Received 24 March 2022
Revised 8 April 2022
Accepted 21 April 2022

Inferences on location parameters based on independent multivariate skew normal distributions

Ziwei Ma
Department of Mathematics, University of Tennessee at Chattanooga, Chattanooga, Tennessee, USA
Tonghui Wang
Department of Mathematical Sciences, New Mexico State University, Las Cruces, New Mexico, USA
Zheng Wei
Department of Mathematics and Statistics, Texas A & M University-Corpus Christi, Corpus Christi, Texas, USA, and
Xiaonan Zhu
Department of Mathematics, University of North Alabama, Florence, Alabama, USA

Abstract

Purpose – The purpose of this study is to extend the classical noncentral F-distribution under normal settings to noncentral closed skew F-distribution for dealing with independent samples from multivariate skew normal (SN) distributions.

Design/methodology/approach – Based on generalized Hotelling's T^2 statistics, confidence regions are constructed for the difference between location parameters in two independent multivariate SN distributions. Simulation studies show that the confidence regions based on the closed SN model outperform the classical multivariate normal model if the vectors of skewness parameters are not zero. A real data analysis is given for illustrating the effectiveness of our proposed methods.

Findings – This study's approach is the first one in literature for the inferences in difference of location parameters under multivariate SN settings. Real data analysis shows the preference of this new approach than the classical method.

Research limitations/implications – For the real data applications, the authors need to remove outliers first before applying this approach.

Practical implications – This study's approach may apply many multivariate skewed data using SN fittings instead of classical normal fittings.

Originality/value – This paper is the research paper and the authors' new approach has many applications for analyzing the multivariate skewed data.

Keywords Confidence regions, Pivotal method, Location parameter, Multivariate skew normal family, Hotelling's T^2

Paper type Research paper

1. Introduction

Although the normal distribution is a standard assumption for modeling observations in general, practitioners and researchers prefer more flexible models that account for the

The authors would like to thank reviewers for their valuable suggestions and comments, which improved the manuscript.

Asian Journal of Economics and Banking
Vol. 6 No. 2, 2022
pp. 270-281
Emerald Publishing Limited
e-ISSN: 2633-7991
p-ISSN: 2615-9821
DOI 10.1108/AJEB-03-2022-0034

non-normality when the data collected in finance and econometric fields. The family of skew normal (SN) distributions, introduced by Azzalini (1985) for the univariate case, Azzalini and Valle (1996) for the multivariate case and Chen and Gupta (2005) for the matrix variate case, becomes a popular parametric family in statistical analysis of real data which account for asymmetry. There are several successful applications using SN, like modeling skewness premium of a financial asset by Carmichael and Coën (2013), addressing "wrong skewness" problems in stochastic frontier models by Wei *et al.* (2021). Here just list a few, an updated review was given by Adcock and Azzalini (2020).

Based on the definition given in Arellano-Valle *et al.* (2005), a p-dimensional random vector $\boldsymbol{Y}$ is said to be SN distributed with the location parameter vector $\boldsymbol{\mu} \in \Re^p$, the scale parameter matrix Σ (a $p \times p$ positive definite matrix), and the shape parameter vector $\boldsymbol{\lambda} \in \Re^p$, denoted as $\boldsymbol{Y} \sim SN_p(\boldsymbol{\mu}, \Sigma, \boldsymbol{\lambda})$, if its probability density function (pdf) is given by

$$f_Y(\boldsymbol{y}) = 2\phi_p(\boldsymbol{y};\boldsymbol{\mu},\Sigma)\Phi\left(\boldsymbol{\lambda}'\Sigma^{-1/2}(\boldsymbol{y}-\boldsymbol{\mu})\right), \qquad \boldsymbol{y} \in \Re^p, \tag{1}$$

where $\phi_p(\cdot;\boldsymbol{\mu},\Sigma)$ is the pdf of the p-dimensional normal distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ, and $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the univariate standard normal distribution. Extensions of Equation (1) are investigated by many researchers (see Azzalini and Capitanio, 1999; Wang *et al.*, 2009; Young *et al.*, 2016; Li *et al.*, 2018).

For the univariate SN family, constructing plausibility regions for skewness parameter was discussed by Zhu *et al.* (2017) using inferential models (IMs). The joint plausibility regions for location parameter and skewness parameter were studied by Ma *et al.* (2018) when scale parameter is known using IMs, and the joint plausibility regions for location parameter and scale parameter were constructed by Zhu *et al.* (2018) when skewness parameter is given. For multivariate SN model, the confidence regions for location parameter are obtained by Ma *et al.* (2019). In this work, we study the difference of location parameters based on independent multivariate SN distributions so that the generalized Hotelling's T^2, and noncentral closed skew F-distributions are used. Under the assumption of equal but unknown scale parameters, the confidence regions for differences of location parameters of the multivariate SN model are proposed. Simulation studies show that the proposed confidence regions have higher relative coverage frequency rates than those in classical normal model for skewed data.

The organization of this paper is listed below. In Section 2, the definition of matrix variate SN distribution is introduced and some useful properties of sampling distribution on difference of sample means are derived. In Section 3, the confidence regions on the difference of location parameters by pivotal method are proposed when scale parameters from two populations are assumed to be equal but unknown. A group of simulation studies, which illustrate the effectiveness of our proposed methods, are given in Section 4, followed by a real data example in Section 5. The conclusion is given in Section 6.

2. Matrix variate SN distributions and sampling distributions

Let $M_{n\times k}$ be the set of all $n \times k$ matrices over the real field $\Re$ and $\Re^n = M_{n\times 1}$. The transpose of a matrix A is denoted as A'. The $n \times n$ identity matrix is denoted as I_n, the constant vector $(1,\ldots,1)' \in \Re^n$ is denoted as $\mathbf{1}_n$, and $\bar{J}_n = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$. For $B = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n)' \in M_{n\times k}$ with $\boldsymbol{b}_i \in \Re^k$, let $\mathrm{Vec}(B) = (\boldsymbol{b}_1', \boldsymbol{b}_2', \ldots, \boldsymbol{b}_n')' \in \Re^{nk}$. For positive definite matrix $T \in M_{n\times n}$, we use T^{-1}, $T^{1/2}$ and $T^{-1/2}$ to denote, respectively, the inverse, symmetric square root of T, and symmetric square root of T^{-1}. For $B \in M_{m\times n}$, $C \in M_{n\times p}$, we use $B \otimes C$ to denote the Kronecker product of B and C. Through this paper, $N(0, 1)$ represents the standard normal distribution and bold phase letters represent vectors.

