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 $X_{k1}, \ldots, X_{kn_k}$ 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:

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

*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 $n = n_1 + n_2$. 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. The 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 $A = [\mathrm{diag}(\Sigma)]^{-1}$, 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 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

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

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. Ma et al. (2015) considered a factor model and proposed a asymptotical 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 $\{X_{k1}, \dots, X_{kn_k}\}$, $k = 1, 2$ 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}, $i = 1, 2, \dots, n_k$ 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, $D_k = diag(\lambda_{k1}^{\frac{1}{2}}, \dots, \lambda_{kr_k}^{\frac{1}{2}})$ 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.

Finally, let $\tau = (n_1 + n_2)/(n_1 n_2)$, S be the pooled sample covariance.

$$S = \frac{1}{n-2} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (X_{ki} - \bar{X}_k)(X_{ki} - \bar{X}_k)^T = \frac{(n_1 - 1) S_1 + (n_2 - 1) S_2}{n-2}, \quad (3)$$

where

$$S_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (X_{ki} - \bar{X}_k)(X_{ki} - \bar{X}_k)^T, \quad (4)$$

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

3. Methodology

In this section, we consider testing the hypotheses (1) with equal covariance matrices.

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.

3.1. Motivation

In high dimensional setting, many test procedures for hypotheses (1) is based on a statistic $T(X)$ which estimates $(\mu_1 - \mu_2)^T A(\mu_1 - \mu_2)$. Usually, $T(X)$ satisfies $\mathrm{E}T = 0$ under null hypothesis and $\mathrm{E}T > 0$ under alternative. To determine the critical value, the asymptotic distribution of T need to be derived, say

$$\frac{T - \mathrm{E}T}{\sqrt{\mathrm{Var}(T)}} \xrightarrow{\mathcal{L}} N(0,1).$$

Since $\mathrm{Var}(T)$ may depend on parameters, a ratio consistent estimator $\widehat{\mathrm{Var}(T)}$ of $\mathrm{Var}(T)$ is necessary. Then the rejection region of a level α test can be defined as $T(X) \geq \widehat{\mathrm{Var}(T)}^{\frac{1}{2}} z_{1-\alpha}$ where $z_{1-\alpha}$ is the $1-\alpha$ quantile of $N(0,1)$. The asymptotic power of the test is

$$\Phi(\frac{\mathrm{E}T}{\sqrt{\mathrm{Var}(T)}} - z_{1-\alpha}).$$

Thus, a general idea to boost the power of test is to reduce the variance $\mathrm{Var}(T)$ while the mean $\mathrm{E}(T)$ varies relatively little.

Now we revisit T_{BS} and T_{CQ} which are both based on the estimation of $\|\mu_1 - \mu_2\|^2$. Denote the spectral decomposition of Σ by $\Sigma = \sum_{i=1}^p \lambda_i p_i p_i^T$ with $\lambda_1 \geq \cdots \geq \lambda_p$, where p_i, $i = 1, \ldots, p$, are orthonormal p dimensional vectors. The main body of both T_{BS} and T_{CQ} is

$$\frac{n_1 n_2}{n_1 + n_2} \sum_{i=1}^p (\bar{X}_1 - \bar{X}_2)^T p_i p_i^T (\bar{X}_1 - \bar{X}_2), \tag{5}$$

which is a sum of p independent terms. Since $\sqrt{n_1 n_2/(n_1+n_2)}(\bar{X}_1 - \bar{X}_2)$ is distributed as $N(0,\Sigma)$, the variance of $n_1 n_2/(n_1+n_2)(\bar{X}_1 - \bar{X}_2)^T p_i p_i^T (\bar{X}_1 - \bar{X}_2)$ is $2\lambda_i^2$ which decreases in i. By our previous argument, if a few leading terms with significantly large variance are removed, the modified test will be more powerful.

