JIRSS (2020)
Vol. 19, No. 02, pp 119-131
DOI:10.52547/jirss.19.2.119

On the Canonical-Based Goodness-of-fit Tests for Multivariate Skew-Normality

Saeed Darijani [1], **Hojatollah Zakerzadeh** [1], **Hamzeh Torabi** [1].

[1] Department of Statistics, Faculty of Mathematics, Yazd University, Iran.

Received: 05/03/2020, Revision received: 18/06/2020, Published online: 18/07/2020

Abstract. It is well-known that the skew-normal distribution can provide an alternative model to the normal distribution for analyzing asymmetric data. The aim of this paper is to propose two goodness-of-fit tests for assessing whether a sample comes from a multivariate skew-normal (MSN) distribution. We address the problem of multivariate skew-normality goodness-of-fit based on the empirical Laplace transform and empirical characteristic function, respectively, using the canonical form of the MSN distribution. Applications with Monte Carlo simulations and real-life data examples are reported to illustrate the usefulness of the new tests.

Keywords. Multivariate Skew-Normal Distribution, Canonical Form, Laplace Transform, Characteristic Function.

MSC: 62H05, 62E15, 62G30.

1 Introduction

The normality-based models are initially used for analyzing data in various areas of sciences due to the mathematical as well as statistical properties and computational convenience. However, the assumption of multivariate normality is often violated as the data have a non-normal features (e.g. strong skewness and heavy tails). In the matter of skewed data, Azzalini (1985) and Azzalini and Capitanio (1999) proposed the univariate and multivariate skew-normal (MSN) distributions, respectively. The probability density function (pdf) of a d-dimensional random vector $\boldsymbol{Z}$ with the standard

Corresponding Author: Saeed Darijani (saeed_darijani@yahoo.com)
Hojatollah Zakerzadeh (hzaker@yazd.ac.ir)
Hamzeh Torabi (htorabi@yazd.ac.ir).

MSN distribution is

$$f(\boldsymbol{z};\boldsymbol{\lambda},\bar{\boldsymbol{\Omega}}) = 2\phi_d\left(\boldsymbol{z};\bar{\boldsymbol{\Omega}}\right)\Phi\left(\boldsymbol{\lambda}^\top \boldsymbol{z}\right), \quad \boldsymbol{z},\boldsymbol{\lambda} \in \mathbb{R}^d, \tag{1.1}$$

where $\phi_d\left(\cdot;\bar{\boldsymbol{\Omega}}\right)$ is the pdf of d-dimensional normal distribution with mean vector $\mathbf{0}$ and correlation matrix $\bar{\boldsymbol{\Omega}}$, denoted by $N_d(\mathbf{0},\bar{\boldsymbol{\Omega}})$, and $\Phi(\cdot)$ represents the cumulative distribution function (cdf) of $N_1(0,1)$. The notation $\boldsymbol{Z} \sim MSN(\bar{\boldsymbol{\Omega}},\boldsymbol{\lambda})$is used in the following to indicate that the random vector $\boldsymbol{Z}$ has pdf (1.1). It can be easily seen that (1.1) tends to the pdf of multivariate normal distribution as $\boldsymbol{\lambda}$ approaches zero. To assume a MSN distribution for data possessing some level of skewness, one may need to assess the skew-normality assumption for obtaining a valid and accurate results in the analysis. Although various statistical tests were introduced to check the skewness of a data set (see, e.g., Jarque and Bera (1987)), the literature on testing the MSN distribution against other distributions is not very extensive. Suppose $\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_n$ are independent and identically random vectors of size n coming from a d-dimensional generic cdf $F_d(\boldsymbol{z};\boldsymbol{\theta})$ with parameter $\boldsymbol{\theta}$. Then, we are going to focus on developing two tests for the multivariate skew-normality with the hypothesis

$$\begin{cases} H_0: & \boldsymbol{Z} \text{ follows a MSN distribution with some parameters;} \\ H_1: & \boldsymbol{Z} \text{ follows a distribution other than the MSN distribution.} \end{cases} \tag{1.2}$$

To cope with the test of (1.2), one may formally be interested in three issues Balakrishnan et al. (2014). These issues are that the test procedure should (1) have appropriate statistical performances; (2) have feasible and straightforward computational algorithm and (3) be mathematically elegant for generalizing to higher dimension. Recently, Mateu-Figueras et al. (2007) and Meintanis (2007) proposed two tests for the univariate skew-normality assumption whereas Meintanis and Hlavka (2010) explored a statistical test for testing multivariate skew-normality based on the moment generating function. The proposed test of Meintanis and Hlavka (2010) is just computationally available for the bivariate case and the bootstrap resampling should be used for obtaining its distribution. Moreover, Balakrishnan et al. (2014) exploited the canonical form of the MSN distribution Azzalini and Capitanio (1999) to construct tests for multivariate skew-normality. The proposed tests of Balakrishnan et al. (2014) have some dimensional restrictions. Furthermore, their significance level depends on the sample size. To possess three aforementioned issues, the main objective of this contribution is to formulate some tests for the multivariate skew-normality. Our proposed tests are based on the canonical form of MSN distribution, on some statistical relationships of the MSN distribution with the gamma and Cauchy models, and on the empirical Laplace transform and empirical characteristic function.

