

# Test for assessing multivariate normality which is available for high-dimensional data

Takayuki Yamada<sup>1,\*</sup>

*The Institute of Statistical Mathematics*

Tetsuto Himeno<sup>2</sup>

*Seikei University*

---

## Abstract

We proposed a test for assessing multivariate normality of the high-dimensional data which the dimension is larger than the sample size. The classical tests based on the sample measures of multivariate skewness and kurtosis defined by Mardia (1970) or Srivastava (1984) do not work for the high-dimensional case. The proposed test does not require explicit conditions on the relationship between the data dimension and sample size. An application of the proposed test is assessing multivariate normality of gene data, which we demonstrate as an numerical study.

*Keywords:* Assessing multivariate normality, High-dimension, Multivariate kurtosis.

*AMS 2010 subject classification:* primary 62H15, secondary 62E20

---

## 1. Introduction

Let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  be a random sample drawn from a population. We assume these  $p$ -dimensional observation vectors to the following models:

$$\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2} \mathbf{z}_i, \quad i = 1, \dots, N, \quad (1)$$

where  $\mathbf{z}_1, \dots, \mathbf{z}_N$  are independently and identically distributed (i.i.d.) as a distribution  $F$  with mean  $E[\mathbf{z}] = \mathbf{0}$  and covariance matrix  $\text{Var}(\mathbf{z}) = \mathbf{I}_p$ . The interest is whether  $F$  is multivariate normal distribution  $N_p(\mathbf{0}, \mathbf{I}_p)$  or not. If  $F = N_p(\mathbf{0}, \mathbf{I}_p)$ , multivariate analysis based on the normality gets meaningful.

---

\*Corresponding author

*Email addresses:* yma@ism.ac.jp (Takayuki Yamada), t-himeno@st.seikei.ac.jp (Tetsuto Himeno)

<sup>1</sup>Risk Analysis Research Center, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

<sup>2</sup>Department of Computer and Information Science, Faculty of Science and Technology, Seikei University, 3-3-1 Kichijoji-Kitamachi, Musashino-shi, Tokyo 180-8633, Japan

In a seminal paper Mardia [5], he defined a multivariate skewness  $\beta_{1,p}$  and a kurtosis  $\beta_{2,p}$  as

$$\begin{aligned}\beta_{1,p} &= E \left[ ((\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}))^3 \right], \\ \beta_{2,p} &= E \left[ ((\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}))^2 \right],\end{aligned}$$

where  $\mathbf{x}$  and  $\mathbf{y}$  are i.i.d. as a distribution with mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ . When the multivariate normality holds,

$$\beta_{1,p} = 0 \quad \text{and} \quad \beta_{2,p} = p(p+2).$$

Multivariate kurtosis is often defined as  $\gamma_{2,p} = \beta_{2,p} - p(p+2)$  in order to be 0 for the case that the distribution is multivariate normal. Srivastava [7] gave other definition of the multivariate skewness and kurtosis. Mardia [5] proposed estimators of  $\beta_{1,p}$  and  $\beta_{2,p}$  as

$$\begin{aligned}b_{1,p} &= \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N ((\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}))^3, \\ b_{2,p} &= \frac{1}{N} \sum_{i=1}^N ((\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}))^2,\end{aligned}$$

where  $\bar{\mathbf{x}}$  and  $\mathbf{S}$  are the sample mean and the unbiased estimator of  $\boldsymbol{\Sigma}$ , respectively, which are defined as follows:

$$\bar{\mathbf{x}} = \sum_{i=1}^N \mathbf{x}_i / N, \quad \mathbf{S} = \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' / n, \quad n = N - 1.$$

As their applications, he introduced the test of assessing multivariate normality based on these estimators. Srivastava [7] also proposed the method for assessing multivariate normality based on his estimators. For reviews of testing the multivariate normality, see e.g., Henze [3] or Mecklin and Mundfrom [6]. Koizumi et al. [4] proposed multivariate Jarque-Bera tests, which is an omnibus test using estimators of Mardia's, multivariate skewness and kurtosis, and also using Srivastava's multivariate skewness and kurtosis. Their proposed tests perform well when the sample size  $N$  is much larger than the dimension  $p$ .

