Statistical Papers (2024) 65:511–555
https://doi.org/10.1007/s00362-023-01397-1

REGULAR ARTICLE

A multivariate modified skew-normal distribution

Sagnik Mondal[1] · **Reinaldo B. Arellano-Valle**[2] · **Marc G. Genton**[1]

Received: 10 December 2021 / Revised: 31 October 2022 / Published online: 13 February 2023

Abstract
We introduce a multivariate version of the modified skew-normal distribution, which contains the multivariate normal distribution as a special case. Unlike the Azzalini multivariate skew-normal distribution, this new distribution has a nonsingular Fisher information matrix when the skewness parameters are all zero, and its profile log-likelihood of the skewness parameters is always a non-monotonic function. We study some basic properties of the proposed family of distributions and present an expectation-maximization (EM) algorithm for parameter estimation that we validate through simulation studies. Finally, we apply the proposed model to the univariate frontier data and to a trivariate wind speed data, and compare its performance with the Azzalini skew-normal model.

Keywords EM algorithm · Fisher information matrix · Modified skew-normal · Skew-elliptical · Skew-generalized normal · Skew-normal

1 Introduction

Because the assumption of the Gaussian model is not appropriate in many statistical applications, despite its appealing stochastic properties, more options for parametric models are needed, especially for multivariate cases. For this reason, a large number of parametric models have been introduced that are more flexible compared to the

This research was supported by King Abdullah University of Science and Technology (KAUST).

✉ Sagnik Mondal
sagnik.mondal@kaust.edu.sa

Reinaldo B. Arellano-Valle
reivalle@mat.uc.cl

Marc G. Genton
marc.genton@kaust.edu.sa

[1] Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

[2] Department of Statistics, Pontificia Universidad Católica de Chile, Santiago, Chile

Gaussian model. Among them, an important and popular one is the skew-normal distribution introduced by Azzalini (1985), which instigated the development of other skew-normal and skew-elliptical models. For more insight on such univariate and multivariate flexible parametric models and their applications, readers are referred to the books by Azzalini and Capitanio (2014) and Genton (2004).

Since the Azzalini skew-normal ($\mathcal{ASN}$) distribution integrates skewness into the model with only one additional parameter, compared to the Gaussian model, and due to its simple interpretation, it has been popular in applications. In the univariate case, a random variable Z follows an $\mathcal{ASN}$ distribution with skewness parameter $\lambda \in \mathbb{R}$ if it has a probability density function (pdf)

$$f_Z(z) = 2\phi(z)\Phi(\lambda z), \quad z \in \mathbb{R}, \tag{1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and the distribution function, respectively, of the standard Gaussian distribution. The parameter λ induces skewness in the distribution. The sign of λ dictates the sign of the skewness, the magnitude of λ dictates the extent of the skewness in the distribution and it reduces to the standard Gaussian distribution when λ becomes zero. The location-scale $\mathcal{ASN}$ model is described by the distribution of $X = \xi + \omega Z$, where Z has the pdf as given in (1), which we denote by $X \sim \mathcal{ASN}(\xi, \omega, \lambda)$, with location parameter $\xi \in \mathbb{R}$ and scale parameter $\omega > 0$. The $\mathcal{ASN}$ distribution has been studied extensively. Various theoretical properties of the $\mathcal{ASN}$ distribution have been derived in Henze (1986), Pewsey (2000), Azzalini (2005), and Gómez et al. (2007), among many more. The $\mathcal{ASN}$ distribution was firstly extended to the multivariate case in Azzalini and Dalla-Valle (1996) and posteriorly studied systematically in Azzalini and Capitanio (1999) and in Azzalini and Capitanio (2003). In the context of applications, the $\mathcal{ASN}$ distribution and its multivariate extension have been used in plenty of situations. For example, Ghosh et al. (2007) used the multivariate $\mathcal{ASN}$ distribution as the joint distribution of the random effects in the random effect model, Lin et al. (2007) applied the $\mathcal{ASN}$ distribution in finite mixture models, and Lachos et al. (2010) considered the $\mathcal{ASN}$ distribution for linear mixed models.

The $\mathcal{ASN}$ distribution can capture both symmetry and asymmetry in the data adequately in many applications, demonstrating its flexibility, and it performs well for moderately skewed datasets. Despite its good properties, two main inferential drawbacks of the location-scale $\mathcal{ASN}$ model are: 1) its Fisher information matrix is singular when the skewness parameter λ is set to be zero, and as a consequence, one cannot use the asymptotic normality of the maximum likelihood estimators to test $H_0 : \lambda = 0$ versus $H_1 : \lambda \neq 0$ with a Wald-type test, which would have been useful for checking whether the data were symmetric or not; 2) the maximum likelihood estimate of the skewness parameter λ can sometimes be infinite, if the data are heavily skewed, and the inference regarding the skewness parameter will then be difficult. Indeed, in that case, it is a boundary point of the parameter space of the skewness parameter, hence it will not satisfy the assumptions of the standard asymptotic distribution theory of the maximum likelihood estimator. Then, the problem of testing $H_0 : \lambda = 0$ versus $H_1 : \lambda \neq 0$ requires a special methodology described by Rotnitzky et al. (2000) and by Chiogna (2005). Although, in the univariate $\mathcal{ASN}$ case, a likelihood ratio test (LRT) can be used to test $H_0 : \lambda = 0$ versus $H_1 : \lambda \neq 0$, this does not extend to

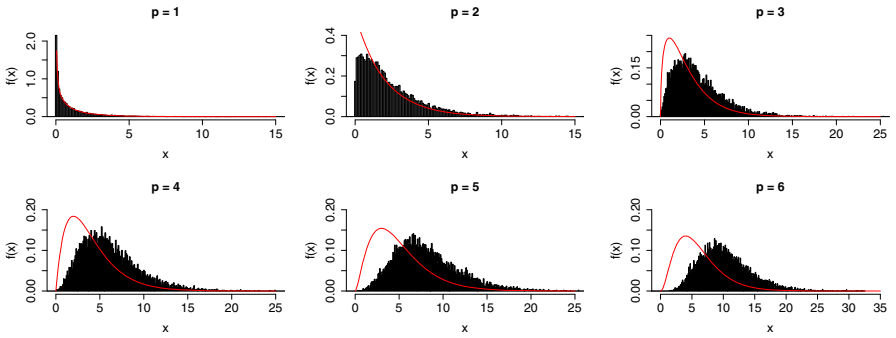


Fig. 1 The histograms of the LRT statistic under the null hypothesis $H_0 : \mathcal{ASN}_p(\mathbf{0}, \mathbf{I}_p, \boldsymbol{\lambda} = \mathbf{0})$ for different values of $p = 1, \ldots, 6$, along with the density of χ^2_p

the p-dimensional multivariate $\mathcal{ASN}$ case. Indeed, one can check by simulations that the usual χ^2_p asymptotic distribution of the LRT test statistic under H_0 is not valid. In Fig. 1, we have plotted the histogram of the LRT statistic based on samples of size 1000 from the null distribution with 10, 000 replicates for different values of p. Along with the histogram, we have plotted the corresponding χ^2_p probability density. Figure 1 shows that the density only matches with the histogram at $p = 1$. The distribution of the LRT statistic for $p > 1$ is still an open problem for this case. One other approach to handle the testing problem is to use the centralized parameterization of the $\mathcal{ASN}$ distribution. The univariate centralized $\mathcal{ASN}$ was introduced by Azzalini and Capitanio (1999) and was later extended to the multivariate regime by Arellano-Valle and Azzalini (2008). The idea of centralized parameterization is to have the mean vector as the location parameter, the covariance matrix as the scale parameter, and the vector consisting of the component-wise skewness as the skewness parameter. Moreover, the transformation from direct to centralized parameterization is one-to-one. However, in the multivariate case, optimizing the log-likelihood of the centralized $\mathcal{ASN}$ is more difficult compared to the direct parameterization of the $\mathcal{ASN}$, because the centralized parameters are interdependent. Due to that, the feasibility of a set of central parameters has to be checked by testing the positive definiteness of the scale matrix of the direct parameterization. In addition, the centralized parameterization does not guarantee the existence of the MLE of the skewness parameter.

To address the problem of infinite maximum likelihood estimator of λ, Sartori (2006) used a modified likelihood function for the $\mathcal{ASN}$ family that always results in a finite estimate for λ. Azzalini and Arellano-Valle (2013) formulated a proposal based on the idea of a penalized likelihood for finding the maximum likelihood estimates, which ensures the finiteness and the consistency of the skewness parameter estimate. Similar work has been done by Jin et al. (2016) under the theme of the finite mixture model of the $\mathcal{ASN}$ distribution. Both these inferential problems can be handled if a Bayesian $\mathcal{ASN}$ model is considered, as shown by Bayes and Branco (2007).

Arrué et al. (2016) addressed the aforementioned problem of singularity by introducing the modified skew-generalized normal ($\mathcal{MSGN}$) distribution. This distribution is obtained from a reparametrization of the so-called skew-generalized-normal

($\mathcal{SGN}$) distribution introduced by Arellano-Valle et al. (2004). The density function of a random variable Z following an $\mathcal{MSGN}$ distribution, with skewness parameter λ and shape parameter ν, is given by

$$f_Z(z) = 2\phi(z)\Phi\left(\frac{\lambda z}{\sqrt{\nu + z^2}}\right), \quad z \in \mathbb{R},$$

with $\lambda \in \mathbb{R}$ and $\nu > 0$, and we denote $Z \sim \mathcal{MSGN}(\lambda, \nu)$. The $\mathcal{MSGN}(\lambda, \nu)$ density reduces to the $\mathcal{N}(0, 1)$ density when $\lambda = 0$ or $\nu \to \infty$. The usual location-scale $\mathcal{MSGN}$ model possesses a nonsingular Fisher information matrix, when $\lambda = 0$ for any $\nu > 0$. To be more comparable to the $\mathcal{ASN}$ family, Arrué et al. (2016) suggested using the $\mathcal{MSGN}(\lambda, 1)$ distribution and named it the modified skew-normal ($\mathcal{MSN}$) distribution. The univariate $\mathcal{MSN}$ distribution has been used in a few applications in statistics. For instance, the $\mathcal{MSN}$ distribution has been applied by Arellano-Valle et al. (2017) to measure the divergence of a density from Gaussianity, and Arrué et al. (2020) considered the $\mathcal{MSN}$ distribution for suggesting a new type of Birnbaum–Saunders model.

Although the $\mathcal{MSN}$ distribution provides an alternative option to model skewed data besides the $\mathcal{ASN}$ distribution, with the additional advantage over the $\mathcal{ASN}$ distribution of non-singularity of the Fisher information matrix when the skewness parameter λ is zero, the maximum likelihood estimator of λ can still be infinite in some situations. This issue has been discussed by Arrué et al. (2016) in detail and was dealt with by considering a modified likelihood function, similar to Sartori (2006). However, all the issues discussed above have been studied in the univariate case only.

In this paper, we provide an alternative parameterization of the $\mathcal{MSN}$ distribution and extend it to the multivariate case. We call it the multivariate modified skew-normal distribution. It has a nonsingular Fisher information matrix when the skewness parameter is set to zero, and the maximum likelihood estimator of the skewness parameter is always finite.

The rest of the paper is as follows: In Sect. 2, we discuss the multivariate $\mathcal{ASN}$ family and introduce an alternative parameterization for it. Then, we formally define the multivariate modified skew-normal distribution. In Sect. 3, we study some of its basic properties, its likelihood function, and its Fisher information matrix. In Sect. 4, we derive an EM algorithm for the parameter estimation of the new family. We provide some simulation studies in Sect. 5 and data applications in Sect. 6. Finally, we conclude the paper by discussing some future research topics in Sect. 7.

2 Multivariate modified skew-normal distribution

In this section, we formally define the multivariate modified skew-normal distribution based on the multivariate skew-normal distribution with alternative parameterization. Some of the basic properties of the latter are in Appendix A. This parameterization is used by Azzalini and Dalla-Valle (1996), up to minor differences. The same parameterization has also been used by Adcock and Shutes (2001), Adcock (2004), and Adcock

(2005). We first revisit the family of multivariate $\mathcal{ASN}$ distributions and investigate the effect of its alternative parameterization.

2.1 The multivariate skew-normal with alternative parameterization

The multivariate $\mathcal{ASN}$ distribution introduced by Azzalini and Dalla-Valle (1996) was defined posteriorly by Azzalini and Capitanio (2003) as follows. A random vector $\boldsymbol{X} \in \mathbb{R}^p$ is said to have a multivariate $\mathcal{ASN}$ distribution with location parameter $\boldsymbol{\xi} \in \mathbb{R}^p$, symmetric positive definite scale parameter $\boldsymbol{\Omega} \in \mathbb{R}^{p\times p}$, and skewness parameter $\boldsymbol{\alpha} \in \mathbb{R}^p$, if its multivariate probability density function (mpdf) is given by

$$f_X(\boldsymbol{x}) = 2\phi_p(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Omega})\, \Phi\{\boldsymbol{\alpha}^\top \boldsymbol{\omega}^{-1}(\boldsymbol{x} - \boldsymbol{\xi})\}, \quad \boldsymbol{x} \in \mathbb{R}^p, \tag{2}$$

where $\phi_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the mpdf of a p-dimensional normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, and $\boldsymbol{\omega} = \text{diag}(\boldsymbol{\Omega})^{1/2}$. We denote this by $\boldsymbol{X} \sim \mathcal{ASN}_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha})$.

The $\mathcal{ASN}_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha})$ distribution can be reparameterized by using the relations $\boldsymbol{\Omega} = \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top$ and $\boldsymbol{\alpha} = (1 + \boldsymbol{\eta}^\top \boldsymbol{\Psi}^{-1}\boldsymbol{\eta})^{-1/2}\, \boldsymbol{\omega}\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}$, where $\boldsymbol{\Psi} \in \mathbb{R}^{p\times p}$ is a symmetric positive definite matrix, $\boldsymbol{\eta} \in \mathbb{R}^p$ and $\boldsymbol{\omega} = \text{diag}\left(\sqrt{\psi_{11} + \eta_1^2}, \ldots, \sqrt{\psi_{pp} + \eta_p^2}\right)$, with Ψ_{ii} and η_i being the ith diagonal element of $\boldsymbol{\Psi}$ and $\boldsymbol{\eta}$, respectively, for $i = 1, \ldots, p$. Conversely, by letting $\boldsymbol{\omega} = \text{diag}(\boldsymbol{\Omega})^{1/2}$, $\bar{\boldsymbol{\Omega}} = \boldsymbol{\omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\omega}^{-1}$ and $\boldsymbol{\delta} = (1 + \boldsymbol{\alpha}^\top \bar{\boldsymbol{\Omega}}\boldsymbol{\alpha})^{-1/2}\bar{\boldsymbol{\Omega}}\boldsymbol{\alpha}$, we have $\boldsymbol{\Psi} = \boldsymbol{\omega}(\bar{\boldsymbol{\Omega}}^{-1} + \boldsymbol{\alpha}\boldsymbol{\alpha}^\top)^{-1}\boldsymbol{\omega} = \boldsymbol{\omega}(\bar{\boldsymbol{\Omega}} - \boldsymbol{\delta}\boldsymbol{\delta}^\top)\boldsymbol{\omega}$ and $\boldsymbol{\eta} = \boldsymbol{\omega}\boldsymbol{\delta}$. With this alternative parameterization, the mpdf of $\boldsymbol{X}$ from Eq. (2) is

$$f_X(\boldsymbol{x}) = 2\phi_p\left(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top\right) \Phi\left\{\frac{\boldsymbol{\eta}^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{x} - \boldsymbol{\xi})}{\sqrt{1 + \boldsymbol{\eta}^\top \boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}}\right\}, \quad \boldsymbol{x} \in \mathbb{R}^p. \tag{3}$$

Definition 1 (*Multivariate skew-normal* ($\mathcal{SN}$) *distribution*) A random vector $\boldsymbol{X} \in \mathbb{R}^p$ with mpdf given by Eq. (3) is said to have a multivariate skew-normal distribution with location parameter $\boldsymbol{\xi} \in \mathbb{R}^p$, scale parameter $\boldsymbol{\Psi} \in \mathbb{R}^{p\times p}$, a symmetric and positive definite matrix, and skewness parameter $\boldsymbol{\eta} \in \mathbb{R}^p$, and denoted by $\boldsymbol{X} \sim \mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$.

We discuss the benefits of using this alternative parameterization for defining the $\mathcal{SN}$ distribution compared to the $\mathcal{ASN}$ distribution in Appendix A.

2.2 The multivariate modified skew-normal distribution

The multivariate $\mathcal{SN}$ distribution is a reformulation of the multivariate $\mathcal{ASN}$ distribution by considering an alternative parameterization, which solves the problem of the infinite maximum likelihood estimator of the skewness parameter. However, as in the multivariate $\mathcal{ASN}$ case, the Fisher information matrix in the multivariate $\mathcal{SN}$ model at symmetry still remains singular. To have a modified version of the multivariate $\mathcal{ASN}$ model without the two aforementioned peculiarities, we next introduce a multivariate version of the $\mathcal{MSN}$ distribution in terms of the same parameterization of the multivariate $\mathcal{SN}$ distribution.

Definition 2 (*Multivariate modified skew-normal* ($\mathcal{MSN}$) *distribution*) A random vector $\boldsymbol{X} \in \mathbb{R}^p$ is said to have a multivariate modified skew-normal distribution with location parameter $\boldsymbol{\xi} \in \mathbb{R}^p$, scale parameter $\boldsymbol{\Psi} \in \mathbb{R}^{p\times p}$, a symmetric and positive definite matrix, and skewness parameter $\boldsymbol{\eta} \in \mathbb{R}^p$ if its mpdf is given by

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = 2\phi_p\left(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top\right) \times \Phi\left\{\frac{1}{\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}} \frac{\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}{\sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^\top(\boldsymbol{\Psi}+\boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}}\right\}, \quad \boldsymbol{x} \in \mathbb{R}^p, \quad (4)$$

which we denote by $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$.

To have some insight into how we arrive at this new multivariate distribution, let us consider a p-variate random vector $\boldsymbol{Z} \in \mathbb{R}^p$ with mpdf

$$f_{\boldsymbol{Z}}(z) = 2\phi_p(z; \boldsymbol{0}, \bar{\boldsymbol{\Omega}})\Phi\left(\frac{\boldsymbol{\alpha}^\top z}{\sqrt{1+z^\top\bar{\boldsymbol{\Omega}}^{-1}z}}\right), \quad z \in \mathbb{R}^p, \quad (5)$$

where $\boldsymbol{\alpha} \in \mathbb{R}^p$ is the skewness parameter and $\bar{\boldsymbol{\Omega}} \in \mathbb{R}^{p\times p}$, a positive definite correlation matrix. Equation (5) is a simple and natural multivariate extension of the pdf of the univariate $\mathcal{MSN}$ distribution considered by Arrué et al. (2016), and it becomes effectively an mpdf. In fact, let $w(z) = (1+z^\top\bar{\boldsymbol{\Omega}}^{-1}z)^{-1/2}\boldsymbol{\alpha}^\top z$. Since $w(-z) = -w(z)$, $\phi_p(\cdot)$ is a valid mpdf and $\Phi(\cdot)$ is a valid cdf, then by Proposition 1.1 from Azzalini and Capitanio (2014), it can be easily established that the function $f_{\boldsymbol{Z}}(z)$ given in Eq. (5) is indeed an mpdf. Now define the location-scale random vector $\boldsymbol{X} = \boldsymbol{\xi} + \boldsymbol{\omega}\boldsymbol{Z}$, where $\boldsymbol{\xi} \in \mathbb{R}^p$, $\boldsymbol{\omega} = \text{diag}(\omega_1, \ldots, \omega_p)$ such that $\omega_i > 0$ for all $i = 1, \ldots, p$, and $\boldsymbol{\Omega} = \boldsymbol{\omega}\bar{\boldsymbol{\Omega}}\boldsymbol{\omega}$. Finally, as in the definition of the multivariate $\mathcal{SN}$ distribution, we again adopt the parameterization $\boldsymbol{\Omega} = \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top$, $\boldsymbol{\omega} = \text{diag}(\boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)^{1/2}$, and $\boldsymbol{\alpha} = (1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta})^{-1/2}\boldsymbol{\omega}\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}$. Thus, the random vector $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$.

3 Properties of the multivariate $\mathcal{MSN}$ distribution

3.1 Basic properties

We next list various properties of the multivariate $\mathcal{MSN}$ distribution. All proofs can be found in Appendix B.

1) $\mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{0})$ is equivalent to $\mathcal{N}_p(\boldsymbol{\xi}, \boldsymbol{\Psi})$: This fact is easily verifiable from Eq. (4) by setting $\boldsymbol{\eta} = \boldsymbol{0}$. When $\boldsymbol{\eta}$ is the zero vector, the mpdf of the multivariate $\mathcal{MSN}$ distribution is elliptically contoured, as it is the mpdf of the multivariate normal distribution, and when $\boldsymbol{\eta}$ is a non-zero vector, the contours of the mpdf of the distribution are shifted towards $\boldsymbol{\eta}$ and lose the elliptical shape. The effect of the skewness parameter $\boldsymbol{\eta}$ on the $\mathcal{MSN}$ distribution is illustrated in the contour plots of two bivariate $\mathcal{MSN}$ distributions with the same location and scale parameters but with different skewness parameters $\boldsymbol{\eta} = (0, 0)^\top$ and $\boldsymbol{\eta} = (10, -5)^\top$ in Fig. 2.

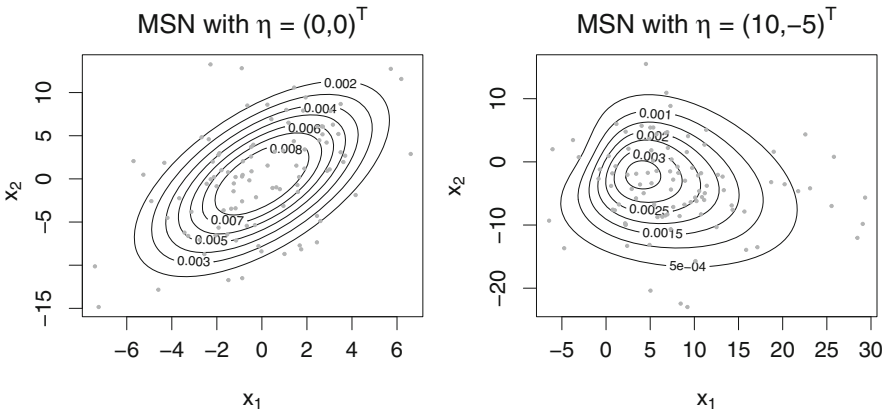


Fig. 2 Contour plot of the mpdf of the $\mathcal{MSN}_2$ distribution with $\boldsymbol{\xi} = (0, 0)^\top$, $\boldsymbol{\Psi} = \begin{pmatrix} 10 & 12 \\ 12 & 40 \end{pmatrix}$ and $\boldsymbol{\eta} = (0, 0)^\top$ and with $\boldsymbol{\eta} = (10, -5)^\top$ with 100 simulated observations from the respective distributions

2) Stochastic representation of the multivariate $\mathcal{MSN}$ distribution: Now we present a stochastic representation of the $\mathcal{MSN}$ distribution. As we show in the later part, this stochastic representation helps us obtain the results regarding the hierarchical representation and the distribution of affine transformations of a $\mathcal{MSN}$ random vector, among other properties of this family. To simplify the notation in the following results, we use the working parametrization $(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$, where as defined earlier

$$\boldsymbol{\Omega} = \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top \quad \text{and} \quad \boldsymbol{\lambda} = \frac{1}{\sqrt{1 - \boldsymbol{\eta}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{\eta}}} \boldsymbol{\Omega}^{-1} \boldsymbol{\eta} = \frac{1}{\sqrt{1 + \boldsymbol{\eta}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{\eta}}} \boldsymbol{\Psi}^{-1} \boldsymbol{\eta}. \quad (6)$$

Proposition 1 *Let* $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$. *Then*

$$\boldsymbol{X} \stackrel{d}{=} \boldsymbol{\xi} + \frac{1}{\sqrt{1 + \boldsymbol{Y}^\top \boldsymbol{\Omega} \boldsymbol{Y}}} \boldsymbol{\Omega} \boldsymbol{Y}\, T + \boldsymbol{U},$$

where $T \sim \mathcal{HN}(0, 1)$, $\boldsymbol{U}|\boldsymbol{Y} = \boldsymbol{y} \sim \mathcal{N}_p\left(\boldsymbol{0}, (\boldsymbol{\Omega}^{-1} + \boldsymbol{y}\boldsymbol{y}^\top)^{-1}\right)$, $\boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})$, *and* T *is independent of* $\boldsymbol{U}$ *and* $\boldsymbol{Y}$.

