Journal of the Korean Statistical Society (1999), **28***: 4, pp 479–488*

A Note on the Chi-Square Test for Multivariate Normality Based on the Sample Mahalanobis Distances

Cheolyong Park[1]

Abstract

Moore and Stubblebine(1981) suggested a chi-square test for multivariate normality based on cell counts calculated from the sample Mahalanobis distances. They derived the limiting distribution of the test statistic only when equiprobable cells are employed. Using conditional limit theorems, we derive the limiting distribution of the statistic as well as the asymptotic normality of the cell counts. These distributions are valid even when equiprobable cells are not employed. We finally apply this method to a real data set.

Keywords: Ancillary statistics; Basu's theorem; Conditional limit theorems; Radon-Nikodym theorem; Wishart distribution

1. INTRODUCTION

Moore and Stubblebine(1981) suggested a chi-square test for multivariate normality based on the sample Mahalanobis distances

$$\hat{d}_i = \sqrt{(X_i - \bar{X})^t S_x^{-1}(X_i - \bar{X})},\ i = 1, 2, \ldots, n,$$

where $\bar{X}$ and S_x are the sample mean vector and sample covariance matrix of a p-variate random sample $X_1, \ldots, X_n$ and "t" is a notation used for transpose. Their chi-square test is based on the frequencies of the squared Mahalanobis distances $\hat{d}_i^2$ falling into fixed disjoint intervals in real line. Their argument for the asymptotic multivariate normality of the cell counts was not rigorous but could be justified by the central limit theorem for empirical processes due to Dudley(1978). They derived an exact form of the limiting distribution of the chi-square test statistic only when equiprobable intervals are chosen; i.e. the intervals are chosen such that

[1] Department of Statistics, Keimyung University, Taegu 704-701

the probability of each squared Mahalanobis distance belonging to any interval is equal.

The equiprobable way of forming intervals is commonly employed in practice since it is perfectly definite and unique given that the number of intervals is fixed. However, it may well result in a loss of sensitivity at the extremes of the range of the squared Mahalanobis distances $\hat{d}_i^2$ and thus it will not necessarily increase the power of the chi-square test statistic (see p.456-458 in Kendall and Stuart(1979) for details). This means that we do not need to confine ourselves to the equiprobable intervals which are not necessarily optimal in terms of the power of the test. Therefore, from both theoretical and practical points of view, it will be of great importance to derive the limiting distribution of the test statistic which can be applied to unequally probable intervals as well as to the equiprobable intervals.

In this paper, we will use conditional limit theorems to derive the limiting distribution of the chi-square test statistic as well as the asymptotic multivariate normality of the cell counts. We can provide an exact form of the limiting distribution of the chi-square test statistic even when the intervals are not chosen equiprobable. We also provide an exact form of the asymptotic joint distribution of the cell counts. We can find that two exact forms of the limiting and the asymptotic joint distributions do not depend on the parameters since the cell counts are based on ancillary statistics $\hat{d}_i^2$.

In Section 2, we introduce Moore and Stubblebine's method in detail and then present main results on the asymptotic joint distribution of the cell counts and on the limiting distribution of the chi-square test statistic. In Section 3, we provide some necessary lemmas and then prove the main results. In Section 4, we provide an example of application to a real data set in which the test with unequally probable cells performs better than the one with the equiprobable cells.

2. THE METHOD AND MAIN RESULTS

Before presenting main results we will define some notations. Most of notations will be the same as those in Moore and Stubblebine(1981). Unless otherwise noted, vectors will be column vectors, but for convenience they will be written in text as row vectors.

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N_p(\mu, \Sigma)$ for some μ and nonsingular Σ. Let $\theta = (\mu, \Sigma)$ be the parameter of the distribution and $\theta_0 = (0, I)$ be a particular parameter with $\mu = 0, \Sigma = I$. The maximum likelihood estimator

(MLE) of θ is denoted by $\hat{\theta}_n = (\bar{X}, S_x)$, where we use n for the denominator of the sample covariance matrix S_x.

