Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.4 pp. 1013~1021

The Rao-Robson Chi-Squared Test for Multivariate Structure[1)]

Cheolyong Park[2)]

Abstract

Huffer and Park (2002) proposed a chi-squared test for multivariate structure. Their test detects the deviation of data from mutual independence or multivariate normality. We will compute the Rao-Robson chi-squared version of the test, which is easy to apply in practice since it has a limiting chi-squared distribution. We will provide a self-contained argument that it has a limiting chi-squared distribution. We study the accuracy in finite samples of the limiting distribution. We finally compare the power of our test with those of other popular normality tests in an application to a real data.

Keywords : Multivariate structure, testing independence.

1. Introduction

Huffer and Park (2002) proposed a test that might detect the presence of 'interesting' multivariate structure in large data sets. The basic idea there is that a data set is 'trivial' if its coordinates are independent. In such a data set, the multivariate structure is entirely determined by the marginal distributions. More generally, a data set is trivial if there exists some linear transformation which converts it into a new data set with independent coordinates. For example, there is a linear transformation which converts the multivariate normal distribution into a distribution with independent normal coordinates and so the multivariate normal distribution is trivial.

Their approach, put briefly, is as follows: Given a data set, a simple linear transformation is employed which converts it into a new data set with the

---

1) This research has been conducted by the Bisa Research Grant of Keimyung University in 2002.

2) Associate Professor, Department of Statistics, Keimyung University, Taegu 704-701
E-mail : cypark1@kmucc.kmu.ac.kr

covariance matrix (at least approximately) equal to the identity matrix. After transforming the data, the hypothesis of independence is tested by discretizing each coordinate based on sample quantiles of the coordinate and analyzing the resulting categorical data as a contingency table. The cell counts in this contingency table are compared with those expected under independence and, if a formal test statistic is desired, the usual chi-squared test of independence is employed.

Their chi-squared test statistic is Pearson-Fisher type and so its limiting distribution is not an exact chi-squared distribution, which is well known since the work of Chernoff and Lehmann (1954). Therefore the test is not easy to apply in practice since its asymptotic p-value is not easy to compute. Even though it is possible to compute the tail probability of a weighted chi-squared distribution (see Imhof (1961), Solomon and Stephens (1977), Farebrother (1990) among others), we will employ the generalized Wald's method which makes the resulting test statistic have a limiting chi-squared distribution. This approach is proposed by Rao and Robson (1974) and independently by Nikulin (1973) and such test statistic is denoted by the Rao-Robson statistic in this paper.

In Section 2, we briefly describe the procedure in Huffer and Park (2002) and then compute the Rao-Robson chi-squared test statistic. We will also show directly that it has a limiting chi-squared distribution. In Section 3, we provide an illustrative example of application to a real data, in which our method is compared with other popular tests of multivariate normality. Moreover, we provide a simulation study to check the accuracy in finite samples of the limiting distribution.

2. The Procedure and Main Result

We will first give some brief remarks on notation. We will use I and 0 to denote an identity matrix and a column vector or matrix of zeros respectively. The dimensions will usually be clear from context, but will be specified by subscripts if necessary. Unless otherwise noted vectors will be column vectors, but for convenience they will be written in text as row vectors.

The p-dimensional multivariate distribution of mean vector μ and covariance matrix Σ and the chi-squared distribution with f degrees of freedom will be dented by $N_p(\mu, \Sigma)$ and $\chi^2(f)$, respectively. The probability density function and cumulative distribution function of the (univariate) standard normal distribution will be by ϕ and Φ, respectively.

Let $Y_1, Y_2, \cdots, Y_n$ be a random sample from $N_p(\mu, \Sigma)$ with nonsingular Σ. Then the sample mean vector is given by $\overline{Y} = \sum_{i=1}^{n} Y_i/n$ and the sample

covariance matrix is given by $S=\sum_{i=1}^{n}(Y_i-\overline{Y})(Y_i-\overline{Y})^t/(n-1)$, where 't' denotes transpose. Here we assume that $n>p$, so that S is nonsingular.

