Journal of Computational and Applied Mathematics 235 (2011) 4729–4735

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

A JAVA program for the multivariate Z_p and C_p tests and its application

Güvenç Arslan [a,*], İlknur Özmen [b]

[a] Department of Mathematics, Izmir University of Economics, 35330, Izmir, Turkey

[b] Department of Statistics and Computer Sciences, Baskent University, 06810, Ankara, Turkey

ARTICLE INFO

Article history:
Received 18 November 2009
Received in revised form 30 June 2010

Keywords:
Multivariate normality
Z_p and C_p test statistics
Simulation
JAVA program

ABSTRACT

The multivariate normality assumption is used in many multivariate statistical analyses. It is, therefore, important to assess the validity of this assumption. The main aim of this study is to develop a JAVA program for applying the recently developed Z_p and C_p test statistics. The application and results of the program are illustrated on two real data sets.

1. Introduction

Both the univariate normality and the multivariate normality assumptions are common and important in many statistical models and methodologies for data analysis. There are many varieties of tests for univariate normality in the literature and many software packages (e.g., SPSS, Minitab, SAS) are available. Unfortunately the application of multivariate normality tests is still limited or even not available in many software packages. One of the software packages that contains a multivariate normality test is the R software, which includes the p-variate version of the Shapiro–Wilk statistic W.

The problem of utilization of the univariate normality tests to assess multivariate normality has been studied by many researchers in recent years. Liang et al. [1], for example, developed three simple quantile–quantile (Q–Q) plots for providing supplementary evidence in detecting a possible departure from the multivariate normality assumption in high-dimensional data analysis. They illustrate how to employ the plots in practice on the Iris data. Szekely and Rizzo [2] proposed a new test of multivariate normality when population parameters are estimated from the sample and present Monte Carlo power comparisons to assess the empirical power performance of the new test. Sürücü [3] gives the results of a simulation study of the power properties of some of the prominent goodness of fit tests (Shapiro–Wilk statistic W, correlation statistic R, combined statistic C, Anderson–Darling statistic $\hat{A}$, Kolmogorov–Smirnov statistic $\hat{D}$, Tiku statistic Z^*). Sürücü [4] examined the Z_p, C_p, R_p, W_p statistics and measure of skewness $b_{1,p}$ for testing multivariate normality and has investigated their power properties by simulation. In addition he has tested the introduced multivariate normality tests on the Iris Setosa plants' data. Liang et al. [1] proposed a new way to generalize the Shapiro–Wilk statistic for testing high-dimensional normality with small sample size (n). They present Monte Carlo studies to investigate the empirical performance of a generalized W statistic for the cases of small n, and give applications of the generalized W statistic on two real data sets.

Some contributions of this paper are: (a) Development of a JAVA program for testing a possible departure from multivariate normality assumption based on the Z_p and C_p statistics. (b) Direct calculation of the expected values of the standardized normal order statistics ($\mu_{i:n}$). (c) Direct calculation of the values of the coefficients $a_{i:n}$ in the W statistic. (d) The developed JAVA program for testing multivariate normality is easy to use and apply. We hope that this program will be helpful for interested users in applications of multivariate normality tests since similar tests in software packages are quite limited.

* Corresponding author.
E-mail addresses: guvenc.arslan@izmirekonomi.edu.tr, guvenc.arslan@gmail.com (G. Arslan).

0377-0427/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2010.08.035

After presenting the univariate W, R, Z and C statistics in Section 2 the multivariate Z_p and C_p statistics are presented in Section 3. Section 4 presents some of the computations which are necessary for the multivariate Z_p and C_p statistics. An application of the JAVA program for multivariate normality on two real data sets is then given in Section 5.

2. The univariate W, R, Z and C statistics

Suppose $x_1, \ldots, x_n$ is a random sample from a normal distribution $N(\mu, \sigma^2)$, with unknown parameters μ and σ^2, and let $x_{1:n} \leq \cdots \leq x_{n:n}$ denote the order statistics of the observed values of this random sample.

The Shapiro–Wilk [5] statistic W is defined by

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{i:n}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad 0 < W < \infty, \tag{2.1}$$

where $\sum_{i=1}^{n} a_i^2 = 1$, and $\bar{x} = \sum_{i=1}^{n} x_i/n$. The values of the coefficients a_i $(1 \leq i \leq n)$ are tabulated for $n < 50$ in [6]. In this study, however, the a_i (for any $n > 1$) are calculated by the program; details are given in Section 4. Note that $\sum_{i=1}^{n} a_i x_{i:n}$ is actually the best linear unbiased estimator (BLUE) of σ. Small values of W lead to the rejection of the univariate normality assumption. We also note that the null distribution of the W statistic is not known.

