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 matrices are assumed to be spiked, which often arises in practice. We propose a new test procedure through projection on the orthogonal complement of principal space. 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 matrices are spiked. Even when the covariance matrices are not spiked, the new test is acceptable.

Keywords: high dimension, mean test, orthogonal complement of principal space, spiked covariance

---

1. Introduction

Suppose that $X_{k,1}, \dots, X_{k,n_k}$ are independent identically distributed (i.i.d.) as $N_p(\mu_k, \Sigma)$, where μ_k and Σ are unknown, $k = 1, 2$. 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, high dimensional setting is adopted, i.e., the dimension p varies as n increase, where $n = n_1 + n_2$ is the total sample size. Testing hypotheses (1) is important in many applications, including biology, finance and economics. Quite often, these data have strong correlations between variables. When strong correlations exist, covariance matrices are often spiked in the sense that a few eigenvalues are distinctively larger than the others. This paper is devoted to testing hypotheses (1) in high dimensional setting with spiked covariance.

A classical test for hypotheses (1) is Hotelling's T^2 test. Hotelling's test statistic is $(\bar{X}_1 - \bar{X}_2)^T S^{-1}(\bar{X}_1 - \bar{X}_2)$, where $S = (n-2)^{-1}\sum_{k=1}^2 \sum_{i=1}^{n_k}(X_{k,i} - \bar{X}_k)(X_{k,i} - \bar{X}_k)^T$ is the pooled sample covariance matrix. However, Hotelling's test 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 low power of Hotelling's test is due to that S is a poor estimator of Σ when p is large compared with n. See Chen and Qin (2010) and the references therein. In high dimensional setting, many test statistics in the literatures are based on an estimator of $(\mu_1 - \mu_2)^T A(\mu_1 - \mu_2)$ for a given positive definite matrix A. For example, Bai and Saranadasa (1996) proposed a test based on

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

which is 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

$$\begin{aligned}
T_{CQ} &= \frac{\sum_{i\neq j}^{n_1} X_{1i}^T X_{1j}}{n_1(n_1-1)} + \frac{\sum_{i\neq j}^{n_2} X_{2i}^T X_{2j}}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} X_{1i}^T X_{2j}}{n_1 n_2} \\
&= \|\bar{X}_1 - \bar{X}_2\|^2 - \frac{1}{n_1}\mathrm{tr}S_1 - \frac{1}{n_2}\mathrm{tr}S_2,
\end{aligned}$$

where $S_k = (n_k - 1)^{-1}\sum_{i=1}^{n_k}(X_{k,i} - \bar{X}_k)(X_{k,i} - \bar{X}_k)^T$ is the sample covariance of the sample k, $k = 1, 2$. Statistic T_{CQ} is also an unbiased estimator of $\|\mu_1 - \mu_2\|^2$. Choosing $A = [\mathrm{diag}(\Sigma)]^{-1}$, Srivastava and Du (2008) proposed a test based on

$$T_S = (\bar{X}_1 - \bar{X}_2)^T[\mathrm{diag}(S)]^{-1}(\bar{X}_1 - \bar{X}_2),$$

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

As Ma et al. (2015) pointed out, however, these test procedures may not be valid if strong correlations exist, i.e., Σ is far away from diagonal matrix. For example, the assumption

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

adopted by Chen and Qin (2010) can be violated when $\Sigma = (1-c)I_p + c\mathbf{1}_p\mathbf{1}_p^T$ where $-1/(p-1) < c < 1$, I_p is the p dimensional identity matrix and $\mathbf{1}_p$ is the p dimensional vector with elements 1. To characterize strong correlations, Ma et al. (2015) considered a factor model and proposed a parameter bootstrap procedure to adjust Chen and Qin (2010)'s critical value.

Strong correlations between variables do exist in practice. In gene expression analysis, genes are correlated due to genetic regulatory networks (see Thulin (2014)). Chen et al. (2011) pointed out that in terms of pathway analysis in proteomic studies, test level can not be guaranteed if correlations are incorrectly assumed to be absent. As Ma et al. (2015) argued, there're strong correlations between different stock returns since they are all affected by the market index.

Incorrectly assuming the absence of correlation between variables will result in level inflation and low power for a test procedure. A class of test procedures is proposed through random projection. See, for example, Lopes et al. (2011), Thulin (2014) and Srivastava et al. (2015). The idea is to project data on some random lower-dimensional subspaces. It has been shown that these procedures perform well under strong correlations.

In many situations, the correlations are determined by a small number of factors. As a consequence, Σ is spiked. See, for example, Cai et al. (2013) and Shen et al. (2013). The random projection methods imply that test procedures are improved when data are projected on certain subspaces. We will see that the ideal subspace is the orthogonal complement of the principal space. Fortunately, the principal space can be estimated consistently even in high dimensional setting by the theory of principal component analysis (PCA). With the assumption of spiked covariance model, we propose a new test procedure through projection on the (estimated) ideal subspace. The asymptotic distribution of the test statistic is derived and hence asymptotic power is given. We will see that the test is more powerful than T_{CQ}. Moreover, even there's no strong correlation showing up, we prove that the new test performs equally well as T_{CQ} does. The idea is also generalized to the unequal variance setting and similar results still hold.

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. All the technical details are in appendix.

2. Some considerations about Chen and Qin (2010)'s test

Let $X_{k,1}, \ldots, X_{k,n_k}$ be i.i.d. observations with common distribution $N(\mu_k, \Sigma)$, $k = 1, 2$. Throughout the paper, we assume $p \to \infty$ as $n \to \infty$ and $n_1/n_2 \to \xi \in (0, +\infty)$, that is, we consider high dimensional and balanced data.

Now we introduce some notations that will be used in this paper. For random variable ξ and η, we write $\xi \sim \eta$ to denote they have the same distribution. Let $\mathcal{L}(\xi)$ be the distribution of ξ and $\mathcal{L}(\xi|\eta)$ be the conditional distribution of ξ given η. We denote by "$\xrightarrow{a.s.}$", "$\xrightarrow{P}$" and "$\xrightarrow{\mathcal{L}}$" the almost surely convergence, convergence in probability and weak convergence. For nonrandom positive sequence $\{a_n\}$ and $\{b_n\}$, $a_n \asymp b_n$ represents $a_n = O(b_n)$ and $b_n = O(a_n)$ as $n \to \infty$. We denote by $\|\cdot\|$ and $\|\cdot\|_F$ the operator and Frobenius norm of matrix, separately. For $p \geq q$, define

$$\mathbb{O}_{p\times q} = \{O |\, O \text{ is } p \times q \text{ column orthonormal matrix }\}.$$

2.1. *The applicable range*

In Chen and Qin (2010), the asymptotic normality of T_{CQ} is derived under condition (2). To determine the critical value of T_{CQ}, they used the limiting normal distribution to approximate the finite sample distribution. The resulting test can asymptotically preserve the test level if and only if the asymptotic normality is correct. While Chen and Qin (2010) only proved (2) is sufficient for the asymptotic normality of T_{CQ}, one may ask: is (2) also necessary for asymptotic normality of T_{CQ}?

