Addressing non-normality in multivariate analysis using the t-distribution

Felipe Osorio[1*], Manuel Galea[2], Claudio Henríquez[2] and Reinaldo Arellano-Valle[2]

[1*]Departamento de Matemática, Universidad Técnica Federico Santa María, Avenida España 1680, Valparaíso, Chile.
[2]Departamento de Estadística, Pontificia Universidad Católica de Chile, Avenida Vicuña Mackena 4860, Santiago, Chile.

*Corresponding author(s). E-mail(s): felipe.osorios@usm.cl; Contributing authors: mgalea@mat.uc.cl; cnhenriquez@mat.uc.cl; reivalle@mat.uc.cl;

Abstract

The main aim of this paper is to propose a set of tools for assessing non-normality taking into consideration the class of multivariate t-distributions. Assuming second moment existence, we consider a reparameterized version of the usual t distribution, so that the scale matrix coincides with covariance matrix of the distribution. We use the local influence procedure and the Kullback-Liebler divergence measure to propose quantitative methods to evaluate deviations from the normality assumption. In addition, the possible non-normality due to the presence of both skewness and heavy tails is also explored. Our findings based on two real datasets are complemented by a simulation study to evaluate the performance of the proposed methodology on finite samples.

Keywords: Kullback-Liebler divergence, Local influence, Negentropy, Outliers

1 Introduction

The need to analyze multivariate continuous observations arises frequently in several areas of knowledge. In particular, most statistical inferences associated with estimation and hypothesis testing in this context are based on the assumption of normality (see, for instance, Anderson, 2003; Härdle and Simar, 2012, among others). Although most of these developments are due to the simplicity and usefulness of the normal distribution, it is well known that statistical inference based on this distribution is susceptible to the presence of atypical data. The fourth-order moments of the normal distribution are determined by the first and second moments of the distribution, making it impossible to adjust the kurtosis of the observations. Specifically, it is necessary to consider alternative distributions that allow us to circumvent this limitation of the multivariate normal distribution. In this direction, several authors have suggested the use distributions belonging to the elliptically contoured class. For example, in Fang and Zhang (1990) and Gupta et al. (2013) the definition, properties and statistical inference for multivariate elliptical models are described in detail, with particular emphasis on dependent elliptical models. Whereas in this paper, we concentrate on the independent multivariate t model, which would provide a robust estimation procedure against possible outliers in data (see, for instance Lange et al., 1989). Moreover, the t distribution incorporates an additional parameter that enables the modeling of data with high kurtosis.

The study of non-normality using entropy-based measures is rather limited, with Gómez-Villegas et al. (2011) being a notable exception. This work proposes a set of tools for the detection of non-normality based on the multivariate t distribution. The local influence procedure and Kullback-Liebler divergence are used as measures to evaluate deviations from normality associated with a sample of observations from a continuous population following a multivariate t distribution. Specifically, we consider a reparameterization of the t distribution with a finite second moment. This enables a more direct comparison with the normal distribution (see Sutradhar, 1993 and Bolfarine and Galea, 1996). In addition, a very simple graphical procedure to evaluate the model assumptions is proposed taking advantage of the distribution associated with the Mahalanobis distances under the multivariate t-distribution. This provides a methodology to quantify the impact of non-normality on the statistical modeling by providing a robust approach based on a reparametrization of the multivariate t-distribution. Although the paper focuses on measures of non-normality due to kurtosis, we also include some measures of non-normality due to possible skewness present in the observations.

The paper unfolds as follows. In Section 2, we present the multivariate t distribution which is based on the work by Fiorentini et al. (2003), who used a reparameterization of the degrees of freedom, as suggested by Lange et al. (1989). An explicit expression for the Fisher information matrix is presented. Two procedures for assessing non-normality using measures based on the Kullback-Liebler divergence and the local influence approach are described in detail in Section 3, whereas Section 4 a quantile-quantile (QQ) plot of transformed Mahalanobis distances is proposed. Section 5 is dedicated to illustrating the methodology used in the paper considering two real datasets from environmental and biological areas, and we report the results of a simulation study to evaluate the performance of the proposed methodology on finite samples. Finally, Section 6 draws conclusions and discusses perspectives for future work. Some details of the results presented throughout the paper are described in the Appendices.

2 The multivariate *t* distribution

We say that a p-dimensional random vector $\boldsymbol{X} = (X_1, \dots, X_p)^\top$ has a multivariate t distribution, with a $\boldsymbol{\mu}$ mean vector, covariance matrix $\boldsymbol{\Sigma}$, and $0 \le \eta < 1/2$ shape parameter, if its pdf is given by

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = K_p(\eta)|\boldsymbol{\Sigma}|^{-1/2}\big(1 + c(\eta)\delta^2\big)^{-\frac{1}{2\eta}(1+\eta p)}, \qquad \boldsymbol{x} \in \mathbb{R}^p, \tag{1}$$

where

$$K_p(\eta) = \Big(\frac{c(\eta)}{\pi}\Big)^{p/2} \frac{\Gamma(\frac{1}{2\eta}(1+\eta p))}{\Gamma(\frac{1}{2\eta})}, \tag{2}$$

in which $c(\eta) = \eta/(1-2\eta)$ and $\delta^2 = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ is the Mahalanobis distance. Besides, we asume that $\boldsymbol{\Sigma}$ is positive definite throughout the paper. If a random vector has a density function (1), we shall denote $\boldsymbol{X} \sim \mathsf{T}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \eta)$. The multivariate t distribution parameterization given in (1) is introduced mainly because $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ correspond to the mean vector and covariance matrix, respectively. On the other hand, η is the inverse of the degrees of freedom, namely, the shape parameter, which allows the kurtosis of the distribution to be adjusted. Using this mean and variance parameterization, Sutradhar (1993) proposed a $C(\alpha)$ test for testing of the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$, and Bolfarine and Galea (1996) used this parameterization in comparative calibration. The parameterization of the degrees of freedom used in this work was originally suggested by Lange et al. (1989), who stated that "…and inferences about ν, the degrees of freedom parameter, itself might be improved by transforming to $1/\nu$ or $\log(\nu)$". Later, Fiorentini et al. (2003) and more recently Galea et al. (2020), used density (1) to model financial data. Note that this parameterization enables a direct comparison between the maximum likelihood (ML) estimators of the mean vector and covariance matrix and the obtained versions under the normal model, which would in turn simplify the interpretation of those parameter estimates. Some properties of the multivariate t distribution given in (1) are deferred to the Supplementary Material.