The argument is also supported by the likelihood ratio test. If Σ is known, the LRT is based on

$$(\bar{X}_1 - \bar{X}_2)^T \Sigma^{-1} (\bar{X}_1 - \bar{X}_2) = \frac{n_1 n_2}{n_1 + n_2} \sum_{i=1}^p \lambda_i^{-1} (\bar{X}_1 - \bar{X}_2)^T p_i p_i^T (\bar{X}_1 - \bar{X}_2). \tag{6}$$

The difference between (5) and (6) is the weights λ_i^{-1}. Unfortunately, λ_i's are hard to precisely estimate in high dimensional setting. See Bai and Silverstein (2010) for detail. Nevertheless, it's possible to identify which λ_i's are large. LRT implies the corresponding terms should have small weights, which coincides with our previous idea.

Under Assumption 2, $\Sigma = V\Lambda V^T + \sigma^2 I_p$. The eigenvalues of Σ are $\lambda_1 + \sigma^2, \ldots, \lambda_r + \sigma^2, \sigma^2, \ldots, \sigma^2$. The eigenvectors corresponding to the first r eigenvalues are columns of V. Follow our previous argument, if the principal subspace VV^T is known, we project X_{ki} on the orthogonal complement space $\tilde{V}\tilde{V}^T$ and invoke the statistic of Chen and Qin (2010). We define the following statistic

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

the asymptotic normality of T_1 can be obtained by Chen and Qin (2010)'s Theorem 1.

Proposition 1. *Under Assumptions 1-4 and local alternative, that means,* $\frac{n}{p}\|\mu_1 - \mu_2\|^2 \to 0$*, we have*

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

Remark 1. The asymptotic variance of T_1 is of order $\tau^2 p$ while the asymptotic variance of T_{CQ} is of order $\tau^2 p^{2\beta}$ by Chen and Qin (2010)'s Theorem 1. The asymptotic variance is reduced significantly if $\beta > 1/2$ and p is sufficiently large.

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 $\hat{V}_i$ and $\hat{\tilde{V}}_i$ the first r and last $p-r$ eigenvectors of S_i respectively, $i = 1, 2$. As estimators of their population counterparts, these simple statistics actually reach the optimal convergence rate. See Cai et al. (2013).

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

Define

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

We propose our new test statistic as

$$Q = \frac{T_2}{\hat{\sigma}^2\sqrt{2\tau^2 p}}, \tag{7}$$

where $\hat{\sigma}^2$ is a ratio consistent estimator of σ^2. In next section, it will be proved that the asymptotic distribution under null of Q is $N(0,1)$. We reject the null hypothesis when Q is larger than the upper α quantile of $N(0,1)$.

Remark 2. Compared with random projection method, our projection is determined by the structure of S_1, S_2 and S. We don't project multiple times as random projection method did, which leads to reproducibility in practice.

Remark 3. The statistic T_2 is invariant under shift transformation, that is, T_2 is invariant when adding a vector to X_{1i} and X_{2j} simultaneously: $X_{1i} \mapsto X_{1i} + \mu$ and $X_{2j} \mapsto X_{2j} + \mu$, $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$.

Remark 4. 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. Therefore, without loss of generality, we will assume that r is known.

Theoretical results will show that the asymptotic variance of T_2 is significantly smaller than T_{CQ}. On the other hand, the new test statistic estimates

$\|\tilde{V}(\mu_1 - \mu_2)\|^2$. Then the superiority of the new test will be established if

$$\frac{\|\tilde{V}(\mu_1 - \mu_2)\|}{\|\mu_1 - \mu_2\|} \approx 1. \tag{9}$$

Unfortunately, (9) is not always the case since there always exists some $\tilde{V}$ and $\mu_1 - \mu_2$ such that $\|\tilde{V}(\mu_1 - \mu_2)\| = 0$. However, (9) 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 probability distribution, it can be established that

$$\frac{\|\tilde{V}(\mu_1 - \mu_2)\|}{\|\mu_1 - \mu_2\|} \xrightarrow{P} 1.$$

Such assumption of μ_k will be used in Theorem 1.

In order to formulate a test procedure, σ^2 needs to be estimated. Note that σ^2 can be written as

$$\sigma^2 = \sum_{i=r+1}^{p} \lambda_i(\Sigma). \tag{10}$$

We estimate it by sample version:

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

If estimator $\hat{\sigma}^2$ of σ^2 is ratio consistent, the asymptotic distribution of (7) will not change if we replace σ^2 by $\hat{\sigma}^2$ due to Slutsky's theorem.

