-2}\hat{\sigma}^2_*\\
=&\boldsymbol{\lambda}_i+\frac{p+n-r-2}{n-2}\sigma^2+O_P\Big(\max\big(\frac{p^\beta}{\sqrt{n}},1\big)\Big)-\frac{p+n-r-2}{n-2}\sigma^2-O_P\Big(\frac{n+p}{n}\max\big(\frac{1}{\sqrt{np}},\frac{1}{p}\big)\Big)\\
=&\boldsymbol{\lambda}_i+O_P\Big(\max\big(\frac{p^\beta}{\sqrt{n}},1\big)\Big),
\end{aligned}$$

which proves (9). □

Proof of Theorem 3. Proposition 2 implies that $\hat{\sigma}^2_*\xrightarrow{P}\sigma^2$ and $\hat{\boldsymbol{\lambda}}_i/\boldsymbol{\lambda}_i\xrightarrow{P}1$. Hence

$$\tau\sum_{i=1}^r\frac{p\hat{\sigma}^2_*}{n\hat{\boldsymbol{\lambda}}_i+(n+p)\hat{\sigma}^2_*}\hat{\boldsymbol{\lambda}}_i=\tau\frac{p}{n}\sum_{i=1}^r\frac{\hat{\sigma}^2_*}{1+\frac{(n+p)}{n\hat{\boldsymbol{\lambda}}_i}\hat{\sigma}^2_*}=\tau\frac{p}{n}\sum_{i=1}^r\frac{\sigma^2(1+o_P(1))}{1+\frac{(n+p)}{n\boldsymbol{\lambda}_i}\sigma^2(1+o_P(1))}.$$

Since $p/n^2 \to 0$ implies $p/(n\boldsymbol{\lambda}_i) \to 0$, we have

$$\tau \sum_{i=1}^{r} \frac{p\hat{\sigma}_*^2}{n\hat{\boldsymbol{\lambda}}_i + (n+p)\hat{\sigma}_*^2} \hat{\boldsymbol{\lambda}}_i = r\tau \frac{p}{n}\sigma^2(1 + o_P(1)) = o_P(\sqrt{\tau^2 p}).$$

In view of (8), we have

$$\begin{aligned}
&\tau(p-r)\hat{\sigma}_*^2 = \tau(p-r)\sigma^2 + O_P\Big(\frac{p}{n}\max\big(\frac{1}{\sqrt{np}}, \frac{1}{p}\big)\Big) \\
=&\tau(p-r)\sigma^2 + O_P\Big(\frac{\sqrt{p}}{n}\max\big(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{p}}\big)\Big) = \tau(p-r)\sigma^2 + o_P(\sqrt{\tau^2 p}).
\end{aligned}$$

Thus,

$$\frac{T_2 - \|\tilde{\mathbf{V}}^T(\mu_1-\mu_2)\|^2}{\sigma^2\sqrt{2\tau^2 p}} = \frac{\|\hat{\tilde{\mathbf{V}}}^T(\bar{X}_1 - \bar{X}_2)\|^2 - \|\tilde{\mathbf{V}}^T(\mu_1-\mu_2)\|^2 - \tau(p-r)\sigma^2}{\sigma^2\sqrt{2\tau^2 p}} + o_P(1),$$

We write

$$\begin{aligned}
&\frac{\|\hat{\tilde{\mathbf{V}}}^T(\bar{X}_1 - \bar{X}_2)\|^2 - \|\tilde{\mathbf{V}}^T(\mu_1-\mu_2)\|^2 - \tau(p-r)\sigma^2}{\sigma^2\sqrt{2\tau^2 p}} \\
=&\frac{1}{\sigma^2\sqrt{2\tau^2 p}}(P_1 + P_2 + P_3),
\end{aligned}$$

where

$$\begin{aligned}
P_1 &= \|\hat{\tilde{\mathbf{V}}}^T\big((\bar{X}_1 - \mu_1) - (\bar{X}_2 - \mu_2)\big)\|^2 - \tau(p-r)\sigma^2, \\
P_2 &= 2(\mu_1-\mu_2)^T\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T\big((\bar{X}_1 - \mu_1) - (\bar{X}_2 - \mu_2)\big), \\
P_3 &= \|\hat{\tilde{\mathbf{V}}}^T(\mu_1-\mu_2)\|^2 - \|\tilde{\mathbf{V}}^T(\mu_1-\mu_2)\|^2.
\end{aligned}$$