To answer this question, we first note that T_{CQ} is a quadratic form of standard normal random vector. To see this, let $Z_{k,i} = \Sigma^{-1/2} X_{k,i}$ so that $Z_{k,i} \sim N_p(0, I_p)$. Write all $Z_{k,i}$ in a long vector form $Z = (Z_{1,1}^T, \ldots, Z_{1,n_1}^T, Z_{2,1}^T, \ldots, Z_{2,n_2}^T)^T$. Then T_{CQ} is a quadratic form of Z and $T_{CQ} = Z^T \big(B_n \otimes \Sigma\big) Z$, where $\otimes$ is the Kronecker product,

$$B_n = \begin{pmatrix} \frac{1}{n_1(n_1-1)}(n_1 \gamma_1 \gamma_1^T - I_{n_1}) & -\frac{1}{\sqrt{n_1 n_2}} \gamma_1 \gamma_2^T \\ -\frac{1}{\sqrt{n_1 n_2}} \gamma_2 \gamma_1^T & \frac{1}{n_2(n_2-1)}(n_2 \gamma_2 \gamma_2^T - I_{n_2}) \end{pmatrix},$$

γ_1 is an n_1 dimensional vector with all elements equal to $1/\sqrt{n_1}$ and γ_2 is an n_2 dimensional vector with all elements equal to $1/\sqrt{n_2}$.

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

Proposition 1. *Suppose Y_n is a k_n dimensional standard normal random vector and A_n is a $k_n \times k_n$ symmetric matrix. Then as $n \to \infty$, a necessary and sufficient condition for*

$$\frac{Y_n^T A_n Y_n - \mathrm{E} Y_n^T A_n Y_n}{[\mathrm{Var}(Y_n^T A_n Y_n)]^{1/2}} \xrightarrow{\mathcal{L}} N(0,1) \tag{3}$$

is that

$$\frac{\lambda_{\max}(A_n^2)}{\mathrm{tr}(A_n^2)} \to 0. \tag{4}$$

To apply Proposition 1 for T_{CQ}, we need to calculate the eigenvalues of $B_n \otimes \Sigma$ which in turn relies on the eigenvalues of B_n. It can be seen that the eigenvalues of 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,

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

and

$$\lambda_{\max}\Big((B_n \otimes \Sigma)^2\Big) = \lambda_{\max}(B_n^2)\lambda_{\max}(\Sigma^2) = \Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)^2 \lambda_{\max}(\Sigma^2).$$

It can be seen that

$$\frac{\lambda_{\max}\Big(\big(B_n \otimes \Sigma\big)^2\Big)}{\operatorname{tr}\big(B_n \otimes \Sigma\big)^2} \to 0$$

is in turn equivalent to $\lambda_{\max}(\Sigma^2)/\operatorname{tr}\Sigma^2 \to 0$. From the inequality

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

we can see that $\lambda_{\max}^2(\Sigma)/\operatorname{tr}(\Sigma^2) \to 0$ is equivalent to (2). By Lemma 1, we can assert that (2) is a necessary and sufficient condition for

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

2.2. Asymptotic distribution under spiked covariance

While condition (2) is necessary for the validity of Chen and Qin (2010)'s test, it requires the eigenvalues of Σ to be concentrated around their average, which can be violated in many applications. For example, in a class of applications, the correlations between variables are mainly driven by several common factors so that Σ has a few eigenvalues which are much larger than the others. See, for example, Jung and Marron (2009), Cai et al. (2013) and Fan and Wang (2017). It is meaningful to consider testing problem for data of such type. To characterize such correlations between variables, we consider the following spiked covariance structure:

Assumption 1. *The covariance matrix Σ has structure $\Sigma = V\Lambda V^T + \sigma^2 I_p$, where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$, $\lambda_1 \geq \cdots \geq \lambda_r > 0$, r is a known number, V is a $p \times r$ orthonormal matrix and $\sigma^2 > 0$. As n, p tend to infinity, r, σ^2 are fixed and Λ satisfies*

$$\kappa p^\beta \geq \lambda_1 \geq \cdots \geq \lambda_r \geq \kappa^{-1} p^\beta.$$

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

The covariance structure in Assumption 1 is commonly adopted in PCA study. See, for example, Cai et al. (2013). The column space of V, the eigenspace of Σ corresponding to the large eigenvalues, is called principal space. Since V

is an orthonormal matrix, VV^T is the orthogonal projection matrix on the principal space. Let $\tilde{V}$ be a $p \times (p-r)$ full column rank orthonormal matrix orthogonal to columns of V. Although such $\tilde{V}$ is not unique, the projection matrix $\tilde{V}\tilde{V}^T = I - VV^T$ is unique and is the projection matrix on the orthogonal complement of principal space. Spiked covariance also has connection to factor model. In fact, the factor model in Ma et al. (2015) corresponds to $\beta = 1$ in Assumption 1.

Under Assumption 1, it is easy to see that condition (2) is equivalent to $\beta < 1/2$. Our previous arguments assert that the asymptotic distribution of T_{CQ} won't be normal when $\beta \geq 1/2$. Under Assumption 1, we can go further and give the asymptotic distribution of T_{CQ}. In fact, we have the following theorem:

Theorem 1. *Under Assumption 1, suppose $\lambda_i/p^\beta \to l_i \in (0, +\infty)$, $i = 1, \ldots, r$. Let $\tau = 1/n_1 + 1/n_2$, we further assume $(\tau p^\beta)^{-1/2}\big(V^T(\mu_1 - \mu_2)\big)_i \to \zeta_i \in (-\infty, +\infty)$, $i = 1, \ldots, r$, and $(\tau p^\beta)^{-1}\|\tilde{V}^T(\mu_1 - \mu_2)\|^2 \to \zeta^* \in [0, +\infty)$. We have*

$$\frac{1}{\tau p^\beta} T_{CQ} \xrightarrow{\mathcal{L}} \sum_{i=1}^r (l_i Z_i + \zeta_i)^2 + \zeta^* + \sqrt{2}\sigma^2 \delta_{\{\frac{1}{2}\}}(\beta)\epsilon - \sum_{i=1}^r l_i,$$

where $Z_1, \ldots, Z_r$ and ϵ are i.i.d. $N(0,1)$ random variables, $\delta_{\{\frac{1}{2}\}}(\beta)$ equals to 1 if $\beta = 1/2$ and 0 otherwise.

It is possible to adjust the critical value of T_{CQ} such that the resulting test has correct level at least asymptotically. For example, one can use permutation method and the resulting test is exact. See, for example, Lehmann and Romano (2005) Chapter 15. See also Ma et al. (2015) for a parametric bootstrap method. Theorem 1 can be used to derive the asymptotic power function of the adjusted test procedure. Let $F(x; \zeta_1, \ldots, \zeta_r, \zeta^*)$ be the cumulative distribution function of $\sum_{i=1}^r (l_i Z_i + \zeta_i)^2 + \zeta^* + \sqrt{2}\sigma^2 \delta_{\{\frac{1}{2}\}}(\beta)\epsilon - \sum_{i=1}^r l_i$. If the adjusted critical value C^* of T_{CQ}, which may depend on data, satisfies

$$F\Big(\frac{C^*}{\tau p^\beta}; 0, \ldots, 0, 0\Big) \xrightarrow{P} 1 - \alpha,$$

the resulting test will have correct level asymptotically. In this way, the asymptotic power is

$$\lim_{n,p\to\infty} \Pr(T_{CQ} \geq C^*) = 1 - F(F^{-1}(1-\alpha; 0, \ldots, 0, 0); \zeta_1, \ldots, \zeta_r, \zeta^*),$$

where α is the nominal level. In particular, the test will have trivial asymptotic power if $\zeta_1 = \cdots = \zeta_r = \zeta^* = 0$, which is equivalent to $\sum_{i=1}^r \zeta_i^2 + \zeta^* = 0$, or

$$\frac{1}{\tau p^\beta}\|\mu_1 - \mu_2\|^2 \to 0.$$

Conversely, to make T_{CQ} have non-trivial power, $\|\mu_1 - \mu_2\|^2$ is at least of order τp^β.