It is noted that  $b_{1,p}$  and  $b_{2,p}$  are defined for the case that  $n \geq p$ . When  $p \geq N$ ,  $\mathbf{S}$  becomes singular, and so these estimators cannot be defined. In this paper, we define other measure  $\gamma_{2,p}^{(h)}$  like the multivariate kurtosis, which is as follows.

$$\gamma_{2,p}^{(h)} = \kappa_{11},$$

where

$$\kappa_{ij} = E[\mathbf{z}'\boldsymbol{\Sigma}^i \mathbf{z} \mathbf{z}'\boldsymbol{\Sigma}^j \mathbf{z}] - 2 \operatorname{tr} \boldsymbol{\Sigma}^{i+j} - \operatorname{tr} \boldsymbol{\Sigma}^i \operatorname{tr} \boldsymbol{\Sigma}^j$$

for positive integers  $i, j$  and  $\mathbf{z} \sim F$ . When  $\boldsymbol{\Sigma} = \mathbf{I}_p$  and  $p = 1$ ,  $\gamma_{2,1}^{(h)} = \gamma_{2,1}$ . When  $F = N_p(\mathbf{0}, \mathbf{I}_p)$ ,  $\gamma_{2,p}^{(h)} = \gamma_{2,p} = 0$ . Himeno and Yamada [2] proposed the unbiased estimator of  $\kappa_{11}$  as

$$\hat{\kappa}_{11} = \frac{-1}{(N-2)(N-3)} \{2(N-1)^2 \operatorname{tr} \mathbf{S}^2 + (N-1)^2 (\operatorname{tr} \mathbf{S})^2 - N(N+1)Q\},$$

where

$$Q = \frac{1}{N-1} \sum_{i=1}^N \{(\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}})\}^2.$$

Assume the following asymptotic framework A1 and assumptions A2 and A3:

$$\text{A1 : } p \rightarrow \infty, n = N - 1 \rightarrow \infty \text{ and } n/p \rightarrow c_0 \in (0, \infty);$$

$$\text{A2 : } a_i := \operatorname{tr} \boldsymbol{\Sigma}^i / p \rightarrow a_{i0} \in (0, \infty) \quad \text{for } i = 1, \dots, 6;$$

$$\text{A3 : } E[(\mathbf{y}'\boldsymbol{\Sigma}\mathbf{z})^4] = o(p^4), \quad \kappa_{22} = o(p^4), \quad \mathbf{y} \text{ and } \mathbf{z} \text{ are i.i.d. as } F.$$

Himeno and Yamada [2] has shown that if all elements of  $\mathbf{z} \sim F$  are i.i.d. and the eighth moment of the element is finite, then  $\kappa_{11} = O(p)$  under A1 and A2. The consistency of the unbiased estimator  $\hat{\kappa}_{11}/p$  has also been shown under A1, A2 and A3. We deal with the testing problem for the null hypothesis

$$H_0 : F = N_p(\mathbf{0}, \mathbf{I}_p)$$

against the alternative hypothesis that  $F$  is not  $N_p(\mathbf{0}, \mathbf{I}_p)$ . The multivariate normality is assessed by evaluating the value of  $\gamma_{2,p}^{(h)} = \kappa_{11}$ , i.e., the null hypothesis  $H_0$  is rejected by verifying  $\kappa_{11} \neq 0$ . We will propose the testing statistic based on the estimator  $\hat{\kappa}_{11}/p$ .