2.1 Score function and Fisher information

Consider $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ as a random sample from a multivariate t distribution given by (1). In this case, the log-likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\mu}^\top, \boldsymbol{\phi}^\top, \eta)^\top$ with $\boldsymbol{\phi} = \operatorname{vech}(\boldsymbol{\Sigma})$ indicating different elements in the matrix $\boldsymbol{\Sigma}$ (see for instance, Magnus and Neudecker, 1999) takes the form

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}), \tag{3}$$

where the ith component of the log-likelihood function is given by

$$\mathcal{L}_i(\boldsymbol{\theta}) = \log K_p(\eta) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2\eta}(1+\eta p)\log\big(1 + c(\eta)\delta_i^2\big),$$

with $K_p(\eta)$ defined in Equation (2) and $\delta_i^2 = (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$ for $i = 1, \dots, n$. From (3), we can see that the ML estimator for $\boldsymbol{\theta}$ does not have an explicit form, and thus, iterative procedures must be used. The EM algorithm for the parameterization defined in this work is reviewed in the Supplementary Material.

We have that the score function takes the form $\boldsymbol{U}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\theta})$, where an individual score can be written in a partitioned form as $\boldsymbol{U}_i(\boldsymbol{\theta}) = (\boldsymbol{U}_i^\top(\boldsymbol{\mu}), \boldsymbol{U}_i^\top(\boldsymbol{\phi}), U_i(\eta))^\top$, with

$$\boldsymbol{U}_i(\boldsymbol{\mu}) = v_i \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}), \tag{4}$$

$$\boldsymbol{U}_i(\boldsymbol{\phi}) = \frac{1}{2}\boldsymbol{D}_p^\top \operatorname{vec}(v_i \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}), \tag{5}$$

$$U_i(\eta) = \frac{1}{2\eta^2}\Big\{\psi\Big(\frac{1}{2\eta}\Big) - \psi\Big(\frac{1+p\eta}{2\eta}\Big) + c(\eta)(p - v_i\delta_i^2) + \log q_i\Big\}, \tag{6}$$

where $v_i = (1/\eta + p)/(1/c(\eta) + \delta_i^2) = (1+p\eta)q_i^{-1}/(1-2\eta)$ corresponds to a weight function that is a decreasing function of the distance δ_i^2, (Kent et al., 1994; Kent and Tyler, 1991; Maronna, 1976), $q_i = 1 + c(\eta)\delta_i^2$, for $i = 1, \dots, n$, $\psi(z) = \mathrm{d}\log\Gamma(z)/\,\mathrm{d}z$ is the digamma function, and $\boldsymbol{D}_p \in \mathbb{R}^{p^2 \times p(p+1)/2}$ denotes the duplication matrix of order p, (Magnus and Neudecker, 1999).

Therefore, the Fisher matrix information for $\boldsymbol{\theta}$ based on the log-likelihood defined in (3) assumes the form (see Appendix A)

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\mathcal{I}}_{\mu\mu}(\boldsymbol{\theta}) & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\mathcal{I}}_{\phi\phi}(\boldsymbol{\theta}) & \boldsymbol{\mathcal{I}}_{\phi\eta}(\boldsymbol{\theta}) \\ \boldsymbol{0} & \boldsymbol{\mathcal{I}}_{\phi\eta}^\top(\boldsymbol{\theta}) & \mathcal{I}_{\eta\eta}(\boldsymbol{\theta}) \end{pmatrix}, \tag{7}$$

where

$$\boldsymbol{\mathcal{I}}_{\mu\mu}(\boldsymbol{\theta}) = c_\mu(\eta)\boldsymbol{\Sigma}^{-1},$$

$$\boldsymbol{\mathcal{I}}_{\phi\phi}(\boldsymbol{\theta}) = \frac{1}{4}\boldsymbol{D}_p^\top\{2c_\phi(\eta)(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{N}_p + (c_\phi(\eta) - 1)(\operatorname{vec}\boldsymbol{\Sigma}^{-1})(\operatorname{vec}\boldsymbol{\Sigma}^{-1})^\top\}\boldsymbol{D}_p,$$

$$\boldsymbol{\mathcal{I}}_{\phi\eta}(\boldsymbol{\theta}) = -\frac{c(\eta)(p+2)}{(1+p\eta)(1+(p+2)\eta)}\boldsymbol{D}_p^\top \operatorname{vec}\boldsymbol{\Sigma}^{-1},$$

$$\mathcal{I}_{\eta\eta}(\boldsymbol{\theta}) = \frac{1}{4\eta^4}\Big\{\psi'\Big(\frac{1}{2\eta}\Big) - \psi'\Big(\frac{1+p\eta}{2\eta}\Big) + 2pc(\eta)^2\Big(\frac{4(p+2)\eta^2 - p\eta - 1}{(1+p\eta)(1+(p+2)\eta)}\Big)\Big\},$$

in which $\boldsymbol{N}_p = \frac{1}{2}(\boldsymbol{I}_{p^2} + \boldsymbol{K}_p)$, and $\boldsymbol{K}_p$ is the commutation matrix of order $p^2 \times p^2$, (Magnus and Neudecker, 1999), $c_\mu(\eta) = c_\phi(\eta)/(1-2\eta)$, $c_\phi(\eta) = (1+p\eta)/(1+(p+2)\eta)$ and $\psi'(z) = \mathrm{d}\,\psi(z)/\,\mathrm{d}z$ is the trigamma function (Abramowitz and Stegun, 1970, p. 260). Note that $c_\mu(\eta)$ and $c_\phi(\eta) \to 1$ when $\eta \to 0$. In addition, we have that $\boldsymbol{N}_p\boldsymbol{D}_p = \boldsymbol{D}_p$, thus recovering the expressions corresponding to the normal case. Also, as $\eta \to 0$, we have that $c(\eta) \to 0$ and as shown in Appendix A, $\mathcal{I}_{\eta\eta}(\boldsymbol{\theta}) \to p(p+2)(2p+5)/2$. Consequently,

