A Two Sample Test for High Dimensional Data with Applications to Gene-set Testing

Chen, Song Xi and Qin, Yingli

Peking University, University of Waterloo

2010

Submitted to the Annals of Statistics

A TWO SAMPLE TEST FOR HIGH DIMENSIONAL DATA WITH APPLICATIONS TO GENE-SET TESTING

By Song Xi Chen*, Ying-Li Qin*

Iowa State University and Peking University, Iowa State University†

We proposed a two sample test for means of high dimensional data when the data dimension is much larger than the sample size. The classical Hotelling's T^2 test does not work for this "large p, small n" situation. The proposed test does not require explicit conditions on the relationship between the data dimension and sample size. This offers much flexibility in analyzing high dimensional data. An application of the proposed test is in testing significance for sets of genes, which we demonstrate in an empirical study.

1. Introduction. High dimensional data are increasingly encountered in many applications of statistics, and most prominently in biological and financial studies. A common feature for high dimensional data is that, while the data dimension is high, the sample size is relatively small. This is the so-called "large p, small n" phenomena where $p/n \to \infty$; here p is the data dimension and n is the sample size. The high data dimension ("large p") alone has created needs to renovate and rewrite some of the conventional multivariate analysis procedures. And these needs only get much greater for "large-p small-n" situations.

A specific "large p, small n" situation arises in simultaneously testing large number of hypotheses, which is largely motivated in identification of significant genes in microarray and genetic sequence studies. A natural question is how many hypotheses can be tested simultaneously. This paper tries to answer this question in the context of two sample simultaneous test for means. Consider two random samples $X_{i1}, \cdots, X_{in_i} \in R^p$ for $i = 1$ and 2, which have means $\mu_1 = (\mu_{11}, \cdots, \mu_{1p})^T$ and $\mu_2 = (\mu_{21}, \cdots, \mu_{2p})^T$, and covariance matrices Σ_1 and Σ_2, respectively. We consider testing a high dimensional hypothesis

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2. \tag{1.1}$$

*The authors acknowledge support from NSF grants SES-0518904, DMS-0604563 and DMS-0714978.

AMS 2000 subject classifications: Primary 62H15, 60K35; secondary 62G10

Keywords and phrases: high dimension; gene-set testing;, large p small n; martingale central limit theorem; multiple comparison.

The hypothesis H_0 consists of p marginal hypotheses $H_{0l} : \mu_{1l} = \mu_{2l}$ for $l = 1, \cdots, p$ regarding the means on each data dimension.

There have been a series of important studies on the high dimensional problem. van der Laan and Bryan (2001) show that the sample mean of p-dimensional data can consistently estimate the population mean uniformly across p dimensions if $\log(p) = o(n)$ for bounded random variables. In a major generalization, Kosorok and Ma (2007) consider uniform convergence for a range of univariate statistics constructed for each data dimension, which includes the marginal empirical distribution, sample mean and sample median. They establish the uniform convergence across p dimensions when $\log(p) = o(n^{1/2})$ or $\log(p) = o(n^{1/3})$ depending on the nature of the marginal statistics. Fan, Hall and Yao (2007) evaluate approximating the overall level of significance for simultaneous testing of means. They demonstrate that the bootstrap can accurately approximate the overall level of significance if $\log(p) = o(n^{1/3})$ when the marginal tests are performed based on the normal or the t-distributions. See also Fan, Peng and Huang (2005) and Huang, Wang and Zhang (2005) for high dimensional estimation and testing in semiparametric regression models.

In an important work, Bai and Saranadasa (1996) proposed using $||\bar{X}_1 - \bar{X}_2||$ to replace $(\bar{X}_1 - \bar{X}_2)^T S_n^{-1} (\bar{X}_1 - \bar{X}_2)$ in the Hotelling's T^2-statistic, where $\bar{X}_1$ and $\bar{X}_2$ are the two sample means, S_n is the pooled sample covariance by assuming $\Sigma_1 = \Sigma_2 = \Sigma$, and $|| \cdot ||$ denotes the Euclidean norm in R^p. They established the asymptotic normality of the test statistics and showed that it has attractive power property when $p/n \to c < \infty$ and under some restriction on the maximum eigenvalue of Σ. However, the requirement of p and n being of the same order is too restrictive to be used in the "large p, small n" situation.

To allow simultaneous testing for ultra high dimensional data, we construct a test which allows p being arbitrarily large independent of the sample size, as long as, in the case of common covariance Σ, $tr(\Sigma^4) = o\{tr^2(\Sigma^2)\}$, where $tr(\cdot)$ is the trace operator of a matrix. The above condition on Σ is trivially true for any p if either all the eigenvalues of Σ are bounded or the largest eigenvalue is of smaller order of $(p-b)^{1/2} b^{-1/4}$ where b is the number of unbounded eigenvalues. We establish the asymptotic normality of a test statistic, which leads to a two sample test for high dimensional data.

Testing significance for gene-sets rather than a single gene is a latest development in genetic data analysis. A critical need for gene-set testing is to have a multivariate test that is applicable for a wide range of data dimensions (the number of genes in a set). It requires P-values for all gene-sets to allow procedures based on either Bonferroni correction or the False Discovery Rate

(Benjamini and Hochberg, 1995) to take into account the multiplicity in the test. We demonstrate in this paper how to use the proposed test for testing significance for gene-sets. An advantage of the proposed test is in its readily producing P-values for significance of each gene-set under study so that the multiplicity of multiple testing can be taken into consideration.

The paper is organized as follows. We outline in Section 2 the framework of the two sample test for high dimensional data and introduce the proposed test statistic. Section 3 provides the theoretical properties of the test. How to apply the proposed test for significance of gene-sets is demonstrated in Section 4, which includes an empirical study on an Acute Lymphoblastic Leukemia data set. Results of simulation studies are reported in Section 5. All the technical details are given in Section 6.

2. Test Statistic. Suppose we have two independent and identically distributed random samples in R^p:

$$\{X_{i1}, X_{i2}, \cdots, X_{in_i}\} \overset{\text{iid}}{\sim} F_i \quad \text{for} \quad i = 1 \text{ and } 2$$

where F_i is a distribution in R^p with mean μ_i and covariance Σ_i. A well pursued interest in high dimensional data analysis is to test if the two high dimensional populations have the same mean or not, namely

$$H_0 : \mu_1 = \mu_2 \text{ v.s. } H_1 : \mu_1 \neq \mu_2. \tag{2.1}$$

The above hypotheses consist of p marginal hypotheses regarding the means of each data dimension. An important question from the point view of multiple testing is that how many marginal hypotheses can be tested simultaneously. The works of van der Laan and Bryan (2001), Kosorok and Ma (2007) and Fan, Hall and Yao (2007) are designed to address the question. The existing results show that p can reach the rate of $e^{\alpha n^\beta}$ for some positive constants α and β. In establishing a rate of the above form, both van der Laan and Bryan (2001) and Kosorok and Ma (2007) assumed that the marginal distributions of F_1 and F_2 are all supported on bounded intervals.

Hotelling's T^2 test is the conventional test for the above hypothesis when the dimension p is fixed and is less than $n =: n_1 + n_2 - 2$, and when $\Sigma_1 = \Sigma_2 = \Sigma$ say. Its performance for high dimensional data is evaluated in Bai and Saranadasa (1996) when $p/n \to c \in [0, 1)$, which reveals a decreasing power as c gets larger. A reason for this negative effect of high dimension is due to having the inverse of the covariance matrix in the T^2 statistic. While standardizing by the covariance brings benefits for data with a fixed dimension, it becomes a liability for high dimensional data. In particular, the

sample covariance matrix S_n may not converge to the population covariance when p and n are of the same order. Indeed, Yin, Bai and Krishnaiah (1988) showed that when $p/n \to c$, the smallest and the largest eigenvalues of the sample covariance S_n do not converge to the respective eigenvalues of Σ. The same phenomena, but on the weak convergence of the extreme eigenvalues of the sample covariance, are found in Tracy and Widom (1996). When $p > n$, Hotelling's T^2 statistic is not defined as S_n may not be invertible.

Our proposed test is motivated by Bai and Saranadasa (1996), who propose testing hypothesis (2.1) under $\Sigma_1 = \Sigma_2 = \Sigma$ based on

$$M_n = (\bar{X}_1 - \bar{X}_2)'(\bar{X}_1 - \bar{X}_2) - \tau tr(S_n) \tag{2.2}$$

where $S_n = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{N_i}(X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)'$ and $\tau = \frac{n_1+n_2}{n_1 n_2}$. The key feature of the Bai and Saranadasa proposal is removing S_n^{-1} in Hotelling's T^2 since having S_n^{-1} is no longer beneficial when $p/n \to c > 0$. The subtraction of $tr(S_n)$ in (2.2) is to make $E(M_n) = ||\mu_1 - \mu_2||^2$. The asymptotic normality of M_n was established and a test statistic was formulated by standardizing M_n with an estimate of its standard deviation.

The main conditions assumed in Bai-Saranadasa's test are

$$p/n \to c < \infty \quad \text{and} \quad \lambda_p = o(p^{1/2}); \tag{2.3}$$

$$n_1/(n_1 + n_2) \to k \in (0, 1) \quad \text{and} \quad (\mu_1 - \mu_2)'\Sigma(\mu_1 - \mu_2) = o\{tr(\Sigma^2)/n\} \tag{2.4}$$

where λ_p denotes the largest eigenvalue of Σ.

A careful study of M_n statistic reveals that the restrictions on p and n, and on λ_p in (2.3) are needed to control terms $\sum_{j=1}^{n_i} X_{ij}'X_{ij}$, $i = 1$ and 2, in $||\bar{X}_1 - \bar{X}_2||^2$. However, these two terms are not useful in the testing. To appreciate this point, let us consider

$$T_n =: \frac{\sum_{i\neq j}^{n_1} X_{1i}'X_{1j}}{n_1(n_1-1)} + \frac{\sum_{i\neq j}^{n_2} X_{2i}'X_{2j}}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} X_{1i}'X_{2j}}{n_1 n_2}$$

after removing $\sum_{j=1}^{n_i} X_{ij}'X_{ij}$ for $i = 1$ and 2 from $||\bar{X}_1 - \bar{X}_2||^2$. Elementary derivations show that

$$E(T_n) = ||\mu_1 - \mu_2||^2.$$

Hence, T_n is basically all we need for testing. Bai and Saranadasa used $tr(S_n)$ to offset the two diagonal terms. However, $tr(S_n)$ itself impose demands on the dimensionality too.

A derivation in the appendix shows that under H_1 and the second condition in (2.4)

$$Var(T_n) = \left\{\frac{2}{n_1(n_1-1)}tr(\Sigma_1^2) + \frac{2}{n_2(n_2-1)}tr(\Sigma_2^2) + \frac{4}{n_1 n_2}tr(\Sigma_1\Sigma_2)\right\}\{1 + o(1)\}$$

where the $o(1)$ term vanishes under H_0.