3) Hierarchical representation of the multivariate $\mathcal{MSN}$ distribution: In the next corollary, we give one hierarchical representation of the $\mathcal{MSN}$ based on the last proposition. This hierarchical representation is used later for implementation of an EM algorithm to find the maximum likelihood estimates of the parameters of the $\mathcal{MSN}$ model.

Corollary 1 *Let* $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$. *A hierarchical representation of* $\boldsymbol{X}$ *is given by*

$$\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}, W = w \sim \mathcal{N}_p\left(\boldsymbol{\xi} + w \frac{1}{1 + \boldsymbol{y}^\top \boldsymbol{\Omega} \boldsymbol{y}} \boldsymbol{\Omega} \boldsymbol{y}, (\boldsymbol{\Omega}^{-1} + \boldsymbol{y}\boldsymbol{y}^\top)^{-1}\right),$$

$$W|\boldsymbol{Y}=\boldsymbol{y}\sim\mathcal{TN}\left(0,1+\boldsymbol{y}^{\top}\boldsymbol{\Omega}\boldsymbol{y};(0,\infty)\right),$$
$$\boldsymbol{Y}\sim\mathcal{N}_p(\boldsymbol{\lambda},\boldsymbol{\Omega}^{-1}).$$

4) Some posterior moments of the missing random quantities $\boldsymbol{Y}$ and W: The results in the next two corollaries will be used directly to establish the EM algorithm.

Corollary 2 *If* $(\boldsymbol{X},\boldsymbol{Y},W)$ *have joint mpdf as given in Eq.* (21)*, then*

$$\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x})=\boldsymbol{\lambda}+W_{\phi}\left\{\frac{\boldsymbol{\lambda}^{\top}(\boldsymbol{x}-\boldsymbol{\xi})}{\sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}}\right\}\times\frac{1}{\sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi}),$$

$$\begin{aligned}\mathbb{E}\left(\boldsymbol{Y}\boldsymbol{Y}^{\top}|\boldsymbol{X}=\boldsymbol{x}\right)&=\boldsymbol{\lambda}\boldsymbol{\lambda}^{\top}+\boldsymbol{\Omega}^{-1}+W_{\phi}\left\{\frac{\boldsymbol{\lambda}^{\top}(\boldsymbol{x}-\boldsymbol{\xi})}{\sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}}\right\}\\&\quad\times\frac{1}{\sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}}\\&\quad\times\left\{\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})\boldsymbol{\lambda}^{\top}+\boldsymbol{\lambda}(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}\right.\\&\quad\left.-\frac{\boldsymbol{\lambda}^{\top}(\boldsymbol{x}-\boldsymbol{\xi})}{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}\right\},\end{aligned}$$

$$\mathbb{E}(W\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x})=\mathbb{E}(W|\boldsymbol{X}=\boldsymbol{x})\boldsymbol{\lambda}+\frac{\{\mathbb{E}(W^2|\boldsymbol{X}=\boldsymbol{x})-\boldsymbol{\lambda}^{\top}(\boldsymbol{x}-\boldsymbol{\xi})\mathbb{E}(W|\boldsymbol{X}=\boldsymbol{x})\}}{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi}),$$

$$\mathbb{E}(W|\boldsymbol{X}=\boldsymbol{x})=(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\lambda}+\sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}\times W_{\phi}\left\{\frac{\boldsymbol{\lambda}^{\top}(\boldsymbol{x}-\boldsymbol{\xi})}{\sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}}\right\},$$

$$\mathbb{E}(W^2|\boldsymbol{X}=\boldsymbol{x})=\{(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\lambda}\}^2+\{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})\}+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\lambda}\sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}\times W_{\phi}\left\{\frac{\boldsymbol{\lambda}^{\top}(\boldsymbol{x}-\boldsymbol{\xi})}{\sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}}\right\},$$

where $W_{\phi}(\cdot)=\phi(\cdot)/\Phi(\cdot)$.

5) Expectation and covariance of the multivariate $\mathcal{MSN}$ distribution: The above hierarchical representation is also useful to derive the expectation and covariance of

the multivariate $\mathcal{MSN}$ distribution. Note that even moments of the $\mathcal{MSN}$ are the same as those of the normal distribution due to the invariance property (Arellano-Valle and Genton 2010).

Proposition 2 *The expectation and the covariance matrix of* $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ *are, respectively:*

$$\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\xi} + \sqrt{\frac{2}{\pi}}\mathbb{E}\left(\frac{1}{\sqrt{1+\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y}}}\boldsymbol{\Omega}\boldsymbol{Y}\right),$$

$$\mathbb{V}\mathrm{ar}(\boldsymbol{X}) = \boldsymbol{\Omega} - \frac{2}{\pi}\mathbb{E}\left(\frac{1}{\sqrt{1+\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y}}}\boldsymbol{\Omega}\boldsymbol{Y}\right)\mathbb{E}\left(\frac{1}{\sqrt{1+\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y}}}\boldsymbol{\Omega}\boldsymbol{Y}\right)^\top,$$

where $\boldsymbol{Y} \sim \mathcal{N}_p\left(\boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1}\right)$, *with* $\boldsymbol{\lambda}$ *and* $\boldsymbol{\Omega}$ *as in* (6).

The explicit forms of expectation and the covariance of the $\mathcal{MSN}$ distribution are not available since they involve the term $\mathbb{E}\left(\frac{1}{\sqrt{1+\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y}}}\boldsymbol{\Omega}\boldsymbol{Y}\right)$ which has no explicit form. However, writing $\boldsymbol{Y} = \boldsymbol{\Omega}^{-1/2}(\alpha\bar{\boldsymbol{\lambda}} + \boldsymbol{Z})$, where $\alpha = \sqrt{\boldsymbol{\lambda}^\top\boldsymbol{\Omega}\boldsymbol{\lambda}}$, $\alpha\bar{\boldsymbol{\lambda}} = \boldsymbol{\Omega}^{1/2}\boldsymbol{\lambda}$, with $\|\bar{\boldsymbol{\lambda}}\| = 1$, and $\boldsymbol{Z} \sim \mathcal{N}_p(\boldsymbol{0}, \mathrm{I}_p)$, we have

$$\begin{aligned}
\mathbb{E}\left(\frac{\boldsymbol{\Omega}\boldsymbol{Y}}{\sqrt{1+\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y}}}\right) &= \boldsymbol{\Omega}^{1/2}\mathbb{E}\left(\frac{\boldsymbol{Z}}{\sqrt{1+\alpha^2+2\alpha\bar{\boldsymbol{\lambda}}^\top\boldsymbol{Z}+\boldsymbol{Z}^\top\boldsymbol{Z}}}\right) \\
&\quad + \boldsymbol{\lambda}\mathbb{E}\left(\frac{1}{\sqrt{1+\alpha^2+2\alpha\bar{\boldsymbol{\lambda}}^\top\boldsymbol{Z}+\boldsymbol{Z}^\top\boldsymbol{Z}}}\right) \\
&= \mathbb{E}\left(\frac{Z_1}{\sqrt{1+(\alpha+Z_1)^2+\sum_{i=2}^p Z_i^2}}\right)\boldsymbol{\Omega}^{1/2}\boldsymbol{1}_p \\
&\quad + \mathbb{E}\left(\frac{1}{\sqrt{1+(\alpha+Z_1)^2+\sum_{i=2}^p Z_i^2}}\right)\boldsymbol{\lambda},
\end{aligned}$$

since $Z_1, \ldots, Z_p \overset{iid}{\sim} \mathcal{N}_1(0,1)$ and $\bar{\boldsymbol{\lambda}}^\top\boldsymbol{Z} \overset{d}{=} Z_1$. Note that $W = \sum_{i=2}^p Z_i^2 \sim \chi^2_{p-1}$ and is independent of Z_1. Thus, to compute the required expectation we only need to solve numerically two-dimensional integrals.

6) Moment generating function of the $\mathcal{MSN}$ distribution: Here we provide the multivariate moment generating function (mmgf) of the $\mathcal{MSN}$ distribution.

Proposition 3 *The mmgf of* $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ *is*

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\{M_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{t})\}$$
$$= 2\exp\left(\boldsymbol{t}^\top\boldsymbol{\xi} + \frac{1}{2}\boldsymbol{t}^\top\boldsymbol{\Omega}\boldsymbol{t}\right)\mathbb{E}\left\{\Phi\left(\frac{\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{t}}{\sqrt{1+\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y}}}\right)\right\}, \quad \boldsymbol{t}\in\mathbb{R}^p, \tag{7}$$

where $\boldsymbol{Y}\sim\mathcal{N}_p(\boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})$*, and* $\boldsymbol{\Omega}$ *and* $\boldsymbol{\lambda}$ *are as in Eq.* (6).

7) Cumulative distribution function of the $\mathcal{MSN}$ distribution: We provide the multivariate cumulative distribution function (mcdf) of the $\mathcal{MSN}$ distribution here.

Proposition 4 *The mcdf of* $\boldsymbol{X}\sim\mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ *is*

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{E}\left[\Phi_{p+1}\left\{\begin{pmatrix}\boldsymbol{x}\\0\end{pmatrix}; \begin{pmatrix}\boldsymbol{\xi}\\0\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Omega} & -\boldsymbol{\Omega}\boldsymbol{Y}\\ -\boldsymbol{Y}^\top\boldsymbol{\Omega} & 1+\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y}\end{pmatrix}\right\}\right], \quad \boldsymbol{x}\in\mathbb{R}^p, \tag{8}$$

where $\boldsymbol{Y}\sim\mathcal{N}_p(\boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})$*, and* $\boldsymbol{\Omega}$ *and* $\boldsymbol{\lambda}$ *are as in Eq.* (6).

8) Affine transformation of the $\mathcal{MSN}$ distribution: The next proposition gives us the distribution of a linear transformation of a $\mathcal{MSN}$ random vector.

Proposition 5 *Let* $\boldsymbol{X}\sim\mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ *and consider an affine transformation* $\widetilde{\boldsymbol{X}} = \boldsymbol{a}+\boldsymbol{B}\boldsymbol{X}$*,* $\widetilde{\boldsymbol{X}}\in\mathbb{R}^k$*, for some vector* $\boldsymbol{a}\in\mathbb{R}^k$ *and for some full row rank matrix* $\boldsymbol{B}\in\mathbb{R}^{k\times p}$*,* $k\le p$*, and* $\boldsymbol{a}$ *and* $\boldsymbol{B}$ *are constant. Then, the distribution of* $\widetilde{\boldsymbol{X}}$ *has an mpdf given by*

$$f_{\widetilde{\boldsymbol{X}}}(\widetilde{\boldsymbol{x}}) = 2\phi_p\left(\widetilde{\boldsymbol{x}}; \boldsymbol{a}+\boldsymbol{B}\boldsymbol{\xi}, \boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top\right)$$
$$\times\int_{\mathbb{R}^p}\Phi\left[\frac{\boldsymbol{\lambda}_B(\boldsymbol{y})^\top(\widetilde{\boldsymbol{x}}-\boldsymbol{a}-\boldsymbol{B}\boldsymbol{\xi})}{\sqrt{1+(\widetilde{\boldsymbol{x}}-\boldsymbol{a}-\boldsymbol{B}\boldsymbol{\xi})^\top(\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top)^{-1}(\widetilde{\boldsymbol{x}}-\boldsymbol{a}-\boldsymbol{B}\boldsymbol{\xi})}}\right]$$
$$\times\,\phi_p\left(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1}\right)\mathrm{d}\boldsymbol{y},$$

for $\widetilde{\boldsymbol{x}}\in\mathbb{R}^p$*, and where* $\boldsymbol{\lambda}_B(\boldsymbol{y}) = \dfrac{(\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top)^{-1}\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{\lambda}}{\sqrt{1+\boldsymbol{y}^\top\{\boldsymbol{\Omega}-\boldsymbol{\Omega}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top)^{-1}\boldsymbol{B}\boldsymbol{\Omega}\}\boldsymbol{y}}}$*, with* $\boldsymbol{\Omega}$ *and* $\boldsymbol{\lambda}$ *being the same as in Eq.* (6)*. In particular, if the rank of* $\boldsymbol{B}$ *is* $k=p$*, then* $\widetilde{\boldsymbol{X}}\sim\mathcal{MSN}_p(\boldsymbol{a}+\boldsymbol{B}\boldsymbol{\xi}, \boldsymbol{B}\boldsymbol{\Psi}\boldsymbol{B}^\top, \boldsymbol{B}\boldsymbol{\eta})$.

9) Skewness and kurtosis of the $\mathcal{MSN}$ distribution: To measure the skewness and kurtosis of the multivariate $\mathcal{MSN}$ distribution, one can use the Mardia measure of multivariate skewness and kurtosis $\beta_{1,p}$ and $\beta_{2,p}$ (Mardia 1970). We could not find any closed form of $\beta_{1,p}$ or $\beta_{2,p}$ but for $p=1$, closed forms for $\sqrt{\beta_{1,1}}$ and $\beta_{2,1}$ are given by Arrué et al. (2016). It has been shown by Arrué et al. (2016) that $-0.9952 < \sqrt{\beta_{1,1}} < 0.9952$ and $3\le\beta_{2,1}\le 3.869$, which are exactly the same for $\mathcal{SN}_1(0, 1, \eta)$. This observation leads us to the conjecture that, although the $\mathcal{MSN}$ distribution uses a different skewing function from the $\mathcal{SN}$ distribution, the range of skewness and kurtosis remain the same.

10) Marginal distributions of the $\mathcal{MSN}$ distribution: In the following corollary we give the mpdf of the marginal $\boldsymbol{X}_1$ of $\boldsymbol{X}$ according to the partition in Eq. (17). The

marginal mpdf of $\boldsymbol{X}_2$ can be obtained in a similar way. For this, we consider the induced partition on the working parameters $(\boldsymbol{\lambda}, \boldsymbol{\Omega})$ defined in Eq. (6) as $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\top, \boldsymbol{\lambda}_2^\top)^\top$ and $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_{ij})$, with $\boldsymbol{\Omega}_{ij} = \boldsymbol{\Psi}_{ij} + \boldsymbol{\eta}_i \boldsymbol{\eta}_j^\top$, $i, j = 1, 2$, and also on the latent random vector $\boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})$ as $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top)^\top$, with $\boldsymbol{Y}_i \sim \mathcal{N}_{p_i}(\boldsymbol{\lambda}_i, \boldsymbol{\Omega}_{ii\cdot j}^{-1})$ and $\boldsymbol{\Omega}_{ii\cdot j} = \boldsymbol{\Omega}_{ii} - \boldsymbol{\Omega}_{ij}\boldsymbol{\Omega}_{jj}^{-1}\boldsymbol{\Omega}_{ji}$, $i, j = 1, 2, i \neq j$. Moreover, from an extensive but simple matrix algebra, we can show that $\boldsymbol{\Omega}_{ii\cdot j} = \boldsymbol{\Psi}_{ii\cdot j} + \boldsymbol{\eta}_{i\cdot j}\boldsymbol{\eta}_{i\cdot j}^\top$, where $\boldsymbol{\Psi}_{ii\cdot j} = \boldsymbol{\Psi}_{ii} - \boldsymbol{\Psi}_{ij}\boldsymbol{\Psi}_{jj}^{-1}\boldsymbol{\Psi}_{ji}$ and $\boldsymbol{\eta}_{i\cdot j} = \dfrac{1}{\sqrt{1 + \boldsymbol{\eta}_j^\top \boldsymbol{\Psi}_{jj}^{-1}\boldsymbol{\eta}_j}}(\boldsymbol{\eta}_i - \boldsymbol{\Psi}_{ij}\boldsymbol{\Psi}_{jj}^{-1}\boldsymbol{\eta}_j)$, $i, j = 1, 2, i \neq j$.

Corollary 3 *Let $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ be partitioned as in Eq. (17). Then, the marginal $\boldsymbol{X}_1$ of $\boldsymbol{X}$ has mpdf given by*

$$f_{X_1}(\boldsymbol{x}_1) = 2\phi_{p_1}(\boldsymbol{x}_1; \boldsymbol{\xi}_1, \boldsymbol{\Omega}_{11}) \int_{\mathbb{R}^{p_2}} \Phi\left\{ \frac{\boldsymbol{\lambda}_{1\cdot 2}^\top (\boldsymbol{x}_1 - \boldsymbol{\xi}_1)}{\sqrt{1 + (\boldsymbol{x}_1 - \boldsymbol{\xi}_1)^\top \boldsymbol{\Omega}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\xi}_1)}} \right\}$$
$$\times \phi_{p_2}\left(\boldsymbol{y}_2; \boldsymbol{\lambda}_2, \boldsymbol{\Omega}_{22\cdot 1}^{-1}\right) \mathrm{d}\boldsymbol{y}_2, \quad \boldsymbol{x}_1 \in \mathbb{R}^{p_1},$$

where $\boldsymbol{\Omega}_{11} = \boldsymbol{\Psi}_{11} + \boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top$ *and* $\boldsymbol{\lambda}_{1\cdot 2} = \sqrt{\dfrac{1 + \boldsymbol{\eta}^\top \boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}{1 + \boldsymbol{y}_2^\top \boldsymbol{\Omega}_{22\cdot 1}\boldsymbol{y}_2}} \dfrac{1}{\sqrt{1 + \boldsymbol{\eta}_1^\top \boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\eta}_1}} \boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\eta}_1$, *with* $\boldsymbol{\Omega}_{22\cdot 1} = \boldsymbol{\Psi}_{22\cdot 1} + \boldsymbol{\eta}_{2\cdot 1}\boldsymbol{\eta}_{2\cdot 1}^\top$, $\boldsymbol{\Psi}_{22\cdot 1} = \boldsymbol{\Psi}_{22} - \boldsymbol{\Psi}_{21}\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Psi}_{12}$ *and* $\boldsymbol{\eta}_{2\cdot 1} = \dfrac{1}{\sqrt{1 + \boldsymbol{\eta}_1^\top \boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\eta}_1}}(\boldsymbol{\eta}_2 - \boldsymbol{\Psi}_{21}\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\eta}_1)$.

11) Log-concavity of the mpdf of the $\mathcal{MSN}$ distribution:

Proposition 6 *The mpdf of the $\mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ is not always log-concave.*

The log-concave nature of the density function ensures that the log-likelihood function has a unique maxima. Although the mpdf of the $\mathcal{MSN}$, unlike the $\mathcal{SN}$, is not always log-concave, that does not mean the log-likelihood function does not have a unique maxima.

12) Limit behavior of the $\mathcal{MSN}_1(\xi, \Psi, \eta)$ distribution as $\Psi \to 0$: The following proposition provides the behavior of the univariate $\mathcal{MSN}_1(\xi, \Psi, \eta)$ distribution as the scale parameter $\Psi \to 0$:

Proposition 7 *The $\mathcal{MSN}_1(\xi, \Psi, \eta)$ distribution converges to $\mathcal{TN}\left(\xi, \eta^2; (\xi, \infty)\right)$ as $\Psi \to 0$ for $\eta > 0$ and converges to $\mathcal{TN}\left(\xi, \eta^2; (-\infty, \xi)\right)$ as $\Psi \to 0$ for $\eta < 0$.*

The last proposition states that the univariate $\mathcal{MSN}$ family includes the truncated normal distribution as a limiting case. The truncated normal distribution is also a special case of the univariate $\mathcal{ASN}$ distribution when the skewness parameter $\alpha \to \pm\infty$. Therefore, if we fit the $\mathcal{ASN}_1$ distribution to some extremely skewed univariate data, we can get an infinite estimate of the skewness parameter, while with the $\mathcal{MSN}_1$ distribution, we will get an estimate of the scale parameter close to 0, but as established

in the Proposition 8 below in Subsect. 3.2, the estimate of the skewness parameter will always be finite. Moreover, a statement similar to Proposition 7 cannot be made for dimensions larger than one. This is because the $\mathcal{MSN}$ distribution is defined in the multivariate case assuming that the scale matrix $\boldsymbol{\Psi}$ is positive definite. Hence, if we consider the multivariate extension of Proposition 7, we end up with a positive semi-definite scale matrix $\boldsymbol{\Psi}$. From this point, it is not possible to extend Proposition 7 to the multivariate setup, since it involves the inversion of $\boldsymbol{\Psi}$.

3.2 Properties of the likelihood function of the $\mathcal{MSN}$ model

In this section we present some important aspects of the likelihood function of the $\mathcal{MSN}$ model. First, we show that, as in the $\mathcal{SN}$ case, the maximum likelihood estimator of the skewness parameter $\boldsymbol{\eta}$ obtained from the $\mathcal{MSN}$ model is always finite. Next, we show that, unlike the $\mathcal{SN}$ model, for the $\mathcal{MSN}$ model the Fisher information matrix is nonsingular at $\boldsymbol{\eta} = \mathbf{0}$.

1) Profile likelihood function of the skewness parameter in the $\mathcal{MSN}$ model: For the $\mathcal{MSN}$ model, the likelihood function for $(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ based on a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ is

$$L(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta}) = \prod_{i=1}^{n} 2\phi_p\left(\boldsymbol{x}_i; \boldsymbol{\xi}, \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top\right) \Phi\left\{\frac{1}{\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}} \frac{\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi})}{\sqrt{1+(\boldsymbol{x}_i-\boldsymbol{\xi})^\top(\boldsymbol{\Psi}+\boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi})}}\right\},$$

which becomes the profile likelihood function for the skewness parameter $\boldsymbol{\eta}$ when we fix $\boldsymbol{\xi}$ and $\boldsymbol{\Psi}$. Next proposition shows that the profile likelihood function for the skewness parameter $\boldsymbol{\eta}$ always gets maximized at some finite point.

Proposition 8 *The maximum likelihood estimator of the skewness parameter $\boldsymbol{\eta}$ of the $\mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ distribution is always finite.*

2) Fisher information matrix of the $\mathcal{MSN}$ model: Now, we obtain the Fisher information matrix of the $\mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ model at $\boldsymbol{\eta} = \mathbf{0}$. In order to test that the skewness parameter $\boldsymbol{\eta} = \mathbf{0}$, the regular asymptotic theory can also be applied to study the asymptotic behavior of the maximum likelihood estimators under the null hypothesis of symmetry when the Fisher information matrix is nonsingular at $\boldsymbol{\eta} = \mathbf{0}$. To derive the Fisher information matrix, we use the fact that the $\mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ family is a special case of the generalized skew-elliptical family. Similar to the $\mathcal{SN}$ family, it can be easily established that the $\mathcal{MSN}$ belongs to the generalized skew-elliptical family with location parameter $\boldsymbol{\xi}$, dispersion matrix $\boldsymbol{\Omega}$, density generator $\phi_p(\cdot)$, and skewing function $\pi(\boldsymbol{z}) : \mathbb{R}^p \to [0, 1]$ with $\pi(\boldsymbol{z}) = \Phi(\boldsymbol{\gamma}^\top \boldsymbol{z}/\sqrt{1+\boldsymbol{z}^\top\boldsymbol{z}})$, where $\boldsymbol{\Omega}$ and $\boldsymbol{\gamma}$ are the same as defined below Eq. (18), just by rewriting its mpdf by replacing $\boldsymbol{\eta}$ and $\boldsymbol{\Psi}$ using the relationships indicated there. The derivation of the Fisher information matrix at $\boldsymbol{\eta} = \mathbf{0}$ is very much influenced by Ley and Paindaveine (2010), where the authors derived the Fisher information matrix for the generalized skew-elliptical family, when the skewness parameter is set to zero. In our case, exactly the same steps are followed as Ley and Paindaveine (2010), since we are dealing with a special case of the aforementioned paper.