We now describe briefly the procedure in Huffer and Park (2002). We first spherize the original data so that the sample mean vector is zero and the sample covariance matrix is identity. In other words, the transformed data is obtained by

$$Z_i=(z_{i1},\cdots,z_{ip})=R(S)(Y_i-\overline{Y}) \quad (i=1,2,\cdots,n)$$

where $R(S)$ is such that $R(S)SR^t(S)=I$. The quantity $R(S)$ is the function of the original data only through the sample covariance matrix S. There are many such choices of $R(S)$ that can spherize the original data. For example, $R(S)$ can be chosen as a low triangular matrix with positive diagonal elements based on Gram-Schmidt procedure. Two other popular choices are the rotation based on the principal components and the rotation using the square root matrix $R(S)=S^{-1/2}$. We can choose any rotation method since the distribution of the transformed data does not depend on the choices of rotations (see lemma 3.1 of Huffer and Park for details).

We next obtain a grouped data by binning the transformed data based on the sample quantiles of each coordinate of transformed data. Here, the same number of boundaries is set for each coordinate and the boundaries must be set such that adjacent boundaries form equiprobable intervals. In other words, we use the sample quantiles of each column to assign the values in that column into, say d groups (labeled $1,2,\cdots,d$) of equal size n/d. (If n is not divisible by d, the group sizes will not be exactly equal.)

We now form a contingency table from the grouped data. This contingency table contains d^p cells and n observations are distributed among these d^p cells. We use $\pi=(\pi_1,\pi_2,\cdots,\pi_p)$, with $1\le\pi_i\le d_i$ for each i, to denote a particular cell in our table. Let $\hat{\xi}_{i,j}$ be the (j/d)-th sample quantile of the values $z_{1i},z_{2i},\cdots,z_{ni}$ in the i-th coordinate of the transformed data. We set $\hat{\xi}_{i,0}=-\infty$ and $\hat{\xi}_{i,d}=\infty$. Then, for each cell π, the cell count $u_{n\pi}$ is given by

$$u_{n\pi}=\sum_{k=1}^{n} I\{\hat{\xi}_{i,\pi_i-1}<z_{ki}\le\hat{\xi}_{i,\pi_i},\ \text{for } 1\le i\le p\},$$

where $I\{\cdot\}$ is an indicator function and z_{ki} is the i-th coordinate of the k-th transformed vector Z_k. As a measure for the degree of departure from the multivariate structure, we use the chi-squared statistic X^2 defined by

$$X^2=\sum_{\pi}\frac{(u_{n\pi}-n/d^p)^2}{n/d^p}.$$

Note that this is the usual chi-squared test statistic for testing total

independence in a multi-way contingency table and that, under independence or multivariate normality, the expected number of observations in any given cell is approximately n/d^p.

The choice of d in our procedure is somewhat arbitrary. We generally prefer to have as many cells as possible without allowing the average cell count n/d^p to be too small. If we wish to use the limiting distribution of the chi-squared statistic for testing purposes, the usual guidelines apply: the limiting distribution is fairly accurate when $n/d^p \geq 5$. Our simulation study in next section shows that if the number of cells is large enough, it is reasonably good even for $n/d^p = 1$. Since the number of cells grows rapidly with p, for high dimensional data sets we are often forced to use small values of d in order to avoid extremely small average cell counts.

We now present the limiting distribution of the vector of cell counts $u_{n\pi}$ and of the chi-squared statistic X^2 under multivariate normality. First, we introduce various matrices which are need in the statements of our results. Let $U_n = (u_{n\pi})$ be the $d^p \times 1$ vector of cell counts. For easy representation of results, we assume the elements of the U_n are arranged in such a way that the corresponding cell vectors $\pi = (\pi_1, \pi_2, \cdots, \pi_p)$ are in a standard order; i.e. the first coordinate π_1 changes form 1 to d_1 the fastest, the second coordinate π_2 changes the second fastest, and so on.

Now, for $i = 1, \cdots, p;\ j = 1, \cdots, d_i,$ define $\xi_j = \Phi^{-1}(j/d)$ and $\psi_j = \phi(\xi_{j-1}) - \phi(\xi_j)$. We define D_1 to be a $d^p \times p$ matrix whose i-th column is given by $e_d \otimes \cdots \otimes \psi \otimes \cdots \otimes e_d$, where e_d is the vector of d ones and $\psi = (\psi_1, \cdots, \psi_p)$ is located in the i-th position of the Kronecker product. Let D_3 be the $d^p \times (p(p-1)/2)$ matrix obtained from D_1 such that the $p(p-1)/2$ columns of D_3 are all the possible products of two distinct columns from D_1.

Lastly, we define the $d^p \times d$ matrices $E_1, E_2, \cdots, E_p$ as follows: $E_i = (\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{id})$ with ε_{ij} a vector of length d^p formed by d^{p-i} repetitions of the vector

$$(0_{(j-1)\times d^{i-1}}, e_{d^{i-1}}, 0_{(j-1)\times d^{i-1}})$$

where 0_k and e_k are vectors of k zeros and ones, respectively.

