

RESEARCH ARTICLE | OCTOBER 17 2018

Goodness-of-fit test of multivariate t -distribution with Monte Carlo simulation using R **FREE**

Ibnu Santoso ✉; Tiodora Hadumaon Siagian; Erni Tri Astuti; Azka Ubaidillah



AIP Conf. Proc. 2021, 060018 (2018)

<https://doi.org/10.1063/1.5062782>



Boost Your Optics and Photonics Measurements

Lock-in Amplifier

Zurich Instruments

Find out more

Boxcar Averager

Goodness-of-Fit Test of Multivariate t -Distribution with Monte Carlo Simulation Using R

Ibnu Santoso^{1,a)}, Tiodora Hadumaon Siagian^{1,b)}, Erni Tri Astuti^{1,c)} and Azka Ubaidillah^{1,d)}

¹*Polytechnic of Statistics STIS, Jakarta, Indonesia 13330*

^{a)}Corresponding author: ibnu@stis.ac.id;

^{b)}theo@stis.ac.id; ^{c)}erni@stis.ac.id; ^{d)}azka@stis.ac.id;

Abstract. Many methods exist in the literature for testing multivariate normality. They can be grouped into graphical and numerical methods. While it is known that the multivariate t -distribution is more realistic for modelling empirical data than the multivariate normal distribution due to its heavier tail, unfortunately there are only a few of goodness-of-fit tests for the multivariate t -distribution. One example is based on the Rényi entropy of order λ , which can be used in comparing the shape, densities, and measuring the heaviness of tails of the distributions. However, the application of this method is a bit complicated and time-consuming. Another method is based on Skewness and Kurtosis; it uses Monte Carlo Simulation for multivariate t -distribution testing. Hypothesis testing with Monte Carlo simulation is considered to have advantages over other classical hypothesis testing statistics because it is simpler and faster to compute. For multivariate t -distribution testing, the latter method is easier to understand, simpler, and easier to apply than earlier methods. Implementing this method in MATLAB would still require some MATLAB toolbox functions. On the other hand, the statistical environment R is an open source software package and a powerful tool for statistical data analysis and graphical representation. R provides the opportunity for many individuals to improve its code and add functions. This study aims to illustrate the procedure of a goodness-of-fit test of the multivariate t -distribution with Monte Carlo simulation using R as an open source alternative to MATLAB. Based on the p -values of the Skewness and Kurtosis tests (both univariate and multivariate) on the generated multivariate t -distribution data, it is shown that the simulation data can more accurately be assumed to follow a multivariate t -distribution. It is expected that this study can be beneficial for other researchers who want to do goodness-of-fit tests of a multivariate t -distribution.

Keywords: Kurtosis, Monte Carlo Simulation, Multivariate t -Distribution, R, Skewness, Testing of fit.

INTRODUCTION

Most multivariate data analysis is based on the assumption that the data follow a multivariate normal distribution as it is mathematically tractable. Consequently, many methods for testing multivariate normality exist in the literature, such as graphical methods (i.e., Q-Q plot) and numerical methods. However, many empirical datasets provide strong evidence that they do not follow a normal distribution.¹ As both the normal distribution and the multivariate t -distribution are from the family of elliptically symmetric distributions, the multivariate t -distribution has become a famous alternative to the multivariate normal distribution since it is more realistic for modelling empirical data due to its heavier tail, and its application is believed to produce more robust statistical inference in multivariate analysis.

Although the multivariate t -distribution is considered more realistic to model real world multivariate data than the multivariate normal distribution due to its heavier tail, there are relatively few tests of the fit to the multivariate t -distribution. One example is a test to verify whether a dataset was sampled from specific elliptical distributions based on Song's measure of kurtosis, proposed by Batsidis and Zografos.² Song proposed a general measure of the shape of a distribution which is based on the Rényi entropy of order λ .³ According to Song,³ this measure has a similar role as a kurtosis measure in comparing the shape, densities and measuring the heaviness of tails of the distributions, but it measures more than the traditional kurtosis. Then Zografos derived Song's measure of kurtosis for the elliptic family of multivariate distribution.⁴ However, the application of this method is a bit complicated and time-consuming.

