doi:10.20944/preprints202102.0544.v1

Article

Normality Testing of High-dimensional Data Based on Principle Component and Jarque–Bera Statistics

Ya-nan Song[1,2] and Xuejing Zhao [1]*

[1] School of Mathematics and Statistics, Lanzhou University, P. R. CHINA;
[2] School of Mathematics and Statistics, Xi'an Jiaotong University, P. R. CHINA;
Email: songyn@stu.xjtu.edu.cn; zhaoxj@lzu.edu.cn

Abstract: The testing of high-dimensional normality is an important issue and has been intensively studied in the literature. It depends on the variance–covariance matrix of the sample and numerous methods have been proposed to reduce the complexity of the variance-covariance matrix. Principle component analysis (PCA) has been widely used in high dimensions, since it can project the high-dimensional data into lower dimensional orthogonal space. The normality of the reduced data can then be evaluated by Jarque–Bera (JB) statistics in each principle direction. We propose a combined statistics—the summation of one-way JB statistics upon the independence of each principle direction—to test the multivariate normality of data in high dimensions. The performance of the proposed method is illustrated by the significance level and empirical power of the simulated normal and non-normal data. Two real examples show the validity of our proposed method.

Keywords: Principal component; Jarque–Bera statistic; Normality testing; Empirical power; Simulation

1. Introduction

Normality plays an important role in statistical analysis, and there are numerous methods for normality testing presented in the literature. Koziol [1] and Slate [2] used the properties of normal distribution function to assess multivariate normality. [3] checked normality using a class of goodness-of-fit tests, and this kind of method was also discussed in [4,5]. Various statistics have also been used in recent years, such as the Cramér-Von Mises(CM) statistic [5], skewness and kurtosis [6], sample entropy [7], Shapiro–Wilk's W statistic [8], and the Kolmogorov-Smirnov(KS) statistic (see also in [9–11]).

It is noticed that many studies of the aforementioned statistics are based on univariate normality, while the practical research we concentrate on is based on multivariate normality. Therefore, generalization should be used to enlarge the conclusions from univariate to multivariate. This is a common practice in multivariate normality testing when some useful statistics are adopted. Projection methods such as principle component analysis (PCA) can be exploited to obtain such achievement, as described in [8,12]. Convenient principle component analysis can project a high dimensional dataset into several lower dimensions in independent directions, then statistical tests in each direction can be summarized together to give a total test for multivariate normality, using the fact that the joint probability distribution is the product of all marginal probability distributions for independent variables. With the help of these orthogonal projections, the dimension can be reduced and the computation can be more efficient.

In this paper, the Jarque–Bera statistic, a combination of skewness and kurtosis, instead of the two statistics, as in [8], is investigated to test the normality in each principle direction. Then, a new kind of statistic JB_{sum} is constructed to test the high-dimensional normality. The performance of the proposed method and its power of testing are illustrated based on some high-dimensional simulated data.

This paper is organized as follows. Section 2 provides the theory of principle component analysis and gives the methodologies of statistical inference for multivariate normality. In Section 3, some simulated examples of normal data and non-normal data are used to illustrate the efficiency of our proposed method. Two real examples are then investigated in Section 4 to verify the methods' effectiveness.

2. High-dimensional Normality Test Based on PC-type JB statistic

For observed data $\mathbf{X} = (x_{ij})_{n\times p}$ with sample size n and dimension p, the principle component analysis reduces the dimension of p-variate random vector $\mathbf{X}$ through linear combinations, and it searches the linear combinations with larger spread among the observed value of $\mathbf{X}$, i.e., the larger variances. Specifically, it searches for the orthogonal directions $\omega_i(i = 1, 2, \ldots, p)$, which satisfy

$$\begin{aligned} &\omega = \arg\max_{\omega} \mathrm{Var}(\mathbf{X}\omega) = \arg\max_{\omega} \omega^T \mathrm{Var}(\mathbf{X})\omega \\ s.t. \quad &\omega^T\omega = 1 \end{aligned} \tag{1}$$

Denoted by $\mathbf{\Sigma}$, the covariance matrix of $\mathbf{X}$, the eigenvalue λ_i and principle components $\omega_i(i = 1, 2, \ldots, p)$ can be obtained by spectral decomposition of the covariance matrix $\mathbf{\Sigma}$. Therefore, the observed data can be projected to the archived lower-dimension space $\{\omega_1, \omega_2, \ldots, \omega_p\}$ by $\mathbf{z}_i = \mathbf{X}\omega_i$, which gives the projected observed matrix $\mathbf{z}$.