3. Methodology

3.1. Motivation

Note that the main part of T_{CQ} is $\tau\|\bar{X}_1 - \bar{X}_2\|^2$, which is also the main part of T_{BS} and Ma et al. (2015)'s method. It can be written as the sum of two terms

$$\tau\|V^T(\bar{X}_1 - \bar{X}_2)\|^2 + \tau\|\tilde{V}^T(\bar{X}_1 - \bar{X}_2)\|^2. \tag{5}$$

Under the null hypotheses, we have

$$\operatorname{Var}\left(\tau\|V^T(\bar{X}_1-\bar{X}_2)\|^2\right) = \sum_{i=1}^r 2(\lambda_i+\sigma^2)^2, \quad \operatorname{Var}\left(\|\tilde{V}^T(\bar{X}_1-\bar{X}_2)\|^2\right) = 2\sigma^4(p-r).$$

The ratio of the two variance is

$$\frac{\sum_{i=1}^r 2(\lambda_i + \sigma^2)^2}{2\sigma^4(p-r)} \asymp p^{2\beta-1},$$

which tends to ∞ as $p \to \infty$ for $\beta > 1/2$. While $\tau\|V^T(\bar{X}_1 - \bar{X}_2)\|^2$ has relative large variance, it only involves the signal from r dimension. This motivates us to drop the first part of (5) and only use the second part. After adjustment of expectation, we define the following statistic

$$T_1 = \|\tilde{V}^T(\bar{X}_1 - \bar{X}_2)\|^2 - \frac{1}{n_1}\operatorname{tr}(\tilde{V}^T S_1 \tilde{V}) - \frac{1}{n_2}\operatorname{tr}(\tilde{V}^T S_2 \tilde{V}).$$

Proposition 2 shows that T_1 has normal limit distribution.

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

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

In another point of view, T_1 is obtained by transforming $X_{k,i}$ to $\tilde{V}^T X_{k,i}$ ($i = 1, \dots, n_k$, $k = 1, 2$) and then invoking the statistic of Chen and Qin (2010). A class of test procedures have been proposed through random projection to lower dimensional space. See, for example, Lopes et al. (2011), Thulin (2014) and Srivastava et al. (2015). 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. This raises the question: is there an optimal projection which is nonrandomized?

Under null hypothesis, we have that

$$\tilde{V} = \underset{O \in \mathbb{O}_{p\times(p-r)}}{\arg\min} \ \mathrm{Var}\left(\|O^T(\bar{X}_1 - \bar{X}_2)\|^2\right).$$

Thus, transformation by $\tilde{V}$ is optimal in the sense of variance reduction. Based on $\tilde{V}^T X_{ki}$, the likelihood ratio test statistic for hypothesis (1) is then $\|\tilde{V}^T(\bar{X}_1 - \bar{X}_2)\|^2$ which coincides with our proposal. In this view, T_1 can be regarded as a restricted likelihood ratio test.

3.2. New test

We denote by $\hat{V}$ and $\hat{\tilde{V}}$ the first r and last $p-r$ eigenvectors of S respectively. Similarly, we denote by $\hat{V}$ and $\hat{\tilde{V}}_k$ the first r and last $p-r$ eigenvectors of S_k respectively, $k = 1, 2$. As estimators of their population counterparts, these simple statistics actually reach the optimal convergence rate (See Cai et al. (2013)).

Note that T_1 relies on the subspace $\tilde{V}\tilde{V}^T$ which is unknown and thus should be estimated. The first part of T_1, $\|\tilde{V}^T(\bar{X}_1 - \bar{X}_2)\|^2$, can be directly estimated by $\|\hat{\tilde{V}}^T(\bar{X}_1 - \bar{X}_2)\|^2$. Note that $n_1^{-1}\mathrm{tr}(\tilde{V}^T S_1 \tilde{V})$, the second part of T_1, only involves sample one. We estimate it by $n_1^{-1}\mathrm{tr}(\hat{\tilde{V}}_1^T S_1 \hat{\tilde{V}}_1)$. Similarly, we estimate the third part of T_1 by $n_2^{-1}\mathrm{tr}(\hat{\tilde{V}}_2^T S_2 \hat{\tilde{V}}_2)$. Define

$$T_2 = \|\hat{\tilde{V}}^T(\bar{X}_1 - \bar{X}_2)\|^2 - \frac{1}{n_1}\mathrm{tr}(\hat{\tilde{V}}_1^T S_1 \hat{\tilde{V}}_1) - \frac{1}{n_2}\mathrm{tr}(\hat{\tilde{V}}_2^T S_2 \hat{\tilde{V}}_2).$$

The asymptotic result of Proposition 2 involves σ^2. In order to formulate a test procedure by asymptotic distribution, σ^2 needs to be consistently estimated. Note that σ^2 can be written as $\sigma^2 = (p-r)^{-1}\sum_{i=r+1}^{p}\lambda_i(\Sigma)$. So it can be estimated by

$$\hat{\sigma}^2 = \frac{1}{p-r}\sum_{i=r+1}^{p}\lambda_i(S).$$

Now we propose our new test statistic as

$$Q = \frac{T_2}{\hat{\sigma}^2\sqrt{2\tau^2 p}}. \tag{6}$$

In the next section, it will be proved that the asymptotic null distribution of Q is $N(0,1)$. Thus, the null hypothesis is rejected when Q is larger than the upper α quantile of $N(0,1)$.

Remark 1. When both samples are simultaneously transformed by shift and orthogonal transformation, the statistic T_2 is invariant. More precisely, T_2 is invariant under the following transformation:

$$X_{1,i} \mapsto OX_{1,i} + \mu \text{ and } X_{2,j} \mapsto OX_{2,j} + \mu,\ i = 1,\ldots,n_1,\ j = 1,\ldots,n_2,$$

where $\mu \in \mathbb{R}^p$ and $O \in \mathbb{O}_{p\times p}$.

Theoretical results will show that the asymptotic variance of T_2 is significantly smaller than T_{CQ}. Since the new test statistic estimates $\|\tilde{V}^T(\mu_1-\mu_2)\|^2$, the superiority of the new test will be established if

$$\frac{\|\tilde{V}^T(\mu_1-\mu_2)\|}{\|\mu_1-\mu_2\|} \approx 1. \tag{7}$$

Obviously, (7) is not always the case since there always exists some $\tilde{V}$ and $\mu_1-\mu_2$ such that $\|\tilde{V}^T(\mu_1-\mu_2)\| = 0$. However, (7) is reasonable since $\tilde{V}\tilde{V}^T$ is nearly an identity matrix in the sense that $\|I_p - \tilde{V}\tilde{V}^T\|_F^2/\|I_p\|_F^2 = r/p \to 0$.

In bayesian framework, if we assume that the elements of μ_k are independently generated from certain prior distribution, it can be established that $\|\tilde{V}(\mu_1-\mu_2)\|/\|\mu_1-\mu_2\| \xrightarrow{P} 1$.

3.3. Theoretical results

In this section, we study the asymptotic behavior of the new test procedure.

We first give a result of the convergence rate of $\hat{\sigma}^2$. In particular, this result shows that $\hat{\sigma}^2$ is a consistent estimator of σ^2. Our proof relies on the Weyl's inequality.

Proposition 3. *Under Assumptions 1, we have*

$$\hat{\sigma}^2 = \sigma^2 + O_P\Big(\frac{\max(n,p)}{np}\Big).$$

In the construction of T_2, we replace $\tilde{V}$ by $\hat{\tilde{V}}$. The asymptotic property of T_2 is closely related the asymptotic property of $\hat{\tilde{V}}\hat{\tilde{V}}^T$ as an estimator of $\tilde{V}\tilde{V}^T$. However, $\hat{\tilde{V}}\hat{\tilde{V}}^T$ can not always consistently estimate $\tilde{V}\tilde{V}^T$ in high dimensional setting. In fact, Cai et al. (2014)'s Theorem 5 implies that it is possible only when $p^{1-\beta}/n \to 0$, see Lemma 2 in appendix. The asymptotic normality of T_2 requires a stronger condition.