3. Main Results. We assume, like Bai and Saranadasa (1996), the following general multivariate model

$$X_{ij} = \Gamma_i Z_{ij} + \mu_i \quad \text{for } j = 1, \cdots, n_i,\ i = 1 \text{ and } 2 \tag{3.1}$$

where each Γ_i is a $p \times m$ matrix for some $m \geq p$ such that $\Gamma_i \Gamma_i' = \Sigma_i$, and $\{Z_{ij}\}_{j=1}^{n_i}$ are m-variate independent and identically distributed (IID) random vectors satisfying $E(Z_{ij}) = 0$, $Var(Z_{ij}) = I_m$, the $m \times m$ identity matrix. Furthermore, if we write $Z_{ij} = (z_{ij1}, \ldots, z_{ijm})'$, we assume $E(z_{ijk}^4) = 3 + \Delta < \infty$, and

$$E\left(z_{ijl_1}^{\alpha_1} z_{ijl_2}^{\alpha_2} \cdots z_{ijl_q}^{\alpha_q}\right) = E(z_{ijl_1}^{\alpha_1}) E(z_{ijl_2}^{\alpha_2}) \cdots E(z_{ijl_q}^{\alpha_q}) \tag{3.2}$$

for a positive integer q such that $\sum_{l=1}^{q} \alpha_l \leq 8$ and $l_1 \neq l_2 \neq \cdots \neq l_q$. Here Δ describes the difference between the fourth moments of z_{ijl} and $N(0,1)$. Model (3.1) says that X_{ij} can be expressed as a linear transformation of a m-variate Z_{ij} with zero mean and unit variance that satisfies (3.2). Model (3.1) is similar to factor models in multivariate analysis. However, instead of having the number of factors $m < p$ in the conventional multivariate analysis, we require $m \geq p$. This is to allow the basic characteristics of the covariance Σ_i, for instance its rank and eigenvalues, are not affected by the transformation. The rank and eigenvalues would be affected if $m < p$. The fact that m is arbitrary offers much flexibility in generating a rich collection of dependence structure. Condition (3.2) means that each Z_{ij} has a kind of pseudo-independence among its components $\{z_{ijl}\}_{l=1}^{m}$. Obviously, if Z_{ij} does have independent components, then (3.2) is trivially true.

We do not assume $\Sigma_1 = \Sigma_2$, as it is a rather strong assumption, and most importantly such an assumption is harder to be verified for high dimensional data. Testing certain special structures of the covariance matrix when p and n are of the same order have been considered in Ledoit and Wolf (2002) and Schott (2005).

We assume

$$n_1/(n_1 + n_2) \to k \in (0,1) \quad \text{as} \quad n \to \infty \tag{3.3}$$

$$(\mu_1 - \mu_2)' \Sigma_i (\mu_1 - \mu_2) = o[n^{-1} tr\{(\Sigma_1 + \Sigma_2)^2\}] \quad \text{for } i = 1 \text{ or } 2 \tag{3.4}$$

which generalize (2.4) to unequal covariances. Condition (3.4) is obviously satisfied under H_0 and implies that the difference between μ_1 and μ_2 is small relative to $n^{-1} tr\{(\Sigma_1 + \Sigma_2)^2\}$ so that a workable expression for the variance of T_n under H_0 and the specified local alternative can be derived. It can be viewed as a high dimensional version of the local alternative hypotheses.

When p is fixed, if we use a standard test for two population means, for instance Hotelling's T^2 test, the local alternative hypotheses has the form of $\mu_1 - \mu_2 = \tau n^{-1/2}$ for a non-zero constant vector $\tau \in R^p$. The Hotelling's test has non-trivial power under such local alternatives (Anderson, 2000). If we assume each components of $\mu_1 - \mu_2$ is the same, say δ, then the local alternatives imply $\delta = O(n^{-1/2})$ for a fixed p. When the difference is $o(n^{-1/2})$, Hotelling's test has non-power beyond the level of significance.

To gain insight on (3.4) for high dimensional situations, let us assume all the eigen-values of Σ_i are bounded above from infinity and below away from zero so that $\Sigma_i = I_p$ is a special case of such regime. Let us also assume like above each component of $\mu_1 - \mu_2$ is the same as a fixed δ, namely $\mu_{1l} - \mu_{2l} = \delta$ for $l = 1, \cdots, p$. Then, (3.4) implies $\delta = o(n^{-1/2})$ which is a smaller order than $\delta = O(n^{-1/2})$ for fixed p case. This can be understood as the high dimensional data ($p \to \infty$) contain more data information which allows finer resolution in differentiating the two means in each component than that in the fixed p case.

To understand the performance of the test when (3.4) is not valid, we reverse the local alternative condition (3.4) to

$$n^{-1}tr\{(\Sigma_1 + \Sigma_2)^2\} = o\{(\mu_1 - \mu_2)'\Sigma_i(\mu_1 - \mu_2)\} \quad \text{for } i = 1 \text{ or } 2, \tag{3.5}$$

implying the Mahanalobis distance between μ_1 and μ_2 is a larger order than that of $n^{-1}tr\{(\Sigma_1+\Sigma_2)^2\}$. This condition can be viewed as a version of fixed alternatives. We will establish asymptotic normally of T_n under either (3.4) and (3.5) in Theorem 1.

The condition we impose on p to replace the first part of (2.3) is

$$tr(\Sigma_i\Sigma_j\Sigma_l\Sigma_h) = o[tr^2\{(\Sigma_1 + \Sigma_2)^2\}] \quad \text{for } i, j, l, h = 1 \text{ or } 2, \tag{3.6}$$

as $p \to \infty$. To appreciate this condition, consider the case of $\Sigma_1 = \Sigma_2 = \Sigma$. Then, (3.6) becomes

$$tr(\Sigma^4) = o\{tr^2(\Sigma^2)\}. \tag{3.7}$$

Let $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_p$ be the eigenvalues of Σ. If all eigenvalues are bounded, then (3.7) is trivially true. If, otherwise, there are b unbounded eigenvalues with respect to p and the remaining $p - b$ eigenvalues are bounded above by a finite constant M such that $(p - b) \to \infty$ and $(p - b)\lambda_1^2 \to \infty$, then sufficient conditions for (3.7) are

$$\lambda_p = o\{(p - b)^{1/2}\lambda_1 b^{-1/4}\} \quad \text{or} \quad \lambda_p = o\{(p - b)^{1/4}\lambda_1^{1/2}\lambda_{p-b+1}^{1/2}\} \tag{3.8}$$

where b can be either bounded or diverging to infinity and the smallest eigen-value λ_1 can converge to zero. To appreciate these, we note that

$$\frac{tr(\Sigma^4)}{tr^2(\Sigma^2)} \leq \frac{(p-b)M^4 + b\lambda_p^4}{(p-b)^2\lambda_1^4 + b^2\lambda_{p-b+1}^4 + 2(p-b)b\lambda_1^2\lambda_{p-b+1}^2},$$

Hence, the ratio converges to 0 under either condition in (3.8).

The following theorem establishes the asymptotic normality of T_n.

Theorem 1. Under the assumptions (3.1), (3.2), (3.3), (3.6) and either (3.4) or (3.5),

$$\frac{T_n - ||\mu_1 - \mu_2||^2}{\sqrt{Var(T_n)}} \xrightarrow{\mathrm{d}} N(0,1) \quad \text{as } p \to \infty \text{ and } n \to \infty.$$

The asymptotic normality is attained without imposing any explicit restriction between p and n directly. The only restriction on the dimension is (3.6) or (3.7). As the discussion given just before Theorem 1 suggests, (3.7) is satisfied provided that the number of divergent eigenvalues of Σ are not too many and the divergence is not too fast. The reason for attaining this in the case of high data dimension is because the statistic T_n is univariate, despite the hypothesis H_0 is of high dimensional. This is different from using a high dimensional statistic. Indeed, Portnoy (1986) considers the central limit theorem for the p-dimensional sample mean $\bar{X}$ and finds that the central limit theorem is not valid if p is not a smaller order of $\sqrt{n}$.

As shown in Section 6.1, $Var(T_n) = \sigma_n^2\{1 + o(1)\}$ where, under (3.4)

$$\sigma_n^2 =: \sigma_{n1}^2 = \tfrac{2}{n_1(n_1-1)}tr(\Sigma_1^2) + \tfrac{2}{n_2(n_2-1)}tr(\Sigma_2^2) + \tfrac{4}{n_1 n_2}tr(\Sigma_1\Sigma_2), \tag{3.9}$$

and under (3.5)

$$\sigma_n^2 =: \sigma_{n2}^2 = \frac{4}{n_1}(\mu_1-\mu_2)'\Sigma_1(\mu_1-\mu_2) + \frac{4}{n_2}(\mu_1-\mu_2)'\Sigma_2(\mu_1-\mu_2). \tag{3.10}$$

In order to formulate a test procedure based on Theorem 1, σ_{n1}^2 in (3.9) needs to be estimated. Bai and Saranadasa (1996) used the following estimator for $tr(\Sigma^2)$ under $\Sigma_1 = \Sigma_2 = \Sigma$:

$$\widehat{tr(\Sigma^2)} = \tfrac{n^2}{(n+2)(n-1)}\{trS_n^2 - \tfrac{1}{n}(trS_n)^2\}.$$

Motivated by the benefits of excluding terms like $\sum_{j=1}^{n_i} X_{ij}'X_{ij}$ in the formulation of T_n, we propose the following estimator of $tr(\Sigma_i^2)$ and $tr(\Sigma_1\Sigma_2)$:

$$\widehat{tr(\Sigma_i^2)} = \{n_i(n_i-1)\}^{-1}tr\{\Sigma_{j\neq k}^{n_i}(X_{ij} - \bar{X}_{i(j,k)})X_{ij}'(X_{ik} - \bar{X}_{i(j,k)})X_{ik}'\},$$

and

$$tr(\widehat{\Sigma_1\Sigma_2}) = (n_1n_2)^{-1}tr\{\Sigma_{l=1}^{n_1}\Sigma_{k=1}^{n_2}(X_{1l} - \bar{X}_{1(l)})X'_{1l}(X_{2k} - \bar{X}_{2(k)})X'_{2k}\}$$

where $\bar{X}_{i(j,k)}$ is the i-th sample mean after excluding X_{ij} and X_{ik}, and $\bar{X}_{i(l)}$ is the i-th sample mean without X_{il}. These are similar to the idea of cross-validation, in that when we construct the deviations of X_{ij} and X_{ik} from the sample mean, both X_{ij} and X_{ik} are excluded from the sample mean calculation. By doing so, the above estimators $tr(\widehat{\Sigma_i^2})$ and $tr(\widehat{\Sigma_1\Sigma_2})$ can be written as the trace of sums of products of independent matrices. We also note that subtraction of only one sample mean per observation is needed in order to avoid term like $||X_{ij}||^4$ which is harder to control asymptotically without an explicit assumption between p and n.

The next theorem shows that the above estimators are ratio-consistent to $tr(\Sigma_i^2)$ and $tr(\Sigma_1\Sigma_2)$, respectively.

Theorem 2. Under the assumptions (3.1)-(3.4) and (3.6), for $i = 1$ or 2,

$$\frac{tr(\widehat{\Sigma_i^2})}{tr(\Sigma_i^2)} \xrightarrow{\text{p}} 1 \quad \text{and} \quad \frac{tr(\widehat{\Sigma_1\Sigma_2})}{tr(\Sigma_1\Sigma_2)} \xrightarrow{\text{p}} 1 \quad \text{as } p \text{ and } n \to \infty.$$

A ratio-consistent estimator of σ_{n1}^2 under H_0 is

$$\hat{\sigma}_{n1}^2 = \tfrac{2}{n_1(n_1-1)}tr(\widehat{\Sigma_1^2}) + \tfrac{2}{n_2(n_2-1)}tr(\widehat{\Sigma_2^2}) + \tfrac{4}{n_1n_2}tr(\widehat{\Sigma_1\Sigma_2}).$$

This together with Theorem 1 leads to the test statistic

$$Q_n = T_n/\hat{\sigma}_{n1} \xrightarrow{d} N(0,1) \quad \text{as } p \text{ and } n \to \infty$$

under H_0. The proposed test with an α level of significance rejects H_0 if $Q_n > \xi_\alpha$ where ξ_α is the upper α quantile of $N(0,1)$.