Definition 2.1. Ye *et al.* (2014) *The $n \times p$ random matrix Y is said to have a* SN ***matrix distribution*** *with location matrix M, scale matrix $V \otimes \Sigma$ and skewness* *parameter matrix* $\boldsymbol{\gamma}\otimes\boldsymbol{\lambda}'$, *denoted by* $Y\sim SN_{n\times p}(M,V\otimes\Sigma,\boldsymbol{\gamma}\otimes\boldsymbol{\lambda}')$, *if* $\boldsymbol{y}\equiv\mathrm{Vec}(Y)\sim SN_{np}(\boldsymbol{\mu},V\otimes\Sigma,\boldsymbol{\gamma}\otimes\boldsymbol{\lambda})$, *where* $M\in M_{n\times p}$, $V\in M_{n\times n}$, $\boldsymbol{\mu}=\mathrm{Vec}(M)$, $\boldsymbol{\gamma}\in\Re^n$ *and* $\boldsymbol{\lambda}\in\Re^p$

Suppose that $X_1\in M_{n_1\times p}$ and $X_2\in M_{n_2\times p}$ are two independent sample matrices such that

$$X_i\sim SN_{n_i\times p}\left(\mathbf{1}_{n_i}\otimes\boldsymbol{\mu}_i',I_{n_i}\otimes\Sigma_i,\mathbf{1}_{n_i}\otimes\boldsymbol{\lambda}_i'\right)\tag{2}$$

for $i=1,2$. We are interested in analyzing the difference vector $\boldsymbol{\mu}_d=\boldsymbol{\mu}_1-\boldsymbol{\mu}_2$. The sampling distributions of the sample mean and sample covariance matrix are given by Ma *et al.* (2019) in following Lemma.

Lemma 2.1. Ma *et al.* (2019) *Let* $Y\sim SN_{n\times p}(\mathbf{1}_n\otimes\boldsymbol{\mu}',I_n\otimes\Sigma,\mathbf{1}_n\otimes\boldsymbol{\lambda}')$. *Then,*

$$\bar{Y}=\left(\frac{1}{n}\mathbf{1}_n'Y\right)'\sim SN_p\left(\boldsymbol{\mu},\frac{\Sigma}{n},\sqrt{n}\,\boldsymbol{\lambda}\right)$$

and

$$(n-1)S=Y'\left(I_n-\bar{J}_n\right)Y\sim W_p(n-1,\Sigma)$$

are independently distributed, where $W_p(n-1,\Sigma)$ *represents the p-dimensional Wishart distribution with* $n-1$ *degrees of freedom and the mean* Σ.

By Lemma 2.1, we have

$$\bar{\boldsymbol{X}}_i=\left(\frac{1}{n_i}\mathbf{1}_{n_i}'X_i\right)'\sim SN_p\left(\boldsymbol{\mu}_i,\frac{\Sigma_i}{n_i},\sqrt{n_i}\,\boldsymbol{\lambda}_i\right)\tag{3}$$

and

$$(n_i-1)S_i=X_i'(I_{n_i}-\bar{J}_{n_i})X_i\sim W_p(n_i-1,\Sigma_i)\tag{4}$$

for $i=1,2$. It is natural to use the statistic $\bar{\boldsymbol{X}}_d=\bar{\boldsymbol{X}}_1-\bar{\boldsymbol{X}}_2$ to inference on $\boldsymbol{\mu}_d$.

The difference between two independent SN distributed random vectors follows a closed SN distribution, which is reviewed below.

Definition 2.2. *(Gonzalez-Farias et al.* (Gonzalez-Farias *et al.*, 2004) *A random vector* $\boldsymbol{Y}\in\Re^p$ *is said to have* ***closed*** SN ***distribution (CSN)****, denoted as* $CSN_{p,q}(\boldsymbol{\mu},\Sigma,D,\boldsymbol{v},\Delta)$, *if its pdf is*

$$f_{p,q}(\boldsymbol{y};\boldsymbol{\mu},\Sigma,D,\boldsymbol{v},\Delta)=C\phi_p(\boldsymbol{y};\boldsymbol{\mu},\Sigma)\Phi_q(D(\boldsymbol{y}-\boldsymbol{\mu});\boldsymbol{v},\Delta),\qquad \boldsymbol{y}\in\Re^p,$$

where $C^{-1}=\Phi_q(0;\boldsymbol{v},\Delta+D\Sigma D')$, $p\geq1$, $q\geq1$, $\boldsymbol{\mu}\in\Re^p$, $\Sigma\in M_{p\times p}^+$, $D\in M_{q\times p}$, $\boldsymbol{v}\in\Re^q$, $\Delta\in M_{q\times q}^+$ *and* $\phi_k(\cdot;\boldsymbol{\eta},\Omega)$, $\Phi_k(\cdot;\boldsymbol{\eta},\Omega)$ *are the pdf and cdf of a k-dimensional normal distribution.*

For simplicity, we assume that $\boldsymbol{\nu}=\mathbf{0}$ so that $\boldsymbol{Y}\sim CSN_{p,q}(\boldsymbol{\mu},\Sigma,D,\Delta)$. The following two properties of CSN can lead to the distribution of $\bar{\boldsymbol{X}}_d$.