Assumption 2. *Assume $p/n^2 \to 0$.*

When $\beta = 1/2$, Assumption 2 is equivalent to $p^{1-\beta}/n$, otherwise 2 is stronger than $p^{1-\beta}/n$. The following theorem establishes the asymptotic normality of T_2.

Theorem 2. *Under Assumptions 1 and 2, suppose*

$$\frac{n}{\sqrt{p}}\|\mu_1-\mu_2\|^2 = O(1),$$

we have

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

Now the power function of the new test can be obtained immediately.

Corollary 1. *Under the conditions of Theorem 2, the asymptotic power function of the new test is*

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

Recall that T_{CQ} has trivial power if $\|\mu_1-\mu_2\|^2=o(\tau p^\beta)$. Note that the power function of the new test do not affected by β. As a result, when $\beta>1/2$, the new test tends to be much more powerful.

The proof of Theorem 2 implies that when Assumption 2 is not satisfied, the asymptotic normality is invalid. In this case, permutation method can be used to determine the critical value. We will see from simulation results that the new test has good power behavior even if p is much large than n.

4. Numerical studies

4.1. Simulation results

In this section, we report the simulation performance of the proposed test and compare it with T_{CQ} and T_S.

In our simulation studies, samples are generated from the model described in Assumption 1, where $V\in\mathbb{O}_{p\times r}$ is randomly generated from Haar invariant distribution, $\lambda_{k,i}$ equals p^β plus a random error from $U(0,1)$ (Uniform distribution between 0 and 1) and $\sigma_k^2=1$, $k=1,2$. In unequal variance case, (V_1,Λ_1) and (V_2,Λ_2) are independently generated. We take nominal level $\alpha=0.05$.

First, we simulate the level of the new test. We set factor number $r=2$. Samples are repeatedly generated 1000 times to calculate empirical level. For comparison, we also give corresponding 'oracle' level which is calculated by 'statistics' $T_1/(\sigma^2\sqrt{2p\tau^2})$ in equal variance case and $T_1/\sqrt{\sigma_n^2}$ in unequal variance case. The results for equal variance case and unequal variance case are listed in Table 1 and 2, respectively. From the results, we can find that for small n and p, even oracle level is not satisfied. Level of the new test is a little inflated compared with oracle level. In all cases, the empirical level tends to be more close to 0.05 as n increases.

Table 1: Test level simulation. Equal variance case.

		β=0.5		β=1		β=2	
n	p	NEW	ORACLE	NEW	ORACLE	NEW	ORACLE
300	200	0.075	0.062	0.079	0.062	0.074	0.070
300	400	0.074	0.065	0.061	0.044	0.046	0.040
300	600	0.058	0.041	0.070	0.052	0.071	0.055
300	800	0.066	0.047	0.071	0.052	0.062	0.048
600	200	0.061	0.055	0.052	0.051	0.058	0.056
600	400	0.051	0.048	0.051	0.042	0.059	0.051
600	600	0.061	0.058	0.056	0.054	0.051	0.047
600	800	0.053	0.046	0.060	0.050	0.056	0.048

Next, we simulate the empirical power of the new test. The simulation results of Ma et al. (2015) have showed that the level of the Chen and Qin (2010)'s test can't be guaranteed when covariance is spiked. To be fair, critical values are all determined by permutation method. We set $\Sigma_1 = \Sigma_2$, under which permutation method can produce exact test procedures, see Lehmann and Romano (2005)'s Example 15.2.2. We permutate the sample 100 times to determine the critical value. We repeat the test procedure 500 times to obtain empirical power. We plot the empirical power versus signal-to-noise ratio SNR $= \|\mu_1 - \mu_2\|^2/(\sigma^2\sqrt{2\tau^2 p})$. The results are illustrated in figure 1 and 2, where 'NEW', 'CQ' and 'S' represent the new test, Chen and Qin (2010)'s test and Srivastava and Du (2008)'s test. From the results, we can find that when Σ is spiked, the new test outperforms T_{CQ} substantially; when Σ is not spiked, all three tests have similar performance.

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

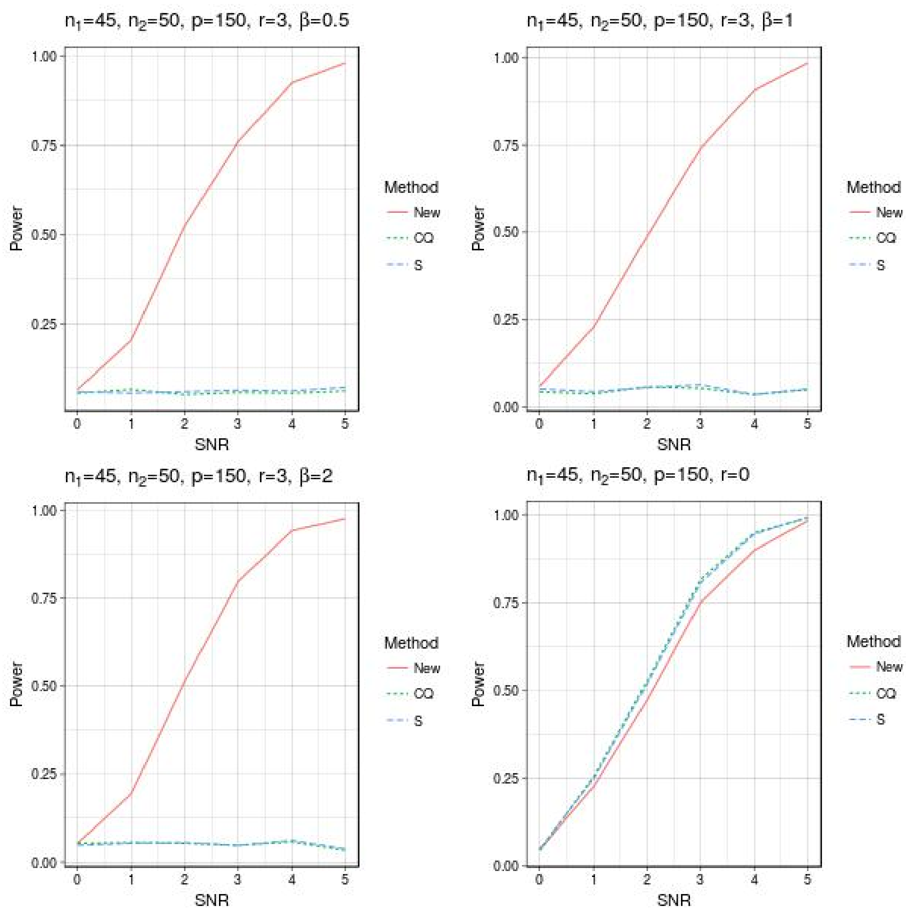

Figure 1: Empirical power simulation.