To prove the theorem, it suffices to show that

$$\frac{P_1}{\sigma^2\sqrt{2\tau^2 p}} \xrightarrow{\mathcal{L}} N(0,1), \quad \frac{P_2}{\sigma^2\sqrt{2\tau^2 p}} \xrightarrow{P} 0 \quad \text{and} \quad \frac{P_3}{\sigma^2\sqrt{2\tau^2 p}} \xrightarrow{P} 0.$$

First we deal with P_2. Let ϵ be any fixed positive number. We have

$$\Pr\Big(\frac{P_2}{\sigma^2\sqrt{2\tau^2 p}} > \epsilon\Big) = \mathrm{E}\big[\Pr\big(P_2 > \epsilon\sigma^2\sqrt{2\tau^2 p}\big|\mathbf{S}\big)\big].$$

Since the conditional probability $\Pr\big(P_2 > \epsilon\sigma^2\sqrt{2\tau^2 p}\big|\mathbf{S}\big)$ is bounded, by dominated convergence theorem, we only need to prove $\Pr\big(P_2 > \epsilon\sigma^2\sqrt{2\tau^2 p}\big|\mathbf{S}\big) \xrightarrow{P} 0$.

Note that $\bar{X}_1$, $\bar{X}_2$, and $\mathbf{S}$ are mutually independent and $\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T$ only depends on $\mathbf{S}$. We have

$$\begin{aligned}
&\Pr\left(P_2 > \epsilon\sigma^2\sqrt{2\tau^2 p}\big|\mathbf{S}\right) \le \frac{1}{2\epsilon^2\sigma^4\tau^2 p}\,\mathrm{E}\left(P_2^2\big|\mathbf{S}\right)\\
=&\frac{1}{2\epsilon^2\sigma^4\tau^2 p}4\tau(\mu_1-\mu_2)^T\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T(\mu_1-\mu_2)\\
\le&\frac{2}{\epsilon^2\sigma^4\tau p}\lambda_1(\hat{\tilde{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\tilde{\mathbf{V}}})(\mu_1-\mu_2)^T\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T(\mu_1-\mu_2)\\
\le&\frac{2}{\epsilon^2\sigma^4\tau p}\|\mu_1-\mu_2\|^2\lambda_1(\hat{\tilde{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\tilde{\mathbf{V}}})\\
=&O(\frac{1}{\sqrt{p}})\lambda_1(\hat{\tilde{\mathbf{V}}}^T(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T+\sigma^2\mathbf{I}_p)\hat{\tilde{\mathbf{V}}})\\
\le&O(\frac{1}{\sqrt{p}})\big(\kappa p^\beta\lambda_1(\hat{\tilde{\mathbf{V}}}^T\mathbf{V}\mathbf{V}^T\hat{\tilde{\mathbf{V}}})+\sigma^2\big).
\end{aligned}$$

But

$$\lambda_1(\hat{\tilde{\mathbf{V}}}^T\mathbf{V}\mathbf{V}^T\hat{\tilde{\mathbf{V}}}) = \|\mathbf{V}^T\hat{\tilde{\mathbf{V}}}\|^2 = \|\mathbf{V}\mathbf{V}^T - \hat{\mathbf{V}}\hat{\mathbf{V}}^T\|^2 = O_P\Big(\frac{p}{p^\beta n}\Big),$$

where the second equality follows from Golub and Van Loan (2013), Theorem 2.5.1 and the last equality follows from Lemma 6. Thus,

$$\Pr\left(P_2 > \epsilon\sigma^2\sqrt{2\tau^2 p}\big|\mathbf{S}\right) = O(\frac{1}{\sqrt{p}})\big(O_P(\frac{p}{n})+\sigma^2\big) = O(1)(O_P(\frac{\sqrt{p}}{n})+\frac{\sigma^2}{\sqrt{p}}) = o_P(1).$$

