Behavior Research Methods (2020) 52:939–946
https://doi.org/10.3758/s13428-019-01291-5

A method of generating multivariate non-normal random numbers with desired multivariate skewness and kurtosis

Wen Qu[1] · **Haiyan Liu**[2] · **Zhiyong Zhang**[1]

Published online: 26 August 2019

Abstract

In social and behavioral sciences, data are typically not normally distributed, which can invalidate hypothesis testing and lead to unreliable results when being analyzed by methods developed for normal data. The existing methods of generating multivariate non-normal data typically create data according to specific univariate marginal measures such as the univariate skewness and kurtosis, but not multivariate measures such as Mardia's skewness and kurtosis. In this study, we propose a new method of generating multivariate non-normal data with given multivariate skewness and kurtosis. Our approach allows researchers to better control their simulation designs in evaluating the influence of multivariate non-normality.

Keywords Multivariate non-normal data · Multivariate skewness · Multivariate kurtosis · Random number generation

Introduction

In social and behavioral sciences, the normality of data is assumed in most statistical methods. Nonetheless, data are rarely normally distributed in practice. Therefore, the statistical inferences may not be valid, and the results may not be reliable any more when procedures developed for normal data are used to analyze non-normal data (Cain, Zhang, & Yuan, 2017; Micceri, 1989). Many studies in the literature investigated the consequences of the violation of the normality assumption and proposed some alternative procedures to analyze non-normal data. For instance, Bradley (1980) showed that robustness of statistical procedures such as the classical Z, t, and F tests suffered from the non-normality of data. Non-parametric tests and procedures have won appreciation of researchers because they do not rely on the data distribution and therefore, the violation of normality does not directly disqualify data analysis (Hollander & Wolfe, 2015).

In the literature, discussions on non-normality mainly focus on the univariate case; whereas the consequences of deviation from the multivariate normality are less explored. However, the analysis of multivariate data is routinely conducted in social and behavioral sciences research. Therefore, it is important to understand the influence of the multivariate non-normality on the multivariate analysis, which can be done through Monte Carlo simulations. To conduct such simulations, one needs to generate multivariate data with the control of the degree of non-normality. In the literature, most non-normal data generators are developed for univariate data, such as the third-order polynomial power method (the power method; Fleishman, 1978), the fifth-order polynomial method (Headrick, 2002), and the g-h distribution method (Field & Genton, 2012).

The existing methods typically generate multivariate data according to specific univariate marginal measures such as the univariate skewness and kurtosis, but not multivariate measures such as Mardia's (1970) skewness and kurtosis. For example, the widely used simulation method proposed by Vale and Maurelli (1983) (VM) was built on Fleishman's (1978) polynomial approach. In addition to generating data for each variable with specific first four moments, their method also controls for a correlation matrix that allows researchers to have a desired multivariate data

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-019-01291-5) contains supplementary material, which is available to authorized users.

✉ Wen Qu
wqu@nd.edu

[1] Department of Psychology, University of Notre Dame, Corbett Family Hall, Notre Dame, IN 46556, USA

[2] Psychological Sciences, University of California, Merced, CA, USA

structure. This method is very popular in the moment-based modeling area, such as structural equation modeling (SEM). However, some researchers have questioned the generalization of this method. Foldnes and Grønneberg (2015) derived the mathematical distribution of the VM approach and showed that even though the approach could generate multivariate data with user-specified marginal skewness and kurtosis, the generated data might not be truly multivariate non-normal. Astivia and Zumbo (2015) have shown that the Vale and Maurelli method has downward bias. In one of their later papers, Astivia and Zumbo (2018) also found the multiplicity solution issue of the Fleishman's polynomial-related method, which means that there are multiple possible solutions for the polynomial coefficients (a, b, c, and d). This issue might lead to the difference in the analysis even with the same inputs. To remedy the drawback, researchers have developed other methods. Mair, Satorra, and Bentler (2012), for example, introduced a multivariate approach based on copulas that could also generate data with a pre-specified variance-covariance matrix. Foldnes and Olsson (2016) presented a method using linear combinations of independent generator variables. Additionally, Lee and Kaplan (2018) developed a generator for the multivariate ordinal data based on entropy procedures.