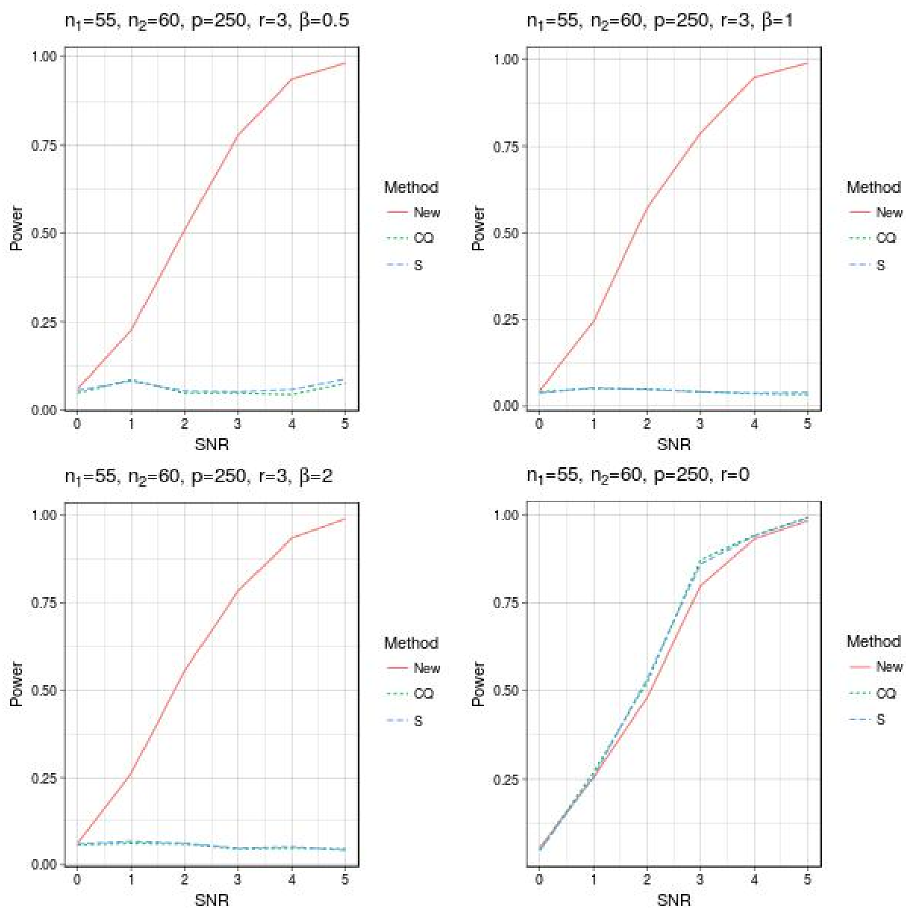

Figure 2: Empirical power simulation.

Table 2: Test level simulation. Unequal variance case.

		β=0.5		β=1		β=2	
n	p	NEW	ORACLE	NEW	ORACLE	NEW	ORACLE
300	200	0.066	0.058	0.057	0.054	0.054	0.050
300	400	0.063	0.047	0.063	0.047	0.069	0.052
300	600	0.070	0.058	0.091	0.059	0.086	0.053
300	800	0.069	0.040	0.097	0.055	0.083	0.056
600	200	0.054	0.056	0.049	0.049	0.052	0.051
600	400	0.060	0.054	0.067	0.059	0.060	0.055
600	600	0.041	0.034	0.069	0.062	0.055	0.049
600	800	0.077	0.063	0.066	0.058	0.071	0.058

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 $\Sigma_1 = \Sigma_2$. The first eigenvaule of 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 set $r = 1$ and perform our new test. The p value is 0.149, which is obtained by 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 drop big variance terms from T_{CQ} 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 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

$$\hat{r} = \mathrm{argmax}_{l \le R} \frac{\lambda_l(S)}{\lambda_{l+1}(S)}, \tag{8}$$

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

Our theoretical results rely on the assumption $\sqrt{p}/n \to 0$. In the situation of small sample 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.

Appendix

Lemma 1 (Weyl's inequality)**.** *Let H and P be two symmetric $n \times n$ matrices and $M = H + P$. If $r + s - 1 \le i \le j + k - n$, we have*

$$\lambda_j(H) + \lambda_k(P) \le \lambda_i(M) \le \lambda_r(H) + \lambda_s(P).$$

See, for example, Horn and Johnson (1985) Theorem 4.3.1.

Lemma 2 (Convergence rate of principal space estimation)**.** *Under the Assumption 1, we have*

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

Proof. Theorem 5 of Cai et al. (2013) asserts that sample principal subspace $\hat{V}\hat{V}^T$ is a minimax rate optimal estimator of VV^T, namely, it reaches the minimax convergence rate

$$E\|\hat{V}\hat{V}^T - VV^T\|_F^2 \asymp r \wedge (p-r) \wedge \frac{r(p-r)}{(n_1+n_2-2)p^\beta} \tag{9}$$

as long as the right hand side tends to 0. It's obvious that the right hand side of (9) is of order $p^{1-\beta}/n$. $\square$

Lemma 3 (Bai-Yin's law)**.** *Suppose* $B_n = \frac{1}{q}ZZ^T$ *where* Z *is* $p \times q$ *random matrix composed of i.i.d. random variables with zero mean, unit variance and finite fourth moment. As* $q \to \infty$ *and* $\frac{p}{q} \to c \in [0, \infty)$*, the largest and smallest non-zero eigenvalues of* B_n *converge almost surely to* $(1+\sqrt{c})^2$ *and* $(1-\sqrt{c})^2$*, respectively.*

Remark 2. Lemma 3 is known as the Bai-Yin's law (Bai and Yin (1993)). As in Remark 1 of Bai and Yin (1993), the smallest non-zero eigenvalue is the $p-q+1$ smallest eigenvalue of B for $c > 1$.

Corollary 2. *Suppose that* W_n *is a* $p\times p$ *matrix distributed as* $\mathrm{Wishart}_p(n, I_p)$ *where* $\mathrm{Wishart}_p(m, \Psi)$ *is the* p *dimensional Wishart distribution with parameter* Ψ *and* m *degrees of freedom. Then as* $n, p \to \infty$*,*

$$\lambda_1(W_n) = O_P(\max(n, p)).$$

Proof. Since $[0, +\infty]$ is compact, for every subsequance $\{n_k\}$ of $\{n\}$, there is a further subsequance $\{n_{k_l}\}$ along which $p/n \to c \in [0, +\infty]$.

If $c \in [0, +\infty)$, by Lemma 3, we have that

$$\frac{\lambda_1(W_{n_{k_l}})}{n_{k_l}} \xrightarrow{P} (1+c)^2.$$

Hence the conclusion holds along this subsequance. If $c = +\infty$, suppose $W_n = Z_n Z_n^T$ where Z_n is a $p \times n$ matrix with all elements independently distributed as $N(0,1)$. Then

$$\frac{\lambda_1(W_{n_{k_l}})}{p} = \frac{\lambda_1(Z_{n_{k_l}}^T Z_{n_{k_l}})}{p} \xrightarrow{P} 1$$

by Lemma 3, which proves the conclusion along the subsequance. Now the conclusion holds by a standard subsequance argument. $\square$

Proof of Proposition 1. Let $\lambda_1(A_n) \geq \cdots \geq \lambda_{k_n}(A_n)$ be the eigenvalues of A_n, then

$$\frac{Y_n^T A_n Y_n - \mathrm{E} Y_n^T A_n Y_n}{\left[\mathrm{Var}(Y_n^T A_n Y_n)\right]^{1/2}} = \sum_{i=1}^{k_n} \frac{\lambda_i(A_n)}{\left[2\mathrm{tr}(A_n^2)\right]^{1/2}} (Z_{ni}^2 - 1), \tag{10}$$

where Z_{ni}'s ($i = 1, \ldots, k_n$) are independent standard normal random variables.