Next we deal with P_3. Note that

$$\begin{aligned}
&|P_3| = \big|(\mu_1-\mu_2)^T(\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T - \tilde{\mathbf{V}}\tilde{\mathbf{V}}^T)(\mu_1-\mu_2)\big| \le \|\mu_1-\mu_2\|^2\|\hat{\tilde{\mathbf{V}}}\hat{\tilde{\mathbf{V}}}^T - \tilde{\mathbf{V}}\tilde{\mathbf{V}}^T\|\\
=&\|\mu_1-\mu_2\|^2\|\hat{\mathbf{V}}\hat{\mathbf{V}}^T - \mathbf{V}\mathbf{V}^T\| = O(\frac{\sqrt{p}}{n})\sqrt{O_P(\frac{p}{p^\beta n})} = o_P(\frac{\sqrt{p}}{n}).
\end{aligned}$$

Hence

$$\frac{P_3}{\sigma^2\sqrt{2\tau^2 p}} = O(\frac{n}{\sqrt{p}})P_3 = o_P(1).$$

Now we prove the asymptotic normality of P_1. To make clear the mode of convergence, we need a metric for weak convergence. For two distribution function F and G, the Levy metric ρ of F and G is defined as

$$\rho(F,G) = \inf\{\epsilon : F(x-\epsilon)-\epsilon \le G(x) \le F(x+\epsilon)+\epsilon \quad \text{for all } x\}.$$

It's well known that $\rho(F_n, F) \to 0$ if and only if $F_n \xrightarrow{\mathcal{L}} F$.

Since the conditional distribution of $\hat{\hat{\mathbf{V}}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)$ given $\mathbf{S}$ is $N(0,\tau\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}})$, we have that

$$\tau^{-1}\big\|\hat{\hat{\mathbf{V}}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\big\|^2\sim\sum_{i=1}^{p-r}\lambda_i(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}})\xi_i^2, \tag{22}$$

where $\xi_1,\ldots,\xi_{p-r}$ are i.i.d. standard normal random variables which are independent of $\hat{\hat{\mathbf{V}}}$. In view of Lemma 1, the asymptotic distribution of P_1 relies on the asymptotic behavior of $\lambda_i(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}})$, $i=1,\ldots,p-r$. As we have shown,

$$\lambda_1(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}})\le\kappa p^\beta\lambda_1(\hat{\hat{\mathbf{V}}}^T\mathbf{V}\mathbf{V}^T\hat{\hat{\mathbf{V}}})+\sigma^2=\kappa p^\beta\|\mathbf{V}\mathbf{V}^T-\hat{\mathbf{V}}\hat{\mathbf{V}}^T\|^2+\sigma^2. \tag{23}$$

On the other hand, for $i=r+1,\ldots,p-r$, we have

$$\lambda_i(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}})=\lambda_i(\hat{\hat{\mathbf{V}}}^T\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\hat{\hat{\mathbf{V}}})+\sigma^2=\sigma^2, \tag{24}$$

where the last equality follows from $\operatorname{Rank}(\hat{\hat{\mathbf{V}}}^T\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\hat{\hat{\mathbf{V}}})\le\operatorname{Rank}(\mathbf{V})=r$. It follows from (23) and (24) that

$$\begin{aligned}\operatorname{tr}(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}})^2=&O_P\Big(\big(p^\beta\|\mathbf{V}\mathbf{V}^T-\hat{\mathbf{V}}\hat{\mathbf{V}}^T\|^2+1\big)^2\Big)+(p-2r)\sigma^4\\=&\big(\frac{p}{n}+1\big)^2O_P(1)+(p-2r)\sigma^4=p\sigma^4(1+o_P(1)).\end{aligned} \tag{25}$$

This, combined with (23), yields

$$\frac{\lambda_1^2(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}})}{\operatorname{tr}(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}})^2}=O_P\Big(p^{-1}\big(p^\beta\|\mathbf{V}\mathbf{V}^T-\hat{\mathbf{V}}\hat{\mathbf{V}}^T\|^2+1\big)^2\Big)=O_P\Big(\frac{(p/n+1)^2}{p}\Big)=o_P(1).$$

Then for every subsequence $\{n(k)\}$ of $\{n\}$, there's a further subsequence $\big\{n\big(k(l)\big)\big\}$ along which

$$\frac{\lambda_1^2(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}})}{\operatorname{tr}(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}})^2}\xrightarrow{a.s.}0.$$