Theorems 1 and 2 allow us to discuss the power properties of the proposed test. The discussion is made under (3.4) and (3.5) respectively. The power under the local alternative (3.4) is

$$\beta_{n1}(||\mu_1 - \mu_2||) = \Phi\left(-\xi_\alpha + \frac{nk(1-k)||\mu_1 - \mu_2||^2}{\sqrt{2tr\{\tilde{\Sigma}(k)^2\}}}\right), \tag{3.11}$$

where $\tilde{\Sigma}(k) = (1-k)\Sigma_1 + k\Sigma_2$ and Φ is the standard normal distribution function. The power of Bai-Saranadasa test has the same form if $\Sigma_1 = \Sigma_2$ and if p and n are of the same order.

The power under (3.5) is

$$\beta_{n2}(||\mu_1-\mu_2||)=\Phi\left(-\frac{\sigma_{n1}}{\sigma_{n2}}\xi_\alpha+\frac{||\mu_1-\mu_2||^2}{\sigma_{n1}}\right)=\Phi\left(\frac{||\mu_1-\mu_2||^2}{\sigma_{n1}}\right)$$

as $\sigma_{n1}/\sigma_{n2}\to 0$. Substitute the expression for σ_{n1}, we have

$$\beta_{n2}(||\mu_1-\mu_2||)=\Phi\left(\frac{nk(1-k)||\mu_1-\mu_2||^2}{\sqrt{2tr\{\tilde{\Sigma}(k)^2\}}}\right). \tag{3.12}$$

Both (3.11) and (3.12) indicate that the proposed test has non-trivial power under the two cases of the alternative hypothesis as long as

$$n||\mu_1-\mu_2||^2/\sqrt{tr\{\tilde{\Sigma}(k)^2\}}$$

does not vanish to 0 as n and $p\to\infty$. The flavor of the proposed test is different from tests formulated by combining p marginal tests on H_{0l} (defined after (1.1)) for $l=1,\ldots,p$. The test statistics of such tests are usually constructed via $\max_{1\le l\le p}T_{nl}$, where T_{nl} is a marginal test statistic for H_{0l}. This is the case of Kosorok and Ma (2007) and Fan, Hall and Yao (2007). A condition on p and n is needed to ensure (i) the convergence of $\max_{1\le l\le p}T_{nl}$, and (ii) p can reach an order of $exp(\alpha n^\beta)$ for positive constants α and β. Usually some additional assumptions are needed: for instance, Kosorok and Ma (2007) assumed each component of the random vector has compact support for testing means.

Naturally, if the number of significant univariate hypotheses ($\mu_{1l}\neq\mu_{2l}$) is a lot less than p, which is the so-called sparsity scenario, a simultaneous test like the one we propose may encounter a loss of power. This is actually quantified by the power expression (3.11). Without loss of generality, suppose that each μ_i can be partitioned as $(\mu_i^{(1)'},\mu_i^{(2)'})'$ so that under $H_1:\mu_1^{(1)}=\mu_2^{(1)}$ and $\mu_1^{(2)}\neq\mu_2^{(2)}$, where $\mu_i^{(1)}$ is of p_1 dimensional and $\mu_i^{(2)}$ is of p_2 dimensional and $p_1+p_2=p$. Then, $||\mu_1-\mu_2||=p_2\delta^2$ for some positive constant δ^2. Suppose that λ_{m_0} be the smallest non-zero eigenvalue of $\tilde{\Sigma}(k)$. Then, under the local alternative (3.4), the asymptotic power is bounded above and below by

$$\Phi\left(-\xi_\alpha+\frac{nk(1-k)p_2\delta^2}{\sqrt{2p}\lambda_p}\right)\le\beta(||\mu_1-\mu_2||)\le\Phi\left(-\xi_\alpha+\frac{nk(1-k)p_2\delta^2}{\sqrt{2(p-m_0)}\lambda_{m_0}}\right).$$

If p is very large relative to n and p_2 under both high dimensionality and sparsity, so that $nk(1-k)p_2\eta^2/\sqrt{2(p-m_0)}\to 0$, the test could endure

low power. With this in mind, we check on the performance of the test under sparsity in simulation studies in Section 5. The simulations showed that the proposed test had a robust power and was in fact can be more powerful than tests based on multiple comparison with either the Bonferroni and False Discovery Rate (FDR) procedures. We note here that, due to the multivariate nature of the test and the hypothesis, the proposed test cannot identify which components are significant after the null multivariate hypothesis is rejected. Additional follow-up procedures have to be employed for that purpose. The proposed test come very handy when the interest is to identify significant groups of components, like sets of genes as illustrated in Section 4. The above discussion can be readily extended to the case of (3.5) due to the similarity in the two power functions.

The proposed two sample test can be modified for paired observations $\{(Y_{i1}, Y_{i2})\}_{i=1}^{n}$ where Y_{i1} and Y_{i2} are two measurements of p-dimensions on a subject i before and after a treatment. Let $X_i = Y_{i2} - Y_{i1}$, $\mu = E(X_i)$ and $\Sigma = Var(X_i)$. This is effectively a one sample problem with high dimensional data. The hypothesis of interest is

$$H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu \neq 0.$$

We can use $F_n = \Sigma_{i\neq j}^{n_1} X_i' X_j / \{n(n-1)\}$ as the test statistic. It is readily shown that $E(F_n) = \mu'\mu$ and $Var(F_n) = \frac{2}{n_1(n_1-1)} tr(\Sigma_1^2)\{1 + o(1)\}$ under both H_0 and H_1 if we assume a condition similar to (3.4) so that $\mu'\Sigma\mu = o\{n^{-1}tr(\Sigma^2)\}$. And the asymptotic normality of F_n by adding $tr(\Sigma^4) = o\{tr^2(\Sigma^2)\}$, a variation of (3.6), can be established by utilizing part of the proof on the asymptotic normality of T_n. The $tr(\Sigma^2)$ can be ratio-consistently estimated with n_1 replaced by n in $\widehat{tr(\Sigma_1^2)}$, which leads to a ratio-consistent variance estimation for F_n. Then, the test and its power can be written out in similar ways as those for the two sample test.

When $p = O(1)$, which may be viewed as having finite dimension, the asymptotic normality as conveyed in Theorem 1 may not be valid anymore. It may be shown under Conditions (3.1)-(3.4) without (3.6), as condition (3.6) is no longer relevant when p is bounded, the test statistic $(n_1 + n_2)T_n$ converges to $\sum_{l=1}^{2p} \eta_l \chi_{1,l}^2$, where $\{\chi_{1,l}^2\}_{l=1}^{2p}$ are independent χ_1^2 distributed random variables and $\{\eta_l\}_{l=1}^{2p}$ is a set of constants. Theorem 2 remains valid when p is bounded under (3.1)-(3.4). The proposed can still be used for testing in this situation of bounded dimension with estimated critical values via estimation of $\{\eta_l\}_{l=1}^{2p}$. However, people may like to use a test specially cantered for such case, for instance, the Hotelling's test.

4. Gene-set Testing. Identifying sets of genes which are significant with respect to certain treatments is a latest development in genetics research; see Barry, Nobel and Wright (2005), Recknor, Dettleton and Reecy (2007), Efron and Tibshrini (2007) and Newton *et.al* (2007). Biologically speaking, each gene does not function individually in isolation. Rather, one gene tends to work with other genes to achieve certain biological tasks.

Suppose that $\mathcal{S}_1, \cdots, \mathcal{S}_q$ be q sets of genes, where the gene-set $\mathcal{S}_g$ consists of p_g genes. Let $F_{1\mathcal{S}_g}$ and $F_{2\mathcal{S}_g}$ be the distribution functions corresponding to $\mathcal{S}_g$ under the treatment and control, and $\mu_{1\mathcal{S}_g}$ and $\mu_{2\mathcal{S}_g}$ be their respective means. The hypothesis of interest is

$$H_{0g} : \mu_{1\mathcal{S}_g} = \mu_{2\mathcal{S}_g} \quad \text{for} \quad g = 1, \cdots, q.$$

The gene sets $\{\mathcal{S}_g\}_{g=1}^{q}$ can overlap as a gene can belong to several functional groups, and p_g, the number of genes in a set, can range from a moderate to a very large number. So, there are issues of both multiplicity and high dimensionality in gene-set testing.

We propose applying the proposed test for significance of each gene-set $\mathcal{S}_g$ when p_g is large. When p_g is of low dimension, the Hotelling's test may be used. Let pv_g, $g = 1, \cdots, q$ be the P-values obtained from these tests. To control the overall family-wise error rate, we can employ the Bonferroni procedure; to control FDR, we can use Benjamini and Hochberg (1995)'s method or its variations as in Benjamini and Yekutieli (2001) and Storey, Taylor and Siegmund (2004). These lead to control of the family-wise error rate or FDR in the context of gene-sets testing. In contrast, tests based on univariate testing have difficulties in producing P-values for gene-sets.

Acute Lymphoblastic Leukemia (ALL) is a form of leukemia, a cancer of white blood cells. The ALL data (Chiaretti et al., 2004) contains microarray expressions for 128 patients with either T-cell or B-cell type Leukemia. Within the B-cell type leukemia, there are two sub-classes representing two molecular classes: the BCR/ABL class and NEG class. The data set has been analyzed by Dudoit, Keles and van der Laan (2006) using a different technology.

Gene-sets are technically defined in Gene Ontology (GO) system that provides structured and controlled vocabularies producing names of gene-sets (also called GO terms). There are three groups of Gene ontologies of interest: Biological Processes (BP), Cellular Components (CC) and Molecular Functions (MF). We carried out preliminary screening for gene-filtering using the approach in Gentleman et al. (2005), which left 2391 genes for analysis. There are 575 unique GO terms in BP category, 221 in MF and 154 in CC for the ALL data. The largest gene-set contains 2059 genes in

BP, 2112 genes in MF and 2078 genes in CC; and the GO terms of the three categories share 1861 common genes. We are interested in detecting differences in the expression levels of gene-sets between the BCR/ABL molecular sub-class ($n_1 = 42$) and the NEG molecular sub-class ($n_2 = 37$) for each of the three categories.

We applied the proposed two sample test with 5% significant level to test each of the gene-sets in conjunction with the Bonferroni correction to control the family-wise error rate at 0.05 level. It was found that there were 259 gene-sets declared significant in the BP group, 110 in the MF group and 53 in the CC group. Figure 1 displays the histograms of the P-values and the values of test statistic Q_n for the three gene-categories. It shows a strong non-uniform distribution of the P-values with a large number of P-values clustered near 0. At the same time, the Q_n-value plots indicate the average Q_n-values were much larger than zero. These explain the large number of significant gene-sets detected by the proposed test.

The number of the differentially expressed gene-sets may seem to be high. This was mainly due to overlapping gene-sets. To appreciate this point, we computed for each (say i-th) significant gene-set, the number of other significant gene-sets which overlapped with it, say b_i; and obtained the average of $\{b_i\}$ and their standard deviation. The average number of overlaps (standard deviation) for BP group was 198.9(51.3), 55.6 (25.2) for MF, and 41.6 (9.5) for CC. These number are indeed very high and reveals the gene-sets and their P-values were highly dependent.