For a given θ, and $0 = c_0 < c_1 < \cdots < c_M = \infty$, define cells

$$E_i(\theta) = \left\{x \in R^p : c_{i-1} \leq (x-\mu)^t \Sigma^{-1}(x-\mu) < c_i\right\} \tag{2.1}$$

and $E_i = E_i(\theta_0) = \{x \in R^p : c_{i-1} \leq x^t x < c_i\}$. Let $N_i(\theta)$ denote the number of $X_1, \ldots, X_n$ falling in $E_i(\theta)$ and let $N(\theta) = (N_1(\theta), \ldots, N_M(\theta))$ denote the M-vector of the cell counts. For M data-dependent cells $E_{in} = E_i(\hat{\theta}_n)$, $i = 1, \ldots, M$, we define $N_{in} = N_i(\hat{\theta}_n)$ to denote the cell count belonging to E_{in} and $N_n = N(\hat{\theta}_n)$ to denote the M-vector of the cell counts. Let

$$p_i(\theta, \theta_1) = P_\theta(X_1 \in E_i(\theta_1)) \tag{2.2}$$

be the cell probability for $E_i(\theta_1)$ under θ, then $p_i(\theta) = p_i(\theta, \hat{\theta}_n)$ is the cell probability for E_{in} under θ. Let $p_{in} = p_i(\hat{\theta}_n)$ be the estimated cell probability corresponding to the cell count N_{in} and $p_n = (p_{1n}, \ldots, p_{Mn})$ be the M-vector of the estimated cell probability. It is easy to verify that p_n does not depend on n since $p_{in} = F_p(c_i) - F_p(c_{i-1})$ where F_p is the cumulative distribution function of the $\chi^2(p)$ distribution.

For a given vector $y = (y_1, y_2, \ldots, y_m)$, we define the diagonal matrix $D(y)$ and the vector of square root values $\sqrt{y}$ to be

$$D(y) = \operatorname{diag}(y_1, y_2, \ldots, y_m), \quad \sqrt{y} = (\sqrt{y_1}, \sqrt{y_2}, \ldots, \sqrt{y_m}).$$

Then Pearson chi-square test statistic for multivariate normality is

$$X^2(\hat{\theta}_n) = (N_n - np_n)^t \left\{D(np_n)\right\}^{-1} (N_n - np_n). \tag{2.3}$$

To derive the limiting distribution of $X^2(\hat{\theta}_n)$, we first need to derive the asymptotic joint distribution of N_n. For data-dependent, non-rectangular cells like E_{in}, we can employ the central limit theorem for empirical processes due to Dudley(1978) to show the asymptotic multivariate normality of N_n. Pollard(1978) derived the limiting distribution of various chi-square test statistics based on the Central Limit Theorem. Moore and Stubblebine(1981) derived the limiting distribution of $X^2(\hat{\theta}_n)$ based on Pollard's result without verifying regularity conditions in details.

We will derive the asymptotic multivariate normality of N_n based on a conditional limit theorem due to Holst(1981). By utilizing the fact that the squared Mahalanobis distances $\hat{d}_i^2$ are ancillary (see Lemma 3.1 in the next Section for proof), we can provide exact forms of both the asymptotic joint distribution of N_n and the limiting distribution of $X^2(\hat{\theta}_n)$.

Here are main results on the asymptotic joint distribution of N_n and the limiting distribution of $X^2(\hat{\theta}_n)$.

Theorem 2.1.

$$n^{-1/2}(N_n - np_n) \xrightarrow{d} N_M(0, A^*) \quad as \quad n \to \infty$$

where $A^* = D(p_n) - p_n p_n^t - 2B^*B^{*t}$, $B^* = (d_1 1_p, \ldots, d_M 1_p)^t$, 1_p *is the p-vector of ones, and*

$$d_i = \left(c_{i-1}^{p/2} e^{-c_{i-1}/2} - c_i^{p/2} e^{-c_i/2}\right) b_p/2$$

$$b_p = \begin{cases} [p(p-2)\cdots 2]^{-1} & p \ even \\ (2/\pi)^{1/2}[p(p-2)\cdots 1]^{-1} & p \ odd. \end{cases}$$

Corollary 2.1.

$$X^2(\hat{\theta}_n) \xrightarrow{d} W_1 + \lambda W_2 \quad as \quad n \to \infty$$

where W_1 *and* W_2 *are independent chi-square variates with degrees of freedom* $M-2$ *and* 1, *respectively and* $\lambda = 1 - 2pd^*$ *with* $d^* = \sum_{i=1}^{M}(d_i^2/p_{in})$.