Despite their usefulness, none of these methods allows the direct control of the multivariate non-normality measures. Multivariate skewness and kurtosis have been shown to directly impact statistical analysis. For example, Yuan, Bentler, and Zhang (2005) noted that a robust procedure might be necessary for reliable SEM inferences when a sample has a large multivariate kurtosis. More recently, Cain et al. (2017) conducted a meta-analysis study on the multivariate non-normality of the data used in 254 published studies and found that the type I error rates of testing the model fit were remarkably higher in factor analysis when the multivariate normality was violated. Generating multivariate non-normal data with desired multivariate measures is the first step to understanding the type and severity of non-normality. This is because it relates to both multivariate skewness and kurtosis on analysis procedures.

Generating multivariate non-normal random data requires the understanding of the definition of the non-normality and the relationship between univariate and multivariate data. Mardia (1970) introduced the measures of population multivariate skewness and kurtosis as the natural extension of the univariate ones. In the univariate case, with non-zero skewness, the distribution is asymmetry. When the excess kurtosis is not 0 (excess kurtosis equals to kurtosis minus 3), the distribution density function is different from a normal distribution. Similarly, Mardia's multivariate kurtosis indicates whether the tails are heavy or light in comparison to those of the multivariate normal distribution (DeCarlo, 1997). On the other hand, the Mardia's skewness is still a measure of symmetry, but cannot take negative values. Higher values indicate severer asymmetry.

To date, there is no available method for researchers to directly specify both multivariate skewness and kurtosis for multivariate non-normal data generation. To fill the gap, we introduce a new method of generating multivariate non-normal data with specific multivariate measures. This approach allows researchers to better control their simulation design in evaluating the influence of the multivariate non-normality. More over, this technique will allow for a better understanding of the relationship between multivariate non-normality and the marginal univariate non-normality.

The rest of the paper is organized as follows. We first propose a new generating method and introduce an R package for the implementation of the method. We then present a simulation study and the results with various conditions. We conclude the study with a summary of our method.

Method

Data model

To generate the non-normal data, we specify the following data model. We use a vector $\mathbf{x}$ of p variables as,

$$\mathbf{x} = r\mathbf{A}\boldsymbol{\xi}, \tag{1}$$

and each marginal x_i as,

$$x_i = r\sum_{j=1}^{q} a_{ij}\xi_j, \tag{2}$$

where $\boldsymbol{\xi} = (\xi_1, ..., \xi_q)$ is a vector containing q independent random variables. Each of the variables ξ_j has the first four ordered moments $E(\xi_j) = 0$, $E(\xi_j^2) = 1$, $E(\xi_j^3)$, and $E(\xi_j^4)$. $\mathbf{A} = (a_{ij})$ is a $p \times q$ matrix of rank p ($p \leq q$), and $\mathbf{A}\mathbf{A}^t = \Sigma = cov(\mathbf{x})$, and r is a random variable, which is independent of $\boldsymbol{\xi}$, with the first four ordered moments $E(r)$, $E(r^2)$, $E(r^3)$, and $E(r^4)$.

The ordered moments are a set of quantitative measures describing the shape of a distribution. When the ordered moments are normalized, they become the standardized moments (or central moments). Skewness and kurtosis are the third and fourth standardized moments. The ordered and standardized moments can convert to each other as long as mean and variance are provided.

According to the definition of Mardia (1970), the population multivariate skewness (β_1) and kurtosis (β_2) of $\mathbf{x}$ based on our model are computed as,

$$\beta_1 = E\{[(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})]^3\} = [E(r^3)]^2 \sum_{j=1}^{q}[E(\xi_j^3)]^2, \tag{3}$$

$$\begin{aligned}\beta_2 &= E\{[(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]^2\} \\ &= E(r^4)[\sum_{j=1}^{q} E(\xi_j^4) + p(p-1)]. \end{aligned} \tag{4}$$[1]

Using these formulas, population multivariate skewness and kurtosis can be calculated when univariate measures[2], the ordered moments of r and ξ_j, are given. The standardized multivariate kurtosis formula, centering β_2 by $p(p+2)$, has been obtained in Yuan, Zhang, and Zhao (2017). However, the solution of univariate measures cannot be uniquely obtained based on these formulas from specified multivariate measures.