Finally, we carried out back-testing for the same hypothesis by randomly splitting the 42 BRC/ABL class into two sub-class of equal sample size and testing for mean differences. This set-up led to the situation of H_0. Figures 2 reports the P-values and Q_n-values for the three Gene Ontology groups. We note that the distributions of the P-values were much closer to the uniform distribution than Figure 1. It is observed that the histograms of Q_n-values were centered close to zero and were much closer to the normal distribution than their counterparts in Figure 1, which were reassuring.

5. Simulation Studies. In this section, we report results from simulation studies which were designed to evaluate the performance of the proposed two sample test for high dimensional data. For comparison, we also conducted the test proposed by Bai and Saranadasa (1996) (BS test), and two tests based on multiple comparison procedures by employing the Bonferroni and the FDR control (Benjamini and Hochberg, 1995). Both procedures control the family-wise error rate at a level of significance α which coincides with the significance for the proposed test and the BS test. In the two mul-

tiple comparison procedures, we conducted univariate two sample t-tests for univariate hypotheses $H_{0l} : \mu_{1l} = \mu_{2l}$ versus $\mu_{1l} \neq \mu_{2l}$ for $l = 1, 2, \cdots, p$.

Two simulation models for X_{ij} were considered. One had a moving average structure that allows a general dependent structure; the other could allocate the the alternative hypotheses sparsely which enable us to evaluate the performance of the tests under sparsity.

5.1. *Moving Average Model.* The first simulation model has the following moving average structure:

$$X_{ijk} = \rho_1 Z_{ijk} + \rho_2 Z_{ijk+1} + \cdots + \rho_p Z_{ijk+p-1} + \mu_{ij}$$

for $i = 1$ and 2, $j = 1, 2, \cdots, n_i$ and $k = 1, 2, \cdots, p$, where $\{Z_{ijk}\}$ were respectively IID random variables. We considered two distributions for the innovations $\{Z_{ijk}\}$. One was a centralized Gamma$(4, 1)$ so that it has zero mean, and the other was $N(0, 1)$.

For each distribution of $\{Z_{ijk}\}$, we considered two configurations of dependence among components of X_{ij}. One had weaker dependence with $\rho_l = 0$ for $l > 3$. This prescribed a "two dependence" moving average structure where X_{ijk_1} and X_{ijk_2} are dependent only if $|k_1 - k_2| \leq 2$. The $\{\rho_l\}_{l=1}^3$ were generated independently from $U(2, 3)$, which were $\rho_1 = 2.883$, $\rho_2 = 2.794$ and $\rho_3 = 2.849$ and were kept fixed throughout the simulation. The second configuration had all ρ_l's generated from $U(2, 3)$, and were again keep fixed throughout the simulation. We call this the "full dependence case". The above dependence structures assigned equal covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma$ and allows a meaningful comparison with BS test.

Without loss of generality, we fixed $\mu_1 = 0$ and chose μ_2 in the same fashion as Benjamini and Hochberg (1995). Specifically, the percentage of true null hypotheses $\mu_{1l} = \mu_{2l}$ for $l = 1, \cdots, p$ were chosen to be 0%,25%,50%,75%, 95% and 99% and 100% , respectively. Experimenting 95% and 99% was designed to gain information on the performance of the test when $\mu_{1l} \neq \mu_{2l}$ were sparse. It provided empirical checks on the potential concerns on the power of the simultaneous high dimensional test as made at the end of Section 3. At each percentage level of true null, three patterns of allocation were considered for the non-zero μ_{2l} in $\mu_2 = (\mu_{21}, \cdots, \mu_{2p})'$: (i) the equal allocation where all the non-zero μ_{2l} were equal; (ii) linearly increasing and (iii) linearly decreasing allocations as specified in Benjamini and Hochberg (1995). To make the power comparable among the configurations of H_1, we set $\eta =: ||\mu_1 - \mu_2||^2/\sqrt{tr(\Sigma^2)} = 0.1$ throughout the simulation. We chose $p = 500$ and 1000 and $n = [20 \log(p)] = 124$ and 138, respectively.

Tables 1 and 2 report the empirical power and size of the four tests with Gamma innovations at 5% nominal significance level or family-wise error

rate or FDR based on 5000 simulations. The results for the Normal innovations had similar pattern, and are not reported here. The simulation results in Tables 1 and 2 can be summarized as follows. The proposed test was much more powerful than Bai-Saranadasa test for all cases considered in the simulation, while maintaining a reasonable size approximation to the nominal 5% level. Both the proposed test and Bai-Saranadasa test were more powerful than the two tests based on the multiple univariate testing using the Bonferroni and FDR procedures. This is a little expected as both the proposed and Bai-Saranadasa test are designed to test for the entire p-dimensional hypotheses while the multiple testing procedures are targeted at the individual univariate hypothesis. What is surprising is that when the percentage of true null was high at 95% and 99%, the proposed test still were much more powerful than the two multiple testing procedures for all three allocations of the non-zero components in μ_2. It is observed that the sparsity (95% and 99% true null) does reduce the power of the proposed test a little. However, the proposed test still enjoyed good power, especially comparing with the other three tests. We also observe that when there was more dependence among multivariate components of the data vectors in the full dependence model, there was a drop in the power for each of the test. The power of the tests based on the Bonferroni and FDR procedures were alarmingly low and were only slightly larger than the nominal significance level.

We also collected information on the quality of $tr(\Sigma^2)$ estimation. Table 3 reports empirical averages and standard deviation of $\widehat{tr(\Sigma^2)}/tr(\Sigma^2)$. It shows that the proposed estimator for $tr(\Sigma^2)$ has much smaller bias and standard deviation than those proposed in Bai and Saranadasa (1996) in all cases, and provides an empirical verification for Theorem 2.

5.2. *Sparse Model.* An examination of the previous simulation setting reveals that the strength of the "signals" $\mu_{2l} - \mu_{1l}$ corresponding to the alternative hypotheses were low relative to the level of noise (variance), which may not be a favorable situation for the two tests based on multiple univariate testing. To gain more information on the performance of the tests under sparsity, we considered the following simulation model such that

$$X_{1il} = Z_{1il} \quad \text{and} \quad X_{2il} = \mu_l + Z_{2il} \quad \text{for } l = 1, \ldots, p$$

where $\{Z_{1il}, Z_{2il}\}_{l=1}^{p}$ are mutually independent $N(0,1)$ random variables, and the "signals"

$$\mu_l = \varepsilon\sqrt{2\log(p)} \quad \text{for } l = 1, \ldots, q = [p^c] \text{ and} \quad \mu_l = 0 \quad \text{for } l > q$$

for some $c \in (0,1)$. Here q is the number of significant alternative hypotheses. The sparsity of the hypotheses is determined by c: the smaller the c is, the more sparse the alternative hypotheses with $\mu_l \neq 0$. This simulation model is similar to the one used in Abramovich, Benjamini, Donoho and Johnstone (2006).

According to (3.11), the power of the proposed test has asymptotic power

$$\beta(||\mu||) = \Phi\left(-\xi_\alpha + \frac{np^{(c-1/2)}\varepsilon^2 log(p)}{2\sqrt{2}}\right)$$

which indicates that the test has a much reduced power if $c < 1/2$ with respect to p. We, therefore, chose $p = 1000$ and $c = 0.25, 0.35, 0.45$ and 0.55 respectively, which led to $q = 6, 11, 22$, and 44 respectively. We call $c = 0.25, 0.35$ and 0.45 the sparse cases.

In order to prevent trivial powers of α or 1 in the simulation, we set $\varepsilon = 0.25$ for $c = 0.25$ and 0.45; and $\varepsilon = 0.15$ for $c = 0.35$ and 0.55. Table 4 summarizes the simulations results based on 500 simulations. It shows that in the extreme sparse cases of $c = 0.25$, the FDR and Bonferroni tests did had lower power than the proposed test. The power were largely similar among the three tests for $c = 0.35$. However, when the sparsity was moderated to $c = 0.45$, the proposed test started to surpass the FDR and Bonferroni procedures. The gap in power performance was further increased when $c = 0.55$. Table 5 reports the quality of the variance estimation in Table 5, which shows the proposed variance estimators incurs very little bias and variance for even very small sample sizes of $n_1 = n_2 = 10$.

6. Technical Details.

6.1. *Derivations for $E(T_n)$ and $Var(T_n)$.* As

$$T_n = \frac{\Sigma_{i\neq j}^{n_1} X_{1i}' X_{1j}}{n_1(n_1-1)} + \frac{\Sigma_{i\neq j}^{n_2} X_{2i}' X_{2j}}{n_2(n_2-1)} - 2\frac{\Sigma_{i=1}^{n_1}\Sigma_{j=1}^{n_2} X_{1i}' X_{2j}}{n_1 n_2},$$

it is straight forward to show that $E(T_n) = \mu_1'\mu_1 + \mu_2'\mu_2 - 2\mu_1'\mu_2 = ||\mu_1 - \mu_2||^2$.

Let $P_1 = \frac{\Sigma_{i\neq j}^{n_1} X_{1i}' X_{1j}}{n_1(n_1-1)}$, $P_2 = \frac{\Sigma_{i\neq j}^{n_2} X_{2i}' X_{2j}}{n_2(n_2-1)}$ and $P_3 = -2\frac{\Sigma_{i=1}^{n_1}\Sigma_{j=1}^{n_2} X_{1i}' X_{2j}}{n_1 n_2}$. It can be shown that

$$Var(P_1) = \frac{2}{n_1(n_1-1)} tr(\Sigma_1^2) + \frac{4\mu_1'\Sigma_1\mu_1}{n_1},$$

$$Var(P_2) = \frac{2}{n_2(n_2-1)} tr(\Sigma_2^2) + \frac{4\mu_2'\Sigma_2\mu_2}{n_2} \quad \text{and}$$

$$Var(P_3) = \frac{4}{n_1 n_2} tr(\Sigma_1\Sigma_2) + \frac{4\mu_2'\Sigma_1\mu_2}{n_1} + \frac{4\mu_1'\Sigma_2\mu_1}{n_2}.$$

Because two samples are independent, $Cov(P_1, P_2) = 0$. Also,

$$Cov(P_1, P_3) = -\frac{4\mu_1'\Sigma_1\mu_2}{n_1} \quad \text{and} \quad Cov(P_2, P_3) = -\frac{4\mu_1'\Sigma_2\mu_2}{n_2}.$$

In summary,

$$\begin{aligned} Var(T_n) = \quad & \frac{2}{n_1(n_1-1)} tr(\Sigma_1^2) + \frac{2}{n_2(n_2-1)} tr(\Sigma_2^2) + \frac{4}{n_1 n_2} tr(\Sigma_1\Sigma_2) \\ & + \frac{4}{n_1}(\mu_1-\mu_2)'\Sigma_1(\mu_1-\mu_2) + \frac{4}{n_2}(\mu_1-\mu_2)'\Sigma_2(\mu_1-\mu_2). \end{aligned}$$

Thus, under H_0

$$Var(T_n) = \sigma_{n1}^2 =: \frac{2}{n_1(n_1-1)} tr(\Sigma_1^2) + \frac{2}{n_2(n_2-1)} tr(\Sigma_2^2) + \frac{4}{n_1 n_2} tr(\Sigma_1\Sigma_2).$$

Under $H_1 : \mu_1 \neq \mu_2$, with (3.4),

$$Var(T_n) = \sigma_{n1}^2\{1 + o(1)\};$$

and with (3.5),

$$Var(T_n) = \sigma_{n2}^2\{1 + o(1)\}$$

where $\sigma_{n2} = \frac{4}{n_1}(\mu_1-\mu_2)'\Sigma_1(\mu_1-\mu_2) + \frac{4}{n_2}(\mu_1-\mu_2)'\Sigma_2(\mu_1-\mu_2)$.