## 2. Asymptotic null distribution

We derive the limiting null distribution of  $\hat{\kappa}_{11}/p$ . From model (1) it can be expressed that

$$\begin{aligned}
\hat{\kappa}_{11} &= \frac{1}{N} \sum_{i=1}^N (z_i' \Sigma z_i)^2 - \frac{2}{N(N-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^N (z_i' \Sigma z_j)^2 - \frac{1}{N(N-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^N z_i' \Sigma z_i z_j' \Sigma z_j \\
&\quad - \left\{ \frac{4}{N(N-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^N z_i' \Sigma z_i z_i' \Sigma z_j - \frac{4}{N(N-1)(N-2)} \sum_{\substack{i,j,k=1 \\ i \neq j, j \neq k, k \neq i}}^N z_i' \Sigma z_i z_i' \Sigma z_k \right\} \\
&\quad + \frac{8}{N(N-1)(N-2)} \sum_{\substack{i,j,k=1 \\ i \neq j, j \neq k, k \neq i}}^N z_i' \Sigma z_j z_i' \Sigma z_k \\
&\quad - \frac{6}{N(N-1)(N-2)(N-3)} \sum_{\substack{i,j,k,\ell=1 \\ i \neq j \neq k \neq \ell, k \neq i \neq \ell \neq j}}^N z_i' \Sigma z_j z_k' \Sigma z_\ell.
\end{aligned}$$

From the assumption of  $H_0 : F = N_p(0, \Sigma)$ , the expression in the curly brace and the last two summations converge in probability to 0, respectively, under the asymptotic framework A1 and the assumption A2, and so we have

$$\begin{aligned}
\frac{N^{1/2}}{p} \hat{\kappa}_{11} &= \frac{1}{pN^{1/2}} \sum_{i=1}^N (z_i' \Sigma z_i)^2 - \frac{2}{pN^{1/2}(N-1)} \sum_{i \neq j}^N (z_i' \Sigma z_j)^2 \\
&\quad - \frac{1}{pN^{1/2}(N-1)} \sum_{i \neq j}^N z_i' \Sigma z_i z_j' \Sigma z_j + o_p(1)
\end{aligned}$$

under A1 and A2, where we abbreviate the notation  $\sum_{\substack{i,j=1 \\ i \neq j}}^N$  as  $\sum_{i \neq j}^N$ , simply. By subtracting the mean, it can be expressed as

$$\begin{aligned}
\frac{N^{1/2}}{p} \hat{\kappa}_{11} &= \frac{1}{pN^{1/2}} \sum_{i=1}^N \{(z_i' \Sigma z_i)^2 - 2 \operatorname{tr} \Sigma^2 - (\operatorname{tr} \Sigma)^2\} - \frac{2a_1}{N^{1/2}} \sum_{i=1}^N (z_i' \Sigma z_i - \operatorname{tr} \Sigma) \\
&\quad - \frac{1}{pN^{1/2}(N-1)} \sum_{i \neq j}^N (z_i' \Sigma z_i - \operatorname{tr} \Sigma)(z_j' \Sigma z_j - \operatorname{tr} \Sigma) \\
&\quad - \frac{2}{pN^{1/2}(N-1)} \sum_{i \neq j}^N \{(z_i' \Sigma z_j)^2 - \operatorname{tr} \Sigma^2\} + o_p(1). \tag{2}
\end{aligned}$$

For the third and fourth terms, the following convergences hold under A1 and A2:

$$\begin{aligned} \frac{1}{p^2 N(N-1)^2} \text{Var} \left( \sum_{i \neq j}^N (\mathbf{z}'_i \boldsymbol{\Sigma} \mathbf{z}_i - \text{tr} \boldsymbol{\Sigma})(\mathbf{z}'_j \boldsymbol{\Sigma} \mathbf{z}_j - \text{tr} \boldsymbol{\Sigma}) \right) &= \frac{8a_2^2}{(N-1)} \rightarrow 0, \\ \frac{4}{p^2 N(N-1)^2} \text{Var} \left( \sum_{i \neq j}^N \{(\mathbf{z}'_i \boldsymbol{\Sigma} \mathbf{z}_j)^2 - \text{tr} \boldsymbol{\Sigma}^2\} \right) &= \frac{16}{(N-1)} \left\{ a_2^2 + \frac{(2N-1)a_4}{p} \right\} \rightarrow 0. \end{aligned}$$