This fact, together with (22) and Lemma 1, implies that along $\big\{n\big(k(l)\big)\big\}$ we have

$$\rho\big(\mathcal{L}(Y_n|\,\mathbf{S}),N(0,1)\big)\xrightarrow{a.s.}0, \tag{26}$$

where

$$Y_n=\frac{\|\hat{\hat{\mathbf{V}}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\|^2-\tau\operatorname{tr}(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}})}{\sqrt{2\tau^2\operatorname{tr}(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}})^2}},$$

and $\mathcal{L}(Y_n|\,\mathbf{S})$ is the conditional distribution of Y_n given $\mathbf{S}$. By the definition of weak convergence, (26) implies that for every continuous bounded function $f(\cdot)$, $\mathrm{E}[f(Y_n)|\,\mathbf{S}] \xrightarrow{a.s.} \mathrm{E}[f(\xi^*)]$ along $\{n\big(k(l)\big)\}$, where ξ^* is a standard normal random variable. By dominated convergence theorem, $\mathrm{E}[f(Y_n)] \to \mathrm{E}[f(\xi^*)]$ along $\{n\big(k(l)\big)\}$. This implies that $Y_n \xrightarrow{\mathcal{L}} N(0,1)$ along $\{n\big(k(l)\big)\}$. Thus, for every subsequence of $\{n\}$, there is a further subsequence along which $Y_n \xrightarrow{\mathcal{L}} N(0,1)$. This means $Y_n \xrightarrow{\mathcal{L}} N(0,1)$, or

$$\frac{\|\hat{\hat{\mathbf{V}}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\|^2-\tau\mathrm{tr}(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}})}{\sqrt{2\tau^2\mathrm{tr}(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}})^2}} \xrightarrow{\mathcal{L}} N(0,1).$$

By (23) and (24), we have

$$\begin{aligned}\mathrm{tr}(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}}) =& \sum_{i=1}^{r}\lambda_i(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}}) + \sum_{i=r+1}^{p-r}\lambda_i(\hat{\hat{\mathbf{V}}}^T\mathbf{\Sigma}\hat{\hat{\mathbf{V}}}) \\ =&O_P(\frac{p}{n}+1)+(p-2r)\sigma^2=(p-r)\sigma^2+o_P(\sqrt{p}).\end{aligned} \tag{27}$$

It follows from (25), (27) and Slutsky's theorem that

$$\frac{\|\hat{\hat{\mathbf{V}}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\|^2-\sigma^2\tau(p-r)}{\sigma^2\sqrt{2\tau^2 p}} \xrightarrow{\mathcal{L}} N(0,1).$$

This completes the proof. □

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 11471035, 11471030.

References

Ahn, S.C., Horenstein, A.R., 2013. Eigenvalue ratio test for the number of factors. Econometrica 81, 1203–1227. doi:10.3982/ecta8968.

Anderson, T.W., 1963. Asymptotic theory for principal component analysis. The Annals of Mathematical Statistics 34, 122–148.

Bai, Z., Saranadasa, H., 1996. Effect of high dimension: by an example of a two sample problem. Statistica Sinica 6, 311–329.

Birnbaum, A., Johnstone, I.M., Nadler, B., Paul, D., 2013. Minimax bounds for sparse PCA with noisy high-dimensional data. The Annals of Statistics 41, 1055–1084. doi:10.1214/12-aos1014.

Cai, T., Ma, Z., Wu, Y., 2015. Optimal estimation and rank detection for sparse spiked covariance matrices. Probability Theory & Related Fields 161, 781–815.

Cai, T.T., Ma, Z., Wu, Y., 2013. Sparse PCA: Optimal rates and adaptive estimation. The Annals of Statistics 41, 3074–3110. doi:10.1214/13-aos1178.

Chen, L.S., Paul, D., Prentice, R.L., Wang, P., 2011. A regularized hotelling'sT2test for pathway analysis in proteomic studies. Journal of the American Statistical Association 106, 1345–1360. doi:10.1198/jasa.2011.ap10599.

Chen, S.X., Qin, Y.L., 2010. A two-sample test for high-dimensional data with applications to gene-set testing. The Annals of Statistics 38, 808–835. doi:10.1214/09-aos716.

