On Schott's and Mao's Test Statistics for Independence of Normal Random Vectors

Shuhua Chang[1], **Yongcheng Qi**[2]

[1]Coordinated Innovation Center for Computable Modeling in Management Science, Tianjin University of Finance and Economics, Tianjin 300222, PR China.
Email: szhang@tjufe.edu.cn

[2]Department of Mathematics and Statistics, University of Minnesota Duluth, 1117 University Drive, Duluth, MN 55812, USA.
Email: yqi@d.umn.edu

Abstract. Consider a random sample of n independently and identically distributed p-dimensional normal random vectors. A test statistic for complete independence of high-dimensional normal distributions, proposed by Schott (2005), is defined as the sum of squared Pearson's correlation coefficients. A modified test statistic has been proposed by Mao (2014). Under the assumption of complete independence, both test statistics are asymptotically normal if the limit $\lim_{n\to\infty} p/n$ exists and is finite. In this paper, we investigate the limiting distributions for both Schott's and Mao's test statistics. We show that both test statistics, after suitably normalized, converge in distribution to the standard normal as long as both n and p tend to infinity. Furthermore, we show that the distribution functions of the test statistics can be approximated very well by a chi-square distribution function with $p(p-1)/2$ degrees of freedom as n tends to infinity regardless of how p changes with n.

Keywords: High dimension; complete independence; normal distribution; limiting distribution

1 Introduction

In classical multivariate analysis, statistical methods have been developed mainly for data from designed experiments and dimensions of the data are fixed or very small compared with the sample size. Nowadays, new technology has generated various types of high-dimensional data sets such as financial data, consumer data, modern manufacturing data, multimedia data, hyperspectral image data, internet data, microarray and DNA data. A common feature for all these datasets is that their dimensions can be very large compared with their sample sizes. See, e.g., Schott (2001, 2005, 2007), Ledoit and Wolf (2002), Fan, Peng and Huang (2005), Bai et al. (2009), Chen et al (2010), Chen and Qin (2010), Fujikoshi et al. (2010), Bühlmann and van de Geer (2011), Jiang et al (2012), Srivastava and Reid (2012).

Throughout the paper, $N_p(\mu, \mathbf{\Sigma})$ denotes the p-dimensional normal distribution with mean vector μ and covariance matrix $\mathbf{\Sigma}$, and $\mathbf{I}_p$ denotes the $p \times p$ identity matrix. We assume that $\mathbf{\Sigma}$ is positive definite. Write $\mathbf{\Sigma} = (\sigma(i,j))_{1 \le i,j \le p}$. Then, $\mathbf{\Gamma} = (\rho_{ij})_{1 \le i,j \le p}$ is the correlation matrix of $\mathbf{\Sigma}$ given by $\rho_{ij} = \sigma(i,j)/\sqrt{\sigma(i,i)\sigma(j,j)}$.

Assume that a p-dimensional random vector $\mathbf{x} = (x_1, \cdots, x_p)'$ has a distribution $N_p(\mu, \mathbf{\Sigma})$. We are interested in testing whether the p components $x_1, x_2, \cdots, x_p$ are independent or equivalently testing whether the covariance matrix $\mathbf{\Sigma}$ is diagonal. Then, the test can be written as

$$H_0 : \mathbf{\Gamma} = \mathbf{I}_p \quad \text{vs} \quad H_a : \mathbf{\Gamma} \neq \mathbf{I}_p. \tag{1.1}$$

In literature, (1.1) is known as the test of complete independence.

Let $\mathbf{x}_1, \cdots, \mathbf{x}_n$ be i.i.d. from $N_p(\mu, \mathbf{\Sigma})$. Write

$$\mathbf{x}_k = (x_{k1}, \cdots, x_{kp})', \qquad k = 1, \cdots, n.$$

Define

$$r_{ij} = \frac{\sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ki} - \bar{x}_i)^2 \cdot \sum_{k=1}^n (x_{kj} - \bar{x}_j)^2}}, \tag{1.2}$$

where $\bar{x}_i = \frac{1}{n}\sum_{k=1}^n x_{ki}$ and $\bar{x}_j = \frac{1}{n}\sum_{k=1}^n x_{kj}$. Then, $\mathbf{R}_n := (r_{ij})_{p \times p}$ is the sample correlation matrix based on the p-dimensional random vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n$.

In classic multivariate analysis when p is a fixed integer, the likelihood method is a nice approach to test (1.1). From Bartlett (1954) or Morrison (2005), the likelihood ratio test statistic is a function of the determinant of $\mathbf{R}_n$. When $p = p_n$ depends on n and $p_n \to \infty$, the limiting distribution of the likelihood ratio test statistic has been obtained in Jiang and Yang (2013), Jiang, Bai and Zheng (2013) and Jiang and Qi (2015), and the likelihood ratio method can still be used to test (1.1). However, the likelihood ratio method fails when $p \geq n$, since the sample correlation matrix $\mathbf{R}_n$ is singular and the corresponding test statistic is degenerate. A natural requirement for non-singularity of $\mathbf{R}_n$ is $p < n$.

Schott (2005) considers the following test statistic

$$t_{np} = \sum_{1\leq j<i\leq p} r_{ij}^2.$$

Assume that the null hypothesis of (1.1) holds and $\lim_{n\to\infty} p/n = \gamma \in (0,\infty)$. Schott (2005) proves that $t_{np} - \frac{p(p-1)}{2(n-1)}$ converges in distribution to a normal distribution with mean 0 and variance γ^2, that is,

$$t_{np}^* := \frac{t_{np} - \frac{p(p-1)}{2(n-1)}}{\tau_{np}} \xrightarrow{d} N(0,1), \tag{1.3}$$

where $\tau_{np}^2 = \frac{p(p-1)(n-2)}{(n-1)^2(n+1)}$.