Let $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ and denote by $\boldsymbol{l}_{\boldsymbol{\xi},\boldsymbol{\Psi},\boldsymbol{\eta}}(\boldsymbol{X})$ the respective score vector for $(\boldsymbol{\xi}^\top, \text{vech}(\boldsymbol{\Psi})^\top, \boldsymbol{\eta}^\top)^\top$. At $\boldsymbol{\eta} = \mathbf{0}$, the score vector, $\boldsymbol{l}_{\boldsymbol{\xi},\boldsymbol{\Psi},\mathbf{0}}(\boldsymbol{X})$, is given by

$$\left(\boldsymbol{\Psi}^{-1}(\boldsymbol{X}-\boldsymbol{\xi}), \frac{1}{2}\boldsymbol{D}_p^\top(\boldsymbol{\Psi}\otimes\boldsymbol{\Psi})^{-1}\text{vec}\{(\boldsymbol{X}-\boldsymbol{\xi})(\boldsymbol{X}-\boldsymbol{\xi})^\top - \boldsymbol{\Psi}^{-1}\}, \sqrt{\frac{2}{\pi}}\frac{\boldsymbol{\Psi}^{-1}(\boldsymbol{X}-\boldsymbol{\xi})}{\sqrt{1+(\boldsymbol{X}-\boldsymbol{\xi})^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{X}-\boldsymbol{\xi})}}\right)^\top.$$

From the score vector at $\boldsymbol{\eta} = \mathbf{0}$, we can see clearly that the components are not linearly related. Thus, the Fisher information matrix, which is the covariance matrix of the score vector, is nonsingular as well when $\boldsymbol{\eta} = \mathbf{0}$. In fact, let $\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ be the Fisher information matrix of $(\boldsymbol{\xi}^\top, \text{vech}(\boldsymbol{\Psi})^\top, \boldsymbol{\eta}^\top)^\top$ for the $\mathcal{MSN}$ model. Then, at $\boldsymbol{\eta} = \mathbf{0}$, this matrix becomes

$$\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \mathbf{0}) = \int_{\mathbb{R}^p} \boldsymbol{l}_{\boldsymbol{\xi},\boldsymbol{\Psi},\mathbf{0}}(\boldsymbol{x})\boldsymbol{l}^\top_{\boldsymbol{\xi},\boldsymbol{\Psi},\mathbf{0}}(\boldsymbol{x})\phi_p(\boldsymbol{x};\boldsymbol{\xi},\boldsymbol{\Psi})\mathrm{d}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{i}_{\boldsymbol{\xi}\boldsymbol{\xi}} & \mathbf{0} & \boldsymbol{i}_{\boldsymbol{\xi}\boldsymbol{\eta}} \\ \mathbf{0} & \boldsymbol{i}_{\boldsymbol{\Psi}\boldsymbol{\Psi}} & \mathbf{0} \\ \boldsymbol{i}_{\boldsymbol{\xi}\boldsymbol{\eta}} & \mathbf{0} & \boldsymbol{i}_{\boldsymbol{\eta}\boldsymbol{\eta}} \end{bmatrix}, \tag{9}$$

where $\boldsymbol{i}_{\boldsymbol{\xi}\boldsymbol{\xi}} = \boldsymbol{\Psi}^{-1}$, $\boldsymbol{i}_{\boldsymbol{\xi}\boldsymbol{\eta}} = \sqrt{\frac{2}{\pi}}\boldsymbol{\Psi}^{-1/2}\mathbb{E}\left(\frac{\boldsymbol{Z}\boldsymbol{Z}^\top}{\sqrt{1+\boldsymbol{Z}^\top\boldsymbol{Z}}}\right)\boldsymbol{\Psi}^{-1/2}$, $\boldsymbol{i}_{\boldsymbol{\eta}\boldsymbol{\eta}} = \frac{2}{\pi}\boldsymbol{\Psi}^{-1/2}\mathbb{E}\left(\frac{\boldsymbol{Z}\boldsymbol{Z}^\top}{1+\boldsymbol{Z}^\top\boldsymbol{Z}}\right)\boldsymbol{\Psi}^{-1/2}$, $\boldsymbol{i}_{\boldsymbol{\Psi}\boldsymbol{\Psi}} = \frac{1}{2}\boldsymbol{D}_p^\top(\boldsymbol{\Psi}\otimes\boldsymbol{\Psi})\boldsymbol{D}_p$ and $\boldsymbol{Z} \sim \mathcal{N}_p(\mathbf{0}, \mathrm{I}_p)$. The zero blocks in $\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \mathbf{0})$ can easily be obtained by noticing that the score in $\text{vech}(\boldsymbol{\Psi})$ is symmetric with respect to $\boldsymbol{X}-\boldsymbol{\xi}$, whereas the scores in $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are anti-symmetric with respect to the same quantity.

Proposition 9 *The Fisher information matrix of the* $\mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ *as given in Eq.* (9) *is nonsingular at* $\boldsymbol{\eta} = \mathbf{0}$.

4 Inference for the $\mathcal{MSN}$ distribution

4.1 EM algorithm

The log-likelihood function of the $\mathcal{MSN}$ distribution has a structure too complicated to be directly maximized in order to obtain the maximum likelihood estimators of the parameters. In this section, we provide an EM algorithm to find the maximum likelihood estimates of the $\mathcal{MSN}$ model.

We use the hierarchical representation of the augmented vector $(\boldsymbol{X}, \boldsymbol{Y}, W)$ for the EM algorithm as given in Corollary 1. To fit the $\mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ model, we consider the alternative parameterization $(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$, as in Corollary 1, and then we estimate these parameters. The maximum likelihood estimators of $(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ can then be easily found by transforming the maximum likelihood estimators for $(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$ accordingly.

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample of size n from $\mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$ and let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ and $W_1, \ldots, W_n$ be the corresponding latent variables. Since $\boldsymbol{\Omega}$ is a positive definite matrix, $\boldsymbol{\Omega}$ can be written as $\boldsymbol{C}^\top \boldsymbol{C}$, where $\boldsymbol{C}$ is a $p \times p$ nonsingular matrix. Let $\boldsymbol{\theta} = \{\boldsymbol{\xi}^\top, \text{vec}(\boldsymbol{C})^\top, \boldsymbol{\lambda}^\top\}^\top$ and define $\boldsymbol{X}_c = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top, \boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_n^\top, W_1, \ldots, W_n)^\top$. After some algebra, we find that the complete log-likelihood for $\boldsymbol{\theta}$ when $\boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n$ are observed is

$$\begin{aligned}\ell_c(\boldsymbol{\theta}|\boldsymbol{X}_c) = n\log 2 - n\frac{(2p+1)}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\xi})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{x}_i - \boldsymbol{\xi}) \\ - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\lambda})^\top \boldsymbol{\Omega}(\boldsymbol{y}_i - \boldsymbol{\lambda}) - \frac{1}{2}\sum_{i=1}^{n}\left\{w_i - \boldsymbol{y}_i^\top(\boldsymbol{x}_i - \boldsymbol{\xi})\right\}^2.\end{aligned}$$

Given the current estimates $\widehat{\boldsymbol{\theta}}_{(k)} = \{\widehat{\boldsymbol{\xi}}_{(k)}^\top, \text{vec}(\widehat{\boldsymbol{C}}_{(k)})^\top, \widehat{\boldsymbol{\lambda}}_{(k)}^\top\}$, the E-step is given by the Q-function

$$\begin{aligned}Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{(k)}) &= \mathbb{E}\{\ell_c(\boldsymbol{\theta}|\boldsymbol{X}_c)|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \widehat{\boldsymbol{\theta}}_{(k)}\} \\ &= c - \frac{1}{2}\sum_{i=1}^{n}[(\boldsymbol{x}_i - \boldsymbol{\xi})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{x}_i - \boldsymbol{\xi}) + \text{tr}(\boldsymbol{\Omega}\widehat{\boldsymbol{P}}_{(k)i}) - 2\boldsymbol{\lambda}^\top \boldsymbol{\Omega}\widehat{\boldsymbol{q}}_{(k)i} \\ &\quad + \boldsymbol{\lambda}^\top \boldsymbol{\Omega}\boldsymbol{\lambda} + \widehat{r}_{(k)i} - 2\widehat{\boldsymbol{s}}_{(k)i}^\top(\boldsymbol{x}_i - \boldsymbol{\xi}) + (\boldsymbol{x}_i - \boldsymbol{\xi})^\top \widehat{\boldsymbol{P}}_{(k)i}(\boldsymbol{x}_i - \boldsymbol{\xi})],\end{aligned} \tag{10}$$

where $c = n\log(2) - \dfrac{n(2p+1)}{2}\log(2\pi)$, $\widehat{\boldsymbol{P}}_{(k)i} = \mathbb{E}(\boldsymbol{Y}_i \boldsymbol{Y}_i^\top|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \widehat{\boldsymbol{\theta}}_{(k)})$, $\widehat{\boldsymbol{q}}_{(k)i} = \mathbb{E}(\boldsymbol{Y}_i|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \widehat{\boldsymbol{\theta}}_{(k)})$, $\widehat{r}_{(k)i} = \mathbb{E}(W_i^2|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \widehat{\boldsymbol{\theta}}_{(k)})$, and $\widehat{\boldsymbol{s}}_{(k)i} = \mathbb{E}(W_i \boldsymbol{Y}_i|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \widehat{\boldsymbol{\theta}}_{(k)})$, which can be computed using the results in Corollary 2.

We need to maximize $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{(k)})$ with respect to $\boldsymbol{\theta}$ in the M-step. To do this we differentiate $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{(k)})$ with respect to $\boldsymbol{\theta}$ and then equate it to $\boldsymbol{0}$ to get the following $(k+1)$-estimation equation:

$$\frac{\partial Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{(k)})}{\partial \boldsymbol{\xi}} = \sum_{i=1}^{n}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}_i - \boldsymbol{\xi}) - \sum_{i=1}^{n}\left\{\widehat{\boldsymbol{s}}_{(k)i} - \widehat{\boldsymbol{P}}_{(k)i}(\boldsymbol{x}_i - \boldsymbol{\xi})\right\} = \boldsymbol{0},$$

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{(k)})}{\partial \boldsymbol{C}} &= \sum_{i=1}^{n}[(\boldsymbol{C}^\top)^{-1}(\boldsymbol{x}_i - \boldsymbol{\xi})(\boldsymbol{x}_i - \boldsymbol{\xi})^\top(\boldsymbol{C}^\top\boldsymbol{C})^{-1} - \{\boldsymbol{C}\widehat{\boldsymbol{P}}_{(k)i} - \boldsymbol{C}(\widehat{\boldsymbol{q}}_{(k)i}\boldsymbol{\lambda}^\top \\ &\quad + \boldsymbol{\lambda}\widehat{\boldsymbol{q}}_{(k)i}^\top) + \boldsymbol{C}\boldsymbol{\lambda}\boldsymbol{\lambda}^\top\}] = \boldsymbol{0},\end{aligned} \tag{11}$$

$$\frac{\partial Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{(k)})}{\partial \boldsymbol{\lambda}} = \sum_{i=1}^{n}\{\boldsymbol{\Omega}\widehat{\boldsymbol{q}}_{(k)i} - \boldsymbol{\Omega}\boldsymbol{\lambda}\} = \boldsymbol{0}. \tag{12}$$

From Eq. (12), we obtain $\widehat{\boldsymbol{\lambda}}_{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{q}}_{(k)i} = \bar{\boldsymbol{q}}_k$. Putting $\widehat{\boldsymbol{\lambda}}_{(k+1)} = \bar{\boldsymbol{q}}_k$ in place of $\boldsymbol{\lambda}$ in Eq. (11) and then premultiplying $\boldsymbol{C}^{-1}$, we get

$$\sum_{i=1}^{n}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi})(\boldsymbol{x}_i-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1} = \sum_{i=1}^{n}\left(\widehat{\boldsymbol{P}}_{(k)i} - \bar{\boldsymbol{q}}_k\bar{\boldsymbol{q}}_k^{\top}\right).$$

The solution for $\boldsymbol{\xi}$ and $\boldsymbol{\Omega}$ from the equations

$$\sum_{i=1}^{n}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi}) = \sum_{i=1}^{n}\left\{\widehat{\boldsymbol{s}}_{(k)i} - \widehat{\boldsymbol{P}}_{(k)i}(\boldsymbol{x}_i-\boldsymbol{\xi})\right\},$$
$$\sum_{i=1}^{n}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi})(\boldsymbol{x}_i-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1} = \sum_{i=1}^{n}\left(\widehat{\boldsymbol{P}}_{(k)i} - \bar{\boldsymbol{q}}_k\bar{\boldsymbol{q}}_k^{\top}\right),$$

gives $\widehat{\boldsymbol{\xi}}_{(k+1)}$ and $\widehat{\boldsymbol{\Omega}}_{(k+1)}$. We cannot solve these equations analytically when the dimension $p > 1$.

Another approach is to solve for $\boldsymbol{\Omega}$, after fixing $\boldsymbol{\xi}$ in Eq. (11) and get

$$\boldsymbol{\Omega}(\boldsymbol{\xi}) = \boldsymbol{A}_k^{-1/2}\left[\boldsymbol{A}_k^{1/2}\left\{\boldsymbol{S} + (\bar{\boldsymbol{x}}-\boldsymbol{\xi})(\bar{\boldsymbol{x}}-\boldsymbol{\xi})^{\top}\right\}\boldsymbol{A}_k^{1/2}\right]^{1/2}\boldsymbol{A}_k^{-1/2}, \tag{13}$$

where $\boldsymbol{A}_k = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\boldsymbol{P}}_{(k)i} - \bar{\boldsymbol{q}}_k\bar{\boldsymbol{q}}_k^{\top})$ and $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\top}$. Then we put $\boldsymbol{\lambda} = \widehat{\boldsymbol{\lambda}}_{(k+1)} = \bar{\boldsymbol{q}}_k$ and $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\xi})$ in the expression of $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{(k)})$ from Eq. (10) to obtain

$$\begin{aligned}
Q(\boldsymbol{\xi}|\widehat{\boldsymbol{\theta}}_{(k)}) \propto & -\frac{1}{2}\sum_{i=1}^{n}[(\boldsymbol{x}_i-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}(\boldsymbol{\xi})^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi}) + \mathrm{tr}\{\boldsymbol{\Omega}(\boldsymbol{\xi})\widehat{\boldsymbol{P}}_{(k)i}\} - \bar{\boldsymbol{q}}_k^{\top}\boldsymbol{\Omega}(\boldsymbol{\xi})\bar{\boldsymbol{q}}_k \\
& + 2\widehat{\boldsymbol{s}}_{(k)i}^{\top}\boldsymbol{\xi} - 2\boldsymbol{x_i}^{\top}\widehat{\boldsymbol{P}}_{(k)i}\boldsymbol{\xi} + \boldsymbol{\xi}^{\top}\widehat{\boldsymbol{P}}_{(k)i}\boldsymbol{\xi}] \\
= & -\frac{n}{2}[\mathrm{tr}\{\boldsymbol{\Omega}(\boldsymbol{\xi})^{-1}\boldsymbol{S}\} + (\bar{\boldsymbol{x}}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}(\boldsymbol{\xi})^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\xi}) + \mathrm{tr}\{\boldsymbol{\Omega}(\boldsymbol{\xi})\bar{\boldsymbol{P}}_k\} \\
& - \bar{\boldsymbol{q}}_k^{\top}\boldsymbol{\Omega}(\boldsymbol{\xi})\bar{\boldsymbol{q}}_k + 2\bar{\boldsymbol{s}}_k^{\top}\boldsymbol{\xi} - 2\boldsymbol{P}_{\boldsymbol{x}_k}^{\top}\boldsymbol{\xi} + \boldsymbol{\xi}^{\top}\bar{\boldsymbol{P}}_k\boldsymbol{\xi}],
\end{aligned}$$

where $\bar{\boldsymbol{s}}_k = \frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{s}}_{(k)i}$, $\bar{\boldsymbol{P}}_k = \frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{P}}_{(k)i}$, and $\boldsymbol{P}_{\boldsymbol{x}_k} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{P}}_{(k)i}\boldsymbol{x}_i$. To get $\widehat{\boldsymbol{\xi}}_{(k+1)}$, we maximize $Q(\boldsymbol{\xi}|\widehat{\boldsymbol{\theta}}_{(k)})$ numerically with respect to $\boldsymbol{\xi}$, and we obtain $\widehat{\boldsymbol{\Omega}}_{(k+1)} = \boldsymbol{\Omega}(\widehat{\boldsymbol{\xi}}_{(k+1)})$, from Eq. (13). We stop when the difference between the log-likelihood for $\widehat{\boldsymbol{\theta}}_{(k)}$ and $\widehat{\boldsymbol{\theta}}_{(k+1)}$, i.e., $\ell(\widehat{\boldsymbol{\theta}}_{(k+1)}) - \ell(\widehat{\boldsymbol{\theta}}_{(k)})$, becomes very small, where

$$\ell(\boldsymbol{\theta}) = n\log(2) - \frac{np}{2}\log(2\pi) - \frac{n}{2}\log\{\det(\boldsymbol{\Omega})\} - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\xi})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{x}_i - \boldsymbol{\xi}) + \sum_{i=1}^{n}\Phi\left\{\frac{\boldsymbol{\lambda}^\top(\boldsymbol{x}_i - \boldsymbol{\xi})}{\sqrt{1 + (\boldsymbol{x}_i - \boldsymbol{\xi})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{x}_i - \boldsymbol{\xi})}}\right\}.$$

The EM algorithm is thus given as follows:

Step 1: For given estimates $\boldsymbol{\xi}_{(k)}$, $\boldsymbol{\Omega}_{(k)}$ and $\boldsymbol{\lambda}_{(k)}$, obtain the function $Q(\boldsymbol{\xi}|\widehat{\boldsymbol{\theta}}_{(k)})$.

Step 2: Numerically obtain $\widehat{\boldsymbol{\xi}}_{(k+1)}$, which maximizes $Q(\boldsymbol{\xi}|\widehat{\boldsymbol{\theta}}_{(k)})$, and obtain $\widehat{\boldsymbol{\Omega}}_{(k+1)} = \boldsymbol{\Omega}(\widehat{\boldsymbol{\xi}}_{(k+1)})$ from Eq. (13) and $\widehat{\boldsymbol{\lambda}}_{(k+1)} = \bar{\boldsymbol{q}}_k$.

Step 3: Repeat Step 1 and Step 2 until the difference between the log-likelihood for $\widehat{\boldsymbol{\theta}}_{(k)}$ and $\widehat{\boldsymbol{\theta}}_{(k+1)}$ becomes very small.

Step 4: Finally, obtain the estimates of $\boldsymbol{\Psi}$ and $\boldsymbol{\eta}$ from the relation $\boldsymbol{\Psi} = \boldsymbol{\Omega} - \dfrac{\boldsymbol{\Omega}\boldsymbol{\lambda}\boldsymbol{\lambda}^\top\boldsymbol{\Omega}}{1 + \boldsymbol{\lambda}^\top\boldsymbol{\Omega}\boldsymbol{\lambda}}$ and $\boldsymbol{\eta} = \dfrac{\boldsymbol{\Omega}\boldsymbol{\lambda}}{\sqrt{1 + \boldsymbol{\lambda}^\top\boldsymbol{\Omega}\boldsymbol{\lambda}}}$.

To maximize $Q(\boldsymbol{\xi}|\widehat{\boldsymbol{\theta}}_{(k)})$, we minimize $-Q(\boldsymbol{\xi}|\widehat{\boldsymbol{\theta}}_{(k)})$ using the R Core Team (2021) function `nlm`, assuming $-Q(\boldsymbol{\xi}|\widehat{\boldsymbol{\theta}}_{(k)})$ is convex. Although this claim has not been proved, it is not contradicted in our simulation study. Since we have to perform a p-dimensional numerical optimization in each maximization step of the EM algorithm, the use of the `nlm` function gives a shorter run time. Another important point to consider is the initialization of the parameters for the EM algorithm. If we start with initial parameters very far from the original ones, the algorithm may converge to an undesirable local point. Here, we use the sample mean vector and the sample covariance matrix for the initial parameter estimates of $\boldsymbol{\xi}$ and $\boldsymbol{\Omega}$, and for the initial estimates of $\boldsymbol{\lambda}$, we use the componentwise skewness measure. These initial estimates work well when the absolute values of all the skewness parameters are not very large, otherwise the initial estimates of the parameters become very far from the actual parameters. Moreover, in Proposition 6 it has been established that the log-likelihood function for the $\mathcal{MSN}$ distribution is not always log-concave. Hence, there is a chance that the EM algorithm might converge to a local maximum. In order to avoid this problem, one can start the EM algorithm with multiple initial parameters and see which one yields the highest likelihood value in the end. We have provided a simulation study in the section "S1" in the supplementary material to justify the last statement. We have also shared the link of the R code for fitting the $\mathcal{MSN}$ distribution to a dataset with this proposed EM algorithm in the section "S3" in the supplementary material.

4.2 Testing for multivariate normality

Since the Fisher information matrix of the $\mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ family is nonsingular when $\boldsymbol{\eta} = \boldsymbol{0}$, we can test whether a dataset has a multivariate normal distribution or not by fitting a $\mathcal{MSN}$ model and then using the regular asymptotic theory to test the hypothesis an $H_0 : \boldsymbol{\eta} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{\eta} \neq \boldsymbol{0}$. A classical testing procedure is given as

follows: Let $\widehat{\boldsymbol{\xi}}$, $\widehat{\boldsymbol{\Psi}}$ and $\widehat{\boldsymbol{\eta}}$ be the maximum likelihood estimators of $\boldsymbol{\xi}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\eta}$, respectively, based on a random sample of size n from the p-variate $\mathcal{MSN}$ distribution. Then, using the asymptotic normality of the maximum likelihood estimators, we have

$$\left(\widehat{\boldsymbol{\xi}}, \operatorname{vech}(\widehat{\boldsymbol{\Psi}}), \widehat{\boldsymbol{\eta}}\right)^{\top} \approx \mathcal{N}_{2p+p(p+1)/2}\left((\boldsymbol{\xi}, \operatorname{vech}(\boldsymbol{\Psi}), \boldsymbol{\eta})^{\top}, \frac{1}{n}\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})^{-1}\right),$$

which holds whatever the value of the skewness parameter $\boldsymbol{\eta} \in \mathbb{R}^p$. From this, we have

$$\widehat{\boldsymbol{\eta}} \approx \mathcal{N}_p\left(\boldsymbol{\eta}, \frac{1}{n}\boldsymbol{I}^{\boldsymbol{\eta}\boldsymbol{\eta}}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})\right),$$

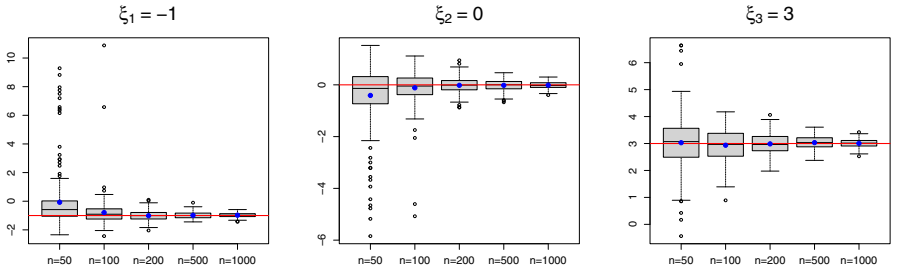
where $\boldsymbol{I}^{\boldsymbol{\eta}\boldsymbol{\eta}}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ is the last $p \times p$ block diagonal matrix of $\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})^{-1}$ and $\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ can be estimated with $\boldsymbol{I}(\widehat{\boldsymbol{\xi}}, \widehat{\boldsymbol{\Psi}}, \widehat{\boldsymbol{\eta}})$. Under H_0, $n\widehat{\boldsymbol{\eta}}^{\top}\{\boldsymbol{I}^{\boldsymbol{\eta}\boldsymbol{\eta}}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})\}^{-1}\widehat{\boldsymbol{\eta}} \approx \chi^2_p$, and using this we can compute the p-value of the testing problem as $1 - F_{\chi^2_p}[n\widehat{\boldsymbol{\eta}}^{\top}\{\boldsymbol{I}^{\boldsymbol{\eta}\boldsymbol{\eta}}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})\}^{-1}\widehat{\boldsymbol{\eta}}]$, where $F_{\chi^2_p}(\cdot)$ is the distribution function of a χ^2_p random variable.