If 4 holds, then

$$\begin{aligned}
&\sum_{i=1}^{k_n} \mathrm{E}\Big[\frac{\lambda_i^2(A_n)}{2\mathrm{tr}(A_n^2)}(Z_{ni}^2 - 1)^2 \Big\{\frac{\lambda_i^2(A_n)}{2\mathrm{tr}(A_n^2)}(Z_{ni}^2 - 1)^2 \geq \epsilon\Big\}\Big] \\
\leq &\sum_{i=1}^{k_n} \frac{\lambda_i^2(A_n)}{2\mathrm{tr}(A_n^2)} \mathrm{E}\Big[(Z_{n1}^2 - 1)^2 \Big\{\frac{\lambda_{\max}(A_n^2)}{2\mathrm{tr}(A_n^2)}(Z_{n1}^2 - 1)^2 \geq \epsilon\Big\}\Big] \\
= &\frac{1}{2}\mathrm{E}\Big[(Z_{n1}^2 - 1)^2 \Big\{\frac{\lambda_{\max}(A_n^2)}{2\mathrm{tr}(A_n^2)}(Z_{n1}^2 - 1)^2 \geq \epsilon\Big\}\Big] \to 0.
\end{aligned}$$

Hence 3 follows by Lindeberg's central limit theorem.

Conversely, if 3 holds, we will prove that there is a subsequence of $\{n\}$ along which 4 holds. Then 4 will hold by a standard contradiction argument.

Denote $c_{ni} = \lambda_i(A_n)/\left[2\mathrm{tr}(A_n^2)\right]^{1/2}$ ($i = 1, \ldots, k_n$), we have $c_{ni} \in [-\sqrt{2}/2, \sqrt{2}/2]$. Since 3 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. For $t \in (-1, 1)$, we have

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

Denote $b_{nl} = \sum_{j=1}^{k_n}(c_{nj})^l$, $n = 1, 2, \cdots$ and $l = 3, 4, \cdots$. For $l \geq 3$, $\big|\sum_{j=1}^{k_n}(c_{nj})^l\big| \leq \big|\sum_{j=1}^{k_n}(c_{nj})^2\big| = 1/2$. 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. Apply dominated convergence theorem to this subsequence we have $\log \mathrm{E} \exp\big(it\sum_{j=1}^{k_n} c_{nj}(Z_{nj}^2 - 1)\big) \to$

$-\frac{1}{2}t^2 + \frac{1}{2}\sum_{l=3}^{+\infty} b_l \frac{1}{l}(i2t)^l$ for $t \in (-1/2, 1/2)$. By the property of power series, we have $b_l = 0$ for $l \geq 3$. Then 4 follows by noting that $b_{n4} \geq \max_j (c_{nj})^4$. $\square$

The rest of the Appendix is devoted to the proof of propositions and theorems in the paper.

Proof of Theorem 1. Note that $(n_k - 1)S_k \sim \text{Wishart}_p(n_k - 1, \Sigma)$, $k = 1, 2$, we have

$$\mathrm{E}\Big(\frac{1}{n_1}\operatorname{tr} S_1 + \frac{1}{n_2}\operatorname{tr} S_2\Big) = \tau \operatorname{tr}\Sigma,$$

and

$$\begin{aligned}
&\operatorname{Var}\Big(\frac{1}{n_1}\operatorname{tr} S_1 + \frac{1}{n_2}\operatorname{tr} S_2\Big) = \Big(\frac{2}{n_1^2(n_1-1)} + \frac{2}{n_2^2(n_2-1)}\Big)\operatorname{tr}\Sigma^2 \\
=&O\Big(\frac{1}{n^3}(p^{2\beta}+p)\Big) = O\Big(\frac{p^{2\beta}}{n^3}\Big).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\frac{1}{n_1}\operatorname{tr} S_1 + \frac{1}{n_2}\operatorname{tr} S_2 = \tau \operatorname{tr}\Sigma + O_P\Big(\frac{1}{n\sqrt{n}}p^\beta\Big) \\
=&\tau\sum_{i=1}^r(\lambda_i+\sigma^2) + \tau(p-r)\sigma^2 + O_P\Big(\frac{1}{n\sqrt{n}}p^\beta\Big) \\
=&\tau p^\beta\sum_{i=1}^r l_i + \tau(p-r)\sigma^2 + o_P\Big(\frac{1}{n}p^\beta\Big).
\end{aligned}$$

Thus,

$$\frac{1}{\tau p^\beta}\Big(\frac{1}{n_1}\operatorname{tr} S_1 + \frac{1}{n_2}\operatorname{tr} S_2\Big) = \sum_{i=1}^r l_i + p^{1-\beta}\sigma^2 + o_P(1). \tag{11}$$

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

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

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

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

By the assumptions of the theorem, we have that

$$\frac{1}{\tau p^\beta}\|V^T(\bar{X}_1-\bar{X}_2)\|^2 \xrightarrow{\mathcal{L}} \sum_{i=1}^{r}(l_i Z_i+\zeta_i)^2. \tag{12}$$

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

$$\begin{aligned}&\|\tilde{V}^T(\bar{X}_1-\bar{X}_2)\|^2 = \left\|\tilde{V}^T(\mu_1-\mu_2)+\tilde{V}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\right\|^2\\ =&\|\tilde{V}^T(\mu_1-\mu_2)\|^2+\left\|\tilde{V}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\right\|^2+2(\mu_1-\mu_2)^T\tilde{V}\tilde{V}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big).\end{aligned} \tag{13}$$

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

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

For the intersection term, we have

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

It follows that

$$\frac{1}{\tau p^\beta}\Big(\left\|\tilde{V}^T(\bar{X}_1-\bar{X}_2)\right\|^2-\sigma^2\tau(p-r)-\left\|\tilde{V}^T(\mu_1-\mu_2)\right\|^2\Big) \xrightarrow{\mathcal{L}} \sqrt{2}\sigma^2\delta_{\{\frac{1}{2}\}}(\beta)\epsilon. \tag{14}$$

Combining (11) (12) and (14) leads to

$$\begin{aligned}&\frac{1}{\tau p^\beta}T_{CQ}=\frac{1}{\tau p^\beta}\big(\|\bar{X}_1-\bar{X}_2\|^2-\frac{1}{n_1}\operatorname{tr}S_1-\frac{1}{n_2}\operatorname{tr}S_2\big)\\ =&\frac{1}{\tau p^\beta}\|V^T(\bar{X}_1-\bar{X}_2)\|^2+\frac{1}{\tau p^\beta}\big(\|\tilde{V}^T(\bar{X}_1-\bar{X}_2)\|^2-\sigma^2\tau(p-r)-\|\tilde{V}^T(\mu_1-\mu_2)\|^2\big)\\ &-\frac{1}{\tau p^\beta}\Big(\frac{1}{n_1}\operatorname{tr}S_1+\frac{1}{n_2}\operatorname{tr}S_2\Big)+\frac{\sigma^2(p-r)}{p^\beta}+\frac{1}{\tau p^\beta}\|\tilde{V}^T(\mu_1-\mu_2)\|^2\\ =&\sum_{i=1}^{r}(l_iZ_i+\zeta_i)^2+\sqrt{2}\sigma^2\delta_{\{\frac{1}{2}\}}(\beta)\epsilon-(\sum_{i=1}^{r}l_i+p^{1-\beta}\sigma^2)+\frac{\sigma^2(p-r)}{p^\beta}+\zeta^*+o_P(1)\\ \xrightarrow{\mathcal{L}}&\sum_{i=1}^{r}(l_iZ_i+\zeta_i)^2+\zeta^*+\sqrt{2}\sigma^2\delta_{\{\frac{1}{2}\}}(\beta)\epsilon-\sum_{i=1}^{r}l_i.\end{aligned}$$

This completes the proof.

