Tests for High-Dimensional Covariance Matrices

Song Xi CHEN, Li-Xin ZHANG, and Ping-Shou ZHONG

We propose tests for sphericity and identity of high-dimensional covariance matrices. The tests are nonparametric without assuming a specific parametric distribution for the data. They can accommodate situations where the data dimension is much larger than the sample size, namely the "large p, small n" situations. We demonstrate by both theoretical and empirical studies that the tests have good properties for a wide range of dimensions and sample sizes. We applied the proposed test on a microarray dataset on Yorkshire Gilts and tested for the covariance structure for the expression levels for sets of genes.

KEY WORDS: Gene-set testing; High data dimension; Identity test; Large p, small n; Sphericity test.

1. INTRODUCTION

High-dimensional data are increasingly encountered in statistical applications with the most prominent ones coming from biology and finance. In genomic studies the data dimension can be a lot larger than the sample size. When this happens, some of the conventional multivariate procedures may not necessarily work since these procedures are justified under a framework where the sample size n tends to infinity while the dimension p remains fixed.

In this paper, we consider tests for covariance matrices of multivariate distributions when p can be much larger than the sample size n. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed p-dimensional random vectors with covariance $\Sigma = \mathrm{var}(X_i)$. Our interest is to test two structures for the covariance:

$$H_0 : \Sigma = \sigma^2 I_p \quad \text{vs.} \quad H_1 : \Sigma \neq \sigma^2 I_p \tag{1.1}$$

and

$$H_0 : \Sigma = I_p \quad \text{vs.} \quad H_1 : \Sigma \neq I_p, \tag{1.2}$$

where I_p is the p-dimensional identity matrix and σ^2 is a unknown but finite positive constant. The identity hypothesis in (1.2) covers the hypothesis $H_0 : \Sigma = \Sigma_0$ for a specific known invertible covariance matrix Σ_0. Conventional tests for covariance based on the likelihood ratio (Anderson 2003) cannot be used without modification when the data dimension is larger than the sample size, since the sample covariance is no longer invertible with probability one.

The practical needs for testing the above hypotheses come from several areas of statistical applications, in particular from microarray analysis and the associated large-scale multiple testing. A common assumption made when analyzing the microarray data is the so-called column-wise or gene-wise independence, namely independence in the expression levels among different genes. Hence, it is important to carry out tests on the covariance structure of the data columns before high-dimensional statistical procedures are employed. Having said these, we note that the issue of data dependence has drawn increasing attention in the context of multiple testing, which include Benjamini and Yekutieli (2001) who showed that the false discovery rate (FDR) procedure proposed in Benjamini and Hochberg (1995) under independence is applicable under the positive regression dependence. There have been also a substantial set of research works on inference for means of high-dimensional distributions either in the context of multiple tests as in van der Laan and Bryan (2001), Donoho and Jin (2004), Fan, Hall, and Yao (2007), Kosorok and Ma (2007), and Hall and Jin (2008), or in the context of simultaneous multivariate testing as in Bai and Saranadasa (1996), Schott (2007), and Chen and Qin (2010). See also Huang, Wang, and Zhang (2005), Fan, Peng, and Huang (2005), and Zhang and Huang (2008) for inference on high-dimensional conditional means. Some of these works accommodate column-wise dependence.

Naturally, tests for the hypotheses (1.1) and (1.2) would be based on the sample covariance matrix as in the conventional formulations of John (1971) and Nagao (1973) and the likelihood ratio statistic (Anderson 2003). For inference on high-dimensional covariance matrices, there has been an array of works on the convergence of the sample covariance matrices based on the spectral analysis of large-dimensional random matrices (Bai and Yin 1993; Bai, Silverstein, and Yin 1998); see Bai and Silverstein (2005) for a comprehensive summary. These studies show that even under the modest $p/n \to c$ for a positive constant c, the smallest and the largest eigenvalues of the sample covariance matrix do not converge to their population counterparts. Hence the sample covariance fails to converge to the population covariance Σ. While this is discouraging, we will show in this paper that consistent test procedures can still be constructed for high-dimensional covariance.

The research in this paper is motivated by the important work of Ledoit and Wolf (2002) who examined two conventional tests (John 1971, 1972 and Nagao 1973) for the above hypotheses for high-dimensional normally distributed random vectors when $p/n \to c$ for a finite constant c. They established the asymptotic normality of the test statistics and found the sphericity test is robust under $p/n \to c$, whereas the identity test needs to be modified under high dimensionality. In this paper, new

Song Xi Chen is Professor, Department of Statistics, Iowa State University, Ames, IA 50011-1210 (E-mail: *songchen@iastate.edu*) and Chair Professor, Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University, Beijing 100651, China (E-mail: *csx@gsm.pku.edu.cn*). Li-Xin Zhang is Assistant Professor, School of Mathematics, Jilin University, Jilin 130012, China (E-mail: *jinwenlixin@163.com*). Ping-Shou Zhong is a Graduate Student, Department of Statistics, Iowa State University, Ames, IA 50011-1210 (E-mail: *pszhong@iastate.edu*). This research was supported in part by grants from NSF-DMS 0604563 and 0714978. We thank the Editor, the Associate Editor, and two referees for insightful comments and suggestions which have improved the presentation of the paper. We would also like to thank Mr. Long Qu from Animal Science Department at Iowa State University for making the Yorkshire Gilts data available.

© 2010 American Statistical Association
Journal of the American Statistical Association
June 2010, Vol. 105, No. 490, Theory and Methods
DOI: 10.1198/jasa.2010.tm09560

tests for hypotheses (1.1) and (1.2) are proposed without the normality assumption and without specifying an explicit relationship between p and n as long as $p \to \infty$ as $n \to \infty$ and $\mathrm{tr}(\Sigma^4) = o\{\mathrm{tr}^2(\Sigma^2)\}$. Hence, the tests accommodate the "large p, small n" situations. The underlying reason for such accommodation is due to more accurate and reliable estimators of $\mathrm{tr}(\Sigma)$ and $\mathrm{tr}(\Sigma^2)$ under both nonnormality and high data dimensionality.

The paper is structured as follows. Section 2 introduces the basic data structure, hypotheses, the estimators, and the assumptions. The test procedures are proposed in Section 3 after establishing the asymptotic normality of the test statistics, followed by results on the consistency of the tests. Section 4 reports simulation studies. An empirical study on a biological data employing the proposed test procedures is given in Section 5. All the technical details are deferred to the Appendix.

2. PRELIMINARIES

Let us now introduce the test statistics for testing the two hypotheses (1.1) and (1.2). For (1.1), a scaled distance measure between $\sigma^{-2}\Sigma$ and I_p is

$$\frac{1}{p}\mathrm{tr}\left[\left(\frac{\Sigma}{(1/p)\mathrm{tr}(\Sigma)} - I_p\right)^2\right] = p\frac{\mathrm{tr}(\Sigma^2)}{\mathrm{tr}^2(\Sigma)} - 1. \tag{2.1}$$

The case for (1.2) is similar by considering

$$\frac{1}{p}\mathrm{tr}(\Sigma - I_p)^2 = \frac{1}{p}\mathrm{tr}(\Sigma^2) - \frac{2}{p}\mathrm{tr}(\Sigma) + 1 \tag{2.2}$$

as a distance measure between Σ and I_p.

For classical fixed dimension normally distributed data, John (1971, 1972) considered a test for sphericity based on

$$U_n' = \frac{1}{p}\mathrm{tr}\left[\left(\frac{S_n}{(1/p)\,\mathrm{tr}(S_n)} - I_p\right)^2\right],$$

where S_n is the sample covariance matrix. And the test for the identity hypothesis (John 1972 and Nagao 1973) is based on

$$V_n' = \frac{1}{p}\mathrm{tr}(S_n - I_p)^2.$$

Basically, these two statistics can be understood by replacing $\mathrm{tr}(\Sigma)$ and $\mathrm{tr}(\Sigma^2)$ by their estimators $\mathrm{tr}(S_n)$ and $\mathrm{tr}(S_n^2)$ respectively in (1.1) and (1.2). While direct substitution of Σ by S_n brings invariance and optimal testing properties as shown in John (1971, 1972) and Nagao (1973) for normally distributed data when p is fixed, tests based on these statistics may not work for high-dimensional data as demonstrated in Ledoit and Wolf (2002).

For normally distributed random vectors, Ledoit and Wolf (2002) evaluate the above two test statistics when the dimension p is increased at the same rate as n so that $p/n \to c$ for a finite c (p can be larger than n). They showed that while the test for sphericity based on U_n' is not affected by a diverging p, the identity test based on V_n' is affected by the increasing dimensionality. They proposed a modification to V_n' that works when $p/n \to c$.

In this paper we propose new tests for the sphericity and identity hypotheses without the normal distribution assumption and under much relaxed conditions for the growth rate of p. In particular, p can be a larger order of n, hence accommodating the so-called "large p, small n" situations. The key for our proposal is to have more durable and accurate estimators for $\mathrm{tr}(\Sigma)$ and $\mathrm{tr}(\Sigma^2)$. Behind our proposal is the observation that while using $\mathrm{tr}(S_n)$ and $\mathrm{tr}(S_n^2)$ to estimate $\mathrm{tr}(\Sigma)$ and $\mathrm{tr}(\Sigma^2)$ is intuitive, these estimators have unnecessary terms which slow down the convergence considerably when the dimension is high.