For each $\mathbf{z}_i$, the skewness and kurtosis can be calculated by

$$S_k(\mathbf{z}_i) = \frac{\frac{1}{n}\sum_{j=1}^{n}\left(z_{ij} - \bar{z}_i\right)^3}{\left(\frac{1}{n}\sum_{j=1}^{n}\left(z_{ij} - \bar{z}_i\right)^2\right)^{3/2}} \tag{2}$$

$$K_u(\mathbf{z}_i) = \frac{\frac{1}{n}\sum_{j=1}^{n}\left(z_{ij} - \bar{z}_i\right)^4}{\left(\frac{1}{n}\sum_{j=1}^{n}\left(z_{ij} - \bar{z}_i\right)^2\right)^{2}} \tag{3}$$

where $\bar{z}_i$ stands for the sample mean. Then, the univariate JB statistic can be given by

$$JB(\mathbf{z}_i) = \frac{n}{6}\left(S_k^2(\mathbf{z}_i) + \frac{(K_u(\mathbf{z}_i) - 3)^2}{4}\right) \tag{4}$$

To test the normality of high-dimensional data, $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_r)$, define

$$JB_{sum}(\mathbf{z}) = \sum_{i=1}^{r} JB(\mathbf{z}_i), \tag{5}$$

where r stands for the number of principle components ultimately selected, which satisfies:

$$\sum_{i=1}^{r}\lambda_i \Big/ \sum_{i=1}^{p}\lambda_i \leq 1 - s.$$

Considering the hypothesis:

$$\mathrm{H}_0 : \text{the data is normally distributed; v.s. } \mathrm{H}_1 : \text{the data is nonnormally distributed}$$

Under the null hypothesis H_0, the JB statistic will be asymptotically $\chi^2(2)$ distributed[13], then the JB_{sum} will be asymptotically $\chi^2(2r)$ distributed. For a given significance α, the critical region will be

$$R_1(Z) = \{Z | JB_{sum}(Z) > \chi^2_\alpha(2r)\}. \tag{6}$$

Upon JB_{sum}, an exact critical region $R(X)$ can be deduced, and therefore the testing can be implemented based on these critical regions.

Evaluating the performance of the proposed PC-type Jarque–Bera testing depends on 1) whether the orthogonal axes are chosen due to the cumulative proportion; and 2) whether the hypothesis is rejected or accepted. Composed by the well known power function, the error will be:

$$Power = \begin{cases} \alpha & \text{with } \mathrm{H}_0 \\ s + (1-s)(1-\beta) = s\beta + (1-\beta) & \text{with } \mathrm{H}_1 \end{cases}, \tag{7}$$

where α is the probability of a Type-I error and β is the probability of a Type-II error. Therefore, we can see that the power is a non-decreasing function of the parameter s.

3. Numerical Simulations

To evaluate the performance of the aforementioned testing, some simulation experiments are carried out in this section.

3.1. Normally distributed data

A series of normally distributed data were investigated with different data dimension p and different sample size n. Let $n \times p$ simulated data matrix $\mathbf{X}_{n\times p} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \mathbf{0}$. Consider two kinds of covariance matrix:

(I) $\boldsymbol{\Sigma} = \rho^{I_{(|i-j|\neq 0)}}$;

(II) $\boldsymbol{\Sigma} = 0.5\rho^{I_{(|i-j|\neq 0)}} + 0.5\rho^{|i-j|}$.

Table 1 and Table 2 describe the significance level of the PC-type JB testing JB_{sum} compared with Mardia's method Z^*_{M1}[19], Kauyuki's method MJB^*_m[17], and Rie's method ZNT[18] in these two different situations, respectively.

From the table above we can conclude that the significance level of JB_{sum} is close to the specified value whenever p/n is large or small. For a given sample size n, with the increase in dimension p, Z^*_{M1}, MJB^*_m and ZNT perform poorly, whereas JB_{sum} still performs well.

3.2. Non-Normally distributed data

In this part, non-normal datasets are simulated to evaluate the performance of the proposed method according to empirical power. The performance is evaluated in three databases as follows:

(III) $Shifted\ \chi^2(1)$: every variable in $\mathbf{X}_{n\times p}$ was centralized, with independently identical distribution $\chi^2(1)$.

(IV) $Shifted\ exp(1)$: every variable in $\mathbf{X}_{n\times p}$ was centralized, with independently identical distribution $exp(1)$.

(V) $N(0,1) + \chi^2(2)$: the first $[p/2]$ variables in $\mathbf{X}_{n\times p}$ are from $N(0,1)$ distribution, while the last $p - [p/2]$ variables independently identically distributed from $\chi^2(2)$, where $[p/2]$ stands for the integer part of $p/2$.

The performance of JB_{sum} compared with the S_k-type statistics χ^2_{sk}, S_{kmax} [14], K_u-type statistics χ^2_{ku}, K_{umax} [14], Mardia's statistics Z^*_{M1}, Z^*_{M2} [19], Srivastava's statistics Z^*_{S1}, Z^*_{S2} [17], Kazuyuki's statistic $mJBM$ [16] and Rie's statistic ZNT [18] are illustrated in Figures 1–5. Since JB_{sum}, χ^2_{sk}, and χ^2_{ku} are based on the sum of χ^2, we call them sum-type. S_{kmax} and K_{umax} come from the maximum of χ^2, and thus we call them max-type.