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) \to \text{block diag}\big(\boldsymbol{\Sigma}^{-1},\ \tfrac{1}{2}\boldsymbol{D}_p^\top(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\boldsymbol{D}_p,\ \tfrac{1}{2}p(p+2)(2p+5)\big),$$

as $\eta \to 0$. Therefore, the Fisher information matrix is no singular in the neighborhood of normality, which allows to maintain the usual asymptotic behavior of the ML estimator.

3 Measuring non-normality

In multivariate analysis it is usual to specify a parametric working model, for the sake of simplicity we consider the class of continuous and elliptically symmetric distributions, with mean vector $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$ and shape parameter η;

$$\mathcal{C}_g = \{f(\boldsymbol{x};\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\mu} \in \mathbb{R}^p \text{ and } \boldsymbol{\Sigma} \in \mathcal{S}^+\},$$

generated by g, where $f(\boldsymbol{x};\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = K_g(\eta)|\boldsymbol{\Sigma}|^{-1/2}g(\delta^2)$, g is a known real-valued function having a strictly negative derivative, $K_g(\eta)$ is a normalizing constant, $\delta^2 = (\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ represents the Mahalanobis distance and $\mathcal{S}^+$ is the class of symmetric and positive-definite matrices of dimension $p \times p$. The multivariate normal and t distributions, considered in this paper, belong to $\mathcal{C}_g$. In this section we use some statistical tools to evaluate non-normality in the class $\mathcal{C}_g$, with emphasis on the multivariate t distribution. Specifically, we use divergence measures to describe the discrepancy between the normal and t distributions. Additionally, we use the local influence methodology to assess the sensitivity of the ML estimator to small perturbations of the normality, using the multivariate t distribution, which is called t-perturbation.

3.1 The Kullback-Leibler divergence

Let $\boldsymbol{Z} \in \mathbb{R}^p$ be a random vector with probability density function (pdf) $f_Z(\boldsymbol{z})$. The Shannon (1948) entropy, is given by

$$H(\boldsymbol{Z}) = -\,\mathrm{E}[\log f_Z(\boldsymbol{z})] = -\int_{\mathbb{R}^p}\{\log f_Z(\boldsymbol{z})\}f_Z(\boldsymbol{z})\,\mathrm{d}\,\boldsymbol{z}.$$

A direct calculation (see Lemma 3 in Supplementary Material), shows that for the normal and t distributions the Shannon entropy has explicit form and are given in the following lemma,

Lemma 1 *If* $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $\boldsymbol{Y} \sim \mathsf{T}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \eta)$*, then their Shannon entropies are given by*

(i) $H(\boldsymbol{X}) = \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{p}{2}(1 + \log 2\pi)$.

(ii) $H(\boldsymbol{Y}) = -\log K_p(\eta) + \frac{1}{2}|\boldsymbol{\Sigma}| + \frac{1}{2\eta}(1 + \eta p)\Big\{\psi\Big(\frac{1+\eta p}{2\eta}\Big) - \psi\Big(\frac{1}{2\eta}\Big)\Big\}$.

Suppose now that $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^p$ are two random vectors with pdf's $f_X(\boldsymbol{x})$ and $f_Y(\boldsymbol{y})$, respectively, which are assumed to have the same support. Related to the entropy concept we can also find divergence measures between the distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$. The most well-known of these measures is the so called Kullback-Leibler (KL) divergence proposed by Kullback (1951) as

$$d_{KL}(f_X, f_Y) = \int_{\mathbb{R}^p} f_X(\boldsymbol{x}) \log\left\{\frac{f_X(\boldsymbol{x})}{f_Y(\boldsymbol{x})}\right\} \mathrm{d}\,\boldsymbol{x} = \mathrm{E}_X\left[\log\left\{\frac{f_X(\boldsymbol{X})}{f_Y(\boldsymbol{X})}\right\}\right],$$

which measures the divergence of f_Y from f_X and where the expectation is defined with respect to the pdf $f_X(\boldsymbol{x})$ of the random vector $\boldsymbol{X}$. Although the KL divergence measures the distance between two densities, it is not a distance measure. The KL from f_X to f_Y is generally not the same as the KL from f_Y to f_X. This is a quasi-metric. Indeed, which is relevant from the statistical point of view is that, $d_{KL}(f_X, f_Y) \geq 0$ and $d_{KL}(f_X, f_Y) = 0$ if and only if $f_X = f_Y$.

A concept related to the entropy is the *Negentropy*. In statistics the negative entropy or negentropy is used to quantify the non-normality of a random vector and measures the difference in entropy between a given distribution and the normal distribution, both with the same vector mean and covariance matrix. The Negentropy is always nonnegative and vanishes if and only if the random vector has a normal distribution. Negentropy of a random vector $\boldsymbol{Y}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, is defined as (Contreras-Reyes and Arellano-Valle, 2012; Gao and Zhang, 2010),

$$H_N(\boldsymbol{Y}) = H(\boldsymbol{X}) - H(\boldsymbol{Y}),$$

where $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The following lemma summarizes these results for normal and t distributions.