Let

$$Y_{1,n} = \frac{1}{n}\sum_{i=1}^{n} X_i'X_i \qquad \text{and} \qquad Y_{3,n} = \frac{1}{P_n^2}\sum_{i\neq j} X_i'X_j.$$

Here and below, we denote $P_n^r = n!/(n-r)!$. Then $\mathrm{E}(Y_{1,n}) = \mathrm{tr}(\Sigma) + \mu'\mu$ and $\mathrm{E}(Y_{1,n} - Y_{3,n}) = \mathrm{tr}(\Sigma)$, and thus

$$T_{1,n} = Y_{1,n} - Y_{3,n} \tag{2.3}$$

is an unbiased estimator of $\mathrm{tr}(\Sigma)$. Similarly, define

$$Y_{2,n} = \frac{1}{P_n^2}\sum_{i\neq j}(X_i'X_j)^2, \qquad Y_{4,n} = \frac{1}{P_n^3}\sum_{i,j,k}^{*} X_i'X_jX_j'X_k,$$

and

$$Y_{5,n} = \frac{1}{P_n^4}\sum_{i,j,k,l}^{*} X_i'X_jX_k'X_l.$$

Here and elsewhere in this paper $\sum^*$ denotes summation over mutually different indices. For example, $\sum_{i,j,k}^*$ means summation over $\{(i,j,k): i\neq j, j\neq k, k\neq i\}$. An unbiased estimator of $\mathrm{tr}(\Sigma^2)$ is

$$T_{2,n} = Y_{2,n} - 2Y_{4,n} + Y_{5,n}.$$

Besides unbiasedness, both $T_{1,n}$ and $T_{2,n}$ are invariant with respect to the location transformation that transforms X_i to $X_i + c$ for an arbitrary constant vector c. This allows us to assume in the rest of the paper that $\mu = 0$.

The test statistics we use for testing the sphericity hypothesis (1.1) and the identity hypothesis (1.2) are, respectively,

$$U_n = p\left(\frac{T_{2,n}}{T_{1,n}^2}\right) - 1 \qquad \text{and} \qquad V_n = \frac{1}{p}T_{2,n} - \frac{2}{p}T_{1,n} + 1.$$

To facilitate our analysis, like Bai and Saranadasa (1996), we assume that the observations X_i follow a multivariate model.

Assumption 1. Suppose $X_1, X_2, \ldots, X_n$ are independent and identically (IID) distributed p-dimensional random vectors such that

$$X_i = \Gamma Z_i + \mu \quad \text{for } i = 1, 2, \ldots, n, \tag{2.4}$$

where μ is a p-dimensional constant vector, Γ is a $p\times m$ constant matrix with $m \geq p$ so that $\Gamma\Gamma^T = \Sigma$, and $Z_1, Z_2, \ldots, Z_n$ are IID m-dimensional random vectors such that $\mathrm{E}(Z_1) = 0$ and $\mathrm{var}(Z_1) = I_m$. Write $Z_1 = (z_{11}, \ldots, z_{1m})^T$. We assume that each z_{1l} has uniformly bounded 8th moment, and there exists a finite constant Δ such that for $l = 1, \ldots, m$, $\mathrm{E}(z_{1l}^4) = 3 + \Delta$ and for any integers $\ell_\nu \geq 0$ with $\sum_{\nu=1}^q \ell_\nu = 8$,

$$\mathrm{E}\left(Z_{1i_1}^{\ell_1} Z_{1i_2}^{\ell_2} \cdots Z_{1i_q}^{\ell_q}\right) = \mathrm{E}\left(z_{1i_1}^{\ell_1}\right)\mathrm{E}\left(z_{1i_2}^{\ell_2}\right)\cdots\mathrm{E}\left(z_{1i_q}^{\ell_q}\right) \tag{2.5}$$

whenever $i_1, i_2, \ldots, i_q$ are distinct indices.

This model, employed earlier in Bai and Saranadasa (1996), maintains that the observations X_i are linearly generated by m-variate Z_i whose components are largely white noise. The latter is reflected by (2.5). We note that m, the dimension of Z_i, is arbitrary as long as $m \ge p$. This offers much flexibility in generating X_i with the given covariance Σ. The distribution of Z_i is unspecified, and hence is nonparametric.

The following assumption specifies the asymptotic framework for the high-dimensional inference.

Assumption 2. As $n \to \infty, p = p(n) \to \infty, \operatorname{tr}(\Sigma^2) \to \infty$ and $\operatorname{tr}(\Sigma^4)/\operatorname{tr}^2(\Sigma^2) \to 0$.

By applying Hölder's inequality, $\operatorname{tr}(\Sigma^2)/\operatorname{tr}^2(\Sigma) \le \operatorname{tr}(\Sigma^4)/\operatorname{tr}^2(\Sigma^2)$. Hence, Assumption 2 implies

$$\operatorname{tr}(\Sigma^2)/\operatorname{tr}^2(\Sigma) \to 0. \tag{2.6}$$

A similar reasoning shows that

$$\operatorname{tr}(\Sigma^3)/\{\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma^2)\} \to 0 \tag{2.7}$$

is valid under Assumption 2 as well.

In Assumption 2, we do not specify explicitly the growth rate of p relative to n as commonly made in existing works on high-dimensional data, but rather only requires that $\operatorname{tr}(\Sigma^4)$ grows at a slower rate than $\operatorname{tr}^2(\Sigma^2)$. This is in fact a weak proposition. To appreciate this, we note that if all the eigenvalues of Σ are bounded away from zero and infinity, Assumption 2 is trivially true for any p as long as $p \to \infty$. Some of the commonly encountered covariance structures satisfy Assumption 2. Let us first consider a correlation matrix

$$\Sigma = \left(\rho^{|j-i|}\right)_{p\times p} \tag{2.8}$$

for some $\rho \in (-1, 1)$. It prescribes that X_{1i} and X_{1j} are less correlated when j and i are further apart. It can be shown after some algebra that

$$\operatorname{tr}(\Sigma^2) = p/(1-\rho^2) + \rho^2(\rho^{2p}-1)/(1-\rho^2)^2 = O(p)$$

and

$$\operatorname{tr}(\Sigma^4) = 2\sum_{k=1}^{p-1}(p-k)(k+1)^2\rho^{2k} + p(1+\rho^2+7\rho^4-\rho^6)/(1-\rho^2)^3 + O(1).$$

It may be checked that $\operatorname{tr}(\Sigma^4) = O(p)$. Hence, $\operatorname{tr}(\Sigma^4) = o\{\operatorname{tr}^2(\Sigma^2)\}$.

Now consider another

$$\Sigma = \left(\sigma_i\sigma_j\rho^{|j-i|}\right)_{p\times p}, \tag{2.9}$$

where $\sigma_l^2 = \operatorname{var}(X_{1l})$ are the marginal variance for $l = 1,\ldots,p$. If $\{\sigma_l^2\}_{l=1}^p$ are uniformly bounded away from infinity and zero respectively. Then, it can be shown, based on the above result for the correlation matrix (2.8), that Assumption 2 is also satisfied. In fact, Assumption 2 is valid even when some σ_l^2 either diverges to the infinity or converges to 0 as long as the speed of divergence/convergence is not too fast relative to p's increase. The same can be said for

$$\Sigma = \left(\sigma_i\sigma_j\rho^{|j-i|}I(|j-i|\le d)\right)_{p\times p},$$

where d is a positive integer. This Σ implies a correlation structure where any two components in X_i are uncorrelated if they are more than d apart. As $\operatorname{tr}(\Sigma^4)$ is bounded by the corresponding term for matrix (2.9), it is $O(p)$. It can be shown that $\operatorname{tr}(\Sigma^2) = O(p)$ as well. Hence Assumption 2 holds for the Σ.

Let $\lambda_1,\ldots,\lambda_p$ be the eigenvalues of Σ. Ledoit and Wolf (2002) assume that both $p^{-1}\sum_{l=1}^p \lambda_l$ and $p^{-1}\sum_{l=1}^p(\lambda_l-\alpha)^2$ are free of p and n in additional to assuming $p/n \to c$ and $p^{-1}\sum_{l=1}^p \lambda_l^j$ converges for $j = 3$ and 4, respectively. Assumption 2 simplifies these conditions.

3. MAIN RESULTS

We first introduce notations which quantify the asymptotic variance of the test statistics. Let $A = \Gamma^T\Gamma$ and

$$\sigma_{1,n}^2 = \frac{4}{n^2} + \frac{8}{n}\operatorname{tr}\left[\left(\frac{\Sigma^2}{\operatorname{tr}(\Sigma^2)} - \frac{\Sigma}{\operatorname{tr}(\Sigma)}\right)^2\right] + \frac{4\Delta}{n}\operatorname{tr}\left[\left(\frac{A^2}{\operatorname{tr}(\Sigma^2)} - \frac{A}{\operatorname{tr}(\Sigma)}\right)\circ\left(\frac{A^2}{\operatorname{tr}(\Sigma^2)} - \frac{A}{\operatorname{tr}(\Sigma)}\right)\right],$$

$$\sigma_{2,n}^2 = \frac{4}{n^2}\operatorname{tr}^2(\Sigma^2) + \frac{8}{n}\operatorname{tr}(\Sigma^2-\Sigma)^2 + \frac{4\Delta}{n}\operatorname{tr}[(A^2-A)\circ(A^2-A)].$$

Here we define, for two matrices $C = (c_{ij})$ and $B = (b_{ij})$, $C \circ B = (c_{ij}b_{ij})$.