It is worth noting that the same test statistic t_{np} is also proposed by Srivastava (2005). Srivastava (2005, 2006) also considers a test statistic which is based on the Fisher's z-transformation and originally proposed by Chen and Mudholkar (1990):

$$Q_{np} = \frac{(n-3)\sum_{1\leq j<i\leq p} z_{ij}^2 - \frac{1}{2}p(p-1)}{\sqrt{p(p-1)}},$$

where $z_{ij} = \frac{1}{2}\log\frac{1+r_{ij}}{1-r_{ij}}$. From Srivastava (2005), such a test has not been designed for large p. Instead, Srivastava (2005) proposes a test statistic T_3 which is related to the sample covariances only. See Srivastava (2005, 2006) for details. Under certain conditions, Srivastava (2005) shows that T_3 converges in distribution to the standard normal under the null hypothesis in (1.1). A simulation study in Srivastava (2006) indicates that Q_{np} statistic is inferior as the test does not give a consistent nominal level when n and p are close.

Very recently, Mao (2014) proposes a new test for complete independence. The new test statistic is closely related to Schott's test and is defined by

$$T_{np} = \sum_{1\leq j<i\leq p} \frac{r_{ij}^2}{1-r_{ij}^2}.$$

It has been proved in Mao (2014) that T_{np} is asymptotically normal under the null hypothesis of (1.1) and assumption that $\lim_{n\to\infty} p/n = \gamma \in (0,\infty)$.

In this paper, we will remove the condition imposed on p and assume only that $p = p_n \to \infty$ as $n \to \infty$. We will show that both T_{np} and t_{np} are asymptotically normal. We also establish a unified chi-square approximation for the distributions of T_{np} and t_{np} regardless of how p changes with n.

The rest of the paper is organized as follows. The main results of the paper are given in section 2 and their proofs are postponed until section 4. A simulation study to compare the performance of several different approaches is reported in section 3.

2 Main Results

Our main results include three theorems. We first obtain the limiting distribution of the test statistic T_{np} in a larger range for p, and then establish a unified chi-square approximation for all $p \ge 2$. The corresponding limiting distributions of t_{np} are given in the third theorem.

The first theorem states that Mao's (2014) test statistic T_{np} is asymptotically normal as long as $p = p_n \to \infty$ as $n \to \infty$.

Theorem 2.1. *Assume $p = p_n \to \infty$ as $n \to \infty$. Then, under the null hypothesis of* (1.1)

$$T_{np}^* := \frac{T_{np} - \frac{p(p-1)}{2(n-4)}}{\sigma_{np}} \xrightarrow{d} N(0,1) \tag{2.1}$$

as $n \to \infty$, where

$$\sigma_{np}^2 = \frac{p(p-1)(n-3)}{(n-4)^2(n-6)}.$$

We expect that the limiting distribution of $(n-4)T_{np}$ is chi-squared with $p(p-1)/2$ degrees of freedom when p is fixed, and this will be confirmed in the following theorem. For applications there seems a gap in the limiting distributions of the test statistic T_{np} as one has to distinguish whether $p = p_n$ converges or diverges. Instead, under linear transformation we define a slightly different statistic as follows

$$T_{np}^c = \sqrt{p(p-1)}T_{np}^* + \frac{1}{2}p(p-1) = \sqrt{\frac{n-6}{n-3}}(n-4)T_{np} + \frac{1}{2}p(p-1)\left(1 - \sqrt{\frac{n-6}{n-3}}\right). \tag{2.2}$$

The statistic T_{np}^c can fill this gap. Our second theorem reveals that the chi-square distribution can be used to approach the distribution of T_{np}^c no matter how $p = p_n$ changes with n.

Theorem 2.2. *Let $p = p_n$, $n \geq 1$ be a sequence of positive integers with $p_n \geq 2$ for all large n. Then, under the null hypothesis of* (1.1) *we have*

$$\sup_x |P(T_{np}^c \leq x) - P(\chi^2_{p(p-1)/2} \leq x)| \to 0 \qquad \text{as } n \to \infty. \tag{2.3}$$

Theorem 2.2 implies that T_{np}^c converges in distribution to a chi-square distribution with $p(p-1)/2$ degrees of freedom uniformly over $p \geq 2$ as $n \to \infty$, that is, the superium of the left-hand side of (2.3) over $p \geq 2$ converges to zero as $n \to \infty$.

In Theorem 2.3 below, we extend Schott's statistic t_{np} in the same manner. We will show that the central limit theorem (1.3) holds for all large p and a chi-square approximation can also be applied to t_{np} for small p. Now we define

$$t_{np}^c = \sqrt{p(p-1)}t_{np}^* + \frac{1}{2}p(p-1) = \sqrt{\frac{n+1}{n-2}}(n-1)t_{np} + \frac{1}{2}p(p-1)(1 - \sqrt{\frac{n+1}{n-2}}). \tag{2.4}$$

Theorem 2.3. *(i) If $p = p_n \to \infty$ as $n \to \infty$, then* (1.3) *holds under the null hypothesis of* (1.1).
(ii) Let $p = p_n$ be a sequence of positive integers with $p_n \geq 2$ for all large n. Then, under the null hypothesis of (1.1) *we have*

$$\sup_x |P(t_{np}^c \leq x) - P(\chi^2_{p(p-1)/2} \leq x)| \to 0 \qquad \text{as } n \to \infty. \tag{2.5}$$

Assume $\alpha \in (0,1)$. Let z_α and $\chi^2_\alpha(p(p-1)/2)$ denote the α level critical values for the standard normal distribution and the chi-squared distribution with $p(p-1)/2$ degrees of freedom, respectively.