6.2. *Asymptotic Normality of T_n.* We note that $T_n = T_{n1} + T_{n2}$ where

$$\begin{aligned} T_{n1} \quad = \quad & \frac{\Sigma_{i\neq j}^{n_1}(X_{1i}-\mu_1)'(X_{1j}-\mu_1)}{n_1(n_1-1)} + \frac{\Sigma_{i\neq j}^{n_2}(X_{2i}-\mu_2)'(X_{2j}-\mu_2)}{n_2(n_2-1)} \\ & -2\frac{\Sigma_{i=1}^{n_1}\Sigma_{j=1}^{n_2}(X_{1i}-\mu_1)'(X_{2j}-\mu_2)}{n_1 n_2}. \end{aligned} \tag{6.1}$$

and

$$\begin{aligned} T_{n2} = & \frac{\Sigma_{i=1}^{n_1}(X_{1i}-\mu_1)'(\mu_1-2\mu_2)}{n_1} + \frac{\Sigma_{i=1}^{n_2}(X_{2i}-\mu_2)'(\mu_2-2\mu_1)}{n_2} \\ & + \mu_1'\mu_1 + \mu_2'\mu_2 - 2\mu_1'\mu_2. \end{aligned}$$

It is easy to show that $E(T_{n1}) = 0$ and $E(T_{n2}) = ||\mu_1-\mu_2||^2$, and

$$\begin{aligned} & Var(T_{n2}) \\ & = 4n_1^{-1}(\mu_1-\mu_2)'\Sigma_1(\mu_1-\mu_2) + 4n_2^{-1}(\mu_2-\mu_1)'\Sigma_2(\mu_2-\mu_1). \end{aligned}$$

Let σ_n^2 be the leading order $Var(T_n)$ as given in the subsection 6.1.

Under (3.4), as

$$Var(\frac{T_{n2} - ||\mu_1 - \mu_2||^2}{\sigma_{n1}}) = o(1),$$

$$\frac{T_n - ||\mu_1 - \mu_2||^2}{\sqrt{Var(T_n)}} = \frac{T_{n1}}{\sigma_{n1}} + o_p(1). \tag{6.2}$$

Under (3.5),

$$\frac{T_n - ||\mu_1 - \mu_2||^2}{\sqrt{Var(T_n)}} = \frac{T_{n2} - ||\mu_1 - \mu_2||^2}{\sigma_{n2}} + o_p(1). \tag{6.3}$$

As T_{n2} are independent sample averages, its asymptotic normality is readily attainable as showed later. The main task of the proof is for the case under (3.4) when T_{n1} is the contributor of the asymptotic distribution. From (6.1), in the derivation for the asymptotic normality of T_{n1}, we can assume without loss of generality $\mu_1 = \mu_2 = 0$.

Let $Y_i = X_{1i}$ for $i = 1, \cdots, n_1$ and $Y_{j+n_1} = X_{2j}$ for $j = 1, \cdots, n_2$, and for $i \neq j$

$$\phi_{ij} = \begin{cases} n_1^{-1}(n_1 - 1)^{-1} Y_i' Y_j, \text{ if } i, j \in \{1, 2, \cdots, n_1\}, \\ -n_1^{-1} n_2^{-1} Y_i' Y_j, \text{ if } i \in \{1, 2, \cdots, n_1\} \text{ and } j \in \{n_1 + 1, \cdots, n_1 + n_2\}, \\ n_2^{-1}(n_2 - 1)^{-1} Y_i' Y_j, \text{ if } i, j \in \{n_1 + 1, \cdots, n_1 + n_2\}. \end{cases}$$

Define $V_{nj} = \sum_{i=1}^{j-1} \phi_{ij}$ for $j = 2, 3, \cdots, n_1 + n_2$, $S_{nm} = \sum_{j=2}^{m} V_{nj}$ and $\mathcal{F}_{nm} = \sigma\{Y_1, Y_2, \cdots, Y_m\}$ which is the σ algebra generated by $\{Y_1, Y_2, \cdots, Y_m\}$. Now

$$T_n = 2 \sum_{j=2}^{n_1+n_2} V_{nj}.$$

Lemma 1: For each n, $\{S_{nm}, \mathcal{F}_{nm}\}_{m=1}^{n}$ is the sequence of zero mean and a square integrable martingale.

Proof. It's obvious that $\mathcal{F}_{nj-1} \subseteq \mathcal{F}_{nj}$, for any $1 \leq j \leq n$ and S_{nm} is of zero mean and square integrable. We only need to show $E(S_{nq}|\mathcal{F}_{nm}) = S_{nm}$ for any $q \geq m$. We note that
If $j \leq m \leq n$, then $E(V_{nj}|\mathcal{F}_{nm}) = \Sigma_{i=1}^{j-1} E(\phi_{ij}|\mathcal{F}_{nm}) = \Sigma_{i=1}^{j-1} \phi_{ij} = V_{nj}$.
If $j > m$, then $E(\phi_{ij}|\mathcal{F}_{nm}) = E(Y_i' Y_j|\mathcal{F}_{nm})$.
If $i > m$, as Y_i and Y_j are both independent of $\mathcal{F}_{nm}$,

$$E(\phi_{ij}|\mathcal{F}_{nm}) = E(\phi_{ij}) = 0.$$

If $i \le m, E(\phi_{ij}|\mathcal{F}_{n,m}) = E(Y_i'Y_j|\mathcal{F}_{n,m}) = Y_i E(Y_j') = 0$. Hence,

$$E(V_{nj}|\mathcal{F}_{n,m}) = 0.$$

In summary, for $q > m$, $E(S_{nq}|\mathcal{F}_{nm}) = \Sigma_{j=1}^{q} E(V_{nj}|\mathcal{F}_{nm}) = \Sigma_{j=1}^{m} V_{nj} = S_{nm}$. This completes the proof of the lemma. □

Lemma 2: Under Condition (3.4),

$$\Sigma_{j=2}^{n_1+n_2} E[V_{nj}^2|\mathcal{F}_{n,j-1}] \xrightarrow{P} \tfrac{1}{4}\sigma_{T_n}^2.$$

PROOF. Note that

$$\begin{aligned} E(V_{nj}^2|\mathcal{F}_{nj-1}) &= E\{(\sum_{i=1}^{j-1} Y_i'Y_j)^2|\mathcal{F}_{nj-1}\} = E(\sum_{i_1,i_2=1}^{j-1} Y_{i_1}'Y_jY_j'Y_{i_2}|\mathcal{F}_{nj-1}) \\ &= \sum_{i_1,i_2=1}^{j-1} Y_{i_1}' E(Y_jY_j'|\mathcal{F}_{nj-1})Y_{i_2} = \sum_{i_1,i_2=1}^{j-1} Y_{i_1}' E(Y_jY_j')Y_{i_2} \\ &= \sum_{i_1,i_2=1}^{j-1} Y_{i_1}' \frac{\tilde{\Sigma}_j}{\tilde{n}_j(\tilde{n}_j-1)} Y_{i_2} \end{aligned}$$

where $\tilde{\Sigma}_j = \Sigma_1, \tilde{n}_j = n_1$, for $j \in [1, n_1]$ and $\tilde{\Sigma}_j = \Sigma_2$, $\tilde{n}_j = n_2$, if $j \in [n_1+1, n_1+n_2]$.
Define

$$\eta_n = \sum_{j=2}^{n_1+n_2} E(V_{nj}^2|\mathcal{F}_{nj-1})$$

Then,

$$\begin{aligned} E(\eta_n) &= \frac{tr(\Sigma_1^2)}{2n_1(n_1-1)} + \frac{tr(\Sigma_2^2)}{2n_2(n_2-1)} + \frac{tr(\Sigma_1\Sigma_2)}{(n_1-1)(n_2-1)} \\ &= \tfrac{1}{4}\sigma_{T_n}^2. \end{aligned} \tag{6.4}$$

Now consider

$$\begin{aligned} E(\eta_n^2) &= E\{\sum_{j=2}^{n_1+n_2} \sum_{i_1,i_2=1}^{j-1} Y_{i_1}' \frac{\tilde{\Sigma}_j}{\tilde{n}_j(\tilde{n}_j-1)} Y_{i_2}\}^2 \\ &= E\{2 \sum_{2\le j_1<j_2}^{n_1+n_2} \sum_{i_1,i_2=1}^{j_1-1} \sum_{i_3,i_4=1}^{j_2-1} Y_{i_1}' \frac{\tilde{\Sigma}_{j_1}}{\tilde{n}_{j_1}(\tilde{n}_{j_1}-1)} Y_{i_2} Y_{i_3}' \frac{\tilde{\Sigma}_{j_2}}{\tilde{n}_{j_2}(\tilde{n}_{j_2}-1)} Y_{i_4} \\ &\quad + \sum_{j=2}^{n_1+n_2} \sum_{i_1,i_2=1}^{j-1} \sum_{i_3,i_4=1}^{j-1} Y_{i_1}' \frac{\tilde{\Sigma}_j}{\tilde{n}_j(\tilde{n}_j-1)} Y_{i_2} Y_{i_3}' \frac{\tilde{\Sigma}_j}{\tilde{n}_j(\tilde{n}_j-1)} Y_{i_4}\} \\ &= 2E(A) + E(B), \quad \text{say,} \end{aligned} \tag{6.5}$$

where

$$A = \sum_{2\le j_1<j_2}^{n_1+n_2}\sum_{i_1,i_2=1}^{j_1-1}\sum_{i_3,i_4=1}^{j_2-1} Y'_{i_1}\frac{\tilde{\Sigma}_{j_1}}{\tilde{n}_{j_1}(\tilde{n}_{j_1}-1)}Y_{i_2}Y'_{i_3}\frac{\tilde{\Sigma}_{j_2}}{\tilde{n}_{j_2}(\tilde{n}_{j_2}-1)}Y_{i_4},$$

$$B = \sum_{j=2}^{n_1+n_2}\sum_{i_1,i_2=1}^{j-1}\sum_{i_3,i_4=1}^{j-1} Y'_{i_1}\frac{\tilde{\Sigma}_{j}}{\tilde{n}_{j}(\tilde{n}_{j}-1)}Y_{i_2}Y'_{i_3}\frac{\tilde{\Sigma}_{j}}{\tilde{n}_{j}(\tilde{n}_{j}-1)}Y_{i_4}. \tag{6.6}$$