Obviously, we can also use the classical likelihood ratio statistic for testing $H_0 : \boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{0})$ against $H_1 : \boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$, since the information matrix is non-singular under the null hypothesis. Under H_0, the maximum likelihood estimators become $\widetilde{\boldsymbol{\xi}} = \bar{\boldsymbol{X}}$, the sample mean vector, $\widetilde{\boldsymbol{\Psi}} = \boldsymbol{S}_X$, the sample covariance matrix, and $\widetilde{\boldsymbol{\eta}} = \boldsymbol{0}$. Thus, denoting by $\Lambda = \Lambda(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ the likelihood ratio statistic, the regular asymptotic theory implies, under the null hypothesis, that $-2\log\Lambda = -2\{\log L(\widetilde{\boldsymbol{\xi}}, \widetilde{\boldsymbol{\Psi}}, \boldsymbol{0}) - \log L(\widehat{\boldsymbol{\xi}}, \widehat{\boldsymbol{\Psi}}, \widehat{\boldsymbol{\eta}})\} \approx \chi^2_p$, and we can again compute the p-value in this case as $1 - F_{\chi^2_p}(-2\log\Lambda_{\text{obs.}})$, where $\Lambda_{\text{obs.}} = \Lambda(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is the observed value of Λ.

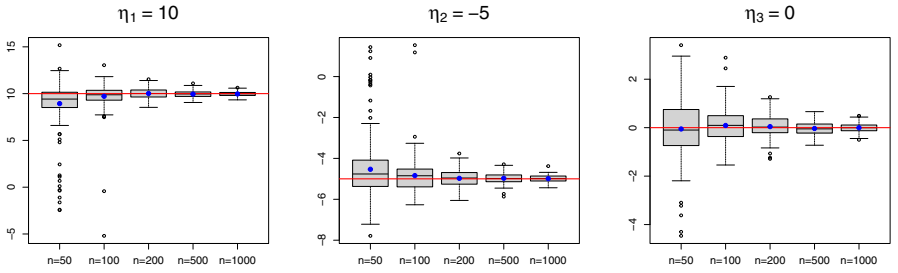
5 Simulation study

In this section, we provide the results of various simulation studies conducted to check the performance of the suggested EM algorithm. The first experiment is to verify the consistency of the estimates obtained from the algorithm. Here, we generate observations from a 3-dimensional $\mathcal{MSN}$ distribution with location parameter $\boldsymbol{\xi} = (-1, 0, 3)^{\top}$, scale parameter $\boldsymbol{\Psi} = \begin{pmatrix} 1 & 1.2 & -0.9 \\ 1.2 & 4 & -0.6 \\ -0.9 & -0.6 & 9 \end{pmatrix}$, and skewness parameter $\boldsymbol{\eta} = (10, -5, 0)^{\top}$ for different sample sizes $n = 50$, 100, 200, 500, and 1000. We then estimate the parameters using our proposed EM algorithm based on the simulated observations.

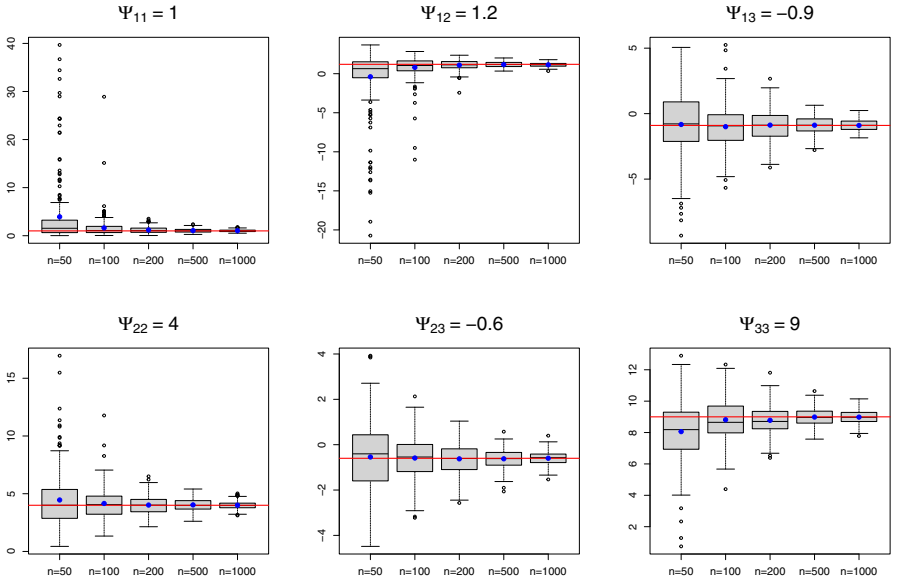
We repeat the process 200 times for each of the three sample sizes. The estimated parameter values are plotted in boxplots (Fig. 3) for different sample sizes. From the boxplots, we can see that the variances of the estimated parameter values become smaller as the sample size increases. The same can be said for the mean square error as well. From these, we can conclude that the EM algorithm provides consistent



(a) Boxplots of the estimates of $\boldsymbol{\xi}$



(b) Boxplots of the estimates of $\boldsymbol{\eta}$



(c) Boxplots of the estimates of $\boldsymbol{\Psi}$

Fig. 3 Boxplots of the parameter estimates of a $\mathcal{MSN}_3$ distribution obtained from the EM algorithm for different sample sizes n: (a) $\boldsymbol{\xi}$; (b) $\boldsymbol{\eta}$; (c) $\boldsymbol{\Psi}$. The red line indicates the true parameter value, and the blue point indicates the mean of the estimated parameter values. (Color figure online)

estimates of the parameters, as it should since the EM algorithm eventually yields maximum likelihood estimates of the parameters.

To check the performance of the EM algorithm under different circumstances, we conduct two separate studies: one with diagonal scale parameter $\boldsymbol{\Psi}$ and another one with equicorrelated $\boldsymbol{\Psi}$. Again, we generate observations from $\mathcal{MSN}_3(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ of sample size $n = 250$, with $\boldsymbol{\xi} = (0, 0, 0)^\top$, $\boldsymbol{\eta} = (10, -5, 0)^\top$, $\boldsymbol{\Psi} = \text{diag}(1, 4, 9)$ for the diagonal scale parameter, and $\boldsymbol{\Psi} = (1 - \rho)\text{I}_3 + \rho\mathbf{1}\mathbf{1}^\top$, $\rho = 0.4$, for the equicorrelated scale parameter. We estimate the parameters in both scenarios 200 times, with the estimated parameter values presented in boxplots in Fig. 4. From the results in the boxplots, we conclude that the suggested EM algorithm performs well under various circumstances. These simulation studies indicate that the suggested EM algorithm is indeed working correctly.

We provide the result of another simulation study regarding the estimation of the parameters when the skewness parameter $\boldsymbol{\eta}$ is zero. In this study, we generate observations from $\mathcal{MSN}_3(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ with $\boldsymbol{\xi} = (0, 0, 0)^\top$, $\boldsymbol{\eta} = (0, 0, 0)^\top$, and $\boldsymbol{\Psi} = \begin{pmatrix} 1 & 0.2 & -0.7 \\ 0.2 & 1 & -0.5 \\ -0.7 & -0.5 & 1 \end{pmatrix}$ with sample sizes 200, 500, and 1000. We find the MLEs of the parameters and repeat the experiment 200 times. The MLE boxplots for different sample sizes are given in Fig. 5. The boxplots show that in scenarios when the skewness parameter is zero, the MLEs of the parameters seem reasonable. The elements of the scale matrix are sometimes getting over or under-estimated. However, the estimates seem appropriate when the sample size is increased. Along with the boxplots of the estimates of $\boldsymbol{\eta}$ and $\boldsymbol{\Psi}$, we also provide the boxplots of the estimates of $\boldsymbol{\lambda}$ and $\boldsymbol{\Omega}$ in Fig. 6. The boxplots of the parameter estimates of $\boldsymbol{\Omega}$ and $\boldsymbol{\lambda}$ show that the estimates behave like the MLEs as it should. We have provided the results of a simulation study when the skewness parameter is close to zero in the section "S2" in the supplementary material.

Next, we performed another experiment, where we fit the $\mathcal{MSN}$ and $\mathcal{ASN}$ models to the data simulated from various models other than the $\mathcal{MSN}$ and $\mathcal{ASN}$ models. We generated bivariate observations of different sizes (50, 100, 200, 500, and 1000) from Clayton, Frank, and Gumbel copulas, with the value of Kendall's tau being 0.6, -0.5, and 0.5, respectively. Next, we transformed the generated observations from the uniform scale to the Gaussian scale with marginal means 1 and 0, and marginal standard deviations 2 and 1.2 for all three copula models. Moreover, we generated observations of different sizes (50, 100, 200, 500, and 1000) from a bivariate generalized hyperbolic distribution (see Chapter 3.2 McNeil et al. 2015). The stochastic representation of $\boldsymbol{X}$ following a generalized hyperbolic distribution is

$$\boldsymbol{X} \overset{\text{d}}{=} \boldsymbol{\mu} + W\boldsymbol{\gamma} + \sqrt{W}\boldsymbol{Z}, \tag{14}$$

where $\boldsymbol{\mu}, \boldsymbol{\gamma} \in \mathbb{R}^p$, $\boldsymbol{Z} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, and W has a Generalized Inverse Gaussian ($\mathcal{GIG}(\lambda, \chi, \psi)$) distribution. We use $\boldsymbol{\mu} = (1, 0)^\top$, $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 0 \\ 0 & 1.2 \end{pmatrix}\begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}\begin{pmatrix} 2 & 0 \\ 0 & 1.2 \end{pmatrix}$, $\boldsymbol{\gamma} = (-0.2, 0.3)^\top$, and $W \sim \mathcal{GIG}(\lambda = 0.5, \chi = 0.1, \psi = 2)$ for gen-

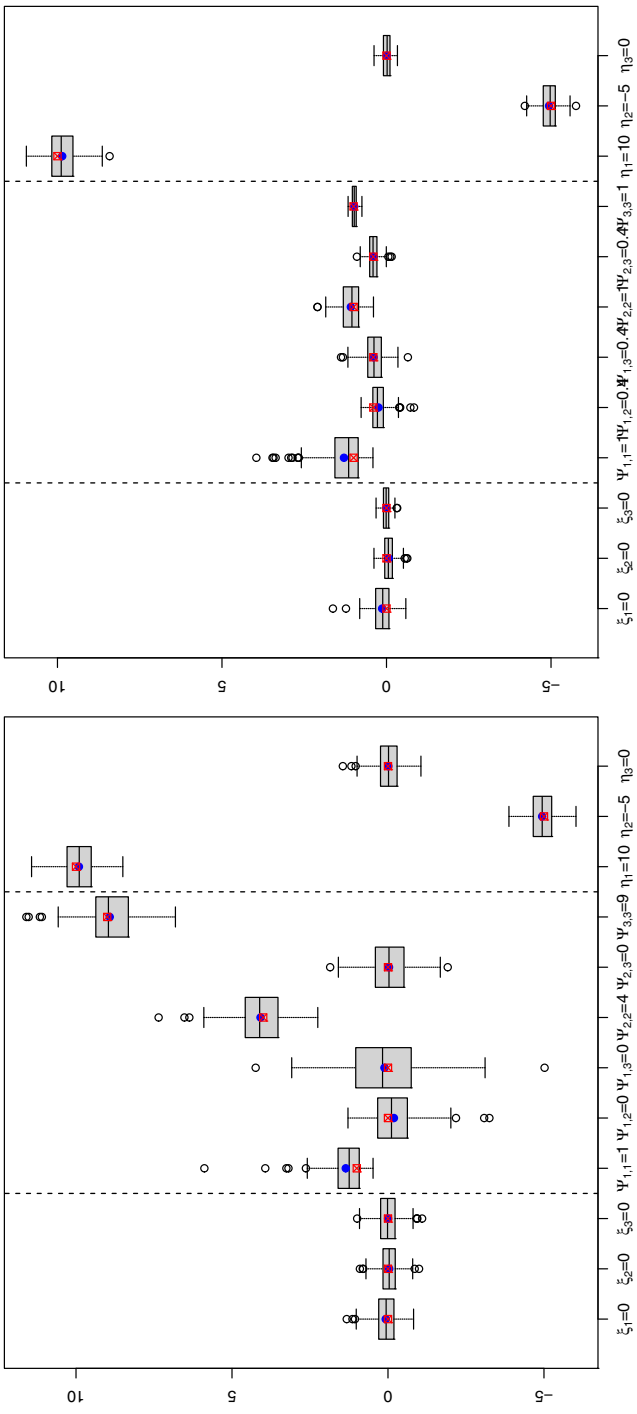
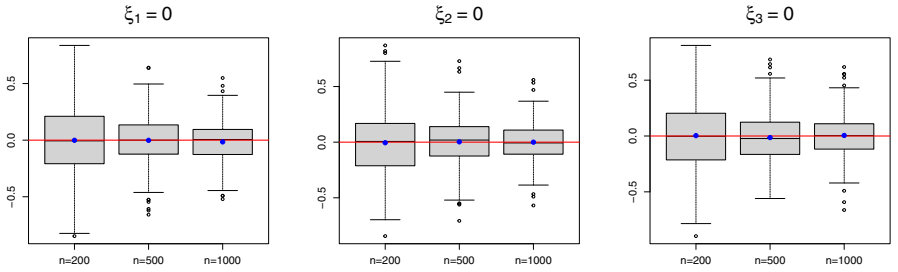
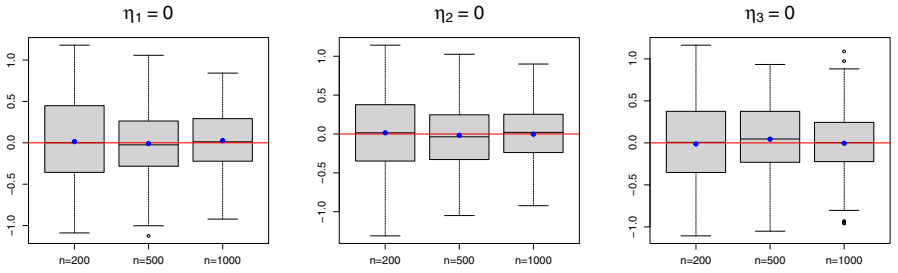


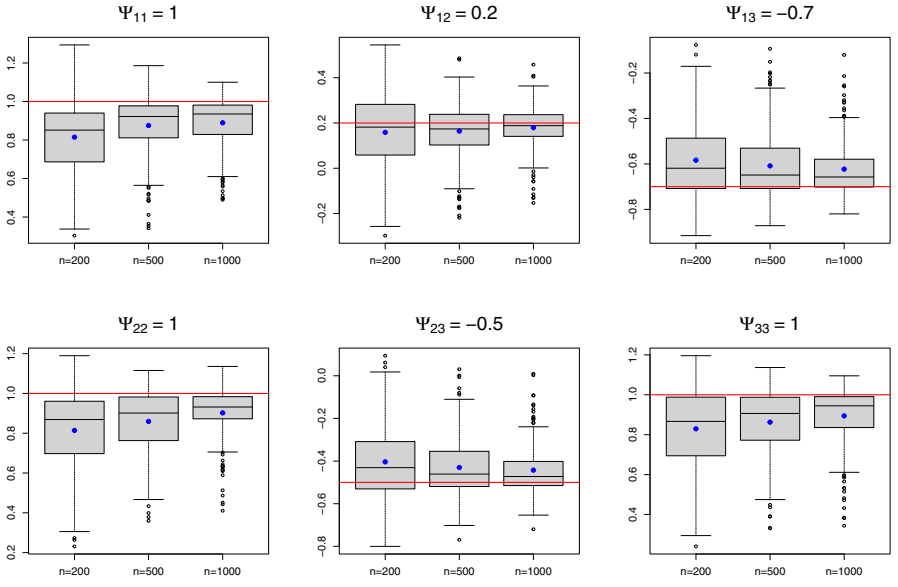
Fig. 4 Boxplots of the parameter estimates of a $\mathcal{MSN}_3$ distribution (a) with diagonal $\mathbf{\Psi}$; and (b) with equicorrelated $\mathbf{\Psi}$, estimated with the EM algorithm. The red square indicates the true parameter value, and the blue point indicates the mean of the estimated parameter values. (Color figure online)



(a) Boxplots of the estimates of $\boldsymbol{\xi}$

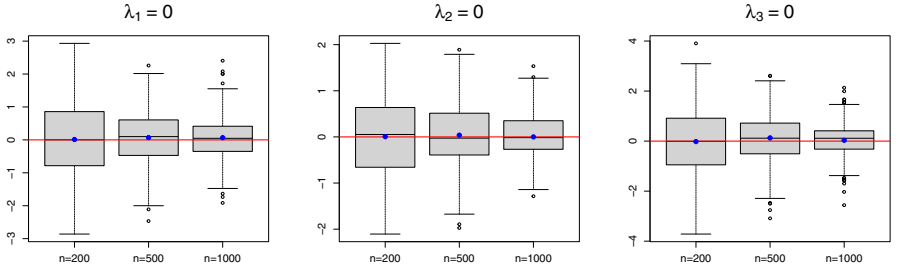


(b) Boxplots of the estimates of $\boldsymbol{\eta}$

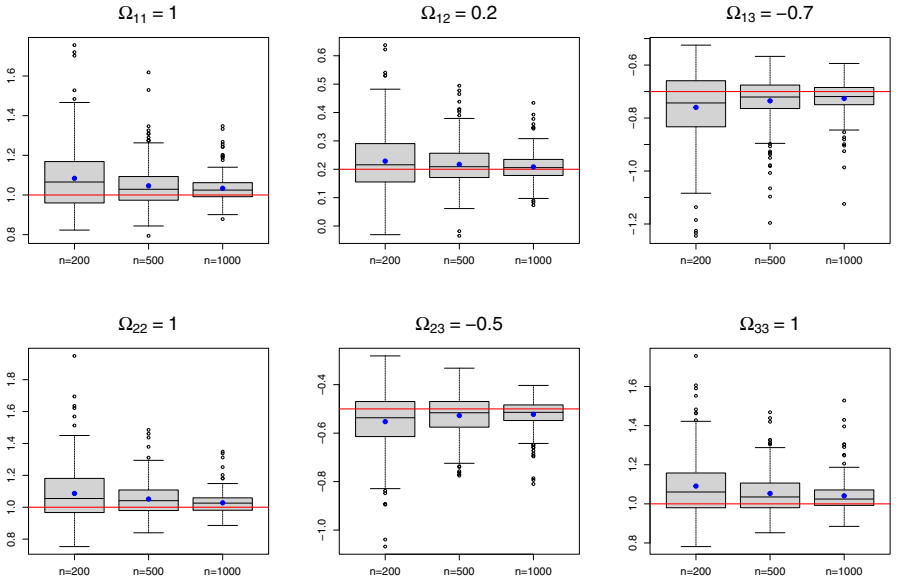


(c) Boxplots of the estimates of $\boldsymbol{\Psi}$

Fig. 5 Boxplots of the parameter estimates of a $\mathcal{MSN}_3$ distribution when $\boldsymbol{\eta}$ is exactly zero for different sample sizes n: (a) $\boldsymbol{\xi}$; (b) $\boldsymbol{\eta}$; (c) $\boldsymbol{\Psi}$. The red line indicates the true parameter value, and the blue point indicates the mean of the estimated parameter values. (Color figure online)



(a) Boxplots of the estimates of $\boldsymbol{\lambda}$



(b) Boxplots of the estimates of $\boldsymbol{\Omega}$

Fig. 6 Boxplots of the parameter estimates of a $\mathcal{MSN}_3$ distribution with $\boldsymbol{\xi}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\lambda}$ parameterization, when $\boldsymbol{\lambda}$ is exactly zero for different sample sizes n: (a) $\boldsymbol{\lambda}$; (b) $\boldsymbol{\Omega}$. The red line indicates the true parameter value, and the blue point indicates the mean of the estimated parameter values. (Color figure online)

erating the observations. We fit the $\mathcal{MSN}$ and the $\mathcal{ASN}$ models to the simulated observations from the four aforementioned models. We repeated the procedure 100 times for each model and each sample size. We compared the fitted $\mathcal{MSN}$ and $\mathcal{ASN}$ models by the AIC. In Table 1, we present the percentages of the experiments where the $\mathcal{MSN}$ model outperforms the $\mathcal{ASN}$ model for different models and different sample sizes.

The results in Table 1 show that the $\mathcal{ASN}$ works better mostly on the observations based on the Clayton copula as the sample size increases. We cannot make a similar statement for the other models. In fact, the $\mathcal{MSN}$ model is more appropriate for the simulated observations from the Gumbel model and the generalized hyperbolic model

Table 1 The percentages when the $\mathcal{MSN}$ model outperforms the $\mathcal{ASN}$ model in terms of AIC, when we fit the two models on simulated data from different models and for different sample sizes

Percentages when $\mathcal{MSN}$ Outperforms $\mathcal{ASN}$	Clayton copula	Frank copula	Gumbel copula	Generalized hyperbolic
$n = 50$	32	43	44	68
$n = 100$	50	51	67	91
$n = 200$	57	49	92	98
$n = 500$	37	46	98	99
$n = 1000$	33	54	99	100

for larger sample sizes. In conclusion, none of the $\mathcal{ASN}$ or $\mathcal{MSN}$ model is always better than the other.

6 Data applications

6.1 Univariate frontier data

We fit the $\mathcal{MSN}$ model to the frontier data from Azzalini and Capitanio (1999), which is a dataset consisting of 50 observations from $\mathcal{ASN}_1(0, 1, 5)$. An interesting feature of the dataset is that the maximum likelihood estimate of the skewness parameter for the $\mathcal{ASN}$ model is infinite. In Fig. 7a, we plot the fitted $\mathcal{MSN}$ density to the data. The estimated Ψ for the $\mathcal{MSN}$ model is 1.97×10^{-5}, and hence from Proposition 7, the fitted $\mathcal{MSN}$ density behaves essentially like a truncated normal density. This is also the case when the skewness parameter tends to infinity for the $\mathcal{ASN}$ model. Essentially, both the fitted $\mathcal{MSN}$ and $\mathcal{ASN}$ models for this dataset are the same. The main difference is that the profile log-likelihood function of the skewness parameter is monotonically increasing for the $\mathcal{ASN}$ model, whereas it is non-monotonic for the $\mathcal{MSN}$.

In Fig. 7b, the profile log-likelihood of the skewness parameter η for the $\mathcal{MSN}$ model is displayed, showing that it is, in fact, a non-monotonic function, attaining its maximum at a unique point. We conducted the testing for $H_0 : \eta = 0$ versus $H_1 : \eta \neq 0$ using the testing procedure discussed in Subsect. 4.2. The p-value for the test is 0, and, based on it, we reject H_0. This should be the case, as the fitted $\mathcal{MSN}$ model is basically a $\mathcal{TN}(-0.11, 1.24^2; (-0.11, \infty))$ model and is very far from a Gaussian model. The p-value of the LRT for the same test based on the $\mathcal{MSN}$ model is 1.3854×10^{-5}, hence it also rejects normality.

6.2 Trivariate wind speed data

We consider trivariate wind speed data produced by Yip (2018) with the Weather Research and Forecasting (WRF) model. The data consist of $n = 156$ wind speed vectors of dimension $p = 3$ representing bi-weekly mid-day windspeed during the period 2009–2014 at three locations near the wind farm Dumat Al Jandal currently

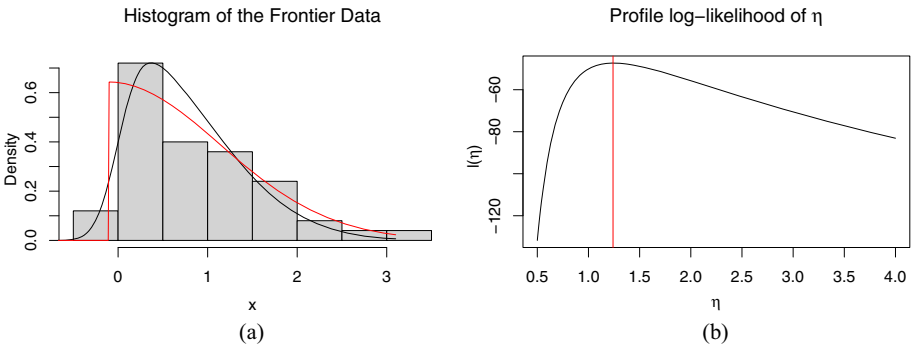


Fig. 7 (a) The red curve is the fitted $\mathcal{MSN}(-0.11, 1.97 \times 10^{-5}, 1.24)$ density function to the frontier data (histogram); the black curve is the original density function that generated the data. (b) The profile log-likelihood of η for the frontier data at the point $\widehat{\xi} = -0.11$ and $\widehat{\Psi} = 1.97 \times 10^{-5}$; the red line corresponds to $\eta = 1.24$ where the profile log-likelihood of η is maximized . (Color figure online)

Table 2 The p-values of different tests for testing the marginal Gaussianity of wind vectors

Test	Var 1	Var 2	Var 3
LRT ($\mathcal{MSN}$)	0.471	0.953	0.959
Wald-type test ($\mathcal{MSN}$)	0.320	0.740	0.789
Skewness test	0.301	0.418	0.276
Kurtosis test	0.871	0.825	0.868

under construction in Saudi Arabia. There is no serial dependence according to a Ljung-Box test, therefore the data are treated as a random sample from a trivariate distribution.