The following theorems establish the asymptotic normality of U_n and V_n respectively.

Theorem 1. Under Assumptions 1–2,

$$\sigma_{1,n}^{-1}\left[\left(\frac{U_n+1}{p}\right)\left(\frac{\operatorname{tr}^2(\Sigma)}{\operatorname{tr}(\Sigma^2)}\right) - 1\right] \xrightarrow{D} \mathrm{N}(0,1). \tag{3.1}$$

In particular, under the null hypothesis in (1.1), $nU_n \xrightarrow{D} \mathrm{N}(0,4)$.

Theorem 2. Under Assumptions 1–2,

$$\sigma_{2,n}^{-1}[pV_n - \operatorname{tr}(\Sigma - I_p)^2] \xrightarrow{D} \mathrm{N}(0,1). \tag{3.2}$$

In particular, under the null hypothesis in (1.2), $nV_n \xrightarrow{D} \mathrm{N}(0,4)$.

Based on the asymptotic normality under the respective null hypothesis, our proposed α-level test for sphericity rejects H_0 in (1.1) if $\frac{1}{2}nU_n \ge z_\alpha$ where z_α is the upper α quantile of $\mathrm{N}(0,1)$. Similarly, the α-level test for the identity hypothesis rejects the H_0 in (1.2) if $\frac{1}{2}nV_n \ge z_\alpha$. Interestingly, both tests have the same form of rejection region.

We note that U_n is invariant under linear rotation transformations, namely the value U_n remains unchanged for $X_i' = aAX_i + c$ where a is a constant, c is a vector of constants and A is an orthogonal matrix. At the same time, V_n is invariant under the shift and orthogonal rotation transformation $X_i' = AX_i + c$ where A and c have the same qualifications as above. Hence, the sphericity and identity tests we have proposed are invariant under linear rotational and rotational and shift transformations, respectively.

To evaluate the power of the sphericity test, we define

$$\delta_{1,n} = 1 - \frac{\operatorname{tr}^2(\Sigma)}{p\operatorname{tr}(\Sigma^2)} \quad \text{and} \quad \delta_{2,n} = \operatorname{tr}\left[\left(\frac{\Sigma^2}{\operatorname{tr}(\Sigma^2)} - \frac{\Sigma}{\operatorname{tr}(\Sigma)}\right)^2\right]. \tag{3.3}$$

Let $\sigma_S^2(\lambda) = \sum_{i=1}^p \lambda_i^2/p - (\sum_{i=1}^p \lambda_i/p)^2$ be the "variance" of the eigenvalues $\{\lambda_i\}_{i=1}^p$ of Σ. Then, $\delta_{1,n} = p\operatorname{tr}^{-1}(\Sigma^2)\sigma_S^2(\lambda)$ measures the departure from the null sphericity hypothesis in (1.1). Specifically, under H_1 of the sphericity hypothesis (1.1), $0 < \delta_{1,n} < 1$. To appreciate this point, we note by applying the Schwarz inequality, $\operatorname{tr}(\Sigma) \le \sqrt{p\operatorname{tr}(\Sigma^2)}$ and so $\delta_{1,n} \ge 0$. Since Σ is nonnegative definite, $\delta_{1,n} \le 1$ is trivially true. Furthermore, $\delta_{1,n} = 1$ if and only if $\Sigma = 0$, which contradicts the first part of Assumption 2. Also, $\delta_{1,n} = 0$ if and only if $\operatorname{tr}(\Sigma/\operatorname{tr}(\Sigma) - I_p/p)^2 = 0$, which is true under the null hypothesis of sphericity.

To describe the power of the test for the identity hypothesis, define

$$\rho_{1,n} = \frac{1}{p}\operatorname{tr}[(\Sigma - I_p)^2] \quad \text{and} \quad \rho_{2,n} = \frac{\operatorname{tr}(\Sigma^2)}{n\operatorname{tr}(\Sigma - I_p)^2}. \tag{3.4}$$

We note that $\rho_{1,n} = \sum_{i=1}^p (\lambda_i - 1)^2/p$ that measures the variation of the eigenvalues from 1 under the identity hypothesis. Clearly, $\rho_{1,n} = 0$ if and only if the identity hypothesis $\Sigma = I_p$ is true.

Let $\beta_{1,n} = P(\frac{1}{2}nU_n \ge z_\alpha | \Sigma \ne \sigma^2 I_p)$ and $\beta_{2,n} = P(\frac{1}{2}nV_n \ge z_\alpha | \Sigma \ne I_p)$, which are respectively the power of the sphericity and identity tests. The following theorem specifies the lower bound for the powers of the tests.

Theorem 3. Under the Assumptions 1–2, there exists finite positive constants $C_1^{(1)}$, $C_1^{(2)}$, $C_2^{(1)}$, and $C_2^{(2)}$ such that

$$\liminf_n \beta_{1,n} \ge 1 - \limsup_n \Phi\left(\frac{-C_1^{(1)} + C_1^{(2)}((1-\delta_{1,n})/(n\delta_{1,n}))}{\sqrt{1/(n^2\delta_{1,n}^2) + \delta_{2,n}/(n\delta_{1,n}^2)}}\right) \tag{3.5}$$

and

$$\liminf_n \beta_{2,n} \ge 1 - \limsup_n \Phi\left(-\frac{C_2^{(1)}}{\sqrt{n^2\rho_{2,n}^2 + n\rho_{2,n}}} + \frac{C_2^{(2)}}{n^2\rho_{1,n}\rho_{2,n}}\right). \tag{3.6}$$

The next theorem asserts the consistency of the two proposed tests.

Theorem 4 (Consistency). Under Assumptions 1–2, as $n \to \infty$:

(i) under H_1 in the sphericity hypothesis (1.1), if $n\delta_{1,n} \to \infty$, then $\delta_{2,n}/(n\delta_{1,n}^2) \to 0$ and $\beta_{1,n} \to 1$;
(ii) under H_1 in the identity hypothesis (1.2), if $n\rho_{1,n} \to \infty$, then $\rho_{2,n} \to 0$ and $\beta_{2,n} \to 1$.

As $\delta_{1,n}$ and $\rho_{1,n}$ measure the departure from the null hypotheses for the sphericity and identity hypotheses, Theorem 4 maintains that as long as the measures of the departure from the null hypotheses $\delta_{1,n}$ and $\rho_{1,n}$ are not shrinking faster than $1/n$, the tests are consistent. These results generalize those of Ledoit and Wolf (2002), who establish the consistency of their tests under the assumptions of $\delta_{1,n}$ and $\rho_{1,n}$ are fixed with respect to n, in which cases both $n\delta_{1,n}$ and $n\rho_{1,n}$ automatically diverge.

As mentioned in the Introduction, the proposed test for identity can be used to test for $H_0: \Sigma = \Sigma_0$ for a known covariance invertible Σ_0. This can be carried out by first transforming X_i to $\Sigma_0^{-1/2}X_i$, and applied the identity test on the transformed data. The idea of formulating the test statistics for sphericity and identity hypotheses can be used for testing $H_0: \Sigma = \operatorname{diag}\{\sigma_1^2, \ldots, \sigma_p^2\}$ for positive constants $\{\sigma_l^2\}_{l=1}^p$, which implies components of X_i are uncorrelated. As H_0 is valid if and only if $\operatorname{tr}(\Sigma^2) = \sum_{l=1}^p \sigma_l^4$, we could construct a test based on $T_{2,n} - \hat{S}_p$ where

$$\hat{S}_p = \sum_{l=1}^p \left\{ (P_n^2)^{-1} \sum_{i,j}^* X_{il}^2 X_{jl}^2 - 2(P_n^3)^{-1} \sum_{i,j,k}^* X_{il}X_{kl}X_{jl}^2 + (P_n^4)^{-1} \sum_{i,j,k,m}^* X_{il}X_{jl}X_{kl}X_{ml} \right\}$$

since $T_{2,n}$ and $\hat{S}_p$ are unbiased estimators of $\operatorname{tr}(\Sigma^2)$ and $S_p =: \sum_{l=1}^p \sigma_l^2$ respectively. The asymptotic property of the test statistic would be more involved than the test statistics for the sphericity and identity hypotheses. We would not pursue it due to a limited space.

4. SIMULATION RESULTS

We report results from simulation studies which were designed to evaluate the performance of the proposed sphericity and identity tests for the covariance matrix. For comparison purposes, we also conducted the tests proposed by Ledoit and Wolf (2002) (LW tests).

For the sphericity test $H_0: \Sigma = \sigma^2 I_p$, we generated p-dimensional independent and identical multivariate random vectors $\{X_i\}_{i=1}^n$ which were generated following the multivariate model in Assumption 1. In particular, we considered two scenarios with respect to the innovation random vector Z_i, the mean μ and the Γ matrix:

(I) Z_i were m-dimensional normal random vector with mean $\mathbf{0}_m$ and covariance I_m; $\Gamma = I_p$, $\mu = \mu_0\mathbf{1}_p$.
(II) $Z_i = (Z_{i1}, \ldots, Z_{im})'$ consisted of IID random variables Z_{ij} which were standardized Gamma(4, 0.5) random variables so that they had zero mean and unit variance; $\Gamma = I_p$ and $\mu = \mu_0\mathbf{1}_p$.

In both scenarios, we chose the dimension of Z_i be the same with that of X_i, namely $m = p$, and $\mu_0 = 2$ under the null hypothesis.