Note that exact forms are given for the asymptotic joint distribution of N_n and for the limiting distribution of $X^2(\hat{\theta}_n)$. Moore and Stubblebine(1981) provided a form of the asymptotic joint distribution of N_n but it is not as precise as Theorem 2.1 since it depends on the limiting value of $\hat{\theta}_n$. Also they provided the exact form of the limiting distribution of $X^2(\hat{\theta}_n)$ only when equiprobable cells are employed; i.e. $p_{in} = 1/M$ for all i.

3. PROOFS

In this section, we will prove Theorem 2.1 and then Corollary 2.1. As noted in Section 1, a conditional limit theorem in exponential families due to Holst(1981) is an essential tool to derive Theorem 2.1. Corollary 2.1 is proved easily from Theorem 2.1.

Proof of Theorem 2.1

To prove Theorem 2.1, we will use the multivariate conditional limit theorem by Park(1995), an extension of a univariate conditional limit theorem due to Holst(1981). While deriving the result, we will utilize the following lemma.

Lemma 3.1. *The squared Mahalanobis distances $\hat{d}_i^2$ are ancillary and independent of $\hat{\theta}_n = (\bar{X}, S_x)$.*

Proof: Define $U_i = \Sigma^{-1/2}(X_i - \mu)$ for each i, then $U_1, \ldots, U_n$ is a random sample from $N_p(0, I)$ and thus the distribution of U_i's are free of θ. Let $\bar{U}$ and S_u denote the sample mean vector and sample covariance matrix of $U_1, \ldots, U_n$. Then the distribution of statistics

$$(X_i - \bar{X})^t S_x^{-1}(X_i - \bar{X}) = (U_i - \bar{U})^t S_u^{-1}(U_i - \bar{U})$$

is free of θ and so $\hat{d}_i^2$'s are ancillary. Since $\hat{\theta}_n$ is a sufficient and complete statistic for θ, $\hat{d}_i^2$'s are independent of $\hat{\theta}_n$ by Basu's theorem. This completes the proof. □

By the above lemma, the vector N_n of cell counts are ancillary and thus we have

$$\begin{aligned}\mathcal{L}_\theta(N_n) &= \mathcal{L}_{\theta_0}(N_n) = \mathcal{L}_{\theta_0}(N_n | \bar{X} = 0, S_x = I) \\ &= \mathcal{L}_{\theta_0}(N(\theta_0) | \bar{X} = 0, S_x = I) \qquad (3.1)\end{aligned}$$

for a fixed parameter $\theta_0 = (0, I)$, where the last equality holds since $N_n = N(\hat{\theta}_n)$ is equal to $N(\theta_0)$, given that $\bar{X} = 0, S_x = I$. To apply the conditional theorem by Park(1995), the conditions given in (3.1) need to be given in terms of the canonical sufficient statistics for the $N_p(\mu, \Sigma)$ distribution. Thus we define some notations: For any p-vector $x = (x_1, \ldots, x_p)$, we define column vectors $s(x) = (x, d(x), r(x))$ and $u(x) = (I(x \in E_1), \ldots, I(x \in E_M))$ where $d(x) = (x_1^2, \ldots, x_p^2)$, $r(x) = (x_1 x_2, \ldots, x_{p-1} x_p)$ and E_i's are defined just after (2.1). With these notations, we can see that the canonical sufficient statistics is $\sum_{i=1}^n s(X_i)$ and $N(\theta_0) = \sum_{i=1}^n u(X_i)$. Since $\{\bar{X} = 0, S_x = I\}$ is equivalent to $\{\sum_i s(X_i)/n = (0_p, 1_p, 0_{p(p-1)/2})\}$, the equation (3.1) is equal to

$$\mathcal{L}_{\theta_0}\left(\sum_{i=1}^n u(X_i) \mid \sum_{i=1}^n s(X_i)/n = (0_p, 1_p, 0_{p(p-1)/2})\right).$$