Lemma 2 *If* $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $\boldsymbol{Y} \sim \mathsf{T}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \eta)$*, then*

(i) $d_{KL}(f_N, f_T) = -\log K_p(\eta) - \frac{p}{2}(1+\log 2\pi) + \frac{1}{2\eta}(1+p\eta)\,\mathrm{E}_N\{\log(1+c(\eta)\delta^2)\}$,

(ii) $d_{KL}(f_T, f_N) = \frac{p}{2}(1+\log 2\pi) + \log K_p(\eta) - \frac{1}{2\eta}(1+\eta p)\Big\{\psi\Big(\frac{1+\eta p}{2\eta}\Big) - \psi\Big(\frac{1}{2\eta}\Big)\Big\}$,

(iv) $H_N(\boldsymbol{Y}) = \frac{p}{2}(1 + \log 2\pi) + \log K_p(\eta) - \frac{1}{2\eta}(1 + \eta p)\Big\{\psi\Big(\frac{1+\eta p}{2\eta}\Big) - \psi\Big(\frac{1}{2\eta}\Big)\Big\}$,

where $\mathrm{E}_N\{\log(1+c(\eta)\delta^2)\} = \mathrm{E}\{\log(1+Y)\}$, *with* $Y \sim \mathsf{Gamma}\left(\frac{p}{2}, \frac{1}{2c(\eta)}\right)$.

Note that in this case $d_{KL}(f_T, f_N) = H_N(\boldsymbol{Y})$. Using a Taylor series expansion to the second order it is possible to show that, $d_{KL}(f_N, f_T) \approx \mathcal{I}_{\eta\eta}(\boldsymbol{\theta})\eta^2/2$, for $0 \leq \eta < 1/2$, with $\mathcal{I}_{\eta\eta}(\boldsymbol{\theta})$ being the Fisher information matrix for η given in (7).

Using the properties of the ML estimator and the delta method (Serfling, 2009) we have the following result on the asymptotic distribution of $\widehat{\eta}$ and of the negentropy, $H_N(\boldsymbol{Y})$.

Lemma 3 *The asymptotic distribution of* $\widehat{\eta}$ *and* $h(\widehat{\eta}) = H_N(\boldsymbol{Y})\big|_{\eta=\widehat{\eta}}$ *is respectively, given by*

$$\sqrt{n}(\widehat{\eta} - \eta) \xrightarrow{D} \mathsf{N}(0, \sigma_\eta^2),$$

$$\sqrt{n}(h(\widehat{\eta}) - h(\eta)) \xrightarrow{D} \mathsf{N}(0, [h'(\eta)]^2\sigma_\eta^2),$$

with $\sigma_\eta^2 = (\mathcal{I}_{\eta\eta}(\boldsymbol{\theta}) - \boldsymbol{\mathcal{I}}_{\phi\eta}^\top(\boldsymbol{\theta})\boldsymbol{\mathcal{I}}_{\phi\phi}^{-1}(\boldsymbol{\theta})\boldsymbol{\mathcal{I}}_{\phi\eta}(\boldsymbol{\theta}))^{-1}$, *and*

$$h'(\eta) = \frac{1}{2\eta^2}\Big\{pc(\eta) + \frac{1+p\eta}{2\eta}\Big(\psi'\Big(\frac{1+p\eta}{2\eta}\Big) - \psi'\Big(\frac{1}{2\eta}\Big)\Big)\Big\}.$$

3.2 Local influence approach

The method of local influence was introduced by Cook (1986) as a general tool for assessing the influence of local departures from the assumptions underlying the statistical model. A perturbation scheme is introduced into the postulated model (working model) through a perturbation vector, and an influence measure is constructed using the basic geometric idea of curvature of the likelihood displacement surface. In our case, we assume that the working model corresponds to the multivariate normal distribution and we perturbed normality using the multivariate t distribution, generating a distribution with heavier tails than normal. In fact, the density of the postulated model for a sample of size n is given by

$$f_N(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^n (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp(-\delta_i^2/2),$$

where $\delta_i^2 = (\boldsymbol{x}_i - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$, $i = 1, \ldots, n$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n\times p}$. Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top$ be a perturbation vector varying in $\Lambda \subseteq \mathbb{R}^n$ and $\mathcal{M} = \{f_T(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\eta}) : \boldsymbol{\eta} \in \Lambda\}$, the perturbed model, where

$$f_T(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\eta}) = \prod_{i=1}^n K_p(\eta_i)|\boldsymbol{\Sigma}|^{-1/2}\big(1 + c(\eta_i)\delta_i^2\big)^{-\frac{1}{2\eta_i}(1+\eta_i p)} \qquad (8)$$

The corresponding perturbed log-likelihood function is given by

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\eta}) = \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}|\eta_i), \tag{9}$$

where

$$\mathcal{L}_i(\boldsymbol{\theta}|\eta_i) = \log K_p(\eta_i) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2\eta_i}(1+\eta_i p)\log\left(1 + c(\eta_i)\delta_i^2\right),$$

with $K_p(\eta)$ defined in Equation (2). The influence of the t-perturbation $\boldsymbol{\eta}$ on the ML estimator can be evaluated by the likelihood displacement given by $LD(\boldsymbol{\eta}) = 2[\mathcal{L}(\widehat{\boldsymbol{\theta}}) - \mathcal{L}(\widehat{\boldsymbol{\theta}}_\eta)]$, where $\mathcal{L}(\boldsymbol{\theta})$, denotes the log-likelihood function of the postulated model, $\widehat{\boldsymbol{\theta}}$ is the ML estimator of $\boldsymbol{\theta}$ in the normal model and $\widehat{\boldsymbol{\theta}}_\eta$ is the ML estimator of $\boldsymbol{\theta}$ in under the multivariate t distribution, that is the perturbed model $\mathcal{M}$. Cook (1986) proposes to study the local behavior of $LD(\boldsymbol{\eta})$ around $\boldsymbol{\eta}_0$ and shows that the normal curvature C_l of $LD(\boldsymbol{\eta})$ at $\boldsymbol{\eta}_0$ in direction of some unit vector $\boldsymbol{l}$, is given by $C_l = C_l(\boldsymbol{\theta}) = 2|\boldsymbol{l}^\top \boldsymbol{\Delta}^\top \ddot{\boldsymbol{L}}^{-1} \boldsymbol{\Delta}\boldsymbol{l}|$, with $\|\boldsymbol{l}\| = 1$, where $\ddot{\boldsymbol{L}} = \partial^2 \mathcal{L}_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ and $\boldsymbol{\Delta} = \partial^2\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\eta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\eta}^\top$ are evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\eta} = \boldsymbol{\eta}_0$, where $\boldsymbol{\eta}_0$ is the vector of null perturbation.

