Biometrika (2005), **92**, 4, *pp.* 951–956

Printed in Great Britain

Miscellanea

Testing for complete independence in high dimensions

By JAMES R. SCHOTT

Department of Statistics and Actuarial Science, University of Central Florida, Orlando, Florida 32816-2370, U.S.A.

jschott@ucf.edu

Summary

A simple statistic is proposed for testing the complete independence of random variables having a multivariate normal distribution. The asymptotic null distribution of this statistic, as both the sample size and the number of variables go to infinity, is shown to be normal. Consequently, this test can be used when the number of variables is not small relative to the sample size and, in particular, even when the number of variables exceeds the sample size. The finite sample size performance of the normal approximation is evaluated in a simulation study and the results are compared to those of the likelihood ratio test.

Some key words: High-dimensional data; Independence of random variables.

1. Introduction

Often applications of multivariate analysis involve a large number of variables m. However, most inference procedures in multivariate analysis are based on asymptotic theory which has the sample size $N = n + 1$ going to infinity while m is fixed. Consequently, these procedures are not likely to be very accurate when m is of the same order of magnitude as n. In these situations, it would be better to use an inference procedure which is based on asymptotic theory as both n and m go to infinity. In particular, we would have n and m go to infinity with m/n converging to a constant $\gamma \in (0, \infty)$. Some examples of work on inference problems in this high-dimensional setting can be found in Bai & Saranadasa (1996), Dempster (1958, 1960), Ledoit & Wolf (2002), Saranadasa (1993) and Schott (2006).

In this paper, we consider a test for the independence of the variables comprising the $m \times 1$ vector y having a multivariate normal distribution with covariance matrix Σ. Suppose we have a random sample, $y_1, \ldots, y_N$, which is used to compute the usual unbiased sample covariance matrix S and the sample correlation matrix R. A test for complete independence tests that the covariance matrix is diagonal or, equivalently, that $P = I_m$, where P denotes the population correlation matrix. The likelihood ratio test rejects this hypothesis of complete independence for small values of $|R|$; see for example Morrison (2005, § 1.9). In particular, we would reject the hypothesis of complete independence if

$$w_{nm} = -\left(n - \frac{2m+5}{6}\right) \log |R|$$

exceeds the appropriate quantile from the $\chi^2_{m(m-1)/2}$ distribution. Clearly, this procedure is not valid for high-dimensional data since $|R| = 0$ whenever $m > n$. Other tests for complete independence have been developed. For instance, a procedure based on Fisher's z-transformation of the correlation coefficients was proposed by Chen & Mudholkar (1989). While their procedure is still based on asymptotic theory as only n goes to infinity, Chen & Mudholkar (1990) showed that it performs quite well for large values of m. However, this requires the approximation of the null distribution of the test statistic through a fitting process that uses approximations for the first three moments of the test statistic.

The goal of this paper is to develop a simpler test procedure specifically designed for high-dimensional data. Our test is based on the sample correlation matrix. An analogous test based on the sample covariance matrix can also be easily constructed. However, as a referee has pointed out, such a test would suffer from lack of invariance under scale changes of the variables.

2. A test for complete independence

If ρ_{ij} is the (i,j)th element of the correlation matrix P, then the hypothesis of complete independence can be written as $H_0: \rho_{ij} = 0$ $(i > j)$. A simple and intuitively plausible statistic for testing this hypothesis is the sum of squared r_{ij}'s for $i > j$. Since $E(r_{ij}^2) = n^{-1}$ under H_0 when $i \neq j$,

$$E\left(\sum_{i=2}^{m}\sum_{j=1}^{i-1} r_{ij}^2\right) = \frac{m(m-1)}{2n}$$

and so a statistic with mean 0 under the null hypothesis is given by

$$t_{nm} = \sum_{i=2}^{m}\sum_{j=1}^{i-1} r_{ij}^2 - \frac{m(m-1)}{2n}.$$

If h, i, j and k are distinct integers, then, under H_0,

$$E(r_{ij}^4) = 3/\{n(n+2)\}, \quad E(r_{ij}^2 r_{ik}^2) = E(r_{hi}^2 r_{jk}^2) = n^{-2}, \tag{1}$$

and this leads to

$$\sigma_{t_{nm}}^2 = \mathrm{var}(t_{nm}) = \frac{m(m-1)(n-1)}{n^2(n+2)},$$

so that $t_{nm}/\sigma_{t_{nm}}$ will have mean 0 and variance 1 if the complete-independence hypothesis holds.