The correlation statistic R is given by

$$R = 1 - \hat{\rho}^2, \quad 0 < R < 1 \tag{2.2}$$

where $\hat{\rho}$ is the estimated value of the product moment correlation coefficient between $x_{i:n}$ and $\mu_{i:n}$ [7,8]. $x_{i:n}$, $(1 \leq i \leq n)$ are the order statistics of a random sample of size n from the normal distribution and $z_{i:n} = (x_{i:n} - \mu)/\sigma$ are the corresponding standardized normal order statistics. $\mu_{i:n}$, $(1 \leq i \leq n)$ are the expected values of the standardized normal order statistics, i.e. $E[z_{i:n}]$. The correlation coefficient statistics can be used for testing any assumed density of type $(1/\sigma)f((x - \mu)/\sigma)$. For ease of computation, in several studies, the $\mu_{i:n}$ are obtained by the population quantiles

$$\mu_{i:n} = F_0^{-1}(i/n + 1), \quad (1 \leq i \leq n) \tag{2.3}$$

where $F_0(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{z^2}{2}} dz$ is the cumulative distribution function of the standard normal distribution. The expected values of the standardized normal order statistics $(\mu_{i:n})$ can be obtained from Harter [9]. Details on how the $\mu_{i:n}$ are calculated in the program are given in Section 4. Note that the null distribution of R is not known, and therefore, its percentage points are generally determined empirically by a Monte Carlo simulation. Large values of R lead to the rejection of the normality assumption.

The [10,11] statistic is defined by

$$Z = \frac{2\sum_{i=1}^{n-1}(n - 1 - i)G_i}{(n-2)\sum_{i=1}^{n-1} G_i}, \quad 0 < Z < \infty \tag{2.4}$$

where

$$G_i = \frac{x_{i+1:n} - x_{i:n}}{\mu_{i+1:n} - \mu_{i:n}} \tag{2.5}$$

are the generalized sample spacings. Like R, the Z statistic can be used for testing any location–scale distribution. For large n (≥ 10), the null distribution of Z is normal with common variance V, i.e. $N(1, V)$. The common variance V depends only on n. However, V is not well approximated unless n is very large $(n > 100)$. The values of V are, therefore, obtained by simulation. Some values of $\sqrt{V}$, obtained by simulation for some n (10, 20, 30, 40, 50, 70, and 100) for testing normality are given by Sürücü [4]. For intermediate values of n, the values $\sqrt{V}$ can be obtained by linear interpolation. In this study, values for V are obtained by simulation for any $n(> 1)$.

The Shapiro–Wilk statistic W is known to be overall the most powerful test against skew and short-tailed symmetric alternatives whereas the correlation statistic R is known to be overall the most powerful against long-tailed symmetric alternatives. Sürücü [4,3] proposed a new statistic defined by

$$C = 1 - \{[1 + \alpha_1(\alpha_2 - 1)W + \alpha_1(1 - \alpha_2)(1 - R)]\}, \tag{2.6}$$

where $\alpha_1 > 0, \alpha_2 < 1$. The coefficients α_1 and α_2 are calculated from the equations

$$\alpha_1 = \exp(-(\gamma_1/0.6)^5) \quad \text{and} \quad \alpha_2 = \exp(-(\gamma_2/3.5)^5). \tag{2.7}$$

$\sqrt{\gamma_1}$ and γ_2 are the sample skewness and kurtosis, respectively. The coefficients α_1 and α_2 are determined empirically to achieve the overall highest power. The combined statistic C is a weighted sum of the Shapiro–Wilk statistic W and the correlation statistic R, the weights being determined by the sample skewness and kurtosis.

Since distributions of W and R are not known, [3] uses a four moment F approximation to the null distribution of the C statistic. Large values of C lead to the rejection of univariate normality.