To mimic the "large p, small n" situation, we chose the dimensionality p according to $p = c_1\exp(n^\eta) + c_2$, where $\eta = 0.4$ for $(c_1, c_2) = (1, 10)$ and $(c_1, c_2) = (2, 0)$ respectively. These allowed p growing at an exponential rate of n. The sample sizes were $n = 20, 40, 60, 80$ which assigned $p = 38, 89, 181, 331$

for $(c_1, c_2) = (1, 10)$ and $p = 55, 159, 343, 642$ for $(c_1, c_2) = (2, 0)$. All p's were larger than the corresponding sample sizes. All the simulation results were based on 1000 simulations and the nominal significant level of the tests was 5%.

To evaluate the power of the sphericity test, two different forms of alternatives were considered in the simulations. In the first alternative, we generated multivariate random vectors Z_i as scenarios (I) and (II) given above under the null hypothesis, and set $\Gamma = \mathrm{diag}(\sqrt{2}\mathbf{1}_{[vp]}, \mathbf{1}_{p-[vp]})$, where $[x]$ denotes the integer truncation of x. We chose $\mu = 2\mathbf{1}_p$ in scenario (I) and $\mu = 2(\sqrt{2}\mathbf{1}'_{[vp]}, \mathbf{1}'_{p-[vp]})'$ in scenario (II). Two levels of v were considered: $v = 0.125$ and 0.25, which yielded diagonal covariance Σ which had respectively 12.5% and 25% of its diagonal elements being 2 whereas the rest were 1.

In the second alternative, we set $m = p + 1$ and $\Gamma = (\sqrt{1-\rho}I_p, \sqrt{2\rho}\mathbf{1}_p)$ for both scenarios; and set $\mu = 2\mathbf{1}_p$ for scenario (I) and $\mu = 2(\sqrt{1-\rho} + \sqrt{2\rho})\mathbf{1}_p$ for scenario (II). Therefore, the second alternative had $\Sigma = (1-\rho)I_p + 2\rho\mathbf{1}_p\mathbf{1}'_p$, the so-called compound symmetric covariance structure. This was an alternative structure we contemplated in the case study that will be reported in the next section. In our simulation, we set $\rho = 0.10$ and 0.15 respectively.

The above settings for evaluating the size and power of the sphericity test were also used for the identity test for $H_0: \Sigma = I_p$. As the simulation results for the identity test followed very much similar patterns to those of the sphericity test, we only report the simulation results for the sphericity test.

Table 1 reports empirical sizes of the proposed sphericity test and the test of Ledoit and Wolf (2002) (LW test) for both scenarios of distributions. These empirical sizes were applicable for both settings of the alternatives. Tables 2 and 3 report the empirical power of the proposed test and LW test for the two alternative settings respectively. We observe from Table 1 that under the scenario (I) of normal distribution, both the proposed test and the LW test had similar empirical sizes. The empirical size of the tests were converging to the nominal level as both p and n increase together. These were all assuring. However, under the scenario (II) of Gamma distribution, the LW test encountered serious size distortion while the proposed test still had reasonable sizes. This is understandable as the LW test was constructed based on the normality assumption and utilized the fact that the sample covariance is Wishart distributed. It was for this reason we do not report the power of the LW test in Tables 2 and 3 under the second scenario of the Gamma distribution. We note that there was some slight size distortion for the proposed tests when either p or n was small ($p = 38$ or $n = 20$) under the Gamma distribution scenario. This could be understood as the test is both asymptotic and nonparametric. However, as p and n both were increased, the sizes of the proposed test were quite close to the nominal 5%.

The power results in Tables 2 and 3 showed the proposed test and the LW test under normality had quite good power, and they were largely comparable. The powers of the proposed test were largely dependent on (i) the sample size n, and (ii) the variation percentage measures v and ρ as they determine $\delta_{1,n}$ and $\delta_{2,n}$, the two quantities which determined the asymptotic power of the test as shown in Theorems 3 and 4. We find the powers in Table 2 were less affected by the increased dimensionality, as compared to Table 3. We were a little surprised that despite the fact that the LW test was justified under $p/n \to c$, it had quite resilient power when p was a lot larger than n for the normal distribution scenario. The powers under the second alternative as reported in Table 3 increased much faster than those under the first alternative reported in Table 2 as the sample size and the dimension were increased. And when ρ was increased from 0.1 to 0.15 under the second alternative, many entries of the empirical powers of the tests approach to 1. This could be viewed

Table 1. Empirical sizes of the proposed sphericity test and the LW test for $H_0: \Sigma = \sigma^2 I_p$ at 5% significance for normal and gamma random vectors

	LW test n				Proposed test n			
p	20	40	60	80	20	40	60	80
				(a) Normal random vectors				
38	0.045	0.055	0.054	0.054	0.061	0.061	0.060	0.063
55	0.051	0.042	0.054	0.056	0.070	0.050	0.062	0.056
89	0.054	0.051	0.047	0.068	0.066	0.054	0.054	0.072
159	0.061	0.054	0.046	0.045	0.068	0.065	0.044	0.048
181	0.048	0.051	0.049	0.049	0.062	0.057	0.052	0.052
331	0.061	0.062	0.047	0.052	0.078	0.069	0.059	0.059
343	0.044	0.062	0.056	0.053	0.062	0.073	0.062	0.060
642	0.062	0.051	0.042	0.050	0.080	0.059	0.041	0.055
				(b) Gamma random vectors				
38	0.174	0.199	0.173	0.165	0.092	0.078	0.060	0.056
55	0.177	0.168	0.198	0.168	0.083	0.065	0.068	0.048
89	0.173	0.183	0.157	0.174	0.088	0.068	0.049	0.046
159	0.188	0.185	0.206	0.192	0.078	0.063	0.064	0.055
181	0.174	0.187	0.172	0.185	0.058	0.059	0.059	0.062
331	0.177	0.186	0.189	0.160	0.084	0.050	0.064	0.042
343	0.206	0.191	0.164	0.176	0.084	0.070	0.048	0.056
642	0.167	0.188	0.191	0.184	0.064	0.060	0.057	0.062

Table 2. Empirical powers of the proposed sphericity tests and the LW test for $H_0 : \Sigma = \sigma_0^2 I_p$ versus $H_a : \Sigma = \sigma_0^2 I_p + \sigma_1^2 A_p$ at 5% significance with $A_p = \mathrm{diag}(\mathbf{1}_{[vp]}, \mathbf{0}_{p-[vp]})$ and $\sigma_0^2 = \sigma_1^2 = 1$

	LW test n				Proposed test n			
p	20	40	60	80	20	40	60	80
(a) Normal random vectors								
$v = 0.125$								
38	0.203	0.400	0.671	0.830	0.230	0.418	0.669	0.840
55	0.179	0.447	0.699	0.860	0.210	0.476	0.709	0.863
89	0.215	0.452	0.764	0.918	0.232	0.473	0.775	0.923
159	0.188	0.489	0.752	0.924	0.232	0.508	0.757	0.928
181	0.212	0.507	0.774	0.931	0.239	0.512	0.772	0.920
331	0.212	0.509	0.779	0.936	0.246	0.517	0.776	0.931
343	0.228	0.514	0.793	0.942	0.260	0.513	0.785	0.942
642	0.202	0.503	0.805	0.949	0.241	0.503	0.797	0.946
$v = 0.250$								
38	0.318	0.682	0.906	0.978	0.337	0.696	0.912	0.977
55	0.288	0.698	0.910	0.990	0.332	0.700	0.917	0.989
89	0.309	0.697	0.933	0.997	0.337	0.706	0.938	0.997
159	0.307	0.725	0.949	0.994	0.342	0.725	0.952	0.997
181	0.324	0.733	0.950	0.996	0.349	0.738	0.941	0.997
331	0.332	0.730	0.942	0.999	0.360	0.740	0.947	0.999
343	0.311	0.742	0.945	0.996	0.336	0.746	0.945	0.997
642	0.323	0.724	0.969	0.995	0.339	0.709	0.972	0.996
(b) Gamma random vectors								
	$v = 0.125$				$v = 0.250$			
38	0.212	0.414	0.615	0.780	0.332	0.639	0.871	0.976
55	0.224	0.434	0.645	0.827	0.341	0.651	0.890	0.973
89	0.253	0.477	0.743	0.889	0.351	0.682	0.918	0.987
159	0.242	0.477	0.759	0.917	0.325	0.715	0.926	0.991
181	0.243	0.531	0.781	0.908	0.328	0.717	0.933	0.994
331	0.262	0.514	0.790	0.947	0.331	0.717	0.940	0.994
343	0.240	0.531	0.775	0.938	0.338	0.725	0.954	0.994
642	0.262	0.534	0.796	0.947	0.343	0.747	0.951	0.995

as an empirical indication of the proposed test being consistent.

5. AN EMPIRICAL STUDY

In genetic microarray analysis, a common practice for testing differentially expressed genes is to apply a t-test on each individual gene and then use a multiple comparison procedure to control the Family Wise Error Rate (FWER) or the False Discover Rate (FDR). The latest development in biological studies focus on sets of genes, as genes tend to work together to achieve certain biological tasks. Identifying significant gene sets, instead of individual genes, for certain treatments under evaluation is gaining substantial interests; see Barry, Nobel, and Wright (2005), Efron and Tibshrini (2007), and Newton et al. (2007). Gene sets are defined under the Gene Ontology (GO) system that gives structured and controlled vocabularies which produce names of gene sets (also called GO terms). Regardless, our interest is rested on genes or set of genes, the correlation structure of the high-dimensional data can have significant implications on the statistical procedures used in the analyses for genetic data.