Derivations given in Chen and Qin (2008) show

$$\begin{aligned}2E(A) = \Big\{&\frac{tr^2(\Sigma_1^2)}{4n_1^2(n_1-1)^2}+\frac{tr^2(\Sigma_2^2)}{4n_2^2(n_2-1)^2}+\frac{tr(\Sigma_1^2)tr(\Sigma_1\Sigma_2)}{n_1^2(n_1-1)(n_2-1)}\\ &+\frac{tr(\Sigma_2^2)tr(\Sigma_1\Sigma_2)}{(n_1-1)n_2(n_2-1)}+\frac{tr^2(\Sigma_2\Sigma_1)}{n_1n_2(n_1-1)(n_2-1)}\\ &+\frac{tr(\Sigma_1^2)tr(\Sigma_2^2)}{2n_1(n_1-1)n_2(n_2-1)}\Big\}\{+o(1)\}.\end{aligned}$$

and $E(B)=o(\sigma_{T_n}^2)$. Hence, from (6.5) and (6.6),

$$\begin{aligned}E(\eta_n^2) = &\Big\{\frac{tr^2(\Sigma_1^2)}{4n_1^2(n_1-1)^2}+\frac{tr^2(\Sigma_2^2)}{4n_2^2(n_2-1)^2}+\frac{tr(\Sigma_1^2)tr(\Sigma_1\Sigma_2)}{n_1^2(n_1-1)(n_2-1)}+\\ &+\frac{tr(\Sigma_2^2)tr(\Sigma_1\Sigma_2)}{(n_1-1)n_2(n_2-1)}+\frac{tr^2(\Sigma_2\Sigma_1)}{n_1n_2(n_1-1)(n_2-1)}\\ &+\frac{tr(\Sigma_1^2)tr(\Sigma_2^2)}{2n_1(n_1-1)n_2(n_2-1)}\Big\}+o(\sigma_{T_n}^4).\end{aligned} \tag{6.7}$$

Based on (6.4) and (6.7),

$$Var(\eta_n)=E(\eta_n^2)-E^2(\eta_n)=o(\sigma_{T_n}^4). \tag{6.8}$$

Combine (6.4) and (6.8), we have

$$\sigma_{T_n}^{-2}E\Big\{\Sigma_{j=1}^{n_1+n_2}E(V_{nj}^2|\mathcal{F}_{n,j-1})\Big\}=\sigma_{T_n}^{-2}E(\eta_n)=\tfrac{1}{4},\ \text{and}$$

$$\sigma_{T_n}^{-4}Var\Big\{\Sigma_{j=1}^{n_1+n_2}E(V_{nj}^2|\mathcal{F}_{n,j-1})\Big\}=\sigma_{T_n}^{-4}Var(\eta_n)=o(1).$$

This completes the proof of Lemma 2. □

Lemma 3 Under the condition (3.4),

$$\sum_{j=2}^{n_1+n_2}\sigma_{T_n}^{-2}E\{V_{nj}^2I(|V_{nj}|>\epsilon\sigma_{T_n})|F_{nj-1}\}\xrightarrow{p}0.$$

PROOF. We note that

$$\sum_{j=2}^{n_1+n_2} \sigma_{T_n}^{-2} E\{V_{nj}^2 I(|V_{nj}| > \epsilon\sigma_{T_n})|F_{nj-1}\} \le \sigma_{T_n}^{-q}\epsilon^{2-q} \sum_{j=1}^{n_1+n_2} E(V_{nj}^q|F_{nj-1});$$

for some $q > 2$. By choosing $q = 4$, the conclusion of the lemma is true if we can show

$$E\{\sum_{j=2}^{n_1+n_2} E(V_{nj}^4|F_{nj-1})\} = o(\sigma_{T_n}^4). \tag{6.9}$$

We notice that

$$\begin{aligned} E\{\sum_{j=2}^{n_1+n_2} E(V_{nj}^4|F_{nj-1})\} &= \sum_{j=1}^{n_1+n_2} E(V_{nj}^4) = \sum_{j=1}^{n_1+n_2} E(\sum_{i=1}^{j-1} Y_i'Y_j)^4 \\ &= \sum_{j=2}^{n_1+n_2} \sum_{i_1,i_2,i_3,i_4}^{j-1} E(Y_{i_1}'Y_jY_{i_2}'Y_jY_{i_3}'Y_jY_{i_4}'Y_j). \end{aligned}$$

The last term can be decomposed as $3Q + P$ where

$$Q = \sum_{j=2}^{n_1+n_2} \sum_{s\neq t}^{j-1} E(Y_j'Y_sY_s'Y_jY_j'Y_tY_t'Y_j)$$

and $P = \Sigma_{j=2}^{n_1+n_2}\Sigma_{s=1}^{j-1}E(Y_s'Y_j)^4$. Now (6.9) is true if $3Q + P = o(\sigma_{T_n}^4)$.

Note that

$$\begin{aligned} Q &= \sum_{j=2}^{n_1+n_2} \sum_{s\neq t}^{j-1} E\{tr(Y_jY_j'Y_tY_t'Y_jY_j'Y_sY_s')\} \\ &= O(n^{-4})\{\sum_{j=2}^{n_1} \sum_{s\neq t}^{j-1} E(Y_j'\Sigma_1Y_jY_j'\Sigma_1Y_j) + \sum_{j=n_1+1}^{n_1+n_2} \sum_{s\neq t}^{j-1} E(Y_j'\Sigma_tY_jY_j'\Sigma_sY_j)\} = o(\sigma_{T_n}^4). \end{aligned}$$

The last equation follows the similar procedure in Lemma 2 under (3.4).

It remains to show $P = \Sigma_{j=2}^{n_1+n_2}\Sigma_{s=1}^{j-1}E(Y_s'Y_j)^4 = o(\sigma_{T_n}^4)$. Note that

$$\begin{aligned} P &= \sum_{j=2}^{n_1+n_2} \sum_{s=1}^{j-1} E(Y_s'Y_j)^4 = \sum_{j=2}^{n_1} \sum_{s=1}^{j-1} E(Y_s'Y_j)^4 + \sum_{j=n_1+1}^{n_1+n_2} \sum_{s=1}^{j-1} E(Y_s'Y_j)^4 \\ &= O(n^{-8})\{\sum_{j=2}^{n_1} \sum_{s=1}^{j-1} E(X_{1s}'X_{1j})^4 + \sum_{j=n_1+1}^{n_1+n_2} \sum_{s=1}^{n_1} E(X_{1s}'X_{2j-n_1})^4 \\ &\quad + \sum_{j=n_1+1}^{n_1+n_2} \sum_{s=n_1+1}^{j-1} E(X_{2s-n_1}'X_{2j-n_1})^4\} \\ &= O(n^{-8})(P_1 + P_2 + P_3), \end{aligned}$$

where $P_1 = \sum_{j=2}^{n_1}\sum_{s=1}^{j-1} E(X'_{1s}X_{1j})^4$, $P_2 = \sum_{j=n_1+1}^{n_1+n_2}\sum_{s=1}^{n_1} E(X'_{1s}X_{2j-n_1})^4$ and

$$P_3 = \sum_{j=n_1+1}^{n_1+n_2}\sum_{s=n_1+1}^{j-1} E(X'_{2s-n_1}X_{2j-n_1})^4.$$

Let us consider $E(X'_{1s}X_{2j-n_1})^4$. Define $\Gamma_1'\Gamma_2 =: (v_{ij})_{m\times m}$ and note the following facts which will be used repeatedly in the rest of the appendix,

$$\begin{aligned}
&\sum_{i,j=1}^{m} v_{ij}^4 \le (\sum_{i,j=1}^{m} v_{ij}^2)^2 = tr^2(\Gamma_1'\Gamma_2\Gamma_2'\Gamma_1) = tr^2(\Sigma_2\Sigma_1),\\
&\sum_{i=1}^{m}\sum_{j_1\neq j_2}^{m} (v_{ij_1}^2 v_{ij_2}^2) \le (\sum_{i,j=1}^{m} v_{ij}^2)^2 = tr^2(\Sigma_2\Sigma_1),\\
&\sum_{i_1\neq i_2}^{m}\sum_{j_1\neq j_2}^{m} v_{i_1j_1}v_{i_1j_2}v_{i_2j_1}v_{i_2j_2} \le \sum_{i_1\neq i_2}^{m} v_{i_1i_2}^{(2)}v_{i_1i_2}^{(2)} \le \sum_{i_1i_2=1}^{m} v_{i_1i_2}^{(2)}v_{i_1i_2}^{(2)},\\
&\sum_{i_1i_2=1}^{m} v_{i_1i_2}^{(2)}v_{i_1i_2}^{(2)} = \sum_{i_1=1}^{m} v_{i_1i_1}^{(4)} = tr(\Gamma_1'\Sigma_2\Gamma_1\Gamma_1'\Sigma_2\Gamma_1) = tr(\Sigma_2\Sigma_1)^2,
\end{aligned}$$

where $\Gamma_1'\Sigma_2'\Gamma_1 = (v_{ij}^{(2)})$ and $(\Gamma_1'\Sigma_2\Gamma_1)^2 = (v_{ij}^{(4)})_{m\times m}$.

From (3.1),

$$\begin{aligned}
E(X'_{1s}X_{2j-n_1})^4 &= \sum_{i=1}^{m}\sum_{j'=1}^{m}(3+\Delta)^2 v_{ij'}^4 + \sum_{i=1}^{m}(3+\Delta)\sum_{j_1\neq j_2}^{m} v_{ij_1}^2 v_{ij_2}^2\\
&\quad + \sum_{j'=1}^{m}(3+\Delta)\sum_{i_1\neq i_2}^{m} v_{i_1j}^2 v_{i_2j}^2 + 9\sum_{i_1\neq i_2}^{m}\sum_{j_1\neq j_2}^{m} v_{i_1j_1}v_{i_1j_2}v_{i_2j_1}v_{i_2j_2}\\
&= O\{tr^2(\Sigma_2\Sigma_1)\} + O\{tr(\Sigma_2\Sigma_1)^2\}.
\end{aligned}$$

Then we conclude

$$\begin{aligned}
O(n^{-8})P_2 &= \sum_{j=n_1+1}^{n_1+n_2}\sum_{s=1}^{n_1}\Big[O\{tr^2(\Sigma_2\Sigma_1)\} + O\{tr(\Sigma_2\Sigma_1)^2\}\Big]\\
&= O(n^{-5})\Big[O\{tr^2(\Sigma_2\Sigma_1)\} + O\{tr(\Sigma_2\Sigma_1)^2\}\Big] = o(\sigma_{T_n}^4).
\end{aligned}$$

We can also prove that $O(n^{-8})P_1 = o(\sigma_{T_n}^4)$ and $O(n^{-8})P_3 = o(\sigma_{T_n}^4)$ by going through the similar procedure. This completes the proof of the lemma. □

6.3. **Proof of Theorem 1.**

PROOF. We note equations (6.2) and (6.3) under conditions (3.4) and (3.5) respectively. Based on Corollary 3.1 of Hall and Heyde (1980), Lemma 1, Lemma 2 and Lemma 3, it can be concluded that $T_n 1/\sigma_{n1} \xrightarrow{d} N(0,1)$. This implies the desired asymptotic normality of T_n under (3.4). Under (3.5), as T_{n2} is the sum of two independent averages, its asymptotic normality can be attained by following the standard means. Hence the theorem is proved. □

6.4. **Proof of Theorem 2.**

PROOF. We only present the proof for the ratio consistency of $tr(\widehat{\Sigma_1^2})$ as the proofs of the other two follow the same route. We want to show

$$E\{tr(\widehat{\Sigma_1^2})\} = tr(\Sigma_1^2)\{1+o(1)\} \text{ and } Var\{tr(\widehat{\Sigma_1^2})\} = o\{tr^2(\Sigma_1^2)\}. \tag{6.10}$$

For notation simplicity, we denote X_{1j} as X_j and Σ_1 as Σ, since we are effectively in a one sample situation.