3. Multivariate Z_p and C_p statistics

The multivariate statistics Z_p, C_p, R_p and W_p are p-variate versions of the univariate Z, C, R and W statistics, respectively. Sürücü [4] has compared the powers of the four statistics (Z_p, C_p, R_p, W_p) and also of the measure of skewness $b_{1,p}$, for $p = 2$, $p = 4$ and $n = 10$, $n = 20$, $n = 50$ at a 10% significance level for different families of alternative distributions. He used Monte Carlo simulation with a size of 100 000 and showed that the C_p statistic is overall the most powerful and effective test against skew, long-tailed as well as short-tailed symmetric alternatives. He also showed that the Z_p statistic is particularly powerful against skew alternatives.

Although there are many kinds of tests for multivariate normality in the literature (e.g., measures of multivariate skewness and kurtosis, W_p, R_p, Q–Q plots), we developed a JAVA program to apply the recently developed powerful Z_p and C_p statistics.

The Z_p statistic is a p-variate version of the univariate statistic Z based on sample spacings [10,11]. The C_p statistic is a p-variate version of the univariate statistic C introduced by Sürücü [4,3].

The initial step in both of the multivariate statistics Z_p and C_p is to transform the $X_1, \ldots, X_p$ random variables having a p-variate normal distribution by the following linear combinations

$$
\begin{aligned}
&X_1\\
&X_2 - \beta_{2.1}X_1\\
&X_3 - \beta_{31.2}X_1 - \beta_{32.1}X_2\\
&X_4 - \beta_{41.23}X_1 - \beta_{42.13}X_2 - \beta_{43.12}X_3\\
&\vdots\\
&X_p - \beta_{p1.q_1}X_1 - \beta_{p2.q_2}X_2 - \cdots - \beta_{p(p-1).q_{p-1}}X_{p-1}.
\end{aligned}
$$

We denote the partial regression coefficients of X_k on X_j with the other $(p-2)$ variables held fixed by $\beta_{kj.q_j}$, where q_j denotes variables other than those in the primary subscripts. The estimations of the partial regression coefficients can be calculated from either ordinary least squares or using Eq. (3.1).

$$\hat{\beta}_{kj.q_j} = -\frac{\sigma_k}{\sigma_j}\frac{C_{kj}}{C_{kk}}. \tag{3.1}$$

In Eq. (3.1), σ_k and σ_j are the standard deviations of X_i and X_j, respectively, and C_{kj} is the cofactor of the (k, j)th element in the correlation matrix [6].

For a random sample of size $n(x_{1i}, \ldots, x_{pi}, 1 \le i \le n)$, we consider the random observations which are uncorrelated with each other as follows

$$
\begin{aligned}
&y_{1i} = x_{1i}\\
&y_{2i} = x_{2i} - \hat{\beta}_{2.1}x_{1i}\\
&y_{3i} = x_{3i} - \hat{\beta}_{31.2}x_{1i} - \hat{\beta}_{32.1}x_{2i}\\
&y_{4i} = x_{4i} - \hat{\beta}_{41.23}x_{1i} - \hat{\beta}_{42.13}x_{2i} - \hat{\beta}_{43.12}x_{3i}\\
&\vdots\\
&y_{pi} = x_{pi} - \hat{\beta}_{p1.q_1}x_{1i} - \hat{\beta}_{p2.q_2}x_{2i} - \cdots - \hat{\beta}_{p(p-1).q_{p-1}}x_{(p-1)i}.
\end{aligned}
$$

The p-variate version of the Z_p and C_p are obtained by applying the Z_j and $C_j (1 \le j \le p)$ statistics to the corresponding variables in the data set.

The Z_p statistic is given by

$$Z_p = \sum_{j=1}^{p}\left(\frac{Z_j - 1}{\sqrt{V}}\right)^2, \quad 0 < Z_p < \infty.$$

The null distributions of the Z_j random variables are asymptotically ($n \to \infty$) normal with $N(1,\ V)$. Large values of Z_p lead to the rejection of multivariate normality. The null distribution of Z_p is asymptotically chi-square with p degrees of freedom.

To determine the accuracy of the asymptotic distribution for small n, [4] gives some simulated values, based on 10 000 Monte Carlo runs, of the probability

$$P(Z_p \geq \chi^2_{0.90;p}).$$

Here $\chi^2_{0.90;p}$ denotes the 90th percentile of a chi-square distribution with p degree of freedom ($p = 2$).

Since goodness of fit tests are usually performed at a 10% significance level, Sürücü does not reproduce values for any other significance level. In this study we also used the same level of significance; that is $\alpha = 0.10$.