The dataset we analyzed came from an experiment conducted by Department of Animal Science in Iowa State University (for more detail of this experiment, see Lkhagvadorj et al. 2009). In this experiment, 24 six-month-old Yorkshire Gilts are equally divided into four groups (blocks). There are two genotypes with 12 pigs each. Two different diet treatments were randomly assigned to each of the 12 pigs in each genotype. One has no restriction on the amount of feed consumed by the pigs, and the other had feed abstained for three days. For each pig in the study, the microarray gene expressions were measured. The aim of the study was to identify treatment effects on the gene-expression levels. What we are interested is testing the correlation structure among the genes within each gene set.

There are 24,123 genes in the microarray data and 4538 gene sets with the number of genes ranging from 2 to 2000. For GO terms with dimension less than 10, we applied the sphericity and equality tests proposed by John (1971) and Nagao (1973) assuming p is fixed. For the other GO terms with dimensions at least 10, we applied the proposed test procedures given in Section 3.

Denote $\mathcal{S}_1, \ldots, \mathcal{S}_q$, $q = 4538$, for the gene sets to be studied, and the gene set $\mathcal{S}_g$ has p_g genes. To make the marginal

Table 3. Empirical powers of the proposed sphericity tests and the LW test for $H_0: \Sigma = \sigma_0^2 I_p$ versus $H_a: \Sigma = (1-\rho)\sigma_0^2 I_p + \rho\sigma_1^2 \mathbf{1}_p \mathbf{1}_p'$ at 5% significance with $\sigma_0^2 = 1$ and $\sigma_1^2 = 2$

	LW test n				Proposed test n			
p	20	40	60	80	20	40	60	80
				(a) Normal random vectors				
				$\rho = 0.10$				
38	0.094	0.117	0.151	0.220	0.105	0.128	0.168	0.228
55	0.107	0.185	0.264	0.329	0.128	0.198	0.275	0.336
89	0.142	0.258	0.413	0.596	0.164	0.261	0.426	0.614
159	0.225	0.488	0.724	0.882	0.256	0.522	0.733	0.887
181	0.265	0.574	0.810	0.931	0.282	0.589	0.809	0.927
331	0.489	0.847	0.977	0.996	0.495	0.865	0.971	0.995
343	0.512	0.865	0.978	1.000	0.522	0.866	0.976	1.000
642	0.728	0.982	0.999	1.000	0.743	0.983	0.999	1.000
				$\rho = 0.15$				
38	0.303	0.606	0.832	0.931	0.326	0.613	0.840	0.932
55	0.474	0.794	0.948	0.993	0.475	0.807	0.947	0.992
89	0.634	0.935	0.995	0.999	0.663	0.937	0.993	0.999
159	0.860	0.995	1.000	1.000	0.850	0.994	1.000	1.000
181	0.884	1.000	1.000	1.000	0.889	0.999	1.000	1.000
331	0.975	1.000	1.000	1.000	0.976	1.000	1.000	1.000
343	0.958	1.000	1.000	1.000	0.962	1.000	1.000	1.000
642	0.999	1.000	1.000	1.000	0.997	1.000	1.000	1.000
				(b) Gamma random vectors				
	$\rho = 0.10$				$\rho = 0.15$			
38	0.103	0.118	0.176	0.261	0.307	0.614	0.802	0.900
55	0.138	0.190	0.280	0.350	0.428	0.760	0.925	0.968
89	0.186	0.295	0.404	0.557	0.625	0.927	0.987	0.995
159	0.267	0.469	0.730	0.821	0.779	0.980	1.000	1.000
181	0.311	0.541	0.799	0.889	0.833	0.991	1.000	1.000
331	0.485	0.805	0.950	0.988	0.945	0.999	1.000	1.000
343	0.508	0.828	0.954	0.993	0.955	0.998	0.999	1.000
642	0.707	0.967	0.998	1.000	0.994	1.000	1.000	1.000

variance homogeneous across genes, we applied a variance stabilization transformation as given in Huber et al. (2002). Let y_{ijkl}^g be the p_g dimensional vector of gene expression levels (after the transformation) measured for the gth gene set, the lth pig in the ith treatments, and jth block with kth genotype. We assume the following factorial design model:

$$y_{ijkl}^g = \tau^g + \mu_i^g + \beta_j \mathbf{1}_{p_g} + \alpha_k \mathbf{1}_{p_g} + \eta_G \mathbf{1}_{p_g} + \epsilon_{ijkl}^g, \qquad (5.1)$$

where τ^g and μ_i^g $(i = 1, 2)$ are p_g dimensional vectors denoting the intercepts and the treatment effects respectively, β_j $(j = 1, \ldots, 4)$, α_k $(k = 1, 2)$, and η_G are univariate denoting the block, the genotype, and the gene-set effects respectively, ϵ_{ijkl}^g is the vector of residuals, and $\mathbf{1}_{p_g}$ is the p_g dimensional vector of 1's. The treatment and genotype effects μ_i^g and α_k are treated as fixed, while the block effect and the GO term effects are random. In addition, due to the factorial design, μ_i^g and α_k are subjected to constraints $\mu_2^g = \alpha_2 = 0$.

We are interested in testing for the covariance structure of $\eta_G \mathbf{1}_{p_g} + \epsilon_{ijkl}^g$. It is not too restrictive to assume $\mathrm{var}(\epsilon_{ijkl}^g) = \sigma_\epsilon^2 I_{p_g}$ for a positive constant σ_ϵ^2. Hence, the covariance of $\eta_G \mathbf{1}_{p_g} + \epsilon_{ijkl}^g$ is

$$\Sigma =: \sigma_\epsilon^2 I_{p_g} + \sigma_{\eta_G}^2 \mathbf{1}_{p_g} \mathbf{1}_{p_g}^T,$$

say, where $\sigma_{\eta_G}^2 = \mathrm{var}(\eta_G)$. Let $\Sigma_{\mathcal{S}_g}$ be the covariance matrix corresponding to the gene set $\mathcal{S}_g$. We want to test the sphericity and identity hypotheses $H_{0a}: \Sigma_{\mathcal{S}_g} = \sigma_\epsilon^2 I_{p_g}$ and $H_{0b}: \Sigma_{\mathcal{S}_g} = I_{p_g}$ for $g = 1, \ldots, q$. Both hypotheses will rule out the gene set effect as both imply $\sigma_{\eta_G}^2 = 0$.

To remove the treatment, block and the genotype effects, we estimate μ_i^g and τ^g by

$$\hat{\mu}_i^g = \frac{1}{n(i)} \sum_{j,k,l} y_{ijkl}^g - \frac{1}{n(2)} \sum_{j,k,l} y_{2jkl}^g \qquad \text{and}$$

$$\hat{\tau}^g = \frac{1}{n(2)} \sum_{j,k,l} y_{2jkl}^g,$$

where $n(i)$ is the number of observations in treatment i. Then, estimate β and α by fitting the linear model

$$y_{ijkl}^g - \hat{\mu}_i^g - \hat{\tau}^g = \beta_j I_{p_g} + \alpha_k \mathbf{1}_{p_g} + e_{ijkl}^g. \qquad (5.2)$$

Here e_{ijkl}^g are the "residuals" representing $\eta_G \mathbf{1}_{p_g} + \epsilon_{ijkl}^g$ in model (5.1) via the least square regression.

The histograms of the p-values of the sphericity and identity tests based on the estimated residuals from (5.2) for the gene sets are displayed in Figure 1. When we controlled the FDR at

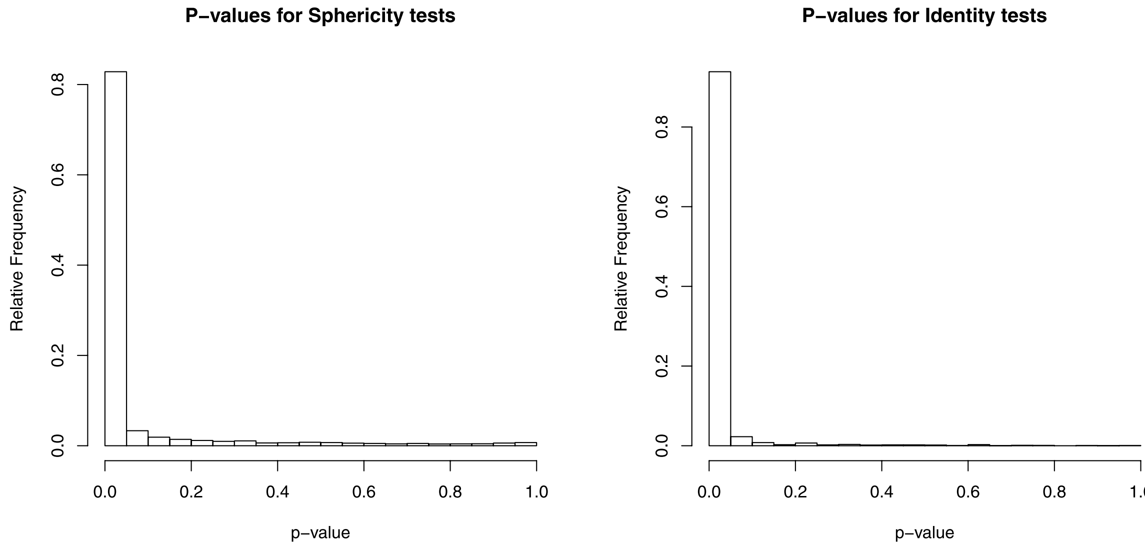

Figure 1. Histograms of p-values for the tests H_{0a} and H_{0b}.