Our main result establishes the asymptotic normality of t_{nm} as n and m approach infinity in such a way that

$$\lim(m/n) = \gamma \in (0, \infty). \tag{2}$$

To be more precise, we could write n and m as n_h and m_h so that each depends on a common index $h = 1, 2, \ldots,$ and then $\lim(m_h/n_h) = \gamma$ as $h \to \infty$. Similarly, we could write the correlation matrix as P_h since it also depends on the index h, because its dimensions are $m_h \times m_h$. However, for notational convenience, the dependence of these quantities on h will be suppressed throughout this paper. Note that, under condition (2),

$$\lim \sigma_{t_{nm}}^2 = \lim \frac{m(m-1)(n-1)}{n^2(n+2)} = \gamma^2.$$

THEOREM 1. *Suppose that the sample correlation matrix R has been computed from a random sample from a multivariate normal distribution with correlation matrix P. If $P = I_m$ and condition* (2) *holds, then t_{nm} converges in distribution to a normal random variable with mean* 0 *and variance γ^2.*

3. Some simulation results

The performance of the null approximating distribution of t_{nm} was investigated by way of simulation. Estimates of the actual significance levels were obtained from 5000 independent simulations with the nominal significance level $\alpha = 0{\cdot}05$. Both n and m ranged over the values 4, 8, 16, 32, 64, 128 and 256.

The simulation results for the test of H_0 based on t_{nm} are given in Table 1. The normal approximation generally yields inflated significance levels; however, in no case encountered in our study was the estimated significance level grossly higher than 0·05. As expected, the approximation improves as n and m increase. To compare these results with the likelihood ratio test, we have tabulated the significance levels for w_{nm} under the same settings in Table 2. The chi-squared approximation is particularly poor when $m = n$. As expected, for fixed m it improves as n increases, but the rate of improvement decreases as m increases.

Table 1: *Simulation study. Estimated significance levels for* t_{nm} *when* $\alpha = 0{\cdot}05$

m	$n=4$	$n=8$	$n=16$	$n=32$	$n=64$	$n=128$	$n=256$
4	0·062	0·063	0·066	0·069	0·071	0·072	0·071
8	0·062	0·061	0·060	0·060	0·059	0·065	0·062
16	0·065	0·060	0·055	0·060	0·057	0·056	0·055
32	0·066	0·060	0·060	0·056	0·050	0·054	0·055
64	0·072	0·054	0·050	0·056	0·051	0·046	0·057
128	0·065	0·068	0·058	0·052	0·050	0·050	0·050
256	0·065	0·055	0·054	0·057	0·047	0·047	0·056

Table 2: *Simulation study. Estimated significance levels for* w_{nm} *when* $\alpha = 0{\cdot}05$

m	$n=4$	$n=8$	$n=16$	$n=32$	$n=64$	$n=128$	$n=256$
4	0·132	0·060	0·053	0·052	0·050	0·054	0·052
8		0·421	0·084	0·058	0·056	0·055	0·050
16			0·722	0·083	0·055	0·052	0·048
32				0·989	0·135	0·060	0·056
64					0·995	0·192	0·068
128						1·000	0·465
256							1·000

In a second set of simulations, power estimates were obtained. Whereas Tables 1 and 2 were based on simulated $y \sim N_m(0, \Sigma)$ with $\Sigma = I_m$, the second simulation study used $\Sigma = (1-\rho)I_m + \rho 1_m 1_m'$, where $\rho = 0{\cdot}1$ and 1_m denotes the $m \times 1$ vector with each component equal to 1. The results for the test based on t_{nm} are given in Table 3. As expected, the power increases as n increases and increases as m increases. The values in the upper triangular portion of the table are generally larger than those in the lower triangular portion, indicating that the power increases at a faster rate in n than it does in m. The corresponding power estimates for w_{nm} are given in Table 4. Comparing with Table 3, we see that w_{nm} has larger power estimates only for cases in which the estimated significance level is substantially larger than the nominal 0·05. For $(m, n) = (8, 16)$, $(16, 32)$ and $(32, 64)$, w_{nm} yields smaller power estimates than t_{nm} even though the corresponding estimated significance levels of w_{nm} in Table 2 exceed those of t_{nm} in Table 1. The values in the upper right-hand portion of Table 4 are smaller than those in the same portion of Table 3, but this is probably because w_{nm} has slightly smaller significance levels than t_{nm} for these cases.