All of these methods are studied in 2000 simulated data. Figure 1– 5 show the comparisons of the power of different dimensions p and various sample sizes n.

(1) Figure 1 indicates that in the case of $p = 5$, Z^*_{M1}'s performance is best in all three cases. Though Z^*_{M2} performs well in Case I and Case II, it is not as good in Case V. Comparatively, Z^*_{S1}, χ^2_{sk} and JB_{sum} perform similarly well and better than χ^2_{ku} and K_{umax}.

Table 1: Significance level of PC-type Jarque–Bera (JB) testing for normally distributed data for Case-I compared with other methods

	$\alpha = 0.01$				$\alpha = 0.05$				$\alpha = 0.10$			
n	Z^*_{M1}	MJB^*_m	ZNT	JB_{sum}	Z^*_{M1}	MJB^*_m	ZNT	JB_{sum}	Z^*_{M1}	MJB^*_m	ZNT	JB_{sum}
$p = 5$												
$n = 25$	0.0195	0.0310	0.0005	0.0215	0.0605	0.0695	0.0205	0.0345	0.1090	0.1140	0.0635	0.0455
$n = 50$	0.0195	0.0280	0.0025	0.0250	0.0735	0.0815	0.0330	0.0500	0.1250	0.1305	0.0765	0.0725
$n = 100$	0.0205	0.0290	0.0105	0.0245	0.0675	0.0765	0.0510	0.0530	0.1205	0.1160	0.0970	0.0865
$n = 200$	0.0130	0.0140	0.0040	0.0275	0.0560	0.0630	0.0465	0.0680	0.1080	0.1085	0.0985	0.0985
$n = 500$	0.0110	0.0125	0.0070	0.0255	0.0540	0.0590	0.0485	0.0690	0.1110	0.1075	0.0900	0.1110
$p = 30$												
$n = 25$	0.1880	0.1880	0.3030	0.0215	0.1900	0.1900	0.3050	0.0345	0.1905	0.1905	0.3050	0.0450
$n = 50$	0.0005	0.0025	0.0000	0.0265	0.0195	0.0215	0.0000	0.0570	0.0540	0.0570	0.0000	0.0805
$n = 100$	0.0195	0.0220	0.0005	0.0335	0.0670	0.0685	0.0105	0.0700	0.1060	0.1070	0.0345	0.1020
$n = 200$	0.0175	0.0180	0.0030	0.0265	0.0755	0.0760	0.0295	0.0715	0.1390	0.1400	0.0715	0.1085
$n = 500$	0.0115	0.0120	0.0115	0.0250	0.0550	0.0560	0.0425	0.0680	0.1030	0.1040	0.0870	0.1130
$p = 50$												
$n = 25$	0.1890	0.1890	0.3000	0.0190	0.1900	0.1900	0.3000	0.0340	0.1905	0.1905	0.3000	0.0470
$n = 50$	0.1215	0.1215	0.1435	0.0255	0.1280	0.1280	0.1505	0.0495	0.1340	0.1340	0.1540	0.0725
$n = 100$	0.0050	0.0050	0.0000	0.0340	0.0350	0.0375	0.0005	0.0630	0.0850	0.0850	0.0030	0.0900
$n = 200$	0.0265	0.0270	0.0010	0.0295	0.0655	0.0660	0.0130	0.0660	0.1175	0.1170	0.0370	0.1000
$n = 500$	0.0180	0.0195	0.0060	0.0210	0.0595	0.0595	0.0330	0.0590	0.1130	0.1130	0.0815	0.1035
$p = 100$												
$n = 25$	0.1840	0.1840	0.3075	0.0135	0.1840	0.1840	0.3075	0.0225	0.1840	0.1840	0.3075	0.0410
$n = 50$	0.1260	0.1260	0.2055	0.0210	0.1265	0.1265	0.2060	0.0420	0.1270	0.1270	0.2060	0.0625
$n = 100$	0.0875	0.0875	0.1070	0.0285	0.0940	0.0940	0.1090	0.0560	0.0960	0.0960	0.1105	0.0845
$n = 200$	0.0080	0.0080	0.0000	0.0305	0.0355	0.0355	0.0000	0.0640	0.0715	0.0720	0.0005	0.0945
$n = 500$	0.0225	0.0225	0.0030	0.0160	0.0820	0.0820	0.0230	0.0615	0.1375	0.1375	0.0550	0.1030
$p = 200$												
$n = 25$	0.1730	0.1730	0.3055	0.0145	0.1735	0.1735	0.3055	0.0240	0.1735	0.1735	0.3055	0.0330
$n = 50$	0.1540	0.1540	0.2295	0.0225	0.1540	0.1540	0.2295	0.0395	0.1540	0.1540	0.2295	0.0580
$n = 100$	0.1235	0.1235	0.1815	0.0310	0.1235	0.1235	0.1815	0.0660	0.1235	0.1235	0.1815	0.0905
$n = 200$	0.0835	0.0835	0.0915	0.0205	0.0855	0.0855	0.0940	0.0540	0.0875	0.0875	0.0945	0.0905
$n = 500$	0.0090	0.0090	0.0000	0.0165	0.0505	0.0505	0.0000	0.0585	0.1095	0.1095	0.0010	0.0960

