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

*Corresponding author

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

1. Introduction

Suppose that $X_{k1}, \ldots, X_{kn_k}$ are independent identically distributed (i.i.d.) as $N_p(\mu_k, \Sigma_k)$, where μ_k and Σ_k are unknown, $k = 1, 2$. We consider the hypothesis testing problem:

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

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.

If $\Sigma_1 = \Sigma_2 = \Sigma$ is unknown, 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 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_1 and S_2 are sample covariance matrices. Statistic T_{CQ} is also an unbiased estimator of $\|\mu_1 - \mu_2\|^2$. Choosing $A = [\text{diag}(\Sigma)]^{-1}$, Srivastava and Du (2008) proposed a test based on

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

where $\text{diag}(A)$ 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

$$\text{tr}(\Sigma^4) = o[\text{tr}^2\{(\Sigma)^2\}] \tag{2}$$

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 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. Then Σ is spiked (see Cai 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, the model and some assumptions are given. In Section 3, we propose a test procedure under $\Sigma_1 = \Sigma_2$. Section 4 exploits properties of the test. In Section 5, we generalize our test procedure to the situation of $\Sigma_1 \neq \Sigma_2$. In Section 6, simulations are carried out and a real data example is given. Section 7 contains some discussion. All the technical details are in appendix.

2. Model and Assumptions

Let $\{X_{k1}, \ldots, X_{kn_k}\}$, $k = 1, 2$ be two independent random samples from p dimensional normal distribution with means μ_1 and μ_2 respectively.

Assumption 1. *Assume $p \to \infty$ as $n \to \infty$. Furthermore, assume two samples are balanced, that is,*

$$\frac{n_1}{n_2} \to \xi \in (0, +\infty).$$

To characterize correlations between p variables, we consider spiked covariance structure which is adopted by PCA study. See Cai et al. (2013) and the references given there.

Assumption 2. *Suppose X_{ki}, $i = 1, 2, \ldots, n_k$ and $k = 1, 2$ are generated by following model*

$$X_{ki} = \mu_k + V_k D_k U_{ki} + Z_{ki},$$

where U_{ki}'s are i.i.d. random vectors distributed as r_k dimensional standard normal distribution with r_k fixed, $D_k = diag(\lambda_{k1}^{\frac{1}{2}}, \ldots, \lambda_{kr_k}^{\frac{1}{2}})$ with $\lambda_{k1} \geq \cdots \geq \lambda_{kr_k} > 0$, V_k is a $p \times r_k$ orthonormal matrix, Z_{ki}'s are i.i.d. random vectors distributed as $N_p(0, \sigma_k^2 I_p)$ independent of U_{ki}'s and $\sigma_k^2 > 0$, $k = 1, 2$.

Then $X_{ki} \sim N(\mu_k, \Sigma_k)$, where $\Sigma_k = \mathrm{Var}(X_{ki}) = V_k \Lambda_k V_k^T + \sigma_k^2 I_p$, $\Lambda_k = D_k^2 = diag(\lambda_{k1}, \ldots, \lambda_{kr_k})$. From Assumption 2, $V_k V_k^T$ is the orthogonal projection matrix on the column space of V_k. Let $\tilde{V}_k$ be a $p \times (p - r_k)$ full column rank orthonormal matrix orthogonal to columns of V_k. Note that $\tilde{V}_k$ may not be unique. But the projection matrix $\tilde{V}_k \tilde{V}_k^T$ is unique because $\tilde{V}_k \tilde{V}_k^T = I - V_k V_k^T$.

Assumption 3. *Assume that there is some constant $\kappa > 0$ and $\beta \geq \frac{1}{2}$ such that*

$$\kappa p^\beta \geq \lambda_{k1} \geq \cdots \geq \lambda_{kr_k} \geq \kappa^{-1} p^\beta.$$

The restriction $\beta \geq 1/2$ breaks down the Condition (2). If $\beta < 1/2$, Condition (2) is meet and Chen and Qin (2010)'s method is valid. Hence $\beta = 1/2$ is the boundary of the scope between T_{CQ} and our new test. The case $\beta = 1$

corresponds to the factor model in paper Ma et al. (2015) with some restrictions of parameters.

Throughout the paper, let $\tau = (n_1 + n_2)/(n_1 n_2)$, S be the pooled sample covariance:

$$S = \frac{1}{n-2}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(X_{ki} - \bar{X}_k)(X_{ki} - \bar{X}_k)^T = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n-2},$$

where

$$S_k = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(X_{ki} - \bar{X}_k)(X_{ki} - \bar{X}_k)^T$$

is the sample covariance of the sample k, $k = 1, 2$.

Denote by $S = \hat{U}\hat{\Lambda}\hat{U}^T$ the eigen decomposition of S, where $\hat{U} = (\hat{u}_1, \ldots, \hat{u}_p)$ and $\hat{\Lambda} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$, $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$. Let $\hat{V} = (\hat{u}_1, \ldots, \hat{u}_r)$ and $\hat{\bar{V}} = (\hat{u}_{r+1}, \ldots, \hat{u}_p)$ be the first r and last $p - r$ eigenvectors of S respectively. Then $\hat{U} = (\hat{V}, \hat{\bar{V}})$. We similar define $\hat{U}_k$, $\hat{\Lambda}_k = \mathrm{diag}(\hat{\lambda}_{k1}, \ldots, \hat{\lambda}_{kp})$, $\hat{V}_k$ and $\hat{\bar{V}}_k$, $k = 1, 2$.

We write $\xi \sim \eta$ to denote the random variable ξ and η have the same distribution. For nonrandom positive sequence $\{a_n\}$ and $\{b_n\}$, $a_n \asymp b_n$ represents $a_n \geq cb_n$ and $a_n \leq Cb_n$ for some positive c, C for every n.

We denote by $\|\cdot\|$ and $\|\cdot\|_F$ the operator and Frobenius norm of matrix, separately.

For a symmetric matrix A, we define $\lambda_i(A)$ to be the ith largest eigenvalue of A and $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ to be the maximal and minimal eigenvalues respectively. We denote by $\mathrm{tr}(A)$ the trace of A. We use $A_{[a:b,:]}$ and $A_{[:,c:d]}$ to represent a-to-b-th rows and c-to-d-th columns of A.

The notations $\xrightarrow{P}$ and $\xrightarrow{\mathcal{L}}$ are used to denote convergence in probability and weak convergence respectively.

3. Test Statistic

In this section, we describe our new test procedure for hypotheses (1). For simplicity, we now work on equal covariance setting and unequal covariance setting will be considered latter.

Assumption 4. *Assume* $V_1 = V_2$, $D_1 = D_2$, $\Lambda_1 = \Lambda_2$, $\sigma_1 = \sigma_2$ *and* $r_1 = r_2$.

To simplify notations, the subscript k of Σ_k, V_k, D_k, Λ_k, σ_k and r_k are dropped.

In high dimensional setting, it is well-known that S is singular when $p \geq n-1$. As a result, Hotelling's T^2 statistic can not be defined. Since Σ has $p(p+1)/2$ independent parameters, it is hard to estimate when p is large compared with n. Therefore, for most recent work of high dimensional variance estimation, some additional assumptions, e.g. sparsity or low-rank, are adopted to regularize parameter space. For some recent development of this direction, see Fan et al. (2015a).

Some existing tests for hypothesis (1) can be regarded as the likelihood ratio test (LRT) under restricted covariance matrix. In fact, $\|\bar{X}_1 - \bar{X}_2\|^2$, the main body of both T_{BS} and T_{CQ}, is the LRT statistic assuming that $\Sigma = \sigma^2 I_p$ where σ^2 is unknown. And T_S is the LRT statistic assuming that Σ is a diagonal matrix with unknown diagonal elements. Although these methods are proved to be valid in more general setting, assumptions like (2) are often adopted. In many applications, assumption (2) may not be realistic due to the presence of common factors between variables. We derive a test statistic suitable for such applications.

Consider the following restriction for Σ

$$\lambda_{r+1}(\Sigma) = \cdots = \lambda_p(\Sigma) = \sigma^2, \tag{3}$$

where r is a known number and $\sigma^2 > 0$ is unknown. Under (3), Anderson (1986) proved that the maximum likelihood estimator (MLE) of Σ is $(n-2)n^{-1}\hat{\Sigma}$, where

$$\hat{\Sigma} = \sum_{i=1}^{r} \hat{\lambda}_i \hat{u}_i \hat{u}_i^T + \hat{\sigma}^2 \hat{\tilde{V}} \hat{\tilde{V}}^T,$$

and $\hat{\sigma}^2 = (p-r)^{-1} \sum_{i=r+1}^{p} \hat{\lambda}_i$.

Surprisingly, $\hat{\Sigma}$ is invertible even if $p \geq n-1$ and

$$\hat{\Sigma}^{-1} = \sum_{i=1}^{r} \hat{\lambda}_i^{-1} \hat{u}_i \hat{u}_i^T + \hat{\sigma}^{-2} \hat{\tilde{V}} \hat{\tilde{V}}^T.$$

Thus, the LRT for hypothesis (1) always exists. And LRT reject the null hypothesis when

$$T_{LRT} \overset{def}{=} \frac{1}{\tau} (\bar{X}_1 - \bar{X}_2)^T \hat{\Sigma}^{-1} (\bar{X}_1 - \bar{X}_2)$$

is large.