Now the asymptotic joint distribution of the vector U_n of cell counts is given as follows:

Lemma 1. (Huffer and Park, 2002) If $Y_1, Y_2, \cdots, Y_n$ is a random sample from

$N_p(\mu, \Sigma)$ where Σ is nonsingular, then

$$(U_n - \frac{n}{d^p} e_{d^p})/\sqrt{n/d^p} \xrightarrow{d} N_{d^p}(0, A) \quad \text{as} \quad n \to \infty$$

where

$$A = I + e_{d^p} e_{d^p}^t (p-1)/d^p - \sum_{i=1}^{p} \{E_i E_i^t / d^{p-1}\} - D_3 D_3^t / d^{p-4}.$$

Since the limiting variance A of $V_n \equiv (U_n - n/d^p e)/\sqrt{n/d^p}$ does not depend on unknown parameter (μ, Σ), the Rao-Robson chi-squared test statistic is given by the form $V_n^t A^- V_n$, where A^- is a generalized inverse of A. By the general Wald's method (see p. 173 of Rao and Mitra (1971) for example), it is well known that

$$V_n^t A^- V_n \xrightarrow{d} X^2(\text{rank}(A)) \quad \text{as} \quad n \to \infty.$$

We will not use this result since we can easily provide a self-contained direct proof that it holds for the Moore-Penrose inverse $A^\dagger$ that is easy to compute. Here is main result on the Rao-Robson chi-squared test statistic:

Theorem 1. Under the assumption of Lemma 1, the Rao-Robson chi-squared test statistic is given by

$$T_n = V_n^t A^\dagger V_n = X^2 + \frac{d^{4-p}}{1-d^2 c} \sum_{k<l} \left(\sum_{i=1}^{p} \sum_{j=1}^{p} \psi_i \psi_j V_{ij}^{(kl)} \right)^2$$

and it has the limiting $X^2(d^p - 1 - p(d-1))$ distribution, where $c = (\sum_{i=1}^{p} \psi_i^2)^2$ and $V_{ij}^{(kl)}$ is the marginal sum of (i, j)-th category of (k, l)-th coordinate of V_n, i.e.

$$V_{ij}^{(kl)} = \sum_{S(i,j,k,l)} V_{n\pi} \text{ with } S(i, j, k, l) = \{\pi = (\pi_1, \cdots, \pi_p) : \pi_k = i, \pi_l = j\}.$$

Proof: Define $\Pi = I + e_{d^p} e_{d^p}^t (p-1)/d^p - \sum_{i=1}^{p} E_i E_i^t / d^{p-1}$ and $\Omega = D_3 D_3^t / (c d^{p-2})$, then, by simple algebra, it is easy to show that $\Pi\Omega = \Omega$ and that Π and Ω are idempotent matrices of ranks $d^p - 1 - p(d-1)$ and $p(p-1)/2$, respectively. Therefore the asymptotic variance A of V_n can be expressed as $\Pi(I - d^2 c \Omega)$ whose Moore-Penrose inverse in given by

$$(I - d^2 c \Omega)^{-1} \Pi^\dagger = \left(I + \frac{d^2 c}{1 - d^2 c} \Omega \right) \Pi = \Pi + \frac{d^2 c}{1 - d^2 c} \Omega.$$

This shows that

$$\begin{aligned} T_n &= V_n^t A^+ V_n = V_n^t \Pi V_n + \frac{d^2 c}{1-d^2 c} V_n^t \Omega V_n \\ &= V_n^t V_n + \frac{d^{4-p}}{1-d^2 c} V_n^t D_3 D_3^t V_n \\ &= X^2 + \frac{d^{4-p}}{1-d^2 c} \sum_{k<l} \left(\sum_{i=1}^{p} \sum_{j=1}^{p} \psi_i \psi_j V_{ij}^{(kl)} \right)^2, \end{aligned}$$

where the second equality holds since $\Pi V_n = V_n$ and the fourth equality holds since

$$V_n^t D_3 = (\sum_i \sum_j \psi_i \psi_j V_{ij}^{(12)}, \cdots, \sum_i \sum_j \psi_i \psi_j V_{ij}^{(p-1,p)}).$$