Table 2: Significance level of the PC-type JB testing for normally distributed data for Case-II compared with other methods

	$\alpha = 0.01$				$\alpha = 0.05$				$\alpha = 0.10$			
n	Z^*_{M1}	MJB^*_m	ZNT	JB_{sum}	Z^*_{M1}	MJB^*_m	ZNT	JB_{sum}	Z^*_{M1}	MJB^*_m	ZNT	JB_{sum}
$p = 5$												
$n = 25$	0.0245	0.0385	0.0005	0.0230	0.0735	0.0795	0.0260	0.0390	0.1195	0.1230	0.0665	0.0490
$n = 50$	0.0190	0.0275	0.0050	0.0255	0.0620	0.0700	0.0320	0.0505	0.1070	0.1090	0.0780	0.0705
$n = 100$	0.0195	0.0250	0.0055	0.0260	0.0715	0.0770	0.0370	0.0585	0.1215	0.1215	0.0925	0.0860
$n = 200$	0.0150	0.0220	0.0115	0.0230	0.0580	0.0660	0.0420	0.0565	0.1070	0.1060	0.0970	0.0925
$n = 500$	0.0095	0.0110	0.0090	0.0205	0.0515	0.0595	0.0460	0.0675	0.1010	0.1040	0.0915	0.1075
$p = 30$												
$n = 25$	0.1985	0.1985	0.3480	0.0180	0.2010	0.2010	0.3500	0.0300	0.2015	0.2015	0.3525	0.0430
$n = 50$	0.0015	0.0020	0.0000	0.0310	0.0170	0.0175	0.0000	0.0535	0.0470	0.0495	0.0000	0.0735
$n = 100$	0.0235	0.0250	0.0015	0.0260	0.0670	0.0680	0.0145	0.0580	0.1115	0.1110	0.0400	0.0920
$n = 200$	0.0225	0.0230	0.0035	0.0245	0.0715	0.0725	0.0345	0.0670	0.1300	0.1290	0.0740	0.1090
$n = 500$	0.0140	0.0140	0.0075	0.0205	0.0685	0.0680	0.0460	0.0640	0.1235	0.1240	0.0870	0.1150
$p = 50$												
$n = 25$	0.2605	0.2605	0.4230	0.0150	0.2610	0.2610	0.4240	0.0295	0.2620	0.2620	0.4265	0.0435
$n = 50$	0.1075	0.1075	0.1615	0.0245	0.1130	0.1130	0.1690	0.0470	0.1175	0.1175	0.1720	0.0710
$n = 100$	0.0065	0.0070	0.0000	0.0260	0.0475	0.0480	0.0005	0.0570	0.0875	0.0885	0.0030	0.0810
$n = 200$	0.0155	0.0155	0.0015	0.0275	0.0645	0.0645	0.0135	0.0615	0.1185	0.1190	0.0400	0.0925
$n = 500$	0.0240	0.0240	0.0100	0.0180	0.0705	0.0705	0.0420	0.0710	0.1210	0.1210	0.0775	0.1130
$p = 100$												
$n = 25$	0.2600	0.2600	0.4315	0.0145	0.2600	0.2600	0.4315	0.0305	0.2600	0.2600	0.4315	0.0400
$n = 50$	0.1970	0.1970	0.3345	0.0265	0.1975	0.1975	0.3345	0.0490	0.1975	0.1975	0.3345	0.0650
$n = 100$	0.0845	0.0845	0.1310	0.0295	0.0900	0.0900	0.1355	0.0605	0.0930	0.0930	0.1385	0.0865
$n = 200$	0.0075	0.0080	0.0000	0.0260	0.0395	0.0395	0.0000	0.0665	0.0740	0.0740	0.0005	0.1020
$n = 500$	0.0250	0.0250	0.0040	0.0210	0.0755	0.0760	0.0210	0.0665	0.1295	0.1295	0.0515	0.1025
$p = 200$												
$n = 25$	0.3050	0.3050	0.4800	0.0110	0.3050	0.3050	0.4805	0.0225	0.3050	0.3050	0.4810	0.0340
$n = 50$	0.2500	0.2500	0.4035	0.0265	0.2510	0.2510	0.4040	0.0515	0.2510	0.2510	0.4040	0.0740
$n = 100$	0.1475	0.1475	0.2465	0.0245	0.1475	0.1475	0.2465	0.0560	0.1475	0.1475	0.2465	0.0785
$n = 200$	0.0650	0.0650	0.0960	0.0250	0.0685	0.0685	0.0980	0.0545	0.0710	0.0710	0.0985	0.0830
$n = 500$	0.0100	0.0100	0.0005	0.0195	0.0570	0.0570	0.0010	0.0615	0.1055	0.1055	0.0030	0.1025