5%, there were 3730 gene sets are significant for the sphericity tests and 4251 gene sets are significant for the identity test. These numbers constitute a high proportion of the total number of gene sets 4538, and indicate that the column-wise independence assumption are not reasonable. That more gene sets were tested significant under identity hypothesis is understandable as the identity hypothesis is more restrictive than the sphericity hypothesis. It is natural to expect that the genes within a gene set are correlated. The study confirms that there were substantial amounts of gene-wise dependence in this dataset. Hence, any procedures which are affected by the dependence has to be used with great care.

APPENDIX

We first report two propositions which contains some basic results which will be used throughout in the appendix. The proofs are elementary and are available in a technical report which is available on request.

Proposition A.1. Under Assumption 1, we have:

(i) for any $m \times m$ symmetric matrices B_1 and B_2,

$$\mathrm{E}\{(Z_1'B_1Z_1)(Z_1'B_2Z_1)\} = \mathrm{tr}(B_1)\mathrm{tr}(B_2) + 2\,\mathrm{tr}(B_1B_2) + \Delta\,\mathrm{tr}(B_1 \circ B_2); \quad \text{(A.1)}$$

(ii) let $A = \Gamma'\Gamma$. Then

$$\mathrm{E}\{(Z_1'AZ_2)^4\} = 3\,\mathrm{tr}^2(\Sigma^2) + 6\mathrm{tr}(\Sigma^4) + 6\Delta\,\mathrm{tr}(A^2 \circ A^2) + \Delta^2 \sum_{i,k=1}^{m} (A_{ik})^4; \quad \text{(A.2)}$$

(iii) for any $m \times m$ positive definite matrix B, there exists a constant C such that

$$\mathrm{E}\big\{(Z_1'BZ_1 - \mathrm{tr}(B))^4\big\} \le C\,\mathrm{tr}^2(B^2); \quad \text{(A.3)}$$

(iv) let A_{ik} be the (i,k)th entry of $A = \Gamma'\Gamma$. Then $\sum_{i,k}(A_{ik})^4 \le \mathrm{tr}(\Sigma^4)$.

Proposition A.2. Let $A = \Gamma'\Gamma$. Then: (i) $\mathrm{var}(Y_{1,n}) = 2n^{-1}\,\mathrm{tr}(\Sigma^2) + \Delta n^{-1}\,\mathrm{tr}(A \circ A)$; (ii) $\mathrm{var}(Y_{2,n}) = 4n^{-2}\,\mathrm{tr}^2(\Sigma^2) + 8n^{-1}\,\mathrm{tr}(\Sigma^4) + 4\Delta n^{-1}\,\mathrm{tr}(A^2 \circ A^2) + O(n^{-3}\,\mathrm{tr}^2(\Sigma^2) + n^{-2}\mathrm{tr}(\Sigma^4))$; (iii) $\mathrm{var}(Y_{3,n}) = 2/[n(n-1)]\mathrm{tr}(\Sigma^2)$; (iv) $\mathrm{cov}(Y_{1,n}, Y_{2,n}) = 4n^{-1}\,\mathrm{tr}(\Sigma^3) + 2\Delta n^{-1} \times \mathrm{tr}(A^2 \circ A)$; (v) $\mathrm{cov}(Y_{1,n}, Y_{3,n}) = 0$; (vi) $\mathrm{var}(Y_{4,n}) = 2n^{-3}\,\mathrm{tr}^2(\Sigma^2) + 2n^{-2}\,\mathrm{tr}(\Sigma^4) + O(n^{-4}\,\mathrm{tr}^2(\Sigma^2) + n^{-3}\mathrm{tr}(\Sigma^4))$; (vii) $\mathrm{var}(Y_{5,n}) = 8n^{-4} \times \mathrm{tr}^2(\Sigma^2) + O(n^{-5}\,\mathrm{tr}^2(\Sigma^2) + n^{-4}\mathrm{tr}(\Sigma^4))$.

Proposition A.3. (i) Let $Z_n = a_nY_{1,n} + b_nY_{2,n}$, for $n = 1, 2, \ldots$, where a_n, b_n are arbitrary real numbers. Then under Assumptions 1–2, as $n \to \infty$

$$\frac{Z_n - \mathrm{E}Z_n}{\mathrm{var}(Z_n)} \xrightarrow{D} \mathrm{N}(0, 1).$$

(ii) Let $\mathrm{var}[(Y_{1,n}, Y_{2,n})'] = B_n$. Then

$$B_n^{-1/2}(Y_{1,n} - \mu_{1,n}, Y_{2,n} - \mu_{2,n})' \xrightarrow{D} \mathrm{N}(0, I_2),$$

where $\mu_{1,n}$ and $\mu_{2,n}$ are, respectively, the means of $Y_{1,n}$ and $Y_{2,n}$.

Proof. We only need to show (i) as the conclusion in (ii) is a direct consequence of (i) by theorem 29.4 of Billingsley (1995). To show (i), we need to use the martingale central limit theorem. For that purpose, let $\mathcal{F}_0 = \{\emptyset, \Omega\}$, $\mathcal{F}_k = \sigma\{X_1, \ldots, X_k\}$, $k = 1, 2, \ldots, n$. Let $\mathrm{E}_k(\cdot)$ denote the conditional expectation of given $\mathcal{F}_k$ $[\mathrm{E}_0(\cdot) = \mathrm{E}(\cdot)]$. Write $Z_n - \mathrm{E}Z_n = \sum_{k=1}^{n} D_{n,k}$, where $D_{n,k} = (\mathrm{E}_k - \mathrm{E}_{k-1})Z_n$. Then for every n, $\{D_{n,k}, 1 \le k \le n\}$ is a martingale difference sequence with respect to the σ-fields $\{\mathcal{F}_k, 1 \le k \le n\}$.

By the martingale central limit theorem (Billingsley 1995, p. 476), it suffices to show that, letting $\sigma_{n,k}^2 = \mathrm{E}_{k-1}(D_{n,k}^2)$, as $n \to \infty$,

$$\frac{\sum_{k=1}^{n} \sigma_{n,k}^2}{\mathrm{var}(Z_n)} \xrightarrow{P} 1 \qquad \text{and} \qquad \frac{\sum_{k=1}^{n} \mathrm{E}(D_{n,k}^4)}{\mathrm{var}^2(Z_n)} \to 0. \quad \text{(A.4)}$$

We first show the first part of (A.4). As it is true $\mathrm{E}(\sum_{k=1}^{n} \sigma_{n,k}^2) = \mathrm{var}(Z_n)$, we only show $\mathrm{var}(\sum_{k=1}^{n} \sigma_{n,k}^2) = o(\mathrm{var}^2(Z_n))$. From (i), (ii), and (iv) of Proposition A.2,

$$\mathrm{var}(Z_n) = \frac{2}{n}\,\mathrm{tr}(\Xi_n^2) + \frac{\Delta}{n}\mathrm{tr}(\Lambda_n \circ \Lambda_n) + \frac{4b_n^2}{n^2}(1 + O(n^{-1}))\mathrm{tr}^2(\Sigma^2),$$

where $\Xi_n = a_n\Sigma + 2b_n\Sigma^2$ and $\Lambda_n = (a_nA + 2b_nA^2)$. It hence follows

$$\mathrm{var}^2(Z_n) \ge K \max\{b_n^2 n^{-3}\,\mathrm{tr}(\Xi_n^2)\mathrm{tr}^2(\Sigma^2), b_n^4 n^{-4}\,\mathrm{tr}^4(\Sigma^2)\} \quad \text{(A.5)}$$

for some constant K for all large n.

To evaluate $\sum_{k=1}^{n}\sigma_{n,k}^2$, define $u_{n,k}=(\mathrm{E}_k-\mathrm{E}_{k-1})Y_{1,n}$ and $v_{n,k}=(\mathrm{E}_k-\mathrm{E}_{k-1})Y_{2,n}$. Then $D_{n,k}=a_n u_{n,k}+b_n v_{n,k}$ so that $\sigma_{n,k}^2=a_n^2\mathrm{E}_{k-1}(u_{n,k}^2)+2a_nb_n\mathrm{E}_{k-1}(u_{n,k}v_{n,k})+b_n^2\mathrm{E}_{k-1}(v_{n,k}^2)$. In the next, let us denote $Q_{k-1}=\sum_{i=1}^{k-1}(X_iX_i'-\Sigma)$, $M_{k-1}=\Gamma' Q_{k-1}\Gamma$, $\xi_{2,k}=\sum_{i=1}^{k-1}\{X_i'\Sigma^2X_i-\mathrm{tr}(\Sigma^3)\}$, and $\xi_{3,k}=\sum_{i=1}^{k-1}\{X_i'\Sigma^3X_i-\mathrm{tr}(\Sigma^4)\}$.