4. Theoretical results

In this section, we derive some properties of the new test statistic.

First we consider the case when the eigenvalues of Σ is bounded, i.e., there is no clear correlation between variables. In many practical problems, the alternative is 'dense', i.e., under H_1 the signals in $\mu_1 - \mu_2$ spread out over a large number of co-ordinates. See Cai et al. (2014). Similar to bayesian models, we assume a normal prior distribution for μ_k to characterize 'dense' alternative. When the eigenvalues of Σ is bounded, spike variance model is not valid. Hence some estimators in our test procedure make no sense. Particularly, if $\hat{r}$ is estimated by (8). the $\hat{r}$ is nothing but a random integer not greater than R and $\hat{V}\hat{V}^T$ is just a random projection. In this case, the difference of our test statistic and Chen and Qin (2010)'s is small. 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 1. *Assume $X_{ki} \sim N(\mu_k, \Sigma)$, $i = 1, \ldots, n_k$, $k = 1, 2$. Suppose that assumption 1 holds, $0 < c \leq \lambda_p(\Sigma) \leq \lambda_1(\Sigma) \leq C < \infty$ where c and C are constant, each element of μ_k is independently generated by $N(0, (n_k\sqrt{p})^{-1}\psi)$ for $k = 1, 2$, ψ is a constant and $\hat{r}$ is estimated by (8). If $p \to \infty$ as $n \to \infty$ and $p = o(n^2)$, then we have*

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

Now we establish the asymptotic normality of the new test statistic under spiked covariance model. We first give a result of the convergence rate of $\hat{\sigma}^2$.

Proposition 2. *Under Assumptions 1-4, we have that*

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

Proposition 2 implies that $\hat{\sigma}^2$ is a consistent estimator of σ^2.

To derive the asymptotic normality of the new test statistic, we require the following relationship of n and p.

Assumption 5. *Assume $\sqrt{p}/n \to 0$.*

Theorem 2. *Under Assumptions 1-5, if the local alternative holds, that is,*

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

then

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

Remark 5. The Assumption 5 is a strong condition. However, it may not be able to relax it. In fact, the asymptotic normality of T_2's main term $\|\hat{\tilde{V}}^T(\bar{X}_1 - \bar{X}_2)\|^2$ requires

$$\frac{\lambda_1\big((\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})^2\big)}{\mathrm{tr}\big((\hat{\tilde{V}}^T\Sigma\hat{\tilde{V}})^2\big)} \xrightarrow{P} 0, \tag{11}$$

see Lemma 4 in appendix. And (11) is equivalent to Assumption 5 by Lemma 2 in appendix.

By Proposition 2 and Theorem 2, the power function of the new test can be obtained immediately.

Corollary 1. *Under Assumptions 1-5, if we reject the null hypothesis when Q is larger than $1-\alpha$ quantile of $N(0,1)$, then the asymptotic power function of our test is*

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

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.

When Assumption 5 doesn't met, the asymptotic normality are not valid. 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.

5. Unequal Variance

In this section, we concern the situation with unequal covariance matrices. Assume $\{X_{11}, \dots, X_{1n_1}\}$ and $\{X_{21}, \dots, X_{2n_2}\}$ 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 3. *Under Assumptions 1-3 and 5, 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 6. 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. Numerical studies

6.1. Simulation results

Our simulation study focus on equal variance case. We generate X_{ki} by the model in Assumption 2, where each element of U_{ki} and Z_{ki} are generated from $N(0,1)$. V is a random orthonormal matrix. We generate λ_i as p^β plus a random error from $U(0,1)$.

First we simulate the level of the new test. The nominal level $\alpha = 0.05$ and we set $r = 2$. Samples are repeatedly generated 1000 times to calculate empirical level. For comparison, we also give corresponding 'oracle' level which is calculated by 'statistic' $T_1/(\sigma^2\sqrt{2p\tau^2})$ whose asymptotic normality can be guaranteed by Theorem 1 in Chen and Qin (2010). The results are listed in Table 1. 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 and it performs better when n is larger.