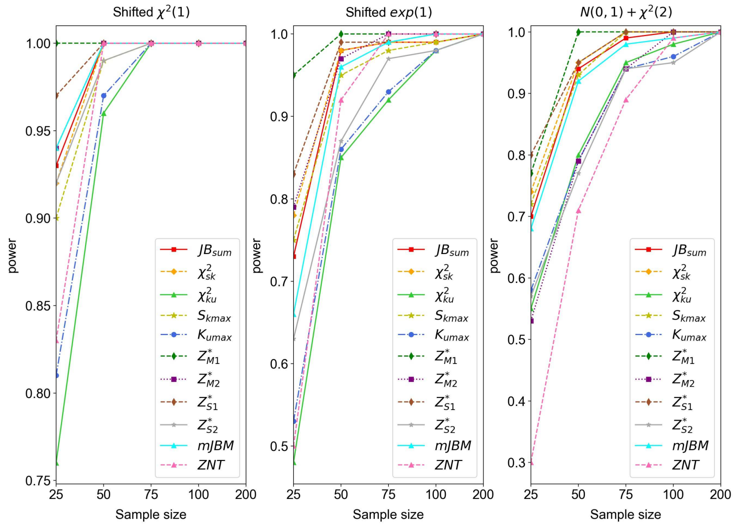

Figure 1. Empirical power of proposed PC-type JB testing compared with other methods (p=5)

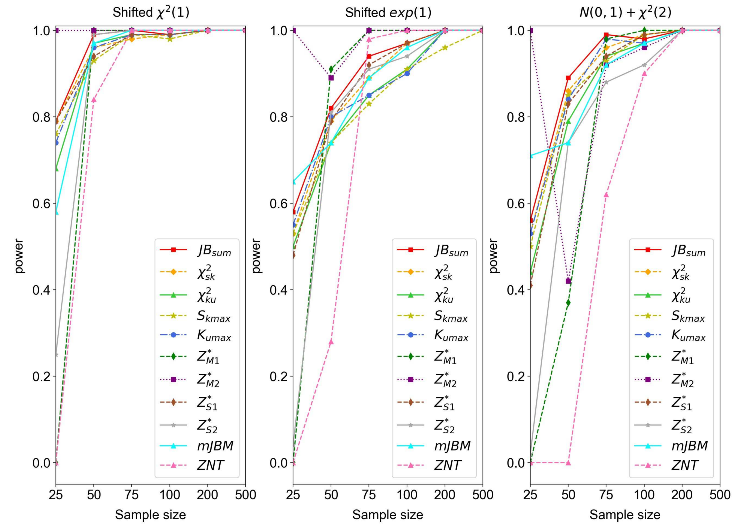

Figure 2. Empirical power of proposed PC-type JB testing compared with other methods (p=30)

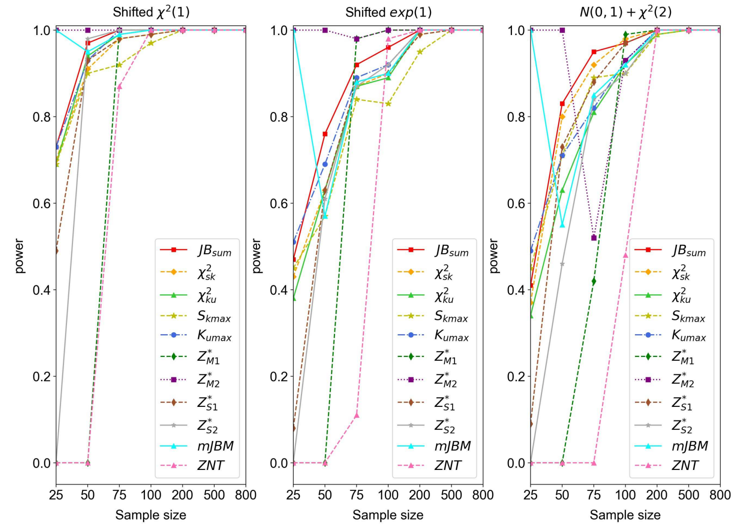

Figure 3. Empirical power of proposed PC-type JB testing compared with other methods (p=50)