Although the solution is not on a one-to-one basis, multiple solutions are available that share the same multivariate measures. Since we only care about the measures at the multivariate level rather than the univariate level (or the marginal level), to remedy the lack of uniqueness of the solution, we establish one from multivariate to univariate by applying some constraints.

First, we set r to be a constant 1 for convenience because it is only a scale factor. Second, the number of variables in $\boldsymbol{\xi}$ is set to be the same as the number of variables in $\mathbf{X}$, so that $p = q$. Additionally, ξ_1 to ξ_p are set to be independent and identically distributed (i.i.d.). Thus, the 3rd- and 4th-ordered moments are the same for all ξ_j, which are defined as $E(\xi^3)$ and $E(\xi^4)$. With these constraints, the multivariate skewness and kurtosis above become,

$$\beta_1^* = E\{[(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})]^3\} = p[E(\xi^3)]^2, \tag{5}$$

$$\beta_2^* = E\{[(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]^2\} = pE(\xi^4) + p(p-1). \tag{6}$$

When multivariate measures (β_1^* and β_2^*) and number of variables (p) are provided, $E(\xi^3)$ and $E(\xi^4)$ can be computed through Equations (5) and (6). Once we have the moments for all ξ_j (i.e., $E(\xi) = 0$, $E(\xi^2) = 1$, $E(\xi^3)$ and $E(\xi^4)$), we can generate random numbers of $\boldsymbol{\xi}$ and then transform them to multivariate random numbers of $\mathbf{x}$ by Eq. 1.

We acknowledge that relaxing some of the constraints would give researchers more control of the univariate measures (e.g., allowing different univariate measures or setting different scaling factors). However, based on a small-scale simulation, we found that they did not influence the behavior of the multivariate measures much (see Table 2 in the supplementary materials). Therefore, we use the constraints for convenience in this study.

For each ξ_j, we use a modified power method to generate random non-normal numbers. The widely used power method is proposed by Fleishman (1978), which is to generate non-normal data through a polynomial transformation

$$Y = a + bZ + cZ^2 + dZ^3, \tag{7}$$

where Z comes from a standard normal distribution. With the information of the first four desired standardized moments (mean, variance, skewness γ_1, and kurtosis γ_2) of Y, Fleishman (1978) derived four equations to obtain the coefficients a, b, c, and d through Newton's method.

In our study, Y is replaced by each ξ_j as

$$\xi_j = a + bZ + cZ^2 + dZ^3. \tag{8}$$

Instead of including the standardized moments (skewness and kurtosis) as used in Fleishman's method, we use the third- and fourth-ordered moments of ξ_j. Therefore, the four equations of solving the coefficients are revised as

$$a + c = 0, \tag{9}$$

$$b^2 + 6bd + 2c^2 + 15d^2 - 1 = 0, \tag{10}$$

$$72bcd + 6b^2c + 8c^3 + 270cd^2 - E(\xi^3) = 0, \tag{11}$$

$$3b^4 + 60b^2c^2 + 60c^4 + 60b^3d + 936bc^2d + 630b^2d^2 \tag{12}$$

$$+4500c^2d^2 + 3780bd^3 + 10395d^4 - E(\xi^4) = 0. \tag{13}$$

With the value of a, b, c, and d, we first generate random numbers from the standard normal distribution to form the sample Z with size n. Then, the sample of ξ_j is obtained by the polynomial transformation in Eq. 8. Repeatedly sampling Z and conducting transformation for each ξ_j, one gets the multivariate data with $\boldsymbol{\xi} = (\xi_1, ..., \xi_p)$. The final step is to obtain $\mathbf{x}$ by applying the specific covariance matrix $\mathbf{A}$ to $\boldsymbol{\xi}$ following the data model in Eq. 1 with $r = 1$.

In summary, the following procedure can be used.

1. With the user-specified multivariate skewness (β_1^*) and kurtosis (β_2^*) and the number of variables (p), calculate the third- and fourth-ordered moments of ξ_j ($j = 1, 2, \cdots, p$).

[1] According to Mardia's definition, $\mathbf{y}$ is an identical but independent distribution of $\mathbf{x}$. Since the skewness is a measure of asymmetry, the inside of the power 3 could not be a symmetrical measure (like the even power). The multivariate skewness is like the "squared version" of the univariate one.