Now we are ready to apply Corollary 1 of Park(1995) and derive the asymptotic distribution of N_n. Assumption A1 is cleary satisfied since the multivariate normal distributions belong to a regular exponential family (See p. 116 of Barndorff-Neilson (1978) for details). Assumption A2 is satisfied by fixing

$$s_n = (0_p, 1_p, 0_{p(p-1)/2}) \text{ and } \theta_n = \theta_0.$$

Assumption A3 is satisfied since $u^ts(X_1)$ is not equal to some constant with probability 1 for any θ unless $u = 0$. Assumptions A4$'$ and A6 are trivially satisfied since $u(X_1)$ is a vector of M indicator variables and since $\theta_n = \theta_0$. We will now verify Assumption A5$'$ to hold in the following lemma.

Lemma 3.2. *There exists n_0 such that*

$$\int_{R^q} \left| E_\theta \exp\left\{ i\xi^t u(X_1) + i\eta^t s(X_1) \right\} \right|^{n_0} d\eta < \infty,$$

for all $\xi \in R^M$ and for all $\theta = (\mu, \Sigma)$ with arbitrary mean vector μ and positive definite covariance matrix Σ, where $q \equiv 2p + p(p-1)/2$.

Proof: First we will show that, for $n \geq p+1$, $\sum_i (X_i, X_i X_i^t)$ has a bounded density. Note that $\sum_i (X_i, X_i X_i^t)$ is just an alternative way of writing $\sum_i s(X_i)$. Since $\sum_i X_i \sim N(n\mu, n\Sigma)$ and nS_x has the Wishart distribution with parameters $n-1, p, \Sigma$, and since $\sum_i X_i$ and nS_x are independent, the joint density of $(\sum_i X_i, nS_x)$ is the product of the densities of $\sum_i X_i$ and nS_x. Since both $\sum_i X_i$ and nS_x have bounded densities when $n \geq p+1$ (see p. 162 in Johnson and Kotz (1972) for details), the joint density of $(\sum_i X_i, nS_x)$ is bounded. Thus the joint density of $\sum_i (X_i, X_i^t X_i)$ is also bounded since the Jacobian of the transformation from $(\sum_i X_i, nS_x)$ to $\sum_i (X_i, X_i X_i^t)$ is 1.

Fix θ. Let P be the probability measure of $s(X_1)$ under θ and P_k be the restriction of P to the set $A_k \equiv \{s(y) : y \in E_k\}$, i.e. $P_k(B) = P(B \cap A_k)$. Define $p_k \equiv P_k(R^q) = \mathrm{P}_\theta\{X_1 \in E_k\}$ for each k. Then $P = \sum_{k=1}^M P_k$ and $\sum_k p_k = 1$ with $p_k > 0$ for each k, so that $Q_k \equiv p_k^{-1} P_k$ is a probability measure. Thus,

$$P^{*m} = \sum_{k_1,\ldots,k_m} P_{k1} * \cdots * P_{k_m} = \sum_{k_1,\ldots,k_m} p_{k_1}^{-1} Q_{k_1} * \cdots * p_{k_m}^{-1} Q_{k_m}, \tag{3.2}$$

where P^{*m} is the m-fold convolution of P. Take $m = p+1$. Since $Q_k^{*m} \ll P^{*m}$, by the Radon-Nikodym Theorem, there exists a non-negative density f_k such that

$$Q_k^{*m}(A) = \int_A f_k \, dP^{*m}$$

for all $A \in R^q$. Since $\sum_i (X_i, X_i X_i^t)$ has a bounded density, the Radon-Nikodym derivative g of P^{*m} with respective to the Lebesgue measure ν on R^q is bounded and

$$Q_k^{*m}(A) = \int_A f_k g \, d\nu.$$

Since $\int_A (1-p_k^{-m} f_k) dP^{*m} = P^{*m}(A) - p_k^{-m} Q_k^{*m}(A) \geq 0$ for all $A \in R^q$ by (3.2), we have $f_k \leq p_k^m$ and so $f_k g$ is also bounded. Now, by theorem 19.1 of Bhattacharya and Rao (1976), the existence of the bounded density $f_k g$ is equivalent to that the characteristic function corresponding to the probability measure Q_k is in $L^{n_0}(R^q)$ space for some n_0. In other words, there exists n_0 such that

$$\int_{R^q} \Big| E_\theta \exp\{i\eta^t Y_k\} \Big|^{n_0} d\eta < \infty \tag{3.3}$$

for all k, where Y_k is a random vector with the probability measure Q_k.

