应用概率统计　第 36 卷
第 1 期　2020 年 2 月

Chinese Journal of Applied Probability and Statistics
Feb., 2020, Vol. 36, No. 1, pp. 41-58
doi: 10.3969/j.issn.1001-4268.2020.01.004

Empirical Distribution Function Based Statistics for Testing High Dimensional Normality [*]

CUI Jiarong
(*School of Statistics, Renmin University of China, Beijing, 100872, China*)

ZHU Fengyi
(*School of Mathematical Sciences, Peking University, Beijing, 100871, China*)

LIU Jiamin　　XU Wangli[⋆]
(*Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing, 100872, China*)

Abstract: Kolmogorov-Smirnov (KS), Cramer-von Mises (CM) and Anderson-Darling (AD) test, which are based on empirical distribution function (EDF), are well-known statistics in testing univariate normality. In this paper, we focus on the high dimensional case and propose a family of generalized EDF based statistics to test the high-dimensional normal distribution by reducing the dimension of the variable. Not only can we approximate the corresponding critical values of three statistics by Monte Carlo method, we also can investigate the approximate distributions of proposed statistics based on approximate formulas in univariate case under null hypothesis. The Monte Carlo simulation is carried out to demonstrate that the performance of proposed statistics is more competitive than existing methods under some alternative hypotheses. Finally, the proposed tests are applied to real data to illustrate their utility.

Keywords: high-dimensional, eigenprojection, EDF based test

2010 Mathematics Subject Classification: 62G10; 62G30

Citation: CUI J R, ZHU F Y, LIU J M, et al. Empirical distribution function based statistics for testing high dimensional normality [J]. Chinese J Appl Probab Statist, 2020, 36(1): 41–58.

§1. Introduction

The population distribution is the fundamental of statistical modelling and inference. Normal distribution is often assumed in statistical analysis. However, if this assumption is not valid, the statistical inference might lead to wrong conclusions. Thus, normality

[*] The project was supported by the National Natural Science Foundation of China (Grant No: 11971478) and the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities (Grant No: 16JJD910002).

[⋆] Corresponding author, E-mail: wlxu@ruc.edu.cn.

Received April 27, 2018. Revised December 5, 2018.

testing is crucial for both statistical theory and application. There are many literatures investigating the univariate normality testing, and many researchers have made efforts to generalize the univariate normality tests to multivariate or high dimensional cases.

For multivariate normality testing, some statistics are constructed by skewness and kurtosis coefficients or empirical characteristic function (ECF). Among others, Mardia[1] constructed a test of multivariate normality based on the proposed measures of multivariate skewness and kurtosis. Srivastava[2] used principal component analysis (PCA) to develop these two measures in [1]. Csorgo[3] established a maximal deviation statistic to test the composite hypothesis of p-variate normality by Mahalanobis transform of ECF. Baringhaus and Henze[4] proposed an omnibus test statistic by weighted integral of the squared modulus of the difference between the ECF and the characteristic functions of the normal distribution. On the other hand, the empirical distribution function (EDF) based statistics, such as Kolmogorov-Smirnov (KS), Cramer-von Mises (CM) and Anderson-Darling (AD) test, are well-known approaches to test univariate normality. Many literatures have investigated the multivariate normality testing based on these three statistics.

Kolmogorov[5] first proposed KS statistic based on the discrepancy between the empirical distribution F_n and the hypothesized normal distribution Φ in univariate case. For multivariate case, Rosenblatt[6] suggested the generalized discrepancy should be measured not only in joint distribution but in all marginal distributions, then constructed a multivariate KS statistic based upon the Rosenblatt's transformation. Justel et al.[7] used Rosenblatt's transformation to introduce a multivariate KS statistic which is distribution free, and developed an algorithm in the bivariate case. Peacock[8] proposed a different discrepancy between F_n and Φ in two dimensional case, which is based on the different definition of cumulative probability function, and applied the largest differences to construct KS statistic. Fasano and Franceschini[9] improved [8]'s statistic and generalized it to three dimensions.

The CM statistic were constructed based on the integral of the squared difference between F_n and Φ by Cramer[10] and von Mises[11] first. Afterwards, Anderson and Darling[12, 13] constructed the AD statistic by incorporating a weight function into CM statistic. For multivariate case, many literatures made improvements and generalizations for CM and AD statistics. Koziol[14] constructed a new CM type statistic based on the eigenvalues and eigenvector of covariance kernel and investigated its asymptotic distribution. Chiu and Liu[15] modified the CM statistic by incorporating the conception of L_p-star discrepancy. Lewis[16] studied the asymptotic distribution of AD statistic by utilizing empirical approximation based on Monte Carlo simulation. As for theoretical approximation, Giles[17] introduced a standard saddlepoint approximation of AD statistic which performs well in both tails of the distribution.

In high dimensional case, some literatures utilized approaches of dimension reduction to investigate the multivariate normality testing. For example, Srivastava and Hui[18] used PCA first and computed Shapiro-Wilk statistic (SW) for each principal component. Liang et al.[19] proposed the generalized SW statistic by projecting the high-dimensional sample onto some selected eigenvector directions. In this paper, inspired by [19], we develop a new way to generalize EDF based statistics including KS, CM, AD statistics to high-dimensional normality test, the corresponding generalized EDF based statistics were constructed based on each selected direction under null hypothesis by projecting the high-dimensional sample matrix onto some selected eigenvector directions.

The paper is organized as follows. Section 2 provides the theoretical methods for constructing the generalized EDF based statistics for testing high-dimensional distribution and gives its approximate null distributions. In Section 3, Monte Carlo studies are carried out to investigate the empirical performance of the generalized EDF based statistics, and applications on two real data sets are given. We conclude this paper with a brief discussion in the last section.

§2. The Proposed Statistics and the Corresponding Properties

Let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ be an i.i.d. sample with dimension p, and $F(\cdot)$ be the corresponding cumulative distribution function. The hypothesis we consider is

$$H_0: \text{The sample } \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\} \text{ is from multivariate normal distribution } \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1}$$

versus the alternative hypothesis that the sample is not from multivariate normal distribution. The idea is to reduce the p dimension of variable $\boldsymbol{x}$ to one dimension by dimensionality reduction method, then classical method for testing univariate normality, including KS statistic, CM statistic, AD statistic, can be applied. The following theorem from [19] is introduced for dimensionality reduction.

Theorem 1 Let $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)'$ be the $n \times p$ observation matrix, and $\boldsymbol{A}$ is a $(n-1) \times n$ constant matrix satisfying $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}_{n-1}$ and $\boldsymbol{A}\boldsymbol{1}_n = \boldsymbol{0}$, where $\boldsymbol{I}_{n-1}$ stands for a identity matrix with $(n-1)$ dimension and $\boldsymbol{1}_n$ stands for a n dimension column vector of 1.