Table 1: Test level simulation

		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

Then we simulate the empirical power of our test and Chen and Qin (2010)'s 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, we use permutation method to compute critical value. The validity of permutation method can be found in Lehmann and Romano (2005)'s Example 15.2.2. We plot the empirical power versus $\|\mu_1 - \mu_2\|$ when other parameters hold constant. The results are illustrated in figure 1. From the results, we can find that when Σ is spiked, the new test outperforms T_{CQ} substantially; when Σ is not spiked, the new test and T_{CQ} are comparable.

6.2. Real data analysis

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

We assume $\Sigma_1 = \Sigma_2$. 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).

7. 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

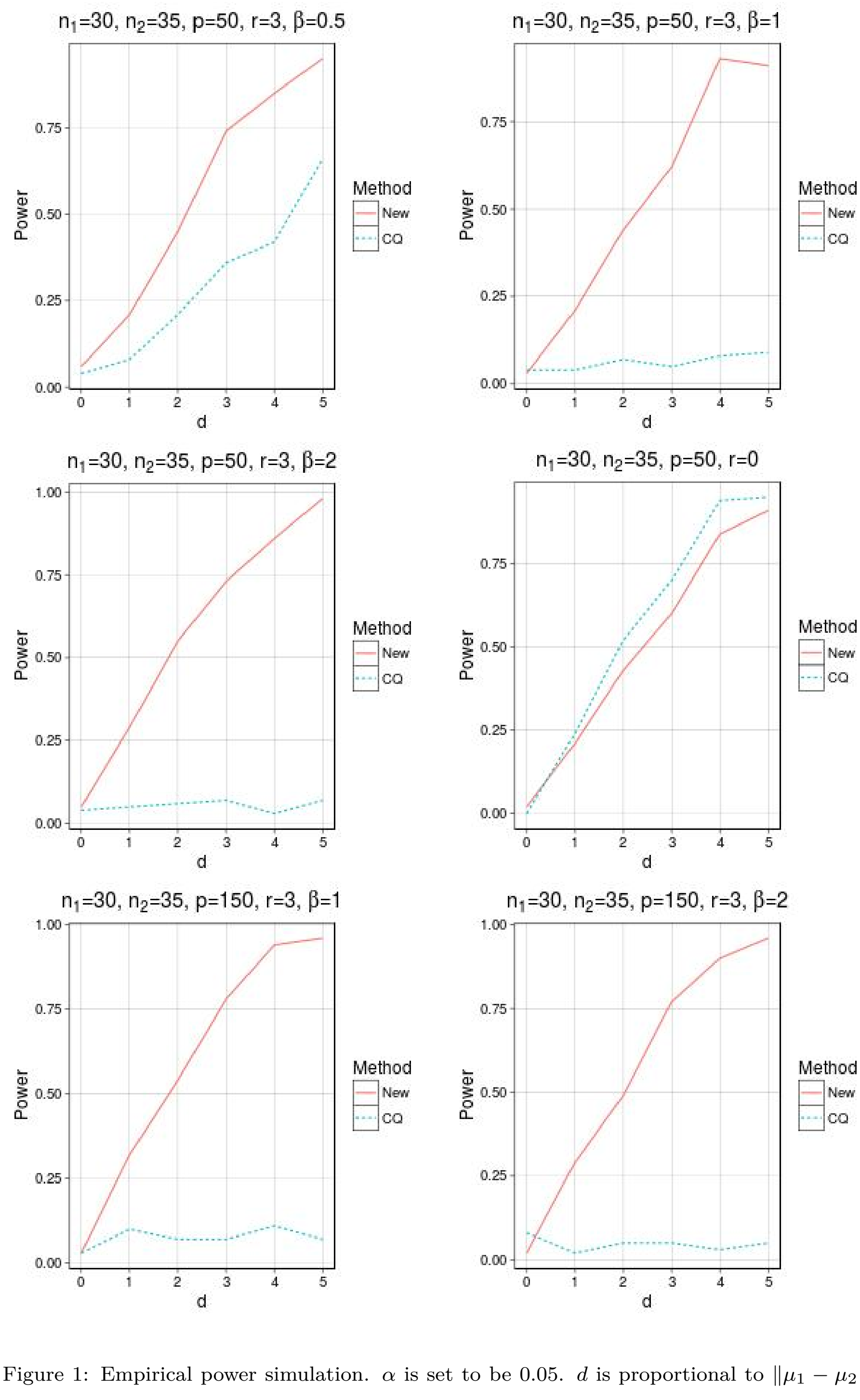

Figure 1: Empirical power simulation. α is set to be 0.05. d is proportional to $\|\mu_1 - \mu_2\|^2$. For each simulation, we do 50 permutations to determine critical value. We generate 100 independent samples to compute empirical power.

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

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 (Convergence rate of principal space estimation)**.** *Under the Assumption 1-4, 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 estimator of VV^T, namely, it reaches the minimax convergence rate

$$E\|\hat{V}\hat{V}^T - VV^T\|_F^2 \asymp r \wedge (p - r) \wedge \frac{r(p - r)}{(n_1 + n_2 - 2)h(\lambda)} \tag{12}$$

as long as the right hand side tends to 0. Here $h(\lambda) = \frac{\lambda^2}{\lambda+1}$, $a_n \asymp b_n$ represents $a_n \geq cb_n$ and $a_n \leq Cb_n$ for some positive c, C for every n. In model of Assumption 2, r is fixed, $\lambda = cp^\beta$. It's obvious that the right hand side of (12) is of order $p^{1-\beta}/n$. We note that it is assumed $\beta \geq \frac{1}{2}$ in Assumption 3, together with $\sqrt{p}/n \to 0$ we have $p^{1-\beta}/n \to 0$. Hence $\hat{V}\hat{V}^T$ reaches the convergence rate.