Using this integrability result it is straightforward to complete the proof of the lemma. First we observe that

$$\begin{aligned} & \Big| E_\theta \exp\{i\xi^t u(X_1) + i\eta^t s(X_1)\} \Big| \\ = & \Big| \sum_{k=1}^{M} \exp\{i\xi_k\} \, E_\theta \Big[\exp\{i\eta^t s(X_1)\} \, I(X_1 \in E_k\Big)\Big] \Big| \\ \leq & \sum_{k=1}^{M} \Big| E_\theta \Big[\exp\{i\eta^t s(X_1)\} \, I(X_1 \in E_k)\Big] \Big| . \end{aligned}$$

Now note that

$$\begin{aligned} E_\theta \Big[\exp\{i\eta^t s(X_1)\} \, I(X_1 \in E_k)\Big] &= \int_{R^q} \exp\{i\eta^t y\} dP_k(y) \\ = p_k \int_{R^q} \exp\{i\eta^t y\} dQ_k(y) &= p_k E_\theta \exp\{i\eta^t Y_k\}. \end{aligned}$$

Thus,

$$\begin{aligned} & \Big| E_\theta \exp\{i\xi^t u(X_1) + i\eta^t s(X_1)\} \Big|^{n_0} \\ \leq & \Big\{ \sum_{k=1}^{M} \Big| E_\theta \Big[\exp\{i\eta^t s(X_1)\} \, I(X_1 \in E_k)\Big] \Big| \Big\}^{n_0} \\ = & \Big\{ \sum_{k=1}^{M} p_k \Big| E_\theta \exp\{i\eta^t Y_k\} \Big| \Big\}^{n_0} \leq \sum_{k=1}^{M} p_k \Big| E_\theta \exp\{i\eta^t Y_k\} \Big|^{n_0} , \end{aligned}$$

which is integrable by (3.3). This completes the proof. □

By Corollary 1 of Park(1995), we have

$$n^{-1/2}(N_n - np_n) \xrightarrow{d} N_M(0, A - BC^{-1}B^t) \quad \text{as } n \to \infty,$$

where

$$A = \mathrm{Cov}_{\theta_0}(u(X_1)),\ B = \mathrm{Cov}_{\theta_0}(u(X_1), s(X_1)),\ C = \mathrm{Cov}_{\theta_0}(s(X_1)).$$

Using the results in Appendix of Moore and Stubblebine(1981), we can easily show that

$$A = D(p_n) - p_n p_n^t,\ B = (B_1, B_2, B_3),\ C = \mathrm{diag}(I_p, 2I_p, I_{p(p-1)/2}),$$

where

$$B_1 = \mathrm{Cov}_{\theta_0}(u(X_1), X_1) = 0,\quad B_3 = \mathrm{Cov}_{\theta_0}(u(X_1), r(X_1)) = 0,$$

and $B_2 = \mathrm{Cov}_{\theta_0}(u(X_1), d(X_1)) = 2B^*$ with B^* defined in this theorem. Therefore we have

$$A^* = A - BC^{-1}B^t = D(p_n) - p_n p_n^t - 2B^* B^{*t},$$

which completes the proof of Theorem 2.1. □

Proof of Corollary 2.1

Since $X^2(\hat{\theta}_n) = (N_n - np_n)^t \left\{D(np_n)\right\}^{-1} (N_n - np_n)$, we need to calculate the eigenvalues of

$$F \equiv \{D(p_n)\}^{-1/2} A^* \{D(p_n)\}^{-1/2} = I - \sqrt{p_n}\sqrt{p_n}^t - 2pd^* EE^t,$$

where $E = \{D(p_n)\}^{-1/2} B^* / \sqrt{pd^*}$ with d^* defined in this corollary. Since $E^t E = 1_p 1_p^t / p$, it is easy to show that EE^t is an idempotent matrix of rank 1 and so is $\sqrt{p_n}\sqrt{p_n}^t$.