[2] In this paper, when it is in the multivariate setting, we refer the moments of x_i as marginal measures, and the univariate measures are the moments of r and ξ_j. The marginal skewness and kurtosis of x_i are $\gamma_1(x_i) = E(r^3)\sum_{j=1}^{q} a_{ij}^3 E(\xi_j^3)/\sigma_{ii}^{3/2}$, $\gamma_2(x_i) = E(r^4)[\sum_{j=1}^{q} a_{ij}^4 (E(\xi_j^4) - 3)/\sigma_{ii}^2 + 3]$ (Yuan & Bentler, 1997).

2. Generate the standardized ξ_j by the modified power method to form $\boldsymbol{\xi}$.
3. Use the Cholesky decomposition to decompose the user-specified correlation matrix (or covariance matrix) to matrix $\mathbf{A}$ and multiply it to $\boldsymbol{\xi}$ ($\mathbf{x} = A\boldsymbol{\xi}$).

Through this process, the generated data $\mathbf{x}$ will have the desired multivariate skewness and kurtosis. One shortcoming of applying the Cholesky decomposition approach is that, after the linear transformations, the population marginal measures of $\mathbf{x}$ (e.g., γ_1 and γ_2) will be different from the original univariate measures of $\boldsymbol{\xi}$. However, unlike the VM method, where the marginal measures are of interest, the focus of our method is the multivariate measures and the marginal measures are nuisance parameters. Therefore, our method does not require an intermediate correlation matrix and can apply the Cholesky decomposition directly.

Limited ranges of the skewness and kurtosis

The power method cannot cover all the possible combinations of univariate skewness and kurtosis. This is because the method does not require the distribution of Y, and thus the moments cannot be analytically derived. However, the range relationship of univariate skewness (γ_1) and kurtosis (γ_2) of Y has been estimated through simulation (Luo, 2011). Based on that relationship, we derived the range relationship between the univariate ξ_j's third and fourth moments with our modified power method, which is

$$E(\xi^4) \geq 1.641[E(\xi^3)]^2 + 1.774. \tag{14}$$

Plugging it into the data model in Eq. 1, the relationship of x_i's skewness and kurtosis is,

$$\gamma_2 \geq \frac{1.641}{\sum a_{ij}^4}\left(\frac{\gamma_1}{\sum a_{ij}^3}\right)^2 - 1.226\sum a_{ij}^4 + 3, \tag{15}$$

which is restricted compared to the theoretical relationship of the general univariate skewness and kurtosis,

$$\gamma_2 \geq \gamma_1^2 + 1. \tag{16}$$

Correspondingly, applying the inequality in Eq. 13 to the multivariate skewness and kurtosis formulas in Eqs. 5 and 6, the relationship of multivariate skewness and kurtosis in our method can be derived as

$$\beta_2^* \geq 1.641\beta_1^* + p(p + 0.774). \tag{17}$$

R package

An R package *mnonr* is developed based on our method to generate multivariate non-normal random numbers with user-specified multivariate skewness and kurtosis as well as the covariance matrix. If the values of the multivariate skewness and kurtosis are beyond the valid range of our method, the users will get a warning message and the allowed ranges. The package not only implements our multivariate method (function: *mnonr*), but also provides the Vale and Maurelli (1983) method (function: *unonr*). In addition, univariate and multivariate skewness and kurtosis significance tests are included (function: *mardia*).

Example

We now illustrate how to generate non-normal data with the *mnonr* package. Suppose the goal is to generate bivariate non-normal data with multivariate skewness $\beta_1^* = 3$ and kurtosis $\beta_2^* = 61$. Both variables have mean 0 and variance 1. The covariance between them is set to be 0.5. In total, we generate 10,000 bivariate random numbers with the desired features.

To generate the data, the R function *mnonr* was used in which we set $n = 10,000$, $p = 2$, $ms = 3$, $mk = 61$, and $Sigma = matrix(c(1, 0.5, 0.5, 1), 2, 2)$. The meaning of each argument is listed below:

- n: the size of random number to generate;
- p: the number of variables;
- ms: the value of multivariate skewness;
- mk: the value of multivariate kurtosis;
- $Sigma$: the covariance matrix of variables.

For illustration, we also calculated the covariance matrix of the generated data and conducted hypothesis testing of the univariate and multivariate skewness and kurtosis through the function *mardia*.