Define a matrix $\boldsymbol{Y} = \boldsymbol{AX} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{n-1})'$, based on theory of eigenvalue-eigenvector, we have

$$\Big(\frac{1}{n-1}\boldsymbol{Y}'\boldsymbol{Y}\Big)\boldsymbol{D} = \boldsymbol{DH},$$

where $\boldsymbol{H} = \mathrm{diag}(h_1, h_2, \cdots, h_p)$ with h_i $(i = 1, 2, \cdots, p)$ being the eigenvalues of $(n-1)^{-1}\boldsymbol{Y}'\boldsymbol{Y}$, and $\boldsymbol{D} = (\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_p)$ with $\boldsymbol{d}_i$ $(i = 1, 2, \cdots, p)$ being the corresponding unit eigenvectors. Without loss of generality, we assume $h_1 \geqslant h_2 \cdots \geqslant h_r > h_{r+1} = \cdots = h_p = 0$, and define the random vectors $\boldsymbol{z}_i = \boldsymbol{Y}\boldsymbol{d}_i$, $i = 1, 2, \cdots, r$. Under the null hypothesis in (1), we have the following two conclusions:

(i) For any finite n, random vector $\boldsymbol{z}_i$ $(i = 1, 2, \cdots, r)$ is a spherical distribution, and the asymptotical distribution of $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \cdots, z_{i(n-1)})'$ is $\mathrm{N}(\boldsymbol{0}, h_i\boldsymbol{I}_{n-1})$;

(ii) Under the condition that $p/n \to 0$, $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_r$ are asymptotically independent.

The proof of Theorem 1 can be found in [20] and we omit the details here. According to Theorem 1, we apply eigenvalue-eigenvector method for dimension reduction to obtain the corresponding random vectors $\boldsymbol{z}_i$ $(i = 1, 2, \cdots, r)$. The idea for constructing test statistics is to apply univariate spherical distribution test based on vectors $\boldsymbol{z}_i$ for each i, and maximize these r quantities to obtain the final statistics for hypothesis in (1). In this paper, we apply KS, CM and AD statistics to test univariate spherical distribution.

2.1 Kolmogorov-Smirnov (KS) Test

We first introduce univariate KS test for one dimension sample $\{x_1, x_2, \cdots, x_n\}$, we want to test whether the sample is from cumulative density function $F(\cdot)$. KS statistic is defined as follow:

$$K_n(\boldsymbol{x}) = \max_{1\leqslant j\leqslant n}\Big\{\max\Big(\Big|F(x_{(j)}) - \frac{j-1}{n}\Big|, \Big|F(x_{(j)}) - \frac{j}{n}\Big|\Big)\Big\}, \tag{2}$$

where $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$ are sorted sample, and $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)'$. Kolmogorov[5] proved the exact distribution of $K_n(\boldsymbol{x})$ when $n \to \infty$, that is

$$\lim_{n\to\infty}\mathsf{P}(K_n(\boldsymbol{x}) \leqslant x) = \sum_{j=-\infty}^{+\infty}(-1)^j\exp(-2nj^2x^2), \qquad 0 < x < +\infty.$$

For a given x, it is difficult to calculate the exact probability because the value of j is from $-\infty$ to $+\infty$. Hence, the approximate distribution of $K_n(\boldsymbol{x})$ had been discussed extensively. For example, Marsaglia et al.[21] proposed the approximation form of $\mathsf{P}(K_n(\boldsymbol{x}) \leqslant x)$ can be

$$\mathsf{P}(K_n(\boldsymbol{x}) \leqslant x) \approx \frac{n!}{n^n}t_{kk}, \tag{3}$$

where t_{kk} is the (k,k) element of a matrix $\boldsymbol{T}=\boldsymbol{M}^n$, $k=\lceil nx\rceil$ with $\lceil a\rceil$ is round-up for a, and $\boldsymbol{M}$ being an $m\times m$ matrix with $m=2k-1$ has the following expression

$$\boldsymbol{M}=\begin{pmatrix} \boldsymbol{v}_{1:m-1} & \boldsymbol{B} \\ v_m & \boldsymbol{v}'_{m-1:1} \end{pmatrix}. \tag{4}$$

Here $\boldsymbol{B}=\{b_{ij}\}$ is a $(m-1)\times(m-1)$ matrix, and its elements b_{ij} $(i,j=1,2,\cdots,m-1)$ equal to 1, 0 or $1/(i-j)!$ which are corresponding to $i=j$, $i<j$ and $i>j$ respectively. The vector $\boldsymbol{v}_{1:m-1}=(v_1,v_2,\cdots,v_{m-1})'$ with $v_j=(1-h^j)/j!$ for $j=1,2,\cdots,m-1$, and v_m in (4) is equal to $(1-2h^m+\max\{0,2h-1\}^m)/m!$, where $h=k-nx$.

In addition, the random variable $F(x_{(j)})$ obeys the same distribution as $t_{(j)}$ under null hypothesis, where $t_{(j)}$'s are ordered statistics from $\mathrm{U}(0,1)$ for $j=1,2,\cdots,n$. Hence, the distribution of statistic $K_n(\boldsymbol{x})$ is identical with $K_n(\boldsymbol{t})$.

$$K_n(\boldsymbol{t})=\max_{1\leqslant j\leqslant n}\left\{\max\left(\left|t_{(j)}-\frac{j-1}{n}\right|,\left|t_{(j)}-\frac{j}{n}\right|\right)\right\}. \tag{5}$$

Based on this consideration, we can investigate the distribution of $K_n(\boldsymbol{t})$ by Monte Carlo method to obtain the distribution of $K_n(\boldsymbol{x})$. Specific steps are as follows:

(K1) Generate the ordered sample $\{t_{(1)},t_{(2)},\cdots,t_{(n)}\}$ from uniformed distribution $\mathrm{U}(0,1)$, and calculate $K_n(\boldsymbol{t})$ by (5);

(K2) Repeat the above step k times, obtain the corresponding statistics $K_n^1(\boldsymbol{t}),K_n^2(\boldsymbol{t}),\cdots,K_n^k(\boldsymbol{t})$;

(K3) For a given x, the estimation of $\mathsf{P}(K_n(\boldsymbol{x})\leqslant x)$ is

$$\widehat{\mathsf{P}}(K_n(\boldsymbol{x})\leqslant x)=\frac{1}{k}\sum_{i=1}^{k}I(K_n^i(\boldsymbol{t})\leqslant x).$$

According to Theorem 1, the projected sample $\boldsymbol{z}_i=(z_{i1},z_{i2},\cdots,z_{i(n-1)})'$ for $i=1,2,\cdots,r$ is from spherical distribution whose accumulative density function is denoted as $\widetilde{F}_{\boldsymbol{z}_i}(\cdot)$, we define KS statistic in (2) for $\boldsymbol{z}_i$ as follow:

$$K_q(\boldsymbol{z}_i)=\max_{1\leqslant j\leqslant q}\left\{\max\left(\left|\widetilde{F}_{\boldsymbol{z}_i}(z_{(ij)})-\frac{j-1}{q}\right|,\left|\widetilde{F}_{\boldsymbol{z}_i}(z_{(ij)})-\frac{j}{q}\right|\right)\right\},$$

where $q=n-1$ and $z_{(i1)}\leqslant z_{(i2)}\leqslant\cdots\leqslant z_{(iq)}$ be the ordered statistics of $z_{i1},z_{i2},\cdots,z_{iq}$. The larger the value of $K_q(\boldsymbol{z}_i)$, the more possibly null hypothesis is rejected. For any $r_0\leqslant r$, we maximize these statistics to obtain the following generalized statistic for hypothesis in (1)

$$\mathrm{GK}_{r_0}=\max_{1\leqslant i\leqslant r_0}\{K_q(\boldsymbol{z}_i)\}. \tag{6}$$

It is evident that null hypothesis is rejected when GK_{r_0} is large. The following theorem states the property of GK_{r_0} in (6).