It is important to study the distribution of this trivariate wind speed vector because it is crucial for understanding wind patterns that will influence the energy production by the nearby wind farm. In particular, it is of interest to assess whether a non-Gaussian distribution is needed instead of a common Gaussian model. We test the univariate Gaussianity of each of the three variables using the likelihood ratio test (LRT) and the Wald-type testing procedure described in Subsect 4.2 based on the $\mathcal{MSN}$ models. The p-values for all the tests for all the variables are reported in Table 2. Based on the p-values we do not reject the marginal Gaussianity for all the three variables at 5% level. In Table 2 we also report the p-values of the skewness and kurtosis test for each of the three variables. At the 5% level, marginal skewness and kurtosis tests do not reject Gaussianity, but Mardia's tests of multivariate skewness and kurtosis reject a trivariate Gaussian distribution with respective p-values 1.858×10^{-3} and 2.220×10^{-16}.

We fit a trivariate $\mathcal{MSN}$ distribution to this data, using the suggested EM algorithm for parameter estimation. In Fig. 8, the bivariate contour plots for the fitted model are presented. We also fitted the $\mathcal{ASN}$ distribution to this data. The optimized log-likelihood value for the $\mathcal{MSN}$ model is -809.76, and the optimized log-likelihood value for the $\mathcal{ASN}$ model is -818.81. Since both models have the same number of parameters, their optimized log-likelihood values are essentially their AIC. In terms of AIC, the $\mathcal{MSN}$ model performs better on this dataset than the $\mathcal{ASN}$ model. From the

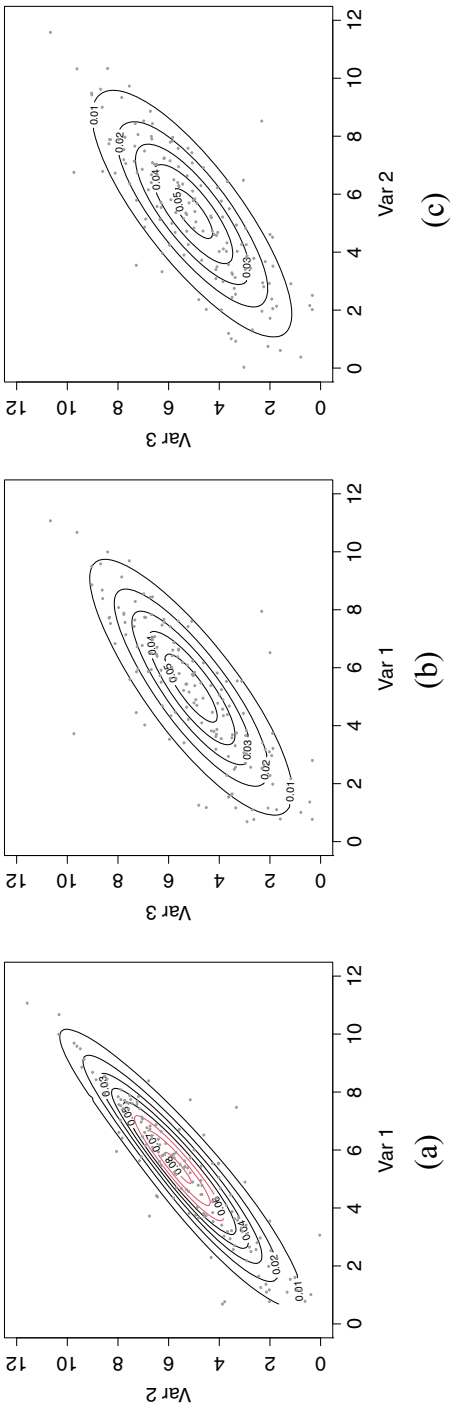


Fig. 8 Contours levels of $\mathcal{MSN}_3$ distribution fitted to wind speed data: (a) Var 1 × Var 2, (b) Var 1 × Var 3, and (c) Var 2 × Var 3

bivariate contours one can suspect that the data are from a trivariate normal distribution. We tested the hypothesis $H_0 : \boldsymbol{\eta} = (0, 0, 0)^\top$ versus $H_1 : \boldsymbol{\eta} \neq (0, 0, 0)^\top$ from the $\mathcal{MSN}$ model using the testing procedure discussed in Subsect 4.2. The p-value is 0.00545, and based on it we reject the hypothesis H_0, that the wind speed data are coming from a trivariate Gaussian model. We also tested the same hypothesis with the LRT based on the $\mathcal{MSN}$ model and the p-value obtained is 0.00019. Hence, we reject the hypothesis that a Gaussian model is suitable.

7 Discussion

In this article, we have introduced a new multivariate version of the modified skew-normal distribution, which is comparable to the multivariate Azzalini skew-normal distribution. It has a finite likelihood function and nonsingular Fisher information matrix when the skewness parameter is zero, unlike the $\mathcal{ASN}$ family. We discussed some of the basic probabilistic properties of the suggested distribution. Despite good properties, the $\mathcal{MSN}$, similar to the $\mathcal{ASN}$, cannot model heavy-tailed data. One future research direction is to study other skewed multivariate distributions such as the skew-t with this parameterization and with the skewing function of the $\mathcal{MSN}$ distribution.

The main advantages of the proposed $\mathcal{MSN}$ distribution over the $\mathcal{ASN}$ distribution are that the $\mathcal{MSN}$ solves the issue of the singularity of the Fisher information matrix when the skewness parameter is zero, and also that its parameterization ensures the MLE of the skewness parameter is always finite. Hence, the $\mathcal{MSN}$ is more suitable for statistical works related to inferring the skewness parameter near zero. However, there are no analytic expressions for the expectation, covariance, skewness, co-skewness, and kurtosis for the $\mathcal{MSN}$ distribution, unlike the $\mathcal{ASN}$ distribution. Moreover, the $\mathcal{MSN}$ distribution is not closed under marginalization and hence not closed under conditioning, which is not the case for the $\mathcal{ASN}$ distribution. This means that the $\mathcal{ASN}$ distribution should be preferred for formal purposes like regression modeling, time-series analysis, spatial modeling, over the $\mathcal{MSN}$ distribution.

We have also provided an EM algorithm for the $\mathcal{MSN}$ family. One issue regarding the EM algorithm is that, as we numerically optimize in each iteration of the maximization step, the algorithm is not efficient when the dimension p of the problem is high. Some other hierarchical representations may resolve this problem. Another research avenue is to improve the initial estimates for the EM algorithm. The current initial estimates of the parameters are naive and may fail when the skewness in the data is very high.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00362-023-01397-1.

Appendix A

We discuss the benefits of using the alternative parameterization for defining the $\mathcal{SN}$ distribution compared to the $\mathcal{ASN}$ distribution. We start with a preliminary stochastic representation of the $\mathcal{SN}$ distribution, from which we can derive most of its main basic properties.

Proposition 10 (Stochastic representation of $\mathcal{SN}$ distribution) *If* $\boldsymbol{X} \sim \mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$, *then* $\boldsymbol{X} \stackrel{d}{=} \boldsymbol{\xi} + T\boldsymbol{\eta} + \boldsymbol{V}$, *where* T *and* $\boldsymbol{V}$ *are independently distributed, with half-normal* T *denoted by* $T \sim \mathcal{HN}(0, 1)$, *and* $\boldsymbol{V} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Psi})$.

Proof Let $\tilde{\boldsymbol{X}} = \boldsymbol{\xi} + T\boldsymbol{\eta} + \boldsymbol{V}$. Since the conditional mpdf of $\tilde{\boldsymbol{X}}|T = t$ is $f_{\tilde{\boldsymbol{X}}|T=t}(\boldsymbol{x}) = \phi_p(\boldsymbol{x}; \boldsymbol{\xi} + t\boldsymbol{\eta}, \boldsymbol{\Psi})$ for $t > 0$, and T has marginal density $f_T(t) = 2\phi(t)I_{(t>0)}$, then for the mpdf of $\tilde{\boldsymbol{X}}$ we have

$$\begin{aligned} f_{\tilde{\boldsymbol{X}}}(\boldsymbol{x}) &= \int_0^\infty f_{\tilde{\boldsymbol{X}}|T=t}(\boldsymbol{x}) f_T(t)\mathrm{d}t = 2\int_0^\infty \phi_p(\boldsymbol{x}; \boldsymbol{\xi} + t\boldsymbol{\eta}, \boldsymbol{\Psi})\phi(t)\mathrm{d}t \\ &= 2\phi_p\left(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top\right) \int_0^\infty \phi\{t; \boldsymbol{\eta}^\top(\boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}(\boldsymbol{x} - \boldsymbol{\xi}), \\ &\quad 1 - \boldsymbol{\eta}^\top(\boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}\boldsymbol{\eta}\}\mathrm{d}t \\ &= 2\phi_p(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)\Phi\left\{\frac{\boldsymbol{\eta}^\top(\boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}(\boldsymbol{x} - \boldsymbol{\xi})}{\sqrt{1 - \boldsymbol{\eta}^\top(\boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}\boldsymbol{\eta}}}\right\}, \quad \boldsymbol{x} \in \mathbb{R}^p, \end{aligned} \tag{15}$$

where we have used the identity (see Lemma 2 in Arellano-Valle et al. (2005))

$$\begin{aligned} \phi_p(\boldsymbol{x}, \boldsymbol{\xi} + t\boldsymbol{\eta}, \boldsymbol{\Psi})\phi(t) &= \phi_p(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top) \\ &\quad \times \phi\{t; \boldsymbol{\eta}^\top(\boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}(\boldsymbol{x} - \boldsymbol{\xi}), 1 - \boldsymbol{\eta}^\top(\boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}\boldsymbol{\eta}\}. \end{aligned}$$

Finally, using the following result:

$$\begin{aligned} (\boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1} &= \boldsymbol{\Psi}^{-1} - \frac{\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}}{1 + \boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}} \Rightarrow \\ \frac{\boldsymbol{\eta}^\top(\boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}(\boldsymbol{x} - \boldsymbol{\xi})}{\sqrt{1 - \boldsymbol{\eta}^\top(\boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}\boldsymbol{\eta}}} &= \frac{\boldsymbol{\eta}^\top\Psi^{-1}(\boldsymbol{x} - \boldsymbol{\xi})}{\sqrt{1 + \boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}}, \end{aligned}$$

it can be easily established that $\tilde{\boldsymbol{X}} \sim \mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$. □

Further basic properties As immediate consequences of the above stochastic representation of a random vector $\boldsymbol{X} \sim \mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$, we have the following basic properties:
1) Expectation and covariance: The mean vector and covariance matrix of $\boldsymbol{X}$ are

$$\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\xi} + \sqrt{\frac{2}{\pi}}\boldsymbol{\eta} \quad \text{and} \quad \mathbb{V}(\boldsymbol{X}) = \boldsymbol{\Psi} + \left(1 - \frac{2}{\pi}\right)\boldsymbol{\eta}\boldsymbol{\eta}^\top. \tag{16}$$

2) Distribution of an affine transformation: For any fixed vector $\boldsymbol{a} \in \mathbb{R}^q$ and any fixed matrix $\boldsymbol{B} \in \mathbb{R}^{q\times p}$ of full row rank and $q \leq p$, we have $\boldsymbol{a} + \boldsymbol{B}\boldsymbol{X} \stackrel{d}{=} \boldsymbol{a} + \boldsymbol{B}\boldsymbol{\xi} + T\boldsymbol{B}\boldsymbol{\eta} + \boldsymbol{B}\boldsymbol{V} \sim \mathcal{SN}_q(\boldsymbol{a}+\boldsymbol{B}\boldsymbol{\xi}, \boldsymbol{B}\boldsymbol{\Psi}\boldsymbol{B}^\top, \boldsymbol{B}\boldsymbol{\eta})$, since, by assumption, T and $\boldsymbol{V}$ are independently distributed, with $T \sim \mathcal{HN}(0, 1)$ and $\boldsymbol{V} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Psi})$.
3) Marginal distributions: Partition now $\boldsymbol{X}$ in two sub-vectors of sizes p_1 and p_2 such that $p_1 + p_2 = p$, with corresponding partitions of the parameters in blocks of matching sizes, as follows

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \quad \boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{pmatrix}, \quad \boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Psi}_{11} & \boldsymbol{\Psi}_{12} \\ \boldsymbol{\Psi}_{21} & \boldsymbol{\Psi}_{22} \end{pmatrix}, \quad \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{pmatrix}. \tag{17}$$

Thus, by using property 2) with $\boldsymbol{a} = \boldsymbol{0}$ and $\boldsymbol{B} = (\mathrm{I}_{p_1}, \boldsymbol{0})$ for $\boldsymbol{X}_1$ and $\boldsymbol{B} = (\boldsymbol{0}, \mathrm{I}_{p_2})$ for $\boldsymbol{X}_2$ it follows for their respective marginals that $\boldsymbol{X}_1 \sim \mathcal{SN}_{p_1}(\boldsymbol{\xi}_1, \boldsymbol{\Psi}_{11}, \boldsymbol{\eta}_1)$ and $\boldsymbol{X}_2 \sim \mathcal{SN}_{p_2}(\boldsymbol{\xi}_2, \boldsymbol{\Psi}_{22}, \boldsymbol{\eta}_2)$.
4) Moment generating function: The multivariate moment generating function (mmgf) of the $\mathcal{SN}$ distribution can be derived in closed form. We present the mmgf of the of $\mathcal{SN}$ distribution in the next proposition.

Proposition 11 *The mmgf of* $\boldsymbol{X} \sim \mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ *is*

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = 2\exp\left(\boldsymbol{t}^\top\boldsymbol{\xi} + \frac{1}{2}\boldsymbol{t}^\top\boldsymbol{\Omega}\boldsymbol{t}\right)\Phi(\boldsymbol{\eta}^\top\boldsymbol{t}), \quad \boldsymbol{t} \in \mathbb{R}^p,$$

where $\boldsymbol{\Omega} = \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top$.

Proof The mpdf of $\boldsymbol{X} \sim \mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ in (15) can be rewritten as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{2}{|\boldsymbol{\Omega}|^{1/2}}\phi_p\left\{\boldsymbol{\Omega}^{-1/2}(\boldsymbol{x}-\boldsymbol{\xi})\right\}\Phi\left\{\boldsymbol{\gamma}^\top\boldsymbol{\Omega}^{-1/2}(\boldsymbol{x}-\boldsymbol{\xi})\right\}, \quad \boldsymbol{x} \in \mathbb{R}^p, \tag{18}$$

where $\phi_p(z) = \phi_p(z; \boldsymbol{0}, \mathrm{I}_p)$, $\boldsymbol{\Omega} = \boldsymbol{\Psi}+\boldsymbol{\eta}\boldsymbol{\eta}^\top$, $\boldsymbol{\gamma} = (1-\boldsymbol{\eta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\eta})^{-1/2}\boldsymbol{\Omega}^{-1/2}\boldsymbol{\eta}$, and, conversely, we have $\boldsymbol{\eta} = (1+\boldsymbol{\gamma}^\top\boldsymbol{\gamma})^{-1/2}\boldsymbol{\Omega}^{1/2}\boldsymbol{\gamma}$ and $\boldsymbol{\Psi} = \boldsymbol{\Omega}-(1+\boldsymbol{\gamma}^\top\boldsymbol{\gamma})^{-1}\boldsymbol{\Omega}^{1/2}\boldsymbol{\gamma}\boldsymbol{\gamma}^\top\boldsymbol{\Omega}^{1/2}$. From Eq. (18) and by using the change of variable $z = \boldsymbol{\Omega}^{-1/2}(\boldsymbol{x}-\boldsymbol{\xi})$ we have that the mmgf of $\boldsymbol{X}$, $M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\{\exp(\boldsymbol{t}^\top\boldsymbol{X})\}$, is given by

$$\begin{aligned}
M_{\boldsymbol{X}}(\boldsymbol{t}) &= \int_{\mathbb{R}^p}\exp(\boldsymbol{t}^\top\boldsymbol{x})2\phi_p(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Omega})\Phi\{\boldsymbol{\gamma}^\top\boldsymbol{\Omega}^{-1/2}(\boldsymbol{x}-\boldsymbol{\xi})\}\mathrm{d}\boldsymbol{x} \\
&= 2\exp(\boldsymbol{t}^\top\boldsymbol{\xi})\int_{\mathbb{R}^p}\exp(\boldsymbol{s}^\top z)\phi_p(z; \boldsymbol{0}, \mathbf{I}_p)\Phi\{\boldsymbol{\gamma}^\top z\}\mathrm{d}z, \quad (\boldsymbol{s} = \boldsymbol{\Omega}^{1/2}\boldsymbol{t}) \\
&= 2\exp(\boldsymbol{t}^\top\boldsymbol{\xi} + \frac{1}{2}\boldsymbol{s}^\top\boldsymbol{s})\int_{\mathbb{R}^p}\phi_p(z; \boldsymbol{s}, \mathbf{I}_p)\Phi\{\boldsymbol{\gamma}^\top z\}\mathrm{d}z \\
&= 2\exp\left(\boldsymbol{t}^\top\boldsymbol{\xi} + \frac{1}{2}\boldsymbol{s}^\top\boldsymbol{s}\right)\mathbb{E}\{\Phi(\boldsymbol{\gamma}^\top\boldsymbol{Z})\}, \quad \boldsymbol{\gamma}^\top\boldsymbol{Z} \sim \mathcal{N}_1(\boldsymbol{\gamma}^\top\boldsymbol{s}, \boldsymbol{\gamma}^\top\boldsymbol{\gamma}) \\
&= 2\exp\left(\boldsymbol{t}^\top\boldsymbol{\xi} + \frac{1}{2}\boldsymbol{s}^\top\boldsymbol{s}\right)\Phi\left(\frac{\boldsymbol{\gamma}^\top\boldsymbol{s}}{\sqrt{1+\boldsymbol{\gamma}^\top\boldsymbol{\gamma}}}\right),
\end{aligned}$$

$$= 2\exp\left(\boldsymbol{t}^\top\boldsymbol{\xi} + \frac{1}{2}\boldsymbol{t}^\top\boldsymbol{\Omega}\boldsymbol{t}\right)\Phi(\boldsymbol{\eta}^\top\boldsymbol{t}).$$

□

5) Cumulative distribution function: In the next proposition we present the exact functional form of the multivariate cumulative distribution function (mcdf) of the $\mathcal{SN}$ distribution.

Proposition 12 *The mcdf of* $\boldsymbol{X} \sim \mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ *is*

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = 2\Phi_{p+1}(\boldsymbol{x}_*; \boldsymbol{\xi}_*, \boldsymbol{\Omega}_*), \quad \textit{with} \quad \boldsymbol{x}_* = (\boldsymbol{x}^\top, 0)^\top, \quad \boldsymbol{x} \in \mathbb{R}^p,$$

where, $\boldsymbol{\xi}_* = (\boldsymbol{\xi}^\top, 0)^\top$ *and* $\boldsymbol{\Omega}_* = \begin{pmatrix} \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top & -\boldsymbol{\eta} \\ -\boldsymbol{\eta}^\top & 1 \end{pmatrix}$.

Proof The mpdf of $\boldsymbol{X}$ is

$$\begin{aligned} f_{\boldsymbol{X}}(z) &= 2\phi_p(z; \boldsymbol{\xi}, \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)\Phi\left\{\frac{\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}(z-\boldsymbol{\xi})}{\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}}\right\} \\ &= 2\phi_p(z; \boldsymbol{\xi}, \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)\int_{-\infty}^{\frac{\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}(z-\boldsymbol{\xi})}{\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}}} \phi(u; 0, 1)\mathrm{d}u \\ &= 2\phi_p(z; \boldsymbol{\xi}, \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)\int_{-\infty}^{0} \phi\left(z_0 + \frac{\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}(z-\boldsymbol{\xi})}{\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}}; 0, 1\right)\mathrm{d}z_0 \\ &= 2\phi_p(z; \boldsymbol{\xi}, \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)\int_{-\infty}^{0} \phi\left(z_0; -\frac{\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}(z-\boldsymbol{\xi})}{\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}}, 1\right)\mathrm{d}z_0 \end{aligned}$$

where we use the change of variable $z_0 = u - \boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}(z-\boldsymbol{\xi})/\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}$. Now, from the marginal-conditional factorization of a $(p+1)$-variate normal mpdf we can write

$$\phi_p(z; \boldsymbol{\xi}, \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)\phi\left(z_0; -\frac{\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}(z-\boldsymbol{\xi})}{\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}}, 1\right) = \phi_{p+1}(z_*; \boldsymbol{\xi}_*, \boldsymbol{\Omega}_{**}),$$

where $z_* = (z^\top, z_0)^\top$ and $\boldsymbol{\Omega}_{**} = \begin{pmatrix} \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top & -\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}}\boldsymbol{\eta}\boldsymbol{\eta} \\ -\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}}\boldsymbol{\eta}\boldsymbol{\eta}^\top & 1 \end{pmatrix}$. Moreover, we have $\boldsymbol{\Omega}_{**} = \boldsymbol{D}_*\boldsymbol{\Omega}_*\boldsymbol{D}_*$, with $D_* = \mathrm{diag}\left(\mathbf{I}_p, \sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}\right)$. Then, the mcdf of $\boldsymbol{X}$ is

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = 2\int_{(-\infty,\boldsymbol{x}]} \phi_p(z; \boldsymbol{\xi}, \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)\Phi\left\{\frac{\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}(z-\boldsymbol{\xi})}{\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}}\right\}\mathrm{d}z$$

$$
\begin{aligned}
&= 2\int_{(-\infty,\boldsymbol{x}]}\int_{-\infty}^{0}\phi_p(z;\boldsymbol{\xi},\boldsymbol{\Psi}+\boldsymbol{\eta}\boldsymbol{\eta}^\top)\phi\left(z_0;-\frac{\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}(z-\boldsymbol{\xi})}{\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}},1\right)\mathrm{d}z_0\mathrm{d}z\\
&= 2\int_{(-\infty,\boldsymbol{x}_*]}\phi_{p+1}(z_*;\boldsymbol{\xi}_*,\boldsymbol{\Omega}_{**})\mathrm{d}z_*\\
&= 2\Phi_{p+1}(\boldsymbol{x}_*;\boldsymbol{\xi}_*,\boldsymbol{\Omega}_{**}),\quad \text{with}\quad \boldsymbol{x}_*=(\boldsymbol{x}^\top,0)^\top,\quad \boldsymbol{x}\in\mathbb{R}^p.
\end{aligned}
$$

But, since $\boldsymbol{x}_*-\boldsymbol{\xi}_*=((\boldsymbol{x}-\boldsymbol{\xi})^\top,0)^\top$, then $\boldsymbol{D}_*^{-1}(\boldsymbol{x}_*-\boldsymbol{\xi}_*)=\boldsymbol{x}_*-\boldsymbol{\xi}_*$ and so

$$
\begin{aligned}
\Phi_{p+1}(\boldsymbol{x}_*;\boldsymbol{\xi}_*,\boldsymbol{\Omega}_{**}) &= \Phi_{p+1}(\boldsymbol{x}_*;\boldsymbol{\xi}_*,\boldsymbol{D}_*\boldsymbol{\Omega}_*\boldsymbol{D}_*)=\Phi_{p+1}(\boldsymbol{D}_*^{-1}(\boldsymbol{x}_*-\boldsymbol{\xi}_*);\boldsymbol{0},\boldsymbol{\Omega}_*)\\
&= \Phi_{p+1}(\boldsymbol{x}_*;\boldsymbol{\xi}_*,\boldsymbol{\Omega}_*).
\end{aligned}
$$

□

Behavior of the likelihood function In the following two results, we show that the alternative parameterization used in (3) for defining the $\mathcal{SN}$ mpdf fixes the problem of the infinite maximum likelihood estimate of the skewness parameter, unlike the original parameterization used in (2) for defining the $\mathcal{ASN}$ mpdf, but it also fails to resolve the problem of the singular Fisher information matrix. In fact, let $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ be an observed random sample from $\mathcal{SN}_p(\boldsymbol{\xi},\boldsymbol{\Psi},\boldsymbol{\eta})$. The corresponding likelihood function for $(\boldsymbol{\xi},\boldsymbol{\Psi},\boldsymbol{\eta})$ is

$$
L(\boldsymbol{\xi},\boldsymbol{\Psi},\boldsymbol{\eta})=\prod_{i=1}^{n}2\phi_p\left(\boldsymbol{x}_i;\boldsymbol{\xi},\boldsymbol{\Psi}+\boldsymbol{\eta}\boldsymbol{\eta}^\top\right)\Phi\left\{\frac{\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi})}{\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}}\right\}. \tag{19}
$$

For fixed $\boldsymbol{\xi}$ and $\boldsymbol{\Psi}$, (19) becomes the profile likelihood function of $\boldsymbol{\eta}$, which we denote by $L(\boldsymbol{\eta})$, $\boldsymbol{\eta}\in\mathbb{R}^p$.