□

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 7. 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 3. *Suppose that W_n is a $p \times p$ matrix distributed as $\text{Wishart}_p(n, I_p)$. Then as $n \to \infty$,*

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

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 distributed as $N(0, 1)$. Then

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

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

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*

$$\frac{X_n^T A_n X_n - \mathrm{E} X_n^T A_n X_n}{\left[\mathrm{Var}(X_n^T A_n X_n)\right]^{1/2}} \xrightarrow{\mathcal{L}} N(0,1) \tag{13}$$

is that

$$\frac{\lambda_{\max}(A_n^2)}{\mathrm{tr}(A_n^2)} \to 0. \tag{14}$$

Remark 8. 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

$$\frac{X_n^T A_n X_n - \mathrm{E} X_n^T A_n X_n}{\left[\mathrm{Var}(X_n^T A_n X_n)\right]^{1/2}} = \sum_{i=1}^{k_n} \frac{\lambda_i(A_n)}{\left[2\mathrm{tr}(A_n^2)\right]^{1/2}} (Z_{ni}^2 - 1), \tag{15}$$

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

If 14 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 13 follows by Lindeberg's central limit theorem.

Conversely, if 13 holds, we will prove that there is a subsequence of $\{n\}$ along which 14 holds. Then 14 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 13 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 14 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 Proposition 1. Since V and $\tilde{V}$ are orthogonal, we have

$$\tilde{V}^T X_{ki} = \tilde{V}^T\mu_i + \tilde{V}^T Z_{ki} \sim N(\tilde{V}^T\mu_k, \sigma^2 I_{p-r}) \quad k = 1, 2 \text{ and } i = 1, \ldots, n_k.$$

Let $\bar{Z}_1$ and $\bar{Z}_2$ be the sample mean of $\{Z_{1i}\}$ and $\{Z_{2i}\}$ respectively. Then

$$\|\tilde{V}^T(\bar{X}_1 - \bar{X}_2)\|^2 = \|\tilde{V}^T(\mu_1 - \mu_2) + \tilde{V}^T(\bar{Z}_1 - \bar{Z}_2)\|^2$$
$$= \|\tilde{V}^T(\mu_1 - \mu_2)\|^2 + \|\tilde{V}^T(\bar{Z}_1 - \bar{Z}_2)\|^2 + 2(\mu_1 - \mu_2)^T\tilde{V}\tilde{V}^T(\bar{Z}_1 - \bar{Z}_2).$$