□

Proof of Proposition 2. Note that

$$
\begin{aligned}
&\|\tilde{V}^T(\bar{X}_1-\bar{X}_2)\|^2 = \left\|\tilde{V}^T(\mu_1-\mu_2)+\tilde{V}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\right\|^2\\
=&\|\tilde{V}^T(\mu_1-\mu_2)\|^2+\left\|\tilde{V}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\right\|^2+2(\mu_1-\mu_2)^T\tilde{V}\tilde{V}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\\
=&\|\tilde{V}^T(\mu_1-\mu_2)\|^2+\left\|\tilde{V}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\right\|^2+o_P(\frac{\sqrt{p}}{n}).
\end{aligned}
\tag{15}
$$

The last equality holds since

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

For $k=1,2$, we have

$$
\frac{1}{n_k}\mathrm{tr}(\tilde{V}^TS_k\tilde{V})\sim\frac{\sigma^2}{n_k(n_k-1)}\chi^2_{(p-r)(n_k-1)}=\sigma^2\frac{p-r}{n_k}\Big(1+O_P(\frac{1}{\sqrt{(p-r)(n_k-1)}})\Big),
$$

where the last equality comes from central limit theorem. It follows that

$$
\frac{1}{n_1}\mathrm{tr}(\tilde{V}^TS_1\tilde{V})+\frac{1}{n_2}\mathrm{tr}(\tilde{V}^TS_2\tilde{V})=\sigma^2\tau(p-r)+o_P(\frac{\sqrt{p}}{n}).\tag{16}
$$

Equation (13) and (16) imply that

$$
\frac{T_1-\|\tilde{V}^T(\mu_1-\mu_2)\|^2}{\sigma^2\sqrt{2\tau^2p}}=\frac{\|\bar{Y}_1-\bar{Y}_2\|^2-\sigma^2\tau(p-r)}{\sigma^2\sqrt{2\tau^2p}}+o_P(1).\tag{17}
$$

Since $\|\tilde{V}^T(\bar{Z}_1-\bar{Z}_2)\|^2\sim\sigma^2\tau\chi^2_{p-r}$, the proposition follows by central limit theorem. □

Proof of Proposition 3. Note that $(n-2)S\sim\mathrm{Wishart}_p(n-2,\Sigma)$. Denote by $\Sigma=UEU^T$ the spectral decomposition of Σ, where $U=(V,\tilde{V})$ is an orthogonal matrix and $E=\mathrm{diag}(\lambda_1+\sigma^2,\ldots,\lambda_r+\sigma^2,\sigma^2,\ldots,\sigma^2)$. Let Z be a $p\times(n-2)$ random matrix with all elements i.i.d. distributed as $N(0,1)$, then

$$
S\sim\frac{1}{n-2}UE^{1/2}ZZ^TE^{1/2}U^T.
$$

Thus,

$$
\hat{\sigma}^2\sim\frac{1}{(p-r)(n-2)}\sum_{i=r+1}^{p}\lambda_i(UE^{1/2}ZZ^TE^{1/2}U^T)=\frac{1}{(p-r)(n-2)}\sum_{i=r+1}^{n-2}\lambda_i(Z^TEZ).
$$

Denote $Z = (Z_{(1)}^T, Z_{(2)}^T)^T$, where $Z_{(1)}$ and $Z_{(2)}$ are the first r rows and last $p-r$ rows of Z. We have

$$Z^T E Z = Z_{(1)}^T(\Lambda + \sigma^2 I_r)Z_{(1)} + \sigma^2 Z_{(2)}^T Z_{(2)}.$$

The first term is of rank r. By Weyl's inequality, we have

$$\sigma^2\lambda_i(Z_{(2)}^T Z_{(2)}) \le \lambda_i(Z^T E Z) \le \sigma^2\lambda_{i-r}(Z_{(2)}^T Z_{(2)}), \quad i = r+1, \ldots, n-2.$$

Thus,

$$\sigma^2 \sum_{i=r+1}^{n-2} \lambda_i(Z_{(2)}^T Z_{(2)}) \le \sum_{i=r+1}^{n-2} \lambda_i(Z^T E Z) \le \sigma^2 \sum_{i=1}^{n-r-2} \lambda_i(Z_{(2)}^T Z_{(2)}).$$

It follows that

$$\begin{aligned}&\Big|\frac{1}{(p-r)(n-2)} \sum_{i=r+1}^{n-2} \lambda_i(Z^T E Z) - \frac{1}{(p-r)(n-2)}\sigma^2 \sum_{i=1}^{n-2} \lambda_i(Z_{(2)}^T Z_{(2)})\Big| \\ \le & r\sigma^2 \frac{1}{(p-r)(n-2)}\lambda_1(Z_{(2)}^T Z_{(2)}).\end{aligned}$$

By Corollary 2, $\lambda_1(Z_{(2)}^T Z_{(2)}) = O_P(\max(n,p))$. Thus,

$$\begin{aligned}&\frac{1}{(p-r)(n-2)} \sum_{i=r+1}^{n-2} \lambda_i(Z^T E Z) \\ =&\frac{1}{(p-r)(n-2)}\sigma^2 \sum_{i=1}^{n-2} \lambda_i(Z_{(2)}^T Z_{(2)}) + O_P(\frac{\max(n,p)}{np}) \\ =&\frac{1}{(p-r)(n-2)}\sigma^2 \mathrm{tr}(Z_{(2)}^T Z_{(2)}) + O_P(\frac{\max(n,p)}{np}) \\ =&\sigma^2 + O_P(\frac{1}{\sqrt{np}}) + O_P(\frac{\max(n,p)}{np}).\end{aligned}$$

The last equality comes from central limit theorem. The theorem follows by noting that

$$O_P(\frac{1}{\sqrt{np}}) = O_P(\frac{\sqrt{np}}{np}) = O_P(\frac{\max(n,p)}{np}).$$

□

Proof of Theorem 2. Note that $\mathrm{tr}(\hat{\hat{V}}^T S_k \hat{\hat{V}}) = \sum_{i=r+1}^p \lambda_i(S_k)$, $k = 1, 2$. Similar to Proposition 3, we have $\mathrm{tr}(\hat{\hat{V}}^T S_k \hat{\hat{V}}) = (p-r)\sigma^2 + O_P(\max(n,p)/n)$,

$k = 1, 2$. Then

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

By Assumption 2, $n^{-1}p^{-1/2}\max(n, p) = \max(p^{-1/2}, p^{1/2}/n) \to 0$. Note that

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

Let

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

To prove the theorem, it suffices to show that

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

We first deal with P_2. To proves the convergence in probability, we only need to prove the convergence in L^2. Note that $\bar{X}_1$, $\bar{X}_2$, and S are mutually independent and $\hat{\tilde{V}}\hat{\tilde{V}}^T$ only depends on S. Thus

$$\begin{aligned}
&\mathrm{E}P_2^2 = \mathrm{E}[\mathrm{E}P_2^2|S] = 4\tau\mathrm{E}[(\mu_1 - \mu_2)^T\hat{\tilde{V}}\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}}\hat{\tilde{V}}^T(\mu_1 - \mu_2)] \\
\leq&4\tau\mathrm{E}[\lambda_1(\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})(\mu_1 - \mu_2)^T\hat{\tilde{V}}\hat{\tilde{V}}^T(\mu_1 - \mu_2)] \leq 4\tau\|\mu_1 - \mu_2\|^2\mathrm{E}[\lambda_1(\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})] \\
=&O(\frac{\sqrt{p}}{n^2})\mathrm{E}[\lambda_1(\hat{\tilde{V}}^T(V\Lambda V^T + \sigma^2 I_p)\hat{\tilde{V}})] \leq O(\frac{\sqrt{p}}{n^2})\big(\kappa p^\beta\mathrm{E}[\lambda_1(\hat{\tilde{V}}^T VV^T\hat{\tilde{V}})] + \sigma^2\big).
\end{aligned}$$