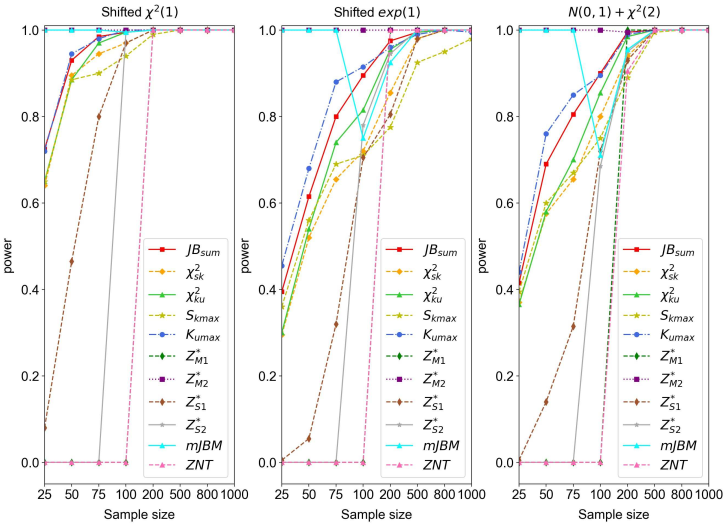

Figure 4. Empirical power of proposed PC-type JB testing compared with other methods (p=100)

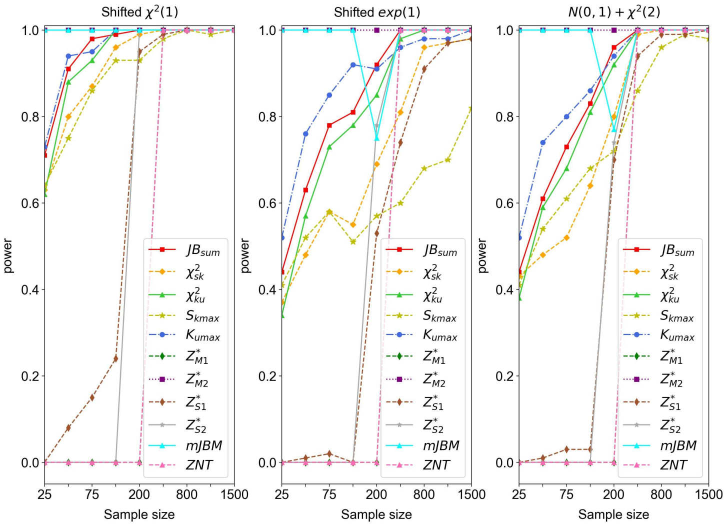

Figure 5. Empirical power of proposed PC-type JB testing compared with other methods (p=200)

(2) In the case of $p = 30$, as in Figure 2, although Z^*_{M1} and Z^*_{M2} perform better than JB_{sum} in Case IV, they do not maintain stable results like JB_{sum} in Case III. In fact, JB_{sum}'s performance is generally better than the other methods mentioned here among all three cases.

(3) In Figure 3, where $p = 50$, JB_{sum}'s performance is best among others except Z^*_{M2}. As in Figure 2, Z^*_{M2} is unstable in Case V when p is close to n. This phenomenon can also be seen in $mJBM$. Combining the information shown in Figure 2, we can see that Z^*_{M1}, Z^*_{M2}, and $mJBM$ are not as stable as JB_{sum}.

(4) With the increase in dimension, as seen in Figure 4, Z^*_{M1} no longer performs as well as before, and $mJBM$ is still not stable enough when n is close to p. Although K_{umax}'s performance is better than JB_{sum}'s at first, it is surpassed by the latter when $n > 100$.

(5) In Figure 5, as in $p = 100$, the power of K_{umax} is initially higher than JB_{sum}, and is eventually surpassed by JB_{sum}. Except for Z^*_{M2} and K_{umax}, JB_{sum}'s performance is the best.

From the phenomenon above, we may conclude that JB_{sum} performs well compared to the other statistics, in that its power is relatively higher than the others and the corresponding simulation results are more stable. Thus, it can be used to test the non-normality of low- or high-dimensional data effectively.

4. Two Real Examples

In this section, we investigated two real examples to illustrate the performance of our proposed method compared with the nine aforementioned existing methods.

4.1. SPECTF Heart data Example

The SPECTF heart dataset [15] provides data on cardiac single proton emission computed tomography (SPECT) images. It describes the diagnosis of cardiac single proton emission computed tomography (SPECT) images, and each patient is classified into two categories: normal and abnormal. The data contain 267 instances, with each instance belonging to a patient along with 44 continuous feature patterns summarized from the original SPECT images. The other attribute is an binary variable that indicates the diagnosis of each patient, with 0 for normal and 1 for abnormal.

In this dataset, we simultaneously evaluate the normality of the whole dataset and each class within it. The testing p-value of each method mentioned above is shown in Table 3.

Let S_0 describe the whole data set and S_1 and S_2 denote the normal class dataset and abnormal class dataset, respectively. We calculate the p-values of our PC-type statistic as well as the S_k-type and K_u-type statistics and other methods mentioned in [16,17] of these three datasets. Since all ten statistics' p-values of data S_0 and S_1 are very close to 0, we will not describe them here, which indicates a non-normal distribution of the whole dataset and abnormal dataset.