Let $\boldsymbol{l}_{\max}$ be the direction of greatest curvature associated with the matrix $\boldsymbol{F} = \boldsymbol{\Delta}^\top \ddot{\boldsymbol{L}}^{-1}\boldsymbol{\Delta}$. Thus, the index plot of $|\boldsymbol{l}_{\max}|$ can reveal if the t-perturbation has influence on $LD(\boldsymbol{\eta})$, in the neighborhood of the normal model ($\boldsymbol{\eta}_0$), Cook (1986). We may also consider the index plot of $C_i = 2|f_{ii}|$ to evaluate the presence of "influential observations", where f_{ii} is the ith element of the diagonal of the matrix $\boldsymbol{F}$. In view that C_l is not invariant under uniform changes of scale, Poon and Poon (1999) proposed the conformal normal curvature $B_l = C_l/\operatorname{tr}(2\boldsymbol{F})$ (see Zhu and Lee, 2001). An interesting property of conformal curvature is that for any unitary direction $\boldsymbol{l}$, it follows that $0 \le B_l \le 1$. We denote by $B_i = 2|f_{ii}|/\operatorname{tr}(2\boldsymbol{F})$ the conformal curvature in the unitary direction with ith entry 1 and all other entries 0. According to Zhu and Lee (2001), the ith observation is potentially influential if $B_i > \bar{B} + 2\operatorname{sd}(B)$, where $\bar{B} = \sum_{i=1}^n B_i/n$ and $\operatorname{sd}(B)$ is the standard deviation of $B_1, \dots, B_n$.

We have that the score function takes the form $\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{\eta}) = \sum_{i=1}^n \boldsymbol{U}_{i\eta}(\boldsymbol{\theta})$, where an individual score can be written in a partitioned form as $\boldsymbol{U}_{i\eta}(\boldsymbol{\theta}) = (\boldsymbol{U}_{i\eta}^\top(\boldsymbol{\mu}), \boldsymbol{U}_{i\eta}^\top(\boldsymbol{\phi}))^\top$, with

$$\begin{aligned}
\boldsymbol{U}_{i\eta}(\boldsymbol{\mu}) &= v_i \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}), \\
\boldsymbol{U}_{i\eta}(\boldsymbol{\phi}) &= \frac{1}{2}\boldsymbol{D}_p^\top \operatorname{vec}\left(v_i\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\right),
\end{aligned}$$

where $v_i = (1/\eta_i + p)/(1/c(\eta_i) + \delta_i^2) = (1 + p\eta_i)/(1 - 2\eta_i + \eta_i\delta_i^2)$ for $i = 1, \dots, n$ corresponds to a weight function that is a decreasing function of the distance δ_i^2, for $i = 1, \dots, n$. Taking derivatives of (9) with respect to η_i and evaluating at $\boldsymbol{\eta}_0 = \mathbf{0}$ and $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, we can write $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \dots, \boldsymbol{\Delta}_n)$, with $\boldsymbol{\Delta}_i = (\boldsymbol{\Delta}_{i\mu}^\top, \boldsymbol{\Delta}_{i\phi}^\top)^\top$, where

$$\begin{aligned}\boldsymbol{\Delta}_{i\mu} &= (p + 2 - \widehat{\delta}_i^2)\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}), \\ \boldsymbol{\Delta}_{i\phi} &= \tfrac{1}{2}(p + 2 - \widehat{\delta}_i^2)\boldsymbol{D}_p^\top \operatorname{vec}\big(\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top \widehat{\boldsymbol{\Sigma}}^{-1}\big),\end{aligned}$$

with $\widehat{\delta}_i^2 = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})$, $i = 1, \dots, n$, $\overline{\boldsymbol{x}} = (1/n)\sum_{i=1}^n \boldsymbol{x}_i$ and $\widehat{\boldsymbol{\Sigma}} = (1/n)\sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top$. It is well known that the sample covariance matrix, $\widehat{\boldsymbol{\Sigma}}$ is positive definite with probability 1 provided that $p \leq n - 1$. If $\eta_i = \eta$ for $i = 1, \dots, n$, and considering a Taylor expansion of order two of $\mathcal{L}(\widehat{\boldsymbol{\theta}}_\eta)$ around $\eta = \eta_0$ it follows that

$$LD_1(\eta) \approx \boldsymbol{\Delta}^\top(-\ddot{\boldsymbol{L}})^{-1}\boldsymbol{\Delta}(\eta - \eta_0)^2, \qquad 0 \leq \eta < 1/2,$$

where

$$\boldsymbol{\Delta} = \sum_{i=1}^n \begin{pmatrix} \boldsymbol{\Delta}_{i\mu} \\ \boldsymbol{\Delta}_{i\phi} \end{pmatrix}.$$

The observed information matrix for $\boldsymbol{\theta} = (\boldsymbol{\mu}^\top, \boldsymbol{\phi}^\top)^\top$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ is the $p(p+3)/2 \times p(p+3)/2$ matrix,

$$-\ddot{\boldsymbol{L}} = n \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\boldsymbol{D}_p^\top(\widehat{\boldsymbol{\Sigma}}^{-1} \otimes \widehat{\boldsymbol{\Sigma}}^{-1})\boldsymbol{D}_p \end{pmatrix}$$