Proposition 13 *The maximum likelihood estimator of the skewness parameter $\boldsymbol{\eta}$ of the $\mathcal{SN}_p(\boldsymbol{\xi},\boldsymbol{\Psi},\boldsymbol{\eta})$ family is always finite.*

Proof We find the limit of $L(\boldsymbol{\eta})$ when some (or all) components of $\boldsymbol{\eta}$ tend to $+\infty$ or $-\infty$. For this, first note from the Cauchy–Schwarz inequality that

$$
\begin{aligned}
&\frac{|\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi})|}{\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}}\\
&\le\sqrt{\frac{\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}}\sqrt{(\boldsymbol{x}_i-\boldsymbol{\xi})^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi})}\le\sqrt{(\boldsymbol{x}_i-\boldsymbol{\xi})^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi})},\quad\forall\,\boldsymbol{\eta}\in\mathbb{R}^p,
\end{aligned}
$$

for $i=1,\ldots,n$. From this inequality, it clearly follows, for all $\boldsymbol{\eta}\in\mathbb{R}^p$ and each $i=1,\ldots,n$, that

$$
\phi_p(\boldsymbol{x}_i;\boldsymbol{\xi},\boldsymbol{\Psi}+\boldsymbol{\eta}\boldsymbol{\eta}^\top)=\frac{\exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\xi})^\top(\boldsymbol{\Psi}+\boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi})\right\}}{(2\pi)^{p/2}|\boldsymbol{\Psi}|^{1/2}\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}}
$$

$$
\begin{aligned}
&= \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\xi})^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi})\right\}}{(2\pi)^{p/2}|\boldsymbol{\Psi}|^{1/2}} \\
&\quad\times\frac{\exp\left[-\frac{1}{2}\frac{\{\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi})\}^2}{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}\right]}{\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}} \\
&\le \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Psi}|^{1/2}\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}},
\end{aligned}
$$

and

$$
0 \le \Phi\left\{-\sqrt{(\boldsymbol{x}_i-\boldsymbol{\xi})^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi})}\right\} \le \Phi\left\{\frac{\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi})}{\sqrt{1+\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}}\right\} \le \Phi\left\{\sqrt{(\boldsymbol{x}_i-\boldsymbol{\xi})^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i-\boldsymbol{\xi})}\right\} \le 1.
$$

These results hold whatever fixed value of $(\boldsymbol{\xi}, \boldsymbol{\Psi})$. Thus, noting also that whenever some (or all) components of $\boldsymbol{\eta}$ tend to $\pm\infty$, then $\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta} \to \infty$, we can easily deduce from the first of the previous inequalities that $\phi_p(\boldsymbol{x}_i; \boldsymbol{\xi}, \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top) \to 0$ for each $i = 1, \ldots, n$ as some (or all) components of $\boldsymbol{\eta}$ tend to $\pm\infty$. Thus, now taking into account the second inequality, we find that $L(\boldsymbol{\eta}) \to 0$ whenever some (or all) components of $\boldsymbol{\eta}$ tend to $\pm\infty$ and whatever fixed value of $(\boldsymbol{\xi}, \boldsymbol{\Psi})$. This result leads us to the conclusion that $L(\boldsymbol{\eta})$ is not a monotonically increasing or decreasing function of any of the components of $\boldsymbol{\eta}$. This means that the profile likelihood of the skewness parameter $\boldsymbol{\eta}$ is always maximized at a finite point for the $\mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ family. □

Remark Proposition 13 shows that the MLE of the skewness parameter of the $\mathcal{SN}$ distribution is always finite for the new parameterization. It was not the case for the $\mathcal{ASN}$ parameterization. So, if we use the new parameterization for a particular data, then we will get a finite MLE of the skewness parameter, whereas we might get infinite MLE for the skewness parameter in the $\mathcal{ASN}$ parameterization. Because these two parameterizations are a one-to-one transformation of each other, due to the invariance property of the MLE, if we transform back the MLE from the new-parameterization to the $\mathcal{ASN}$ parameterization, then we will get back the old results. In other words, although the MLE of the skewness parameter $\boldsymbol{\eta}$ is always finite, it may sometimes correspond to an MLE of the skewness parameter $\boldsymbol{\alpha}$ that is infinite.

In what follows, we use the notation of Magnus and Neudecker (1979) related to the Kronecker product and matrix vectorization. For instance, let $\text{vech}(\boldsymbol{\Psi})$ be the $p(p+1)/2$-subvector of $\text{vec}(\boldsymbol{\Psi})$, where only upper-diagonal entries of $\boldsymbol{\Psi}$ are considered. Also, let $\boldsymbol{K}_p$ be the $p^2 \times p^2$ commutation matrix, i.e., $\boldsymbol{K}_p\text{vec}(\boldsymbol{A}) = \text{vec}(\boldsymbol{A}^\top)$ for any $p \times q$ matrix $\boldsymbol{A}$, and let $\boldsymbol{D}_p$ be the $p^2 \times p(p+1)/2$ duplication matrix, i.e., $\boldsymbol{D}_p\text{vech}(\boldsymbol{A}) = \text{vec}(\boldsymbol{A})$ for any $p \times p$ symmetric matrix $\boldsymbol{A}$.

Proposition 14 *The Fisher information matrix of the $\mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ family is singular when the skewness parameter $\boldsymbol{\eta}$ is set to zero.*

Proof The score vector and Fisher information matrix for the $\mathcal{ASN}_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha})$ family are derived by Arellano-Valle and Azzalini (2008) in terms of the reparametrization

$(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = \boldsymbol{\omega}^{-1}\boldsymbol{\alpha}$. These author also showed that the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$ is singular at $\boldsymbol{\lambda} = \mathbf{0}$:

$$\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \mathbf{0}) = \begin{bmatrix} \boldsymbol{\omega}^{-1} & \mathbf{0} & \sqrt{\frac{2}{\pi}}\mathrm{I}_p \\ \mathbf{0} & \frac{1}{2}\boldsymbol{D}_p^\top(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1})\boldsymbol{D}_p & \mathbf{0} \\ \sqrt{\frac{2}{\pi}}\mathrm{I}_p & \mathbf{0} & \frac{2}{\pi}\boldsymbol{\Omega} \end{bmatrix}.$$

Since the $\mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ family corresponds to a reparametrization of the $\mathcal{ASN}_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha})$ family, its Fisher information matrix becomes $\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta}) = \boldsymbol{J}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})^\top \boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})\, \boldsymbol{J}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$, where $\boldsymbol{J}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$ denotes the Jacobian matrix of the transformation from $(\boldsymbol{\xi}, \mathrm{vech}(\boldsymbol{\Omega}), \boldsymbol{\lambda})$ to $(\boldsymbol{\xi}, \mathrm{vech}(\boldsymbol{\Psi}), \boldsymbol{\eta})$. Thus, since the inverse transformation from $(\boldsymbol{\xi}, \mathrm{vech}(\boldsymbol{\Psi}), \boldsymbol{\eta})$ to $(\boldsymbol{\xi}, \mathrm{vech}(\boldsymbol{\Omega}), \boldsymbol{\lambda})$ turns out to be $\boldsymbol{\xi} = \boldsymbol{\xi}$, $\boldsymbol{\Omega} = \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top$ and $\boldsymbol{\lambda} = (1 + \boldsymbol{\eta}\boldsymbol{\Psi}^{-1}\boldsymbol{\eta})^{-1/2}\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}$, for the Jacobian matrix we have

$$\boldsymbol{J}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda}) = \begin{bmatrix} \frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\xi}^\top} & \frac{\partial \mathrm{vec}(\boldsymbol{\Omega})}{\partial \boldsymbol{\xi}^\top} & \frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\xi}^\top} \\ \frac{\partial \boldsymbol{\xi}}{\partial \mathrm{vec}(\boldsymbol{\Psi})^\top} & \frac{\partial \mathrm{vec}(\boldsymbol{\Omega})}{\partial \mathrm{vec}(\boldsymbol{\Psi})^\top} & \frac{\partial \boldsymbol{\lambda}}{\partial \mathrm{vec}(\boldsymbol{\Psi})^\top} \\ \frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\eta}^\top} & \frac{\partial \mathrm{vec}(\boldsymbol{\Omega})}{\partial \boldsymbol{\eta}^\top} & \frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\eta}^\top} \end{bmatrix} = \begin{bmatrix} \mathrm{I}_p & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{I}_{p(p+1)/2} & \boldsymbol{J}_{23} \\ \mathbf{0} & \boldsymbol{J}_{32} & \boldsymbol{J}_{33} \end{bmatrix},$$

where $\boldsymbol{J}_{23} = \boldsymbol{D}_p^+(\mathrm{I}_p \otimes \boldsymbol{\eta} + \boldsymbol{\eta} \otimes \mathrm{I}_p)$, with $\boldsymbol{D}_p^+ = (\boldsymbol{D}_p^\top \boldsymbol{D}_p)^{-1}\boldsymbol{D}_p^\top$, $\boldsymbol{J}_{32} = (1 + \boldsymbol{\eta}\boldsymbol{\Psi}^{-1}\boldsymbol{\eta})^{-1/2}\{\frac{1}{2}(\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1} \otimes \boldsymbol{\Psi}^{-1}\boldsymbol{\eta}\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}) - (\boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1} \otimes \boldsymbol{\Psi}^{-1})\}\boldsymbol{D}_p$ and $\boldsymbol{J}_{33} = (1 + \boldsymbol{\eta}\boldsymbol{\Psi}^{-1}\boldsymbol{\eta})^{-1/2}(\boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}$. When $\boldsymbol{\eta} = \mathbf{0}$ we have that $\boldsymbol{\lambda} = \mathbf{0}$, $\boldsymbol{\Omega} = \boldsymbol{\Psi}$ and the Jacobian matrix $\boldsymbol{J}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \mathbf{0}) = \mathrm{diag}(\mathrm{I}_p, \mathrm{I}_{p(p+1)/2}, \boldsymbol{\Psi}^{-1})$. Hence, at $\boldsymbol{\eta} = \mathbf{0}$, the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ of the $\mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ family becomes

$$\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \mathbf{0}) = \boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \mathbf{0}) = \begin{bmatrix} \boldsymbol{\Psi}^{-1} & \mathbf{0} & \sqrt{\frac{2}{\pi}}\boldsymbol{\Psi}^{-1} \\ \mathbf{0} & \frac{1}{2}\boldsymbol{D}_p^\top(\boldsymbol{\Psi}^{-1} \otimes \boldsymbol{\Psi}^{-1})\boldsymbol{D}_p & \mathbf{0} \\ \sqrt{\frac{2}{\pi}}\boldsymbol{\Psi}^{-1} & \mathbf{0} & \frac{2}{\pi}\boldsymbol{\Psi}^{-1} \end{bmatrix},$$

which is clearly singular. □

Remark Since $\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \mathbf{0})$ is singular, in the proof that $\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \mathbf{0})$ is also singular it is enough to prove that the Jacobian matrix $\boldsymbol{J}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \mathbf{0})$ is finite (in the matrix sense).

Obviously, the singularity of the Fisher information matrix of the $\mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ family when $\boldsymbol{\eta} = \mathbf{0}$ is due to the fact that the score vectors corresponding to the location vector $\boldsymbol{\xi}$ and the skewness vector $\boldsymbol{\eta}$ are linearly dependent at $\boldsymbol{\eta} = \mathbf{0}$. In fact, the score vector of $\boldsymbol{X} \sim \mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ for $(\boldsymbol{\xi}^\top, \mathrm{vech}(\boldsymbol{\Psi})^\top, \boldsymbol{\eta}^\top)^\top$, at $\boldsymbol{\eta} = \mathbf{0}$, becomes

$$\boldsymbol{l}_{\boldsymbol{\xi},\boldsymbol{\Psi},\mathbf{0}}(\boldsymbol{X}) = \Bigg(\boldsymbol{\Psi}^{-1}(\boldsymbol{X} - \boldsymbol{\xi}),\, \frac{1}{2}\boldsymbol{D}_p^\top(\boldsymbol{\Psi} \otimes \boldsymbol{\Psi})^{-1}\mathrm{vec}\left\{(\boldsymbol{X} - \boldsymbol{\xi})(\boldsymbol{X} - \boldsymbol{\xi})^\top - \boldsymbol{\Psi}\right\}, \sqrt{\frac{2}{\pi}}\boldsymbol{\Psi}^{-1}(\boldsymbol{X} - \boldsymbol{\xi}) \Bigg)^\top. \tag{20}$$

It is evident from (20) that, at $\boldsymbol{\eta} = \mathbf{0}$, the score vectors of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are linearly related. Consequently, the Fisher information matrix, which is the covariance matrix of the score vector, is singular at $\boldsymbol{\eta} = \mathbf{0}$. The score vector in (20) can be obtained from a direct differentiation of the $\mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ log-likelihood function, or from the results in Arellano-Valle and Azzalini (2008) as well as from Ley and Paindaveine (2010) since the $\mathcal{SN}$ family belongs to the generalized skew-normal family (see Genton and Loperfido (2005)). This last fact can be easily verified by the form of the mpdf of $\boldsymbol{X} \sim \mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ in Equation (18), and from there it is clear that $\mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ is a generalized skew-normal distribution with location parameter $\boldsymbol{\xi}$, dispersion matrix $\boldsymbol{\Omega}$, density generator $\phi_p(z)$, and skewing function $\pi(z) : \mathbb{R}^p \to [0, 1]$ with $\pi(z) = \Phi\left(\boldsymbol{\gamma}^\top z\right)$.

Appendix B

Proof of Proposition 1

Let $\tilde{\boldsymbol{X}} = \boldsymbol{\xi} + \dfrac{1}{\sqrt{1 + \boldsymbol{Y}^\top \boldsymbol{\Omega} \boldsymbol{Y}}} \boldsymbol{\Omega} \boldsymbol{Y}\, T + \boldsymbol{U}$. Since, by assumption, T is independent of $(\boldsymbol{U}^\top, \boldsymbol{Y}^\top)^\top$, it is then clear that, conditionally on $\boldsymbol{Y} = \boldsymbol{y}$, $\tilde{\boldsymbol{X}}$ has the same distribution as $\boldsymbol{\xi} + \dfrac{1}{\sqrt{1 + \boldsymbol{y}^\top \boldsymbol{\Omega} \boldsymbol{y}}} \boldsymbol{\Omega} \boldsymbol{y}\, T + \boldsymbol{U}_{\boldsymbol{y}}$, where $\boldsymbol{U}_{\boldsymbol{y}} \overset{d}{=} \boldsymbol{U} | \boldsymbol{Y} = \boldsymbol{y} \sim \mathcal{N}_p\left(\mathbf{0}, (\boldsymbol{\Omega}^{-1} + \boldsymbol{y}\boldsymbol{y}^\top)^{-1}\right)$, independent of T. By Proposition 10, this means $\tilde{\boldsymbol{X}} | \boldsymbol{Y} = \boldsymbol{y} \sim \mathcal{SN}_p\left(\boldsymbol{\xi}, (\boldsymbol{\Omega}^{-1} + \boldsymbol{y}\boldsymbol{y}^\top)^{-1}, \dfrac{1}{\sqrt{1 + \boldsymbol{y}^\top \boldsymbol{\Omega} \boldsymbol{y}}} \boldsymbol{\Omega} \boldsymbol{y}\right)$, and hence, by Proposition 10, the mpdf of $\tilde{\boldsymbol{X}} | \boldsymbol{Y} = \boldsymbol{y}$ becomes $f_{\tilde{\boldsymbol{X}} | \boldsymbol{Y} = \boldsymbol{y}}(\boldsymbol{x}) = 2\phi_p(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Omega}) \Phi\{\boldsymbol{y}^\top (\boldsymbol{x} - \boldsymbol{\xi})\}$. Thus, since the mpdf of $\tilde{\boldsymbol{X}}$ is $f_{\tilde{\boldsymbol{X}}}(\boldsymbol{x}) = \int_{\mathbb{R}^p} f_{\tilde{\boldsymbol{X}} | \boldsymbol{Y}}(\boldsymbol{x} | \boldsymbol{y}) f_{\boldsymbol{Y}}(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}$, we have

$$
\begin{aligned}
f_{\tilde{\boldsymbol{X}}}(\boldsymbol{x}) &= 2\phi_p(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Omega}) \int_{\mathbb{R}^p} \Phi\{\boldsymbol{y}^\top (\boldsymbol{x} - \boldsymbol{\xi})\} \phi_p\left(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1}\right) \mathrm{d}\boldsymbol{y} \\
&= 2\phi_p(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Omega}) \mathbb{E}[\Phi\{\boldsymbol{Y}^\top (\boldsymbol{x} - \boldsymbol{\xi})\}].
\end{aligned}
$$

Now, using the fact that, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}\{\Phi(X)\} = \Phi\left(\dfrac{\mu}{\sqrt{1 + \sigma^2}}\right)$, we get the above result. □

Proof of Corollary 1

Since, by assumption, $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$, we have by Proposition 1 that $\boldsymbol{X}$ can be represented as $\boldsymbol{X} | \boldsymbol{Y} = \boldsymbol{y} \sim \mathcal{SN}_p\left(\boldsymbol{\xi}, (\boldsymbol{\Omega}^{-1} + \boldsymbol{y}\boldsymbol{y}^\top)^{-1}, \dfrac{1}{\sqrt{1 + \boldsymbol{y}^\top \boldsymbol{\Omega} \boldsymbol{y}}} \boldsymbol{\Omega} \boldsymbol{y}\right)$ and $\boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})$. Combining this statement with the stochastic representation of the

$\mathcal{SN}$ distribution as given in Proposition 10, the mpdf of $\boldsymbol{X}$ can be expressed as

$$f_X(\boldsymbol{x}) = 2\int_{\mathbb{R}^p}\int_0^\infty \phi_p\left(\boldsymbol{x}; \boldsymbol{\xi} + \frac{1}{\sqrt{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}}\boldsymbol{\Omega}\boldsymbol{y}\,t, (\boldsymbol{\Omega}^{-1}+\boldsymbol{y}\boldsymbol{y}^\top)^{-1}\right) \times\phi_p(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})\phi(t)\mathrm{d}t\mathrm{d}\boldsymbol{y}.$$

Now, using Lemma 2 in Arellano-Valle et al. (2005) and that $(\boldsymbol{\Omega}^{-1}+\boldsymbol{y}\boldsymbol{y}^\top)^{-1} = \boldsymbol{\Omega} - \frac{1}{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}\boldsymbol{\Omega}\boldsymbol{y}\boldsymbol{y}^\top\boldsymbol{\Omega}$, we have the identity given by

$$\phi_p\left(\boldsymbol{x}; \boldsymbol{\xi} + \frac{1}{\sqrt{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}}\boldsymbol{\Omega}\boldsymbol{y}\,t, \boldsymbol{\Omega} - \frac{1}{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}\boldsymbol{\Omega}\boldsymbol{y}\boldsymbol{y}^\top\boldsymbol{\Omega}\right)\phi(t) = \phi_p(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Omega})\phi\left\{t; \frac{\boldsymbol{y}^\top(\boldsymbol{x}-\boldsymbol{\xi})}{\sqrt{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}}, \frac{1}{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}\right\}.$$

Thus, we have

$$f_X(\boldsymbol{x}) = 2\int_{\mathbb{R}^p}\int_0^\infty \phi_p(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Omega})\phi_p(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})\phi\left\{t; \frac{\boldsymbol{y}^\top(\boldsymbol{x}-\boldsymbol{\xi})}{\sqrt{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}}, \frac{1}{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}\right\}\mathrm{d}t\mathrm{d}\boldsymbol{y}.$$

Considering the transformation $w = \sqrt{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}\,t$, we have, for the joint mpdf of $(\boldsymbol{X}, \boldsymbol{Y}, W)$:

$$f_{\boldsymbol{X},\boldsymbol{Y},W}(\boldsymbol{x}, \boldsymbol{y}, w) = 2\phi_p(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Omega})\phi_p(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})\phi\{w; \boldsymbol{y}^\top(\boldsymbol{x}-\boldsymbol{\xi}), 1\}, \quad \boldsymbol{x}\in\mathbb{R}^p, \boldsymbol{y}\in\mathbb{R}^p, w>0. \tag{21}$$

Again, using Lemma 2 in Arellano-Valle et al. (2005), we have

$$\phi_p(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Omega})\phi\{w; \boldsymbol{y}^\top(\boldsymbol{x}-\boldsymbol{\xi}), 1\} = \phi_p\left\{\boldsymbol{x}; \boldsymbol{\xi} + w\frac{1}{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}\boldsymbol{\Omega}\boldsymbol{y}, (\boldsymbol{\Omega}^{-1}+\boldsymbol{y}\boldsymbol{y}^\top)^{-1}\right\} \times\phi(w; 0, 1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}),$$

and using this result in (21), we get

$$f_{\boldsymbol{X},\boldsymbol{Y},W}(\boldsymbol{x}, \boldsymbol{y}, w) = 2\phi_p\left\{\boldsymbol{x}; \boldsymbol{\xi} + w\frac{1}{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}\boldsymbol{\Omega}\boldsymbol{y}, (\boldsymbol{\Omega}^{-1}+\boldsymbol{y}\boldsymbol{y}^\top)^{-1}\right\} \times\phi_p(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})\phi(w; 0, 1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}) \tag{22}$$

for $\boldsymbol{x}\in\mathbb{R}^p$, $\boldsymbol{y}\in\mathbb{R}^p$, and $w>0$. The rest of the proof is trivial from (22). □

Proof of Corollary 2

From Eq. (21), we get that the joint mpdf of $(\boldsymbol{X}, \boldsymbol{Y})$ is given by

$$\begin{aligned} f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}) &= 2\phi_p(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Omega})\phi_p(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1}) \int_0^\infty \phi\{w; \boldsymbol{y}^\top(\boldsymbol{x} - \boldsymbol{\xi}), 1\}\mathrm{d}w \\ &= 2\phi_p(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Omega})\phi_p(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})\Phi\{\boldsymbol{y}^\top(\boldsymbol{x} - \boldsymbol{\xi})\}, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p. \end{aligned}$$

From this mpdf it follows that the marginal mpdf of $\boldsymbol{X}$ and the conditional mpdf of $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$ are:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = 2\phi_p(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Omega})\Phi\left\{\frac{\boldsymbol{\lambda}^\top(\boldsymbol{x} - \boldsymbol{\xi})}{\sqrt{1 + (\boldsymbol{x} - \boldsymbol{\xi})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\xi})}}\right\}, \quad \boldsymbol{x} \in \mathbb{R}^p,$$

$$\begin{aligned} &f_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}) \\ &= \frac{1}{\Phi\left\{\dfrac{\boldsymbol{\lambda}^\top(\boldsymbol{x} - \boldsymbol{\xi})}{\sqrt{1 + (\boldsymbol{x} - \boldsymbol{\xi})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\xi})}}\right\}}\phi_p(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})\Phi\{\boldsymbol{y}^\top(\boldsymbol{x} - \boldsymbol{\xi})\}, \quad \boldsymbol{y} \in \mathbb{R}^p. \end{aligned}$$