From Chebyshev's inequality, the third and the fourth terms of the right-hand side of the equation (2) converge to 0 in probability, respectively. Thus it is found that

$$\frac{\sqrt{N}}{p} \hat{\kappa}_{11} = \frac{1}{\sqrt{N}} \sum_{i=1}^N \eta_i + o_p(1),$$

where

$$\eta_i = \frac{1}{p} \{(\mathbf{z}'_i \boldsymbol{\Sigma} \mathbf{z}_i)^2 - 2 \text{tr} \boldsymbol{\Sigma}^2 - (\text{tr} \boldsymbol{\Sigma})^2\} - 2a_1(\mathbf{z}'_i \boldsymbol{\Sigma} \mathbf{z}_i - \text{tr} \boldsymbol{\Sigma}).$$

The random variables  $\eta_1, \dots, \eta_N$  are i.i.d. with the mean 0 and the variance  $8a_2^2 + (48/p)a_4$ . Since the variance does not include  $n$ , from the central limit theorem,  $N^{1/2}(1/N) \sum_{i=1}^N \eta_i$  converges in distribution to the normal distribution with mean 0 and variance  $8a_2^2_0$  as  $n, p$  tend to infinity together along the asymptotic framework A1 under the assumption A2. Himeno and Yamada [2] proposed the unbiased and the consistent estimator  $\hat{a}_2$  under A1, A2 and A3, which is as follows.

$$\hat{a}_2 = \frac{N-1}{pN(N-2)(N-3)} \{(N-1)(N-2) \text{tr} \mathbf{S}^2 + (\text{tr} \mathbf{S})^2 - NQ\}.$$

From Slutsky's theorem of convergence in distribution, we obtain the asymptotic null distribution of  $\hat{\kappa}_{11}/p$ , which is given as the following theorem.

**Theorem 1.** *Assume that the null hypothesis  $H_0 : F = N_p(\mathbf{0}, \boldsymbol{\Sigma})$  is true. Under the asymptotic framework A1 and the assumptions A2,  $(N^{1/2} \hat{\kappa}_{11}/p)/(8^{1/2} \hat{a}_2)$  converges in distribution to the standard normal distribution.*

### 3. Numerical studies

The simulation experiments were carried out to see the precision of the approximation for the case that  $N, p = 60, 80, 120, 200$ . We calculated the actual error probability of the first kind when the nominal is 0.05. We did Monte Carlo simulation based on 100,000 iteration. Generate the data based on the model (1). As a covariance matrix, we consider two cases. One is the identity matrix, i.e.,

$\Sigma_1 = \mathbf{I}_p$ , and the other is an positive definite matrix  $\Sigma_2$  as

$$\Sigma_2 = \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_p \end{pmatrix} \begin{pmatrix} 0.1^{|1-1|} & 0.1^{|1-2|} & \dots & 0.1^{|1-p|} \\ 0.1^{|2-1|} & 0.1^{|2-2|} & \dots & 0.1^{|2-p|} \\ \vdots & \vdots & \ddots & \vdots \\ 0.1^{|p-1|} & 0.1^{|p-2|} & \dots & 0.1^{|p-p|} \end{pmatrix} \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_p \end{pmatrix}$$

for  $\sigma_i = 5 + (-1)^{i-1}(p - i + 1)/p$ . We can see from Table 1 that the actual error probability of the first kind is almost monotone decreasing for  $N$  and  $p$ , and larger than 0.05. It should be noted that our approximation method underestimates the percentile points. We also calculated the power for the case that the nominal is 0.05. The distributions of  $\mathbf{z} = (z_1, \dots, z_p)' \sim F$  under alternative hypothesis are considered as follows.