The distribution of Hotelling's T^2 test statistic doesn't depend on unknown parameter. Hence the critical value can be determined by exact distribution. However, even if (3) holds, the distribution of T_{LRT} still depends on parameters. The critical value may be defined by asymptotic distribution or randomization methods.

4. Theoretical results

In this section, we study the behavior of T_{LRT}.

Although T_{LRT} is derived under (3), we study the behavior of T_{LRT} in a more general setting.

Assumption 5. *Suppose Z_{ki}'s are i.i.d. random vectors with common distribution $N_p(0, I_p)$, $i = 1, \ldots, n_k$, $k = 1, 2$. Let $X_{ki} = U_k \Lambda_k^{1/2} Z_{ki}$, where $\Lambda_k = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_{k1} \geq \cdots \geq \lambda_{kp}$ and U_k is a $p \times p$ orthgonal matrix, $i = 1, \ldots, n_k$, $k = 1, 2$. Suppose $c \leq \lambda_{kp} \leq \lambda_{k, r_k+1} \leq C$, where $c > 0$ and $C > 0$ are absolute constants.*

Let V_k and $\tilde{V}_k$ be the first r_k columns and last $p - r_k$ columns of U_k respectively, $k = 1, 2$. Let $\Lambda_{k,(1)} = \operatorname{diag}(\lambda_{k1}, \ldots, \lambda_{kr_k})$ and $\Lambda_{k,(2)} = \operatorname{diag}(\lambda_{k,r+1}, \ldots, \lambda_{kp})$, $k = 1, 2$.

Write

$$T_{LRT} = \frac{1}{\tau} \sum_{i=1}^{r} \hat{\lambda}_i^{-1} \big(\hat{u}_i^T (\bar{X}_1 - \bar{X}_2)\big)^2 + \frac{1}{\tau \hat{\sigma}^2} \|\hat{\tilde{V}}^T (\bar{X}_1 - \bar{X}_2)\|^2 \stackrel{def}{=} T_1 + \hat{\sigma}^{-2} T_2.$$

The randomness of T_{LRT} is mainly due to T_2. T_1 can be regard as a power enhancement term, see Fan et al. (2015b). Hence we deal with T_1 and T_2 separately. The following theorem gives the asymptotic and nonasymptotic results for T_2.

Theorem 1. *Suppose the null hypothesis holds. Let r be fixed as n, p vary. Denote by $Q(1-\alpha)$ the $1-\alpha$ quantile of $T_2 - (p-r)\hat{\sigma}^2$. Then*

(a)

$$Q(1-\alpha) \lesssim \sqrt{p} + \frac{p}{n} + \frac{\lambda_1 \max(p, n)}{\max(n\lambda_r, \max(p, n))}.$$

(b) *Suppose $n, p \to \infty$ and*

$$\frac{p}{n^2} \to 0, \qquad \frac{\lambda_1 \max(p, n)}{\sqrt{p} \max\big(n\lambda_r, \max(p, n)\big)} \to 0. \tag{4}$$

then

$$\frac{T_2 - (p-r)\hat{\sigma}^2}{\sqrt{2 \sum_{i=r+1}^{p} \lambda_i^2}} \xrightarrow{\mathcal{L}} N(0, 1).$$

Remark 1. If $\lambda_1 \asymp \lambda_r$, then (4) is equivalent to $p/n^2 \to 0$. In this case, it can be shown that $p/n^2 \to 0$ is the necessary condition for asymptotic normality

Proof. Under the null hypothesis

$$\hat{\tilde{V}}^T(\bar{X}_1 - \bar{X}_2) = \hat{\tilde{V}}^T U \Lambda^{1/2}(\bar{Z}_1 - \bar{Z}_2).$$

Hence

$$T_2 = \frac{1}{\tau}\|\hat{\tilde{V}}^T U \Lambda^{1/2}(\bar{Z}_1 - \bar{Z}_2)\|^2.$$

By the property of normal random variables, $\{\bar{Z}_1, \bar{Z}_2\}$ are independent of S. Since $\hat{\tilde{V}}$ only depends on S, it is independent of $\{\bar{Z}_1, \bar{Z}_2\}$. It can be easily shown that

$$\begin{aligned}\frac{1}{\tau}\|\hat{\tilde{V}}^T U \Lambda^{1/2}(\bar{Z}_1 - \bar{Z}_2)\|^2 &\sim \sum_{i=1}^{p-r} \lambda_i(\hat{\tilde{V}}^T U \Lambda U^T \hat{\tilde{V}})\xi_i^2 \\ &= \sum_{i=1}^{r} \lambda_i(\hat{\tilde{V}}^T U \Lambda U^T \hat{\tilde{V}})\xi_i^2 + \sum_{i=r+1}^{p-r} \lambda_i(\hat{\tilde{V}}^T U \Lambda U^T \hat{\tilde{V}})\xi_i^2,\end{aligned} \tag{5}$$

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

First we deal with $\sum_{i=1}^{r} \lambda_i(\hat{\tilde{V}}^T U \Lambda U^T \hat{\tilde{V}})\xi_i^2$. By writting out

$$U\Lambda U^T = V\Lambda_{(1)}V^T + \tilde{V}\Lambda_{(2)}\tilde{V}^T,$$

we have matrix inequality

$$\lambda_r V V^T \leq U\Lambda U^T \leq \lambda_1 V V^T + C I_p.$$

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

$$\lambda_r \lambda_i(\hat{\tilde{V}}^T U U^T \hat{\tilde{V}}) \leq \lambda_i(\hat{\tilde{V}}^T U \Lambda U^T \hat{\tilde{V}}) \leq \lambda_1 \lambda_i(\hat{\tilde{V}}^T V V^T \hat{\tilde{V}}) + C.$$

Then $\sum_{i=1}^{r} \lambda_i(\hat{\tilde{V}}^T U \Lambda U^T \hat{\tilde{V}})\xi_i^2$ can be upper and lower bounded

$$\lambda_r \lambda_r(\hat{\tilde{V}}^T V V^T \hat{\tilde{V}}) \sum_{i=1}^{r} \xi_i^2 \leq \sum_{i=1}^{r} \lambda_i(\hat{\tilde{V}}^T U \Lambda U^T \hat{\tilde{V}})\xi_i^2 \leq (\lambda_1 \lambda_1(\hat{\tilde{V}}^T V V^T \hat{\tilde{V}}) + C) \sum_{i=1}^{r} \xi_i^2.$$

Now by Theorem 7 and the independence of $\hat{\hat{V}}$ and ξ_i's, with probability at least $1-3\exp(-t^2/2)$, we have

$$\sum_{i=1}^{r}\lambda_i(\hat{\hat{V}}^TU\Lambda U^T\hat{\hat{V}})\xi_i^2 \le \big(\frac{C\lambda_1}{n\hat{\lambda}_r}(\sqrt{p-r}+\sqrt{n-2}+t)^2+C\big)q_1(r,t), \tag{6}$$

where $q_1(r,t)$ is the $1-\exp(-t^2/2)$ quantile of χ_r^2. Theorem 6 can lower bound $\hat{\lambda}_r$ which appears in the right hand side of (6). More precisely, with probability at least $1-9\exp(-t^2/2)$, we have

$$\hat{\lambda}_r \ge \nu(t,\lambda_r,n,p), \tag{7}$$

where

$$\nu(t,\lambda_r,n,p) = \max\Big(\lambda_r\big(1-\frac{2}{\sqrt{n-2}}(\sqrt{r}+t)\big),$$
$$\frac{c\max(p-r,n-2)}{n-2}\big(1-\sqrt{\frac{2}{(p-r)(n-2)}}t-(r-1)\big(\frac{1}{\sqrt{p-r}}+\frac{1}{\sqrt{n-2}}+\frac{t}{\sqrt{(p-r)(n-2)}}\big)^2\big)\Big).$$

As a conclusion of (6) and (7), as $n,p\to\infty$,

$$\sum_{i=1}^{r}\lambda_i(\hat{\hat{V}}^TU\Lambda U^T\hat{\hat{V}})\xi_i^2 = O_P\Big(\frac{\lambda_1\max(p,n)}{\max\big(n\lambda_r,\max(p,n)\big)}\Big). \tag{8}$$

By (4) and (8),

$$\sum_{i=1}^{r}\lambda_i(\hat{\hat{V}}^TU\Lambda U^T\hat{\hat{V}})\xi_i^2 = o_p\Big(\sqrt{2\sum_{i=r+1}^{p}\lambda_i^2}\Big). \tag{9}$$

Turning to the second term of (5). Since $c\le\lambda_i\le C$, $i=r+1,\dots,p$. By Lemma 3, with probability at least $1-\exp(-t^2/2)$, we have

$$\sum_{i=r+1}^{p-r}\lambda_i(\hat{\hat{V}}^TU\Lambda U^T\hat{\hat{V}})\xi_i^2 \le \sum_{i=r+1}^{p-r}\lambda_i(\hat{\hat{V}}^TU\Lambda U^T\hat{\hat{V}})+\sqrt{2p}Ct+Ct^2. \tag{10}$$