Based on (2.1), an approximate level α test for (1.1) has a critical region or rejection region

$$\mathcal{R}_T^*(\alpha) = \left\{ T_{np} \geq \frac{p(p-1)}{2(n-4)} + z_\alpha \sqrt{\frac{p(p-1)(n-3)}{(n-4)^2(n-6)}} \right\}. \tag{2.6}$$

Based on the chi-square approximation (2.3), an approximate level α test rejects (1.1) in the region

$$\mathcal{R}_T^c(\alpha) = \left\{ T_{np} \geq \frac{p(p-1)}{2(n-4)} \left(\sqrt{\frac{n-3}{n-6}} - 1 \right) + \chi^2_\alpha(p(p-1)/2) \sqrt{\frac{(n-3)}{(n-4)^2(n-6)}} \right\}. \tag{2.7}$$

While based on the normal approximation (1.3) to Schott's test statistic t_{np}, an approximate level α test for (1.1) has a rejection region

$$\mathcal{R}_t^*(\alpha) = \left\{ t_{np} \geq \frac{p(p-1)}{2(n-1)} + z_\alpha \sqrt{\frac{p(p-1)(n-2)}{(n-1)^2(n+1)}} \right\}. \tag{2.8}$$

Similarly, we have an approximate level α rejection region for test (1.1)

$$\mathcal{R}_t^c(\alpha) = \left\{ t_{np} \geq \frac{p(p-1)}{2(n-1)}(1 - \sqrt{\frac{n-2}{n+1}}) + \chi^2_\alpha(p(p-1)/2)\sqrt{\frac{(n-2)}{(n-1)^2(n+1)}} \right\} \tag{2.9}$$

based on the chi-square approximation (2.5).

3 Simulation Study

Mao (2014) has conducted a simulation study and compared the performance of three test statistics including Mao's T_{np}, Schott's t_{np} and Srivastava's T_3. It has been reported in Mao (2014) that Mao's test statistic is comparable to the other two test statistics in terms of the accuracy of sizes of the tests and outperforms in some models under weak dependence.

In this section we will carry out a finite-sample simulation study to compare the performance of Schott's (2005) t_{np} and Mao's (2014) T_{np} based on the normal approximation and the chi-square approximation. We will not simply repeat Mao's (2014) choices. Our focus is on the two test statistics t_{np} and T_{np} which are related to the sample correlations. More specifically, we consider four normalized test statistics: t^*_{np}, T^*_{np}, t^c_{np} and T^c_{np}. Their limiting distributions are determined by (1.3), (2.1), (2.5) and (2.3), respectively, and the corresponding rejection regions for the four tests at level α are given by (2.8), (2.6), (2.9) and (2.7).

Let $\mathbf{\Sigma}_p^{(\rho)}$ denote a $p \times p$ matrix whose diagonal entries are equal to 1 and all off-diagonal entries are equal to ρ, where $\rho \in (-1, 1)$. $\mathbf{\Sigma}_p^{(\rho)}$ is the covariance matrix of a normal random vector with all p components being standard normal random variables and covariances (and correlation coefficients) equal to ρ. A random sample of size n is drawn from multivariate normal distribution $N_p(0, \mathbf{\Sigma}_p^{(\rho)})$ with the different choices for $n = 7$, 15, 30, 60, 100, 200, $p = 3$, 10, 20, 50, 100, 200, and $\rho = 0$, 0.02. For each combination of the choices on n, p and ρ, the simulation experiment is repeated 10000 times so that the sizes and powers of the tests can be estimated very accurately. The type I error $\alpha = 0.05$ is fixed in our simulation study.

When $\rho = 0$, the null hypothesis in (1.1) is true. The estimated sizes for these test statistics are reported in Table 1. When $\rho = 0.02$, the alternative hypothesis in (1.1) is true, and this indicates a weak dependence among the coordinates of a normal random vector. The estimated powers for these test statistics are given in Table 2.

We notice that the rejection rates based on t^*_{np} are always larger than those based on t^c_{np}. This is due to the fact that both test statistics are linear functions of t_{np}, and the rejection regions $\mathcal{R}^c_t$ based on t^c_{np} are contained in the rejection regions $\mathcal{R}^*_t$ based on t^*_{np}. See equations (2.9) and (2.8). Theoretically, we can show that the cutoff values in (2.8) are smaller than those in (2.9) for all large p. The same relationship is true between T^*_{np} and T^c_{np}.

The simulation study when $n = 7$ is suggested by a referee so as to compare the performance of these test statistics in small-n case when p is large. Note that the variance for Mao's statistic T_{np} is infinite when $n \leq 6$ and the normalized test statistic T^*_{np} is not well defined in this case. We observe that the estimated type I errors for Mao's test statistic T^*_{np} are quite close to the nominal level 0.05 when p is large. For Schott's test statistic t^*_{np}, the estimated type I errors are slightly higher than 0.05 but quite stable over p. It would be very useful to investigate the limiting distributions for these test statistics for small or fixed n when p goes to infinity.

The rest of our discussion below will be based on the simulation results in Tables 1 and 2 when $n \geq 15$.