Case 1:  $F$  is the scaled multivariate  $t$  distribution with 500 degrees of freedom, the mean  $\mathbf{0}$  and the covariance matrix  $\mathbf{I}_p$ ;

Case 2: For  $w_1, \dots, w_p$  are i.i.d. as the chi-squared distribution with 10 degrees of freedom,  $z_i = \sqrt{10}(w_i/10 - 1)/\sqrt{2}$ ,  $i = 1, \dots, p$ .

Tables 2-3 indicate values of powers, which are corresponding to Case 1-2, based on Monte Carlo simulation with 100,000 repetition when the nominal is 0.05. We can see from Table 2 that the power is almost monotone increasing for  $N$  and  $p$ . The reason why it occurs is that the discrepancy from the standard normal distribution gets large as  $p$  becomes large when the degrees of freedom of the multivariate  $t$  distribution is fixed. It can be found from Table 3 that the power is almost monotone increasing for  $N$  and is not affected by the value of  $p$ .

Table 1: Actual error probability of the first kind when the nominal level is 0.05.

| $N$ | $p$ | $\Sigma = \Sigma_1 = \mathbf{I}_p$ | $\Sigma = \Sigma_2$ | $N$ | $p$ | $\Sigma = \Sigma_1 = \mathbf{I}_p$ | $\Sigma = \Sigma_2$ |
|-----|-----|------------------------------------|---------------------|-----|-----|------------------------------------|---------------------|
| 60  | 60  | 0.064                              | 0.067               | 120 | 60  | 0.058                              | 0.061               |
|     | 80  | 0.063                              | 0.064               |     | 80  | 0.057                              | 0.058               |
|     | 120 | 0.062                              | 0.065               |     | 120 | 0.058                              | 0.058               |
|     | 200 | 0.061                              | 0.064               |     | 200 | 0.056                              | 0.057               |
| 80  | 60  | 0.061                              | 0.064               | 200 | 60  | 0.057                              | 0.058               |
|     | 80  | 0.058                              | 0.063               |     | 80  | 0.057                              | 0.057               |
|     | 120 | 0.058                              | 0.061               |     | 120 | 0.055                              | 0.056               |
|     | 200 | 0.058                              | 0.060               |     | 200 | 0.054                              | 0.056               |

We apply our test to two dataset:

**Colon data** 2000( $p$ ) genes expression levels are available on 22 ( $N_1$ ) normal colon tissues and 40 ( $N_2$ ) tumor colon tissues, which these data are publically available at “<http://genomcs-pubs.princeton>.”

Table 2: Power for the multivariate T distribution when the nominal is 0.05

| $N$ | $p$ | $\Sigma = \Sigma_1 = \mathbf{I}_p$ | $\Sigma = \Sigma_2$ | $N$ | $p$ | $\Sigma = \Sigma_1 = \mathbf{I}_p$ | $\Sigma = \Sigma_2$ |
|-----|-----|------------------------------------|---------------------|-----|-----|------------------------------------|---------------------|
| 60  | 60  | 0.151                              | 0.145               | 120 | 60  | 0.199                              | 0.187               |
|     | 80  | 0.199                              | 0.188               |     | 80  | 0.278                              | 0.258               |
|     | 120 | 0.307                              | 0.285               |     | 120 | 0.463                              | 0.424               |
|     | 200 | 0.547                              | 0.509               |     | 200 | 0.787                              | 0.746               |
| 80  | 60  | 0.168                              | 0.158               | 200 | 60  | 0.264                              | 0.240               |
|     | 80  | 0.227                              | 0.207               |     | 80  | 0.386                              | 0.356               |
|     | 120 | 0.362                              | 0.333               |     | 120 | 0.637                              | 0.591               |
|     | 200 | 0.645                              | 0.600               |     | 200 | 0.933                              | 0.906               |