To verify if the t-perturbation (see Equation (8)), is an appropriate perturbation scheme, we can use the methodology proposed by Zhu et al. (2007). Let $\boldsymbol{G}(\boldsymbol{\eta}) = \mathrm{E}_\eta(\boldsymbol{U}_\eta \boldsymbol{U}_\eta^\top)$, where $\boldsymbol{U}_\eta$ denoted the score vector under the perturbed log-likelihood function (9). The ith element of the score function is given by

$$U_\eta(\eta_i) = \frac{1}{2\eta_i^2}\Big\{\psi\Big(\frac{1}{2\eta_i}\Big) - \psi\Big(\frac{1 + p\eta_i}{2\eta_i}\Big) + c(\eta_i)(p - v_i\delta_i^2) + \log(1 + c(\eta_i\delta_i^2)\Big\},$$

where $v_i = (1/\eta_i + p)/(1/c(\eta_i) + \delta_i^2)$, for $i = 1, \dots, n$. In our case we have $\boldsymbol{G}(\boldsymbol{\eta}) = \operatorname{diag}(g_{11}(\eta_1), \dots, g_{nn}(\eta_n))$, with

$$\begin{aligned} g_{ii}(\eta_i) = -\frac{1}{2\eta_i^2}\Big\{&\frac{p}{(1 - 2\eta_i)^2}\Big(\frac{1 + p\eta_i(1 - 4\eta_i) - 8\eta_i^2}{(1 + \eta_i p)(1 + (p+2)\eta_i)}\Big) \\ &+ \frac{1}{2\eta_i^2}\Big(\psi'\Big(\frac{1 + p\eta_i}{2\eta_i}\Big) - \psi'\Big(\frac{1}{2\eta_i}\Big)\Big)\Big\},\end{aligned}$$

for $i = 1, \dots, n$. The perturbation $\boldsymbol{\eta}$ is appropriate if it satisfies $\boldsymbol{G}(\boldsymbol{\eta}_0) = c\boldsymbol{I}_n$, with $c > 0$. In our case, $c = p(p+2)(2p+5)/2$, and the t-perturbation is an appropriate perturbation scheme. The evaluation of the $\boldsymbol{G}(\boldsymbol{\eta})$ matrix at $\boldsymbol{\eta} = \boldsymbol{\eta}_0$ should be understood as $\lim_{\eta \to 0^+} \boldsymbol{G}(\boldsymbol{\eta})$.

4 Goodness of fit

Any statistical analysis should include a critical analysis of the model assumptions. Next, we describe a simple graphical device for model checking, using the Mahalanobis distance. We have that, the random variables

$$F_i = \Big(\frac{1}{1-2\eta}\Big)\frac{\delta_i^2}{p} \overset{\text{iid}}{\sim} F(p, 1/\eta)$$

for $i = 1, \dots, n$. Substituting the ML estimators yields $\widehat{F}_i = F_i(\widehat{\boldsymbol{\theta}})$, which has asymptotically the same F distribution as F_i, $i = 1, \dots, n$. Using the Wilson-Hilferty (1931) approximation,

$$z_i = \frac{\Big(1 - \frac{2\eta}{9}\Big)\widehat{F}_i^{1/3} - \Big(1 - \frac{2}{9p}\Big)}{\sqrt{\frac{2\eta}{9}\widehat{F}_i^{2/3} + \frac{2}{9p}}}, \qquad i = 1, \dots, n,$$

which approximately follows a standard normal distribution. Thus, a QQ-plot of the transformed distances $\{z_1, \dots, z_n\}$ can be used to evaluate the fit of the multivariate t-distribution. For $\eta = 0$, the transformed distances are simplified to $z_i = \{\widehat{F}_i^{1/3} - (1 - 2/(9p))\}/\sqrt{2/(9p)}$, and can be used to assess of fit of the multivariate normal distribution. Additionally, the Mahalanobis distance can be used for multivariate outlier detection. In addition, larger than expected values of the modified Mahalanobis distance, $\widehat{F}_i$, $i = 1, \dots, n$, identify outlying cases (see Lange et al., 1989). It should be stressed that these graphical diagnostics can reveal model inadequacy. The technique will be illustrated in Section 5.

5 Numerical experiments

In this section, we study the performance of the proposed methodology by considering three real data sets. In addition, we present a simulation study to evaluate the performance of non-normality detection procedure using the influence diagnostics. The non-normality measures, local influence and goodness of fit for the multivariate t-distribution described in previous sections has been implemented in R, and the codes are available on the github webpage https://github.com/faosorios/nonnormality.

5.1 Monte Carlo study

We reported our findings from a Monte Carlo simulation study, which was designed to evaluate the performance of the local influence procedure described in the previous section on small samples. 1,000 datasets with sample size of $n = 50, 200$ and 1,000 were created from a normal distribution, $\mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $p = 2, 5, 20, 50$ and 100. Based on the simulation study reported by Leal et al.

(2019), we considered

$$\boldsymbol{\mu} = \mathbf{0}, \qquad \boldsymbol{\Sigma} = (1 - 0.95)\boldsymbol{I}_p + 0.95\,\mathbf{1}_p\mathbf{1}_p^\top.$$

It is well known that a single outlying observation can strongly affect the assumption of normality. To introduce an outlier, for each dataset, a single observation of the second variable X_{12} was changed to $X_{12} + \delta$, where $\delta = 0.5, 1.5, 2.0, 2.5, 3.0$ and 3.5. Let $\boldsymbol{l} = |\boldsymbol{l}_{\max}|$, thus we detect the synthetic outlier if l_1 is greater than the following threshold $\bar{l} + 2\,\mathrm{sd}(\boldsymbol{l})$, where $\bar{l}$ and $\mathrm{sd}(\boldsymbol{l})$ are the average and standard deviation of $\boldsymbol{l}$, respectively. Table 1 contains the outlier detection percentages computed using this threshold for different values of δ.