But

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

Then

$$\|\tilde{V}^T(\bar{X}_1 - \bar{X}_2)\|^2 = \|\tilde{V}^T(\mu_1 - \mu_2)\|^2 + \|\tilde{V}^T(\bar{Z}_1 - \bar{Z}_2)\|^2 + o_P(\frac{\sqrt{p}}{n}). \tag{16}$$

Note that $\frac{1}{n_i}\tilde{V}^T S_i \tilde{V} \sim \frac{\sigma^2}{n_i(n_i-1)} Wishart_{p-r}(n_i - 1, I_{p-r})$, $i = 1, 2$. Then

$$\begin{aligned}\frac{1}{n_i}\mathrm{tr}(\tilde{V}^T S_i \tilde{V}) \sim& \frac{\sigma^2}{n_i(n_i-1)}\chi^2_{(p-r)(n_i-1)} \\ =&\sigma^2 \frac{p-r}{n_i}(1 + O_P(\frac{1}{\sqrt{(p-r)(n_i-1)}})),\end{aligned}$$

where the second line holds by central limit theorem. It follows that

$$\frac{1}{n_1}\mathrm{tr}(\tilde{V}^T S_1 \tilde{V}) + \frac{1}{n_2}\mathrm{tr}(\tilde{V}^T S_2 \tilde{V}) = \sigma^2 \tau (p-r) + o_P(\frac{\sqrt{p}}{n}). \tag{17}$$

By (16) and (17), we have

$$\frac{T_1 - \|\tilde{V}^T(\mu_1 - \mu_2)\|^2}{\sigma^2\sqrt{2\tau^2 p}} = \frac{\|\tilde{V}^T(\bar{Z}_1 - \bar{Z}_2)\|^2 - \sigma^2\tau(p-r)}{\sigma^2\sqrt{2\tau^2 p}} + o_P(1). \tag{18}$$

The proposition follows by noting that $\|\tilde{V}^T(\bar{Z}_1 - \bar{Z}_2)\|^2 \sim \sigma^2 \tau \chi^2_{p-r}$. □

Proof Of Proposition 2. Note that $(n-2)S \sim \mathrm{Wishart}_p(n-2, \Sigma)$. Denote by $\Sigma = OEO^T$ the spectral decomposition of Σ, where O 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

$$S \sim \frac{1}{n-2} O E^{1/2} Z Z^T E^{1/2} O^T.$$

Thus,

$$\begin{aligned}\hat{\sigma}^2 \sim& \frac{1}{(p-r)(n-2)} \sum_{i=r+1}^{p} \lambda_i(OE^{1/2}ZZ^T E^{1/2} O^T) \\ =& \frac{1}{(p-r)(n-2)} \sum_{i=r+1}^{n-2} \lambda_i(Z^T E Z).\end{aligned}$$

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

$$Z^T E Z = Z_{(1)}^T E_1 Z_{(1)} + \sigma^2 Z_{(2)}^T Z_{(2)},$$

where $E_1 = \mathrm{diag}(\lambda_1 + \sigma^2, \ldots, \lambda_r + \sigma^2)$. The first term is of rank r. By Weyl's inequality, we have for $i = r+1, \ldots, n-2$ that

$$\sigma^2 \lambda_i(Z_{(2)}^T Z_{(2)}) \le \lambda_i(Z^T E Z) \le \sigma^2 \lambda_{i-r}(Z_{(2)}^T Z_{(2)}).$$