For $i=r+1,\dots,p-r$, we have that

$$\lambda_i(\hat{\hat{V}}^TU\Lambda U^T\hat{\hat{V}}) = \lambda_i(\Lambda^{1/2}U^T\hat{\hat{V}}\hat{\hat{V}}^TU\Lambda^{1/2}) = \lambda_i(\Lambda-\Lambda^{1/2}U^T\hat{V}\hat{V}^TU\Lambda^{1/2}).$$

Since $\mathrm{Rank}(\Lambda^{1/2}U^T\hat{V}\hat{V}^TU\Lambda^{1/2})=r$, by Weyl's inequality, $\lambda_i(\hat{\hat{V}}^TU\Lambda U^T\hat{\hat{V}})$ can be bounded by

$$\lambda_{i+r}\le\lambda_i(\hat{\hat{V}}^TU\Lambda U^T\hat{\hat{V}})\le\lambda_i. \tag{11}$$

By (11), $\sum_{i=r+1}^{p}\lambda_i - rC \le \sum_{i=r+1}^{p-r}\lambda_i(\hat{\tilde{V}}^T U\Lambda U^T\hat{\tilde{V}}) \le \sum_{i=r+1}^{p}\lambda_i$, then (10) implies

$$\sum_{i=r+1}^{p-r}\lambda_i(\hat{\tilde{V}}^T U\Lambda U^T\hat{\tilde{V}})\xi_i^2 \le \sum_{i=r+1}^{p}\lambda_i + \sqrt{2p}Ct + Ct^2, \tag{12}$$

As another conclusion of (11),

$$\sum_{i=r+1}^{p-r}\lambda_{i+r}\xi_i^2 \le \sum_{i=r+1}^{p-r}\lambda_i(\hat{\tilde{V}}^T U\Lambda U^T\hat{\tilde{V}})\xi_i^2 \le \sum_{i=r+1}^{p-r}\lambda_i\xi_i^2.$$

By Lyapunov's central limit theorem, as $p\to\infty$,

$$\frac{\sum_{i=r+1}^{p-r}\lambda_i(\hat{\tilde{V}}^T U\Lambda U^T\hat{\tilde{V}})\xi_i^2 - \sum_{i=r+1}^{p}\lambda_i}{\sqrt{2\sum_{i=r+1}^{p}\lambda_i^2}} \xrightarrow{\mathcal{L}} N(0,1).$$

Together with (9), we can deduce that under (9),

$$\frac{\tau^{-1}\|\hat{\tilde{V}}^T U\Lambda^{1/2}(\bar{Z}_1-\bar{Z}_2)\|^2 - \sum_{i=r+1}^{p}\lambda_i}{\sqrt{2\sum_{i=r+1}^{p}\lambda_i^2}} \xrightarrow{\mathcal{L}} N(0,1). \tag{13}$$

Next we deal with $\hat{\sigma}^2$. By Theorem 8, with probability at least $1-4\exp(-t^2/2)$, we have

$$\Big|\frac{\hat{\sigma}^2}{(p-r)^{-1}\sum_{i=r+1}^{p}\lambda_i}-1\Big| \le \frac{1}{(n-2)\sum_{i=r+1}^{p}\lambda_i}\big(rC(\sqrt{p-r}+\sqrt{n-2}+t)^2+\sqrt{2(n-2)p}Ct+Ct^2\big). \tag{14}$$

It follows that

$$\hat{\sigma}^2 = \frac{\sum_{i=r+1}^{p}\lambda_i}{p-r} + O_p\Big(\frac{1}{\min(n,p)}\Big). \tag{15}$$

By (13), (15), (4) and Slutsky's theorem, we have

$$\begin{aligned}&\frac{\tau^{-1}\|\hat{\tilde{V}}^T U\Lambda^{1/2}(\bar{Z}_1-\bar{Z}_2)\|^2 - (p-r)\hat{\sigma}^2}{\sqrt{2\sum_{i=r+1}^{p}\lambda_i^2}}\\ =&\frac{\tau^{-1}\|\hat{\tilde{V}}^T U\Lambda^{1/2}(\bar{Z}_1-\bar{Z}_2)\|^2 - \sum_{i=r+1}^{p}\lambda_i}{\sqrt{2\sum_{i=r+1}^{p}\lambda_i^2}} + O_p\Big(\frac{\sqrt{p}}{\min(n,p)}\Big) \xrightarrow{\mathcal{L}} N(0,1).\end{aligned} \tag{16}$$

Conclusion (b) is established.

With probability at least $1-17\exp(-t^2/2)$, inequality (6), (7), (12) and (14) are satisfied simultaneously. Choose t^* such that $17\exp(-t^{*2}/2)=\alpha$. Then (6), (7), (12) and (14) imply that

$$Q(1-\alpha) \le \sum_{i=r+1}^{p} \lambda_i + \sqrt{2p}Ct^* + Ct^{*2} + \big(\frac{C\lambda_1}{n\nu(t^*, \lambda_r, n, p)}(\sqrt{p-r} + \sqrt{n-2} + t^*)^2 + C\big)q_1(r, t^*)$$
$$- \Big(\sum_{i=r+1}^{p} \lambda_i - \frac{1}{n-2}\big(rC(\sqrt{p-r} + \sqrt{n-2} + t)^2 + \sqrt{2(n-2)p}Ct + Ct^2\big)\Big)$$
$$\asymp \sqrt{p} + \frac{p}{n} + \frac{\lambda_1 \max(p,n)}{n\nu(t^*, \lambda_r, n, p)} \asymp \sqrt{p} + \frac{p}{n} + \frac{\lambda_1 \max(p,n)}{\max(n\lambda_r, \max(p,n))}.$$

Conclusion (a) is established.

□

Theorem 2. *Let r be fixed as $n, p \to \infty$. Suppose $np^{-1/2}\|\mu_1 - \mu_2\|^2 = O(1)$, $\lambda_1/\sqrt{p} \to \infty$ and (4) holds. Then*

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

Proof. Since

$$\hat{\tilde{V}}^T(\bar{X}_1 - \bar{X}_2) = \hat{\tilde{V}}^T U\Lambda^{1/2}(\bar{Z}_1 - \bar{Z}_2) + \hat{\tilde{V}}^T(\mu_1 - \mu_2),$$

T_2 can be written as the sum of three terms

$$T_2 = \tau^{-1}\|\hat{\tilde{V}}^T U\Lambda^{1/2}(\bar{Z}_1 - \bar{Z}_2)\|^2 + 2\tau^{-1}(\mu_1 - \mu_2)^T\hat{\tilde{V}}\hat{\tilde{V}}^T U\Lambda^{1/2}(\bar{Z}_1 - \bar{Z}_2) + \tau^{-1}\|\hat{\tilde{V}}^T(\mu_1 - \mu_2)\|^2.$$

The first term has normal distribution asymptotically. We now deal with the second term. Note that

$$\tau^{-1}(\mu_1 - \mu_2)^T\hat{\tilde{V}}\hat{\tilde{V}}^T U\Lambda^{1/2}(\bar{Z}_1 - \bar{Z}_2) \sim \sqrt{\tau^{-1}(\mu_1 - \mu_2)^T\hat{\tilde{V}}\hat{\tilde{V}}^T U\Lambda U^T\hat{\tilde{V}}\hat{\tilde{V}}^T(\mu_1 - \mu_2)}\xi,$$

where ξ is a standard normal variable and is independent of $\hat{\tilde{V}}$. Similar to the prove of (9),

$$\sqrt{\tau^{-1}(\mu_1 - \mu_2)^T\hat{\tilde{V}}\hat{\tilde{V}}^T U\Lambda U^T\hat{\tilde{V}}\hat{\tilde{V}}^T(\mu_1 - \mu_2)} \le \sqrt{\tau^{-1}}\|\mu_1 - \mu_2\|\sqrt{\lambda_{\max}(\hat{\tilde{V}}^T U\Lambda U^T\hat{\tilde{V}})}$$
$$\le \sqrt{\tau^{-1}}\|\mu_1 - \mu_2\|\sqrt{\lambda_1\lambda_{\max}(\hat{\tilde{V}}^T V V^T\hat{\tilde{V}}) + C} = o_P(n^{1/2}p^{1/4}\|\mu_1 - \mu_2\|) = o_P(\sqrt{p}). \tag{17}$$

Next we deal with $\tau^{-1}\|\hat{\tilde{V}}^T(\mu_1-\mu_2)\|^2$. Note that

$$\begin{aligned}
&\left|\tau^{-1}\|\hat{\tilde{V}}^T(\mu_1-\mu_2)\|^2-\tau^{-1}\|\tilde{V}^T(\mu_1-\mu_2)\|^2\right|\\
=&\tau^{-1}\left|(\mu_1-\mu_2)^T(\hat{\tilde{V}}\hat{\tilde{V}}^T-\tilde{V}\tilde{V}^T)(\mu_1-\mu_2)\right|\\
\leq&\tau^{-1}\|\hat{\tilde{V}}\hat{\tilde{V}}^T-\tilde{V}\tilde{V}^T\|\|\mu_1-\mu_2\|^2=\tau^{-1}\|\hat{V}\hat{V}^T-VV^T\|\|\mu_1-\mu_2\|^2 \qquad (18)\\
\leq&\tau^{-1}\|\hat{V}\hat{V}^T-VV^T\|_F\|\mu_1-\mu_2\|^2=\tau^{-1}\sqrt{2\operatorname{tr}(\tilde{V}^T\hat{V}\hat{V}^T\tilde{V})}\|\mu_1-\mu_2\|^2\\
\leq&\tau^{-1}\sqrt{2r\lambda_{\max}(\tilde{V}^T\hat{V}\hat{V}^T\tilde{V})}\|\mu_1-\mu_2\|^2.
\end{aligned}$$

Condition $\lambda_1/\sqrt{p}\to\infty$ and (4) require

$$\frac{\max(p,n)}{\max\big(n\lambda_r,\max(p,n)\big)}\to 0,$$

which in turn implies $\lambda_{\max}(\tilde{V}^T\hat{V}\hat{V}^T\tilde{V})\to 0$ in probability by Theorem 7. Then (18) implies

$$\tau^{-1}\|\hat{\tilde{V}}^T(\mu_1-\mu_2)\|^2=\tau^{-1}\|\tilde{V}^T(\mu_1-\mu_2)\|^2+o_P(\sqrt{p}). \qquad (19)$$

The theorem follows by Theorem 1, (17), (19) and Slutsky's theorem.