Lemma 2.2. *(Gonzalez-Farias et al.* (Gonzalez-Farias *et al.*, 2004) *Let* $\boldsymbol{Y}\sim CSN_{p,q}(\boldsymbol{\mu},\Sigma,D,\Delta)$

(1) *For an arbitrary constant* $\boldsymbol{b}\in\Re^p$,

$$\boldsymbol{Y}+\boldsymbol{b}\sim CSN_{p,q}(\boldsymbol{\mu}+\boldsymbol{b},\Sigma,D,\Delta)\tag{5}$$

(2) *For nonzero real number* $c \in \Re$,

$$c\boldsymbol{Y} \sim CSN_{p,q}\left(c\boldsymbol{\mu}, c^2\Sigma, c^{-1}D, \Delta\right); \tag{6}$$

(3) *Let* $\boldsymbol{Y}_i \sim CSN_{p,q_i}(\boldsymbol{\mu}_i, \Sigma_i, D_i, \Delta_i)$*, for* $i = 1, 2$*, be independently distributed. Then,*

$$\boldsymbol{Y}_1 + \boldsymbol{Y}_2 \sim CSN_{p,q_1+q_2}\left(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \Sigma_1 + \Sigma_2, D^*, \Delta^*\right) \tag{7}$$

where

$$D^* = \begin{pmatrix} D_1\Sigma_1(\Sigma_1+\Sigma_2)^{-1} \\ D_2\Sigma_2(\Sigma_1+\Sigma_2)^{-1} \end{pmatrix}, \qquad \Delta^* = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

and

$$\begin{aligned} A_{11} &= \Delta_1 + D_1\Sigma_1D_1' - D_1\Sigma_1(\Sigma_1+\Sigma_2)^{-1}\Sigma_1D_1', \\ A_{22} &= \Delta_2 + D_2\Sigma_2D_2' - D_2\Sigma_2(\Sigma_1+\Sigma_2)^{-1}\Sigma_2D_2', \\ A_{12} &= -D_1\Sigma_1(\Sigma_1+\Sigma_2)^{-1}\Sigma_2D_2'. \end{aligned}$$

In term of CSN, $\bar{\boldsymbol{X}}_i \sim CSN_{p,1}\left(\boldsymbol{\mu}_i, \frac{\Sigma_i}{n_i}, n_i\boldsymbol{\lambda}_i'\Sigma_i^{-1/2}, 1\right)$ for $i = 1, 2$. Thus, we obtain the distribution of $\bar{\boldsymbol{X}}_d$.

Theorem 2.1. Let $\bar{\boldsymbol{X}}_d = \bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2$ *with* $\bar{\boldsymbol{X}}_i \sim CSN_{p,1}\left(\boldsymbol{\mu}_i, \frac{\Sigma_i}{n_i}, n_i\boldsymbol{\lambda}_i'\Sigma_i^{-1/2}, 1\right)$ *for* $i = 1, 2$. *Then,*

$$\bar{\boldsymbol{X}}_d \sim CSN_{p,2}(\boldsymbol{\mu}_d, \Sigma_d, D_d, \Delta_d) \tag{8}$$

where

$$\begin{aligned} &\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \qquad \Sigma_d = \frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2} \\ &D_d = \begin{pmatrix} \boldsymbol{\lambda}_1'\Sigma_1^{1/2}\Sigma_d^{-1} \\ -\boldsymbol{\lambda}_2'\Sigma_2^{1/2}\Sigma_d^{-1} \end{pmatrix}, \quad \Delta_d = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \end{aligned} \tag{9}$$

with

$$\begin{aligned} A_{11} &= 1 + n_1\boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1'\Sigma_1^{1/2}\Sigma_d^{-1}\Sigma_1^{1/2}\boldsymbol{\lambda}_1, \\ A_{22} &= 1 + n_2\boldsymbol{\lambda}_2'\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_2'\Sigma_2^{1/2}\Sigma_d^{-1}\Sigma_2^{1/2}\boldsymbol{\lambda}_2, \\ A_{12} &= A_{21} = \boldsymbol{\lambda}_1'\Sigma_1^{1/2}\Sigma_d^{-1}\Sigma_2^{1/2}\boldsymbol{\lambda}_2. \end{aligned}$$

Proof. By part (2) and (3) of Lemma 2.2, the desired result follows immediately. □

Remark. If $\boldsymbol{\lambda}_2 = \mathbf{0}$, i.e. $\boldsymbol{X}_2$ following multivariate normal distribution with mean $\boldsymbol{\mu}_2$ and covariance $\frac{\Sigma_2}{n_2}$, the distribution of difference $\bar{\boldsymbol{X}}_d$ has the form $\boldsymbol{X}_d \sim CSN_{p,1}(\boldsymbol{\mu}_d, \Sigma_d, \boldsymbol{\lambda}_1'\Sigma_1^{1/2}\Sigma_d^{1/2}, A_{11})$ which can be further expressed as $\boldsymbol{X}_d \sim SN_p\ (\boldsymbol{\mu}_d, \Sigma_d, \Sigma_1^{1/2}\boldsymbol{\lambda}_1/A_{11}^{1/2})$.

Figure 1 presents the contour of bivariate closed SN for various combinations of shape parameter parameters D with different scale parameter $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Specifically, the gray contours show $D = 0$ with $\rho = 0, 0.5$ and -0.5; the green contours show $D = \begin{pmatrix} 0 & 0 \\ 3 & 0 \end{pmatrix}$ with $\rho = 0, 0.5$ and -0.5; the blue contours show $D = \begin{pmatrix} 1 & -5 \\ 0 & 0 \end{pmatrix}$ with $\rho = 0, 0.5$ and -0.5 the red contours show $D = \begin{pmatrix} 1 & -5 \\ 3 & 0 \end{pmatrix}$ with $\rho = 0, 0.5$ and -0.5. From another point of view, these contour plots present bivariate normal distribution (gray), SN distribution (green and blue) and closed SN distribution (red).

3. Inference on difference of location parameters

In this section, the inference on the difference of location parameter is proposed when the scale parameter Σ_1 and Σ_2 are unknown but assumed to be equal, say $\Sigma_1 = \Sigma_2 = \Sigma$. The main result is based on the generalized Hotelling's T^2 under multivariate SN setting.

3.1 Some related distributions

At first, we consider the distribution of $S = (\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)' \Sigma_d^{-1} (\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)$. The following definition and lemma by Zhu *et al.* (2019) are useful to derive the distribution of S.