Noting that $Y_{1,n}=(1/n)X_k'X_k+(1/n)\sum_{i\notin\{k\}}X_i'X_i$. Then,

$$u_{n,k}=(\mathrm{E}_k-\mathrm{E}_{k-1})(1/n)X_k'X_k=(1/n)\{X_k'X_k-\mathrm{tr}(\Sigma)\}. \quad (\mathrm{A.6})$$

Similarly,

$$\begin{aligned} v_{n,k}&=(\mathrm{E}_k-\mathrm{E}_{k-1})\frac{2}{n(n-1)}\sum_{i\notin\{k\}}(X_i'X_k)^2\\ &=\frac{2}{n(n-1)}\{X_k'Q_{k-1}X_k-\mathrm{tr}(Q_{k-1}\Sigma)\}\\ &\quad+\frac{2}{n}\{X_k'\Sigma X_k-\mathrm{tr}(\Sigma^2)\}. \end{aligned} \quad (\mathrm{A.7})$$

Note that $\mathrm{E}_{k-1}(u_{n,k}^2)=C_0$, say, is a constant. Then, from (A.1),

$$\mathrm{E}_{k-1}(u_{n,k}v_{n,k})=\frac{4}{n^2(n-1)}\xi_{2,k}+\frac{2\Delta}{n^2(n-1)}\mathrm{tr}(A\circ M_{k-1})+C_1,$$

$$\begin{aligned}\mathrm{E}_{k-1}(v_{n,k}^2)&=\frac{16}{n^2(n-1)}\xi_{3,k}+\frac{8}{n^2(n-1)^2}\mathrm{tr}(Q_{k-1}\Sigma Q_{k-1}\Sigma)\\ &\quad+\frac{8\Delta}{n^2(n-1)}\mathrm{tr}(A^2\circ M_{k-1})\\ &\quad+\frac{4\Delta}{n^2(n-1)^2}\mathrm{tr}(M_{k-1}\circ M_{k-1})+C_2,\end{aligned}$$

where C_1 and C_2 are two constants. Let $C=a_n^2C_0+C_1+C_2$. Then,

$$\sum_{k=1}^{n}\sigma_{n,k}^2=R_{1,n}+R_{2,n}+R_{3,n}+R_{4,n}+Cn,$$

where

$$R_{1,n}=\frac{8b_n}{n^2(n-1)}\sum_{k=1}^{n}(a_n\xi_{2,k}+2b_n\xi_{3,k}),$$

$$R_{2,n}=\frac{4b_n\Delta}{n^2(n-1)}\sum_{k=1}^{n}\mathrm{tr}(\Lambda_n\circ M_{k-1}),$$

$$R_{3,n}=\frac{8b_n^2}{n^2(n-1)^2}\sum_{k=1}^{n}\mathrm{tr}(M_{k-1}^2),$$

and

$$R_{4,n}=\frac{4\Delta b_n^2}{n^2(n-1)^2}\sum_{k=1}^{n}\mathrm{tr}(M_{k-1}\circ M_{k-1}).$$

We need to show $\mathrm{var}(R_{i,n})=o(\mathrm{var}^2(Z_n))$ for $i=1,\ldots,4$. Using (A.1), for any positive integers ℓ_1,ℓ_2,

$$\mathrm{cov}(X_1'\Sigma^{\ell_1}X_1,X_1'\Sigma^{\ell_2}X_1)=\Delta\,\mathrm{tr}(A^{\ell_1+1}\circ A^{\ell_2+1})+2\,\mathrm{tr}(\Sigma^{\ell_1+\ell_2+2}).$$

It follows that for any $k\le j$,

$$\begin{aligned}&\mathrm{cov}(a_n\xi_{2,k}+2b_n\xi_{3,k},a_n\xi_{2,j}+2b_n\xi_{3,j})\\ &\quad=(k-1)\{2\,\mathrm{tr}(\Xi_n^2\Sigma^4)+\Delta\mathrm{tr}[(\Lambda_nA^2)\circ(\Lambda_nA^2)]\}\\ &\quad\le(k-1)(2+\Delta)\,\mathrm{tr}(\Xi_n^2\Sigma^4)\\ &\quad\le(k-1)(2+\Delta)\,\mathrm{tr}(\Xi_n^2)\mathrm{tr}(\Sigma^4),\end{aligned}$$

where we applied the fact $\mathrm{tr}[(\Lambda_nA^2)\circ(\Lambda_nA^2)]\le\mathrm{tr}(\Xi_n^2\Sigma^4)$. Thus, $\mathrm{var}(R_{1,n})\le Kb_n^2n^{-3}\,\mathrm{tr}(\Sigma^4)\,\mathrm{tr}(\Xi_n^2)$ and hence by (A.5)

$$\frac{\mathrm{var}(R_{1,n})}{\mathrm{var}^2(Z_n)}\le K\frac{\mathrm{tr}(\Sigma^4)}{\mathrm{tr}^2(\Sigma^2)}\to 0.$$

By carrying out similar procedures we can show that the above is true for $R_{l,n}$ for $l=2,3$, and 4, and hence completes the proof for the first part of (A.4).

To show the second part of (A.4), we note from (A.6) and (A.7),

$$D_{n,k}=\frac{1}{n}\{Z_k'\Lambda_nZ_k-\mathrm{tr}(\Lambda_n)\}+\frac{2b_n}{n(n-1)}\{Z_k'M_{k-1}Z_k-\mathrm{tr}(M_{k-1})\}.$$

By (A.3),

$$\begin{aligned}\sum_{k=1}^{n}\mathrm{E}(D_{n,k}^4)&\le 8\Bigg[\frac{1}{n^3}\mathrm{E}\{Z_k'\Lambda_nZ_k-\mathrm{tr}(\Lambda_n)\}^4\\ &\qquad+\frac{16b_n^4}{n^4(n-1)^4}\sum_{k=1}^{n}\mathrm{E}\{Z_1'M_{k-1}Z_1-\mathrm{tr}(M_{k-1})\}^4\Bigg]\\ &\le K(n^{-3}\,\mathrm{tr}^2(\Xi_n^2)+b_n^4n^{-6}\,\mathrm{tr}^4(\Sigma^4))\\ &=O(n^{-1}\,\mathrm{var}^2(Z_n)).\end{aligned} \quad (\mathrm{A.8})$$

This proves the second part of (A.4) and finishes the proof of this proposition.

Proof of Theorem 1

Let $\mu_{1,n}=\mathrm{E}(T_{1,n})$ and $\mu_{2,n}=\mathrm{E}(T_{2,n})$. Then $\mu_{1,n}=\mathrm{tr}(\Sigma)$, $\mu_{2,n}=\mathrm{tr}(\Sigma^2)$. Write

$$\tilde U_n=T_{2,n}/\mu_{2,n}-2T_{1,n}/\mu_{1,n}+1,\qquad \varepsilon_n=(T_{1,n}-\mu_{1,n})/\mu_{1,n}.$$

Then

$$\left(\frac{\mathrm{tr}^2(\Sigma)}{\mathrm{tr}(\Sigma^2)}\right)\left(\frac{U_n+1}{p}\right)-1=\frac{\tilde U_n-\varepsilon_n^2}{(1+\varepsilon_n)^2}.$$

It suffices to show $\varepsilon_n\xrightarrow{P}0$ and $\sigma_{1,n}^{-1}(\tilde U_n-\varepsilon_n^2)\xrightarrow{D}\mathrm{N}(0,1)$.

By (2.3), (i), (iii), and (v) of Proposition A.2,

$$\begin{aligned}\mathrm{var}(\varepsilon_n)&=\frac{1}{\mathrm{tr}^2(\Sigma)}\left[\frac{2}{n}\mathrm{tr}(\Sigma^2)+\frac{\Delta}{n}\mathrm{tr}(A\circ A)+\frac{2}{n(n-1)}\mathrm{tr}(\Sigma^2)\right]\\ &\le\left[\frac{2+\Delta}{n}+\frac{2}{n(n-1)}\right]\frac{\mathrm{tr}(\Sigma^2)}{\mathrm{tr}^2(\Sigma)},\end{aligned} \quad (\mathrm{A.9})$$

where $\mathrm{tr}(\Sigma^2)/\mathrm{tr}^2(\Sigma)\to 0$. Thus, $\varepsilon_n\xrightarrow{P}0$.

Furthermore, $\tilde U_n=[Y_{2,n}/\mu_{2,n}-2Y_{1,n}/\mu_{1,n}+1]+2Y_{3,n}/\mu_{1,n}-2Y_{4,n}/\mu_{2,n}+Y_{5,n}/\mu_{2,n}$. From (i), (ii), and (iv) of Proposition A.2,

$$\mathrm{var}(Y_{2,n}/\mu_{2,n}-2Y_{1,n}/\mu_{1,n}+1)=\sigma_{1,n}^2+O\left(\frac{1}{n^3}\right)+O\left(\frac{1}{n^2}\right)\frac{\mathrm{tr}(\Sigma^2)}{\mathrm{tr}^2(\Sigma)}.$$

Also, it is straightforward to see $\mathrm{var}(Y_{3,n}/\mu_{1,n})=O(\frac{1}{n^2}\frac{\mathrm{tr}(\Sigma^2)}{\mathrm{tr}^2(\Sigma)})=o(\frac{1}{n^2})=o(\sigma_{1,n}^2)$, $\mathrm{var}(Y_{4,n}/\mu_{2,n})=O(\frac{1}{n^3}+\frac{1}{n^2}\frac{\mathrm{tr}(\Sigma^4)}{\mathrm{tr}^2(\Sigma^2)})=o(\frac{1}{n^2})=o(\sigma_{1,n}^2)$, and $\mathrm{var}(Y_{5,n}/\mu_{2,n})=O(\frac{1}{n^3})=o(\sigma_{1,n}^2)$. Hence by (i) of Proposition A.3, $\sigma_{1,n}^{-1}\tilde U_n\xrightarrow{D}\mathrm{N}(0,1)$. Finally, that $\mathrm{E}(\varepsilon_n^2)=o(\frac{1}{n})=o(\sigma_{1,n})$ implies $\sigma_{1,n}^{-1}\varepsilon_n\xrightarrow{P}0$. Thus, $\sigma_{1,n}^{-1}(\tilde U_n-\varepsilon_n^2)\xrightarrow{D}\mathrm{N}(0,1)$.