□

The asymptotic normality of T_2 requires condition (4), which is strong. However, it may not be able to be relaxed. For simplicity, suppose $\lambda_1\asymp\lambda_r$. Then (4) is equivalent to $p/n^2\to 0$. The asymptotic normality of $\|\hat{\tilde{V}}^T(\bar{X}_1-\bar{X}_2)\|^2$, the main major part of T_2, requires

$$\frac{\lambda_1\big((\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})^2\big)}{\operatorname{tr}\big((\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})^2\big)}\xrightarrow{P}0. \qquad (20)$$

See Lemma 4 in appendix. And (20) is equivalent to $p/n^2\to 0$ by Lemma **??** in appendix.

By Proposition **??** and Theorem **??**, the power function of the new test can be obtained immediately.

Corollary 1. *Under Assumptions 1-**??**, if we reject the null hypothesis when Q is larger than $1-\alpha$ quantile of $N(0,1)$, then the asymptotic power* *function of the new test is*

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

Note that the power of T_{CQ} is of the form

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

The relative efficiency of our test with respect to Chen's test is

$$\sqrt{\frac{\mathrm{tr}\Sigma^2}{(p-r)\sigma^4}}\frac{\|\tilde{V}(\mu_1 - \mu_2)\|^2}{\|\mu_1 - \mu_2\|^2} \sim p^{\beta - 1/2}\frac{\|\tilde{V}(\mu_1 - \mu_2)\|^2}{\|\mu_1 - \mu_2\|^2},$$

which is large when $\beta > 1/2$ and $\|\tilde{V}(\mu_1 - \mu_2)\|/\|\mu_1 - \mu_2\|$ is close to 1.

However, this does not mean the new statistic can not be used. In fact, since the samples are exchangeable under null hypothesis, we can always use permutation method to determine the critical value. We will see from simulation results that the new test has good power behavior even in large p small n case.

In practice, it may not be an easy task to check if the covariance matrices are spiked, especially in high dimension setting. When the spiked covariance model is not valid, some estimators in our test procedure make no sense. In particular, if $\hat{r}$ is estimated by (**??**). the $\hat{r}$ is nothing but a random integer not greater than R and $\hat{V}\hat{V}^T$ is just a random projection. Hence it is a natural question how the new test procedure behaves when the spiked covariance model breaks down. We study the asymptotic behavior of the new test procedure in two non-spiked setting.

First we consider the case when the eigenvalues of Σ is bounded. Similar to bayesian models, we assume a normal prior distribution for μ_k to characterize 'dense' alternative. The next theorem shows that the power of our new test is asymptotically the same as Chen and Qin (2010)'s test in this case.

Theorem 3. *Assume $X_{ki} \sim N(\mu_k, \Sigma)$, $i = 1, \ldots, n_k$, $k = 1, 2$. Suppose that Assumptions 1 and* **??** *holds, $0 < c \le \lambda_p(\Sigma) \le \lambda_1(\Sigma) \le C < \infty$ where c and C are constants, each element of μ_k is independently generated by*

$N(0, (n_k\sqrt{p})^{-1}\psi)$ for $k = 1, 2$, where ψ is a constant and $\hat{r} \leq R$ for a positive constant R. Then we have

$$\frac{T_2 - \|\mu_1 - \mu_2\|^2}{\sqrt{2\tau^2 \mathrm{tr}\Sigma^2}} \xrightarrow{\mathcal{L}} N(0, 1).$$

The second setting we consider is the model in Assumption 2 with $r = 0$. In this case, the Assumption **??** can be dropped and we don't need to assume a random μ_k.

Theorem 4. *Under Assumptions 1-4 with factor number $r = 0$, if*

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

and $\hat{r} \leq R$ for a positive constant R, then

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

These results show that the new test procedure is robust against the invalidity of spiked covariance model.

5. Unequal Variance

In this section, we concern the situation with unequal covariance matrices. Assume $\{X_{11}, \ldots, X_{1n_1}\}$ and $\{X_{21}, \ldots, X_{2n_2}\}$ are both generated from the model in Assumption 2. Denote by $\hat{V}_k$ the first r_k eigenvectors of S_k for $k = 1, 2$. With a little abuse of notation, let VV^T be the projection on the sum of column spaces of V_1 and V_2, that is,

$$VV^T = (V_1, V_2)\big((V_1, V_2)^T(V_1, V_2)\big)^+(V_1, V_2)^T.$$

where A^+ is the Moore-Penrose inverse of a matrix A. Similarly, let $\hat{V}\hat{V}^T$ be the projection matrix on the sum of column spaces of $\hat{V}_1$ and $\hat{V}_2$. We define $\tilde{V}\tilde{V}^T = I_p - VV^T$ and $\hat{\tilde{V}}\hat{\tilde{V}}^T = I_p - \hat{V}\hat{V}^T$.

The previous statistic can not be directly used since the principal subspace is different for X_{1i} and X_{2j}. The idea here is to remove all large variance terms from T_{CQ} by projecting data on the space $\tilde{V}\tilde{V}^T$. Thus, we propose a new test statistic as

$$T_3 = \|\hat{\tilde{V}}^T(\bar{X}_1 - \bar{X}_2)\|^2 - \frac{1}{n_1}\mathrm{tr}(\hat{\tilde{V}}_1^T S_1 \hat{\tilde{V}}_1) - \frac{1}{n_2}\mathrm{tr}(\hat{\tilde{V}}_2^T S_2 \hat{\tilde{V}}_2).$$

The theoretical results are parallel to those in equal variance setting.

Theorem 5. *Under Assumptions 1-3 and* **??**, *if*

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

then we have

$$\frac{T_3 - \|\hat{\tilde{V}}^T(\mu_1 - \mu_2)\|^2}{\sqrt{\sigma_n^2}} \xrightarrow{\mathcal{L}} N(0, 1).$$

where $\sigma_n^2 = \frac{2(p-r_1-r_2)}{n_1(n_1-1)}\sigma_1^4 + \frac{2(p-r_1-r_2)}{n_2(n_2-1)}\sigma_2^4 + \frac{4(p-r_1-r_2)}{n_1 n_2}\sigma_1^2\sigma_2^2$.

Remark 2. Even if $\hat{\tilde{V}}_k\hat{\tilde{V}}_k^T$ is an consistent estimator of $\tilde{V}_k\tilde{V}_k^T$ for $k = 1, 2$, $\hat{\tilde{V}}\hat{\tilde{V}}^T$ may not be an consistent estimator of $\tilde{V}\tilde{V}^T$. Nevertheless, the asymptotic normality still holds. However, the centering term should be $\|\hat{\tilde{V}}^T(\mu_1 - \mu_2)\|^2$ and can not be replaced by $\|\tilde{V}^T(\mu_1 - \mu_2)\|^2$.

σ_n^2 can be estimated by ratio consistent estimators of σ_k^2 for $k = 1, 2$. Thus, if n and p are large and $\sqrt{p}/n$ is small, we reject when $T_3/\sqrt{\hat{\sigma}_n^2} > z_{1-\alpha}$.

6. 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 removes big variance terms from T_{CQ} and it's power is boosted substantially. The asymptotic normality of the new statistic is proved and the asymptotic power is given.

In another paper, Zhao and Xu (2016) proved 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, our work shows that in some circumstance, the complement of principal subspace is more useful.

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

We first collect a few useful technical lemmas in Section 6.1. Then we give the proof of Theorem **??**. We develop a non-asymptotic theory for PCA in. HH.

6.1. Technical Lemmas

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

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

Corollary 2. *Let H and P be two symmetric matrices and $M = H + P$. If $\operatorname{rank}(P) < k$, then*

$$\lambda_k(M) \leq \lambda_1(H).$$

Lemma 2 (Davidson-Szarek bound)**.** *Let Z be an $N \times n$ ($N \geq n$) matrix whose entries are independent standard normal random variables. Then for every $t > 0$, with probability at least $1 - 2\exp(-t^2/2)$ one has*

$$\left(\max(\sqrt{N} - \sqrt{n} - t, 0)\right)^2 \leq \lambda_n(ZZ^T) \leq \lambda_1(ZZ^T) \leq (\sqrt{N} + \sqrt{n} + t)^2.$$

See Davidson and Szarek (2001).