Definition 3.1. Zhu *et al.* (2019) *Let* $\boldsymbol{X} \sim CSN_{p,q}(\boldsymbol{\mu}, I_p, D, \Delta_q)$. *The distribution of* $\boldsymbol{X}'\boldsymbol{X}$, *denoted by* $\boldsymbol{X}'\boldsymbol{X} \sim CS\chi_p^2(\lambda, \delta_1, \delta_2, \Delta_q)$, *is called a* ***noncentral closed skew chi-square distribution*** *with degrees of freedom p, noncentrality parameter* $\lambda = \boldsymbol{\mu}'\boldsymbol{\mu}$, *skewness parameters* $\delta_1 = D\boldsymbol{\mu}$, $\delta_2 = DD'$ *and parameter* Δ_q

Lemma 3.1. Zhu *et al.* (2019) *Let* $\boldsymbol{X} \sim CSN_{p,q}(\boldsymbol{\mu}, \Sigma, D, \Delta)$ *and* $Q = \boldsymbol{X}'W\boldsymbol{X}$ *with a nonnegative definite* $W \in M_{p\times p}$. *If* $\Sigma^{1/2}W\Sigma^{1/2}$ *is idempotent of rank k, then* $Q \sim CS\chi_k^2(\lambda, \delta_1, \delta_2, \Omega_q)$, *where* $\lambda = \boldsymbol{\mu}'W\boldsymbol{\mu}$, $\delta_1 = D\Sigma W\boldsymbol{\mu}$, $\delta_2 = D\Sigma W\Sigma D'$ *and* $\Omega_q = \Delta + D(\Sigma - \Sigma W\Sigma)^{-1}D'$

Based on Theorem 2.1 and Lemma 3.1, we obtain the following result.

Proposition 3.1. *Let* $S = (\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)' \Sigma_d^{-1} (\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)$. *Then,* $S \sim CS\chi_p^2(0, \mathbf{0}, \delta_2, \Omega)$ *with*

$$\delta_2 = \tfrac{n_1 n_2}{n_1+n_2}\begin{pmatrix} \boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1 & \boldsymbol{\lambda}_2'\boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2'\boldsymbol{\lambda}_1 & \boldsymbol{\lambda}_2'\boldsymbol{\lambda}_2 \end{pmatrix} \text{ and } \Omega = \begin{pmatrix} 1 + \dfrac{n_1^2\boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1}{n_1+n_2} & \dfrac{n_1 n_2\boldsymbol{\lambda}_2'\boldsymbol{\lambda}_1}{n_1+n_2} \\ \dfrac{n_1 n_2\boldsymbol{\lambda}_2'\boldsymbol{\lambda}_1}{n_1+n_2} & 1 + \dfrac{n_2^2\boldsymbol{\lambda}_2'\boldsymbol{\lambda}_2}{n_1+n_2} \end{pmatrix}.$$

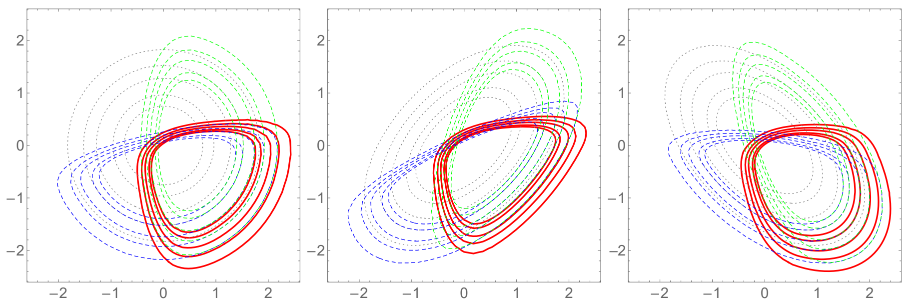

Figure 1.
Contour plot of bivariate normal, SN and close SN distributions with various values of parameters

Proof. From part (i) of Lemma 2.2, we have $\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d \sim CSN_{p,2}(0, \Sigma_d, D_d, \Delta_d)$. Since $\Sigma_d^{1/2}\Sigma_d^{-1}\Sigma_d^{1/2} = I_p$ is an idempotent of rank p, the desired result follows immediately. □

Remark. Comparing with one sample case, the distribution of quantity $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\Sigma^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \chi_p^2$ for $\bar{\boldsymbol{X}} \sim SN_p\left(\boldsymbol{\mu}, \frac{\Sigma}{n}, \sqrt{n}\boldsymbol{\lambda}\right)$ is free of the skewness parameter $\boldsymbol{\lambda}$. Here, the distribution of S follows noncentral closed SN distribution given above which depends on the parameters δ_2 and Ω. Readers are referred to check out Figures 5 and 6 in Zhu *et al.* (2019) for the density curves of $CS\chi^2(0, \mathbf{0}, \delta_2, \Omega)$.

3.2 Confidence region of $\boldsymbol{\mu}_d$

In this subsection, we will extend the Hotelling's T^2 statistic from multivariate normal setting to the multivariate SN setting, called the generalized Hotelling's $T^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)'S_p^{-1}(\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)$, to construct the confidence regions for the difference of location vector where $(n_1 + n_2 - 2)S_p = (n_1 - 1)S_1 + (n_2 - 1)S_2$. First we need to derive the distribution of S_p, then extend the F-distribution to closed skew F-distribution which can describe the distribution of T^2 under the multivariate SN setting.

Proposition 3.2. Let $(n_1 + n_2 - 2)S_p = (n_1 - 1)S_1 + (n_2 - 1)S_2$ with S_1 and S_2 defined by equation (4) Then, $(n_1 + n_2 - 2)S_p \sim W_p(n_1 + n_2 - 2, \Sigma)$

Proof. By Lemma 2.1, $(n_i - 1)S_i \sim W_p(n_i - 1, \Sigma)$ for $i = 1, 2$ are independently distributed. Thus, the well-known properties of Wishart distribution for sums and scale transformation lead to the desired result. □

To obtain the distribution of T^2, we need the following well-known result (Lemma 3.2, Mardia *et al.* (1980), Theorem 3.4.7) and extended version of the F-distribution, called closed skew F-distribution, Definition 3.2, which was introduced by Zhu *et al.* (2019).