Davidson, K.R., Szarek, S.J., 2001. Handbook of the Geometry of Banach Spaces. volume 1. North-Holland, Amsterdam. URL: http://www.sciencedirect.com/science/article/pii/S1874584901800103, doi:http://dx.doi.org/10.1016/S1874-5849(01)80010-3. handbook of the Geometry of Banach Spaces.

Golub, G.H., Van Loan, C.F., 2013. Matrix Computations. Fourth ed., The Johns Hopkins University Press.

Horn, R.A., Johnson, C.R., 2012. Matrix Analysis. 2nd ed., Cambridge University Press, New York.

Jung, S., Marron, J.S., 2009. PCA consistency in high dimension, low sample size context. The Annals of Statistics 37, 4104–4130. doi:10.1214/09-aos709.

Lopes, M., Jacob, L., Wainwright, M.J., 2011. A more powerful two-sample test in high dimensions using random projection, in: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 24. Curran Associates, Inc., pp. 1206–1214.

Ma, Y., Lan, W., Wang, H., 2015. A high dimensional two-sample test under a low dimensional factor structure. Journal of Multivariate Analysis 140, 162–170. doi:10.1016/j.jmva.2015.05.005.

Passemier, D., Li, Z., Yao, J., 2017. On estimation of the noise variance in high dimensional probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79, 51–67. doi:10.1111/rssb.12153.

Srivastava, M.S., Du, M., 2008. A test for the mean vector with fewer observations than the dimension. Journal of Multivariate Analysis 99, 386–402. doi:10.1016/j.jmva.2006.11.002.

Srivastava, R., Li, P., Ruppert, D., 2016. RAPTT: An exact two-sample test in high dimensions using random projections. Journal of Computational and Graphical Statistics 25, 954–970. doi:10.1080/10618600.2015.1062771.

Thulin, M., 2014. A high-dimensional two-sample test for the mean using random subspaces. Computational Statistics & Data Analysis 74, 26–38. doi:10.1016/j.csda.2013.12.003.

Wang, W., Fan, J., 2017. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. The Annals of Statistics 45, 1342–1374. URL: https://doi.org/10.1214/16-AOS1487.

Zhao, J., Xu, X., 2016. A generalized likelihood ratio test for normal mean when p is greater than n. Computational Statistics & Data Analysis 99, 91–104. doi:10.1016/j.csda.2016.01.006.

Table 1: $n_1 = n_2 = 50$

	Normal			Chi-squared			Student's t		
p	200	500	800	200	500	800	200	500	800
$\beta = 0.5$									
New1	0.1630	0.2535	0.3285	0.1715	0.2705	0.3590	0.1845	0.2920	0.3435
New2	0.0800	0.0940	0.1080	0.0815	0.0905	0.1270	0.0840	0.1065	0.1085
New3	0.0635	0.0565	0.0580	0.0610	0.0520	0.0695	0.0595	0.0640	0.0595
oracle	0.0510	0.0470	0.0520	0.0480	0.0410	0.0560	0.0520	0.0525	0.0520
chi	0.0565	0.0535	0.0415	0.0520	0.0515	0.0480	0.0560	0.0420	0.0445
fast	0.0525	0.0485	0.0290	0.0485	0.0395	0.0395	0.0560	0.0365	0.0325
CQ	0.0630	0.0645	0.0575	0.0675	0.0665	0.0750	0.0630	0.0725	0.0610
SD	0.0515	0.0500	0.0440	0.0525	0.0465	0.0580	0.0440	0.0515	0.0405
ljw	0.0535	0.0480	0.0455	0.0510	0.0575	0.0440	0.0440	0.0500	0.0505
$\beta = 1$									
New1	0.1700	0.2725	0.3420	0.1680	0.2730	0.3570	0.1645	0.2725	0.3450
New2	0.0815	0.1060	0.1335	0.0900	0.1105	0.1270	0.0780	0.1080	0.1230
New3	0.0560	0.0585	0.0620	0.0625	0.0600	0.0505	0.0505	0.0585	0.0510
oracle	0.0545	0.0560	0.0565	0.0605	0.0535	0.0485	0.0450	0.0535	0.0385
chi	0.0480	0.0445	0.0570	0.0490	0.0525	0.0525	0.0540	0.0400	0.0425
fast	0.0505	0.0470	0.0590	0.0500	0.0550	0.0530	0.0560	0.0420	0.0435
CQ	0.0710	0.0785	0.0720	0.0725	0.0775	0.0730	0.0770	0.0715	0.0675
SD	0.0200	0.0165	0.0130	0.0225	0.0130	0.0100	0.0215	0.0075	0.0100
ljw	0.0470	0.0520	0.0460	0.0565	0.0470	0.0535	0.0485	0.0440	0.0540
$\beta = 2$									
New1	0.1685	0.2605	0.3290	0.1750	0.2620	0.3550	0.1715	0.2620	0.3600
New2	0.0860	0.1010	0.1215	0.0865	0.1010	0.1225	0.0795	0.1080	0.1245
New3	0.0545	0.0535	0.0470	0.0550	0.0545	0.0585	0.0505	0.0585	0.0540
oracle	0.0535	0.0495	0.0450	0.0520	0.0545	0.0510	0.0505	0.0440	0.0450
chi	0.0510	0.0505	0.0495	0.0480	0.0540	0.0530	0.0485	0.0410	0.0475
fast	0.0540	0.0530	0.0510	0.0500	0.0565	0.0555	0.0520	0.0445	0.0495
CQ	0.0690	0.0760	0.0675	0.0745	0.0665	0.0630	0.0785	0.0745	0.0700
SD	0.0060	0.0010	0.0000	0.0045	0.0010	0.0010	0.0055	0.0010	0.0010
ljw	0.0435	0.0500	0.0535	0.0410	0.0450	0.0430	0.0530	0.0500	0.0515