The outline for the rest of this paper is structured as follows. Section 2 is divided into three parts in which a review of the MSN distribution and its canonical form, the construction of goodness-of-fit test based on empirical Laplace transform (ELT) and formulation of goodness-of-fit test based on the empirical characteristic function are presented. In Section 3, some Monte Carlo (MC) simulation studies are carried out to examine the performance of the proposed tests. For this purpose we consider some specific alternative distributions for the evaluation of the tests' power. We present four real examples for illustrative purposes in Section 4. Finally some concluding remarks are made in Section 5.

2 The Test Statistics

2.1 Canonical form of the MSN Distribution

Let $\boldsymbol{Z} \sim MSN(\boldsymbol{\lambda}, \bar{\boldsymbol{\Omega}})$. The transformed vector $\boldsymbol{Z}^* = \boldsymbol{A}^*\boldsymbol{Z}$ is said to be the canonical form of the MSN distribution if it admits the following pdf

$$f_{\boldsymbol{Z}^*}(\boldsymbol{z};\boldsymbol{\lambda}) = 2\phi_d\left(\boldsymbol{z};\boldsymbol{I}_d\right)\Phi\left(\lambda_* z_m\right), \quad \boldsymbol{z} \in \mathbb{R}^d, \;\; m \in \{1,2,...,d\},$$

where $\boldsymbol{A}^*$ is a $d \times d$ non-singular matrix and $\lambda_* = (\boldsymbol{\lambda}^\top \bar{\boldsymbol{\Omega}} \boldsymbol{\lambda})^{1/2}$ is the only non-zero component of the shape parameter of $\boldsymbol{Z}^*$. It is interesting to note that the pdf of $\boldsymbol{Z}^*$ can be expressed by the product of $d-1$ pdfs of $N_1(0,1)$ and a pdf of standard univariate skew-normal distribution with shape parameter λ_*. Moreover, the marginal univariate components of $\boldsymbol{Z}^*$ are independent. Proposition 4 of Azzalini and Capitanio (1999) ensures that the non-unique $\boldsymbol{A}^*$ exists. However, it is not applicable to construct a test statistics. To use the canonical form, one can consider Capitanio (2012) who defined the canonical representation of MSN distribution in a general form. If $\boldsymbol{Z} \sim MSN(\bar{\boldsymbol{\Omega}}, \boldsymbol{\lambda})$, then the random vector $\boldsymbol{Y} = \boldsymbol{\xi} + \omega\boldsymbol{Z}$ follows MSN distribution with a location parameter $\boldsymbol{\xi}$ and a diagonal matrix ω of scale parameters, such that $\boldsymbol{\Omega} = \omega\bar{\boldsymbol{\Omega}}\omega$ is a covariance matrix, denoted by $\boldsymbol{Y} \sim MSN(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$. The following propositions give constructive approaches to obtain a canonical form for the MSN distributed variable in order to build test statistics.

Proposition 2.1. *Let $\boldsymbol{Y} \sim MSN(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$ and consider the non-singular transform $\boldsymbol{Y}^* = \left(\boldsymbol{C}^{-1}\boldsymbol{P}\right)^\top \omega^{-1}(\boldsymbol{Y} - \boldsymbol{\xi})$, where $\boldsymbol{C}\boldsymbol{C}^\top = \bar{\boldsymbol{\Omega}}$ and $\boldsymbol{P}$ be an orthogonal matrix, with its first column proportional to $\boldsymbol{C}\boldsymbol{\lambda}$. Then, $\boldsymbol{Y}^* \sim MSN(\boldsymbol{0}, \boldsymbol{I}_d, \boldsymbol{\lambda}_{\boldsymbol{Y}^*})$, where $\boldsymbol{\lambda}_{\boldsymbol{Y}^*} = [\lambda_*, 0, ..., 0]^\top$ and $\lambda_* = (\boldsymbol{\lambda}^\top \bar{\boldsymbol{\Omega}} \boldsymbol{\lambda})^{1/2}$.*

Proposition 2.2. *Capitanio (2012) Let $\boldsymbol{Y} \sim MSN(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$ and $\boldsymbol{M} = \boldsymbol{\Omega}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Omega}^{-1/2}$, where $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{Y}$. Let $\boldsymbol{Q}$ be an orthogonal $d \times d$ matrix and $\boldsymbol{\Lambda}$ a $d \times d$ diagonal matrix such that $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ is the spectral decomposition of $\boldsymbol{M}$. Then, for $\boldsymbol{H} = \boldsymbol{\Omega}^{-1/2}\boldsymbol{Q}$, the linear transformation leading $\boldsymbol{Y}$ to the canonical form is $\boldsymbol{Y}^* = \boldsymbol{H}^\top(\boldsymbol{Y} - \boldsymbol{\xi})$.*

2.2 Test Statistics Based on Empirical Laplace Transform

For a sequence of random vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ with a d-dimensional $MSN(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$, define $\boldsymbol{Y}_j^* = [Y_{1j}^*, \ldots, Y_{dj}^*]^\top$, $j = 1, \ldots, n$, where Y_{ij}^* is the linear transformation leading $\boldsymbol{Y}$ to the canonical form, presented in Proposition 2.2. If H_0 in (1.2) is true, then Y_{1j}^* is a univariate SN distributed random variable, and all other variables $Y_{2j}^*, \ldots, Y_{dj}^*$ are independently distributed by the Gaussian model. Then, it can be seen that $\sum_{i=1}^{d}(Y_{ij}^*)^2 \sim \chi_d^2$, where