The R input and output are given below.

```
> mvn.data=mnonr(n=10000, p=2, ms=3, mk=61,
    Sigma=matrix(c(1,0.5,0.5,1),2,2))
> cov(mvn.data)
          [,1]      [,2]
[1,] 1.0795673 0.5435589
[2,] 0.5435589 1.0378786
> mardia(mvn.data)
Sample size:  10000
Number of variables:  2

Marginal skewness and kurtosis
     Skewness    SE_skew Kurtosis    SE_kurt
[1,] 0.9397589 0.02449122 24.99584
    0.04897755
[2,] 1.1900858 0.02449122 17.70874
    0.04897755

Mardia's multivariate skewness and kurtosis
                  b         z p-value
Skewness  3.154189 5256.9822       0
Kurtosis 64.110259  701.3782       0
```

The sample data yield a multivariate skewness 3.15 and multivariate kurtosis 64.11. The covariance matrix is close

to the specified one. It also shows clearly that the marginal univariate skewness and kurtosis for the two variables are different. According to marginal (2), the theoretical skewness and kurtosis for marginal variables are: $\gamma_1(x_1) = 1.22$, $\gamma_2(x_1) = 29.50$, $\gamma_1(x_2) = 0.95$, $\gamma_2(x_2) = 19.56$. The scatter-plot and marginal histograms are shown in Fig. 1. Even though both variables have leptokurtic distributions, x_1 has larger kurtosis than x_2, which shows on the figure that the distribution of x_1 has a fatter tail. This is because when we form $\mathbf{x}$, the transformation $\mathbf{x} = A\boldsymbol{\xi}$ would yield different distributions of each x_j, $j = 1, \cdots, p$, even though the ξ_j, $j = 1, \cdots, p$, are iid.

Simulation study

To evaluate the performance of our method, we conducted the following simulation study by varying the sample sizes, covariances, number of variables, and different combinations of multivariate skewness and kurtosis.

Study design

The sample sizes are set to be 100, 1000, and 10,000. We set the variances all to be 1 and varied the covariance between two variables from low to high with values 0, 0.1, 0.3, 0.5, 0.7, and 0.9. In each condition, the covariances of any two variables are set to be the same. The numbers of variables in the multivariate data are set to be 2, 4, and 6, which are also the number of ξ_j in $\boldsymbol{\xi}$.

The values of multivariate skewness and kurtosis are chosen based on Cain et al. (2017). They provided a descriptive table of Mardia's multivariate skewness and kurtosis values collected from 136 multivariate studies. We choose the minimum, first quartile, median, and third quartile values when the sample sizes are larger than 100. The values of multivariate skewness are $\beta_1^* = 0$, 1, 3, and 15. The multivariate kurtosis values are $\beta_2^* = 10$, 32, 61, and 91.

We deleted some conditions due to the restricted range of multivariate skewness and kurtosis by Eq. 16. In total, 480 conditions are evaluated, and 1000 replications of data are generated under each condition.

Evaluation

We evaluated the performance of our random number generation method by comparing the statistics of the generated data with the population ones used to generate the data. Specifically, the statistics included multivariate