Table 2: $n_1 = n_2 = 100$

	Normal			Chi-squared			Student's t		
p	200	500	800	200	500	800	200	500	800
$\beta = 0.5$									
New1	0.0960	0.1360	0.1560	0.1075	0.1255	0.1640	0.1135	0.1410	0.1660
New2	0.0600	0.0670	0.0770	0.0685	0.0705	0.0790	0.0660	0.0710	0.0760
New3	0.0515	0.0470	0.0550	0.0595	0.0535	0.0535	0.0545	0.0565	0.0515
oracle	0.0475	0.0455	0.0525	0.0555	0.0495	0.0470	0.0535	0.0480	0.0490
chi	0.0480	0.0475	0.0475	0.0600	0.0510	0.0380	0.0430	0.0430	0.0450
fast	0.0510	0.0505	0.0475	0.0660	0.0555	0.0390	0.0480	0.0465	0.0460
CQ	0.0680	0.0610	0.0640	0.0650	0.0575	0.0630	0.0585	0.0705	0.0510
SD	0.0580	0.0520	0.0535	0.0560	0.0500	0.0485	0.0410	0.0615	0.0445
ljw	0.0440	0.0440	0.0460	0.0495	0.0470	0.0475	0.0440	0.0450	0.0505
$\beta = 1$									
New1	0.0975	0.1410	0.1600	0.0885	0.1320	0.1690	0.0895	0.1220	0.1625
New2	0.0645	0.0690	0.0800	0.0545	0.0735	0.0770	0.0585	0.0655	0.0795
New3	0.0530	0.0500	0.0475	0.0445	0.0545	0.0510	0.0455	0.0475	0.0500
oracle	0.0560	0.0495	0.0415	0.0425	0.0565	0.0455	0.0465	0.0465	0.0490
chi	0.0545	0.0480	0.0530	0.0500	0.0370	0.0400	0.0460	0.0405	0.0530
fast	0.0550	0.0500	0.0550	0.0525	0.0415	0.0450	0.0465	0.0450	0.0555
CQ	0.0715	0.0665	0.0880	0.0655	0.0680	0.0730	0.0785	0.0655	0.0740
SD	0.0220	0.0135	0.0085	0.0205	0.0140	0.0095	0.0225	0.0115	0.0100
ljw	0.0525	0.0515	0.0485	0.0420	0.0475	0.0475	0.0485	0.0445	0.0465
$\beta = 2$									
New1	0.1000	0.1285	0.1690	0.1025	0.1260	0.1635	0.0995	0.1215	0.1725
New2	0.0620	0.0740	0.0815	0.0570	0.0730	0.0775	0.0630	0.0670	0.0800
New3	0.0510	0.0500	0.0565	0.0470	0.0525	0.0445	0.0510	0.0465	0.0495
oracle	0.0475	0.0505	0.0520	0.0425	0.0505	0.0450	0.0505	0.0475	0.0525
chi	0.0420	0.0605	0.0530	0.0545	0.0445	0.0485	0.0615	0.0490	0.0430
fast	0.0460	0.0635	0.0580	0.0565	0.0475	0.0525	0.0640	0.0520	0.0455
CQ	0.0855	0.0775	0.0675	0.0615	0.0695	0.0755	0.0750	0.0740	0.0700
SD	0.0055	0.0025	0.0015	0.0030	0.0000	0.0005	0.0025	0.0005	0.0005
ljw	0.0480	0.0495	0.0545	0.0410	0.0455	0.0510	0.0535	0.0435	0.0550