Now it remains to show that T_n has the limiting $X^2(d^p - 1 - p(d-1))$ distribution. Let $0 \le \lambda_1 \le \cdots \le \lambda_{d^p}$ be the eigenvalues of $A = I - (I - \Pi) - d^2 c \Omega$ and let $q_1, \cdots, q_{d^p}$ be the corresponding orthogonal eigenvectors. Since $(I - \Pi)\Omega = 0$ and $I - \Pi$ and Ω are orthogonal idempotent matrices of ranks $1 + p(d-1)$ and $p(p-1)/2$, it is easy to show that nonzero eigenvalues are 1 with multiplicity $d^p - 1 - p(d-1) - p(p-1)/2$ and $1 - d^2 c$ with multiplicity $p(p-1)/2$, i.e.

$$\lambda_1 = \cdots = \lambda_{1+p(d-1)} = 0, \lambda_{2+p(d-1)} = 1 - d^2 c, \lambda_{3+p(d-1)} = \cdots = \lambda_{d^p} = 1.$$

Let $Q = (q_1, \cdots, q_{d^p})$ and $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_{d^p})$, then $A = Q \Lambda Q^t$ by spectral decomposition and its Moore-Penrose inverse is given by $A^+ = Q \Lambda^+ Q^t$ where Λ^+ is the matrix obtained by replacing the nonzero elements of Λ by their reciprocals. Define $W_n = (\Lambda^+)^{1/2} Q^t V_n$, then W_n has the limiting $N_{d^p}(0, \mathrm{diag}(0, I_{d^p - 1 - p(d-1)}))$ by the continuous mapping theorem since V_n has the limiting $N_{d^p}(0, A)$ distribution. This completes the proof. □

3. Simulation and Application

We first provide a small simulation study to check accuracy in finite samples of the limiting distribution of our chi-squared test statistic. We have tried many configurations but almost all of them show almost the same results unless the number d^p of cells is small. Therefore we present the results for the case where $p = 3$, $d = 3$ so that the limiting distribution of the chi-square statistic is the chi-squared distribution with 20 degrees of freedom.

We consider four different sample sizes, $n = 27, 54, 96$ and 196, which have an average of 1, 2, 4 and 8 observations per cell, respectively.

For each sample size n, we generate 500 samples of size n from $N_3(0, I)$

and then calculate the chi-squared statistics for each of them. These 500 values are plotted against the corresponding quantiles of the limiting chi-squared

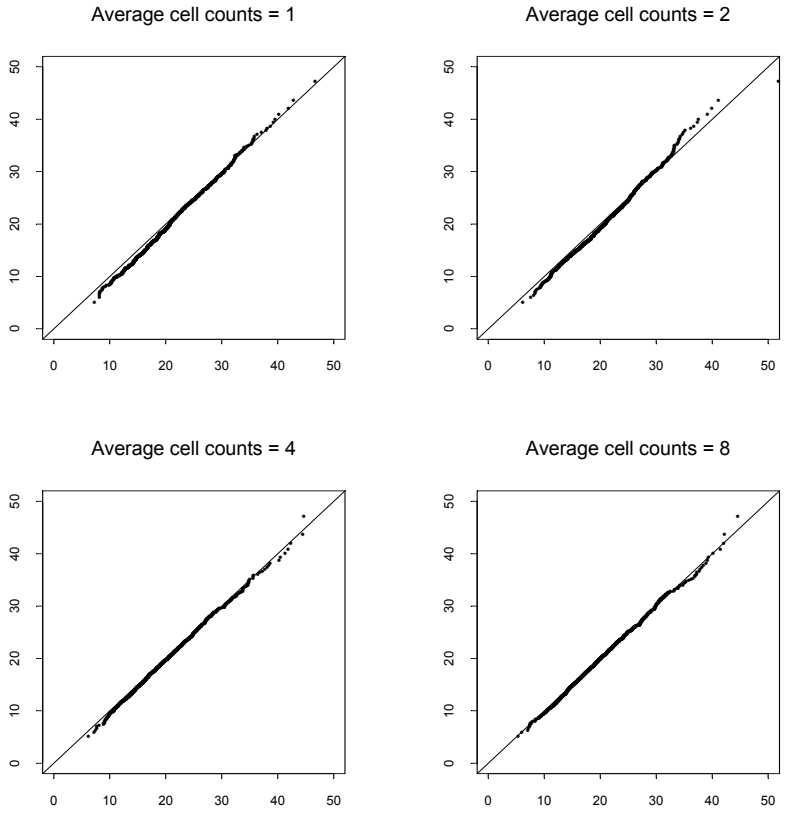

Figure 1. Chi-Squared Probability Plots

distribution. The resulting chi-squared probability plots are displayed in Figure 1.