Lemma 3. *Let $Y_1, \ldots, Y_n$ be i.i.d. standard normal random variables. Let $a_1, \ldots, a_n$ be nonnegative. We set*

$$|a|_\infty = \sup_{i=1,\ldots,n} |a_i|, \quad |a|_2^2 = \sum_{i=1}^n a_i^2.$$

Then, for every $t > 0$, with probability at least $1 - \exp(-t^2/2)$, one has

$$\sum_{i=1}^n a_i Y_i^2 \leq \sum_{i=1}^n a_i + \sqrt{2}|a|_2 t + |a|_\infty t^2.$$

On the other hand, with probability at least $1 - \exp(-t^2/2)$, one has

$$\sum_{i=1}^n a_i Y_i^2 \geq \sum_{i=1}^n a_i - \sqrt{2}|a|_2 t.$$

See Laurent and Massart (2000).

Lemma 4. *Suppose X_n is a k_n dimensional standard normal random vector and A_n is a $k_n \times k_n$ symmetric matrix. Then a necessary and sufficient condition for*

$$\frac{X_n^T A_n X_n - \mathrm{E} X_n^T A_n X_n}{[\mathrm{Var}(X_n^T A_n X_n)]^{1/2}} \xrightarrow{\mathcal{L}} N(0,1) \tag{21}$$

is that

$$\frac{\lambda_{\max}(A_n^2)}{\mathrm{tr}(A_n^2)} \to 0. \tag{22}$$

Remark 3. This lemma is from the Example 5.1 of Jiang (1996). Here we give a proof by characteristic function.

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

$$\frac{X_n^T A_n X_n - \mathrm{E} X_n^T A_n X_n}{[\mathrm{Var}(X_n^T A_n X_n)]^{1/2}} = \sum_{i=1}^{k_n} \frac{\lambda_i(A_n)}{\left[2\mathrm{tr}(A_n^2)\right]^{1/2}} (Z_{ni}^2 - 1), \tag{23}$$

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

If 22 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 21 follows by Lindeberg's central limit theorem.

Conversely, if 21 holds, we will prove that there is a subsequence of $\{n\}$ along which 22 holds. Then 22 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 21 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

$$\log \mathrm{E} \exp\big(it\sum_{j=1}^{k_n} c_{nj}(Z_{nj}^2-1)\big) = -i(\sum_{j=1}^{k_n} c_{nj})t - \frac{1}{2}\sum_{j=1}^{k_n}\log(1-i2c_{nj}t)$$

$$= -i(\sum_{j=1}^{k_n} c_{nj})t + \frac{1}{2}\sum_{j=1}^{k_n}\sum_{l=1}^{+\infty}\frac{1}{l}(i2c_{nj}t)^l = -i(\sum_{j=1}^{k_n} c_{nj})t + \frac{1}{2}\sum_{l=1}^{+\infty}\Big[\sum_{j=1}^{k_n}(c_{nj})^l\Big]\frac{1}{l}(i2t)^l$$

$$= -\frac{1}{2}t^2 + \frac{1}{2}\sum_{l=3}^{+\infty}\Big[\sum_{j=1}^{k_n}(c_{nj})^l\Big]\frac{1}{l}(i2t)^l.$$

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 22 follows by noting that $b_{n4} \geq \max_j (c_{nj})^4$. $\square$

6.2. PCA Theory

In this section, we give some non-asymptotic theory of PCA. Since our main task is to obtain the properties of principal space, we impose less assumptions compared with existing results.

Lemma 5. *Under the assumption of Lemma 2, for every $t > 0$, with probability at least $1 - 3\exp(-t^2/2)$, for every i such that $1 \le i \le n$, we have*

$$\lambda_i(ZZ^T) \ge N\Big(1 - \sqrt{\frac{2}{Nn}}t - (i-1)(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{n}} + \frac{t}{\sqrt{Nn}})^2\Big).$$

Proof. $\mathrm{tr}(ZZ^T)$ is distributed as χ^2_{Nn}. By Lemma 3, for every $t > 0$, with probability at least $1 - \exp(-t^2/2)$, we have that

$$\sum_{j=1}^{i-1} \lambda_j(ZZ^T) + \sum_{j=i}^{n} \lambda_j(ZZ^T) = \mathrm{tr}(ZZ^T) \ge Nn(1 - \sqrt{\frac{2}{Nn}}t),$$

where $1 \le i \le n$. Thus, with probability at least $1 - \exp(-t^2/2)$, for every i such that $1 \le i \le n$, we have

$$\begin{aligned}
\lambda_i(ZZ^T) &\ge \frac{1}{n}\sum_{j=i}^{n} \lambda_j(ZZ^T) \\
&\ge \frac{1}{n}\Big(Nn(1 - \sqrt{\frac{2}{Nn}}t) - \sum_{j=1}^{i-1} \lambda_j(ZZ^T)\Big) \\
&\ge \frac{1}{n}\Big(Nn(1 - \sqrt{\frac{2}{Nn}}t) - (i-1)\lambda_1(ZZ^T)\Big).
\end{aligned}$$

By the above inequality and Lemma 2, with probability at least $1-3\exp(-t^2/2)$, for every i such that $1 \le i \le n$, we have

$$\begin{aligned}
\lambda_i(ZZ^T) &\ge \frac{1}{n}\Big(Nn(1 - \sqrt{\frac{2}{Nn}}t) - (i-1)(\sqrt{N} + \sqrt{n} + t)^2\Big) \\
&= N\Big(1 - \sqrt{\frac{2}{Nn}}t - (i-1)(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{n}} + \frac{t}{\sqrt{Nn}})^2\Big).
\end{aligned}$$

□

Assumption 6. *Suppose that $Z = (Z_1, \ldots, Z_n)$ is an $p \times n$ random matrix whose entries Z_{ij}'s are i.i.d. standard normal random variables, $i = 1, \ldots, p$,*

$j = 1, \dots, n$. *Let the sample matrix be* $X = (X_1, \dots, X_n) = U\Lambda^{1/2}Z$, *where* $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$ *with* $\lambda_1 \geq \cdots \geq \lambda_p$ *and* U *is a* $p \times p$ *orthogonal matrix. Suppose* $c \leq \lambda_p \leq \lambda_{r+1} \leq C$, *where* $c > 0$ *and* $C > 0$ *are absolute constants.*

The sample covariance matrix is $\frac{1}{n}XX^T$. We denote by $\frac{1}{n}XX^T = \hat{U}\hat{\Lambda}\hat{U}^T$ the spectral decomposition of S where $\hat{\Lambda} = \operatorname{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_p)$ and $\hat{U}$ is a orthogonal matrix.

Let u_i be the ith column of U, $i = 1, \dots, p$. Denote $U = (V, \tilde{V})$, where V and $\tilde{V}$ are the first r and last $p - r$ columns of U respectively. Similarly, we define the corresponding part of $\hat{U}$ by $\hat{u}_i$, $\hat{V}$ and $\hat{\tilde{V}}$.

The PCA theory is mainly focus on the convergence properties of $\hat{u}_i$ to it's population counterpart u_i. See Jung and Marron (2009), Shen et al. (2012), Shen et al. (2013) and Fan and Wang (2017) for some recent developements for PCA theory. Here we are mainly interested in the asymptotic properties of $\hat{V}$. Compared to existing results, the consistency results of $\hat{V}$ require less assumptions on the order of $\lambda_1, \dots, \lambda_r$.

Let $Z_{(1)}$ and $Z_{(2)}$ be the first r rows and the last $p-r$ rows of Z respectively. Then $Z_{(1)}$ is an $r \times n$ matrix and $Z_{(2)}$ is an $(p - r) \times n$ matrix. Let $\Lambda_{(1)} = \operatorname{diag}(\lambda_1, \dots, \lambda_r)$ and $\Lambda_{(2)} = \operatorname{diag}(\lambda_{r+1}, \dots, \lambda_p)$. Define $\hat{\Lambda}_{(1)}$ and $\hat{\Lambda}_{(2)}$ in a similar way.

Theorem 6. *Suppose Assumption 6 holds. Let* i *be a fixed number such that* $1 \leq i \leq r$. *Then for every* $t > 0$, *with probability at least* $1 - 9\exp(-t^2/2)$, *we have*

$$\frac{\hat{\lambda}_i}{\lambda_i} \geq \max\Big(1 - \frac{2}{\sqrt{n}}(\sqrt{r} + t),$$
$$\frac{c\max(p-r, n)}{n\lambda_i}\big(1 - \sqrt{\frac{2}{(p-r)n}}t - (i-1)\big(\frac{1}{\sqrt{p-r}} + \frac{1}{\sqrt{n}} + \frac{t}{\sqrt{(p-r)n}}\big)^2\big)\Big),$$

and

$$\frac{\hat{\lambda}_i}{\lambda_i} \leq 1 + \frac{2}{\sqrt{n}}(\sqrt{r} + t) + \frac{1}{n}(\sqrt{r} + t)^2 + \frac{C}{n\lambda_i}(\sqrt{p-r} + \sqrt{n} + t)^2.$$

Proof. The non-zero eigenvalues of $\frac{1}{n}XX^T$ are equal to that of $\frac{1}{n}X^TX$. And

$\frac{1}{n}X^TX$ can be further written as the sum of two quantities

$$\frac{1}{n}X^TX = \frac{1}{n}Z^T\Lambda Z = \frac{1}{n}Z_{(1)}^T\Lambda_{(1)}Z_{(1)} + \frac{1}{n}Z_{(2)}^T\Lambda_{(2)}Z_{(2)} \overset{def}{=} A + B.$$

By Weyl's inequality,

$$\frac{\max(\lambda_i(A), \lambda_i(B))}{\lambda_i} \le \frac{\hat{\lambda}_i}{\lambda_i} \le \frac{\lambda_i(A)}{\lambda_i} + \frac{\lambda_{\max}(B)}{\lambda_i}, \tag{24}$$

where $i = 1, \ldots, r$. We deal with $\lambda_i(A)$ and $\lambda_i(B)$ separately.