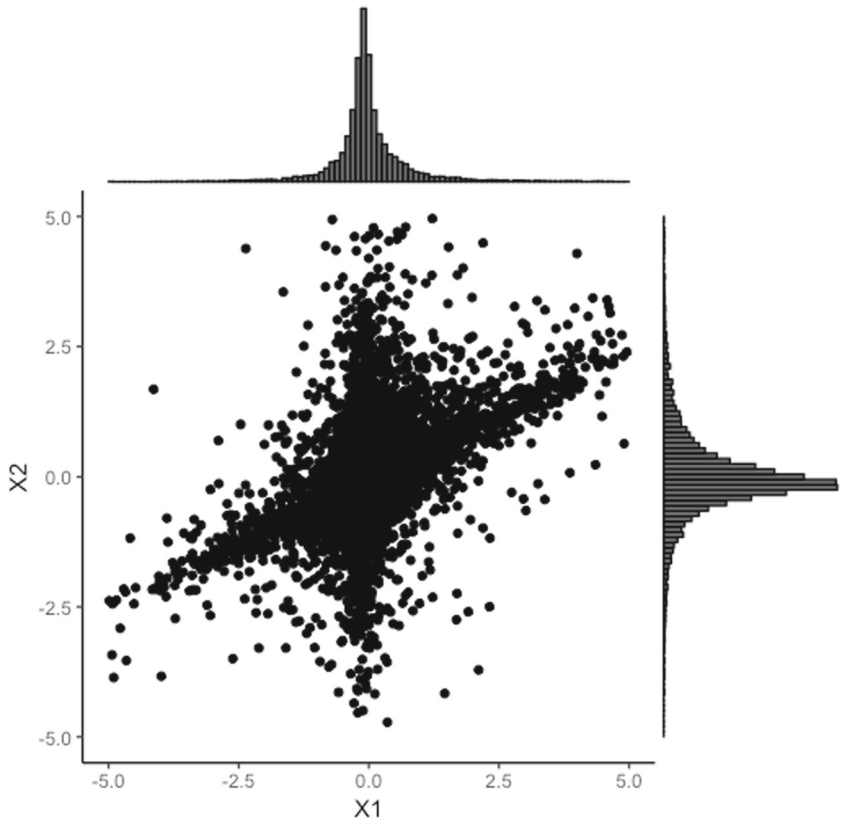

Fig. 1 Scatter-plot and marginal histograms of two-variable multivariate data

skewness β_1^*, multivariate kurtosis β_2^*, ξ_j's third moment $E(\xi)^3$, and ξ_j's fourth moment $E(\xi)^4$. We calculated the bias (B) and relative bias (RB) of the simulation results of the above statistics. The bias is the difference between the mean of the sample statistic value ($\hat{\theta}$) and its corresponding population parameter value (θ); and the relative bias is the proportion of the bias of the population value, which are

$$B = \frac{\sum^N \hat{\theta}}{N} - \theta, \tag{18}$$

$$RB = \frac{B}{\theta} \times 100\% \tag{19}$$

Results

For the sake of space, we only report several representative conditions in Table 1 and the full results are available in the Supplementary Materials. They represent the small ($\beta_1^* = 1$, $\beta_2^* = 32$), medium ($\beta_1^* = 3$, $\beta_2^* = 61$), and large ($\beta_1^* = 15$, $\beta_2^* = 91$) multivariate skewness and kurtosis combinations.

Table 1 Simulation results (partial)

	$B(\hat{\beta}_1^*)$	$B(\hat{\beta}_2^*)$	$RB(\hat{\beta}_1^*)\%$	$RB(\hat{\beta}_2^*)\%$	$B(E(\hat{\xi})^3)$	$B(E(\hat{\xi})^4)$
	p = 2					
(a)$\beta_1^* = 1, \beta_2^* = 32$						
N = 100	3.384	−11.147	338.4	−34.8	0.160	1.021
N = 1000	1.180	−2.853	118.0	−8.9	−0.017	−0.824
N = 10,000	0.180	−0.411	18.0	−1.3	0.007	−0.159
(b)$\beta_1^* = 3, \beta_2^* = 61$						
N = 100	6.781	−29.623	226.0	−48.6	−0.032	1.238
N = 1000	3.353	−9.331	111.8	−15.3	−0.060	−2.267
N = 10,000	0.654	−1.425	21.8	−2.3	0.015	−0.452
(c)$\beta_1^* = 15, \beta_2^* = 91$						
N = 100	1.486	−51.942	9.9	−57.1	−0.125	−9.706
N = 1000	3.775	−17.465	25.2	−19.2	−0.081	−3.658
N = 10,000	1.025	−2.816	6.8	−3.1	0.015	−0.660
	p = 4					
(a)$\beta_1^* = 1, \beta_2^* = 32$						
N = 100	1.066	−10.313	106.6	−32.2	−0.067	−1.377
N = 1000	0.835	−4.332	83.5	−13.5	−0.027	−0.675
N = 10,000	0.227	−0.478	22.7	−1.5	0.002	−0.010
(b)$\beta_1^* = 3, \beta_2^* = 61$						
N = 100	5.834	−17.100	194.5	−28.0	−0.030	−0.761
N = 1000	1.690	−3.570	56.3	−5.9	0.018	−0.143
N = 10,000	0.239	−0.239	8.0	−0.4	0.054	0.005
(c)$\beta_1^* = 15, \beta_2^* = 91$						
N = 100	2.080	−36.198	13.9	−39.8	−0.062	−0.135
N = 1000	1.895	−8.388	12.6	−9.2	0.032	0.332
N = 10,000	0.359	−0.805	2.4	−0.9	0.003	−0.084
	p = 6					
(b)$\beta_1^* = 3, \beta_2^* = 61$						
N = 100	2.554	−15.119	85.1	−24.8	−0.060	−1.265
N = 1000	1.704	−4.756	56.8	−7.8	0.023	1.077
N = 10,000	0.320	−0.606	10.7	−1.0	0.011	0.048
(c)$\beta_1^* = 15, \beta_2^* = 91$						
N = 100	2.673	−20.626	17.8	−22.7	−0.030	0.166
N = 1000	0.771	−3.770	5.1	−4.1	0.003	0.050
N = 10,000	0.175	−0.157	1.2	−0.2	0.000	0.000