Another example is a procedure for testing whether or not the data follows the multivariate t -distribution proposed by Kan and Zhou,¹ who used Mardia's measure of skewness and kurtosis. Kan and Zhou use Monte Carlo Simulation for their multivariate t -distribution testing. Hypothesis testing with Monte Carlo simulation is considered to have advantages over other classical hypothesis testing statistics because it is simpler and faster to compute. This method is considered to be easier to understand, simpler, and easier to apply than the method of Batsidis and Zografos. Unfortunately, implementing the method of Kan and Zhou in MATLAB still requires some MATLAB toolbox functions. On the other hand, the statistical environment R is an open source software package and a powerful tool for statistical data analysis and graphical representation. R provides the opportunity for many individuals to improve the code and add functions. Therefore, this study aims to illustrate the procedure of a goodness-of-fit test for the multivariate t -distribution with Monte Carlo simulation using R as an open source alternative to MATLAB.

The Multivariate t -Distribution

Both the multivariate normal distributions and the multivariate t -distributions are members of the general family of elliptically symmetric distributions. The probability density function of the d -dimensional multivariate Student's t distribution is given by Kotz and Nadarajah,⁵

$$\frac{\Gamma[(v+p)/2]}{\Gamma(v/2)v^{p/2}\pi^{p/2}|\Sigma|^{1/2}} \left[1 + \frac{1}{v}(x-\mu)^T \Sigma^{-1}(x-\mu) \right]^{-(v+p)/2} \quad \text{i)}$$

where x is a random vector with dimension p , Σ is a variance-covariance matrix, μ is the vector of means, v is a positive scalar, and $p \leq 2$.

Goodness-of-Fit Test of Multivariate t -Distribution

Many methods exist for testing multivariate normality. For example, in R there is the 'mvnTest' package (Goodness of Fit Tests for Multivariate Normality), which implements three Wald-type chi-squared tests; the non-parametric Anderson-Darling and Cramer-von Mises tests; the Doornik-Hansen test, the Royston test, and the Henze-Zirkler test.⁶

On the other hand, there are only a few of goodness-of-fit tests for the multivariate t -distribution. A proposed method uses a measure based on the Rényi entropy of order λ , which is something that can be used to compare the shape and densities of distributions as well as measure the heaviness of their tails.⁴ However, the application of this method is a bit difficult and time consuming.

Another method, that easier to understand and apply, uses Monte Carlo Simulation for multivariate- t distribution testing based on Skewness and Kurtosis for financial data analysis.¹ Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The normal distribution has a skewness value of 0. Kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable. The normal distribution has a kurtosis value of 3. The Multivariate Skewness and Multivariate Kurtosis formulas are as follows:⁷

- Multivariate Skewness

$$b_{1,p} = \frac{1}{n^2} \sum_{ij=1}^n [(X_i - \bar{X})' S^{-1} [(X_j - \bar{X}) J]^3 \quad \text{ii)}$$

- Multivariate Kurtosis

$$b_{2,p} = \frac{1}{n} \sum_{i=1}^n [(X_i - \bar{X})' S^{-1} [(X_j - \bar{X})] J]^2 \quad \text{iii)}$$

Monte Carlo Simulation

Monte Carlo Simulation is a class of computational algorithms that rely on repeated random sampling to obtain numerical results. The essential idea is using randomness to solve problems that might be deterministic in principle. This is often used in physical and mathematical problems and is most useful when it is difficult or impossible to use other approaches. Hypothesis testing with Monte Carlo simulation is considered to have advantages over other classical hypothesis testing statistics because it is simpler and faster to compute.

METHODOLOGY

The goodness-of-fit test of the Multivariate t -Distribution with Monte Carlo Simulation has several steps:

1. Input research data and get the dimension.
2. Randomly generate two simulation datasets as many as n times:
 - First dataset follows multivariate-normal distribution;
 - Second dataset follows multivariate t -distribution with three different numbers of degrees of freedom.
 - These two simulation datasets have the same dimension as the research dataset.
3. Compute univariate skewness and kurtosis of research data.
4. Compute multivariate skewness and kurtosis of research data.
5. Compute **p -value** for both univariate and multivariate skewness by computing percentage of skewness value of simulation data **bigger** than skewness value of research data.
6. Compute **p -value** for both univariate and multivariate kurtosis by computing percentage of kurtosis value of simulation data **bigger** than kurtosis value of research data.
7. Hypotheses:
 - Research data follows Multivariate Normal Distribution
 - H_0 : Data follows Multivariate Normal Distribution
 - H_A : Data doesn't follow Multivariate Normal Distribution
 - Research data follows Multivariate t -Distribution
 - H_0 : Data follows Multivariate t -Distribution
 - H_A : Data doesn't follow Multivariate t -Distribution
8. Reject H_0 if p -value < 0.05 (with $\alpha = 5\%$)

COMPARISON OF BENEFITS OF USING MATLAB AND R

MATLAB and R have many features in common. Some of the differences are trivial while others can be troublesome.⁸ One of the biggest advantages of R is that it is free and open source. MATLAB's Statistics and Machine Learning toolbox is not free either.

Conversion from MATLAB to R


The procedures of running goodness-of-fit tests in MATLAB were obtained through private communication with its author.¹ However, some of the equivalents in R to the MATLAB toolbox functions must be searched for. For example, `chi2rnd()` in MATLAB's toolbox is equivalent with `rchisq()` in R. Some functions with the same names behave a bit differently in MATLAB than in R, for example: `mean()` in MATLAB has very broad usage for vectors and matrices, while in R we need to specify what kind of mean we want (`colMeans()` or `rowMeans()` or `mean()` itself). Translation was done by searching R-equivalent syntax by looking at various resources, and testing the output of each function thoroughly.

RESULTS AND EVALUATION

To test the procedure in R, we conducted an experiment with two generated multivariate datasets: multivariate normal data and multivariate t -distributed data, and compared the output.

Multivariate Normal Data Test

Multivariate normal data was generated using the command `mvnrnd (MU, SIGMA, cases)` in MATLAB, with MU the mean vector [0,0,0], SIGMA the covariance [1 0.3 0.2;0.3 1 0.3;0.2 0.3 1], and 100 cases. The generated multivariate normal data was then tested with the goodness-of-fit test procedures in MATLAB and R. Fig. 1 shows the results.

Number of Simulations = 10000											
Variabel	Skewness	p-value				Kurtosis	p-value				
		Normal	df=10	df=8	df=6		Normal	df=10	df=8	df=6	
1	0.258	26.07	44.58	49.54	58.21	3.234	20.77	60.89	69.14	80.98	
2	0.168	46.37	61.55	64.87	71.32	2.772	58.65	87.76	90.75	95.69	
3	-0.111	62.52	73.89	76.13	80.68	2.500	85.56	97.05	97.73	99.12	
Multi	0.512	52.07	87.34	91.39	96.12	14.417	57.75	98.43	99.51	99.86	


Number of Simulations = 10000											
Variabel	Skewness	p-value				Kurtosis	p-value				
		Normal	df=10	df=8	df=6		Normal	df=10	df=8	df=6	
1	0.258	26.47	44.30	49.87	57.47	3.234	20.94	60.17	70.02	80.81	
2	0.168	46.33	61.38	65.74	70.57	2.772	60.14	87.81	91.64	95.74	
3	-0.111	62.63	73.76	76.89	80.10	2.500	85.98	97.28	98.18	99.16	
Multi	0.512	52.10	87.50	92.00	96.24	14.417	57.75	98.51	99.46	99.92	


FIGURE 1. Output Results for Multivariate Normal Data Test in MATLAB and R.

Interpretation:

- Univariate and multivariate skewness shows the data is distributed normally (with skewness close to 0). Furthermore, p -value for all univariate and multivariate skewness shows that the data is distributed normally (with $\alpha = 5\%$): all p -value > 0.05 .
- Univariate kurtosis values are all close to 3. Furthermore, p -value for univariate and multivariate kurtosis shows the data is distributed normally (with $\alpha = 5\%$): all p -value $> 0,05$
- H_0 that data follows multivariate normal distribution is not rejected.
- H_0 that data follows multivariate t -distribution is not rejected either, so one can choose between two options.