Theorem 2 Under null hypothesis in (1), the probability of $\mathrm{GK}_{r_0} \leqslant x$ is

$$\mathsf{P}(\mathrm{GK}_{r_0} \leqslant x) = \mathsf{P}\Big(\max_{1\leqslant i\leqslant r_0} \{K_q(\boldsymbol{z}_i)\} \leqslant x\Big) = \prod_{i=1}^{r_0} \mathsf{P}(K_q(\boldsymbol{z}_i) \leqslant x).$$

Hence,

(i) based on the approximate distribution in (3), $\mathsf{P}(\mathrm{GK}_{r_0} \leqslant x) \approx [(q!/q^q)\cdot t_{kk}]^{r_0}$, where the definition of t_{kk} can be found in (3);

(ii) based on the simulation steps K1−K3, $\mathsf{P}(\mathrm{GK}_{r_0} \leqslant x) \approx \big[k^{-1}\sum\limits_{i=1}^{k} I(K_q^i(\boldsymbol{t}) \leqslant x)\big]^{r_0}$.

Theorem 2 can be proved by the conclusion that $\boldsymbol{z}_i$ are asymptotical independent in Theorem 1.

2.2 Cramer-von Mises (CM) Test

For the one dimension sample $\{x_1, x_2, \cdots, x_n\}$, the univariate CM statistic for distribution test is defined as follow,

$$C_n(\boldsymbol{x}) = n\int_{-\infty}^{+\infty} [F_n(x) - F(x)]^2 \mathrm{d}F(x) = \frac{1}{12n} + \sum_{j=1}^{n}\Big[\frac{2j-1}{2n} - F(x_{(j)})\Big]^2, \tag{7}$$

where $F_n(x)$ is the empirical distribution function of sample $\{x_1, x_2, \cdots, x_n\}$. For this statistic, Anderson and Darling[12] proposed exact distribution by Brownian bridge and Bessel function:

$$\begin{aligned}&\lim_{n\to\infty} \mathsf{P}(C_n(\boldsymbol{x}) \leqslant x)\\ &= \frac{1}{\pi\sqrt{x}}\sum_{j=0}^{\infty}(-1)^j\binom{-1/2}{j}(4j+1)^{1/2}\exp\Big[-\frac{(4j+1)^2}{16x}\Big]K_{1/4}\Big(\frac{(4j+1)^2}{16x}\Big),\end{aligned}$$

where $0 < x < +\infty$ and $K_{1/4}(\cdot)$ is the standard Bessel function.

It's hard to obtain the exact probability from the exact distribution of $C_n(\boldsymbol{x})$. Hence, many approximate distributions of CM statistic are given. Among others, Tiku[22] developed a Chi-square approximation for the distribution of CM statistic with little accuracy losing for upper tail probabilities. Stephens and Maag[23] discussed the exact quantiles and derived the exact distribution only for lower tail probabilities. So far, there are relatively few literatures about approximate distribution of $C_n(\boldsymbol{x})$ in the whole scope of x. Here we utilize Monte Carlo method to simulate the distribution of CM statistic. Similar with KS statistic, the distribution of statistic $C_n(\boldsymbol{x})$ is identical with $C_n(\boldsymbol{t})$ under null hypothesis.

$$C_n(\boldsymbol{t}) = \frac{1}{12n} + \sum_{j=1}^{n}\Big(\frac{2j-1}{2n} - t_{(j)}\Big)^2, \tag{8}$$

where $t_{(j)}$ is defined same as that in $K_n(\boldsymbol{t})$. The steps of simulation are as follows:

(C1) Generate the ordered sample $\{t_{(1)}, t_{(2)}, \cdots, t_{(n)}\}$ from uniformed distribution U$(0,1)$, and calculate $C_n(\boldsymbol{t})$ by (8);

(C2) Repeat the above step k times, obtain the corresponding statistics $C_n^1(\boldsymbol{t}), C_n^2(\boldsymbol{t}), \cdots, C_n^k(\boldsymbol{t})$;

(C3) For a given x, the estimation of $\mathsf{P}(C_n(\boldsymbol{x}) \leqslant x)$ is

$$\widehat{\mathsf{P}}(C_n(\boldsymbol{x}) \leqslant x) = \frac{1}{k} \sum_{i=1}^{k} I(C_n^i(\boldsymbol{t}) \leqslant x).$$

Under the multidimensional case, dimension reduction method is applied to obtain the projected sample $\boldsymbol{z}_i$, and we define CM statistic for each $\boldsymbol{z}_i$ as follow:

$$C_q(\boldsymbol{z}_i) = \frac{1}{12q} + \sum_{j=1}^{q} \Big[\frac{2j-1}{2q} - \widetilde{F}_{\boldsymbol{z}_i}(z_{(ij)})\Big]^2,$$

where $q = n-1$. For any $r_0 \leqslant r$, we construct a family of generalized CM statistics by maximizing these statistics for hypothesis in (1)

$$\mathrm{GC}_{r_0} = \max_{1\leqslant i\leqslant r_0} \{C_q(\boldsymbol{z}_i)\}. \tag{9}$$

The larger the value of GC_{r_0}, the more possibly null hypothesis is rejected. Theorem 3 states the property of GC_{r_0} in (9).

Theorem 3 Under null hypothesis in (1), the probability of $\mathrm{GC}_{r_0} \leqslant x$ is

$$\mathsf{P}(\mathrm{GC}_{r_0} \leqslant x) = \mathsf{P}\Big(\max_{1\leqslant i\leqslant r_0} \{C_q(\boldsymbol{z}_i)\} \leqslant x\Big) = \prod_{i=1}^{r_0} \mathsf{P}(C_q(\boldsymbol{z}_i) \leqslant x).$$

Hence, based on the simulation steps C1 − C3, $\mathsf{P}(\mathrm{GC}_{r_0} \leqslant x) \approx \big[k^{-1} \sum_{i=1}^{k} I(C_q^i(\boldsymbol{t}) \leqslant x)\big]^{r_0}$.

Theorem 3 can be proved similarly as that of Theorem 2.

2.3 Anderson-Darling (AD) Test

For the one dimension sample $\{x_1, x_2, \cdots, x_n\}$, the univariate AD statistic for distribution test is defined as follow:

$$\begin{aligned} A_n(\boldsymbol{x}) &= n \int_{-\infty}^{+\infty} \frac{[F_n(x) - F(x)]^2}{F(x)[1-F(x)]} \mathrm{d}F(x) \\ &= -n - \sum_{j=1}^{n} \frac{2j-1}{n} \{\ln F(x_{(j)}) + \ln[1 - F(x_{(n+1-j)})]\}, \end{aligned} \tag{10}$$

where $F_n(x)$ and $F(x)$ are empirical distribution function of $\{x_1, x_2, \cdots, x_n\}$ and accumulative distribution function under null hypothesis respectively, $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$

are order sample and $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)'$. Anderson and Darling[12] derived the exact distribution of $A_n(\boldsymbol{x})$, it can be expressed as

$$\lim_{n\to\infty} \mathsf{P}(A_n(\boldsymbol{x}) \leqslant x) = \frac{\sqrt{2\pi}}{x} \sum_{j=0}^{\infty} \bigg\{ \binom{-1/2}{j} (4j+1)^{1/2} \exp\Big[-\frac{(4j+1)^2\pi^2}{8x}\Big] \times \int_0^{\infty} \exp\Big[\frac{x}{8(\omega^2+1)} - \frac{(4j+1)^2\pi^2\omega^2}{8x}\Big] \mathrm{d}\omega \bigg\}, \qquad 0 < x < +\infty.$$