In terms of the estimated size, a test is considered to be preferable if its estimated size is close to the nominal level $\alpha = 0.05$. Table 1 indicates that t^*_{np} and T^*_{np} are comparable in terms of the estimated size for tests and the normal approximation yields significantly larger sizes than the nominal level for both t^*_{np} and T^*_{np} when the dimension p is relatively small. The test statistics t^c_{np} and T^c_{np} have much better performance than their competitors t^*_{np} and T^*_{np} when p is small as the chi-square approximation is used to determine the corresponding rejection regions. When p is large, the four test statistics are comparable.

The estimated powers for the four test statistics are recorded in Table 2. From the table, both t^*_{np} and T^*_{np} have slightly larger powers than t^c_{np} and T^c_{np} for small p. This is not surprising since the normal approximation to t_{np} and T_{np} sacrifices the accuracy in the

Table 1: Size of tests ($\rho = 0$)

Test Statistic	$n\backslash p$	3	10	20	50	100	200
t^*_{np}	7	0.0658	0.0621	0.0621	0.0636	0.0621	0.0606
	15	0.0718	0.0583	0.0614	0.0563	0.0574	0.0576
	30	0.0725	0.0611	0.0598	0.0510	0.0560	0.0550
	60	0.0711	0.0593	0.0559	0.0544	0.0519	0.0580
	100	0.0738	0.0611	0.0587	0.0539	0.0542	0.0493
	200	0.0712	0.0606	0.0551	0.0554	0.0506	0.0497
T^*_{np}	7	0.0328	0.0427	0.0462	0.0552	0.0553	0.0538
	15	0.0633	0.0619	0.0605	0.0554	0.0573	0.0578
	30	0.0696	0.0600	0.0599	0.0522	0.0553	0.0541
	60	0.0700	0.0617	0.0568	0.0556	0.0519	0.0575
	100	0.0726	0.0623	0.0597	0.0537	0.0551	0.0490
	200	0.0713	0.0606	0.0561	0.0555	0.0505	0.0494
t^c_{np}	7	0.0442	0.0524	0.0577	0.0623	0.0609	0.0604
	15	0.0478	0.0479	0.0556	0.0537	0.0566	0.0574
	30	0.0497	0.0509	0.0535	0.0490	0.0550	0.0546
	60	0.0501	0.0512	0.0505	0.0516	0.0507	0.0577
	100	0.0526	0.0512	0.0545	0.0525	0.0530	0.0490
	200	0.0514	0.0512	0.0494	0.0531	0.0493	0.0486
T^c_{np}	7	0.0242	0.0371	0.0427	0.0531	0.0544	0.0529
	15	0.0466	0.0539	0.0562	0.0536	0.0558	0.0577
	30	0.0485	0.0509	0.0544	0.0508	0.0546	0.0537
	60	0.0503	0.0515	0.0505	0.0537	0.0510	0.0572
	100	0.0520	0.0524	0.0550	0.0521	0.0535	0.0487
	200	0.0519	0.0512	0.0495	0.0531	0.0493	0.0490

size of the tests when p is small. The performances of the four test statistics are similar when p is large.

In summary, we can conclude that the test statistics t^c_{np} and T^c_{np} are consistently accurate in terms of the size over the whole range of p and achieve satisfactory power compared with Schott's t_{np} and Mao's T_{np} . Our simulation study suggests that the normal

Table 2: Power of tests: $\rho = 0.02$

Test Statistic	$n\backslash p$	3	10	20	50	100	200
t^*_{np}	7	0.0646	0.0657	0.0684	0.0743	0.0779	0.0976
	15	0.0725	0.0656	0.0647	0.0765	0.1003	0.1557
	30	0.0717	0.0693	0.0757	0.1002	0.1598	0.3130
	60	0.0811	0.0805	0.0932	0.1667	0.3206	0.6505
	100	0.0812	0.0902	0.1297	0.2651	0.5714	0.9176
	200	0.0901	0.1255	0.2175	0.5834	0.9413	0.9996
T^*_{np}	7	0.0323	0.0461	0.0519	0.0591	0.0609	0.0738
	15	0.0641	0.0673	0.0661	0.0744	0.1017	0.1505
	30	0.0689	0.0715	0.0756	0.0984	0.1583	0.3096
	60	0.0788	0.0820	0.0923	0.1667	0.3199	0.6511
	100	0.0793	0.0909	0.1303	0.2646	0.5706	0.9173
	200	0.0894	0.1258	0.2183	0.5838	0.9415	0.9996
t^c_{np}	7	0.0460	0.0535	0.0618	0.0722	0.0765	0.0971
	15	0.0494	0.0556	0.0579	0.0742	0.0987	0.1539
	30	0.0513	0.0581	0.0685	0.0977	0.1574	0.3116
	60	0.0594	0.0677	0.0843	0.1610	0.3182	0.6492
	100	0.0578	0.0777	0.1212	0.2594	0.5674	0.9166
	200	0.0649	0.1085	0.2045	0.5763	0.9398	0.9996
T^c_{np}	7	0.0238	0.0397	0.0478	0.0569	0.0601	0.0727
	15	0.0466	0.0577	0.0602	0.0720	0.1005	0.1486
	30	0.0506	0.0604	0.0686	0.0954	0.1566	0.3083
	60	0.0594	0.0703	0.0853	0.1621	0.3172	0.6499
	100	0.0578	0.0778	0.1211	0.2589	0.5672	0.9164
	200	0.0647	0.1076	0.2041	0.5754	0.9400	0.9996

approximation to Schott's t_{np} and Mao's T_{np} are inferior to the chi-square approximation to t^c_{np} and T^c_{np} when p is small. When p is large, the four test statistics under consideration are quite similar in terms of powers and accuracy in the size. Our simulation also confirms the theoretical consistency in using the normal approximation to both t_{np} and T_{np} under the complete independence when $p_n \to \infty$ as $n \to \infty$ regardless of how fast p_n increases with n.