The C_p statistic is given by

$$C_p = \sum_{j=1}^{p} C_j, \quad 0 < C_p < \infty.$$

Large values of C_p lead to the rejection of multivariate normality. Sürücü [4] used a three moment chi-square approximation to the null distribution of C_p as follows:

Let μ'_1 be the mean of a positive random variable X, and μ_2, $\mu_3 (> 0)$, μ_4 be its variance, third and fourth central moments, respectively. If the skewness coefficient $\sqrt{\Gamma_1} = \mu_3/\mu_2^{3/2}$ is positive, and together with the kurtosis coefficient $\Gamma_2 = \mu_4/\mu_2^2$ satisfies the condition $|\Gamma_2 - (3 + 1.5\,\Gamma_1)| \leq 0.5$, then $\chi^2_v = \frac{X+a}{b}$ gives a remarkable accurate approximation to the upper percentage points of X (see [12,13]).

The null distribution of C_p is

$$\chi^2_v = \frac{C_p + a}{b}, \tag{3.2}$$

where χ^2_v has a central chi-square distribution with v degrees of freedom. The values of a, b and v are determined by equating the first three moments on both sides of (3.2);

$$v = \frac{8}{\Gamma_1}, \qquad b = \sqrt{\frac{\mu_2}{2v}} \quad \text{and} \quad a = bv - \mu'_1$$

Sürücü [4] gives some simulated values, based on 10 000 Monte Carlo runs, of the probability

$$P(C_p \geq b\chi^2_{0.90;v} - a). \tag{3.3}$$

Here $\chi^2_{0.90;v}$ denotes the 90th percentile of a three moment chi-square distribution with v degree of freedom ($p = 2$).

4. Some computations for the multivariate Z_p and C_p statistics

In both of the multivariate tests one has to do various intensive computations and apply some simulation techniques. For example, in certain steps one needs to calculate the expected values of the standardized normal order statistics ($\mu_{i:n}$). Though one can use several approximations or tables we calculated these values by using numerical integration; see for example [9] for some traditional approaches and [4] for some other approaches to calculate these values. However, by using numerical integration it is possible to apply the tests for data of any size ($n \geq 10$) and any dimension ($p \geq 2$). The restriction of $n \geq 10$ is used because power studies in [4] show that these tests give accurate results for $n \geq 10$. In this section it is shown how some of the calculations have been implemented in the program. These calculations make it possible to apply these recently developed powerful tests in real applications. In addition, we note that necessary calculations for some needed values of the standard normal distribution, the chi-square distributions and the first four moments are done by using numerical integration. These calculations are taking most of the computation time when the tests are applied.

The partial regression coefficients ($\hat{\beta}_{kj.q_j}$) can be calculated by the program for any $n \geq 2$ and any $p \geq 2$ by using the ordinary least squares formula. Similarly, the common variance (V) used in the Z_p statistic can be calculated by the program for any $n \geq 2$. In addition, simulated values of the probabilities (under normality assumption) in Eq. (3.3), showing the accuracy of the asymptotic distributions, for any $p \geq 2$, can also be obtained if desired by the user.

The formula for the expected values of the standardized normal order statistics, $\mu_{i:n}$ ($1 \leq i \leq n$), is

$$\mu_{i:n} = E[z_{i:n}] = i\binom{n}{i}\int_{-\infty}^{\infty} x[F_0(x)]^{i-1}[1-F_0(x)]^{n-i} f_0(x)\mathrm{d}x. \tag{4.1}$$

Since the integrand in formula (4.1) is very close to 0 for $|x| > 7.6$ (see [9]) the trapezoidal rule for numerical integration can be applied to evaluate these values. The computational results show that this gives accurate results up to at least four decimal places. For given n only half of the values need to be calculated because of symmetry of the $\mu_{i:n}$. We also note that the $\mu_{i:n}$ values are used in the univariate W, R, Z and C tests.

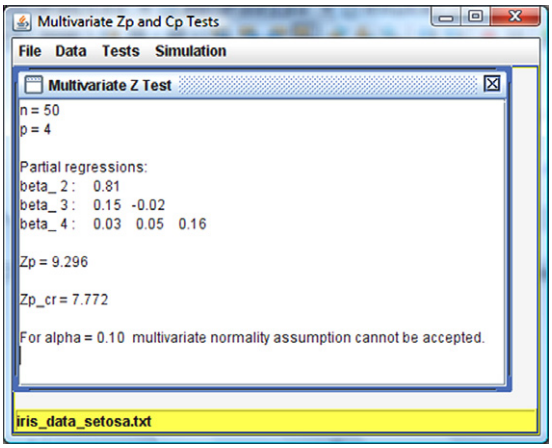

Fig. 1. Z_p test applied to the Iris Setosa data set.