It follows that

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

Hence we have

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

By Corollary 3, $\lambda_1(Z_{(2)}^T Z_{(2)}) = O_P(\max(n,p))$. Hence

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

The last line of the above equality holds since $\mathrm{tr}(Z_{(2)}^T Z_{(2)})$ is the sum of square of the elements of $Z_{(2)}$ and thus central limit theorem can be invoked. 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 1. Our proof starts with the observation that the elements of $\mu_1 - \mu_2$ is distributed as $N(0, \tau p^{-\frac{1}{2}} \psi)$. Hence

$$(\mu_1 - \mu_2)^T \Sigma (\mu_1 - \mu_2) = O(\|\mu_1 - \mu_2\|^2) = O_P(\tau p^{\frac{1}{2}}) = o_P(\tau \, \mathrm{tr}\Sigma^2).$$

Here the second equality holds by law of large number, the first and third equalities are due to boundedness of the eigenvalues of Σ. It follows that every subsequence has a further subsequence along which we have

$$(\mu_1 - \mu_2)^T \Sigma (\mu_1 - \mu_2) = o(\tau \, \mathrm{tr}\Sigma^2)$$

almost surely (a.s.). Let

$$\eta_n = \frac{T_{CQ} - \|\mu_1 - \mu_2\|^2}{\sqrt{2\tau^2 \text{tr}\Sigma^2}},$$

then by Theorem 1 in Chen and Qin (2010),

$$P(\eta_n \leq x | \mu_1, \mu_2) \to \Phi(x) \quad \text{a.s.}$$

along the further subsequence. Therefore,

$$P(\eta_n \leq x | \mu_1, \mu_2) \xrightarrow{P} \Phi(x).$$

We conclude from dominated convergence theorem that $\eta_n \xrightarrow{\mathcal{L}} N(0,1)$. What is left is to show that

$$\frac{T_{CQ} - T_2}{\sqrt{2\tau^2 \text{tr}\Sigma^2}} \xrightarrow{P} 0. \tag{19}$$

We note that

$$\begin{aligned} T_{CQ} - T_2 = & \frac{\sum_{i \neq j}^{n_1} X_{1i}^T \hat{V} \hat{V}^T X_{1j}}{n_1(n_1 - 1)} + \frac{\sum_{i \neq j}^{n_2} X_{2i}^T \hat{V} \hat{V}^T X_{2j}}{n_2(n_2 - 1)} - 2\frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} X_{1i}^T \hat{V} \hat{V}^T X_{2j}}{n_1 n_2} \\ \stackrel{\text{def}}{=} & P_1 + P_2 - 2P_3. \end{aligned}$$

And

$$\frac{P_1}{\sqrt{2\tau^2 \text{tr}\Sigma^2}} = O(1) \frac{\sum_{i \neq j}^{n_1} X_{1i}^T \hat{V} \hat{V}^T X_{1j}}{n_1 \sqrt{p}},$$

which can be further written by

$$\begin{aligned} \frac{\sum_{i \neq j}^{n_1} X_{1i}^T \hat{V} \hat{V}^T X_{1j}}{n_1 \sqrt{p}} = & \frac{n_1(n_1 - 1) \bar{X}_1^T \hat{V} \hat{V}^T \bar{X}_1}{n_1 \sqrt{p}} - \frac{\sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^T \hat{V} \hat{V}^T (X_{1i} - \bar{X}_1)}{n_1 \sqrt{p}} \\ \stackrel{\text{def}}{=} & R_1 - R_2. \end{aligned}$$

Now we deal with R_1. Since $\bar{X}_1 | \mu_1 \sim N(\mu_1, \frac{1}{n}\Sigma)$ and $\mu_1 \sim N(0, \frac{\psi}{n_1\sqrt{p}} I_p)$, we have $\bar{X}_1 \sim N(0, \frac{1}{n_1}(\Sigma + \frac{1}{\sqrt{p}}\psi I_p))$. Hence we have $\hat{V}^T \bar{X}_1 | S \sim N(0, \frac{1}{n}\hat{V}^T(\Sigma + \frac{1}{\sqrt{p}}\psi I_p)\hat{V})$ by the independence of S and $(\mu_1, \bar{X}_1)$. Therefore,

$$\begin{aligned} & E[\bar{X}_1^T \hat{V} \hat{V}^T \bar{X}_1] = EE[\bar{X}_1^T \hat{V} \hat{V}^T \bar{X}_1 | S] \\ = & E[\frac{1}{n_1} \text{tr} \hat{V}^T (\Sigma + \frac{1}{\sqrt{p}} \psi I_p) \hat{V}] = O(\frac{1}{n_1}). \end{aligned}$$