χ^2_d represents the chi-square distribution with d degree of freedom. Therefore, the test of the MSN distribution turns into the test of gamma model with the shape and scale parameters $\vartheta = d/2$ and $\beta = 0.5$, respectively. In this regard, one can use the statistical test of gamma distribution. An efficient goodness-of-fit test for gamma distribution, exploited in this paper to test multivariate skew-normality, is based on the ELT developed by Henze et al. (2012). For a gamma distributed random variable X with parameters ϑ and β, the Laplace transform, theoretically expressed by $l(t) = E[\exp(-tX)]$, is defined as the only solution of the differential equation $(1 + \beta t)y'(t) + \vartheta\beta y(t) = 0$, such that $y(0) = 1$. Under H_0 in (1.2), β and ϑ are assumed to be the parameters of the true underlying gamma distribution, but they should also make sense under a fixed alternative to H_0 (under appropriate conditions). It is pointed out by Henze et al. (2012) that for a given set of random points of size n, the estimator $\hat{\beta}_n = \hat{\beta}_n(X_1, \ldots, X_n)$ and $\hat{\vartheta}_n = \hat{\vartheta}_n(X_1, \ldots, X_n)$ of β and ϑ, respectively, are scale equivariant and converge almost surely to some $\beta > 0$ and $\vartheta > 0$, respectively. Therefore, to achieve scale invariance property, the ELT can be considered as $L_n(t) = \frac{1}{n}\sum_{j=1}^{n} \exp(-tW_j)$, where the scaled data is $W_j = X_j/\hat{\beta}_n$, for $j = 1, 2, \ldots, n$. Note that based on scale invariant property $\hat{\beta}_n(W_1, \ldots, W_n) = 1$. Hence, as $W_1, \ldots, W_n$ are approximately distributed as $\Gamma(\vartheta, 1)$ under H_0, for large n and some $\vartheta > 0$, a test for H_0 can be constructed based on a measure of deviation from zero of the random function

$$D_n(t) = (1 + t)L'_n(t) + \hat{\vartheta}_n L_n(t).$$

Subsequently, for a continuous weight function $w(t)$ that fulfills $\int_0^\infty t^4 w(t)\, dt < \infty$, the test statistic can be defined as $T_n = \int_0^\infty nD_n^2(t)w(t)\, dt$. The null hypothesis H_0 will be rejected if T_n gets large values. As suggested by Henze et al. (2012), we consider two class of weight functions $w_1(t) = \exp\{-at\}$ and $w_2(t) = \exp\{-at^2\}$ for a parameter $a > 0$, that lead, respectively, to the following closed form test statistics:

$$\begin{aligned}
T^{(1)}_{n,a} = \int_0^\infty nD_n^2(t)\exp\{-at\}\, dt = \frac{1}{n}\sum_{j,k=1}^{n}\Bigg[&\frac{W_jW_k - \hat{\vartheta}_n(W_j + W_k) + \hat{\vartheta}_n^2}{W_j + W_k + a} \\
&+ \frac{2W_jW_k - \hat{\vartheta}_n(W_j + W_k)}{(W_j + W_k + a)^2} + \frac{2W_jW_k}{(W_j + W_k + a)^2}\Bigg], \\
T^{(2)}_{n,a} = \int_0^\infty nD_n^2(t)\exp\{-at^2\}\, dt& \\
=\frac{1}{2n}\sqrt{\frac{\pi}{a}}\sum_{j,k=1}^{n}&\left[W_jW_k - \hat{\vartheta}_n(W_j + W_k) + \hat{\vartheta}_n^2\right]\varphi_{jk}(a) \\
&+ \frac{1}{4na}\sum_{j,k=1}^{n}\left[W_jW_k - \hat{\vartheta}_n(W_j + W_k)\right]\left[2 - \frac{\pi}{a}(W_j + W_k)\varphi_{jk}(a)\right] \\
&+ \frac{1}{8na^2}\sum_{j,k=1}^{n} W_jW_k\left[\left\{\frac{\pi}{a}(W_j + W_k)^2 + 2\sqrt{\pi a}\right\}\varphi_{jk}(a) - 2(W_j + W_k)\right],
\end{aligned}$$

where $\hat{\vartheta}_n$ is the estimate of ϑ based on $(W_1, \ldots, W_n)$, and for the error function $errf(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)dt$,

$$\varphi_{jk}(a) = \left[1 - errf\left(\frac{W_j + W_k}{2\sqrt{a}}\right)\right] \exp\left\{\frac{(W_j + W_k)^2}{4a}\right\}.$$

2.3 Test Statistics Based on Empirical Characteristic Function

Using the canonical form presented in Section 2.2, let us define the random variables $W_1, \ldots, W_n$ as

$$W_j = \sum_{i=2}^{d} Y^*_{ij} \Big/ \sqrt{d-1}|Y^*_{1j}|, \qquad j = 1, .., n. \tag{2.1}$$