Lemma 3.2. If $H \sim W_p(m, \Sigma)$, $m > p$, then the ratio $\boldsymbol{a}'\Sigma^{-1}\boldsymbol{a}/\boldsymbol{a}'H^{-1}\boldsymbol{a} \sim \chi^2_{m-p+1}$ for any fixed p-vector $\boldsymbol{a}$

Definition 3.2. Zhu *et al.* (2019) *Let $U_1 \sim CS\chi^2_{n_1}(\lambda, \delta_1, \delta_2, \Delta_m)$, $U_2 \sim \chi^2_{n_2}$, and U_1 and U_2 are independent. The distribution of $F = \frac{U_1/n_1}{U_2/n_2}$ is called the* ***noncentral closed skew F-distribution*** *with degrees of freedom n_1 and n_2, and parameters λ, δ_1, δ_2 and Δ_m, denoted by $F \sim CSF_{n_1,n_2}(\lambda, \delta_1, \delta_2, \Delta_m)$*

Based on above definition, the pdf of noncentral closed skew F-distribution can be obtained below.

Proposition 3.3. Let $F \sim CSF_{n_1,n_2}(\lambda, \delta_1, \delta_2, \Delta_m)$. The pdf of F is given by

$$f_F(x; \lambda, \delta_1, \delta_2, \Delta_m) = \int_0^\infty \frac{n_1 v}{n_2} f_1\left(\frac{n_1 x v}{n_2}; \lambda, \delta_1, \delta_2, \Delta_m\right) f_2(v)dv \tag{10}$$

where $f_1(\cdot; \lambda, \delta_1, \delta_2, \Delta_m)$ and $f_2(\cdot)$ are pdf of $CS\chi^2_{n_1}(\lambda, \delta_1, \delta_2, \Delta_m)$ and $\chi^2_{n_2}$, respectively.

Proof. Let $F = \frac{U_1/n_1}{U_2/n_2}$ with $U_1 \sim CS\chi^2_{n_1}(\lambda, \delta_1, \delta_2, \Delta_m)$, $U_2 \sim \chi^2_{n_2}$ independently distributed. The joint density of (U_1, U_2) is $f_{(U_1,U_2)}(u, v) = f_1(u)f_2(v)$ where $f_1(\cdot)$ and $f_2(\cdot)$ are pdf of U_1 and U_2, respectively. Then change of variables $x = \frac{n_2 u}{n_1 v}$, $h = v$, for $x > 0$, $h > 0$. The Jacobian of this transformation is $n_1 v/n_2$. So integrating with respect to v over $0 < v < \infty$ leads to desired results.

By Lemma 3.2 and Definition 3.2, we obtain the distribution of T^2 as follows.

Theorem 3.1. *For two independently distributed random matrices*

$$X_i \sim SN_{n_i \times p}\left(\mathbf{1}_{n_i} \otimes \boldsymbol{\mu}_i', I_{n_i} \otimes \Sigma, \mathbf{1}_{n_i} \otimes \boldsymbol{\lambda}_i'\right) \qquad \text{for} \qquad i = 1, 2,$$

where $\boldsymbol{\lambda}_i$'s are known. Let $\bar{\boldsymbol{X}}_i$, S_i be given by equation (3) and (4), $\bar{\boldsymbol{X}}_d = \bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2$, and $S_p = (n_1 - 1)S_1 + (n_2 - 1)S_2$. Then, the distribution of $T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)' S_p^{-1} (\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)$ is given by

$$\frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} T^2 \sim CSF_{p, n_1 + n_2 - p - 1}(0, 0, \delta_2, \Omega), \tag{11}$$

where $\delta_2 = \frac{n_1 n_2}{n_1 + n_2}\begin{pmatrix} \boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1 & \boldsymbol{\lambda}_2'\boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2'\boldsymbol{\lambda}_1 & \boldsymbol{\lambda}_2'\boldsymbol{\lambda}_2 \end{pmatrix}$ and $\Omega = \begin{pmatrix} 1 + \dfrac{n_1^2 \boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1}{n_1 + n_2} & \dfrac{n_1 n_2 \boldsymbol{\lambda}_2'\boldsymbol{\lambda}_1}{n_1 + n_2} \\ \dfrac{n_1 n_2 \boldsymbol{\lambda}_2'\boldsymbol{\lambda}_1}{n_1 + n_2} & 1 + \dfrac{n_2^2 \boldsymbol{\lambda}_2'\boldsymbol{\lambda}_2}{n_1 + n_2} \end{pmatrix}$.

Proof. Rewrite T^2 as

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \frac{(\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)' \Sigma^{-1} (\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)}{(\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)' \Sigma^{-1} (\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d) \Big/ (\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)' S_p^{-1} (\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)}.$$

Since S_p and $\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d$ are independent by Lemma 2.1, the conditional distribution of

$$w = (n_1 + n_2 - 2)(\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)' \Sigma^{-1} (\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d) \Big/ (\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)' S_p^{-1} (\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d) \sim \chi^2_{n_1 + n_2 - p - 1}$$

given $\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d$ by Proposition 3.2 and Lemma 3.2. It is clear that $w \sim \chi^2_{n_1 + n_2 - p - 1}$ since the conditional distribution does not depend on $\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d$. On the other hand, since $\Sigma_d^{-1} = \frac{n_1 n_2}{n_1 + n_2}\Sigma^{-1}$, then $\frac{n_1 n_2}{n_1 + n_2}(\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)' \Sigma^{-1} (\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d) = (\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)' \Sigma_d^{-1} (\bar{\boldsymbol{X}}_d - \boldsymbol{\mu}_d)$ which follows closed skew chi-square distribution $CS\chi^2_p(0, \mathbf{0}, \delta_2, \Omega)$ from Proposition 3.1. Therefore, the desired result follows by Definition 3.2.

Based on above results, we construct confidence regions for the difference of location parameter $\boldsymbol{\mu}_d$ by using generalized Hotelling's T^2 as a pivotal statistics.

Theorem 3.2. *Assume two samples satisfying* (2) *with unknown Σ_1 and Σ_2 but $\Sigma_1 = \Sigma_2$ and known $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$. Then, the $100(1-\alpha)\%$ confidence regions for $\boldsymbol{\mu}_d$ is given by*

$$C^P_{\boldsymbol{\mu}_d}(\alpha) = \left\{ \boldsymbol{\mu}_d : \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} T^2 < CSF^2_{p, n_1 + n_2 - p - 1}(1 - \alpha) \right\},$$

where $CSF^2_{p, n_1 + n_2 - p - 1}(1 - \alpha)$ represents the $1 - \alpha$ quantile of $CSF^2_{p, n_1 + n_2 - p - 1}(0, \mathbf{0}, \delta_2, \Omega)$

The following plots present the pdf of noncentral closed skew F-distribution (see Figure 2).