Note that

$$\begin{aligned}
& tr(\widehat{\Sigma^2}) \\
= \; & \{n(n-1)\}^{-1} tr[\sum_{j\neq k}^{n} \{(X_j-\mu)(X_j-\mu)'(X_k-\mu)(X_k-\mu)' \\
& -2(\bar{X}_{(j,k)}-\mu)(X_j-\mu)'(X_k-\mu)(X_k-\mu)'\} \\
& + \sum_{j\neq k}^{n} \{2(X_j-\mu)\mu'(X_k-\mu)(X_k-\mu)' - 2(\bar{X}_{(j,k)}-\mu)\mu'(X_k-\mu)(X_k-\mu)'\} \\
& + \sum_{j\neq k}^{n} \{(\bar{X}_{(j,k)}-\mu)(X_j-\mu)'(\bar{X}_{(j,k)}-\mu)(X_k-\mu)'\} \\
& - \sum_{j\neq k}^{n} \{2(X_j-\mu)\mu'(\bar{X}_{(j,k)}-\mu)(X_k-\mu)' \\
& -2(\bar{X}_{(j,k)}-\mu)\mu'(\bar{X}_{(j,k)}-\mu)(X_k-\mu)'\} \\
& + \sum_{j\neq k}^{n} \{(X_j-\mu)\mu'(X_k-\mu)\mu' - 2(\bar{X}_{(j,k)}-\mu)\mu'(X_k-\mu)\mu'\} \\
& + \sum_{j\neq k}^{n} \{(\bar{X}_{(j,k)}-\mu)\mu'(\bar{X}_{(j,k)}-\mu)\mu'\}] \\
& =: \sum_{l=1}^{10} tr(A_l), \text{say}.
\end{aligned}$$

It is easy to show that $E\{tr(A_1)\} = tr(\Sigma^2)$, $E\{tr(A_i)\} = 0$ for $i = 2, \cdots, 9$ and $E\{tr(A_{10})\} = \mu'\Sigma\mu/(n-2) = o\{tr(\Sigma^2)\}$. The last equation is based on (3.4). This leads to the first part of (6.10). Since $tr(A_{10})$ is non-negative and $E\{tr(A_{10})\} = o\{tr(\Sigma^2)\}$, we have $tr(A_{10}) = o_p\{tr(\Sigma^2)\}$. However, to establish the orders of other terms, we need to derive $Var\{tr(A_i)\}$. We shall only show $Var\{tr(A_1)\}$ here. Derivations for other $Var\{tr(A_i)\}$ are given in Chen and Qin (2008).

Note that

$$\begin{aligned}
&Var\{tr(A_1)\} + tr^2(\Sigma^2) \\
=&E[\frac{1}{n(n-1)}tr\{\sum_{j\neq k}^{n}(X_j-\mu)(X_j-\mu)'(X_k-\mu)(X_k-\mu)'\}]^2 \\
=&\frac{1}{n^2(n-1)^2}E[tr\{\sum_{j_1\neq k_1}^{n}(X_{j_1}-\mu)(X_{j_1}-\mu)'(X_{k_1}-\mu)(X_{k_1}-\mu)'\} \\
&\times tr\{\sum_{j_2\neq k_2}^{n}(X_{j_2}-\mu)(X_{j_2}-\mu)'(X_{k_2}-\mu)(X_{k_2}-\mu)'\}].
\end{aligned}$$

It can be shown, by considering the possible combinations of the subscripts j_1, k_1, j_2 and k_2, that

$$\begin{aligned}
Var\{tr(A_1)\} &= \{n(n-1)\}^{-1}E\{(X_1-\mu)'(X_2-\mu)\}^4 + \\
&\quad \frac{4(n-2)}{n(n-1)}E\Big\{(X_1-\mu)'\Sigma(X_1-\mu)\Big\}^2 + o\{tr^2(\Sigma^2)\} \\
&=: \frac{2}{n(n-1)}B_{11} + \frac{4(n-2)}{n(n-1)}B_{12} + o\{tr^2(\Sigma^2)\}, \qquad (6.11)
\end{aligned}$$

where

$$\begin{aligned}
B_{11} &= E(Z_1'\Gamma'\Gamma Z_2)^4 = E(\sum_{s,t=1}^{m} z_{1s}\nu_{st}z_{2t})^4 \\
&= E(\sum_{s_1,s_2,s_3,s_4,t_1,t_2,t_3,t_4=1}^{m} \nu_{s_1t_1}\nu_{s_2t_2}\nu_{s_3t_3}\nu_{s_4t_4}z_{1s_1}z_{1s_2}z_{1s_3}z_{1s_4}z_{2t_1}z_{2t_2}z_{2t_3}z_{2t_4})
\end{aligned}$$

and

$$\begin{aligned}
B_{12} &= E(Z_1'\Gamma'\Gamma\Gamma'\Gamma Z_1)^2 = E(\sum_{s,t=1}^{m} z_{1s}u_{st}z_{1t})^2 \\
&= E(\sum_{s_1,s_2,t_1,t_2=1}^{m} u_{s_1t_1}u_{s_2t_2}z_{1s_1}z_{1s_2}z_{1t_1}z_{1t_2}).
\end{aligned}$$

Here ν_{st} and u_{st} are respectively the (s,t) element of $\Gamma'\Gamma$ and $\Gamma'\Sigma\Gamma$.

Since $tr^2(\Sigma^2) = (\sum_{s,t=1}^m \nu_{st}^2)^2 = \sum_{s_1,s_2,t_1,t_2=1}^m \nu_{s_1t_1}^2\nu_{s_2t_2}^2$ and $tr(\Sigma^4) = \sum_{t_1,t_2=1}^m u_{t_1t_2}^2$. It can be shown that $A_{11} \le c\,tr^2(\Sigma^2)$ for a finite positive number c and hence $\{n(n-1)\}^{-1}B_{11} = o\{tr^2(\Sigma^2)\}$. It may also be shown that

$$\begin{aligned} B_{12} &= 2\sum_{s,t=1}^m u_{st}^2 + \sum_{s,t=1}^m u_{ss}u_{tt} + \Delta\sum_{s=1}^m u_{ss}^2 \\ &= 2tr(\Sigma^4) + tr^2(\Sigma^2) + \Delta\sum_{s=1}^m u_{ss}^2 \\ &\le (2+\Delta)tr(\Sigma^4) + tr^2(\Sigma^2). \end{aligned}$$

Therefore, from (6.11)

$$\begin{aligned} Var\{tr(A_1)\} &\le \frac{2}{n(n-1)}ctr^2(\Sigma^2) + \frac{4(n-2)}{n(n-1)}\Big\{(2+\Delta)tr(\Sigma^4) + tr^2(\Sigma^2)\Big\} \\ &\quad + \frac{(n-2)(n-3)}{n(n-1)}tr^2(\Sigma^2) - tr^2(\Sigma^2) \\ &= o\{tr^2(\Sigma^2)\}. \end{aligned}$$

This completes the proof. □

Acknowledgements. We are grateful to two reviewers for valuable comments and suggestions which have improved the presentation of the paper. We also thank Dan Nettleton and Peng Liu for useful discussions.

References.

[1] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley.

[2] Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006). Adaptive to unknown sparsity in controlling the false discovery rate. *The Annals of Statistics* **34**, 584-653.

[3] Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* **6** 311–329.

[4] Barry, W., Nobel, A. and Wright, F. (2005). Significance analysis of functional categories in gene expression studies: A structured permutation approach. *Bioinformatics* **21** 1943-1949.

[5] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society series B* **57**, 289-300.

[6] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**, 1165–1188.

[7] Chen, S. X. and Qin, Y.-L. (2008). *A Two Sample Test For High Dimensional Data With Applications To Gene-set Testing*. Research Report, Department of Statistics, Iowa State University.

[8] Chiaretti, S., Li, X.C., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz,J. and Foa, R. (2004) Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood* **103**, No. 7, 2771–2778.

[9] Dudoit, S., Keles, S. and van der Laan, M. (2006). Multiple tests of association with biological annotation metadata. Manuscript.

[10] Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, **1**, 107-129.

[11] Fan, J., Hall, P. and Yao, Q. (2007). To how many simultaneous hypothesis tests can normal, student's t or bootstrap calibration be applied. *Journal of the American Statistical Association*, **102**, 1282-1288.

[12] Fan, J., Peng, H. and Huang, T. (2005). Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency. *Journal of the American Statistical Association*, **100**, 781-796.

[13] Gentleman, R., Irizarry, R.A., Carey, V.J., Dudoit, S. and Huber, W. c(2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer.

[14] Hall, P. and Heyde, C.(1980). *Martingale Limit Theory and Applications*, Academic Press, New York.

[15] Huang, J., Wang, D. and Zhang, C. (2005). A two-way semilinear model for normalization and analysis of cDNA microarray data. *Journal of the American Statistical Association* **100** 814–829.

[16] Kosorok, M. and Ma, S. (2007). Marginal asymptotics for the "large p, small n" paradigm: with applications to microarray data. *The Annals of Statistics*, **35**, 1456-1486.

[17] Ledoit, O. and Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compare to the sample size. *The Annals of Statistics*, **30**, 1081-1102.

[18] Newton, M., Quintana, F., Den Boon, J., Sengupta, S. and Ahlquist, P. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *The Annals of Applied Statistics*, **1**, 85-106.

[19] Portnoy, S. (1986). On the central limit theorem in R^p when $p \to \infty$. *Probability*

Theory and Related Fields, **73**, 571-583.

[20] Recknor, J., Nettleton, D. and Reecy, J. (2007). Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics*, to appear.

[21] Schott, J. R. (2005). Testing for complete independence in high dimensions. *Biometrika*, **92**, 951-956.

[22] Storey, J., Taylor, J. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society series B*, **66**, 187–205.

[23] Tracy, C. and Widom, H. (1996). On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics* **177**, 727–754.

[24] van der Laan, M. and Bryan, J. (2001). Gene expression analysis with the parametric bootstrap. *Biostatistics* **2**, 445–461.

[25] Yin, Y., Bai, Z. and Krishnaiah, P. R. (1988) On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. *Probability Theory and Related Fields* **78**, 509–521.