Since $\sum_i d_i = 0$, two idempotent matrices $\sqrt{p_n}\sqrt{p_n}^t$ and EE^t are orthogonal. This show that eigenvalues of F are 1 with multiplicity $M - 2$, $1 - 2pd^*$ with multiplicity 1, and 0 with multiplicity 1. This completes the proof. □

4. AN EXAMPLE

In this section, we provide an illustrative example of applying the chi-square test to a real data set. The real data are the mineral content data which are presented in table 1.7 of Johnson and Wichern (1992). The table contains 25 cases of six measurements for the mineral content on the dominant and nondominant sides of three bones. We investigate the multivariate normality of the first two variables of the table; the mineral content of the dominant and nondominant sides of radius.

We first focus on the case where the number of cells is three, i.e. $M = 3$. The test with the equiprobable cells leads to the following results; the vector of cell counts is $N_n = (12, 8, 5)$ and the chi-square value is 2.96, so that the asymptotic p-value of the test is greater than .0853 since 2.96 corresponds to the upper .0853 quantile of $\chi^2(1)$. The test with $p_{1n} = 1/5, p_{2n} = p_{3n} = 2/5$ leads to the vector of cell counts $N_n = (2, 18, 5)$ and the chi-square value 10.7. Thus this test has an asymptotic p-value less than .0047 since 10.7 corresponds to the upper .0047 quantile of $\chi^2(2)$. For the case where $M = 4$, we obtain similar results; the equiprobable cells lead to the chi-square value 5.24 with p-value greater than .0728 whereas the cells with $p_{1n} = p_{2n} = 1/5, p_{3n} = p_{4n} = 3/10$ lead to the chi-square value 16.53 with p-value less than .0009.

For $M = 5$, the chi-square test with equiprobable cells performs quite well with p-value less than .0018. We do not consider unequally probable cells since expected counts for some cells are less than 5. Similarly we do not consider the cases where $M > 5$ since average cell counts are less than 5.

In this example, the test with unequally probable cells performs better in detecting deviations from the multivariate normality. When the number of cells is small, the chi-square test is quite sensitive to the choice of cells and thus we may have to try unequally probable cells as well as the equiprobable cells.

References

Barndorff-Nielson, O. (1978). *Information and Exponential Families in Statistical Theory.* John Wiley & Sons, Inc.

Bhattacharya, R.N., and Rao, R.R. (1976). *Normal Approximation and Asymptotic Expansions.* John Wiley & Sons, Inc.

Dudley, R.M. (1978). "Central Limit Theorems for Empirical Measures," *The Annals of Probability* **6**, pp. 899-929.

Holst, L. (1981). "Some Conditional Limit Theorems in Exponential Families," *The Annals of Probability* **9**, 818-30.

Johnson, N.L., and Kotz, S. (1970). *Distributions in Statistics: Continuous Multivariate Distributions.* John Wiley & Sons, Inc.

Johnson, R.A., and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis,* Third Edition. Prentice Hall.

Kendall, S.M., and Stuart, A.S. (1979). *The Advanced Theory of Statistics: Volume 2 Inference and Relationship.* Charles Griffin & Company Ltd.

Moore, D.S., and Stubblebine, J.B. (1981). "Chi-square Tests for Multivariate Normality with Application to Common Stock Prices," *Communications in Statistics - Theory and Methods* **10**, 713-738.

Park, C. (1995). "Some Remarks on the Chi-Squared Test with Both Margins Fixed," *Communications in Statistics - Theory and Methods* **24**, 653-61.

Pollard, D. (1979). "General Chi-square Goodness-of-fit Tests with Data-dependent Cells," *Z. Wahrsch. verw. Gebiete* **50**, 317-332.