It is straightforward to see that each W_j is distributed as the standard Cauchy distribution with pdf $f_{W_j}(w_j) = \left(\pi(1 + w_j^2)\right)^{-1}$. This results in testing for the Cauchy distribution instead of the MSN model. So, using the canonical form of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ and the transformation on data by (2.1), the hypothesis (1.2) transforms to the goodness-of-fit test of Cauchy distribution as

$$\begin{cases} H_0^* : & W \text{ follows a Cauchy distribution;} \\ H_1^* : & W \text{ follows a distribution other than the Cauchy distribution.} \end{cases} \tag{2.2}$$

Now, we can follow the strategy of empirical characteristic function for Cauchy goodness-of-fit test presented by Gürtler and Henze (2000). For the copies of n independent random variables $W_1, \ldots, W_n$, the null hypothesis of interest can be expressed as H_0 : $F \in \mathcal{F} = \{F(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, where $\boldsymbol{\theta} = \left\{(\mu, \sigma);\ \mu \in \mathbb{R},\ \sigma > 0\right\}$ is a two-dimensional parameter space, and $F(\cdot; \boldsymbol{\theta})$ is the cdf of Cauchy distribution, $C(\mu, \sigma)$, given by

$$F(w; \boldsymbol{\theta}) = 0.5 + \pi^{-1} \arctan\left(\frac{w - \mu}{\sigma}\right).$$

Since the hypothesis (2.2) checks whether data from an unknown distribution belongs to the location-scale family $\mathcal{F}$ generated by the standard Cauchy cdf $F_0(x) = F(w; (0, 1))$ and also since $\mathcal{F}$ is closed under affine transformations and the alternatives to H_0 are rarely known in practice, one may be interest to construct affine invariant and consistent test. In this matter, let $\varphi_n(t)$ be the empirical characteristic function of the standardized data $V_j = (W_j - \hat{\mu}_n)/\hat{\sigma}_n,\ \ 1 \leq j \leq n$, defined by

$$\varphi_n(t) = n^{-1} \sum_{j=1}^{n} e^{itV_j},$$

where $\hat{\mu}_n = \hat{\mu}_n(W_1, ..., W_n)$ and $\hat{\sigma}_n = \hat{\sigma}_n(W_1, ..., W_n)$ are estimators for μ and σ, respectively, such that for each $a > 0$ and $b \in \mathbb{R}$,

$$\hat{\mu}_n(aW_1 + b, \ldots, aW_n + b) = a\hat{\mu}_n(W_1, \ldots, W_n) + b,$$

$$\hat{\sigma}_n(aW_1 + b, \ldots, aW_n + b) = a\hat{\sigma}_n(W_1, \ldots, W_n). \tag{2.3}$$

Since $W_1, \ldots, W_n$ and $\varphi_n(t)$ are free of μ and σ, set $\mu = 0$ and $\sigma = 1$ without loss of generality and define the test statistic as

$$D_{n,c} = \int_{-\infty}^{\infty} (\varphi_n(t) - \exp\{-|t|\})^2 \exp\{-c|t|\}dt, \tag{2.4}$$

where c is a fixed positive weighting parameter. We reject H_0 if $D_{n,c}$ gets large values. Note that the test statistic (2.4) is a weighted L^2-distance between $\varphi_n(t)$ and the characteristic function of $C(0, 1)$, $\exp\{-|t|\}$. Moreover, under H_0 and for suitable choice of $\{(\hat{\mu}_n, \hat{\sigma}_n)\}_{n\geq 1}$, $\varphi_n(t)$ converges in probability to $\exp\{-|t|\}$. Then, by (2.4), $D_{n,c}$ can be obtained through the straightforward calculations as

$$D_{n,c} = \frac{2}{n}\sum_{j,k=1}^{n} \frac{c}{c^2 + (V_j - V_k)^2} - 4\sum_{j=1}^{n} \frac{1 + c}{(1 + c)^2 + V_j^2 + (2n/2 + c)}. \tag{2.5}$$

2.4 Computational Aspects

In order to apply our propose tests to the real data, estimation of the parameters $\boldsymbol{\xi}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\lambda}$ are needed for evaluating $\boldsymbol{Y}^*$. In the forthcoming data analyses, we use the maximum likelihood (ML) estimates of $\boldsymbol{\xi}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\lambda}$, which give consistent estimates Balakrishnan et al. (2014). The use of consistent estimates may provide a close approximation to the true null distribution. Also, for the fix n, $\boldsymbol{\xi}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\Omega}$, we use the procedure in Algorithm 2.1 to obtain the critical values. Since the power of a test plays important rule in the statistical inference, we present a convenience procedure in Algorithm 2.2 to compute the power of the proposed tests by using the obtained critical values. We note finally that the process of calculating corresponding p-values of the tests are described in Algorithm 2.3.