The exact distribution of AD statistic in (10) is fairly complex due to the existence of the infinite series. For this reason, there are many literatures to investigate its approximate distribution. For example, Marsaglia and Marsaglia[24] evaluated the asymptotic distribution via series with two-term recursions directly. Grace and Wood[25] presented a more accuracy approximating function for tail probabilities in case of $2 \leqslant x < \infty$ in (10). Here we apply [24]'s approximation when $0 < x < 2$ and [25]'s method when $2 \leqslant x < \infty$ to obtain the approximate expression of $\mathsf{P}(A_n(\boldsymbol{x}) \leqslant x)$ as follows

$$\mathsf{P}(A_n(\boldsymbol{x}) \leqslant x) = \begin{cases} f_1(x) + g_1(n, f_1(x)), & 0 < x < 2; \\ 1 - x^{-0.48897} \exp[g_2(x,n)], & 2 \leqslant x < \infty, \end{cases} \tag{11}$$

here

$$\begin{aligned} f_1(x) &= x^{-1/2} \mathrm{e}^{-1.2337141/x} \{2.00012 + \{0.247105 - \{0.0649821 \\ &\quad - [0.0347962 - (0.0116720 - 0.00168691x)x]x\}x\}x\}, \\ g_2(x,n) &= (0.23945n^{-0.9379} - 0.1201n^{-0.9600} - 1.0002816)x \\ &\quad - 1.437n^{-0.9379} + 1.441n^{-0.9600} - 0.0633101, \end{aligned}$$

and

$$g_1(x,n) = \begin{cases} (0.0037/n^3 + 0.0078/n^2 + 0.00006/n) f_2(x/c(n)), & x < c(n); \\ (0.04213/n + 0.01365/n^2) f_3([x - c(n)]/[0.8 - c(n)]), & c(n) < x < 0.8; \\ f_4(x)/n, & 0.8 < x. \end{cases} \tag{12}$$

The functions $f_2(x)$, $f_3(x)$, $f_4(x)$ and $c(n)$ above are defined as

$$\begin{aligned} f_2(x) &= \sqrt{x}(1-x)(49x - 102), \\ f_3(x) &= -0.00022633 + \{6.54034 - \{14.6538 - [14.458 - (8.259 - 1.91864x)x]x\}x\}x, \\ f_4(x) &= -130.2137 + \{745.2337 - \{1705.091 - [1950.646 - (1116.360 - 255.7844x)x]x\}x\}x, \\ c(n) &= 0.01265 + 0.1757/n. \end{aligned}$$

Based on the conclusion that $F(x_{(j)})$ obeys the same distribution as $t_{(j)}$ under null hypothesis, we also can apply Monte Carlo method to simulate the distribution of $A_n(\boldsymbol{t})$ to obtain the distribution of $A_n(\boldsymbol{x})$, where $A_n(\boldsymbol{t})$ is defined as follow:

$$A_n(\boldsymbol{t}) = -n - \sum_{j=1}^{n} \frac{2j-1}{n}\{\ln t_{(j)} + \ln[1 - t_{(n+1-j)}]\}, \tag{13}$$

where $t_{(j)}$ is defined same as that in $K_n(\boldsymbol{t})$. The distribution of $A_n(\boldsymbol{t})$ can be simulated as follows:

(A1) Generate the ordered sample $\{t_{(1)}, t_{(2)}, \cdots, t_{(n)}\}$ from uniformed distribution $\mathrm{U}(0,1)$, and calculate $A_n(\boldsymbol{t})$ by (13);

(A2) Repeat the above step k times, obtain the corresponding statistics $A_n^1(\boldsymbol{t}), A_n^2(\boldsymbol{t}), \cdots, A_n^k(\boldsymbol{t})$;

(A3) For a given x, the estimation of $\mathsf{P}(A_n(\boldsymbol{x}) \leqslant x)$ is

$$\widehat{\mathsf{P}}(A_n(\boldsymbol{x}) \leqslant x) = \frac{1}{k}\sum_{i=1}^{k} I(A_n^i(\boldsymbol{t}) \leqslant x).$$

Under the multidimensional case, the AD statistics for the projected sample $\boldsymbol{z}_i$ can be defined as

$$A_q(\boldsymbol{z}_i) = -q - \sum_{j=1}^{q} \frac{2j-1}{q}\{\ln \widetilde{F}_{\boldsymbol{z}_i}(z_{(ij)}) + \ln[1 - \widetilde{F}_{\boldsymbol{z}_i}(z_{(i,q+1-j)})]\},$$

where $q = n-1$, and for any $r_0 \leqslant r$, generalized AD statistics for null hypothesis in (1) is constructed as

$$\mathrm{GA}_{r_0} = \max_{1 \leqslant i \leqslant r_0}\{A_q(\boldsymbol{z}_i)\}. \tag{14}$$

We reject null hypothesis with larger GA_{r_0}, and Theorem 4 below states the property of statistics GA_{r_0}.

Theorem 4 Under null hypothesis (1), the probability of GA_{r_0} is

$$\mathsf{P}(\mathrm{GA}_{r_0} \leqslant x) = \mathsf{P}\Big(\max_{1 \leqslant i \leqslant r_0}\{A_q(\boldsymbol{z}_i)\} \leqslant x\Big) = \prod_{i=1}^{r_0} \mathsf{P}(A_q(\boldsymbol{z}_i) \leqslant x).$$

Hence,

(i) based on the approximative distribution in (11),

$$\mathsf{P}(\mathrm{GA}_{r_0} \leqslant x) \approx \begin{cases} [f_1(x) + g_1(n, f_1(x))]^{r_0}, & 0 < x < 2; \\ \{1 - x^{-0.48897}\exp[g_2(x,n)]\}^{r_0}, & 2 \leqslant x < \infty, \end{cases}$$

where the definition of $f_1(x)$, $g_1(n,x)$ and $g_2(x,n)$ can be found in (11);

(ii) based on the simulation steps A1 – A3, $\mathsf{P}(\mathrm{GA}_{r_0} \leqslant x) \approx \big[k^{-1} \sum\limits_{i=1}^{k} I(A_q^i(\boldsymbol{t}) \leqslant x)\big]^{r_0}$.

Theorem 4 can be proved similarly as that for Theorem 2.

Remark 5 According to Theorem 1, the projected sample $\boldsymbol{z}_i$ is from spherical distribution, and corresponding accumulative density function is denoted as $\widetilde{F}_{\boldsymbol{z}_i}(\cdot)$. In this paper, we use its asymptotical distribution $\mathrm{N}(0, \widehat{h}_i)$ to estimate $\widetilde{F}_{\boldsymbol{z}_i}(\cdot)$ in calculating the corresponding statistics, where $\widehat{h}_i$ is the estimator of the parameter h_i.

Based on KS, CM and AD statistic, larger statistics $K_q(\boldsymbol{z}_i)$, $C_q(\boldsymbol{z}_i)$, $A_q(\boldsymbol{z}_i)$ derived by projection on eigenvectors imply more deviation from spherical distribution. Hence, we reject null hypothesis for larger GK_{r_0}, GC_{r_0} and GA_{r_0}. For GK_{r_0} and GA_{r_0}, we can utilize approximate distribution or Monte Carlo method to obtain the corresponding critical value under null hypothesis (1). But for GC_{r_0}, we only can apply Monte Carlo method to obtain critical value under null hypothesis (1).