When the sample size increases, the bias of both univariate and multivariate measures becomes smaller. The performance of ξ_j verified that the modified power method does not affect the accuracy of the power method for generating univariate non-normal data. For multivariate measures, kurtosis tended to be underestimated and skewness tended to be overestimated. Additionally, multivariate kurtosis had smaller relative bias than multivariate skewness.

When comparing simulation results with various covariance settings, we found that both multivariate skewness and kurtosis do not seem to be affected by covariances. The multivariate skewness is only related to the number of variables and the value of $E(\xi)^3$. Similarly, the multivariate kurtosis is influenced by the number of variables and the value of $E(\xi)^4$. Covariance does not play a crucial role in multivariate skewness and kurtosis via our generating method and therefore variance-covariance matrix only affects the marginal measures rather other multivariate measures.

Increasing the number of variables leads to less biased skewness and kurtosis estimates holding other conditions constant.

Conclusions and future directions

In this paper, we proposed a new method for generating multivariate non-normal data. The advantage of our method is that it allows researchers to directly specify both multivariate skewness and kurtosis to better control them. With the data generating model, we established one possible solution to relate multivariate measures to univariate measures: using univariate measures to generate $\boldsymbol{\xi}$ and apply the variance-covariance matrix to produce multivariate $\mathbf{x}$. Our method can help researchers better understand the influence of the multivariate non-normality.

The widely used VM method and our method are both based on Fleishman's polynomial-related approach. Both can generate correlated multivariate random data. However, the two methods also have important differences. The main difference lies in the perspectives of the multivariate non-normality. First of all, the multivariate non-normality can be simply because of the non-normality of the marginal distribution and/or the multivariate distribution. The VM method concentrates on the univariate non-normality without specifically controlling the multivariate non-normality. In contrast, our method focuses on the multivariate non-normality but not controlling the univariate marginal non-normality. For instance, in a bivariate distribution of $\mathbf{x} = (x_1, x_2)'$, through the VM method, researchers can specify the marginal univariate skewness and kurtosis of x_1 and x_2, but not themultivariate measures. With our method, one can directly determine multivariate skewness and kurtosis of $\mathbf{x}$, but with no control of the marginal distribution. The choice of the two methods should be based on the particular research interest.

Because of the use of the power method, our method also inherits the same problems associated with it, such as the Gaussian-like property and multiplicity solution issue. First, as it is discussed by Foldnes and Grønneberg (2015), to evaluate the robustness of Gaussian ML estimation using multivariate data with Gaussian-like property, even with the marginal univariate measures showing severe non-normality, the researchers might get biased results. Without further exploration, we could not identify the degree of the potential impact related to our method. Second, there are different sets of coefficients (a, b, c, d) in the modified power method. For example, in the limited simulation experiment in the supplementary materials (see Table 3), we found that within each parameter (i.e., β_1^*, β_2^*, n, p) setting, there were four sets of possible coefficients. Different starting values will yield different sets of coefficients, which could affect the multivariate skewness and kurtosis. We recommend that the researchers should try different starting values in data generating and our R package provides such an option in addition to the default value.