Algorithm 2.1. Critical Values Computation Algorithm

1. *Simulate a sample of size n from MSN($\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda}$).*
2. *Calculate the ML estimate of the parameters $\boldsymbol{\xi}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\Omega}$.*
3. *Compute the canonical form based on Proposition 2.2.*
4. *Calculate $T^{(1)}_{n,a}$, $T^{(2)}_{n,a}$, and $D_{n,c}$ formulated in Section 2.*
5. *Repeat steps 1-4 B times.*
6. *The value of the critical constant $T^{(1)}_{n,a}$, $T^{(2)}_{n,a}$, and $D_{n,c}$ is determined with the quantiles 100α from the simulation of $T^{(1)}_{n,a}$, $T^{(2)}_{n,a}$, and $D_{n,c}$, where α is a prespecified significance level.*

Algorithm 2.2. Power Value Computation Algorithm

1. *Simulate a sample of size n from alternative distribution.*
2. *Estimate parameters of the MSN distribution via ML approach. Based on the canonical form (Proposition 2.2), compute $T_{n,a}^{(1)}$, $T_{n,a}^{(2)}$, and $D_{n,c}$ as described in Section 2.*
3. *Repeat steps 1-2 B times.*
4. *The power of the test statistics $T_{n,a}^{(1)}$, $T_{n,a}^{(2)}$, and $D_{n,c}$ is determined with the number of $T_{n,a}^{(1)}$, $T_{n,a}^{(2)}$, and $D_{n,c}$ that is grater than the critical constant (obtained in Algorithm 2.1) divided by B.*

Algorithm 2.3. Procedure of Computing p-values of $T_{n,a}^{(1)}$, $T_{n,a}^{(2)}$, and $D_{n,c}$.

1. *Obtain the ML estimate of the MSN parameters $\boldsymbol{\xi}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\Omega}$ for a real data set. and calculate $T_{n,a}^{(1)}$, $T_{n,a}^{(2)}$, and $D_{n,c}$ for the real data set.*
2. *Generate a sample of size n (same size of the real data) from MSN$(\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\Omega}}, \hat{\boldsymbol{\lambda}})$.*
3. *Calculate $T_{n,a}^{(1)}$, $T_{n,a}^{(2)}$, and $D_{n,c}$ based on canonical form presented in Section 2 for the simulated data and denote them $T_{n,a}^{(1)*}$, $T_{n,a}^{(2)*}$, and $D_{n,c}^{*}$.*
4. *Repeat steps 2 and 3 B times.*
5. *The corresponding p-values of $T_{n,a}^{(1)}$, $T_{n,a}^{(2)}$, and $D_{n,c}$ is determined with the number of $T_{n,a}^{(1)*}$, $T_{n,a}^{(2)*}$, and $D_{n,c}^{*}$ that exceed the values $T_{n,a}^{(1)}$, $T_{n,a}^{(2)}$, and $D_{n,c}$ divided by B.*

3 Simulation Study

Examples 3.1. In order to evaluate the true significance level, an artificial analysis is conducted, when data come from a 3-dimensional MSN distribution with various sample sizes n =100, 200 and 300. The presumed parameters are

$$\boldsymbol{\xi} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \qquad \boldsymbol{\Omega} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2.5 & 1 \\ 1 & 1 & 5 \end{bmatrix}, \qquad \boldsymbol{\lambda} = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}. \tag{3.1}$$

The parameters set in (3.1) have a Mardia index of skewness 0.5 and a sample from this distribution would have $\lambda_* \approx 3.3$. So, they can be considered to generate data from the MSN distribution with a moderate skewness. Through fitting the MSN model to the simulated data in each 1000 replication of the trails we obtain ML estimate and compute three tests $T_{n,a}^{(1)}$, $T_{n,a}^{(2)}$, and $D_{n,c}$. Table 1 shows the significance levels of each test obtained for $\alpha = 0.05$. It can be observed from Table 1 that significance levels are very closed to nominal level of $\alpha = 0.05$ for all sample sizes, showing that the proposed tests are less sensitive to the sample size compared with table 1 in Balakrishnan et al. (2014), and also confirming that all tests can achieve the significance level without any restriction.

Table 1: Simulated significance level based on Nominal $\alpha = 0.05$.

	size	100	200	300	100	200	300
		$a = c = 1.0$			$a = c = 1.5$		
$T^{(1)}_{n,a}$	α	0.048	0.049	0.050	0.049	0.051	0.053
$T^{(2)}_{n,a}$	α	0.047	0.051	0.051	0.048	0.050	0.050
$D_{n,c}$	α	0.049	0.050	0.050	0.052	0.051	0.052

Examples 3.2. We conduct a second simulation study aims at investigating the power of the proposed tests under different conveniently alternatives. Following Balakrishnan et al. (2014), we consider two family of distributions: ($C1$) a multivariate skew-t distribution Azzalini and Capitanio (2003) and ($C2$) a mixture of two MSN distributions. To compute the power of tests, Algorithm 2.2 is exploited where the selected number of iterations is $B = 1000$ and samples are generated as follows. For $C1$, the multivariate skew-t distribution with parameters set (3.1), $n = 100$, and various degree of freedoms ν =1, 2, 3, 5, and 10 are considered. As the second group of alternative distributions $C2$, data are generated from a mixture of two MSN distributions with pdf