4 Proofs

Proof of Theorem 2.1.

We will employ a martingale central limit theorem in McLeish (1974). Since some details are somewhat similar to those in Mao (2014), we outline our proof as follows.

Step 1. Express r_{ij} as $r_{ij} = w_i' w_j$, where $w_1, w_2, \cdots, w_{p_n}$ are independent random vectors that are uniformly distributed on the surface of the $(n-1)$-sphere. Let $\mathcal{F}_{n\ell} = \sigma(w_1, w_2, \cdots, w_\ell)$ denote the σ-algebra generated by $\{w_1, w_2, \cdots, w_\ell\}$, see Mao (2014).

Step 2. For $2 \le \ell \le p_n$ set $y_{n\ell} = \sigma_{np_n}^{-1} \sum_{j=1}^{\ell-1} \hat{r}_{\ell j}$, where $\hat{r}_{\ell j} = \frac{r_{\ell j}^2}{1-r_{\ell j}^2} - \frac{1}{n-4}$. Then, $\{y_{n\ell},\ \mathcal{F}_{n\ell},\ 2 \le \ell \le p_n,\ n \ge 6\}$ form an array of martingale differences, see Mao (2014). Note that $T_{np}^* = \sum_{\ell=2}^{p_n} y_{n\ell}$. According to Theorem 2.3 in McLeish (1974), to show (2.1), it suffices to prove the following three conditions:

(a) $\sup\limits_{n \ge n_0} E(\max\limits_{2\le\ell\le p_n} (y_{n\ell})^2) < \infty$ for some n_0;

(b) $\max\limits_{2\le\ell\le p_n} |y_{n\ell}|$ converges to zero in probability;

(c) $\sum\limits_{\ell=2}^{p_n} y_{n\ell}^2$ converges to one in probability.

To verify the above three conditions, we need to show that

$$\sum_{\ell=2}^{p_n} E(y_{n\ell}^4) \to 0 \quad \text{and} \quad E(\sum_{\ell=2}^{p_n} y_{n\ell}^2 - 1)^2 \to 0 \tag{4.1}$$

as $n \to \infty$. The second limit implies condition (c) immediately. The first limit implies condition (a), since

$$E(\max_{2\le\ell\le p_n} (y_{n\ell})^2) \le \sqrt{E(\max_{2\le\ell\le p_n} y_{n\ell}^4)} \le \sqrt{\sum_{\ell=2}^{p_n} E(y_{n\ell}^4)} \to 0.$$

Condition (b) follows from the above equation by using the Markov inequality.

It has been proved in Mao (2014) that

$$E(\prod_{i=1}^4 \hat{r}_{\ell j_i}) = \begin{cases} \frac{12(n-3)(5n^2-27n+40)}{(n-4)^2(n-6)(n-8)(n-10)}, & \text{if } j_1 = j_2 = j_3 = j_4; \\ \frac{4(n-3)^2}{(n-4)^2(n-6)^2}, & \text{if } \{j_1, j_2, j_3, j_4\} \text{ forms two matching pairs;} \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

Note that $\sigma_{np_n} \sim \frac{p_n}{n}$ as $n \to \infty$. Then, we have

$$E(y_{n\ell}^4) = \sigma_{np_n}^{-4} \sum_{1 \le j_1, j_2, j_3, j_4 \le \ell-1} E(\hat{r}_{\ell j_1} \hat{r}_{\ell j_2} \hat{r}_{\ell j_3} \hat{r}_{\ell j_4}) = \sigma_{np_n}^{-4} O\left(\frac{\ell^2}{n^4}\right) = O\left(\frac{\ell^2}{p_n^4}\right)$$

uniformly over $2 \le \ell \le p_n$ as $n \to \infty$. Therefore, $\sum_{\ell=2}^{p_n} E(y_{n\ell}^4) = O(1/p) \to 0$ as $n \to \infty$. This proves the first limit in (4.1). Mao (2014) has shown that $E(\sum_{\ell=2}^{p} y_{n\ell}^2) = 1$ and $\sum_{2 \le i \ne j \le p_n} E(y_{ni}^2 y_{nj}^2) - 1 = -\frac{2\sigma_{np}^{-4}(n-3)^2 p(p-1)(2p-1)}{3(n-4)^4(n-6)^2}$ which is of order p_n^{-1}. Therefore, we have as $n \to \infty$ that

$$E(\sum_{\ell=2}^{p_n} y_{n\ell}^2 - 1)^2 = E(\sum_{\ell=2}^{p_n} y_{n\ell}^4) + \sum_{2 \le i \ne j \le p_n} E(y_{ni}^2 y_{nj}^2) - 1 = O(p_n^{-1}) \to 0,$$

which yields the second limit in (4.1). This completes the proof of the theorem. ■

Proof of Theorem 2.2.