§3. Numerical Study

3.1 Simulation Analysis

In this section, we mainly use Monte Carlo analysis to study the efficiency of the proposed statistics. Different examples are investigated here to illustrate the performance of the proposed statistics. Under null hypothesis, the sample $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ is from multivariate normal distribution $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Two cases of covariance matrix $\boldsymbol{\Sigma}$, which is corresponding to independent and dependent, are selected as: (i) $\boldsymbol{\mu} = \boldsymbol{0}$, $\boldsymbol{\Sigma} = \boldsymbol{I}_p$; (ii) $\boldsymbol{\mu} = \boldsymbol{0}$, $\boldsymbol{\Sigma} = (\rho_{ij})$, where $\rho_{ii} \equiv 1$, $\rho_{ij} = \rho = 0.5$, when $|i-j| = 1$. The simulation results for generalized KS, AD, CM statistics are denoted as GK_{r_0}, GA_{r_0} and GC_{r_0} respectively. For comparison, we include the generalized SW statistic denoted as GW_{r_0} in [19]. The projection dimension r_0 is chosen as $r_0 = \lfloor p/2 \rfloor$ for $p \leqslant 10$; $r_0 = \lfloor p/3 \rfloor$ for $p > 10$ as suggested by [26].

According to Theorem 2 and Theorem 4, there are two methods to approximate the distribution functions $\mathsf{P}\{\mathrm{GK}_{r_0} \leqslant x\}$ and $\mathsf{P}\{\mathrm{GA}_{r_0} \leqslant x\}$ respectively. One method is based on approximate formulas, and the other one is based on Monte Carlo methods. Under null hypothesis and the conditions that sample size $n = 20$, dimension $p = 20, 25, 30$, we plot the probabilities based on two methods with 10 000 replications to observe the difference between them in Figures 1 – 2. Figure 1 (a) – (c) is from an independent covariance matrix while Figure 1 (d) – (f) is from a dependent one. Through Figure 1 (a) – (f), we can obtain that whether the covariance matrix $\boldsymbol{\Sigma}$ is independent or not, and whether n is larger than p

or not, probabilities computed by approximate formulas are close to the probabilities based on Monte Carlo steps. As for Figure 2, there are similar conclusions for $\mathsf{P}\{\mathrm{GA}_{r_0} \leqslant x\}$. Simulation analysis also indicates that the results by two methods are almost the same. Hence, in the following, we only display the results computed by approximate formulas.

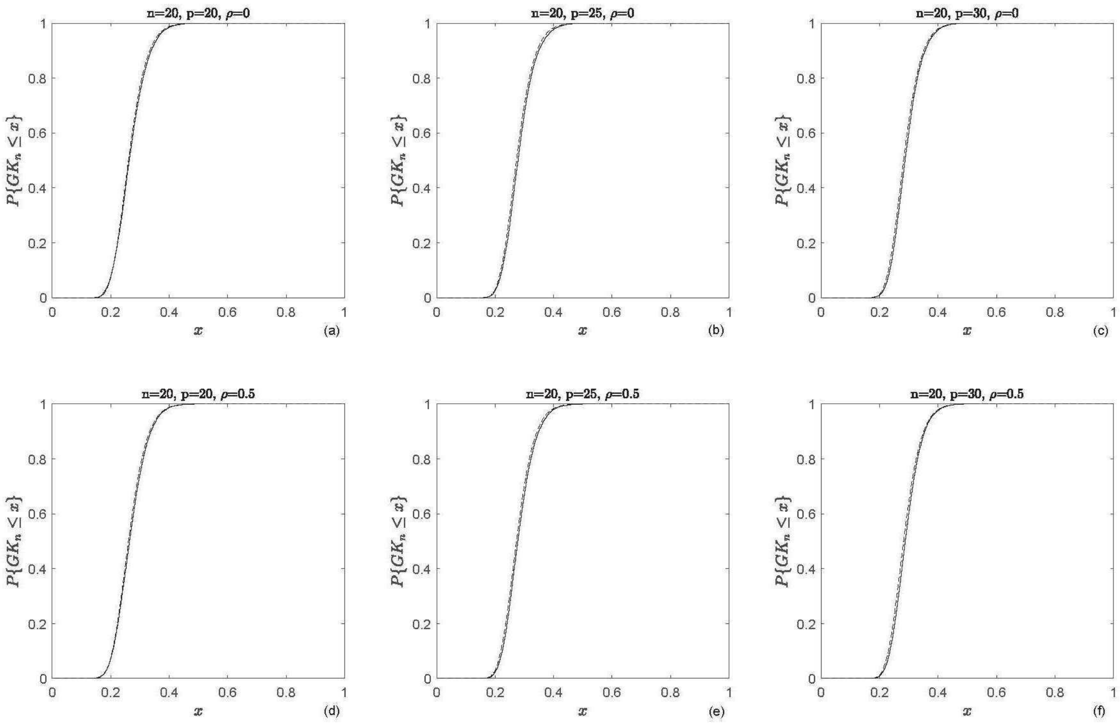

Figure 1 The distribution function of GK_{r_0}. The dashed and solid lines denote the probability by approximate formula and Monte Carlo simulations respectively.

First, we investigate the performance of the tests under null hypothesis at the significance level $\alpha = 0.05$. The simulation results with 10 000 replications are summarized in Table 1 for $n = 20$ and $p = 5, 10, 15, 20, 25, 30$, where the sizes of proposed statistics GK_{r_0}, GA_{r_0} are obtained from the approximate distribution in Theorem 2 and Theorem 4 respectively. As for generalized CM test, we get the critical values of statistics GC_{r_0} by steps C1 − C3 in Section 2.2. As we can see in Table 1, all the test statistics can contain type I error rate.

Under alternative hypothesis, we choose six cases to study the powers performance. The following alternative distributions for $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$, $i = 1, 2, \cdots, n$ are chosen:

- The multivariate t-distribution with degree of freedom 5 and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$: the compo-