Table 3: *Simulation study. Estimated power for t_{nm} when $\alpha = 0{\cdot}05$*

m	$n=4$	$n=8$	$n=16$	$n=32$	$n=64$	$n=128$	$n=256$
4	0·076	0·087	0·123	0·172	0·307	0·534	0·845
8	0·079	0·101	0·177	0·313	0·597	0·903	0·998
16	0·112	0·166	0·310	0·595	0·904	0·997	1·000
32	0·161	0·285	0·557	0·871	0·996	1·000	1·000
64	0·255	0·486	0·797	0·987	1·000	1·000	1·000
128	0·375	0·698	0·946	1·000	1·000	1·000	1·000
256	0·542	0·846	0·990	1·000	1·000	1·000	1·000

Table 4: *Simulation study. Estimated power for w_{nm} when $\alpha = 0{\cdot}05$*

m	$n=4$	$n=8$	$n=16$	$n=32$	$n=64$	$n=128$	$n=256$
4	0·142	0·071	0·088	0·124	0·233	0·447	0·783
8		0·445	0·154	0·224	0·446	0·817	0·993
16			0·793	0·370	0·705	0·982	1·000
32				0·996	0·926	1·000	1·000
64					0·994	1·000	1·000
128						1·000	1·000
256							1·000

4. An example

To illustrate the method developed in this paper, we use some of the biochemical data given in Beerstecher et al. (1950). These data consist of 62 measurements on each of 12 individuals, 8 of whom were controls while the other 4 were alcoholics. We will restrict attention to one subset of the 62 variables, a set of 8 blood serum measurements. For each group, control and alcoholic, we will test the hypothesis of complete independence. Clearly, we will not be able to use the likelihood ratio test for either of the tests since $m = 8$ and $n = 7$ for the control group, while $n = 3$ for the alcoholic group.

For the control group, we find that $t_{7,8} = 3{\cdot}966$ and $\gamma^2 = 0{\cdot}762$. Standardising leads to a z-score of 4·54 which corresponds to a p-value of 0·000003. Thus, we have very strong evidence of some dependence among the eight variables. Turning to the alcoholic group, we obtain $t_{3,8} = 2{\cdot}477$ and $\gamma^2 = 2{\cdot}489$. This yields a z-score of 1·57 with a p-value of 0·0582. Consequently, the evidence of dependence among the variables is not nearly as strong for the alcoholic group.

Acknowledgement

The author would like to thank an associate editor and a referee whose comments on an earlier version of this work led to the contruction of the statistic presented in this paper.

Appendix

Proof of Theorem 1

Note that the (i, j)th element of S can be written as $s_{ij} = n^{-1}\sigma_{ii}^{1/2}\sigma_{jj}^{1/2}z_i'z_j$, where $z_1, \ldots, z_m$ are independently and identically distributed as $N_n(0, I_n)$. As a result, r_{ij} can be expressed as $r_{ij} = u_i'u_j$, where $u_i = (z_i'z_i)^{-1/2}z_i$, and $u_1, \ldots, u_m$ are independently distributed, each having a uniform distribution on the surface of the n-sphere. For $l = 2, \ldots, m$, let

$$X_{nl} = t_{nl} - t_{n,l-1} = \sum_{i=1}^{l-1} r_{li}^2 - \frac{l-1}{n},$$

where $t_{n1}=0$, so that $t_{nm}=\sum_{l=2}^{m} X_{nl}$. If we define the set $\mathscr{F}_{n,l-1}=\{u_1,\ldots,u_{l-1}\}$, then

$$E(r_{li}^2|\mathscr{F}_{n,l-1})=u_i'E(u_lu_l')u_i=u_i'(n^{-1}I_n)u_i=\frac{1}{n},$$

so that $E(X_{nl}|\mathscr{F}_{n,l-1})=0$. Consequently, for each n, $\{t_{nl}, l=2,\ldots,m\}$ is a martingale and $X_{n2},\ldots,X_{nm}$ are martingale differences. As a result, the theorem will follow from Corollary 3.1 of Hall & Heyde (1980, p. 58) if we can show that

$$\sum_l E\{X_{nl}^2 I(|X_{nl}|>\varepsilon)|\mathscr{F}_{n,l-1}\}\to 0, \tag{A1}$$

in probability, for all $\varepsilon>0$, and

$$\sum_l E(X_{nl}^2|\mathscr{F}_{n,l-1})\to\gamma^2, \tag{A2}$$

in probability. Here $I(.)$ denotes the indicator function. Using (1), we find that

$$E\{E(X_{nl}^2|\mathscr{F}_{n,l-1})\}=E(X_{nl}^2)=\frac{2(l-1)(n-1)}{n^2(n+2)},$$

and so it follows that

$$E\left\{\sum_{l=2}^{m}E(X_{nl}^2|\mathscr{F}_{n,l-1})\right\}=\sum_{l=2}^{m}\frac{2(l-1)(n-1)}{n^2(n+2)}=\frac{m(m-1)(n-1)}{n^2(n+2)}$$