4. Simulation study

Simulations are conducted for evaluating the performance of the proposed confidence regions for $\boldsymbol{\mu}_d$ under independent multivariate SN settings using the coverage relative frequency rates. Comparisons of proposed confidence regions with those in classical independent multivariate normal distributions are given.

4.1 Coverage frequencies

To evaluate the proposed confidence regions for difference of location parameters under multivariate SN setting, Monte Carlo simulation studies (each with a number of simulation runs $M = 10{,}000$) are conducted for combinations of various values of sample sizes $(n_1, n_2) = (20,25)$, $(40, 50)$ and $(80,100)$, $(\rho_1, \rho_2) = (\{0.1, 0.5, 0.8\}, \{0.1, 0.5, 0.8\})$, $D_1 = \begin{pmatrix} 1 & 3 \\ 2 & -5 \end{pmatrix}$ and $D_2 = \begin{pmatrix} -2 & 2 \\ 3 & 0 \end{pmatrix}$.

Table 1 shows that our method provides reliable inference about the difference of location parameters with nominal confidence level (95%). To further illustrate the effectiveness of the proposed method, we confidence intervals of the coverage probability presented in the following plot.

From Figure 3, we can see clearly that the pivotal quantity-based closed skew *F*-distribution produce more robust confidence region than that based on *F*-distribution. The coverage relative frequencies, based on the SN model, are close to the nominal confidence level 95% consistently for the combination of different sample sizes, scale parameter and

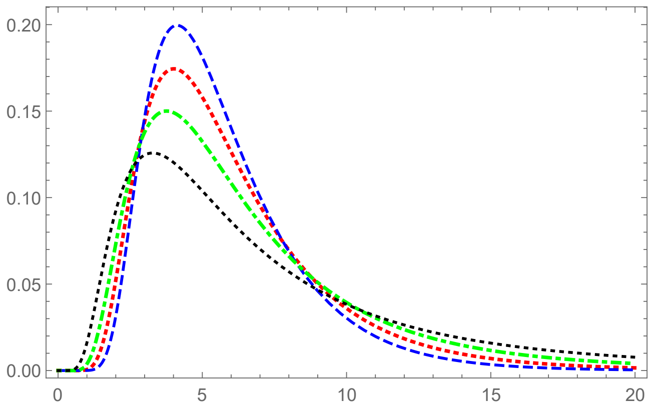

Figure 2. The pdfs of noncentral closed skew $F_{2,n}\left(5, (5,11)^T, \begin{pmatrix} 5 & 11 \\ 11 & 25 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$ distributions with $n = 5, 10, 20$ and 50 (black, green, red and blue)

(n_1, n_2)	$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$	$D_1 = \begin{pmatrix} 1 & 3 \\ 2 & -5 \end{pmatrix}$		$D_2 = \begin{pmatrix} -2 & 2 \\ 3 & 0 \end{pmatrix}$	
		F-distribution	*CSF*-distribution	*F*-distribution	*CSF*-distribution
(20, 25)	$\rho = 0.2$	0.9268	0.9495	0.9224	0.9524
	$\rho = 0.5$	0.9274	0.9511	0.9135	0.9502
	$\rho = 0.8$	0.9206	0.9504	0.9227	0.9520
(40, 45)	$\rho = 0.2$	0.9025	0.9479	0.9008	0.9534
	$\rho = 0.5$	0.9118	0.9451	0.9095	0.9509
	$\rho = 0.8$	0.9036	0.9521	0.9016	0.9524
(80, 100)	$\rho = 0.2$	0.8932	0.9527	0.9080	0.9479
	$\rho = 0.5$	0.8917	0.9463	0.8702	0.9470
	$\rho = 0.8$	0.9035	0.9483	0.8933	0.9448
(200, 250)	$\rho = 0.2$	0.8824	0.9503	0.8883	0.9461
	$\rho = 0.5$	0.88920	0.9533	0.8956	0.9559
	$\rho = 0.8$	0.8998	0.9492	0.8852	0.9546

Table 1. Coverage relative frequencies of confidence regions at confidence level $\alpha = 95\%$ for difference of location parameters $\boldsymbol{\mu}_d$ with various combinations of sample sizes, ρ_1, ρ_2 and D_1, D_2 using Hotelling's T^2 as the pivotal quantity when Σ_1 and Σ_2 are equal but unknown