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \pi f_1(\boldsymbol{x}) + (1 - \pi) f_2(\boldsymbol{x}),$$

where π is a mixture proportion, $f_1(\cdot)$ and $f_2(\cdot)$ are the pdf of $\boldsymbol{X}_1 \sim MSN(\boldsymbol{\xi}_1, \boldsymbol{\Omega}_1, \boldsymbol{\lambda}_1)$ and $\boldsymbol{X}_2 \sim MSN(\boldsymbol{\xi}_2, \boldsymbol{\Omega}_2, \boldsymbol{\lambda}_2)$. For the sample size 100, 200 and 300, and 1000, the presumed parameters in (3.1) are considered for generating X_1 and

$$\boldsymbol{\xi}_2 = \begin{bmatrix} 5 \\ 10 \\ -4 \end{bmatrix}, \qquad \boldsymbol{\Omega}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \boldsymbol{\lambda}_2 = \begin{bmatrix} 5 \\ -5 \\ 6 \end{bmatrix}.$$

Moreover, three mixture proportion 0.5, 0.3 and 0.1 are also chosen to allow different levels of variation from the MSN distribution. For $C1$ alternative hypothesis, the computed power of the proposed tests $T^{(1)}_{n,a}$, $T^{(2)}_{n,a}$, and $D_{n,c}$ are summarized in Table 2. The results depicted in Table 2 show reasonable values for the tests power. Since the multivariate skew-t tends to the MSN model for larger degrees of freedom, it is expected that the power of the test decreases as ν increased. For the second alternative hypothesis $C2$, the power of our proposed tests as well as two tests introduced by Balakrishnan et al. (2014), T^{*} and U, are computed. Table 3 shows the power of the five considered tests for $C2$ case, which shows that both couples (T^{*}, U) and $(T^{(1)}_{n,a}, T^{(2)}_{n,a})$ have a very similar behaviour. It can be seen that, for $\pi = 0.5$, the power of all implemented tests is less than those obtained for $\pi = 0.3$ and 0.1. One may expected this manner since under the considered parameters and $\pi = 0.5$, the mixed distribution gets close to the one-component MSN model. However, it is clear that $T^{(1)}_{n,a}$ and $T^{(2)}_{n,a}$ have the biggest power in almost all choose of sample size and π.

Table 2: Power values when the alternative hypothesis is the multivariate skew-t distribution.

	$T^{(1)}_{n,a}$	$T^{(2)}_{n,a}$	$D_{n,c}$	$T^{(1)}_{n,a}$	$T^{(2)}_{n,a}$	$D_{n,c}$
df		$a = c = 1.0$			$a = c = 1.5$	
1	0.921	0.930	0.832	0.932	0.936	0.870
2	0.764	0.771	0.601	0.780	0.784	0.638
3	0.561	0.569	0.490	0.580	0.585	0.506
5	0.334	0.340	0.322	0.342	0.345	0.330
10	0.257	0.262	0.205	0.264	0.267	0.212

Table 3: Power values when the alternative hypothesis is the mixture of two MSN distributions for three choices of sample size.

Distibution	n		T^{*}	U	$T^{(1)}_{n,a}$	$T^{(2)}_{n,a}$	$D_{n,c}$
SN mixture	100	$\pi = 0.5$	0.757	0.758	0.862	0.865	0.771
		$\pi = 0.3$	0.993	0.991	0.993	0.997	0.983
		$\pi = 0.1$	0.978	0.975	0.984	0.987	0.978
	200	$\pi = 0.5$	0.784	0.783	0.886	0.892	0.841
		$\pi = 0.3$	0.950	0.946	0.982	0.985	0.970
		$\pi = 0.1$	0.882	0.882	0.902	0.910	0.871
	300	$\pi = 0.5$	0.784	0.784	0.836	0.841	0.795
		$\pi = 0.3$	0.901	0.894	0.938	0.941	0.921
		$\pi = 0.1$	0.767	0.765	0.810	0.815	0.779

Examples 3.3. In order to compare our proposed tests with those considered by Meintanis and Hlavka (2010), denoted by MH, and Balakrishnan et al. (2014) the third simulation study is conducted. We only focus on the bivariate case since the MH test is not computationally available for $d > 2$. Aims at comparing significant level of the tests, in each replication of 1000 trials, we simulate samples from a 2-dimensional MSN distribution for the various sample sizes n =100, 200, 300, and with true parameter values

$$\boldsymbol{\xi} = [1\ 2]^{\top}, \qquad \boldsymbol{\Omega} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \tag{3.2}$$

where ρ was chosen as 0.4 and 0.8 and $\boldsymbol{\alpha} = [1\ 2]^{\top}$ and $[4\ 2]^{\top}$. Moreover, we consider three choices of bivariate gamma mixtures to compare tests' power. For this purpose, we simulate x_0, x_1, x_2, x_3 of sizes n =100, 200, or 300, independently from four gamma distributions with parameters β = 1, 2, 1, 2, and ϑ =1, 1, 2, 3, respectively, and obtain y_1, y_2, y_3, as

$$y_1 = x_0 + x_1, \qquad y_2 = x_0 + x_2, \qquad y_3 = x_0 + x_3.$$

Then, the considered bivariate gamma mixtures are: a) (y_1, y_2), b) (y_1, y_3), and c) (y_2, y_3).

Results summarized in Table 4 show that along all sample sizes and all combinations of real parameters, the significance levels of $T^{(1)}_{n,a}$, $T^{(2)}_{n,a}$, and $D_{n,c}$ are very closed to the nominal level of $\alpha = 0.05$. In contrast with the MH, T^{*} and U tests, these results confirm that the significance level of our proposed tests does not critically depend on the parameter and sample size. It can be also observed that the power values of $T^{(1)}_{n,a}$, $T^{(2)}_{n,a}$, and $D_{n,c}$ are much higher than those related to the MH, T^{*} and U tests for all sample sizes.