Also, the conditional multivariate moment generating function of $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$ is

$$\begin{aligned} M_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{t}) &= \mathbb{E}\{\exp(\boldsymbol{t}^\top\boldsymbol{Y})|\boldsymbol{X} = \boldsymbol{x}\} \\ &= \int_{\mathbb{R}^p} \frac{\exp(\boldsymbol{t}^\top\boldsymbol{y})}{\Phi\left\{\dfrac{\boldsymbol{\lambda}^\top(\boldsymbol{x} - \boldsymbol{\xi})}{\sqrt{1 + (\boldsymbol{x} - \boldsymbol{\xi})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\xi})}}\right\}}\phi_p(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})\Phi\{\boldsymbol{y}^\top(\boldsymbol{x} - \boldsymbol{\xi})\}\mathrm{d}\boldsymbol{y} \\ &= \frac{\exp\left(-\dfrac{1}{2}\boldsymbol{\lambda}^\top\boldsymbol{\Omega}\boldsymbol{\lambda}\right)}{\Phi\left\{\dfrac{\boldsymbol{\lambda}^\top(\boldsymbol{x} - \boldsymbol{\xi})}{\sqrt{1 + (\boldsymbol{x} - \boldsymbol{\xi})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\xi})}}\right\}} \\ &\quad \times \int_{\mathbb{R}^p} \frac{1}{(\sqrt{2\pi})^p\{\det(\boldsymbol{\Omega}^{-1})\}^{1/2}} \exp[-\frac{1}{2}\{\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y} - 2(\boldsymbol{\Omega}\boldsymbol{\lambda} + \boldsymbol{t})^\top\boldsymbol{y}\}] \\ &\quad \times \Phi\{\boldsymbol{y}^\top(\boldsymbol{x} - \boldsymbol{\xi})\}\mathrm{d}\boldsymbol{y} \\ &= \frac{\exp\left(\boldsymbol{\lambda}^\top\boldsymbol{t} + \dfrac{1}{2}\boldsymbol{t}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{t}\right)}{\Phi\left\{\dfrac{\boldsymbol{\lambda}^\top(\boldsymbol{x} - \boldsymbol{\xi})}{\sqrt{1 + (\boldsymbol{x} - \boldsymbol{\xi})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\xi})}}\right\}} \\ &\quad \times \Phi\left\{\frac{\boldsymbol{\lambda}^\top(\boldsymbol{x} - \boldsymbol{\xi}) + (\boldsymbol{x} - \boldsymbol{\xi})^\top\boldsymbol{\Omega}^{-1}\boldsymbol{t}}{\sqrt{1 + (\boldsymbol{x} - \boldsymbol{\xi})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\xi})}}\right\}, \quad \boldsymbol{t} \in \mathbb{R}^p, \end{aligned}$$

where the last step follows from Lemma 5.3 in Azzalini and Capitanio (2014). Thus, we have

$$\begin{aligned}
\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}) &= \frac{\partial}{\partial \boldsymbol{t}} M_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{t})\Big|_{\boldsymbol{t}=\boldsymbol{0}} \\
&= \boldsymbol{\lambda} + W_{\phi}\left\{\frac{\boldsymbol{\lambda}^{\top}(\boldsymbol{x}-\boldsymbol{\xi})}{\sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}}\right\} \\
&\quad \times \frac{1}{\sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi}), \\
\mathbb{E}(\boldsymbol{Y}\boldsymbol{Y}^{\top}|\boldsymbol{X}=\boldsymbol{x}) &= \frac{\partial^2}{\partial \boldsymbol{t}\partial \boldsymbol{t}^{\top}} M_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{t})\Big|_{\boldsymbol{t}=\boldsymbol{0}} \\
&= \boldsymbol{\lambda}\boldsymbol{\lambda}^{\top} + \boldsymbol{\Omega}^{-1} + W_{\phi}\left\{\frac{\boldsymbol{\lambda}^{\top}(\boldsymbol{x}-\boldsymbol{\xi})}{\sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}}\right\} \\
&\quad \times \frac{1}{\sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}} \\
&\quad \times \Big\{\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})\boldsymbol{\lambda}^{\top} + \boldsymbol{\lambda}(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1} \\
&\quad - \frac{\boldsymbol{\lambda}^{\top}(\boldsymbol{x}-\boldsymbol{\xi})}{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}\Big\}.
\end{aligned}$$

Now, since $\mathbb{E}(W\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}) = \mathbb{E}\{W\mathbb{E}(\boldsymbol{Y}|W, \boldsymbol{X}=\boldsymbol{x})|\boldsymbol{X}=\boldsymbol{x}\}$, then for the evaluation of this quantity we need the conditional mpdfs of $\boldsymbol{Y}|W=w, \boldsymbol{X}=\boldsymbol{x}$ and $W|\boldsymbol{X}=\boldsymbol{x}$. Again, from (21):

$$f_{W|\boldsymbol{X}=\boldsymbol{x}}(w) \propto \int_{\mathbb{R}^p} \phi_p(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})\phi\{w; \boldsymbol{y}^{\top}(\boldsymbol{x}-\boldsymbol{\xi}), 1\}\mathrm{d}\boldsymbol{y}, \quad w>0,$$

where, from Lemma 2 in Arellano-Valle et al. (2005), the product $\phi_p(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})\phi\{w; \boldsymbol{y}^{\top}(\boldsymbol{x}-\boldsymbol{\xi}), 1\}$ is equal to $\phi\{w; (\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\lambda}, 1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})\}\phi_p[\boldsymbol{y}; \boldsymbol{\lambda}+\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\xi})\{w-\boldsymbol{\lambda}^{\top}(\boldsymbol{x}-\boldsymbol{\xi})\}, \boldsymbol{\Lambda}]$, where $\boldsymbol{\Lambda} = \{\boldsymbol{\Omega}+(\boldsymbol{x}-\boldsymbol{\xi})(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\}^{-1}$. Thus,

$$f_{W|\boldsymbol{X}=\boldsymbol{x}}(w) \propto \phi\{w; (\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\lambda}, 1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})\}, \quad w>0.$$

That is, $W|\boldsymbol{X}=\boldsymbol{x} \sim \mathcal{TN}\{(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\lambda}, 1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi}); (0,\infty)\}$, and hence

$$\begin{aligned}
\mathbb{E}(W|\boldsymbol{X}=\boldsymbol{x}) &= (\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\lambda} + \sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})} \\
&\quad \times W_{\phi}\left\{\frac{\boldsymbol{\lambda}^{\top}(\boldsymbol{x}-\boldsymbol{\xi})}{\sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}}\right\}, \\
\mathbb{E}(W^2|\boldsymbol{X}=\boldsymbol{x}) &= \{(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\lambda}\}^2 + \{1+(\boldsymbol{x}-\boldsymbol{\xi})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})\}
\end{aligned}$$

$$+ (\boldsymbol{x}-\boldsymbol{\xi})^\top \boldsymbol{\lambda} \sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}$$
$$\times W_\phi \left\{ \frac{\boldsymbol{\lambda}^\top (\boldsymbol{x}-\boldsymbol{\xi})}{\sqrt{1+(\boldsymbol{x}-\boldsymbol{\xi})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})}} \right\}.$$

Furthermore, from Eq. (21), we have $f_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},W=w}(\boldsymbol{y}) \propto \phi_p(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})\phi\{w; \boldsymbol{y}^\top(\boldsymbol{x}-\boldsymbol{\xi}), 1\}$, where again, from Lemma 2 in Arellano-Valle et al. (2005), the product $\phi_p(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})\phi\{w; \boldsymbol{y}^\top(\boldsymbol{x}-\boldsymbol{\xi}), 1\}$ is equal to $\phi\{w; (\boldsymbol{x}-\boldsymbol{\xi})^\top\boldsymbol{\lambda}, 1+(\boldsymbol{x}-\boldsymbol{\xi})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})\}\phi_p[\boldsymbol{y}; \boldsymbol{\lambda}+\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\xi})\{w-\boldsymbol{\lambda}^\top(\boldsymbol{x}-\boldsymbol{\xi})\}, \boldsymbol{\Lambda}]$, with $\boldsymbol{\Lambda} = \{\boldsymbol{\Omega}+(\boldsymbol{x}-\boldsymbol{\xi})(\boldsymbol{x}-\boldsymbol{\xi})^\top\}^{-1}$. Therefore,

$$f_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},W=w}(\boldsymbol{y}) = \phi_p[\boldsymbol{y}; \boldsymbol{\lambda}+\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\xi})\{w-\boldsymbol{\lambda}^\top(\boldsymbol{x}-\boldsymbol{\xi})\}, \boldsymbol{\Lambda}], \quad \boldsymbol{y} \in \mathbb{R}^p.$$

That is, $\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}, W=w \sim \mathcal{N}_p(\boldsymbol{\lambda}+\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\xi})\{w-\boldsymbol{\lambda}^\top(\boldsymbol{x}-\boldsymbol{\xi})\}, \boldsymbol{\Lambda})$. Hence,

$$\begin{aligned}
\mathbb{E}(W\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}) &= \mathbb{E}\{W\mathbb{E}(\boldsymbol{Y}|W, \boldsymbol{X}=\boldsymbol{x})|\boldsymbol{X}=\boldsymbol{x}\} \\
&= \mathbb{E}\left[W \left\{ \boldsymbol{\lambda} + \frac{W-\boldsymbol{\lambda}^\top(\boldsymbol{x}-\boldsymbol{\xi})}{1+(\boldsymbol{x}-\boldsymbol{\xi})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})} \boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi}) \right\} \mid \boldsymbol{X}=\boldsymbol{x} \right] \\
&= \mathbb{E}(W|\boldsymbol{X}=\boldsymbol{x})\boldsymbol{\lambda} \\
&\quad + \frac{\{\mathbb{E}(W^2|\boldsymbol{X}=\boldsymbol{x}) - \boldsymbol{\lambda}^\top(\boldsymbol{x}-\boldsymbol{\xi})\mathbb{E}(W|\boldsymbol{X}=\boldsymbol{x})\}}{1+(\boldsymbol{x}-\boldsymbol{\xi})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})} \boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi}).
\end{aligned}$$

□

Proof of Proposition 2

From Proposition 1, we have that $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ can be represented as $\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y} \sim \mathcal{SN}_p\left(\boldsymbol{\xi}, \boldsymbol{\Omega} - \frac{1}{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}\boldsymbol{\Omega}\boldsymbol{y}\boldsymbol{y}^\top\boldsymbol{\Omega}, \frac{1}{\sqrt{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}}\boldsymbol{\Omega}\boldsymbol{y}\right)$ and $\boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})$. Now, using the results in Eq. (16), we get $\mathbb{E}(\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}) = \boldsymbol{\xi} + \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}}\boldsymbol{\Omega}\boldsymbol{y}$ and $\mathbb{V}\mathrm{ar}(\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}) = \boldsymbol{\Omega} - \frac{2}{\pi}\frac{1}{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}\boldsymbol{\Omega}\boldsymbol{y}\boldsymbol{y}^\top\boldsymbol{\Omega}$, where $\boldsymbol{Y} \sim \mathcal{N}_p\left(\boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1}\right)$, so that

$$\mathbb{E}(\boldsymbol{X}) = \mathbb{E}\{\mathbb{E}(\boldsymbol{X}|\boldsymbol{Y})\} = \boldsymbol{\xi} + \sqrt{\frac{2}{\pi}}\mathbb{E}\left(\frac{1}{\sqrt{1+\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y}}}\boldsymbol{\Omega}\boldsymbol{Y}\right)$$

and

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\boldsymbol{X}) &= \mathbb{E}\{\mathbb{V}\mathrm{ar}(\boldsymbol{X}|\boldsymbol{Y})\} + \mathbb{V}\mathrm{ar}\{\mathbb{E}(\boldsymbol{X}|\boldsymbol{Y})\} \\
&= \boldsymbol{\Omega} - \frac{2}{\pi}\mathbb{E}\left(\frac{1}{1+\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y}}\boldsymbol{\Omega}\boldsymbol{Y}\boldsymbol{Y}^\top\boldsymbol{\Omega}\right) + \frac{2}{\pi}\mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{1+\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y}}}\boldsymbol{\Omega}\boldsymbol{Y}\right).
\end{aligned}$$

Thus,

$$\mathbb{V}\mathrm{ar}(\boldsymbol{X}) = \boldsymbol{\Omega} - \frac{2}{\pi}\mathbb{E}\left(\frac{}{\sqrt{1+\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y}}}\boldsymbol{\Omega}\boldsymbol{Y}\right)\mathbb{E}\left(\frac{1}{\sqrt{1+\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y}}}\boldsymbol{\Omega}\boldsymbol{Y}\right)^\top.$$

□

Proof of Proposition 3

Let $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$. Then, by Proposition 1 we have that $\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y} \sim \mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}_{\boldsymbol{y}}, \boldsymbol{\eta}_{\boldsymbol{y}})$, with $\boldsymbol{\Psi}_{\boldsymbol{y}} = \boldsymbol{\Omega} - \boldsymbol{\eta}_{\boldsymbol{y}}\boldsymbol{\eta}_{\boldsymbol{y}}^\top$ and $\boldsymbol{\eta}_{\boldsymbol{y}} = (1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y})^{-1/2}\boldsymbol{\Omega}\boldsymbol{y}$, where $\boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})$, with $\boldsymbol{\lambda} = (1-\boldsymbol{\eta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\eta})^{-1/2}\boldsymbol{\Omega}^{-1}\boldsymbol{\eta}$ and $\boldsymbol{\Omega} = \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top$ as defined in Eq. (6). Hence by adapting the result of the $\mathcal{SN}$-mmgf from Proposition 11 to the conditional mmgf of $\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y} \sim \mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}_{\boldsymbol{y}}, \boldsymbol{\eta}_{\boldsymbol{y}})$, we have

$$\begin{aligned} M_{\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}}(\boldsymbol{t}) &= 2\exp\left(\boldsymbol{t}^\top\boldsymbol{\xi} + \frac{1}{2}\boldsymbol{t}^\top\boldsymbol{\Omega}_{\boldsymbol{y}}\boldsymbol{t}\right)\Phi(\boldsymbol{\eta}_{\boldsymbol{y}}^\top\boldsymbol{t}) \\ &= 2\exp\left(\boldsymbol{t}^\top\boldsymbol{\xi} + \frac{1}{2}\boldsymbol{t}^\top\boldsymbol{\Omega}\boldsymbol{t}\right)\Phi\left(\frac{\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{t}}{\sqrt{1+\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}}\right), \end{aligned}$$

since $\boldsymbol{\Omega}_{\boldsymbol{y}} = \boldsymbol{\Psi}_{\boldsymbol{y}} + \boldsymbol{\eta}_{\boldsymbol{y}}\boldsymbol{\eta}_{\boldsymbol{y}}^\top = \boldsymbol{\Omega}$. Hence, the mmgf of $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ becomes

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\{M_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{t})\} = 2\exp\left(\boldsymbol{t}^\top\boldsymbol{\xi} + \frac{1}{2}t^\top\boldsymbol{\Omega}\boldsymbol{t}\right)\mathbb{E}\left\{\Phi\left(\frac{\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{t}}{\sqrt{1+\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y}}}\right)\right\},$$

where as before $\boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})$, and we note that

$$\mathbb{E}\left\{\Phi\left(\frac{\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{t}}{\sqrt{1+\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y}}}\right)\right\} = \mathbb{E}\left\{\Phi\left(\frac{\boldsymbol{Z}^\top\boldsymbol{t}}{\sqrt{1+\boldsymbol{Z}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{Z}}}\right)\right\},$$

where $\boldsymbol{Z} = \boldsymbol{\Omega}\boldsymbol{Y} \sim \mathcal{N}_p(\bar{\boldsymbol{\lambda}}, \boldsymbol{\Omega})$, with $\bar{\boldsymbol{\lambda}} = \boldsymbol{\Omega}\boldsymbol{\lambda} = (1-\boldsymbol{\eta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\eta})^{-1/2}\boldsymbol{\eta}$. □

Proof of Proposition 4

From Proposition 12, the conditional mcdf of $\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y} \sim \mathcal{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}_{\boldsymbol{y}}, \boldsymbol{\eta}_{\boldsymbol{y}})$ is given by

$$\begin{aligned} F_{\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}}(\boldsymbol{x}) &= \int_{\boldsymbol{u}\le\boldsymbol{x}} 2\phi_p(\boldsymbol{u}; \boldsymbol{\xi}, \boldsymbol{\Omega}_{\boldsymbol{y}})\Phi\left\{\boldsymbol{\gamma}_{\boldsymbol{y}}^\top\boldsymbol{\Omega}_{\boldsymbol{y}}^{-1/2}(\boldsymbol{u}-\boldsymbol{\xi})\right\}\mathrm{d}\boldsymbol{u} \\ &= 2\Phi_{p+1}(\boldsymbol{x}^*; \boldsymbol{\xi}^*, \boldsymbol{\Omega}_{\boldsymbol{y}}^*), \end{aligned}$$

where

$$\boldsymbol{x}^* = \begin{pmatrix} \boldsymbol{x} \\ 0 \end{pmatrix}, \quad \boldsymbol{\xi}^* = \begin{pmatrix} \boldsymbol{\xi} \\ 0 \end{pmatrix},$$
$$\boldsymbol{\Omega}_y^* = \begin{pmatrix} \boldsymbol{\Omega}_y - \boldsymbol{\Omega}_y^{1/2}\boldsymbol{\gamma}_y \\ -\boldsymbol{\gamma}_y^\top \boldsymbol{\Omega}_y^{1/2} \quad 1 + \boldsymbol{\gamma}_y^\top \boldsymbol{\Omega}_y^{-1}\boldsymbol{\gamma}_y \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Omega} & -\boldsymbol{\Omega}\boldsymbol{y} \\ -\boldsymbol{y}^\top\boldsymbol{\Omega} \; 1 + \boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y} \end{pmatrix},$$

since $\boldsymbol{\Omega}_y = \boldsymbol{\Omega}$, $\boldsymbol{\gamma}_y = (1 - \boldsymbol{\eta}_y^\top \boldsymbol{\Omega}_y^{-1}\boldsymbol{\eta}_y)^{-1/2}\boldsymbol{\Omega}_y^{-1/2}\boldsymbol{\eta}_y = \boldsymbol{\Omega}^{1/2}\boldsymbol{y}$ and $\boldsymbol{\eta}_y = (1 + \boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y})^{-1/2}\boldsymbol{\Omega}\boldsymbol{y}$. Therefore, the mcdf of $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ becomes

$$\begin{aligned} F_X(\boldsymbol{x}) &= \mathbb{E}\left[\Phi_{p+1}\left\{\begin{pmatrix} \boldsymbol{x} \\ 0 \end{pmatrix}; \begin{pmatrix} \boldsymbol{\xi} \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Omega} & -\boldsymbol{\Omega}\boldsymbol{Y} \\ -\boldsymbol{Y}^\top\boldsymbol{\Omega} \; 1 + \boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y} \end{pmatrix}\right\}\right] \\ &= \mathbb{E}\left[\Phi_{p+1}\left\{\begin{pmatrix} \boldsymbol{x} \\ 0 \end{pmatrix}; \begin{pmatrix} \boldsymbol{\xi} \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Omega} & -\boldsymbol{Z} \\ -\boldsymbol{Z} \; 1 + \boldsymbol{Z}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{Z} \end{pmatrix}\right\}\right], \end{aligned}$$

where as before $\boldsymbol{Z} = \boldsymbol{\Omega}\boldsymbol{Y} \sim \mathcal{N}_p(\bar{\boldsymbol{\lambda}}, \boldsymbol{\Omega})$, with $\bar{\boldsymbol{\lambda}} = \boldsymbol{\Omega}\boldsymbol{\lambda} = (1 - \boldsymbol{\eta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\eta})^{-1/2}\boldsymbol{\eta}$. □

Proof of Proposition 5

By assumption $\widetilde{\boldsymbol{X}} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}$, with $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$, and therefore from the stochastic representation of $\boldsymbol{X}$, given in Proposition 10, we have $\widetilde{\boldsymbol{X}}$ has stochastic representation given by $\widetilde{\boldsymbol{X}} \overset{d}{=} \boldsymbol{a} + \boldsymbol{B}\boldsymbol{\xi} + \dfrac{1}{\sqrt{1 + \boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{Y}}}\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{Y}\, T + \boldsymbol{B}\boldsymbol{U}$, where $T \sim \mathcal{HN}(0, 1)$ and $\boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})$ are independent, and $\boldsymbol{U}|\boldsymbol{Y} = \boldsymbol{y} \sim \mathcal{N}_p\left(\boldsymbol{0}, (\boldsymbol{\Omega}^{-1} + \boldsymbol{y}\boldsymbol{y}^\top)^{-1}\right)$. Also, by conditioning $\widetilde{\boldsymbol{X}}$ on $\boldsymbol{Y} = \boldsymbol{y}$ from this stochastic representation, we have that $\widetilde{\boldsymbol{X}}|\boldsymbol{Y} = \boldsymbol{y} \sim \mathcal{SN}_k(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{\xi}, \boldsymbol{\Psi}_y, \boldsymbol{\eta}_y)$, with conditional mpdf

$$f_{\widetilde{\boldsymbol{X}}|\boldsymbol{Y}=\boldsymbol{y}}(\tilde{\boldsymbol{x}}) = \phi_p(\tilde{\boldsymbol{x}}; \boldsymbol{a} + \boldsymbol{B}\boldsymbol{\xi}, \boldsymbol{\Psi}_y + \boldsymbol{\eta}_y\boldsymbol{\eta}_y^\top)\Phi\left\{\frac{\boldsymbol{\eta}_y^\top\boldsymbol{\Psi}_y^{-1}(\tilde{\boldsymbol{x}} - \boldsymbol{a} - \boldsymbol{B}\boldsymbol{\xi})}{\sqrt{1 + \boldsymbol{\eta}_y^\top(\boldsymbol{\Psi}_y + \boldsymbol{\eta}_y\boldsymbol{\eta}_y^\top)^{-1}\boldsymbol{\eta}_y}}\right\},$$

where $\boldsymbol{Y} \sim \mathcal{N}_p\left(\boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1}\right)$, $\boldsymbol{\Psi}_y = \boldsymbol{B}(\boldsymbol{\Omega}^{-1} + \boldsymbol{y}\boldsymbol{y}^\top)^{-1}\boldsymbol{B}^\top$ and $\boldsymbol{\eta}_y = \dfrac{1}{\sqrt{1 + \boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{y}}}\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{y}$. Thus, since $f_{\widetilde{\boldsymbol{X}}}(\tilde{\boldsymbol{x}}) = \int_{\mathbb{R}^p} f_{\widetilde{\boldsymbol{X}}|\boldsymbol{Y}}(\tilde{\boldsymbol{x}})\, f_{\boldsymbol{Y}}(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$, we have, after some extensive but straightforward algebra, that the mpdf of $\widetilde{\boldsymbol{X}}$ becomes

$$\begin{aligned} f_{\widetilde{\boldsymbol{X}}}(\tilde{\boldsymbol{x}}) &= 2\phi_p(\tilde{\boldsymbol{x}}; \boldsymbol{a} + \boldsymbol{B}\boldsymbol{\xi}, \boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top)\int_{\mathbb{R}^p} \Phi\left[\frac{\boldsymbol{y}^\top\boldsymbol{\Omega}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top)^{-1}(\tilde{\boldsymbol{x}} - \boldsymbol{a} - \boldsymbol{B}\boldsymbol{\xi})}{\sqrt{1 + \boldsymbol{y}^\top\{\boldsymbol{\Omega} - \boldsymbol{\Omega}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top)^{-1}\boldsymbol{B}\boldsymbol{\Omega}\}\boldsymbol{y}}}\right]\phi_p\left(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1}\right)\mathrm{d}\boldsymbol{y} \\ &= 2\phi_p(\tilde{\boldsymbol{x}}; \boldsymbol{a} + \boldsymbol{B}\boldsymbol{\xi}, \boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top)\mathbb{E}\left(\Phi\left[\frac{\boldsymbol{Y}^\top\boldsymbol{\Omega}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top)^{-1}(\tilde{\boldsymbol{x}} - \boldsymbol{a} - \boldsymbol{B}\boldsymbol{\xi})}{\sqrt{1 + \boldsymbol{Y}^\top\{\boldsymbol{\Omega} - \boldsymbol{\Omega}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^\top)^{-1}\boldsymbol{B}\boldsymbol{\Omega}\}\boldsymbol{Y}}}\right]\right), \quad \tilde{\boldsymbol{x}} \in \mathbb{R}^p, \end{aligned}$$

where $\boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})$.