In this study the values of the Shapiro–Wilk coefficients a_i $(1 \leq i \leq n)$ are calculated for any n greater than 2. For given n let $\boldsymbol{a}^T = (a_1, a_2, \ldots, a_n)$ denote the Shapiro–Wilk coefficients. Then these coefficients are given by the following formula

$$\boldsymbol{a}^T = \frac{\boldsymbol{m}^T \boldsymbol{\Sigma_0}^{-1}}{(\boldsymbol{m}^T \boldsymbol{\Sigma_0}^{-1} \boldsymbol{\Sigma_0}^{-1} \boldsymbol{m})^{\frac{1}{2}}}$$

where $\boldsymbol{m}^T = (m_1, m_2, \ldots, m_n)$, $m_i = \mu_{i:n}$ and $\boldsymbol{\Sigma_0}$ is the variance–covariance matrix of the standardized normal order statistics $Z_{i:n}$ and $Z_{j:n}$.

Using the formula $(\boldsymbol{\Sigma_0})_{ij} = E[Z_{i:n}Z_{j:n}] - \mu_{i:n}\mu_{j:n}$ the covariances can be calculated by computing the $E[Z_{i:n}Z_{j:n}]$ values.

$$E[Z_{i:n}Z_{j:n}] = C_{ij}^n \iint_D z_i z_j [F_0(z_i)]^{i-1} [F_0(z_j) - F_0(z_i)]^{j-i-1} [1 - F_0(z_j)]^{n-j} f_0(z_i) f_0(z_j) \mathrm{d}z_i \mathrm{d}z_j$$

$$C_{ij}^n = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}$$

$$D = \{(z_i, z_j) : -\infty < z_i < z_j < \infty\}.$$

The $E[Z_{i:n}Z_{j:n}]$ values, on the other hand, can be calculated by numerical integration of this double integral with consideration of the region D. The calculation of the matrix $\boldsymbol{\Sigma_0}$ is probably the most time consuming part in all of the computations.

5. Applications

To demonstrate the application of the developed program some examples from the literature have been used. The first data set is a part of the Iris data set – Iris Setosa – which has been used in several studies on multivariate normality tests. This data set consists of 50 examples with four variables: sepal length, sepal width, petal length, petal width. It is known that this data set is not normally distributed. An example output for testing the joint multivariate normality assumption using the program is shown in Fig. 1.

The C_p test leads to the same conclusion as the Z_p test. Actually the C_p test value is 0.143 while the critical value of the test is calculated as 0.135 leading to the rejection of the multivariate normality assumption.

As another example a subset of the data set examined in [14] is considered. This data set has been used as an example by Liang et al. [15] for testing multivariate normality for small n and high dimensionality $(n \leq p)$. Since [4] showed that the assumptions for the null distributions of the tests used in the developed program are quite accurate for $n \geq 10$ and since the data set contains censored data, only a subset has been used. The part of the data set that has been used in the program is given in Table 1. This part actually consists of the data that is not censored. For details about the full data set one may refer to [14].

The output for the C_p test for testing the joint multivariate normality assumption is shown in Fig. 2. Again the Z_p test leads to the same conclusion as the C_p test. The Z_p test value is 5.486 while the critical value of the test is calculated as 10.642 which means that the multivariate normality assumption cannot be rejected.

Liang et al. [15] proposed a generalized Shapiro–Wilk W statistic for testing high-dimensional normality. They conclude that the mice data set can be assumed to be normal for $(n, p) = (8, 11)$ and $(n, p) = (5, 12)$. Note that the results of this study support this conclusion.

Table 1
Uncensored part of the data from Tan et al. [14].