Table 4: Simulated values of significance levels and power when data are respectively generated from the bivariate SN distribution and from the bivariate gamma mixtures model.

		Significance level (for *MSN* generator)				Power (for Gamma mixture model)		
		$\rho = 0.4$	$\rho = 0.4$	$\rho = 0.8$	$\rho = 0.8$	(a)	(b)	(c)
	Test	$\boldsymbol{\lambda} = [1,2]^\top$	$\boldsymbol{\lambda} = [4,2]^\top$	$\boldsymbol{\lambda} = [1,2]^\top$	$\boldsymbol{\lambda} = [4,2]^\top$			
$n = 100$	T^{*}	0.042	0.111	0.047	0.138	0.489	0.345	0.554
	U	0.030	0.092	0.026	0.126	0.476	0.330	0.550
	MH	0.050	0.054	0.036	0.050	0.476	0.464	0.538
	$T^{(1)}_{n,a}$	0.052	0.049	0.046	0.052	0.862	0.868	0.795
	$T^{(2)}_{n,a}$	0.054	0.048	0.055	0.050	0.852	0.855	0.790
	$D_{n,c}$	0.052	0.050	0.051	0.043	0.682	0.679	0.571
$n = 200$	T^{*}	0.043	0.055	0.058	0.045	0.276	0.121	0.257
	U	0.032	0.044	0.035	0.035	0.271	0.114	0.255
	MH	0.061	0.048	0.059	0.052	0.297	0.302	0.406
	$T^{(1)}_{n,a}$	0.051	0.052	0.050	0.050	0.554	0.492	0.596
	$T^{(2)}_{n,a}$	0.053	0.048	0.052	0.050	0.550	0.499	0.590
	$D_{n,c}$	0.045	0.049	0.054	0.055	0.487	0.452	0.561
$n = 300$	T^{*}	0.045	0.051	0.049	0.046	0.159	0.053	0.126
	U	0.047	0.051	0.054	0.041	0.164	0.050	0.129
	MH	0.054	0.049	0.055	0.042	0.199	0.215	0.298
	$T^{(1)}_{n,a}$	0.052	0.051	0.050	0.050	0.412	0.390	0.461
	$T^{(2)}_{n,a}$	0.052	0.052	0.049	0.050	0.408	0.395	0.462
	$D_{n,c}$	0.053	0.049	0.049	0.049	0.371	0.362	0.409

4 Real Data Examples

As illustrative purposes, we apply the proposed methodology on four real-world data sets. We consider, at first, the well-known biomedical data related to the 102 male and 100 female athletes that are collected by the Australian Institute of Sport. Called AIS data, Azzalini and Capitanio (1999) fitted the MSN distribution to a subset of the data for verifying its performance. Moreover, Lin et al. (2014) analyzed the combination of male and female to examine the ability of two-component mixture of the MSN distributions. We focus on four variables: body mass index (bmi), sum of skin folds (ssf), body fat percentage (Bfat) and lean body mass (lbm). In the second real-world example, the leptograpus crabs data set, previously studied by Campbell and Mahon (1974), is used. The crabs data contain 5 biological attributes measured on 200 crabs (100 male and 100 female) from the genus leptograpus. The third considered data is related to the three variables of the Wisconsin Breast Cancer (WDBC) data Mangasarian et al. (1995). In the analysis of WDCB, the tests statistic are examined through three variables, the mean texture, the largest area and the largest smoothness that can be classified as malignant (212 instances) or benign (357 instances). Finally, for the fourth data example, the performance of goodness-of-fit tests for MSN distribution are compared on the five variate open/closed book (OCB) dataset Mardia et al. (1979). The OCB data are recently analyzed by Kim et al. (2016), who considered three new skewed factor models, namely the SN, skew-*t*, and the generalized SN factor models, depending on a selection mechanism of the factors. The OCB data contain the results of five proficiency namely mechanics, vectors, algebra, analysis, and statistics tested on $n = 88$ students.

Table 5: p-values of tests of multivariate skew-normality for three sets of variables from the AIS data set.

	Male			**Female**		
Test	Three variables	Three variables	Four variables	Three variables	Three variables	Four variables
	(bmi, Bfat, lbm)	(ssf, Bfat, lbm)	(bmi, ssf, Bfat, lbm)	(bmi, Bfat, lbm)	(ssf, Bfat, lbm)	(bmi, ssf, Bfat, lbm)
U	0.001	≈ 0	0.097	0.178	0.982	0.185
T^{*}	0.001	≈ 0	≈ 0	0.289	0.790	≈ 1
$T_{n,a}^{(1)}$	0.120	0.078	0.156	0.459	0.991	≈ 1
$T_{n,a}^{(2)}$	0.112	0.075	0.148	0.420	0.986	≈ 1
$D_{n,c}$	0.092	0.064	0.126	0.381	0.885	0.985

Applying the expectation-maximization algorithm Dempster et al. (1977) to fit the MSN distribution to the data and finding its ML estimate, we preform Algorithm 2.3 to obtain the p-values of the tests. Tables 5 and 6 show the p-values of the five tests. Results depicted in Table 5 is related to the AIS data and show that for the three-variable sets (bmi, Bfat, lbm) and (ssf, Bfat, lbm), and for the four-variable set of the male group, there is not enough evidence in favour of the MSN distribution. However, for the female group, most of the considered tests suggest to accept the hypothesis of the multivariate skew-normality distribution.