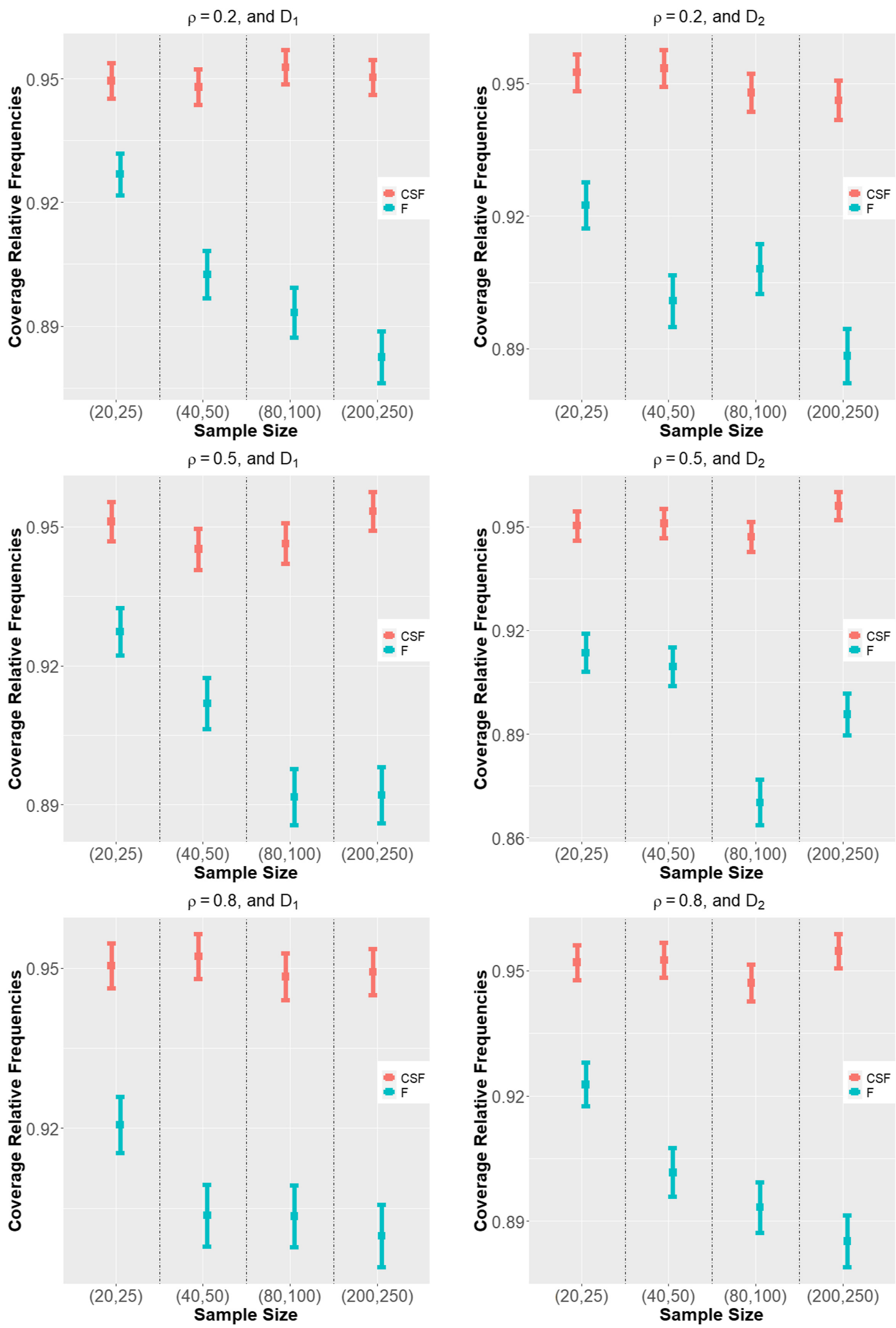

Figure 3.
The confidence intervals of coverage relative frequencies at confidence level $\alpha = 0.95$, red ones based on SN model and blue ones based on normal model with sample size $(n_1, n_2) =$ (20, 25), (40, 50), (80, 100) and (200, 250) (from left to right in each figure), $\rho = 0.2, 0.5$ and 0.8 (in each row), and D_1 and D_2 (from left to right), respectively

skewness parameters. But the coverage relative frequencies, based on the normal model, are lower than the nominal confidence level.

5. Real data example

In this section, we illustrate the effectiveness and applicability of the proposed methods by applying them to Australian Institute of Sport (AIS) data (Cook and Weisberg, 2009). We explore the difference of body mass index (BMI) and lean body mass (LMB) between males and females athletes in AIS data.

The point estimates of the parameters for AIS data are reported in Table 2.

In Figure 4, the scatter plots and contour plots of fitted bivariate SN distributions are presented. Based on our previous work (Azzalini and Valle, 1996), this data set prefers multivariate SN model. So we adopt multivariate SN model as well to explore the difference of location parameters. Using point estimates listed in Table 2 and applying Theorem 2.1, the differences of sample mean has closed SN distribution, $\bar{\boldsymbol{X}}_d \sim CSN_{2,2}(\boldsymbol{\mu}, \Sigma, D, \Delta)$ with estimated parameters

$$\widehat{\boldsymbol{\mu}} = (1.29, 18.13)', \qquad \widehat{\Sigma} = \begin{pmatrix} 0.13 & 0.41 \\ 0.41 & 2.09 \end{pmatrix}$$

$$\widehat{D} = \begin{pmatrix} 3.20 & -4.83 \\ -29.80 & 16.82 \end{pmatrix}, \qquad \widehat{\Delta} = \begin{pmatrix} 28.90 & 101.01 \\ 101.01 & 473.42 \end{pmatrix}.$$

Then, we apply Theorem 3.2 to construct the confidence region of difference of location parameters $\boldsymbol{\mu}_d$. In Figure 5, the confidence regions for the difference of location parameters are given below at 95% confidence level.

	Males	Females
$\widehat{\boldsymbol{\mu}}'$	$(24.47, 79.55)$	$(23.18, 61.42)$
$\widehat{\Sigma}$	$\begin{pmatrix} 4.81 & 18.95 \\ 18.95 & 114.00 \end{pmatrix}$	$\begin{pmatrix} 8.32 & 21.30 \\ 21.30 & 89.96 \end{pmatrix}$
$\widehat{\boldsymbol{\lambda}}'$	$(-0.20, -0.80)$	$(0.88, -2.63)$

Table 2. Point estimates of SN parameters for the males and females AIS data, respectively

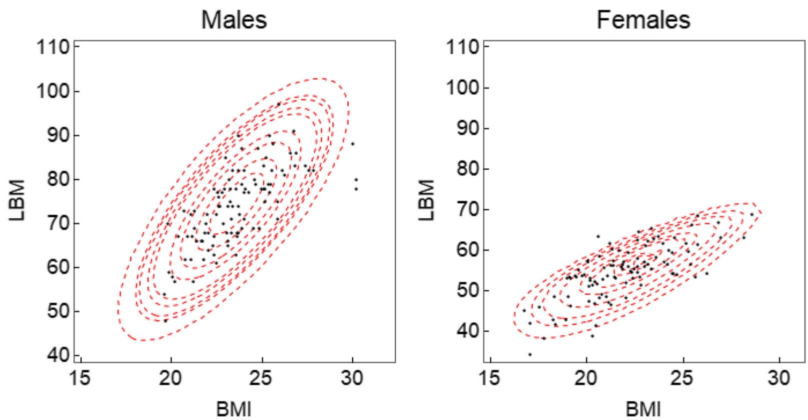

Figure 4. The scatter plots and contour plots for the AIS data of males and females