converges to γ^2. Now $E(r_{li}^4|\mathscr{F}_{n,l-1})=3/\{n(n+2)\}$ and $E(r_{li}^2r_{lj}^2|\mathscr{F}_{n,l-1})=\{1+2(u_i'u_j)^2\}/\{n(n+2)\}$ if $i\neq j$, which leads to

$$E(X_{nl}^2|\mathscr{F}_{n,l-1})=\frac{3(l-1)}{n(n+2)}+\sum_{i=1}^{l-1}\sum_{\substack{j=1\\ j\neq i}}^{l-1}\frac{1+2(u_i'u_j)^2}{n(n+2)}-\frac{(l-1)^2}{n^2}.$$

Thus, upon using the identities

$$E\{(u_i'u_j)^2\}=n^{-1},\quad E\{(u_i'u_j)^4\}=\frac{3}{n(n+2)},\quad E\{(u_i'u_j)^2(u_i'u_k)^2\}=E\{(u_h'u_i)^2(u_j'u_k)^2\}=n^{-2},$$

where h, i, j and k are distinct, and then simplifying, we find that

$$E\left[\left\{\sum_{l=2}^{m}E(X_{nl}^2|\mathscr{F}_{n,l-1})\right\}^2\right]=\frac{m^4}{n^4}+o(1),$$

and this converges to γ^4. It then follows that

$$E\left[\left\{\sum_{l=2}^{m}E(X_{nl}^2|\mathscr{F}_{n,l-1})-\gamma^2\right\}^2\right]=E\left[\left\{\sum_{l=2}^{m}E(X_{nl}^2|\mathscr{F}_{n,l-1})\right\}^2\right]-2\gamma^2E\left\{\sum_{l=2}^{m}E(X_{nl}^2|\mathscr{F}_{n,l-1})\right\}+\gamma^4$$
$$\to\gamma^4-2\gamma^2(\gamma^2)+\gamma^4=0,$$

and this guarantees that (A2) holds. The Lindeberg condition given in (A1) can be established by showing that the stronger Liapounov condition,

$$\sum_l E(X_{nl}^4|\mathscr{F}_{n,l-1})\to 0, \tag{A3}$$

in probability, holds. Now using $E(r_{li}^8)=105/\{n(n+2)(n+4)(n+6)\}$ and

$$E(r_{li}^6)=15/\{n(n+2)(n+4)\},$$

along with the other moments of r_{li} previously identified and the fact that m is $O(n)$, we find that

$$E\left\{\sum_l E(X_{nl}^4|\mathscr{F}_{n,l-1})\right\}=\left\{\sum_l E(X_{nl}^4)\right\}=O(n^{-1}). \tag{A4}$$

Since the quantity in (A4) converges to 0, (A3) holds. This completes the proof. □

REFERENCES

BAI, Z. D. & SARANADASA, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statist. Sinica* **6**, 311–29.

BEERSTECHER, JR., E., SUTTON, H. E., BERRY, H. K., BROWN, W. D., REED, J., RICH, G. B., BERRY, L. J. & WILLIAMS, R. J. (1950). Biochemical individuality. V. Explorations with respect to the metabolic patterns of compulsive drinkers. *Arch. Biochem.* **29**, 27–40.

CHEN, S. & MUDHOLKAR, G. S. (1989). A remark on testing significance of an observed correlation matrix. *Aust. J. Statist.* **31**, 105–10.

CHEN, S. & MUDHOLKAR, G. S. (1990). Null distribution of the sum of squared z-transforms in testing complete independence. *Ann. Inst. Statist. Math.* **42**, 149–55.

DEMPSTER, A. P. (1958). A high dimensional two sample significance test. *Ann. Math. Statist.* **29**, 995–1010.

DEMPSTER, A. P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics* **16**, 41–50.

HALL, P. & HEYDE, C. C. (1980). *Martingale Limit Theory and its Applications*. New York: Academic Press.

LEDOIT, O. & WOLF, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.* **30**, 1081–102.

MORRISON, D. F. (2005). *Multivariate Statistical Methods*, 4th ed. Belmont, CA: Brooks/Cole.

SARANADASA, H. (1993). Asymptotic expansion of the misclassification probabilities of D- and A-criteria for discrimination from two high-dimensional populations using the theory of large-dimensional random matrices. *J. Mult. Anal.* **46**, 154–74.

SCHOTT, J. R. (2006). A high dimensional test for the equality of the smallest eigenvalues of a covariance matrix. *J. Mult. Anal.* To appear.

[*Received April* 2005. *Revised June* 2005]