Now, let $\boldsymbol{Y}_B = \boldsymbol{B\Omega Y}$ and $\boldsymbol{Y}_C = \boldsymbol{CY}$, where $\boldsymbol{C} = \mathrm{I}_p - \boldsymbol{B}^\top(\boldsymbol{B\Omega B}^\top)^{-1}\boldsymbol{B\Omega}$. Note that $\boldsymbol{BC}^\top = \boldsymbol{0}$ and $\boldsymbol{\Omega} - \boldsymbol{\Omega B}^\top(\boldsymbol{B\Omega B}^\top)^{-1}\boldsymbol{B\Omega} = \boldsymbol{C}^\top\boldsymbol{\Omega C}$ and so $\boldsymbol{Y}^\top\{\boldsymbol{\Omega} - \boldsymbol{\Omega B}^\top(\boldsymbol{B\Omega B}^\top)^{-1}\boldsymbol{B\Omega}\}\boldsymbol{Y} = \boldsymbol{Y}^\top\boldsymbol{C}^\top\boldsymbol{\Omega CY} = \boldsymbol{Y}_C^\top\boldsymbol{\Omega Y}_C$. Since $\boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1})$, it follows, from the properties of the multivariate normal distribution, that $\boldsymbol{Y}_B \sim \mathcal{N}_k(\boldsymbol{B\Omega\lambda}, \boldsymbol{B\Omega B}^\top)$ and $\boldsymbol{Y}_C \sim \mathcal{N}_k(\boldsymbol{C\lambda}, \boldsymbol{C\Omega}^{-1}\boldsymbol{C}^\top)$ and they are independent since $\mathrm{cov}(\boldsymbol{Y}_B, \boldsymbol{Y}_C) = \boldsymbol{BC}^\top = \boldsymbol{0}$. In turn, this means that $(\boldsymbol{B\Omega B}^\top)^{-1}\boldsymbol{Y}_B = (\boldsymbol{B\Omega B}^\top)^{-1}\boldsymbol{B\Omega Y} \sim \mathcal{N}_k((\boldsymbol{B\Omega B}^\top)^{-1}\boldsymbol{B\Omega\lambda}, (\boldsymbol{B\Omega B}^\top)^{-1})$ and it is independent of $\boldsymbol{Y}_C^\top\boldsymbol{\Omega Y}_C = \boldsymbol{Y}^\top\{\boldsymbol{\Omega} - \boldsymbol{\Omega B}^\top(\boldsymbol{B\Omega B}^\top)^{-1}\boldsymbol{B\Omega}\}\boldsymbol{Y}$. Thus, for the above expectation, we have, by using the same arguments as in the proof of Proposition 1, that

$$
\begin{aligned}
&\mathbb{E}\left(\Phi\left[\frac{\boldsymbol{Y}^\top\boldsymbol{\Omega B}^\top(\boldsymbol{B\Omega B}^\top)^{-1}(\widetilde{\boldsymbol{x}} - \boldsymbol{a} - \boldsymbol{B\xi})}{\sqrt{1 + \boldsymbol{Y}^\top\{\boldsymbol{\Omega} - \boldsymbol{\Omega B}^\top(\boldsymbol{B\Omega B}^\top)^{-1}\boldsymbol{B\Omega}\}\boldsymbol{Y}}}\right]\right)\\
&= \mathbb{E}\left[\Phi\left\{\frac{\boldsymbol{Y}_B^\top(\boldsymbol{B\Omega B}^\top)^{-1}(\widetilde{\boldsymbol{x}} - \boldsymbol{a} - \boldsymbol{B\xi})}{\sqrt{1 + \boldsymbol{Y}_C^\top\boldsymbol{\Omega Y}_C}}\right\}\right]\\
&= \mathbb{E}\left[\Phi\left\{\frac{\boldsymbol{\lambda}^\top\boldsymbol{\Omega B}^\top(\boldsymbol{B\Omega B}^\top)^{-1}(\widetilde{\boldsymbol{x}} - \boldsymbol{a} - \boldsymbol{B\xi})}{\sqrt{1 + \boldsymbol{Y}^\top\{\boldsymbol{\Omega} - \boldsymbol{\Omega B}^\top(\boldsymbol{B\Omega B}^\top)^{-1}\boldsymbol{B\Omega}\}\boldsymbol{Y}}\sqrt{1 + (\widetilde{\boldsymbol{x}} - \boldsymbol{a} - \boldsymbol{B\xi})^\top(\boldsymbol{B\Omega B}^\top)^{-1}(\widetilde{\boldsymbol{x}} - \boldsymbol{a} - \boldsymbol{B\xi})}}\right\}\right].
\end{aligned}
$$

When $\boldsymbol{B}$ is a nonsingular square matrix, this expectation, which corresponds to the skewing function of the mpdf of $\widetilde{\boldsymbol{X}}$, reduces to $\Phi\left\{\dfrac{\boldsymbol{\lambda}^\top\boldsymbol{\Omega B}^\top(\boldsymbol{B\Omega B}^\top)^{-1}(\widetilde{\boldsymbol{x}} - \boldsymbol{a} - \boldsymbol{B\xi})}{\sqrt{1 + (\widetilde{\boldsymbol{x}} - \boldsymbol{a} - \boldsymbol{B\xi})^\top(\boldsymbol{B\Omega B}^\top)^{-1}(\widetilde{\boldsymbol{x}} - \boldsymbol{a} - \boldsymbol{B\xi})}}\right\}$, thus it follows that $\widetilde{\boldsymbol{X}} \sim \mathcal{MSN}_p(\boldsymbol{a} + \boldsymbol{B\xi}, \boldsymbol{B\Psi B}^\top, \boldsymbol{B\eta})$. □

Proof of Corollary 3

By considering the partitions of $\boldsymbol{X}$, $\boldsymbol{\xi}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\eta}$, as in Eq. (17), the mpdf of $\boldsymbol{X}_1$ can be found using Proposition 5, putting $\boldsymbol{a} = \boldsymbol{0}$ and $\boldsymbol{B} = (\mathrm{I}_{p_1}\ \boldsymbol{0})$ to obtain

$$
\begin{aligned}
f_{X_1}(\boldsymbol{x}_1) &= 2\phi_{p_1}(\boldsymbol{x}_1; \boldsymbol{\xi}_1, \boldsymbol{\Omega}_{11}) \int_{\mathbb{R}^p} \Phi\left\{\frac{(\boldsymbol{\lambda}_1 + \boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\lambda}_2)^\top(\boldsymbol{x}_1 - \boldsymbol{\xi}_1)}{\sqrt{1 + \boldsymbol{y}_2^\top(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})\boldsymbol{y}_2}}\right\}\\
&\quad \times \phi_p\left(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\Omega}^{-1}\right)\mathrm{d}\boldsymbol{y}\\
&= 2\phi_{p_1}(\boldsymbol{x}_1; \boldsymbol{\xi}_1, \boldsymbol{\Omega}_{11}) \int_{\mathbb{R}^{p_2}} \Phi\left\{\frac{(\boldsymbol{\lambda}_1 + \boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\lambda}_2)^\top(\boldsymbol{x}_1 - \boldsymbol{\xi}_1)}{\sqrt{1 + \boldsymbol{y}_2^\top(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})\boldsymbol{y}_2}}\right\}\\
&\quad \times \phi_{p_2}\left(\boldsymbol{y}_2; \boldsymbol{\lambda}_2, \boldsymbol{\Omega}_{22\cdot 1}^{-1}\right)\mathrm{d}\boldsymbol{y}_2,\ \boldsymbol{x}_1 \in \mathbb{R}^{p_1},
\end{aligned}
$$

where $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top)^\top$, with $\boldsymbol{y}_i \in \mathbb{R}^{p_i}$, and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\top, \boldsymbol{\lambda}_2^\top)^\top$, with $\boldsymbol{y}_i, \boldsymbol{\lambda}_i \in \mathbb{R}^{p_i}$, for $i = 1, 2$, and $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_{ij})$, with $\boldsymbol{\Omega}_{ij} = \boldsymbol{\Psi}_{ij} + \boldsymbol{\eta}_i\boldsymbol{\eta}_j^\top$, $i, j = 1, 2$. Finally, by using the relations described at the beginning of the corollary we have, after some algebra, that

$$\frac{\boldsymbol{\lambda}_1 + \boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\lambda}_2}{\sqrt{1 + \boldsymbol{y}_2^\top(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})\boldsymbol{y}_2}} = \frac{1}{\sqrt{1 + \boldsymbol{y}_2^\top\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{y}_2}\sqrt{1 - \boldsymbol{\eta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\eta}}}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\eta}_1 =$$
$$\sqrt{\frac{1 + \boldsymbol{\eta}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}{1 + \boldsymbol{y}_2^\top\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{y}_2}}\frac{1}{\sqrt{1 + \boldsymbol{\eta}_1^\top\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\eta}_1}}\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\eta}_1 = \boldsymbol{\lambda}_{1\cdot2}. \quad \square$$

Proof of Proposition 6

To prove this, we will show that the density function of $\mathcal{MSN}_1(0, 1, \eta)$ is not a log-concave function. The log-density of $\mathcal{MSN}_1(0, 1, \eta)$ is

$$\log\{f(x)\} = \log\left[2\phi(x; 0, 1 + \eta^2)\Phi\left\{\frac{\eta x/\sqrt{1 + \eta^2}}{\sqrt{1 + x^2/(1 + \eta^2)}}\right\}\right].$$

The second derivative of $\log\{f(x)\}$ with respect to x is

$$\frac{\mathrm{d}^2\log\{f(x)\}}{\mathrm{d}x^2} = \frac{1}{1 + \eta^2}\left[-1 - \frac{1}{(1 + t^2)^3}\left\{\frac{\phi(\eta t/\sqrt{1 + t^2})}{\Phi(\eta t/\sqrt{1 + t^2})}\right\}^2 - \frac{1}{\Phi(\eta t/\sqrt{1 + t^2})}\left\{\frac{3t\phi(\eta t/\sqrt{1 + t^2})}{(1 + t^2)^{5/2}} + \frac{\eta^2 t\phi(\eta t/\sqrt{1 + t^2})}{(1 + t^2)^{7/2}}\right\}\right],$$

where $t = x/\sqrt{1 + \eta^2}$. It has been found numerically that the sign of $\dfrac{\mathrm{d}^2\log\{f(x)\}}{\mathrm{d}x^2}$ changes for $x \in (-10, 10)$ for various values of η. Thus, $\log\{f(x)\}$ is not always log-concave. $\square$

Proof of Proposition 7

Let $X \sim \mathcal{MSN}_1(\xi, \Psi, \eta)$. Then, the pdf of X is

$$f_X(x) = 2\phi(x; \xi, \Psi + \eta^2)\Phi\left\{\frac{\eta}{\sqrt{\Psi}}\frac{(x - \xi)}{\sqrt{\Psi + \eta^2 + (x - \xi)^2}}\right\}, \quad x \in \mathbb{R}.$$

For $x < \xi$ and for $\eta > 0$, $\dfrac{\eta}{\sqrt{\Psi}}\dfrac{(x - \xi)}{\sqrt{\Psi + \eta^2 + (x - \xi)^2}} \to -\infty$, and for $x > \xi$, $\dfrac{\eta}{\sqrt{\Psi}}\dfrac{(x - \xi)}{\sqrt{\Psi + \eta^2 + (x - \xi)^2}} \to \infty$, when $\Psi \to 0$. As a consequence, we have, as $\Psi \to 0$: $f_X(x) \to 0$ if $x < \xi$ and $f_X(x) \to 2\phi(x; \xi, \eta^2)$ if $x > \xi$, which completes the proof for $\eta > 0$. The proof for $\eta < 0$ is similar. $\square$

Proof of Proposition 8

The likelihood function for $(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ based on a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from $\boldsymbol{X} \sim \mathcal{MSN}_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ is

$$L(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta}) = \prod_{i=1}^{n} 2\phi_p\left(\boldsymbol{x}_i; \boldsymbol{\xi}, \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top\right) \times \Phi\left\{\frac{1}{\sqrt{1 + \boldsymbol{\eta}^\top \boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}} \frac{\boldsymbol{\eta}^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i - \boldsymbol{\xi})}{\sqrt{1 + (\boldsymbol{x}_i - \boldsymbol{\xi})^\top (\boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}(\boldsymbol{x}_i - \boldsymbol{\xi})}}\right\},$$

which becomes the profile likelihood function for the skewness parameter $\boldsymbol{\eta}$ when $\boldsymbol{\xi}$ and $\boldsymbol{\Psi}$ are fixed. Similar to the $\mathcal{SN}$, it can be argued exactly in the same way as in the proof of Proposition 13 that the maximum likelihood estimator of the skewness parameter is always finite for the $\mathcal{MSN}$ distribution as well. Indeed, as in the $\mathcal{SN}$ case, for the skewing function of the $\mathcal{MSN}$ distribution we have

$$\begin{aligned} 0 &\le \Phi\left\{-\sqrt{(\boldsymbol{x}_i - \boldsymbol{\xi})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i - \boldsymbol{\xi})}\right\} \\ &\le \Phi\left\{\frac{1}{\sqrt{1 + \boldsymbol{\eta}^\top \boldsymbol{\Psi}^{-1}\boldsymbol{\eta}}} \frac{\boldsymbol{\eta}^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i - \boldsymbol{\xi})}{\sqrt{1 + (\boldsymbol{x}_i - \boldsymbol{\xi})^\top (\boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top)^{-1}(\boldsymbol{x}_i - \boldsymbol{\xi})}}\right\} \\ &\le \Phi\left\{\sqrt{(\boldsymbol{x}_i - \boldsymbol{\xi})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i - \boldsymbol{\xi})}\right\} \le 1, \quad \forall\, \boldsymbol{\eta} \in \mathbb{R}^p. \end{aligned}$$

Also, as we already have established in the proof of Proposition 13, $\phi_p(\boldsymbol{x}_i; \boldsymbol{\xi}, \boldsymbol{\Psi} + \boldsymbol{\eta}\boldsymbol{\eta}^\top) \to 0$ for each $i = 1, \ldots, n$ as some (or all) components of $\boldsymbol{\eta}$ tend to $\pm\infty$. Hence, for any fixed value of $\boldsymbol{\xi} \in \mathbb{R}^p$ and $\boldsymbol{\Psi} > 0$, we can say that $L(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta}) \to 0$ whenever some (or all) components of $\boldsymbol{\eta}$ tend to $\pm\infty$. This observation leads us to the conclusion that $L(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\eta})$ is not a monotonically increasing or decreasing function of any of the components of $\boldsymbol{\eta}$. Thus, the profile likelihood of the skewness parameter $\boldsymbol{\eta}$ is always maximized at a finite point for the $\mathcal{MSN}$ family.

Proof of Proposition 9

The non-singularity of the matrix $\boldsymbol{i}_{\boldsymbol{\Psi}\boldsymbol{\Psi}}$ is just a special case of a more general result proven in Hallin and Paindaveine (2006). Thus, the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{0})$ is nonsingular if

$$\boldsymbol{I}_{\xi\eta} = \begin{bmatrix} \boldsymbol{i}_{\xi\xi} & \boldsymbol{i}_{\xi\eta} \\ \boldsymbol{i}_{\xi\eta} & \boldsymbol{i}_{\eta\eta} \end{bmatrix}$$

is nonsingular. Let $\mathbb{E}\left(\frac{\boldsymbol{Z}\boldsymbol{Z}^\top}{\sqrt{1+\boldsymbol{Z}^\top\boldsymbol{Z}}}\right) = \boldsymbol{U}$ and $\mathbb{E}\left(\frac{\boldsymbol{Z}\boldsymbol{Z}^\top}{1+\boldsymbol{Z}^\top\boldsymbol{Z}}\right) = \boldsymbol{V}$. Then, $\boldsymbol{I}_{\boldsymbol{\xi\eta}}$ can be written as

$$\boldsymbol{I}_{\boldsymbol{\xi\eta}} = \boldsymbol{\Psi}^{-1/2}\begin{bmatrix} \mathrm{I}_p & \sqrt{\frac{2}{\pi}}\boldsymbol{U} \\ \sqrt{\frac{2}{\pi}}\boldsymbol{U} & \frac{2}{\pi}\boldsymbol{V} \end{bmatrix}\boldsymbol{\Psi}^{-1/2}.$$

Thus, we conclude that $\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{0})$ is nonsingular iff the matrix $\boldsymbol{V} - \boldsymbol{U}^2$ is nonsingular.

Let $R = |\boldsymbol{Z}|$, $\boldsymbol{W} = \boldsymbol{Z}/R$ and $\boldsymbol{Z}_* = \boldsymbol{Z}/\sqrt{1+R^2}$. Since $\boldsymbol{Z} = R\boldsymbol{W} \sim \mathcal{N}_p(\boldsymbol{0}, \mathrm{I}_p)$, R and $\boldsymbol{W}$ are independent. Also, we know that $\mathbb{E}(\boldsymbol{W}) = \boldsymbol{0}$ and $\mathbb{V}\mathrm{ar}(\boldsymbol{W}) = \mathbb{E}(\boldsymbol{W}\boldsymbol{W}^\top) = (1/p)\mathrm{I}_p$. Now, note that $\boldsymbol{Z}_* = R_*\boldsymbol{W}$, where $R_* = R/\sqrt{1+R^2}$, and so is independent of $\boldsymbol{W}$. Then, we have $\boldsymbol{U} = \mathbb{C}\mathrm{ov}(\boldsymbol{Z}, \boldsymbol{Z}_*) = \mathbb{C}\mathrm{ov}(R\boldsymbol{W}, R_*\boldsymbol{W}) = \mathbb{E}(RR_*)\mathbb{E}(\boldsymbol{W}\boldsymbol{W}^\top) = (1/p)\mathbb{E}(R^2/\sqrt{1+R^2})\mathrm{I}_p$ and $\boldsymbol{V} = \mathbb{V}\mathrm{ar}(\boldsymbol{Z}_*) = \mathbb{V}\mathrm{ar}(R_*\boldsymbol{W}) = \mathbb{E}(R_*^2)\mathbb{E}(\boldsymbol{W}\boldsymbol{W}^\top) = (1/p)E(R_*^2)\mathrm{I}_p$. Hence, $\boldsymbol{V} - \boldsymbol{U}^2$ is positive definite iff $p\mathbb{E}(R_*^2) - \{\mathbb{E}(RR_*)\}^2 > 0$, i.e., iff $p\mathbb{E}\{R^2/(1+R^2)\} - \{\mathbb{E}(R^2/\sqrt{1+R^2})\}^2 > 0$.

Now, by the Cauchy–Schwartz inequality, we get

$$\left\{\mathbb{E}\left(\frac{R^2}{\sqrt{1+R^2}}\right)\right\}^2 \le \mathbb{E}(R^2)\mathbb{E}\left(\frac{R^2}{1+R^2}\right)$$
$$\Rightarrow p\mathbb{E}\{R^2/(1+R^2)\} - \{\mathbb{E}(R^2/\sqrt{1+R^2})\}^2 \ge 0, \text{ as } \mathbb{E}(R^2) = p.$$

Since equality in the previous inequality cannot be achieved in this case, we conclude that $\boldsymbol{V} - \boldsymbol{U}^2$ is positive definite, and, consequently, $\boldsymbol{I}(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{0})$ is also positive definite. □

References

Adcock CJ (2004) Capital asset pricing for UK stocks under the multivariate skew-normal distribution. In: Genton MG (ed) Skew elliptical distributions and their applications: a journey beyond normality. Chapman and Hall, London

Adcock CJ (2005) Exploiting skewness to build an optimal hedge fund with a currency overlay. Eur J Financ 11(5):445–462

Adcock, CJ, Shutes K (2001) Portfolio selection based on the multivariate skew normal distribution. In: A. Skulimowski, Ed., Financial Modelling, Progress & Business Publishers, Krakow, pp 167–177

Arellano-Valle RB, Azzalini A (2008) The centred parametrization for the multivariate skew-normal distribution. J Multivar Anal 99(7):1362–1382

Arellano-Valle RB, Genton MG (2010) An invariance property of quadratic forms in random vectors with a selection distribution, with application to sample variogram and covariogram estimators. Ann Inst Stat Math 62(2):363–381

Arellano-Valle RB, Gómez HW, Quintana FA (2004) A new class of skew-normal distributions. Commun Stat Theory Methods 33(7):1465–1480

Arellano-Valle RB, Bolfarine H, Lachos V (2005) Skew-normal linear mixed models. J Data Sci 3(4):415–438

Arellano-Valle RB, Contreras-Reyes JE, Stehlík M (2017) Generalized skew-normal negentropy and its application to fish condition factor time series. Entropy 19(10):528

Arrué J, Arellano-Valle RB, Gómez HW (2016) Bias reduction of maximum likelihood estimates for a modified skew-normal distribution. J Stat Comput Simul 86(15):2967–2984
Arrué J, Arellano-Valle RB, Gómez HW, Leiva V (2020) On a new type of Birnbaum–Saunders models and its inference and application to fatigue data. J Appl Stat 47(13–15):2690–2710
Azzalini A (1985) A class of distributions which includes the normal ones. Scand J Stat 12:171–178
Azzalini A (2005) The skew-normal distribution and related multivariate families. Scand J Stat 32(2):159–188
Azzalini A, Arellano-Valle RB (2013) Maximum penalized likelihood estimation for skew-normal and skew-t distributions. J Stat Plan Inference 143(2):419–433
Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew normal distribution. J R Stat Soc Ser B Stat Methodol 61(3):579–602
Azzalini A, Capitanio A (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. J R Stat Soc Ser B Stat Methodol 65(2):367–389
Azzalini A, Capitanio A (2014) The skew-normal and related families, vol 3. Cambridge University Press, Cambridge
Azzalini A, Dalla-Valle A (1996) The multivariate skew-normal distribution. Biometrika 83(4):715–726
Bayes CL, Branco MD (2007) Bayesian inference for the skewness parameter of the scalar skew-normal distribution. Braz J Probab Stat 21(2):141–163
Chiogna M (2005) A note on the asymptotic distribution of the maximum likelihood estimator for the scalar skew-normal distribution. Stat Methods Appl 14(3):331–341
Genton MG (2004) Skew-elliptical distributions and their applications: a journey beyond normality. CRC Press, Boca Raton
Genton MG, Loperfido NM (2005) Generalized skew-elliptical distributions and their quadratic forms. Ann Inst Stat Math 57(2):389–401
Ghosh P, Branco MD, Chakraborty H (2007) Bivariate random effect model using skew-normal distribution with application to HIV-RNA. Stat Med 26(6):1255–1267
Gómez HW, Venegas O, Bolfarine H (2007) Skew-symmetric distributions generated by the distribution function of the normal distribution. Environmetrics 18(4):395–407
Hallin M, Paindaveine D (2006) Semiparametrically efficient rank-based inference for shape. I. Optimal rank-based tests for sphericity. Ann Stat 34(6):2707–2756
Henze N (1986) A probabilistic representation of the 'skew-normal' distribution. Scand J Stat 13(4):271–275
Jin L, Xu W, Zhu L, Zhu L (2016) Penalized maximum likelihood estimator for skew normal mixtures. arXiv:1608.01513
Lachos VH, Ghosh P, Arellano-Valle RB (2010) Likelihood based inference for skew-normal independent linear mixed models. Stat Sin 20(1):303–322
Ley C, Paindaveine D (2010) On Fisher information matrices and profile log-likelihood functions in generalized skew-elliptical models. Metron 68(3):235–250
Lin TI, Lee JC, Yen SY (2007) Finite mixture modelling using the skew normal distribution. Stat Sin 17(3):909–927
Magnus JR, Neudecker H (1979) The commutation matrix: some properties and applications. Ann Stat 7(2):381–394
Mardia KV (1970) Measures of multivariate skewness and kurtosis with applications. Biometrika 57(3):519–530
McNeil AJ, Frey R, Embrechts P (2015) Quantitative risk management: concepts, techniques and tools-revised edition. Princeton University Press, Princeton
Pewsey A (2000) Problems of inference for Azzalini's skew-normal distribution. J Appl Stat 27(7):859–870
R Core Team (2021) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/
Rotnitzky A, Cox DR, Bottai M, Robins J (2000) Likelihood-based inference with singular information matrix. Bernoulli 6(2):243–284
Sartori N (2006) Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. J Stat Plan Inference 136(12):4259–4275
Yip CMA (2018) Statistical characteristics and mapping of near-surface and elevated wind resources in the middle east. Ph.D. Thesis, KAUST