The results of Table 6 for the crabs data reveal that our proposed tests show more evidence in favour of the MSN distribution for each group of the male and female, while all the tests suggest to reject the null hypothesis when the data are combined. Similarly, by looking at the test p-values of the WDBC data, one can see sufficient information to accept when the data are may follow the MSN distribution separately for malignant and benign classes. But on all classes, we suggest to reject the multivariate skew-normality assumption since all of the p-values of the considered tests tend to approaches zero. We note here that the joint distributions of the five-variable crabs and three-variable WDBC data sets can be at least bimodal, being a mixture of measures coming from a group of males or females and a group of malignant or benign. It will be of interest to propose a goodness-of-fit test of multivariate skew-normal mixture models based on the canonical form.

Finally, the p-values of the proposed tests included in Table 6 lead us to conclude in favour of the MSN distribution for the OCB data set, which are significantly greater than the p-value of the U and T^* tests.

Table 6: p-values of the tests of multivariate skew-normality for three data sets.

	Crabs data			**WDBC data**			**OCB data**
Test	(Male, Female)	Male	Female	(Malignant, Benign)	Malignant	Benign	
U	≈ 0	0.332	0.395	0.056	0.487	0.974	0.159
T^*	0.001	0.359	0.451	0.053	0.468	0.604	0.174
$T^{(1)}_{n,a}$	≈ 0	0.891	0.926	0.009	0.682	0.725	0.530
$T^{(2)}_{n,a}$	≈ 0	0.907	0.950	0.005	0.664	0.713	0.552
$D_{n,c}$	0.002	0.755	0.794	0.012	0.564	0.694	0.503

5 Concluding Remarks

In this paper, we have dealt with proposing some goodness-of-fit tests for assessing the adequacy of the MSN distribution. To construct the test statistics, we focus on the canonical form of the MSN model and on both empirical Laplace transform and empirical canonical form characteristic functions. Three algorithms for computing the critical values, power values and corresponding p-values of the tests are presented. Numerical results of simulations and real data examples suggest that the proposed tests can work well without any computation and/or dimensional restrictions.

The utility of our current approach can be extended to obtain a test statistic of the multivariate skew-t distribution based on canonical form. Furthermore, another possible extension of the work herein is to consider a test to assess if a sample comes from a mixture of the MSN distributions.

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171–178.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution.*Journal of the Royal Statistical Society: Series B*, **61**, 579-602.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution.*Journal of the Royal Statistical Society: Series B*, **65**, 367-389.

Balakrishnan, N., Capitanio, A. and Scarpa, B. (2014). A test for multivariate skew-normality based on its canonical form.*Journal of Multivariate Analysis*, **128**, 19-32.

Campbell, N.A. and Mahon, R.J. (1974). A multivariate study of variation in two species of rock crab of genus Leptograpsus. *Australian Journal of Zoology*, **22**, 417-425.

Capitanio, A. (2012). On the canonical form of scale mixtures of skew-normal distributions.*Available at arXiv.org:1207.0797*.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1-22.

Gürtler, N., Henze, N. (2000). Goodness-of-fit tests for the Cauchy distribution based on the empirical characteristic function. *Annals of the Institute of Statistical Mathematics*, **52**, 267-286.

Henze, N., Meintanis, S. and Ebner, B. (2012). Goodness-of-fit tests for the gamma distribution based on the empirical laplace transform.*Communication in Statistics-Theory and Method*, **41**, 1543-1556.

Jarque, C.M. and Bera, A.K. (1987). A test for normality of observations and regression residuals.*International Statistical Review/Revue Internationale de Statistique*, 163-172.

Kim H. M., Maadooliat, M. Arellano-Valle R. B. and Genton, M.G. (2016). Skewed factor models using selection mechanisms, *Journal of Multivariate Analysis*, **145**, 162-177.

Lin, T.I., Ho, H.J. and Lee, C.R. (2014). Flexible mixture modelling using the multivariate skew-t-normal distribution.*Statistics and Computing*, **24**, 531-546.

Mangasarian, O.L., Street, W.N. and Wolberg, W.H. (1995). Breast cancer diagnosis and prognosis via linear programming.*Operations Research*, **43**, 570-577.

Mardia, K.V., Kent, J.T. and Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press, INC.

Mateu, G., Puig, P. and Pewsey, A. (2007). Goodness-of-fit tests for the skew-normal distribution when the parameters are estimated from the data.*Communication in Statistics-Theory and Method*, **36**, 1735–1755.

Meintanis, S.G. (2007). A kolmogorov-smirnov type test for skew normal distributions based on the empirical moment generating function.*Journal of Statistical Planning and Inference*, **137**, 2681–2688.

Meintanis, S.G. and Hlavka, Z. (2010). Goodness-of-fit test for bivariate and multivariate skew-normal distributions.*Scandinavian Journal of Statistics*, **37**, 701-714.