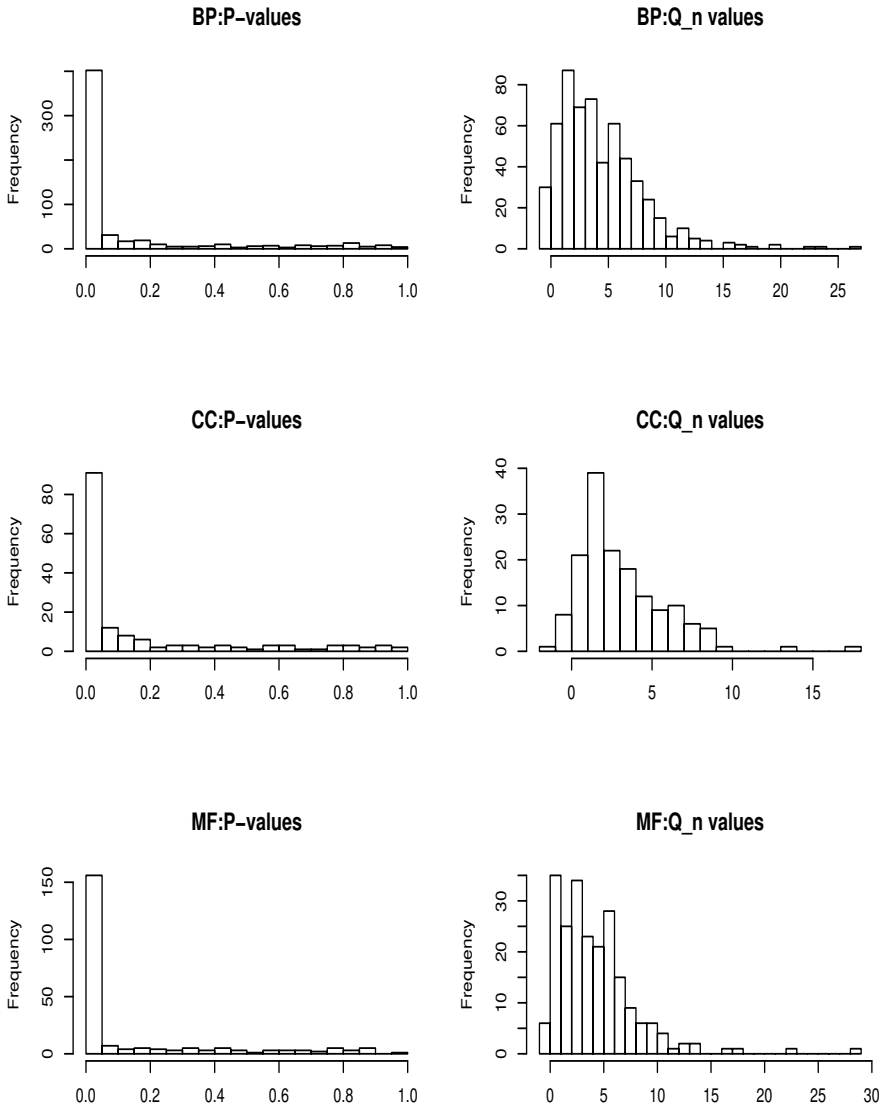

FIG 1. *Two sample tests for differentially expressed gene-sets between BCR/ABL and NEG class ALL: Histograms of P-values (left panels) and Q_n values (right panels) for BP, CC and MF gene categories.*

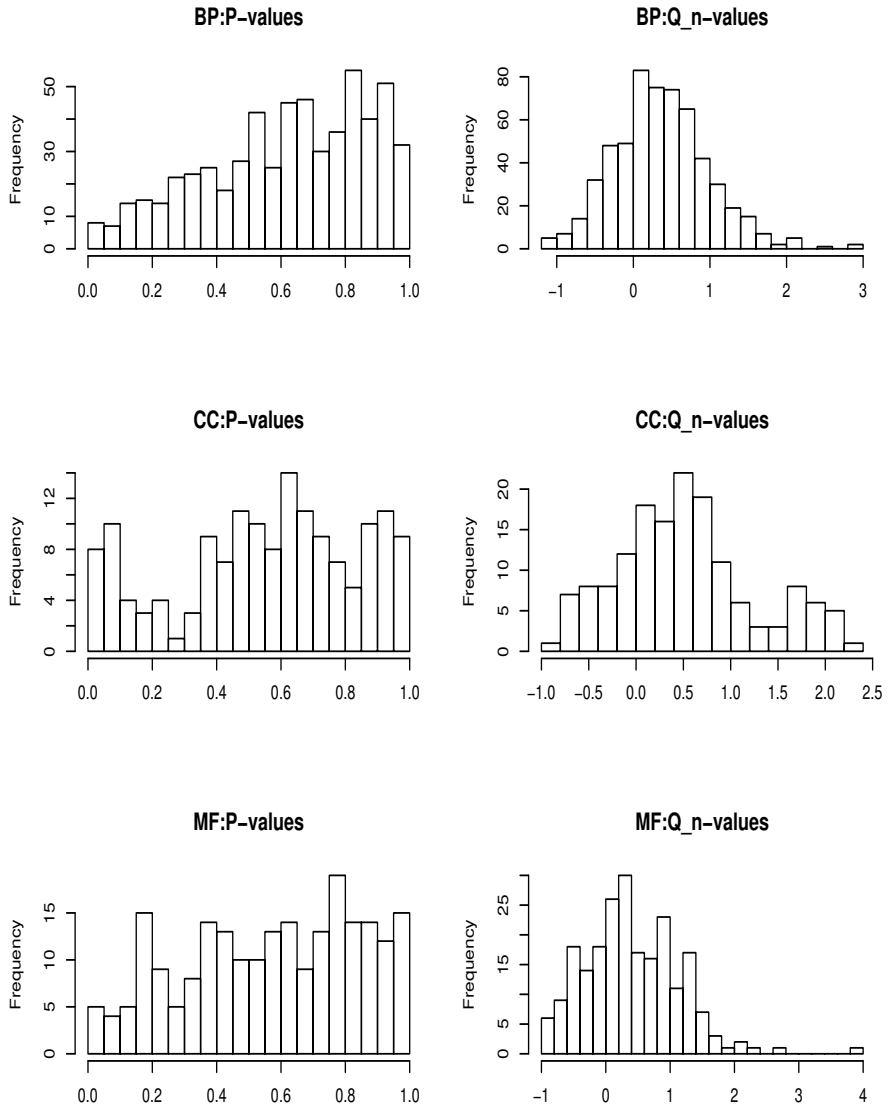

FIG 2. *Back-testing round 2 for differentially expressed gene-sets between two randomly assigned BCR/ABL groups: Histograms of P-values (left panels) and Q_n values (right panels) for BP, CC and MF gene categories.*

Table 1. Empirical Power and Size for the 2-Dependence Model with Gamma Innovation

Type of	% of	$p = 500, n = 124$				$p = 1000, n = 138$			
Allocation	True Null	NEW	BS	Bonf	FDR	NEW	BS	Bonf	FDR
Equal	0%	.511	.399	.13	.16	.521	.413	.11	.16
	25%	.521	.387	.14	.16	.518	.410	.12	.16
	50%	.513	.401	.13	.17	.531	.422	.12	.17
	75%	.522	.389	.13	.18	.530	.416	.11	.17
	95%	.501	.399	.14	.16	.500	.398	.13	.17
	99%	.499	.388	.13	.15	.507	.408	.15	.18
	100%(size)	.043	.043	.040	.041	.043	.042	.042	.042
Increasing	0%	.520	.425	.11	.13	.522	.409	.12	.15
	25%	.515	.431	.12	.15	.523	.412	.14	.16
	50%	.512	.412	.13	.15	.528	.421	.15	.17
	75%	.522	.409	.15	.17	.531	.431	.16	.19
	95%	.488	.401	.14	.15	.500	.410	.15	.17
	99%	.501	.409	.15	.17	.511	.412	.15	.16
	100%(size)	.042	.041	.040	.041	.042	.040	.039	.041
Decreasing	0%	.522	.395	.11	.15	.533	.406	.09	.15
	25%	.530	.389	.11	.15	.530	.422	.11	.17
	50%	.528	.401	.12	.17	.522	.432	.12	.17
	75%	.533	.399	.13	.18	.519	.421	.12	.17
	95%	.511	.410	.12	.15	.508	.411	.15	.18
	99%	.508	.407	.14	.15	.507	.418	.16	.17
	100%(size)	.041	.042	.041	.042	.042	.040	.040	.042

Table 2. Empirical Power and Size for the Full-Dependence Model with Gamma Innovation

Type of	% of	$p = 500, n = 124$				$p = 1000, n = 138$			
Allocation	True Null	NEW	BS	Bonf	FDR	NEW	BS	Bonf	FDR
Equal	0%	.322	.120	.08	.10	.402	.216	.09	.11
	25%	.318	.117	.08	.10	.400	.218	.08	.11
	50%	.316	.115	.09	.11	.409	.221	.09	.10
	75%	.307	.113	.10	.12	.410	.213	.09	.13
	95%	.233	.128	.11	.14	.308	.215	.10	.13
	99%	.225	.138	.12	.15	.316	.207	.11	.12
	100%(size)	.041	.041	.043	.043	.042	.042	.040	.041
Increasing	0%	.331	.121	.09	.12	.430	.225	.10	.11
	25%	.336	.119	.10	.12	.423	.231	.12	.12
	50%	.329	.123	.12	.14	.422	.226	.13	.14
	75%	.330	.115	.12	.15	.431	.222	.14	.15
	95%	.219	.120	.12	.13	.311	.218	.14	.15
	99%	.228	.117	.13	.15	.315	.217	.15	.17
	100%(size)	.041	.040	.042	.043	.042	.042	.040	.042
Decreasing	0%	.320	.117	.08	.11	.411	.213	.08	.10
	25%	.323	.119	.09	.11	.408	.210	.08	.11
	50%	.327	.120	.11	.12	.403	.208	.09	.10
	75%	.322	.122	.12	.12	.400	.211	.12	.13
	95%	.217	.109	.12	.15	.319	.207	.12	.15
	99%	.224	.111	.13	.16	.327	.205	.11	.13
	100%(size)	.042	.043	.039	.041	.042	.211	.040	.041

Table 3. Empirical averages of $tr(\widehat{\Sigma^2})/tr(\Sigma^2)$ with standard deviations in the parentheses.

Type of Innovation	Type of Dependence	$p = 500, n = 124$ NEW	BS	$tr(\Sigma^2)$
Normal	2-Dependence	1.03 (0.015)	1.39 (0.016)	3102
	Full-Dependence	1.008 (0.00279)	.1.17 (0.0032)	35911
Gamma	2-Dependence	1.03 (0.006)	1.10 (0.007)	14227
	Full-Dependence	1.108 (0.0019)	1.248 (0.0017)	152248
		$p = 1000, n = 138$		
		NEW	BS	$tr(\Sigma^2)$
Normal	2-Dependence	0.986 (0.0138)	1.253 (0.0136)	6563
	Full-Dependence	0.995 (0.0026)	1.072 (0.0033)	76563
Gamma	2-Dependence	1.048 (0.005)	1.138 (0.006)	32104
	Full-Dependence	1.088 (0.00097)	1.231 (0.0013)	325879

Table 4. Empirical Power and Size for the Sparse Model.

Sample		ε=.25				ε=.15			
Size		c=.25		c=.45		c=.35		c=.55	
$(n_1 = n_2)$	Methods	Power	Size	Power	Size	Power	Size	Power	Size
10	FDR	.084	.056	.180	.040	.044	.034	.066	.034
	Bonf	.084	.056	.170	.040	.044	.034	.062	.032
	New	.100	.046	.546	.056	.072	.064	.344	.064
20	FDR	.380	.042	.855	.044	.096	.036	.326	.058
	Bonf	.368	.038	.806	.044	.092	.034	.308	.056
	New	.238	.052	.976	.042	.106	.052	.852	.046
30	FDR	.864	.042	1	.060	.236	.048	.710	.038
	Bonfe	.842	.038	.996	.060	.232	.048	.660	.038
	New	.408	.050	.998	.058	.220	.054	.988	.042

Table 5. Average ratios of $\widehat{\sigma_M^2}/\sigma_M^2$ and their Standard Deviation (in parenthesis) for the Sparse Model

Sample	True	ε=.25		ε=.15	
Size	σ_M^2	c=0.25	c=0.45	c=0.35	c=0.55
$n_1 = n_2$=10	84.4	1.003(.0123)	1.005 (.0116)	.998 (.0120)	.999(.0110)
$n_1 = n_2$=20	20.5	1.003(.0033)	1.000 (.0028)	1.003(.0028)	1.002(.0029)
$n_1 = n_2$=30	9.0	.996(.0013)	.998(.0013)	1.004(.0014)	.999(.0013)

Department of Statistics
Iowa State University
Ames, Iowa 50011-1210;
and Guanghua School of Management,
Peking University, Beijing 100871, China
E-mail: songchen@iastate.edu

Department of Statistics
Iowa State University
Ames, Iowa 50011-1210;
qinyl@iastate.edu