First we deal with $\lambda_i(A)$, $i = 1, \ldots, r$. By Corollary 2, we have that

$$\frac{\lambda_i(A)}{\lambda_i} \le \frac{1}{n\lambda_i}\lambda_{\max}\big(Z_{(1)}^T \operatorname{diag}(\underbrace{0, \ldots, 0}_{i-1}, \underbrace{\lambda_i, \ldots, \lambda_i}_{r-i+1})Z_{(1)}\big) = \frac{1}{n}\lambda_{\max}(Z_{[i:r,:]}^T Z_{[i:r,:]}).$$

Then by Lemma 2, with probability at least $1 - 2\exp(-t^2/2)$ we have

$$\frac{\lambda_i(A)}{\lambda_i} \le \frac{1}{n}(\sqrt{n} + \sqrt{r-i+1} + t)^2 \le 1 + \frac{2}{\sqrt{n}}(\sqrt{r} + t) + \frac{1}{n}(\sqrt{r} + t)^2. \tag{25}$$

On the other hand, by Weyl's inequility, we have that

$$\frac{\lambda_i(A)}{\lambda_i} \ge \frac{1}{n\lambda_i}\lambda_i\big(Z_{(1)}^T \operatorname{diag}(\underbrace{\lambda_i, \ldots, \lambda_i}_{i}, \underbrace{0, \ldots, 0}_{r-i})Z_{(1)}\big) = \frac{1}{n}\lambda_{\max}(Z_{[1:i,:]}^T Z_{[1:i,:]}).$$

Again by Lemma 2, with probability at least $1 - 2\exp(-t^2/2)$ we have

$$\frac{\lambda_i(A)}{\lambda_i} \ge \frac{1}{n}\big(\max(\sqrt{n} - \sqrt{i} - t, 0)\big)^2 \ge 1 - \frac{2}{\sqrt{n}}(\sqrt{r} + t). \tag{26}$$

Now we deal with $\lambda_i(B)$. Since $\lambda_i(B) \ge \frac{c}{n}\lambda_i(Z_{(2)}^T Z_{(2)})$, by Lemma 5, with probability at least $1 - 3\exp(-t^2/2)$ we have

$$\lambda_i(B) \ge \frac{c\max(p-r, n)}{n}\Big(1 - \sqrt{\frac{2}{(p-r)n}}t - (i-1)(\frac{1}{\sqrt{p-r}} + \frac{1}{\sqrt{n}} + \frac{t}{\sqrt{(p-r)n}})^2\Big). \tag{27}$$

Since $\lambda_1(B) \le \frac{C}{n}\lambda_1(Z_{(2)}^T Z_{(2)})$, by Lemma 2, with probability at least $1 - 2\exp(-t^2/2)$ we have

$$\lambda_{\max}(B) \le \frac{C}{n}(\sqrt{p-r} + \sqrt{n} + t)^2. \tag{28}$$

The theorem follows by (24), (25), (26), (27), and (28). $\square$

Theorem 7. *Suppose Assumption 6 holds. For every $t > 0$, with probability at least $1 - 2\exp(-t^2/2)$, we have*

$$\lambda_1(\tilde{V}^T \hat{V} \hat{V}^T \tilde{V}) \leq \frac{C}{n\hat{\lambda}_r}(\sqrt{p-r} + \sqrt{n} + t)^2. \tag{29}$$

With probability at least $1 - \exp(-t^2/2)$, we have

$$r - \frac{1}{2}\|\hat{V}\hat{V}^T - VV^T\|_F^2 \leq \frac{\lambda_1}{\hat{\lambda}_r}(r + \sqrt{\frac{2r}{n}}t + \frac{1}{n}t^2). \tag{30}$$

If we further assume $p - r \geq n$, then with probability at least $1 - 2\exp(-t^2/2)$, we have

$$\lambda_r(\tilde{V}^T \hat{V} \hat{V}^T \tilde{V}) \geq \frac{c}{n\hat{\lambda}_1}(\max(\sqrt{p-r} - \sqrt{n} - t, 0))^2. \tag{31}$$

Proof. Since

$$\frac{1}{n}XX^T = \hat{U}\hat{\Lambda}\hat{U}^T = \frac{1}{n}U\Lambda^{1/2}ZZ^T\Lambda^{1/2}U^T,$$

we have

$$\Lambda^{-1/2}U^T\hat{U}\hat{\Lambda}\hat{U}^TU\Lambda^{-1/2} = \frac{1}{n}ZZ^T. \tag{32}$$

We first prove (29). It follows from (32) that

$$\Lambda_{(2)}^{-1/2}\tilde{V}^T\hat{U}\hat{\Lambda}\hat{U}^T\tilde{V}\Lambda_{(2)}^{-1/2} = \frac{1}{n}Z_{(2)}Z_{(2)}^T. \tag{33}$$

The left hand side of (33) equals to $A+B$, where $A = \Lambda_{(2)}^{-1/2}\tilde{V}^T\hat{V}\hat{\Lambda}_{(1)}\hat{V}^T\tilde{V}\Lambda_{(2)}^{-1/2}$ and $B = \Lambda_{(2)}^{-1/2}\tilde{V}^T\hat{\tilde{V}}\hat{\Lambda}_{(2)}\hat{\tilde{V}}^T\tilde{V}\Lambda_{(2)}^{-1/2}$.

Obviously, we have $\lambda_1(A) \leq n^{-1}\lambda_1(Z_{(2)}Z_{(2)}^T)$ and $\lambda_1(A) \geq C^{-1}\hat{\lambda}_r\lambda_1(\tilde{V}^T\hat{V}\hat{V}^T\tilde{V})$. It follows that

$$\lambda_1(\tilde{V}^T\hat{V}\hat{V}^T\tilde{V}) \leq \frac{C}{n\hat{\lambda}_r}\lambda_1(Z_{(2)}Z_{(2)}^T).$$

Then (29) holds by Lemma 2.

If $p - r \geq n$, then $\mathrm{Rank}(A) = r$, $\mathrm{Rank}(B) = n - r$ and $\mathrm{Rank}(A + B) = n$. By Weyl's inequality,

$$\lambda_n(A + B) \leq \lambda_r(A) + \lambda_{n-r+1}(B) = \lambda_r(A).$$

Thus,

$$\frac{1}{n}\lambda_n(Z_{(2)}Z_{(2)}^T) \leq \lambda_r(A) \leq \frac{\hat{\lambda}_1}{c}\lambda_r(\tilde{V}^T\hat{V}\hat{V}^T\tilde{V}),$$

or

$$\lambda_r(\tilde{V}^T\hat{V}\hat{V}^T\tilde{V}) \geq \frac{c}{n\hat{\lambda}_1}\lambda_n(Z_{(2)}Z_{(2)}^T).$$

Then (31) holds by Lemma 2.

Now we prove (30). Equation (32) implies that

$$\Lambda_{(1)}^{-1/2}V^T\hat{U}\hat{\Lambda}\hat{U}^T V\Lambda_{(1)}^{-1/2} = \frac{1}{n}Z_{(1)}Z_{(1)}^T. \tag{34}$$

Note that

$$\begin{aligned}\operatorname{tr}\big(\Lambda_{(1)}^{-1/2}V^T\hat{U}\hat{\Lambda}\hat{U}^T V\Lambda_{(1)}^{-1/2}\big) &\geq \operatorname{tr}\big(\Lambda_{(1)}^{-1/2}V^T\hat{V}\hat{\Lambda}_{(1)}\hat{V}^T V\Lambda_{(1)}^{-1/2}\big)\\ &\geq \frac{\hat{\lambda}_r}{\lambda_1}\Big(r - \frac{1}{2}\|\hat{V}\hat{V}^T - VV^T\|_F^2\Big).\end{aligned} \tag{35}$$

Then

$$\Big(r - \frac{1}{2}\|\hat{V}\hat{V}^T - VV^T\|_F^2\Big) \leq \frac{\lambda_1}{n\hat{\lambda}_r}\operatorname{tr} Z_{(1)}Z_{(1)}^T.$$

Hence (30) holds by Lemma 3.