Proof of Theorem 2

Similar to the proof of Theorem 1, we have $pV_n = (Y_{2,n} - 2Y_{1,n} + p) + 2Y_{3,n} - 2Y_{4,n} + Y_{5,n}$, in which $\mathrm{E}(Y_{2,n} - 2Y_{1,n} + p) = \mathrm{tr}(\Sigma - I)^2$, $\mathrm{var}(Y_{2,n} - 2Y_{1,n} + p) = \sigma_{2,n}^2 + O(n^{-3}\,\mathrm{tr}^2(\Sigma^2) + n^{-2}\,\mathrm{tr}(\Sigma^4))$, and $\mathrm{var}(Y_{3,n}) = o(\sigma_{2,n}^2)$, $\mathrm{var}(Y_{4,n}) = o(\sigma_{2,n}^2)$, $\mathrm{var}(Y_{5,n}) = o(\sigma_{2,n}^2)$. By Proposition A.2, $\sigma_{2,n}^{-1}[pV_n - \mathrm{tr}(\Sigma - I)^2] \xrightarrow{D} \mathrm{N}(0, 1)$. This completes the proof.

Proof of Theorem 3

(i) From Theorem 1,

$$\liminf_n \beta_{1,n} \ge 1 - \limsup_n \Phi\left(\frac{(1-\delta_{1,n})(2z_\alpha/n) - \delta_{1,n}}{\sigma_{1,n}}\right).$$

Let $C_1 = 1/\sqrt{8+4\Delta}$ and $C_2 = 2z_\alpha$. Then,

$$2z_\alpha C_2^{-1}\sqrt{\frac{1}{n^2} + \frac{1}{n}\delta_{2,n}} \le \sigma_{1,n} \le C_1^{-1}\sqrt{\frac{1}{n^2} + \frac{1}{n}\delta_{2,n}}.$$

Hence,

$$\limsup_n \Phi\left(\frac{(1-\delta_{1,n})(2z_\alpha/n) - \delta_{1,n}}{\sigma_{1,n}}\right)$$
$$\le \limsup_n \left[\Phi\left(\frac{(1-\delta_{1,n})(2z_\alpha/n)}{2z_\alpha[C_2]^{-1}\sqrt{1/n^2 + (1/n)\delta_{2,n}}} - \frac{\delta_{1,n}}{C_1^{-1}\sqrt{1/n^2 + (1/n)\delta_{2,n}}}\right)\right]$$

and (3.5) follows.

In this case of (ii),

$$\liminf_n \beta_{2,n} \ge 1 - \limsup_n \Phi\left(\frac{2pz_\alpha}{n\sigma_{2,n}} - \frac{\mathrm{tr}(\Sigma - I)^2}{\sigma_{2,n}}\right).$$

For $C_1 = 1/\sqrt{8+4\Delta}$ and $C_2 = 2z_\alpha$, it is easy to see $\sigma_{2,n} \ge n^{-1}\,\mathrm{tr}(\Sigma^2) = (p/n)\rho_{1,n}\rho_{2,n}$. Furthermore, using the fact that $\mathrm{tr}(\Sigma^2 - \Sigma)^2 \le \mathrm{tr}(\Sigma^2)\,\mathrm{tr}(\Sigma - I)^2$, $\sigma_{2,n} \le C_1^{-1}\,\mathrm{tr}(\Sigma - I)^2\sqrt{\rho_{2,n} + \rho_{2,n}^2}$. Using the same technique as in the proof of (i), we get (3.6).

Proof of Theorem 4

From Theorem 3, all the conclusions in Theorem 4 follow directly except the one that asserts $\rho_{1,n} \to \infty$ if and only if $\rho_{2,n} \to 0$. We first show the necessity part. Suppose $\rho_{1,n} \to \infty$. Denote $\gamma_n = (1/p)\,\mathrm{tr}(\Sigma^2)$. Then $\rho_{1,n}\rho_{2,n} = \gamma_n$. By the way of contradiction, if $\limsup_n \rho_{2,n} = b > 0$ (b may be ∞), there exists a subsequence $\{n_j\}$ such that $\rho_{2,n_j} \to b$. Thus, $\gamma_{n_j} \to \infty$. On the other hand, using $\mathrm{tr}(\Sigma) \le \sqrt{p\,\mathrm{tr}(\Sigma^2)}$, we have $\rho_{2,n_j}^{-1} \ge n(\gamma_{n_j}^{-1/2} - 1)^2$. This implies $\rho_{2,n_j} \to 0$, contradicting the hypothesis $b > 0$.

We can prove the sufficiency part in the same way. Suppose $\rho_{2,n} \to 0$, but $\liminf \rho_{1,n} \to a < \infty$. Then there exists a subsequence $\{n_j\}$ such that $\rho_{1,n_j} \to a$ and so $\gamma_{n_j} \to 0$. On the other hand, using $\mathrm{tr}(\Sigma) \le \sqrt{p\,\mathrm{tr}(\Sigma^2)}$, we have $\rho_{1,n_j} \ge n(\gamma_{n_j}^{1/2} - 1)^2$. This implies $\rho_{1,n_j} \to \infty$, contradicting the hypothesis $a < \infty$.

[Received September 2009. Revised January 2010.]

REFERENCES

Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis*, Hoboken, NJ: Wiley. [810]

Bai, Z., and Saranadasa, H. (1996), "Effect of High Dimension: By an Example of a Two Sample Problem," *Statistica Sinica*, 6, 311–329. [810-812]

Bai, Z. D., and Silverstein, J. W. (2005), *Spectral Analysis of Large Dimensional Random Matrices*, Beijing: Scientific Press. [810]

Bai, Z. D., and Yin, Y. Q. (1993), "Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix," *The Annals of Probability*, 21, 1276–1294. [810]

Bai, Z. D., Silverstein, J. W., and Yin, Y. Q. (1998), "A Note on the Largest Eigenvalue of a Large-Dimensional Sample Covariance Matrix," *The Journal of Multivariate Analysis*, 26, 166–168. [810]

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. [815]

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, Ser. B*, 57, 289–300. [810]

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. [810]

Billingsley, P. (1995), *Probability and Measure* (3rd ed.), New York: Wiley. [817]

Chen, S. X., and Qin, Y.-L. (2010), "A Two Sample Test for High Dimensional Data With Applications to Gene-Set Testing," *The Annals of Statistics*, 38, 808–835. [810]

Donoho, D., and Jin, J. (2004), "Higher Criticism for Detecting Sparse Heterogeneous Mixtures," *The Annals of Statistics*, 32, 962–994. [810]

Efron, B., and Tibshirani, R. (2007), "On Testing the Significance of Sets of Genes," *The Annals of Applied Statistics*, 1, 107–129. [815]

Fan, J., Hall, P., and Yao, Q. (2007), "How Many Simultaneous Hypothesis Tests Can Normal, Student's *t* or Bootstrap Calibration Be Applied," *Journal of the American Statistical Association*, 102, 1282–1288. [810]

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. [810]

Hall, P., and Jin, J. (2008), "Properties of Higher Criticism Under Long-Range Dependence," *The Annals of Statistics*, 36, 381–402. [810]

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. [810]

Huber, W., Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002), "Variance Stablization Applied to Microarray Data Calibration and to the Quantification of Differential Expression," *Bioinformatics*, 18 (Suppl. 1), S96–S104. [816]

John, S. (1971), "Some Optimal Multivariate Tests," *Biometrika*, 59, 123–127. [810,811,815]

——— (1972), "The Distribution of a Statistic Used for Testing Sphericity of Normal Distributions," *Biometrika*, 59, 169–173. [810,811]

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. [810]

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. [810-814]

Lkhagvadorj, S., Qu, L., Cai, W., Couture, O., Barb, C., Hausman, G., Nettleton, D., Anderson, L., Dekkers, J., and Tuggle, C. (2009), "Microarray Gene Expression Profiles of Fasting Induced Changes in Liver and Adipose Tissues of Pigs Expressing the Melanocortin-4 Receptor D298N Variant," *Physiological Genomics*, 38, 98–111. [815]

Nagao, H. (1973), "On Some Test Criteria for Covariance Matrix," *The Annals of Statistics*, 1, 700–709. [810,811,815]

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. [815]

Schott, J. R. (2007), "Some High Dimensional Tests for a One-Way MANOVA," *The Journal of Multivariate Analysis*, 98, 1825–1839. [810]

van der Laan, M., and Bryan, J. (2001), "Gene Expression Analysis With the Parametric Bootstrap," *Biostatistics*, 2, 445–461. [810]

Zhang, C.-H., and Huang, J. (2008), "The Sparsity and Bias of the LASSO Selection in High-Dimensional Linear Regression," *The Annals of Statistics*, 36, 1567–1594. [810]