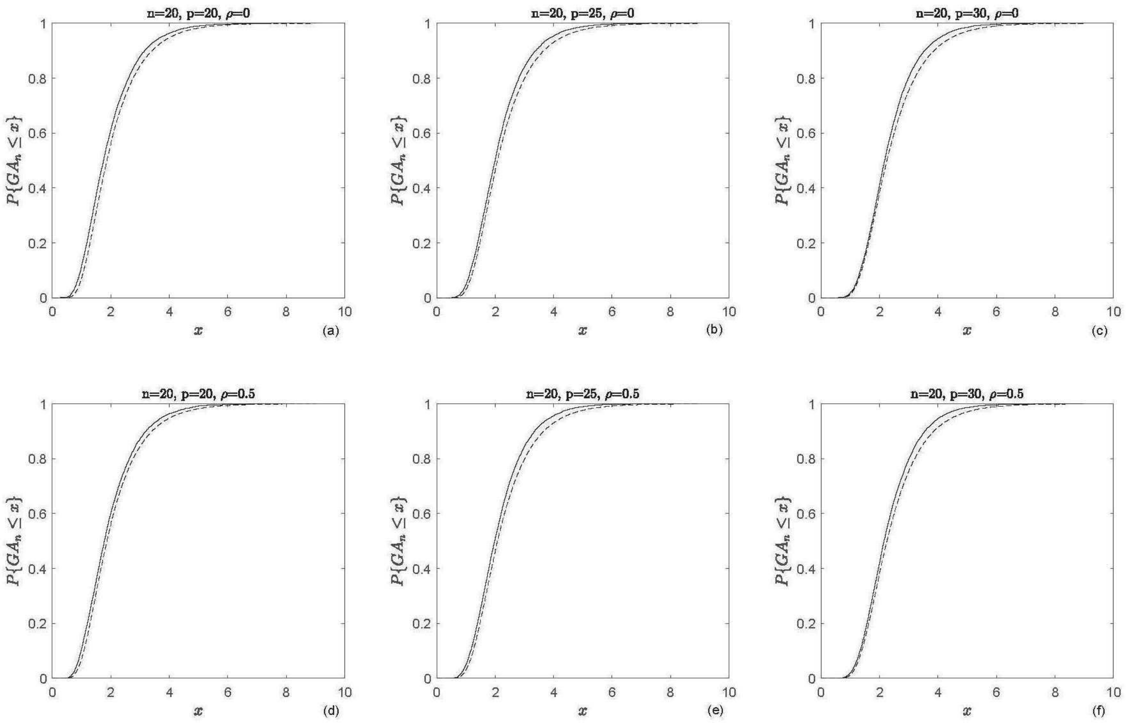

Figure 2 The distribution function of GA_{r_0}. The dashed and solid lines denote the probability by approximate formula and Monte Carlo simulations respectively.

Table 1 Sizes of GK_{r_0}, GA_{r_0}, GC_{r_0} and GW_{r_0} ($n=20$, $\alpha=0.05$)

p	r_0	Σ	GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}	Σ	GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}
5	2	I_p	0.0483	0.0360	0.0299	0.0488	D	0.0434	0.0591	0.0314	0.0457
10	5		0.0515	0.0349	0.0332	0.0516		0.0502	0.0518	0.0518	0.0554
15	5		0.0478	0.0355	0.0236	0.0533		0.0383	0.0545	0.0224	0.0530
20	6		0.0464	0.0418	0.0227	0.0515		0.0506	0.0440	0.0251	0.0544
25	8		0.0512	0.0312	0.0177	0.0537		0.0498	0.0276	0.0242	0.0569
30	10		0.0494	0.0281	0.0195	0.0542		0.0458	0.0577	0.0222	0.0565

*I_p represents identity matrix; D represents the matrix (ρ_{ij}), where $\rho_{ii}\equiv 1$ and $\rho_{ij}=0.5$, when $|i-j|=1$.

nents of $\boldsymbol{x}_i$ are independent identically distributed from $t(5)$ and the definition of multivariate t-distribution can be found in [27].

- The β-generalized normal distribution $\mathrm{N}(\mathbf{0},\boldsymbol{I}_p,1)$: the probability density function of β-generalized normal distribution $\mathrm{N}(\mathbf{0},\boldsymbol{I}_p,\beta)$ is

$$f(x_1,x_2,\cdots,x_p)=\frac{\beta^p\gamma^{p/\beta}}{2^p\Gamma^p(1/\beta)}\cdot\exp\Big(-\gamma\sum_{i=1}^{p}|x_i|^\beta\Big),\qquad (x_1,x_2,\cdots,x_p)'\in R^p,$$

which is given by [28]. Here, both β $(0 < \beta < 2)$ and γ $(\gamma > 0)$ are parameters.

- The centered $\chi^2(1)$ distribution: the components of $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,p})$ are independent identically distributed from $\chi^2(1) - 1$.
- The centered $\Gamma(2,2)$ distribution: the components of $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,p})$ are independent identically distributed from $\Gamma(2,2) - 1$.
- The mixture distribution of centered $\chi^2(1)$ and normal distribution: for $\boldsymbol{x}_i = (x_{i,1}, \cdots, x_{i,\lfloor p/2 \rfloor}, x_{i,(\lfloor p/2 \rfloor+1)}, \cdots, x_{i,p})$, $(x_{i,1}, \cdots, x_{i,\lfloor p/2 \rfloor})$ are independent identically distributed from $\chi^2(1) - 1$ and $(x_{i,(\lfloor p/2 \rfloor+1)}, \cdots, x_{i,p})$ are independent identically distributed from $\mathrm{N}(0,1)$, where $\lfloor p/2 \rfloor$ stands for $p/2$ round-down.
- The associated normal distribution: for $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,p})$, when $p = 2k$, $\widetilde{\boldsymbol{x}}_{i1} = (x_{i,1}, x_{i,2}, \cdots, x_{i,k}) \sim \mathrm{N}(0,1)$, $\widetilde{\boldsymbol{x}}_{i2} = (x_{i,(k+1)}, x_{i,(k+2)}, \cdots, x_{i,p}) = (x_{i,1}^2, x_{i,2}^2, \cdots, x_{i,k}^2) + e$, with $e \sim \mathrm{N}(0,1)$; when $p = 2k+1$, $\boldsymbol{x}_i = (\widetilde{\boldsymbol{x}}_{i1}, \widetilde{\boldsymbol{x}}_{i2}, x_{i,p})$, with $x_{i,p} \sim \mathrm{N}(0,1)$.

The multivariate t-distribution and β-generalized normal distribution are symmetric distributions with thicker tails than multivariate normal distribution. Both the centered $\chi^2(1)$ distribution and centered $\Gamma(2,2)$ distribution are skewed distributions. The last two are mixture distributions involving normal marginals.

The simulation results from the six alternative distributions are summarized in Tables 2 – 3. It can be obtained that when the alternative distribution is shifted $\Gamma(2,2)$ distribution and the β-generalized normal distribution, all the proposed statistics have better power than GW_{r_0}. The power of GW_{r_0} is insensitive to the covariance matrix $\boldsymbol{\Sigma}$. However, the power of proposed statistics are influenced quite slightly by the covariance matrix $\boldsymbol{\Sigma}$. Among the proposed statistics, the powers of GA_{r_0} are significantly higher than the others under all the alternative hypotheses except the mixture distribution of centered $\chi^2(1)$ and normal distribution, where GC_{r_0} is the most effective.

3.2 Real Data Analysis

In this part, the proposed statistics are applied to real data to illustrate the performance of the tests. In the following examples, we test the multivariate normality of real data. Since sample size n is quite small, we compute the p-values of the proposed statistics and GW_{r_0} for $r_0 = 1, 2, \cdots, \lceil \min(n,p)/2 \rceil$, where $\lceil \min(n,p)/2 \rceil$ stands for $\min(n,p)/2$ round-up.