We may see from Table 3 that S_2's corresponding p-values are a little different from the former two sets, in which the p-values of χ^2_{sk}, Z^*_{M1} and MJB^*_M depart from 0. The relatively high p-values motivate us to conduct a detailed survey to investigate the normality of the SPECTF heart data's normal class by selecting some kinds of different variables that belong to a variety of degrees of normality.

In this normal category, we extract some variables and construct a new dataset S_3 from several experiments. The selected variables included in S_3 are $X_2, X_4, X_6, X_7, X_9 \sim X_{12}, X_{14} \sim X_{21}$, $X_{23} \sim X_{28}$, $X_{31} \sim X_{34}$, and $X_{37} \sim X_{43}$. We then compute the p-values of this dataset, and the results are shown in Table 3. It can be seen that all normality testing methods have a relatively high p-value, which demonstrates the multivariate normality of set S_3. For comparison, we constructed another two datasets,

Table 3: p-values of the six statistics of single proton emission computed tomography (SPECT) heart data

data set	χ^2_{sk}	χ^2_{ku}	S_{kmax}	K_{umax}	Z^*_{M1}	Z^*_{S1}	MJB^*_m	MJB^*_s	$mMJB$	JB_{sum}
S_2	0.2076	0.0141	0.0713	0.0000	0.4533	0.0329	0.4553	0.0241	0.1367	0.0138
S_3	0.5345	0.5318	0.6518	0.4207	0.0087	0.2109	0.0066	0.1935	0.4567	0.5560
S_4	0.1956	0.1201	0.3231	0.0728	0.0000	0.0244	0.0000	0.0050	0.0212	0.0780
S_5	0.0096	0.0045	0.0056	0.0038	0.0000	0.0415	0.0000	0.0111	0.0464	0.0003

S_4 and S_5, which consist of several verified normal variables and non-normal variables, respectively. Specifically, S_4 contains the variables $X_3, X_5, X_6 \sim X_8, X_{11} \sim X_{14}, X_{17}, X_{21}, X_{22}, X_{27} \sim X_{32}, X_{35}, X_{36}, X_{38}, X_{40}, X_{43}, and X_{44}$, while S_5 contains variables $X_3 \sim X_8, X_{13}, X_{15}, X_{22}, X_{29}, X_{30}, X_{35}, X_{36}, X_{42}, and X_{44}$. From Table 3 we can see the results of these two sets. This time, the p-values of the ten methods are no longer as high as before, meaning that our method performs well in assessing the normality of normal and non-normal data.

4.2. Body data example

In this part, we analyze the normality of body data investigated in [14] to show the consistency of our method with other existing methods and conclusions before. This data set contains 100 human individuals and each individual has 12 measurements of the human body (see [14] for details). As before, the p-values of the PC-type statistics and the S_k-type, K_u-type, and Kazuyuki's statistics are computed.

Let B_0 describe the whole dataset, and the multivariate normality of it can be investigated by the resulting p-values of each method shown in Table 3. Since all the p-values approach 0, we may conclude that this dataset contains non-normal data. As with the discussion in [14], we also investigate the other six datasets to show the validity of our proposed method, as well as making a comparison with other methods. For convenience, we denote $B_1 = (X_1, X_3, X_8, X_{10}, X_{12})$, $B_2 = (X_1, X_3, X_8, X_{10})$, $B_3 = (X_1, X_8, X_{10}, X_{12})$, $B_4 = (X_3, X_8, X_{10}, X_{12})$, $B_5 = (X_4, X_5, X_6, X_{11})$, and $B_6 = (X_2, X_4, X_6, X_{11})$. From Table 4, we can conclude that the normality testing results of our proposed PC-type statistic JB_{sum} are nearly the same as those for S_k-type statistics, K_u-type statistics, and Kazuyuki's methods. Since B_1, B_2, B_3, and B_4 have multivariate distribution, whereas B_5 and B_6 have non-normal distribution [14], our method is closer to the truth in the sense of higher p-values in multivariate normal situations and lower p-values in non-normal situations compared with the S_k-type and K_u-type statistics.

This phenomenon indicates that when testing real multivariate normal distributed data, our method results in a slightly higher p-value than the compared S_k-type and K_u-type statistics, whereas for non-normal distribution data, our method shows a relatively lower p-value. Thus, our PC-type statistic JB_{sum} constitutes a more effective way of testing normality both in normal data and non-normal data, with more stable testing results.

5. Conclusion

The purpose of this paper is to use a JB-type testing method to test high-dimensional normality. The statistics we proposed here used the generalized statistic JB_{sum} of JB statistics to test normality based on the dimensional reduction performed by PCA.

Through simulated experiments, we find that in both low and high dimensions, JB_{sum} performs well in testing normal and non-normal data, and it is more stable than many other compared methods. Therefore, it can be used to test normality effectively.