□

Theorem 8. *Suppose Assumption 6 holds. Then, with probability at least* $1 - 4\exp(-t^2/2)$*, we have*

$$\Big|\sum_{i=r+1}^{\min(n,p)}\lambda_i(XX^T) - n\sum_{i=r+1}^{p}\lambda_i\Big| \leq rC(\sqrt{p-r}+\sqrt{n}+t)^2 + \sqrt{2np}Ct + Ct^2.$$

Proof. We deal with $Z^T\Lambda Z$ instead of XX^T since $\lambda_i(XX^T) = \lambda_i(Z^T\Lambda Z)$, $i = 1\ldots,\min(n,p)$. And $Z^T\Lambda Z$ can be written as the sum of two quantities

$$Z^T\Lambda Z = Z_{(1)}^T\Lambda_{(1)}Z_{(1)} + Z_{(2)}^T\Lambda_{(2)}Z_{(2)} \overset{def}{=} A + B.$$

Note that $\operatorname{Rank}(A) = r$. By Weyl's inequality, for $i = r+1,\ldots,\min(n,p)$, we have that

$$\lambda_i(B) \leq \lambda_i(Z^T\Lambda Z) \leq \lambda_{i-r}(B).$$

Sum the above inequality over $i = r+1,\ldots,\min(n,p)$,

$$\sum_{i=r+1}^{\min(n,p)}\lambda_i(B) \leq \sum_{i=r+1}^{\min(n,p)}\lambda_i(Z^T\Lambda Z) \leq \sum_{i=1}^{\min(n,p)-r}\lambda_i(B).$$

It implies that

$$\Big|\sum_{i=r+1}^{n}\lambda_i(Z^T\Lambda Z)-\mathrm{tr}(B)\Big|\le r\lambda_1(B).$$

Then

$$\Big|\sum_{i=r+1}^{n}\lambda_i(Z^T\Lambda Z)-n\sum_{i=r+1}^{p}\lambda_i\Big|\le r\lambda_1(B)+\Big|\mathrm{tr}(B)-n\sum_{i=r+1}^{p}\lambda_i\Big|. \tag{36}$$

Note that $\mathrm{tr}(B)=\sum_{i=r+1}^{p}\sum_{j=1}^{n}\lambda_i Z_{ij}^2$, by Lemma 3, with probability at least $1-2\exp(-t^2/2)$, we have

$$\Big|\mathrm{tr}(B)-n\sum_{i=r+1}^{p}\lambda_i\Big|\le\sqrt{2np}Ct+Ct^2. \tag{37}$$

Note that $\lambda_1(B)\le C\lambda_1(Z_{(2)}^T Z_{(2)})$. The conclusion follows by (36), (37) and Lemma 2.

$\square$

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

Proof Of Theorem ??. Note that $\operatorname{tr}(\hat{\tilde{V}}_i^T S_i \hat{\tilde{V}}_i) = \sum_{i=r+1}^p \lambda_i(S_i)$, $i = 1, 2$. Similar to Proposition **??**, we have that $\operatorname{tr}(\hat{\tilde{V}}_i^T S_i \hat{\tilde{V}}_i) = (p-r)\sigma^2 + O_P(\frac{\max(n,p)}{n})$, $i = 1, 2$. Hence

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

By Assumption **??**, $\frac{\max(n,p)}{n\sqrt{p}} = \max(\frac{1}{\sqrt{p}}, \frac{\sqrt{p}}{n}) \to 0$. And

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

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, we only need 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

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

By the following useful relationship

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

and Lemma **??**, we have that

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

As for P_3. To prove the convergence in probability, here we prove the convergence in L^1:

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

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 $\hat{\hat{V}}^T\big((\bar{X}_1-\mu_1)-(\bar{X}_2-\mu_2)\big)$ given S is $N(0, \tau\hat{\hat{V}}^T\Sigma\hat{\hat{V}})$. As we have shown,

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

On the other hand, $\lambda_i(\hat{\hat{V}}^T\Sigma\hat{\hat{V}}) = \sigma^2$ for $i = r+1,\ldots,p-r$. Then

$$(p-2r)\sigma^4 \le \mathrm{tr}(\hat{\hat{V}}^T\Sigma\hat{\hat{V}})^2 \le \big(\frac{p}{n}+1\big)^2 O_P(1) + (p-2r)\sigma^4,$$

or

$$\mathrm{tr}(\hat{\hat{V}}^T\Sigma\hat{\hat{V}})^2 = p\sigma^4(1+o_P(1)). \tag{38}$$

It follows that

$$\frac{\lambda_1^2(\hat{\hat{V}}^T\Sigma\hat{\hat{V}})}{\mathrm{tr}(\hat{\hat{V}}^T\Sigma\hat{\hat{V}})^2} = O_P\Big(\frac{(p/n+1)^2}{p}\Big) = o_P(1). \tag{39}$$

Then for every subsequence of $\{n\}$, there's a further subsequence along which (39) holds almost surely. By Lemma 4, for every subsequence of $\{n\}$, there's a further subsequence along which we have

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

It means that

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

Thus the weak convergence also holds unconditionally:

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

Similar to (38) we have

$$\mathrm{tr}(\hat{\hat{V}}^T\Sigma\hat{\hat{V}})=(p-r)\sigma^2\big(1+O_P\big(\frac{1}{n}+\frac{1}{p}\big)\big). \tag{40}$$

By (38), (40) and Slulsk's theorem,

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

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

Proof Of Theorem 3. By assumption, $\hat{r}\le R$ for some constant R. Similar to the proof of Proposition **??**, in the current context we have that $\mathrm{tr}(\hat{\hat{V}}_iS_i\hat{\hat{V}}_i)=\mathrm{tr}\Sigma+P_P(\frac{\max(n,p)}{n})$, $i=1,2$. It follows that

$$\frac{T_2-\|\mu_1-\mu_2\|^2}{\sqrt{2\tau^2\mathrm{tr}\Sigma^2}}=\frac{\|\hat{\hat{V}}^T(\bar{X}_1-\bar{X}_2)\|^2-\|\mu_1-\mu_2\|^2-\tau\mathrm{tr}\Sigma}{\sqrt{2\tau^2\mathrm{tr}\Sigma^2}}+o_P(1).$$

Since $\bar{X}_i|\mu_i\sim N(\mu_i,\frac{1}{n_i}\Sigma)$ and $\mu_i\sim N(0,\frac{\psi}{n_i\sqrt{p}}I_p)$, we have $\bar{X}_i\sim N(0,\frac{1}{n_i}(\Sigma+\frac{1}{\sqrt{p}}\psi I_p))$, $i=1,2$. Hence we have that $\hat{\hat{V}}^T(\bar{X}_1-\bar{X}_2)|S\sim N(0,\tau\hat{\hat{V}}^T(\Sigma+\frac{1}{\sqrt{p}}\psi I_p)\hat{\hat{V}})$ by the independence of S and $(\mu_1,\mu_2,\bar{X}_1,\bar{X}_2)$. Note that

$$c+\frac{1}{\sqrt{p}}\psi\le\lambda_{\min}(\hat{\hat{V}}^T(\Sigma+\frac{1}{\sqrt{p}}\psi I_p)\hat{\hat{V}})\le\lambda_{\max}(\hat{\hat{V}}^T(\Sigma+\frac{1}{\sqrt{p}}\psi I_p)\hat{\hat{V}})\le C+\frac{1}{\sqrt{p}}\psi.$$

Then by Lemma 4,

$$\frac{\|\hat{\tilde{V}}^T(\bar{X}_1-\bar{X}_2)\|^2-\tau\mathrm{tr}(\hat{\tilde{V}}^T(\Sigma+\frac{1}{\sqrt{p}}\psi I_p)\hat{\tilde{V}})}{\sqrt{2\tau^2\mathrm{tr}(\hat{\tilde{V}}^T(\Sigma+\frac{1}{\sqrt{p}}\psi I_p)\hat{\tilde{V}})^2}}\xrightarrow{\mathcal{L}}N(0,1). \tag{41}$$

It can be easily shown that

$$\frac{\mathrm{tr}(\hat{\tilde{V}}^T(\Sigma+\frac{1}{\sqrt{p}}\psi I_p)\hat{\tilde{V}})^2}{\mathrm{tr}(\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})^2}\xrightarrow{P}1. \tag{42}$$

Next we will show that

$$\frac{\mathrm{tr}(\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})^2}{\mathrm{tr}\Sigma^2}\xrightarrow{P}1. \tag{43}$$

In fact, for $i=1,\ldots,p$ we have

$$\lambda_i(\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})=\lambda_i(\Sigma^{1/2}\hat{\tilde{V}}\hat{\tilde{V}}^T\Sigma^{1/2})\leq\lambda_i(\Sigma). \tag{44}$$

On the other hand, for $i=1,\ldots,p-\hat{r}$ we have that

$$\lambda_i(\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})=\lambda_i(\Sigma^{1/2}(I_p-\hat{V}\hat{V}^T)\Sigma^{1/2})=\lambda_i(\Sigma-\Sigma^{1/2}\hat{V}\hat{V}^T\Sigma^{1/2})\geq\lambda_{i+\hat{r}}(\Sigma), \tag{45}$$

where the last inequality holds by Weyl's inequality and the fact that the rank of $\Sigma^{1/2}\hat{V}\hat{V}^T\Sigma^{1/2}$ is at most $\hat{r}$. By (44) and (45),

$$\sum_{i=\hat{r}+1}^{p}\lambda_i^2(\Sigma)\leq\mathrm{tr}(\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})^2\leq\mathrm{tr}\Sigma^2.$$

Then $|\mathrm{tr}(\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})^2-\mathrm{tr}\Sigma^2|\leq\sum_{i=1}^{\hat{r}}\lambda_i^2(\Sigma)\leq RC^2$. Hence (43) holds. By (41), (42), (43) and Slutsky's theorem,

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

Note that

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

We only need to show the last two terms are negligible. But $\frac{1}{\tau}\|\mu_1-\mu_2\|^2 \sim \frac{\psi}{\sqrt{p}}\chi_p^2 = \sqrt{p}\psi + O_P(1)$ by central limit theorem, then

$$\frac{\frac{p-\hat{r}}{\sqrt{p}}\psi - \frac{1}{\tau}\|\mu_1-\mu_2\|^2}{\sqrt{2\text{tr}\Sigma^2}} = o_P(1).$$

And

$$\frac{\text{tr}(\hat{\hat{V}}\Sigma\hat{\hat{V}}) - \text{tr}\Sigma^2}{\sqrt{2\text{tr}\Sigma}} = o_P(1)$$

by (44) and (45). The proof is completed.