Each plot displays the reference line with slope 1 and intercept 0, which corresponds to the ideal case where empirical and theoretical distributions coincide. Examining the plots, we see that the limiting distribution is an excellent approximation for the cases where average cell counts is 4 and 8 and that it is reasonably good for the cases where average cell counts is 1 and 2.

We next provide an illustrative example of application to geyser data. There are two time series in geyser data and the waiting time between eruptions is used for our example (see Azzalani and Bowman (1990) for details). We compare the power of our method with those of three popular tests for multivariate normality. The skewness and kurtosis tests of Mardia (1970) and the Q_n test (with Cholesky implementation) of Ozturk and Romeu (1992) have been selected for comparison

since they performed quite well in an extensive simulation study of Romeu and Ozturk (1993). We will examine the multivariate structure of the residuals from fitting a time series model to the waiting time.

We use an automatic procedure called **AR** in S-Plus to select one of the best autoregressive models based on the Akaike information criterion. By using Yule-Walker equations to estimate the autoregression coefficients, the procedure have chosen AR(2) model. Both autocorrelation and partial autocorrelation functions are well inside the error bars up to 25 lags and the value of Shapiro and Wilk (1965) W test statistic is 0.9822 with p-value 0.3654. Therefore we could not find unusual pattern by the usual time series diagnostics and a univariate test of normality.

We now examine the residuals by our method and other competitors for multivariate normality. To obtain multivariate observations, we divide the residuals into subseries of three consecutive residuals and take each subseries as an observation. In this way, we obtain 99 multivariate observations. Our method with $d=3$ leads to chi-squared value 34.46 with an (asymptotic) p-value .0232 and $d=4$ leads to chi-squared value 78.01 with an (asymptotic) p-value .0179. Therefore, our method signals there are some multivariate structure in the residuals. However, other competitors could not detect this structure: Marida's skewness and kurtosis have (asymptotic) p-values 0.160 and 0.569, respectively and the test based on Q_n has (asymptotic) p-value 0.601. Therefore, in this example, our test has more power than other tests.

References

1. Azzalini, A., and Bowman, A.W. (1990). A Look at Some Data on the Old Faithful Geyser, *Applied Statistics,* 39, 357-365.
2. Chernoff, H., and Lehmann, E.L. (1954). The Use of Maximum Likelihood Estimates in χ^2 Tests for Goodness of Fit, *Annals of Mathematical Statistics,* 25, 579-586.
3. Farebrother, R.W. (1990). The Distribution of a Quadratic Form in Normal Variables, *Applied Statistics,* 39, 294-309.
4. Huffer, F.W. and Park, C. (2002). The Limiting Distribution of a Test for Multivariate Structure, *Journal of Statistical Planning and Inference,* 105, 417-431.
5. Imhof, J.P. (1961). Computing the Distribution of Quadratic Forms in Normal Variables, *Biometrika,* 48, 419-426.
6. Mardia, K.V. (1970). Measures of Multivariate Skewness and Kurtosis with Applications, *Biometrika,* 57, 519-530.
7. Nikulin, M.S. (1973). Chi-Square Test for Continuous Distributions with

Shift and Scale Parameters, *Theory of Probability and Its Applications,* 18, 559-568.
8. Ozturk, A. and Romeu, J.L. (1992). A New Method for Assessing Multivariate Normality with Graphical Applications, *Communications in Statistics - Simulations and Computation,* 21, 15-34.
9. Rao, C.R., and Mitra, S.K. (1971). *Generalized Inverse of Matrices and Its Applications,* John Wiley & Sons, New York.
10. Rao, C.R., and Robson, D.S. (1974). A Chi-Square Statistic for Goodness-of-Fit Tests within the Exponential Family, *Communications in Statistics,* 3, 1139-1153.
11. Romeu, J.L. and Ozturk, A. (1993). A Comparative Study of Goodness-of-fit Tests for Multivariate Normality. *Journal of Multivariate Analysis,* 46, 309-34.
12. Shapiro, S.S. and Wilk, M.B. (1965). An Analysis of Variance Test for Normality (Complete Samples), *Biometrika,* 52, 591-611.
13. Solomon, H., and Stephens, M.A. (1977). Distribution of a Sum of Weighted Chi-Square Variables, *Journal of the American Statistical Association,* 72, 881-885.

[received date : Sep. 2003, accepted date : Nov. 2003]