From two real examples, we can also see that our proposed method possesses the superiority of stability in performing the normality testing of real datasets, as well as the inclination of detecting the true normality from the perspective of p-values.

Table 4: p-values of the six statistics of body data

data set	χ^2_{sk}	χ^2_{ku}	S_{kmax}	K_{umax}	Z^*_{M1}	Z^*_{S1}	MJB^*_m	MJB^*_s	$mMJB$	JB_{sum}
B_0	0.0005	0.0046	0.0007	0.0007	0.0018	0.0051	0.0014	0.0037	0.0253	0.0000
B_1	0.6148	0.7214	0.5606	0.6502	0.5602	0.5568	0.5632	0.6345	0.9584	0.7879
B_2	0.3568	0.5468	0.3083	0.4704	0.1893	0.2897	0.2303	0.3771	0.8128	0.5087
B_3	0.6069	0.4335	0.5813	0.5116	0.3277	0.5817	0.3309	0.6405	0.8588	0.6097
B_4	0.6447	0.4297	0.5759	0.5776	0.7257	0.5863	0.5275	0.5285	0.5483	0.6280
B_5	0.0109	0.0628	0.0338	0.0422	0.0028	0.0163	0.0014	0.0099	0.0405	0.0048
B_6	0.0538	0.2003	0.1183	0.2662	0.1124	0.0290	0.1252	0.0221	0.0777	0.0533

Acknowledgments

This research was funded by National Natural Science Foundation of China (No. 11971214, 81960309), sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, Ministry of Education of China, and supported by Cooperation Project of Chunhui Plan of the Ministry of Education of China 2018. The authors would also like to thank Edit-in-chief and the referees for their suggestions to improve the paper.

References

1. Koziol, J. A. (1983). On assessing multivariate normality. *Journal of the Royal Statistical Society*, 45(3):358–361.
2. Slate, E. H. (1999). Assessing multivariate nonnormality using univariate distributions. *Biometrika*, 86(1):191–202.
3. Romeu, J. L. and Ozturk, A. (1993). A comparative study of goodness-of-fit tests for multivariate normality. *Journal of Multivariate Analysis*, 46:309–334.
4. Székely, G. J. and Rizzo, M. L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58 – 80.
5. Chiu, S. N. and Liu, K. I. (2009). Generalized cramér-von mises goodness-of-fit tests for multivariate distributions. *Computational Statistics and Data Analysis*, 53(11):3817 – 3834.
6. Small, N. J. H. (1980). Marginal skewness and kurtosis in testing multivariate normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(1):85–87.
7. Zhu, L.-X., Wong, H. L., and Fang, K.-T. (1995). A test for multivariate normality based on sample entropy and projection pursuit. *Journal of Statistical Planning and Inference*, 45(3):373 – 385.
8. Liang, J., Tang, M.-L., and Chan, P. S. (2009). A generalized shapiro-wilk w statistic for testing high-dimensional normality. *Computational Statistics and Data Analysis*, 53(11):3883 – 3891.
9. Doornik, J. A. and Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, 70:927–939.
10. Horswell, R. L. and Looney, S. W. (1992). A comparison of tests for multivariate normality that are based on measures of multivariate skewness and kurtosis. *Journal of Statistical Computation and Simulation*, 42(1-2):21–38.
11. Tenreiro, C. (2011). An affine invariant multiple test procedure for assessing multivariate normality. *Computational Statistics and Data Analysis*, 55(5):1980 – 1992.
12. Liang, J., Li, R., Fang, H., and Fang, K.-T. (2000). Testing multinormality based on low-dimensional projection. *Journal of Statistical Planning and Inference*, 86(1):129 – 141.
13. Jönsson, K. (2011). A robust test for multivariate normality. *Economics Letters*, 113(2):199–201.
14. Liang, J., Tang, M.-L., and Zhao, X. (2019). Testing high-dimensional normality based on classical skewness and kurtosis with a possible small sample size. *Communications in Statistics - Theory and Methods*, 48(23):5719-5732.
15. Dua, D. and Graff, C. (2017). UCI machine learning repository.
16. Kazuyuki, K., Masashi, H., and Tatjana, P. (2014). Modified Jarque-Bera Type Tests for Multivariate Normality in a High-Dimensional Framework. *Journal of Statistical Theory and Practice*, 8(2):382–399.
17. Kazuyuki, K., Naoya, O., and Takashi, S. (2008). On Jarque-Bera tests for assessing multivariate normality. *Journal of Statistics Advances in Theory and Applications*, 1(2):207–220.
18. Rie, E., Zofia, H., Ayako, H, and Takashi, S. (2020). Multivariate normality test using normalizing transformation for Mardia's multivariate kurtosis. *Communications in Statistics - Simulation and Computation*, 49(3):684–698.
19. Mardia, K.V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyá, , Series B*, 36:115–128.