□

Proof Of Theorem 4. Since $r=0$, $X_{ki}=\mu_k+Z_{ki}$, $i=1,\ldots,n_k$ and $k=1,2$. As in the proof of Theorem 3, we only need to prove

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

Independent of data, generate a $p\times p$ random orthogonal matrix with Haar invariant distribution. It can be seen that $\big(O(\bar{Z}_1-\bar{Z}_2), OSO^T\big) \sim \big((\bar{Z}_1-\bar{Z}_2), S\big)$ and are independent of O. But the eigenvectors of OSO^T are $(O\hat{V}, O\hat{\hat{V}})$, thus $\big(O(\bar{Z}_1-\bar{Z}_2), O\hat{\hat{V}}\big) \sim \big((\bar{Z}_1-\bar{Z}_2), \hat{\hat{V}}\big)$. It follows that

$$\begin{aligned}\|\hat{\hat{V}}^T(\bar{Z}_1-\bar{Z}_2)+\hat{\hat{V}}^T(\mu_1-\mu_2)\|^2 &= \|(O\hat{\hat{V}})^T O(\bar{Z}_1-\bar{Z}_2) + (O\hat{\hat{V}})^T O(\mu_1-\mu_2)\|^2\\ &\sim \|\hat{\hat{V}}^T(\bar{Z}_1-\bar{Z}_2)+\hat{\hat{V}}^T O(\mu_1-\mu_2)\|^2.\end{aligned}$$

Note that $O(\mu_1-\mu_2)/\|\mu_1-\mu_2\|$ is uniformly distributed on the unit ball in $\mathbb{R}^p$. Independent of data and O, generate a random variable $R>0$ with $R^2\sim\chi_p^2$. Then

$$\xi \overset{def}{=} R\frac{O(\mu_1-\mu_2)}{\|\mu_1-\mu_2\|} \sim N_p(0_p, I_p).$$

Now we have

$$\begin{aligned}&\|\hat{\hat{V}}^T(\bar{Z}_1-\bar{Z}_2)+\hat{\hat{V}}^T O(\mu_1-\mu_2)\|^2\\ =&\|\hat{\hat{V}}^T(\bar{Z}_1-\bar{Z}_2)\|^2 + \frac{\|\hat{\hat{V}}^T\xi\|^2}{R^2}\|\mu_1-\mu_2\|^2 + \frac{\|\mu_1-\mu_2\|}{R}\xi^T\hat{\hat{V}}\hat{\hat{V}}^T(\bar{Z}_1-\bar{Z}_2).\end{aligned} \tag{46}$$

Since $\hat{\hat{V}}^T(\bar{Z}_1-\bar{Z}_2)|\hat{\hat{V}} \sim N_{p-\hat{r}}(0_{p-\hat{r}}, \tau\sigma^2 I_{p-\hat{r}})$, the asymptotic normality of the first term of (46) follows by central limit theorem:

$$\frac{\|\hat{\hat{V}}^T(\bar{Z}_1-\bar{Z}_2)\|^2 - \tau(p-\hat{r})\sigma^2}{\sigma^2\sqrt{2\tau^2(p-\hat{r})}} \xrightarrow{\mathcal{L}} N(0,1). \tag{47}$$

By the fact that $\hat{\hat{V}}^T\xi|\hat{\hat{V}} \sim N_{p-\hat{r}}(0_{p-\hat{r}}, I_{p-\hat{r}})$ and central limit theorem, we have

$$\|\hat{\hat{V}}^T\xi\|^2 = (p-\hat{r})(1+O_P(\frac{1}{\sqrt{p-\hat{r}}})) = p(1+O_P(\frac{1}{\sqrt{p}})).$$

Also by central limit theorem, $R^2 = p(1+O_P(\frac{1}{\sqrt{p}}))$. Thus for the second term of (46), we have

$$\frac{\|\hat{\hat{V}}^T\xi\|^2}{R^2}\|\mu_1-\mu_2\|^2 = \|\mu_1-\mu_2\|^2+O_P(\frac{1}{\sqrt{p}})\|\mu_1-\mu_2\|^2 = \|\mu_1-\mu_2\|^2+o_P(\sigma^2\sqrt{2\tau^2p}). \tag{48}$$

Now we deal wih the second term of (46). Note that $\xi^T\hat{\hat{V}}\hat{\hat{V}}^T(\bar{Z}_1-\bar{Z}_2)\big|\big(\hat{\hat{V}},(\bar{Z}_1-\bar{Z}_2)\big) \sim N(0,\|\hat{\hat{V}}^T(\bar{Z}_1-\bar{Z}_2)\|^2)$, which implies that

$$\xi^T\hat{\hat{V}}\hat{\hat{V}}^T(\bar{Z}_1-\bar{Z}_2) = O_P(1)\|\hat{\hat{V}}^T(\bar{Z}_1-\bar{Z}_2)\| = O_P(\sqrt{\tau p}).$$

It follows that

$$\frac{\|\mu_1-\mu_2\|}{R}\xi^T\hat{\hat{V}}\hat{\hat{V}}^T(\bar{Z}_1-\bar{Z}_2) = O_P(\sqrt{\tau})\|\mu_1-\mu_2\| = o_P(\sigma^2\sqrt{2\tau^2p}). \tag{49}$$

By (46), (47), (48), (49) and Slutsky's theorem, we have the conclusion

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

□

Proof Of Theorem 5. The method of Theorem **??**'s proof can still work here with some modifications. The term P_3 in THeorem **??**'s proof disappears in the current circumstance. The other two terms can be treated as before if we can show that

$$\lambda_1(\hat{\hat{V}}^T\Sigma_k\hat{\hat{V}}) = O_P(\frac{p}{n}) \quad \text{k=1,2}.$$

In fact,

$$\lambda_1(\hat{\hat{V}}^T\Sigma_k\hat{\hat{V}}) = \lambda_1(\hat{\hat{V}}^TV_kD_k^2V_k^T\hat{\hat{V}})+\sigma^2 \le \kappa p^\beta\lambda_1(\hat{\hat{V}}^TV_kV_k^T\hat{\hat{V}})+\sigma^2.$$

But

$$\lambda_1(\hat{\hat{V}}^TV_kV_k^T\hat{\hat{V}}) = \lambda_1(V_k^T(I_p-\hat{V}\hat{V}^T)V_k) \le \lambda_1(V_k^T(I_p-\hat{V}_k\hat{V}_k^T)V_k).$$

The last inequality holds since $\hat{V}\hat{V}^T$ is the projection on the sum space of $\hat{V}_1\hat{V}_1^T$ and $\hat{V}_2\hat{V}_2^T$ and hence $\hat{V}\hat{V}^T \geq \hat{V}_1\hat{V}_1^T$. Thus,

$$\lambda_1(\hat{\hat{V}}^T V_k V_k^T \hat{\hat{V}}) \leq \frac{1}{2}\|V_k V_k^T - \hat{V}_k \hat{V}_k^T\|_F^2 = O_P(\frac{p}{np^\beta}).$$

Therefore, $\lambda_1(\hat{\hat{V}}^T \Sigma_k \hat{\hat{V}}) = O_P(\frac{p}{n})$. □

Acknowledgements

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

References

Anderson TW. Asymptotic theory for principal component analysis. Annals of Mathematical Statistics 1986;34(1):122–48.

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, 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.

Davidson KR, Szarek SJ. Local operator theory, random matrices and banach spaces. "handbook in Banach Spaces" Vol 2001;:317–66.

Fan J, Liao Y, Liu H. An overview on the estimation of large covariance and precision matrices. Eprint Arxiv 2015a;.

Fan J, Liao Y, Yao J. Power enhancement in high dimensional cross-sectional tests. Econometrica 2015b;83(4):1497.

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

Jiang J. Reml estimation: asymptotic behavior and related topics. Annals of Statistics 1996;24(1):255–86.

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

Laurent B, Massart P. Adaptive estimation of a quadratic functional by model selection. Annals of Statistics 2000;28(5):1302–38.

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. A general framework for consistency of principal component analysis. ArXiv e-prints 2012;arXiv:1211.2671.

Shen D, Shen H, Marron JS. Consistency of sparse pca in high dimension, low sample size contexts. Journal of Multivariate Analysis 2013;115(1):317–333.

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.