Multivariate t Data Test

Multivariate t -distributed data was generated using the command `mvtrnd (C,df,cases)` in MATLAB, with C the correlation matrix valued [1 0.3 0.2;0.3 1 0.3;0.2 0.3 1], $df=4$, cases=100. The generated multivariate t -distributed data was then tested with the goodness-of-fit test procedures in both MATLAB and R. Fig. 2 shows the results.

Number of Simulations = 10000											
Variabel	Skewness	p-value				Kurtosis	p-value				
		Normal	df=10	df=8	df=6		Normal	df=10	df=8	df=6	
1	3.232	0.00	0.01	0.03	0.27	23.224	0.00	0.01	0.06	0.25	
2	1.668	0.00	0.32	0.75	2.28	14.879	0.00	0.06	0.30	0.99	
3	0.739	0.32	5.92	8.89	15.92	6.002	0.02	3.82	7.10	14.81	
Multi	19.238	0.00	0.04	0.10	0.72	54.948	0.00	0.01	0.03	0.37	


Number of Simulations = 10000											
Variabel	Skewness	p-value				Kurtosis	p-value				
		Normal	df=10	df=8	df=6		Normal	df=10	df=8	df=6	
1	3.232	0.00	0.00	0.03	0.28	23.224	0.00	0.01	0.02	0.27	
2	1.668	0.00	0.35	0.81	2.43	14.879	0.00	0.11	0.28	1.17	
3	0.739	0.34	5.90	9.12	15.69	6.002	0.02	3.98	7.36	14.33	
Multi	19.238	0.00	0.04	0.12	0.73	54.948	0.00	0.02	0.05	0.31	

FIGURE 2. Output Results for Multivariate-t Data Test in MATLAB and R.

Interpretation:

- Univariate and multivariate skewness shows that data are not normally distributed. Furthermore, *p-value* for univariate and multivariate skewness shows the data are not normally distributed (at $\alpha = 5\%$) except for variable 3.
- Univariate and multivariate kurtosis are all more than 3. Furthermore, *p-value* for univariate and multivariate kurtosis shows the data are not normally distributed (at $\alpha = 5\%$).
- H_0 where data follows multivariate normal distribution can be rejected because most *p-value* < 0.05 .
- If it is assumed that the data follows a multivariate *t*-distribution with 6, 8, or 10 degrees of freedom, *p-value* keeps increasing. The assumption that the data follows a multivariate *t*-distribution is accepted if $df=6$ (where all *p-value* > 0.05)

CONCLUSION

The goodness-of-fit test procedures were successfully run in R, producing the same output for univariate and multivariate skewness and kurtosis, and similar *p-values* to those of the MATLAB output. Moreover, based on the *p-values* of the Skewness and Kurtosis tests (both univariate and multivariate) on the generated multivariate *t*-distribution data, it can be concluded that it is more accurate to assume that the simulation data follows a multivariate *t*-distribution. For further research, this procedure can be developed into an R package to benefit other researchers who want to do goodness-of-fit tests for the multivariate *t*-distribution in R.

ACKNOWLEDGEMENTS

We would like to thank Prof. Raymond Kan sincerely for providing us procedures of goodness-of-fit test in MATLAB, so we can test and convert it to R procedures.

REFERENCES

1. R. Kan and G. Zhou, Modeling Non-normality Using Multivariate t: Implications for Asset Pricing, 2006. (Accessed from <http://apps.olin.wustl.edu/faculty/zhou/KZ-t-06.pdf>).
2. A. Batsidis and K. Zografos, *J. Multivar. Anal.* **113**, 91-105 (2013).
3. K. S. Song, *J. Stat. Plan. Infer.* (2001).
4. K. Zografos, *J. Multivar. Anal.* **99**, 858-879 (2008).

5. S. Kotz and S. Nadarajah, *Multivariate t Distributions and Their Applications*. (Cambridge University Press, Cambridge, United Kingdom, 2004.)
6. P. Natalya, V. Vassily, M. Rashid and V. Yevgeniy, *Goodness of Fit Tests for Multivariate Normality*, 2016. (Accessed from <https://cran.r-project.org/web/packages/mvnTest/mvnTest.pdf>).
7. K. V. Mardia, *Biometrika*, **57**, 519-530 (1970).
8. J. O. Ramsay, G. Hooker and S. Graves, *Functional Data Analysis with R and MATLAB*. (New York, Springer, 2009).