Table 1 Outlier detection percentage using the t-perturbation to detect non-normality

p	n	δ						
		0.5	1.0	1.5	2.0	2.5	3.0	3.5
2	50	16.9	65.4	96.0	99.9	100.0	100.0	100.0
	200	12.7	61.5	96.2	99.9	100.0	100.0	100.0
	1000	13.8	64.7	95.0	100.0	100.0	100.0	100.0
5	50	15.0	67.9	98.6	100.0	100.0	100.0	100.0
	200	10.3	68.8	98.5	100.0	100.0	100.0	100.0
	1000	9.4	61.9	98.9	100.0	100.0	100.0	100.0
20	50	6.5	27.9	74.0	96.0	99.7	100.0	100.0
	200	4.4	38.2	90.0	99.8	100.0	100.0	100.0
	1000	2.2	34.5	90.9	99.9	100.0	100.0	100.0
50	200	1.2	13.2	57.7	95.5	100.0	100.0	100.0
	1000	0.6	11.1	66.1	98.1	99.9	100.0	100.0
100	200	0.9	4.9	22.3	58.7	89.2	99.6	100.0
	1000	0.3	4.3	33.4	83.6	99.6	100.0	100.0

As expected, the outlier detection percentages improve as δ increases. It should be noted that inasmuch as p increases, the percentage of detected outliers decreases. In Table 1, we have excluded cases $n = 50$ and $p = 50, 100$, because in this situation the sample covariance matrix is not invertible. Our findings suggest that t-perturbation is quite efficient in detecting non-normality caused by outliers. It should be highlighted that our results are consistent with the outcomes of the Monte Carlo experiment developed by Leal et al. (2019).

5.2 Real-life examples

5.2.1 Wind speed data

The wind speed dataset consists of $n = 278$ hourly average wind speed in the Pacific North-West of the United States collected at three meteorological towers approximately located on a line and ordered from west to east: Goodnoe Hills (gh), Kennewick (kw), and Vansycle (vs). The data were collected from

25 February to 30 November 2003 recorded at midnight, a time when wind speeds tend to peak. More information about the data can be found in Azzalini and Genton (2008) (see also, Arellano-Valle et al., 2018).

Following Arellano-Valle et al. (2018), let $\boldsymbol{X}(t)$ the three-dimensional vector of wind speed at the towers (gh, kw and vs) recorded at time $t = 1, \ldots, 278$. Azzalini and Genton (2008) applied a Ljung-Box test to the data that indicated some serial correlation at the Goodnoe Hills tower, but not at the other two towers. Given this result, the authors propose to treat the observations as being independent and identically distributed. A descriptive statistics summary is presented in Table 2. Although the data present moderate levels of negative skewness, in this paper we find that the multivariate symmetric t-distribution presents a good fit, substantially better than the adjust of the normal distribution.

Table 2 Descriptive statistics for the Wind speed dataset.

Variable	Mean vector	Sample covariance matrix			Skewness	Kurtosis
vs	16.980	185.307			-0.849	4.376
gh	12.741	126.959	177.779		-0.692	2.617
kw	14.032	148.180	110.620	297.167	-0.410	3.009
Mardia's skewness: 3.516, and kurtosis: 23.850						

Table 3 Gaussian fit: Wind speed dataset.

Variable	Mean vector	Covariance matrix		
vs	16.980	184.641		
gh	12.741	126.502	177.140	
kw	14.032	147.647	110.222	296.098
log-likelihood: -3254.534				

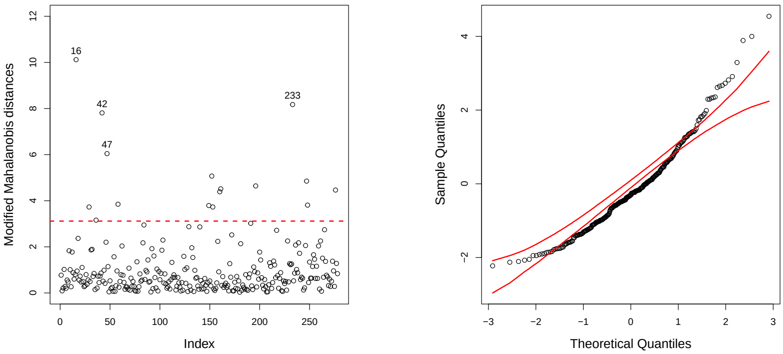
Table 4 Multivariate t fit: Wind speed dataset.

Variable	Mean vector	Covariance matrix		
vs	18.959	217.241		
gh	14.831	175.155	246.069	
kw	16.725	200.799	152.030	353.586
log-likelihood: -3211.186, $\widehat{\eta} = 0.237$				

Tables 3 and 4 show the ML estimators under the normal and t distributions, respectively. The maximum value of the log-likelihood function is also reported. We can see that under the t-distribution the ML estimators of the mean wind speeds and their standard deviations are slightly larger than under the normal distribution. To note the improvement obtained when using a multivariate t-distribution, we considered the following hypothesis test:

$$H_0 : \eta = 0, \qquad \text{against} \qquad H_1 : \eta > 0. \tag{10}$$

The likelihood-ratio statistic is $LR = 86.696$. We should highlight that the asymptotic distribution of the LR test for the previous hypothesis corresponds to a 50:50 mixture of chi-squares with zero and one degree of freedom whose critical value at significance level of 5% is 2.7055 (see, for instance Song et al., 2007). This indicates that the assumed normality is not supported by the data. Alternatively, the assumption of normality can also be evaluated by either using Mardia's test (1974) or by considering the score statistic developed by Fiorentini et al. (2003). Figures 1 and 2 show the QQ-plots and Mahalanobis distances for the two distributions. Some possible outliers are observed, and also it is observed a better adjustment of the multivariate t distribution. This conclusion is also confirmed by the confidence intervals displayed in Table 5, and by the bootstrap distributions of the ML estimators of η and of the negentropy $h(\widehat{\eta}) = H_N(\boldsymbol{Y})$ (see Figure 3). These distributions were generated by using 1000 bootstrap samples, and the solid line corresponds to the normal approximation presented in Lemma 3.