To prove (2.3), it suffices to show that for every sequence of integers $\{n_i,\ \ i \ge 1\}$, there exists its subsequence $\{n_{i(j)},\ j \ge 1\}$ such that (2.3) holds along $\{n_{i(j)}\}$. Here we choose the subsequence so that $p_{n_{i(j)}}$ converges as $j \to \infty$. Since $p_{n_{i(j)}}$'s are integers, the limit of $n_{i(j)}$ is a finite integer p or infinity. Therefore, we need to show that (2.3) holds along any subsequence of integers n_i such that p_{n_i} is a fixed integer p for all large i or $p_{n_i} \to \infty$ as $i \to \infty$. Since the proof of (2.3) along a subsequence is the same as that along the entire sequence, for simplicity, we will show (2.3) under the following conditions:

$$p_n = p \ge 2 \text{ is a fixed integer for all large } n; \tag{4.3}$$

$$p_n \to \infty \quad \text{as } n \to \infty. \tag{4.4}$$

First, we will show under (1.1) and (4.3) that

$$(n-4)T_{np} \xrightarrow{d} \chi^2_{p(p-1)/2} \qquad \text{as } n \to \infty,$$

which implies (2.3) since $T_{np}^c = (1 + o(1))(n-4)T_{np} + o(1)$.

Express $w_j = z_j/(z_j' z_j)^{1/2}$ for $1 \le j \le p$, where $z_j = (z_{j1}, \cdots, z_{j(n-1)})'$, $1 \le j \le p$ are i.i.d. random vectors with $N_{n-1}(0, \mathbf{I}_{n-1})$ distribution. Write $s_{i,j} = z_i' z_j = \sum_{k=1}^{n-1} z_{ik} z_{jk}$. By using the multivariate central limit theorem,

$$\frac{1}{\sqrt{n-1}}(s_{2,1}, s_{3,1}, s_{3,2}, \cdots, s_{p,1}, \cdots, s_{p,(p-1)})' \xrightarrow{d} N_{p(p-1)/2}(0, \mathbf{I}_{p(p-1)/2}) \tag{4.5}$$

as $n \to \infty$, which implies that $\frac{1}{n-1}(s_{2,1}^2, s_{3,1}^2, s_{3,2}^2, \cdots, s_{p,1}^2, \cdots, s_{p,(p-1)}^2)'$ converges in distribution to a random vector whose $p(p-1)/2$ components are independent random variables having a chi-squared distribution with 1 degree of freedom. By the law of large numbers, $\frac{z_i' z_i'}{n-1} = 1 + o_p(1)$ for $i = 1, \cdots, p$, which implies

$$\max_{1 \le i \le p} \left| \frac{z_i' z_i}{n-1} - 1 \right| = o_p(1) \quad \text{ as } n \to \infty.$$

Therefore, we have

$$r_{ij}^2 = s_{i,j}^2 / ((z_i' z_i)(z_j' z_j)) = \frac{s_{i,j}^2}{(n-1)^2}(1 + o_p(1)), \tag{4.6}$$

which implies

$$\frac{(n-4) r_{ij}^2}{1 - r_{ij}^2} = \frac{s_{i,j}^2}{n-1}(1 + o_p(1)),$$

and consequently,

$$(n-4) T_{np} = \frac{\sum\limits_{1 \le j < i \le p} s_{ij}^2}{n-1}(1 + o_p(1)) \xrightarrow{d} \chi_{p(p-1)/2}^2 \quad \text{ as } n \to \infty.$$

Now assume (4.4) and the null hypothesis in (1.1) hold. In this case, we can apply Theorem 2.1 directly. It follows from (2.2) and (2.1) that

$$\frac{T_{np}^c - \frac{p(p-1)}{2}}{\sqrt{p(p-1)}} = T_{np}^* \xrightarrow{d} N(0,1),$$

which implies that

$$\sup_x \left| P\left(\frac{T_{np}^c - \frac{p(p-1)}{2}}{\sqrt{p(p-1)}} \le x \right) - \Phi(x) \right| \to 0 \quad \text{ as } n \to \infty, \tag{4.7}$$

where $\Phi(x)$ is the standard normal cumulative distribution function. Also, notice that a chi-squared random variable with $p(p-1)/2$ degrees of freedom can be written as the sum of $p(p-1)/2$ independent and identically distributed random variables having a chi-squared distribution with 1 degree of freedom. From the classic central limit theorem, we have

$$\frac{\chi_{p(p-1)/2}^2 - \frac{p(p-1)}{2}}{\sqrt{p(p-1)}} \xrightarrow{d} N(0,1),$$

and thus

$$\sup_x \left| P\left(\frac{\chi^2_{p(p-1)/2} - \frac{p(p-1)}{2}}{\sqrt{p(p-1)}} \leq x \right) - \Phi(x) \right| \to 0. \tag{4.8}$$

Therefore, by combining (4.7) and (4.8) and using the triangle inequality we have

$$\begin{aligned}
& \sup_x |P(T^c_{np} \leq x) - P(\chi^2_{p(p-1)/2} \leq x)| \\
= \ & \sup_x \left| P\left(\frac{T^c_{np} - \frac{p(p-1)}{2}}{\sqrt{p(p-1)}} \leq x \right) - P\left(\frac{\chi^2_{p(p-1)/2} - \frac{p(p-1)}{2}}{\sqrt{p(p-1)}} \leq x \right) \right| \\
\leq \ & \sup_x \left| P\left(\frac{T^c_{np} - \frac{p(p-1)}{2}}{\sqrt{p(p-1)}} \leq x \right) - \Phi(x) \right| + \sup_x \left| P\left(\frac{\chi^2_{p(p-1)/2} - \frac{p(p-1)}{2}}{\sqrt{p(p-1)}} \leq x \right) - \Phi(x) \right| \\
\to \ & 0
\end{aligned}$$

as $n \to \infty$. This completes the proof of (2.3). ■

Proof of Theorem 2.3.