As shown in the simulation results, with a small sample size ($n = 100$), the relative bias of the multivariate measures could be very high. With the increasing number of variables (p) and sample sizes, this issue becomes less severe. When merging the data of each marginal univariate variable, a small deviation could lead to a large gap for multivariate data. This drawback is shared with other multivariate data generators relating to the reliability of multivariate measures. As a future direction, we plan to develop a sample size planning method of different multivariate skewness and kurtosis to optimize the generating process.

Since our method only used one approach to generate univariate variables, another related limitation as described in the Method section is that some combinations of skewness and kurtosis cannot be obtained. However, our procedure provides researchers a simple data model to transform multivariate measures to univariate ones. In the future, we will apply other univariate generators to our method in order to improve the empirical performance of the multivariate generator and eliminate the potential problems that are related to the current modified power method such as the solution multiplicity and Gaussian-like property.

Open practices statement The data in this study are based on simulation. None of the data or materials are related to any experiments.

References

Astivia, O. O., & Zumbo, B. D. (2015). A cautionary note on the use of the Vale and Maurelli method to generate multivariate, nonnormal data for simulation purposes. *Educational and Psychological Measurement*, *75*(4), 541–567. https://doi.org/10.1177/0013164414548894

Astivia, O. O., & Zumbo, B. D. (2018). On the solution multiplicity of the Fleishman method and its impact in simulation studies. *British Journal of Mathematical and Statistical Psychology*, *71*, 437–458. https://doi.org/10.1111/bmsp.12126

Bradley, J. V. (1980). Nonrobustness in z, t, and f tests at large sample sizes. *Bulletin of the Psychonomic Society*, *16*(5), 333–336. https://doi.org/10.3758/BF03329558

Cain, M. K., Zhang, Z., & Yuan, K. H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*, 1716–1735. https://doi.org/10.3758/s13428-016-0814-1

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods*, *2*(3), 292–307. https://doi.org/10.1037/1082-989X.2.3.292

Field, C., & Genton, M. G. (2012). The multivariate g-and-h distribution. *Technometrics*, *48*(1), 104–111. https://doi.org/10.1198/004017005000000562

Fleishman, A. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*(4), 521–532. https://doi.org/10.1007/BF02293811

Foldnes, N., & Grønneberg, S. (2015). How general is the Vale–Maurelli simulation approach? *Psychometrika*, *80*(4), 1066–1083. https://doi.org/10.1007/s11336-014-9414-0

Foldnes, N., & Olsson, U. H. (2016). A simple simulation technique for nonnormal data with prespecified skewness, kurtosis, and covariance matrix. *Multivariate Behavioral Research*, *51*, 207–219. https://doi.org/10.1080/00273171.2015.1133274

Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics & Data Analysis*, *40*, 685–711. https://doi.org/10.1016/S0167-9473(02)00072-5

Hollander, M., & Wolfe, D. A. (2015). *Nonparametric statistical methods*, 3rd edn. Wiley.

Lee, Y., & Kaplan, D. (2018). Generating multivariate ordinal data via entropy principles. *Psychometrika*, *83*(1), 156–181. https://doi.org/10.1007/s11336-018-9603-3

Luo, H. (2011). Generation of non-normal data-a study of Fleishman's power method. Dept. of Statistics Uppsala Univ.

Mair, P., Satorra, A., & Bentler, P. (2012). Generating nonnormal multivariate data using copulas: Applications to SEM. *Multivariate Behavioral Research*, *47*(4), 547–565. https://doi.org/10.1080/00273171.2012.692629

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519–530. https://doi.org/10.1093/biomet/57.3.519

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*(3), 465–471. https://doi.org/10.1007/BF02293687

Yuan, K. H., & Bentler, P. (1997). Generating multivariate distributions with specified marginal skewness and kurtosis. In Bandilla, W., & Faulbaum, F. (Eds.) *SoftStat'97-advances in statistical software*, (Vol. 6, pp. 385–391). Stuttgart: Lucius & Lucius.

Yuan, K. H., Bentler, P. M., & Zhang, W. (2005). The effect of skewness and kurtosis on mean and covariance structure analysis: The univariate case and its multivariate implication. *Sociological Methods & Research*, *34*(2), 240–258. https://doi.org/10.1177/0049124105280200

Yuan, K. H., Zhang, Z., & Zhao, Y. (2017). Reliable and more powerful methods for power analysis in structure equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *2*(3), 315–330. https://doi.org/10.1080/10705511.2016.1276836