Table 3: $n_1 = n_2 = 150$

	Normal			Chi-squared			Student's t		
p	200	500	800	200	500	800	200	500	800
$\beta = 0.5$									
New1	0.0825	0.1000	0.1125	0.0830	0.1045	0.1115	0.0875	0.1095	0.1115
New2	0.0580	0.0575	0.0675	0.0625	0.0635	0.0655	0.0555	0.0650	0.0650
New3	0.0520	0.0495	0.0575	0.0560	0.0505	0.0540	0.0495	0.0540	0.0540
oracle	0.0500	0.0450	0.0515	0.0550	0.0510	0.0505	0.0475	0.0520	0.0490
chi	0.0475	0.0545	0.0485	0.0450	0.0415	0.0480	0.0510	0.0570	0.0440
fast	0.0510	0.0595	0.0510	0.0510	0.0440	0.0530	0.0625	0.0645	0.0495
CQ	0.0655	0.0780	0.0710	0.0720	0.0515	0.0635	0.0565	0.0725	0.0660
SD	0.0530	0.0675	0.0600	0.0585	0.0445	0.0585	0.0465	0.0600	0.0560
ljw	0.0540	0.0520	0.0535	0.0545	0.0555	0.0480	0.0470	0.0500	0.0605
$\beta = 1$									
New1	0.0780	0.1000	0.1040	0.0845	0.0890	0.1070	0.0775	0.0910	0.1100
New2	0.0555	0.0615	0.0685	0.0595	0.0570	0.0585	0.0560	0.0620	0.0610
New3	0.0490	0.0520	0.0490	0.0515	0.0465	0.0425	0.0470	0.0520	0.0460
oracle	0.0485	0.0485	0.0545	0.0480	0.0500	0.0395	0.0475	0.0480	0.0420
chi	0.0405	0.0505	0.0435	0.0470	0.0445	0.0515	0.0510	0.0490	0.0450
fast	0.0430	0.0530	0.0455	0.0495	0.0480	0.0530	0.0545	0.0510	0.0470
CQ	0.0785	0.0690	0.0740	0.0720	0.0695	0.0715	0.0730	0.0795	0.0725
SD	0.0265	0.0115	0.0105	0.0220	0.0130	0.0090	0.0145	0.0145	0.0080
ljw	0.0430	0.0605	0.0460	0.0515	0.0580	0.0525	0.0580	0.0595	0.0510
$\beta = 2$									
New1	0.0775	0.0895	0.1165	0.0840	0.0995	0.1235	0.0940	0.1035	0.1145
New2	0.0550	0.0590	0.0735	0.0590	0.0595	0.0740	0.0675	0.0645	0.0655
New3	0.0495	0.0495	0.0565	0.0510	0.0490	0.0560	0.0575	0.0520	0.0520
oracle	0.0465	0.0485	0.0555	0.0505	0.0485	0.0485	0.0555	0.0505	0.0475
chi	0.0495	0.0405	0.0480	0.0460	0.0530	0.0545	0.0475	0.0395	0.0510
fast	0.0525	0.0415	0.0505	0.0490	0.0560	0.0560	0.0490	0.0400	0.0555
CQ	0.0675	0.0705	0.0695	0.0765	0.0655	0.0775	0.0785	0.0720	0.0875
SD	0.0040	0.0020	0.0005	0.0030	0.0000	0.0000	0.0025	0.0010	0.0000
ljw	0.0490	0.0475	0.0490	0.0465	0.0535	0.0480	0.0490	0.0445	0.0490