Example 6 The data is from [29]'s study that examined the multivariate normality on a data set used in a cancer research. There are eleven laboratory mice divided into two groups. The strains and growth situations of these mice are the same, thus they can be treated as

Tabel 2 Powers of GK_{r_0}, GA_{r_0}, GC_{r_0} **and** GW_{r_0} ($n=20$, $\alpha=0.05$, $\boldsymbol{\Sigma}=\boldsymbol{I}_p$)

p	r_0	Multi-t(5)				β-Nor.			
		GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}	GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}
5	2	0.1193	0.3165	0.0836	0.3926	0.3595	0.5158	0.4882	0.2202
10	5	0.1415	0.3329	0.1175	0.5152	0.3101	0.6501	0.8987	0.1787
15	5	0.1902	0.2558	0.0691	0.6220	0.7085	0.6127	0.9948	0.1690
20	6	0.1887	0.2676	0.0641	0.6970	1.0000	0.7467	1.0000	0.1740
25	8	0.2129	0.2776	0.0625	0.7359	1.0000	1.0000	1.0000	0.1725
30	10	0.1886	0.3005	0.0693	0.7757	1.0000	1.0000	1.0000	0.1644
p	r_0	$\chi^2(1)-1$				$\Gamma(2,2)-1$			
		GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}	GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}
5	2	0.3615	0.5309	0.3238	0.7959	0.8322	1.0000	0.6386	0.3733
10	5	0.3756	0.3720	0.4609	0.7520	0.9261	1.0000	0.8101	0.3052
15	5	0.3390	0.4291	0.4032	0.7314	0.9385	1.0000	0.7477	0.2961
20	6	0.2982	0.5136	0.3788	0.7062	0.9572	1.0000	0.6948	0.2689
25	8	0.3115	0.5602	0.3991	0.6671	0.9658	1.0000	0.7592	0.2455
30	10	0.2793	0.4501	0.4315	0.6472	0.9850	1.0000	0.7882	0.2210
p	r_0	$\chi^2(1)+$ Nor.				Assoc-Nor.			
		GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}	GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}
5	2	0.2284	0.2037	0.3135	0.6136	0.2186	0.3579	0.0342	0.4178
10	5	0.2003	0.3233	0.5185	0.6529	0.2294	0.4280	0.0529	0.4482
15	5	0.2169	0.3313	0.4045	0.6491	0.2180	0.4566	0.0228	0.4453
20	6	0.1821	0.3679	0.4121	0.6469	0.2603	0.4576	0.0198	0.4420
25	8	0.2114	0.2626	0.4356	0.6102	0.2731	0.5180	0.0156	0.4278
30	10	0.2090	0.2594	0.4461	0.5944	0.2090	0.6152	0.0176	0.4076

*I_p represents identity matrix.

virtually genetically identical. All of the mice are cultured with subcutaneous transplant of tumor and are given a new anticancer drug irinotecan (CPT-11) with two different doses. One group received 0.4mg/kg CPT-11 and the other one 0.26mg/kg CPT-11. The tumor volumes were measured weekly lasting for 12 weeks. Missing data arise from sixth week because of toxicity or tumor growing too fast which result in the death of the mice. The complete testing data can be found in Table 1 of [29].

We consider two hypothesis test problems. For the first problem, there are 8 mice keep alive by the eleventh week. Then the sample size is $n=8$ and the dimension is $p=11$. For the other one, there are 5 mice keep alive by the twelfth week. The sample size is $n=5$ and the dimension is $p=12$. For the two sets of data, we want to test that whether they are distributed from multivariate normal distribution. These problems are equal to test (1) with $(n,p)=(8,11)$ and $(n,p)=(5,12)$ respectively. The p-values are shown in Table 4.

Tabel 3 Powers of GK_{r_0}, GA_{r_0}, GC_{r_0} and GW_{r_0} ($n = 20$, $\alpha = 0.05$, $\boldsymbol{\Sigma} = \boldsymbol{D}$)

p	r_0	Multi-t(5)				β-Nor.			
		GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}	GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}
5	2	0.1624	0.1575	0.0706	0.3789	0.5624	0.9754	0.9362	0.2202
10	5	0.1651	0.2474	0.1353	0.5161	0.8628	0.9998	0.9993	0.1888
15	5	0.1778	0.1913	0.0619	0.6317	0.9554	0.9969	1.0000	0.1693
20	6	0.2025	0.2186	0.0573	0.6921	1.0000	0.9994	1.0000	0.1817
25	8	0.2119	0.3051	0.0709	0.7380	1.0000	1.0000	1.0000	0.1696
30	10	0.2172	0.2885	0.0813	0.7776	1.0000	1.0000	1.0000	0.1619
p	r_0	$\chi^2(1)-1$				$\Gamma(2,2)-1$			
		GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}	GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}
5	2	0.3814	0.4370	0.3141	0.8011	0.6556	1.0000	0.5856	0.3713
10	5	0.2772	0.6378	0.5040	0.7475	0.9635	1.0000	0.7948	0.3071
15	5	0.2770	0.3592	0.3591	0.7191	0.8315	1.0000	0.7094	0.2947
20	6	0.2895	0.3426	0.4662	0.7111	0.9505	1.0000	0.6545	0.2775
25	8	0.2674	0.6358	0.4934	0.6714	0.9624	1.0000	0.7313	0.2430
30	10	0.2756	0.6328	0.4888	0.6446	0.9916	1.0000	0.8011	0.2197
p	r_0	$\chi^2(1)+$ Nor.				Assoc-Nor.			
		GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}	GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}
5	2	0.2124	0.1485	0.3336	0.6199	0.2184	0.2713	0.0433	0.4273
10	5	0.2348	0.2034	0.4995	0.6627	0.2976	0.5172	0.0341	0.4353
15	5	0.1692	0.1818	0.3835	0.6571	0.1699	0.3637	0.0191	0.4474
20	6	0.2063	0.1954	0.4118	0.6512	0.2708	0.3680	0.0308	0.4418
25	8	0.1813	0.1886	0.4528	0.6213	0.2218	0.2943	0.0197	0.4220
30	10	0.1690	0.2378	0.4519	0.5936	0.2077	0.5251	0.0235	0.4012

*D represents the matrix (ρ_{ij}), where $\rho_{ii} \equiv 1$ and $\rho_{ij} = 0.5$, when $|i-j| = 1$.

Tabel 4 p-values of statistics ($\alpha = 0.05$)

$(n,p) = (8,11)$				
r_0	GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}
1	0.3148	0.4843	0.0290	0.0298
2	0.2567	0.6895	0.0921	0.0587
3	0.4983	0.5804	0.1136	0.0867
4	0.3945	0.9329	0.1944	0.1139
$(n,p) = (5,12)$				
r_0	GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}
1	0.3390	0.6150	0.1541	0.2543
2	0.6321	0.6461	0.2314	0.4439
3	0.7049	0.9674	0.3464	0.5853

It can be concluded obviously, that nearly all the p-values of statistics are larger than the significance level α. These results indicate that the multivariate normality assumption

(1) on both sets of data with $(n,p)=(8,11)$ and $(n,p)=(5,12)$ should not be rejected, which is consistent with [29]'s conclusion.

Example 7 The data is from [30], which is about a study of 19 depressive patients' the absolute theta power of electroencephalogram (EEG) during a six-week therapy. These 9 variables represent the changes of nine selected channels respectively. The complete testing data can be found in Table 1 of [30]. For each patients, there are observation data of 9 selected channels respectively. The sample size is $n=19$ and the dimension is $p=9$. For this set of data, we are concerned with whether they are distributed from multivariate normal distribution. This problem is equal to test (1) with $(n,p)=(19,9)$. The p-values are shown in Table 5.

Tabel 5 p**-values of statistics when** $n=19$, $p=9$ $(\alpha=0.05)$

r_0	GK_{r_0}	GA_{r_0}	GC_{r_0}	GW_{r_0}
1	0.3953	0.3288	0.0405	0.1361
2	0.0073	0.0041	0.0913	0.0025
3	0.0050	0.0067	0.0453	0.0019
4	0.0138	0.0098	0.0400	0.0090
5	0.0158	0.0140	0.0090	0.0021

We can know from the result in Table 5 that nearly all the p-values of statistics are extremely smaller than the significance level α except $r_0=2$ for GC_{r_0} and $r_0=1$ for the other statistics. These results indicate that the multivariate normality assumption (1) on the set of data with $(n,p)=(19,9)$ should be rejected.