Table 3: Power for the independent case when the nominal is 0.05

| $N$ | $p$ | $\Sigma = \Sigma_1 = \mathbf{I}_p$ | $\Sigma = \Sigma_2$ | $N$ | $p$ | $\Sigma = \Sigma_1 = \mathbf{I}_p$ | $\Sigma = \Sigma_2$ |
|-----|-----|------------------------------------|---------------------|-----|-----|------------------------------------|---------------------|
| 60  | 60  | 0.756                              | 0.730               | 120 | 60  | 0.941                              | 0.929               |
|     | 80  | 0.753                              | 0.738               |     | 80  | 0.945                              | 0.937               |
|     | 120 | 0.761                              | 0.748               |     | 120 | 0.950                              | 0.942               |
|     | 200 | 0.768                              | 0.755               |     | 200 | 0.953                              | 0.947               |
| 80  | 60  | 0.841                              | 0.821               | 200 | 60  | 0.993                              | 0.990               |
|     | 80  | 0.845                              | 0.833               |     | 80  | 0.994                              | 0.992               |
|     | 120 | 0.854                              | 0.844               |     | 120 | 0.995                              | 0.994               |
|     | 200 | 0.861                              | 0.851               |     | 200 | 0.995                              | 0.995               |

edu/oncology/affydata/index.html". We preprocessed the data by applying 10 logarithmic transformation.

**Leukemia data** 3571( $p$ ) gene expressions are available on 47 patients suffering from acute lymphoblastic leukemia (ALL,47 cases) and 25 patients suffering from acute myeloid leukemia (AML,25 cases), which these data are publically available at "<http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi>". The dataset are preprocessed by following protocol written in Dudoit et al. [1].

For the colon data, the  $p$ -value of our test is 0.237 for normal colon tissues and 0.063 for tumor colon tissues. This indicates that the multivariate normality assumption on both sets of data cannot be rejected at the usual significance level 5%. For the leukemia data, the  $p$ -value of our test is 0.663 for ALL and is 0.838 for AML. It is not suspected for the conclusion of multivariate analysis for the leukemia data derived under multivariate normal distribution.

#### 4. Concluding and remarks

For the case that the sample size  $N$  is larger than the dimension  $p$ , it has been reported some tests for assessing multivariate normality based on multivariate skewness and kurtosis. However, when  $p > N$ , all these tests cannot be defined. In this paper, we proposed a testing statistic for assessing multivariate normality which is available for the high-dimensional case. Simulation results revealed that our proposed test has good precision for the percentile point approximation for large  $N$  and  $p$ . In addition, power is almost monotone increasing for  $N$ . We applied two dataset of microarray data.

#### Acknowledgement

The first author was supported by grant from Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research, Young Scientists (B), #23740085, 2011-2012.

#### References

- [1] S. Dudoit, J. Fridlyand, T.P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, *J. Amer. Statist. Assoc.* 97 (2002) 77–87.
- [2] T. Himeno, T. Yamada, Estimations for some functions of covariance matrix in high dimension under non-normality, TR No.12-11. Statistical Research Group, Hiroshima University.
- [3] N. Henze, Invariant tests for multivariate normality: a critical review, *Stat. Papers* 43 (2002) 467–506.
- [4] K. Koizumi, N. Okamoto, T. Seo, On Jarque-Bera tests for assessing multivariate normality, *J. Stat., Adv. Theory Appl.* 1 (2009) 207–220.
- [5] K. V. Mardia, Measures of multivariate skewness and kurtosis with applications, *Biometrika* 57 (1970) 519–530.
- [6] C.J. Mecklin, D.J. Mundfrom, An appraisal and bibliography of tests for multivariate normality, *Internat. Statist. Rev.* 72 (2004) 123–138.
- [7] M.S. Srivastava, A measure of skewness and kurtosis and a graphical method for assessing multivariate normality, *Statist. Probab. Lett.* 2 (1984) 263–267.