We will sketch the proof. We continue to use the notation in the proof of Theorem 2.2. As in the proof of Theorem 2.1, write

$$r_{ij} = w_i' w_j, \quad 1 \leq i, j \leq p. \tag{4.9}$$

(i) First, we need to show (1.3), i.e.,

$$t^*_{np_n} = \frac{\sum_{i=2}^{p_n-1} \sum_{j=1}^{i-1} r^2_{ij} - \frac{p_n(p_n-1)}{2(n-1)}}{\tau_{np_n}} \xrightarrow{d} N(0,1), \tag{4.10}$$

under the assumption that $p_n \to \infty$ as $n \to \infty$.

Set

$$z_{n\ell} = \sum_{i=1}^{\ell-1} r^2_{\ell i} - \frac{\ell - 1}{n-1}.$$

Then, $\{z_{n\ell},\ \mathcal{F}_{n\ell},\ 2 \leq \ell \leq p_n,\ n \geq 1\}$ form an array of martingale differences. See, e.g., Schott (2005). It suffices to show that

$$\frac{\sum_{\ell=2}^{p_n} z_{n\ell}}{\tau_{np_n}} \xrightarrow{d} N(0,1). \tag{4.11}$$

We will use martingale approach like that in Schott (2005). In view of Corollary 3.1 in Hall and Heyde (1980), the martingale central limit theorem (4.11) holds if the following two conditions hold:

$$\frac{1}{\tau_{np_n}^2}\sum_{\ell=2}^{p_n} E(z_{n\ell}^2 I(|z_{n\ell}| \geq \varepsilon\tau_{np_n})|\mathcal{F}_{n(\ell-1)}) \to 0 \quad \text{in probability} \tag{4.12}$$

for every $\varepsilon > 0$, and

$$\frac{1}{\tau_{np_n}^2}\sum_{\ell=2}^{p_n} E(z_{n\ell}^2|\mathcal{F}_{n(\ell-1)}) \to 1 \quad \text{in probability.} \tag{4.13}$$

It has been shown in Schott (2005), pp. 955 that

$$E\left(\sum_{\ell=2}^{p_n} E(z_{n\ell}^2|\mathcal{F}_{n(\ell-1)})\right) = \tau_{np_n}^2.$$

Thus, we have

$$\begin{aligned}
\Delta_n : &= \sum_{\ell=2}^{p_n} E(z_{n\ell}^2|\mathcal{F}_{n,\ell-1}) - \tau_{np}^2 \\
&= \frac{2}{(n-1)(n+1)}\sum_{\ell=2}^{p_n}\sum_{i=1}^{\ell-1}\sum_{j=1,j\neq i}^{\ell-1}\left(r_{ij}^2 - \frac{1}{n-1}\right) \\
&= \frac{2}{(n-1)(n+1)}\sum_{\ell=3}^{p_n}\sum_{1\leq i\neq j\leq \ell-1}\left(r_{ij}^2 - \frac{1}{n-1}\right) \\
&= \frac{4}{(n-1)(n+1)}\sum_{\ell=3}^{p_n}\sum_{1\leq j<i\leq \ell-1}\left(r_{ij}^2 - \frac{1}{n-1}\right) \\
&= \frac{4}{(n-1)(n+1)}\sum_{1\leq j<i\leq p_n-1}(p_n - i)\left(r_{ij}^2 - \frac{1}{n-1}\right).
\end{aligned}$$

For $1 \leq j < i \leq p_n$ and $1 \leq t < s \leq p_n$, we can verify from Schott (2005) that

$$E\left(r_{ij}^2 - \frac{1}{n-1}\right)\left(r_{st}^2 - \frac{1}{n-1}\right) = \begin{cases} 0, & \text{if } (i,j) \neq (s,t), \\ \frac{3}{(n-1)(n+1)} - \frac{1}{(n-1)^2} = \frac{2n-4}{(n-1)^2(n+1)}, & \text{if } (i,j) = (s,t). \end{cases}$$

Then, we have

$$\begin{aligned}
E(\Delta_n^2) &= \frac{16}{(n-1)^2(n+1)}\sum_{1\leq j<i\leq p_n-1}(p_n-i)^2\frac{2(n-2)}{(n-1)^2(n+1)} \\
&= \frac{32(n-2)}{(n-1)^4(n+1)^3}\sum_{1<i\leq p_n-1}(p_n-i)^2(i-1) \\
&= O\left(\frac{p_n^4}{n^6}\right),
\end{aligned}$$

which implies that

$$\frac{E(\Delta_n^2)}{\tau_{np_n}^4} = O\left(\frac{1}{n^2}\right) \to 0 \text{ as } n \to \infty. \tag{4.14}$$

Next, we verify that

$$\frac{1}{\tau_{np_n}^4} \sum_{\ell=2}^{p_n} E(z_{n\ell}^4) = o(1) \quad \text{as } n \to \infty. \tag{4.15}$$

Set $q_{\ell i} = r_{\ell i}^2 - \frac{1}{n-1}$. Then

$$z_{n\ell} = \sum_{i=1}^{\ell-1} q_{\ell i}.$$

Note that $r_{\ell i} = w_\ell' w_i$. Conditional on w_ℓ, $r_{\ell 1}, \cdots, r_{\ell(\ell-1)}$ are i.i.d. Set $c_r = E(r_{\ell 1}^{2r} | w_\ell)$, $1 \le r \le 4$. Then

$$c_1 = \frac{1}{n-1},\ c_2 = \frac{3}{(n-1)(n+1)},\ c_3 = \frac{15}{(n-1)(n+1)(n+3)},$$

and

$$c_4 = \frac{105}{(n-1)(n+1)(n+3)(n+5)}.$$

Set $d_r = E(q_{\ell 1}^r | w_\ell) = E((r_{\ell 1}^2 - \frac{1}{n-1})^r | w_\ell)$. Then

$$d_1 = 0, \qquad d_2 = c_2 - \left(\frac{1}{n-1}\right)^2 = \frac{2(n-2)}{(n-1)^2(n+1)},$$

$$d_3 = c_3 - 3c_2\frac{1}{n-1} + 3c_1\frac{1}{(n-1)^2} - \frac{1}{(n-1)^3} = O\left(\frac{1}{n^3}\right),$$

and

$$d_4 = c_4 - 4c_3\frac{1}{n-1} + 6c_2\frac{1}{(n-1)^2} - 4c_1\frac{1}{(n-1)^4} + \frac{1}{(n-1)^4} = O\left(\frac{1}{n^4}\right).$$