The last equality holds because the rank of $\hat{V}$ is at most R which is fixed. It follows that $R_1 \xrightarrow{P} 0$.

$$
\begin{aligned}
R_2 &= \frac{\operatorname{tr}[\hat{V}^T \sum_{i=1}^{n_1}(X_{1i}-\bar{X}_1)(X_{1i}-\bar{X}_1)^T \hat{V}]}{n_1\sqrt{p}} \\
&\leq R\frac{\lambda_1(\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_1)(X_{1i}-\bar{X}_1)^T)}{n_1\sqrt{p}}.
\end{aligned}
$$

Lemma 3 implies that $\lambda_1(\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_1)(X_{1i}-\bar{X}_1)^T) = O_P(\max(n_1,p))$. Therefore, by noting $p = o(n_1^2)$ we have $R_2 \xrightarrow{P} 0$. It follows that $\frac{P_1}{\sqrt{2\tau^2\operatorname{tr}\Sigma^2}} \xrightarrow{P} 0$. Similar arguments lead to $\frac{P_2}{\sqrt{2\tau^2\operatorname{tr}\Sigma^2}} \xrightarrow{P} 0$.

$$
\begin{aligned}
\frac{P_3}{\sqrt{2\tau^2\operatorname{tr}\Sigma^2}} &= O(1)\frac{\sqrt{n_1n_2}\bar{X}_1^T\hat{V}\hat{V}^T\bar{X}_2}{\sqrt{p}} \\
&\leq O(1)\sqrt{\frac{n_1\|\hat{V}^T\bar{X}_1\|^2}{\sqrt{p}}}\sqrt{\frac{n_2\|\hat{V}^T\bar{X}_2\|^2}{\sqrt{p}}},
\end{aligned}
$$

where the inequality is due to Cauchy inequality. By noting the relationship with R_1, the right hand side converges to 0 in probability. The proof is completed.

$\square$

Proof Of Theorem 2. 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 2, 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 5, $\frac{\max(n,p)}{n\sqrt{p}} = \max(\frac{1}{\sqrt{p}}, \frac{\sqrt{p}}{n}) \to 0$. 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

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

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 $\hat{\hat{V}}\hat{\hat{V}}^T$ only depends on S, thus

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

By the following useful relationship

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

and Lemma 2, we have that

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

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

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

or

$$\operatorname{tr}(\hat{\hat{V}}^T\Sigma\hat{\hat{V}})^2 = p\sigma^4(1+o_P(1)). \tag{20}$$

It follows that

$$\frac{\lambda_1^2(\hat{\hat{V}}^T\Sigma\hat{\hat{V}})}{\operatorname{tr}(\hat{\hat{V}}^T\Sigma\hat{\hat{V}})^2} = O_P\Big(\frac{(p/n+1)^2}{p}\Big) = o_P(1). \tag{21}$$

Then for every subsequence of $\{n\}$, there's a further subsequence along which (21) holds almost surely. By Lemma 4, for every subsequence of $\{n\}$, there's a further subsequence along which we have

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

It means that

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

Thus the weak convergence also holds unconditionally:

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

Similar to (20) we have

$$\operatorname{tr}(\hat{\hat{V}}^T\Sigma\hat{\hat{V}}) = (p-r)\sigma^2\big(1+O_P(\frac{1}{n}+\frac{1}{p})\big). \tag{22}$$

By (20), (22) and Slulsk's theorem,

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

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

Proof Of Theorem 3. The method of Theorem 2's proof can still work here with some modifications. The term P_3 in THeorem 2'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,

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

But

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

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,

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

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

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 Z, Silverstein JW. Spectral analysis of large dimensional random matrices. 2nd ed. Springer Series in Statistics. Springer, New York, 2010. doi:10.1007/978-1-4419-0661-8.

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.

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

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.

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.