§4. Conclusions

In this paper, we expand three well-known statistics based on empirical distribution function (EDF) in testing univariate normality to the high dimensional case by projection methods. Since the projection data obey spherical distribution, based on the theory of spherical distribution, we maximize corresponding statistics obtained from each projection direction. We derive the approximate distributions of proposed statistics based on that in univariate case under null hypothesis. By Monte Carlo method, the exact distributions of proposed statistics can be approximated, and two of them can be obtained by approximate formulas as well. Plotting the curves, it is illustrated that the results computed by two methods match well. Note that the approximate independence of projected vectors holds under the condition $p/n\to 0$, hence, this condition is necessary to prove the properties of proposed statistics. Numerical analysis verify that proposed statistics can control the type I error well and are more competitive under some alternative hypotheses with large p. Real data application also demonstrates the effectiveness of the generalized tests.

References

[1] MARDIA K V. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies [J]. *Sankhyā Ser B*, 1974, **36(2)**: 115–128.

[2] SRIVASTAVA M S. A measure of skewness and kurtosis and a graphical method for assessing multivariate normality [J]. *Statist Probab Lett*, 1984, **2(5)**: 263–267.

[3] CSORGO S. Testing for normality in arbitrary dimension [J]. *Ann Statist*, 1986, **14(2)**: 708–723.

[4] BARINGHAUS L, HENZE N. A consistent test for multivariate normality based on the empirical characteristic function [J]. *Metrika*, 1988, **35(1)**: 339–348.

[5] KOLMOGOROV A N. Sulla determinazione empirica di una legge di distribuzione [J]. *Giorn Ist Ital Attuari*, 1933, **4**: 83–91.

[6] ROSENBLATT M. Remarks on a multivariate transformation [J]. *Ann Math Statist*, 1952, **23(3)**: 470–472.

[7] JUSTEL A, PEÑA D, ZAMAR R. A multivariate Kolmogorov-Smirnov test of goodness of fit [J]. *Statist Probab Lett*, 1997, **35(3)**: 251–259.

[8] PEACOCK J A. Two-dimensional goodness-of-fit testing in astronomy [J]. *Monthly Notices Roy Astronom Soc*, 1983, **202(3)**: 615–627.

[9] FASANO G, FRANCESCHINI A. A multidimensional version of the Kolmogorov-Smirnov test [J]. *Monthly Notices Roy Astronom Soc*, 1987, **225(1)**: 155–170.

[10] CRAMÉR H. On the composition of elementary errors [J]. *Scand Actuar J*, 1928, **1928(1)**: 13–74.

[11] VON MISES R. *Wahrscheinlichkeit Statistik und Wahrheit* [M]. Berlin: Springer-Verlag, 1928.

[12] ANDERSON T W, DARLING D A. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes [J]. *Ann Math Statist*, 1952, **23(2)**: 193–212.

[13] ANDERSON T W, DARLING D A. A test of goodness of fit [J]. *J Amer Statist Assoc*, 1954, **49**: 765–769.

[14] KOZIOL J A. A class of invariant procedures for assessing multivariate normality [J]. *Biometrika*, 1982, **69(2)**: 423–427.

[15] CHIU S N, LIU K I. Generalized Cramér-von Mises goodness-of-fit tests for multivariate distributions [J]. *Comput Statist Data Anal*, 2009, **53(11)**: 3817–3834.

[16] LEWIS P A W. Distribution of the Anderson-Darling statistic [J]. *Ann Math Statist*, 1961, **32(4)**: 1118–1124.

[17] GILES D E A. A saddlepoint approximation to the distribution function of the Anderson-Darling test statistic [J]. *Comm Statist Simulation Comput*, 2001, **30(4)**: 899–905.

[18] SRIVASTAVA M S, HUI T K. On assessing multivariate normality based on Shapiro-Wilk W statistic [J]. *Statist Probab Lett*, 1987, **5(1)**: 15–18.

[19] LIANG J J, TANG M L, CHAN P S. A generalized Shapiro-Wilk W statistic for testing high-dimensional normality [J]. *Comput Statist Data Anal*, 2009, **53(11)**: 3883–3891.

[20] LIANG J J, TANG M L. Generalized F-tests for the multivariate normal mean [J]. *Comput Statist Data Anal*, 2009, **53(4)**: 1177–1190.

[21] MARSAGLIA G, TSANG W W, WANG J B. Evaluating Kolmogorov's distribution [J]. *J Stat Softw*, 2003, **8(18)**: 1–4.

[22] TIKU M L. Laguerre series forms of non-central X^2 and F distributions [J]. *Biometrika*, 1965, **52(3-4)**: 415–427.

[23] STEPHENS M A, MAAG U R. Further percentage points for W_N^2 [J]. *Biometrika*, 1968, **55(2)**: 428–430.

[24] MARSAGLIA G, MARSAGLIA J C W. Evaluating the Anderson-Darling distribution [J]. *J Stat Softw*, 2004, **9(2)**: 1–5.

[25] GRACE A W, WOOD I A. Approximating the tail of the Anderson-Darling distribution [J]. *Comput Statist Data Anal*, 2012, **56(12)**: 4301–4311.

[26] LIANG J J, LI R Z, FANG H B, et al. Testing multinormality based on low-dimensional projection [J]. *J Statist Plann Inference*, 2000, **86(1)**: 129–141.

[27] FANG K T, KOTZ S, NG K W. *Symmetric Multivariate and Related Distributions* [M]. London: Chapman and Hall, 1990.

[28] GOODMAN I R, KOTZ S. Multivariate θ-generalized normal distributions [J]. *J Multivariate Anal*, 1973 **3(2)**: 204–219.

[29] TAN M, FANG H B, TIAN G L, et al. Testing multivariate normality in incomplete data of small sample size [J]. *J Multivariate Anal*, 2005, **93(1)**: 164–179.

[30] LÄUTER J, GLIMM E, KROPF S. New multivariate tests for data with an inherent structure [J]. *Biometrical J*, 1996, **38(1)**: 5–23.

基于经验分布函数的高维正态性检验

崔甲蓉 朱枫怡

(中国人民大学统计学院, 北京, 100872) (北京大学数学科学学院, 北京, 100871)

刘佳敏 许王莉

(中国人民大学应用统计中心和统计学院, 北京, 100872)

摘 要: 基于经验分布函数 (EDF) 的 Kolmogorov-Smirnov (KS), Cramer-von Mises (CM) 和 Anderson-Darling (AD) 统计量是单变量正态性检验中常用的统计量. 本文通过变量降维方法, 提出基于 EDF 的广义统计量来检验高维正态性. 通过蒙特卡洛方法模拟了三种统计量的近似临界值, 并基于单变量情形下统计量的近似分布公式研究了广义统计量在原假设下的近似分布. 蒙特卡洛模拟说明在某些备择假设下, 所提出的统计量比现有方法功效更好. 最后, 本章将提出的检验方法应用于实际数据验证统计量的有效性.

关键词: 高维; 特征值分解; EDF 检验统计量

中图分类号: O212.1