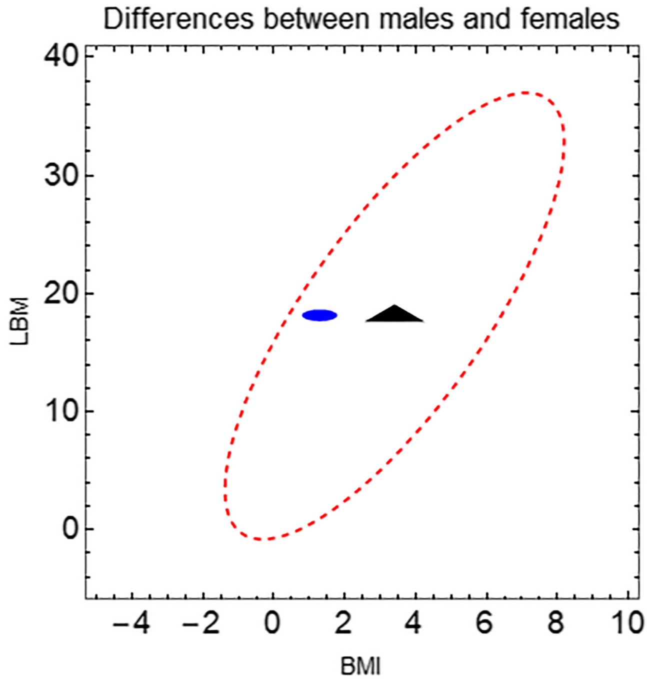

Figure 5.
The confidence region of difference of location parameter $\boldsymbol{\mu}_d$ at confidence level 95% for AIS data between males and females, where the black triangle represents the $\widehat{\boldsymbol{\mu}}_d$, and the blue dot represents the $\boldsymbol{\mu}_d^0$

6. Conclusion

In this work, the difference for location parameters between two independent samples under multivariate SN setting is studied. The construction of confidence region procedure is developed. From the results of simulation studies, the confidence region based on SN model has better performance than normal model in term of relative coverage frequencies to capture the true value when the data are generated from skewed distribution.

References

Adcock, C. and Azzalini, A. (2020), "A selective overview of skew-elliptical and related distributions and of their applications", *Symmetry*, Vol. 12 No. 1, p. 118.

Arellano-Valle, R., Bolfarine, H. and Lachos, V. (2005), "Skew-normal linear mixed models", *Journal of Data Science*, Vol. 3 No. 4, pp. 415-438.

Azzalini, A. (1985), "A class of distributions which included the normal ones", *Scandinavian Journal of Statistics*, Vol. 12 No. 2, pp. 171-178.

Azzalini, A. and Capitanio, A. (1999), "Statistical applications of the multivariate skew normal distribution", *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. 61 No. 3, pp. 579-602.

Azzalini, A. and Valle, A.D. (1996), "The multivariate skew-normal distribution", *Biometrika*, Vol. 83 No. 4, pp. 715-726.

Carmichael, B. and Coën, A. (2013), "Asset pricing with skewed-normal return", *Finance Research Letters*, Vol. 10 No. 2, pp. 50-57.

Chen, J.T. and Gupta, A.K. (2005), "Matrix variate skew normal distributions", *Statistics*, Vol. 39 No. 3, pp. 247-253.

Cook, R.D. and Weisberg, S. (2009), *An Introduction to Regression Graphics*, John Wiley & Sons, New York, Vol. 405.

Gonzalez-Farias, G., Dominguez-Molina, A. and Gupta, A.K. (2004), "Additive properties of skew normal random vectors", *Journal of Statistical Planning and Inference*, Vol. 126 No. 2, pp. 521-534.

Li, B., Tian, W. and Wang, T. (2018), "Remarks for the singular multivariate skew-normal distribution and its quadratic forms", *Statistics and Probability Letters*, Vol. 137, pp. 105-112.

Ma, Z., Chen, Y.-J., Wang, T. and Peng, W. (2019), "The inference on the location parameters under multivariate skew normal settings", in Kreinovich, V., Trung, N. and Thach, N. (Eds), *Beyond Traditional Probabilistic Methods in Economics*, Springer Nature, pp. 146-162.

Ma, Z., Zhu, X., Wang, T. and Autchariyapanitkul, K. (2018), "Joint plausibility regions for parameters of skew normal family", in Krennovich, V., Sriboonchitta, S. and Chakpitak, N. (Eds), *Predictive Econometrics and Big Data*, Springer-Verlag, New York, pp. 233-245.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1980), *Multivariate Analysis (Probability and Mathematical Statistics)*.

Wang, T., Li, B. and Gupta, A.K. (2009), "Distribution of quadratic forms under skew normal settings", *Journal of Multivariate Analysis*, Vol. 100 No. 3, pp. 533-545.

Wei, Z., Zhu, X. and Wang, T. (2021), "The extended skew-normal-based stochastic frontier model with a solution to 'wrong skewness' problem", *Statistics*, pp. 1-20, doi: 10.1080/02331888.2021.2004142.

Ye, R., Wang, T. and Gupta, A.K. (2014), "Distribution of matrix quadratic forms under skew-normal settings", *Journal of Multivariate Analysis*, Vol. 131, pp. 229-239, 00010.

Young, P.D., Harvill, J.L. and Young, D.M. (2016), "A derivation of the multivariate singular skew-normal density function", *Statistics and Probability Letters*, Vol. 117, pp. 40-45.

Zhu, X., Li, B., Wu, M. and Wang, T. (2018), "Plausibility regions on parameters of the skew normal distribution based on inferential models", in Krennovich, V., Sriboonchitta, S. and Chakpitak, N. (Eds), *Predictive Econometrics and Big Data*, Springer-Verlag, New York, pp. 287-302.

Zhu, X., Li, B., Wang, T. and Gupta, A.K. (2019), "Sampling distributions of skew normal populations associated with closed skew normal distributions", *Random Operators and Stochastic Equations*, Vol. 27 No. 2, pp. 75-87.

Zhu, X., Ma, Z., Wang, T. and Teetranont, T. (2017), "Plausibility regions on the skewness parameter of skew normal distributions based on inferential models", in Krennovich, V., Sriboonchitta, S. and Huynh, V. (Eds), *Robustness in Econometrics*, Springer-Verlag, New York, pp. 267-286.

Corresponding author
Tonghui Wang can be contacted at: twang@nmsu.edu