Group	Mouse	Weeks					
		0	1	2	3	4	5
I	1	2.34	2.48	2.04	1.06	1.26	0.91
	2	1.11	1.54	0.81	0.93	1.37	1.00
	3	0.96	0.99	0.99	0.53	0.72	0.36
	4	0.66	0.60	0.49	0.78	1.40	1.33
	5	2.08	2.15	1.97	0.83	0.78	0.26
II	6	1.09	1.04	0.76	0.77	0.72	0.29
	7	0.74	0.93	0.83	0.59	0.60	0.41
	8	0.94	1.12	1.67	2.69	3.51	2.77
	9	1.84	1.99	2.75	4.29	6.41	4.04
	10	1.21	1.41	1.97	2.07	2.98	2.30
	11	1.24	1.32	1.63	2.43	3.00	2.04

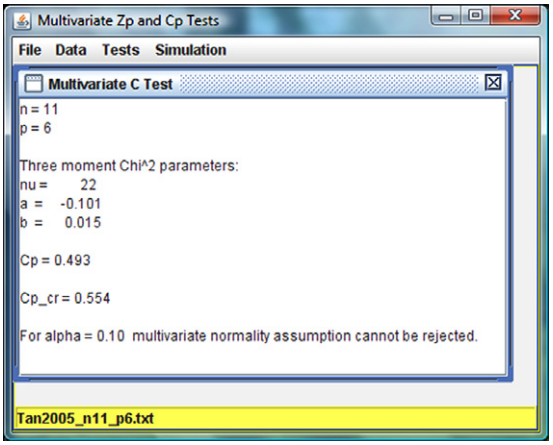

Fig. 2. *Cp* test applied to the mice data set in [14].

6. Conclusions

Since the Z_p and C_p tests are one of the most powerful tests for testing multivariate normality, it is important to have necessary tools in order to apply these tests. We hope that this study will be of help in developing more advanced tools for testing multivariate normality which is of great importance in many statistical methods and applications. In such a tool one should also include tests which have been shown to be valid also for the case of small sample size and/or high dimensionality such as in [14,15].

Remark. Since the size of the source code is quite large and would not fit into several pages they are not given in the paper. Interested readers may obtain the program and source code by e-mail from the corresponding author.

Acknowledgements

The authors thank the referees for their comments and suggestions which helped to improve the presentation of this paper.

References

[1] J. Liang, W.S.Y. Pan, Z.H. Yang, Characterization based *Q*–*Q* plots for testing multinormality, Statistics & Probability Letters 70 (2004) 183–190.
[2] G.J. Szekely, M.L. Rizzo, A new test for multivariate normality, Journal of Multivariate Analysis 93 (2005) 58–80.
[3] B. Sürücü, A power comparison and simulation study for goodness-of-fit tests, Computers & Mathematics with Applications 56 (2008) 1617–1625.
[4] B. Sürücü, Goodness-of-fit tests for multivariate distributions, Communications in Statistics — Theory and Methods 35 (2006) 1319–1331.
[5] S.S. Shapiro, M.B. Wilk, An analysis of variance test for normality (complete samples), Biometrika 52 (1965) 591–611.
[6] E.S. Pearson, H.O. Hartley, Biometrika Tables for Statisticians, vol. 2, Cambridge University Press, Cambridge, 1972.
[7] J.J. Filliben, The probability plot correlation coefficient test for normality, Technometrics 17 (1975) 111–117.
[8] R.M. Smith, L.J. Bain, Correlation type goodness of fit statistics with censored sampling, Communications in Statistics A5 2 (1976) 119–132.
[9] H.L. Harter, Expected values of normal order statistics, Biometrika 48 (1961) 151–165.
[10] M.L. Tiku, Goodness-of-fit statistics based on the spacing of complete or censored samples, Australian Journal of Statistics 22 (1980) 260–275.
[11] M.L. Tiku, Order statistics in goodness-of-fit tests, Communications in Statistics — Theory and Methods 17 (1988) 2369–2387.

[12] E.S. Pearson, Note on an approximation to the distribution of non-central χ^2, Biometrika 46 (1959) 364.
[13] M.L. Tiku, Chi-square approximations for the distributions of goodness of fit statistics U_N^2 and W_N^2, Biometrika 52 (1963) 630–633.
[14] M. Tan, H.-B. Fang, G.-L. Tian, G. Wei, Testing multivariate normality in incomplete data of small sample size, Journal of Multivariate Analysis 93 (2005) 164–179.
[15] J. Liang, M.L. Tang, P.S. Chan, A generalized Shapiro–Wilk W statistic for testing high-dimensional normality, Computational Statistics & Data Analysis 53 (2009) 3883–3891.