By the relationship

$$\lambda_1(\hat{\tilde{V}}^T VV^T\hat{\tilde{V}}) \leq \mathrm{tr}(\hat{\tilde{V}}^T VV^T\hat{\tilde{V}}) = \frac{1}{2}\|VV^T - \hat{V}\hat{V}^T\|_F^2$$

and Lemma 2, we have that

$$\mathrm{E}P_2^2 = O(\frac{\sqrt{p}}{n^2})\big(O(\frac{p}{n}) + \sigma^2\big) = o(\frac{p}{n^2}).$$

Next we deal with P_3. To prove the convergence in probability, we prove the convergence in L^1.

$$
\begin{aligned}
&\mathrm{E}|P_3| = \mathrm{E}\big|(\mu_1-\mu_2)^T(\hat{\tilde{V}}\hat{\tilde{V}}^T - \tilde{V}\tilde{V}^T)(\mu_1-\mu_2)\big| \le \|\mu_1-\mu_2\|^2\mathrm{E}\|\hat{\tilde{V}}\hat{\tilde{V}}^T - \tilde{V}\tilde{V}^T\| \\
=&\|\mu_1-\mu_2\|^2\mathrm{E}\|\hat{V}\hat{V}^T - VV^T\| \le \|\mu_1-\mu_2\|^2\sqrt{\mathrm{E}\|\hat{V}\hat{V}^T - VV^T\|^2} \\
\le&\|\mu_1-\mu_2\|^2\sqrt{\mathrm{E}\|\hat{V}\hat{V}^T - VV^T\|_F^2} = O(\frac{\sqrt{p}}{n})\sqrt{O(\frac{p}{p^\beta n})} = o(\frac{\sqrt{p}}{n}).
\end{aligned}
$$

Now we prove the asymptotic normality of P_1. To make clear the sense 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 $F_n \xrightarrow{\mathcal{L}} F$.

The conditional distribution of $\hat{\tilde{V}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)$ given S is $N(0, \tau\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})$. It can be seen that

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

where $\{\xi_i\}_{i=1}^{p-r}$ are i.i.d. standard normal random variables which are independent of $\hat{\tilde{V}}$. Note that

$$
\lambda_1(\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}}) \le \frac{1}{2}\kappa p^\beta\|VV^T - \hat{V}\hat{V}^T\|_F^2 + \sigma^2.
$$

Hence $\lambda_i(\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}}) = O_P(p/n+1)$, $i = 1,\ldots,r$. Moreover, by Weyl's inequality, $\lambda_i(\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}}) = \sigma^2$, $i = r+1,\ldots,p-r$. Therefore

$$
\mathrm{tr}(\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})^2 = \big(\frac{p}{n}+1\big)^2 O_P(1) + (p-2r)\sigma^4 = p\sigma^4(1+o_P(1)). \tag{18}
$$

It follows that

$$
\frac{\lambda_1^2(\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})}{\mathrm{tr}(\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})^2} = O_P\Big(\frac{(p/n+1)^2}{p}\Big) = o_P(1). \tag{19}
$$

Then for every subsequence of $\{n\}$, there's a further subsequence along which (19) holds almost surely. By Lemma 1, for every subsequence of $\{n\}$, there's a further subsequence along which

$$
\rho\big(\mathcal{L}(Z_n|S), N(0,1)\big) \xrightarrow{a.s.} 0, \tag{20}
$$

where

$$Z_n = \frac{\|\hat{\hat{V}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)\|^2 - \tau \mathrm{tr}(\hat{\hat{V}}^T\Sigma\hat{\hat{V}})}{\sqrt{2\tau^2 \mathrm{tr}(\hat{\hat{V}}^T\Sigma\hat{\hat{V}})^2}}.$$

By the definition of weak convergence, for every continuous bounded function $f(\cdot)$, we have $E[f(Z_n)|S] \xrightarrow{a.s.} E[f(\epsilon)]$ along the subsequence, where $\epsilon \sim N(0,1)$. By dominated convergence theorem, we have $E[f(Z_n)] \to E[f(\epsilon)]$ along the subsequence. Thus, $Z_n \xrightarrow{\mathcal{L}} N(0,1)$ along the subsequence, or equivalently, $\rho(\mathcal{L}(Z_n, N(0,1))) \to 0$ along the subsequence. But this means $Z_n \xrightarrow{\mathcal{L}} N(0,1)$, or

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

Similar to (18) we have

$$\mathrm{tr}(\hat{\hat{V}}^T\Sigma\hat{\hat{V}}) = (p-r)\sigma^2\big(1+O_P(\frac{1}{n}+\frac{1}{p})\big). \tag{21}$$

By (18), (21) and Slutsky's theorem,

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

Now the desired asymptotic properties of P_1, P_2 and P_3 are established, the theorem follows. □

Acknowledgements

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

References

Ahn SC, Horenstein AR. Eigenvalue ratio test for the number of factors. Econometrica 2013;81(3):1203–27. doi:10.3982/ECTA8968.

Bai Z, Saranadasa H. Effect of high dimension: by an example of a two sample problem. Statist Sinica 1996;6(2):311–29.

Bai ZD, Yin Y. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. Annals of Probability 1993;21(3):1275–94.

Cai TT, Liu W, Xia Y. Two-sample test of high dimensional means under dependence. J R Stat Soc Ser B Stat Methodol 2014;76(2):349–72. doi:10.1111/rssb.12034.

Cai TT, Ma Z, Wu Y. Sparse PCA: optimal rates and adaptive estimation. Ann Statist 2013;41(6):3074–110. doi:10.1214/13-AOS1178.

Chen LS, Paul D, Prentice RL, Wang P. A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. J Amer Statist Assoc 2011;106(496):1345–60. doi:10.1198/jasa.2011.ap10599.

Chen SX, Qin YL. A two-sample test for high-dimensional data with applications to gene-set testing. Ann Statist 2010;38(2):808–35. doi:10.1214/09-AOS716.

Durrett R. Probability : theory and examples. Journal of the American Statistical Association 2010;87(418):586.

Fan J, Wang W. Asymptotics of empirical eigen-structure for ultra-high dimensional spiked covariance model. Annals of Statistics 2017;.

Horn RA, Johnson CR. Matrix analysis. University of Albertadownloaded on December at from IEEE Xplorerestrictions Apply 1985;:1–17.

Jung S, Marron JS. Pca consistency in high dimension, low sample size context. Annals of Statistics 2009;37(6):4104–30.

Lehmann EL, Romano JP. Testing statistical hypotheses. 3rd ed. Springer Texts in Statistics. Springer, New York, 2005.

Lopes M, Jacob L, Wainwright MJ. A more powerful two-sample test in high dimensions using random projection. In: Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ, editors. Advances in Neural Information Processing Systems 24. Curran Associates, Inc.; 2011. p. 1206–14.

Ma Y, Lan W, Wang H. A high dimensional two-sample test under a low dimensional factor structure. J Multivariate Anal 2015;140:162–70. doi:10.1016/j.jmva.2015.05.005.

Shen D, Shen H, Marron JS. Consistency of sparse pca in high dimension, low sample size contexts. Journal of Multivariate Analysis 2013;115(1):317–33.

Srivastava MS, Du M. A test for the mean vector with fewer observations than the dimension. Journal of Multivariate Analysis 2008;99(3):386 – 402. doi:http://dx.doi.org/10.1016/j.jmva.2006.11.002.

Srivastava R, Li P, Ruppert D. Raptt: An exact two-sample test in high dimensions using random projections. Journal of Computational and Graphical Statistics 2015;doi:10.1080/10618600.2015.1062771.

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

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