Since

$$E(z_{n\ell}^4) = E(E(z_{n\ell}^4 | w_\ell)) = E(E((\sum_{i=1}^{\ell-1} q_{\ell i})^4 | w_\ell)) = (\ell-1)d_4 + 6(\ell-1)(\ell-2)d_2^2$$

for $2 \le \ell \le p$, we obtain

$$\sum_{\ell=2}^{p_n} E(z_{n\ell}^4) = O\left(\frac{p_n^3}{n^4}\right).$$

Then, it follows that

$$\frac{1}{\tau_{np_n}^4}\sum_{\ell=2}^{p_n} E(z_{n\ell}^4) = O\left(\frac{1}{p_n}\right) \to 0 \quad \text{as } n \to \infty,$$

which implies (4.15). (4.12) and (4.13) can be easily verified from (4.14) and (4.15). Therefore, we obtain (4.11).

(ii) For the proof of (2.5) we can use the arguments in the proof of Theorem 2.2. First, under assumption (4.3), we have from (4.5) and (4.6) that $(n-1)t_{np} \overset{d}{\to} \chi^2_{p(p-1)/2}$ as $n \to \infty$. The rest of the proof follows exactly from the same lines in the proof of Theorem 2.2 by using (1.3). The details are omitted. ■

Acknowledgements: We would like to thank the reviewer for his/her constructive suggestions that have led to improvement in the paper. Chang's research was supported in part by the Major Research Plan of the National Natural Science Foundation of China (91430108), the National Basic Research Program (2012CB955804), the National Natural Science Foundation of China (11171251), and the Major Program of Tianjin University of Finance and Economics (ZD1302).

References

[1] Bai, Z., Jiang, D., Yao, J. and Zheng, S. (2009). Corrections to LRT on large dimensional covariance matrix by RMT. *J. Royal Stat. Soc.*, Ser. B 16, 296-298.

[2] Bartlett, M. S. (1954). A note on multiplying factors for various chi-squared approximations. *J. Royal Stat. Soc.*, Ser. B 16, 296-298.

[3] Bühlmann, P., van de Geer, S. (2011). Statistics For High-dimensional Data: Methods, Theory and Applications. Springer, Heidelberg, New York.

[4] Chen, S. and Mudholkar, G. S. (1990). Null distribution of the sum of squared z-transformations in testing complete independence. *Ann. Inst. Statist. Math.* 42, 149-155.

[5] Chen, S. X. and Qin, Y. L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* 38, 808-835.

[6] Chen, S. X., Zhang, L. and Zhong, P. (2010). Tests for high dimensional covariance matrices. *J. Amer. Stat. Assoc.* 105, 810-819.

[7] Fan, J. Q., Peng, H., Huang, T. (2005). Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency. *J. Amer. Statist. Assoc.* 100, 781-796.

[8] Fujikoshi, Y., Ulyanov, V. V. and Shimizu, R. (2010). Multivariate Statistics: High-dimensional and Large-sample Approximations. Wiley, Hoboken, N.J.

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

[10] Jiang, D., Bai, Z. and Zheng, S. (2013). Testing the independence of sets of large-dimensional variables. *Sci. China Math.* 56, 135-147.

[11] Jiang, D., Jiang, T. and Yang, F. (2012). Likelihood ratio tests for covariance matrices of high-dimensional normal distributions. *J. Stat. Plann. Inference* 142, 2241-2256.

[12] Jiang, T. and Yang, F. (2013). Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. *Ann. Stat.* 41, 2029-2074.

[13] Jiang, T. and Qi, Y. (2015). Likelihood ratio tests for high-dimensional normal distributions. *Scand. J. Statist.* 42, 988-1009.

[14] Ledoit, O. and Wolf, M. (2002). Some hypothesis test for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.* 30, 1081-1102.

[15] Morrison, D. F. (2005). Multivariate Statistical Methods. *Duxbury Press*, 4th Ed.

[16] Mao, G. (2014). A new test of independence for high-dimensional data. *Statist. Probab. Lett.* 93, 14-18.

[17] McLeish, D.L. (1974). Dependent central limit theorems and invariance principles. *Ann. Probab.* 2, 620-628.

[18] Schott, J. R. (2001). Some tests for the equality of covariance matrices. *J. Stat. Plann. Inference* 94, 25-36.

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

[20] Schott, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Comput. Statist. Data Anal.* 51, 6535-6542.

[21] Srivastava, M. S. (2005). Some tests concerning the covariance matrix in high dimensional data. *J. Japan Statist. Soc.* 35, 251-272.

[22] Srivastava, M. S. (2006). Some tests criteria for the covariance matrix with fewer observations than the dimension. *Acta Comment. Univ. Tartu. Math.* 10, 77-93.

[23] Srivastava, M. S. and Reid, N. (2012). Testing the structure of the covariance matrix with fewer observations than the dimension. *J. Multivariate Anal.* 112, 156-171.