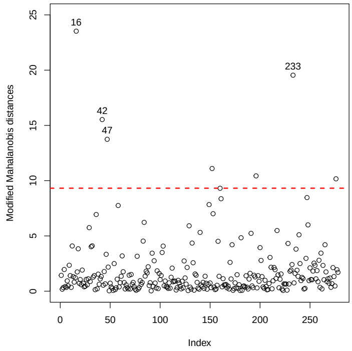


(a) (b)

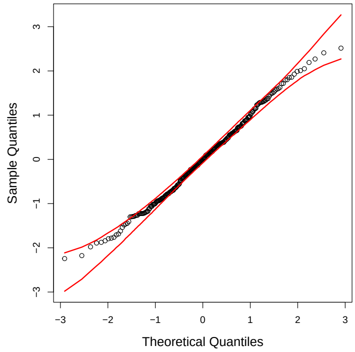
Fig. 1 Wind speed dataset: (a) Mahalanobis distances and (b) QQ-plot of transformed distances, fitted model under the normality assumption.

Table 5 Asymptotic and Bootstrap confidence intervals, for η and $h(\eta)$. Wind speed dataset $(n = 278, \widehat{\eta} = 0.237, h(\widehat{\eta}) = 0.279)$

Method	$CI_n(\eta)$		$CI_n(h(\eta))$	
normal approximation	0.162	0.311	0.054	0.504
bootstrap percentile	0.159	0.305	0.111	0.540
bootstrap pivotal	0.168	0.314	0.018	0.447



(a)



(b)

Fig. 2 Wind speed dataset: (a) Mahalanobis distances and (b) QQ-plot of transformed distances, fitted model under the multivariate t distribution.

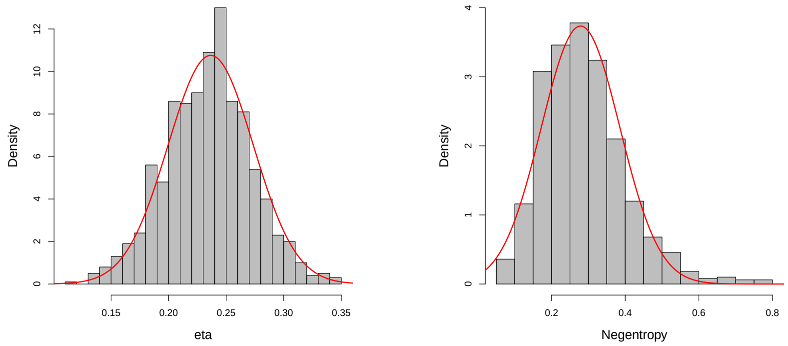
The Figure 4 shows index plots for local influence measures for t-perturbation scheme. Observations 16, 42, 47 and 233 stand out as potentially influential on the assumption of normality. Note that these same observations are also highlighted by the Mahalanobis distances (see Figures 1 and 2). In Table 6, we summarize the results of the fit of the t-distribution without considering these observations.

An interesting comment raised by one of the referees was that the non-normality may also come from the asymmetry present in the data. An alternative to address this scenario is to consider the skew-t distribution (see, for instance Azzalini and Genton, 2008; Gupta et al., 2003). Details on the development of tools for assessing non-normality based on entropy measures are presented in Appendix B. Moreover, the results of Section 4 can be used directly for the construction of QQ-plots with envelopes in order to evaluate the goodness of fit for skew-normal and skew-t models.

Tables 7 and 8 present the fit using the skew-normal and skew-t models defined by the densities presented in Equations (B1) and (B2), respectively. The results reported in Table 9 allow us to refute the assumption of normality, although it should be noted that for this dataset the non-normality is not only due to the asymmetry, but also to the presence of extreme wind speeds. Indeed, the QQ-plot displayed in Figure 5 (a) reveals that the skew-normal model is not an appropriate alternative for modeling this type of data. As mentioned by Azzalini and Genton (2008), the skew-t distribution offers an improvement over the Gaussian model, although compared to the multivariate t-distribution the improvement seems marginal. Indeed, QQ-plots in Figures 2 (b) and 5 (b) are remarkably similar. Apparently, the use of the multivariate t-distribution yields some protection against small levels of skewness. We believe this is a feature of the model that merits further investigation.

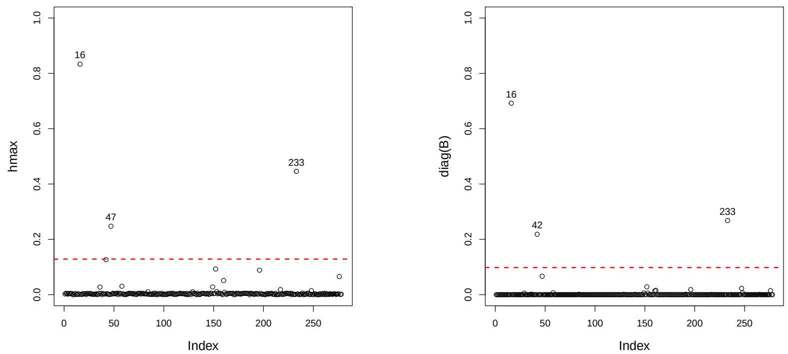
Table 6 Multivariate t fit with observations 16, 42, 47 and 233 removed: Wind speed dataset.

Variable	Mean vector	Covariance matrix		
vs	18.833	188.656		
gh	14.620	154.921	217.412	
kw	16.613	176.538	132.304	310.028
log-likelihood: -3135.325, $\widehat{\eta} = 0.189$				



(a) (b)

Fig. 3 Wind speed dataset: Bootstrap distributions for (a) $\widehat{\eta}$ and (b) $h(\widehat{\eta})$.



(a) (b)

Fig. 4 Wind speed dataset: Indices plot for the t-perturbation, (a) $|\boldsymbol{l}_{\max}|$ and (b) B_i.

Table 7 Skew normal fit: Wind speed dataset.

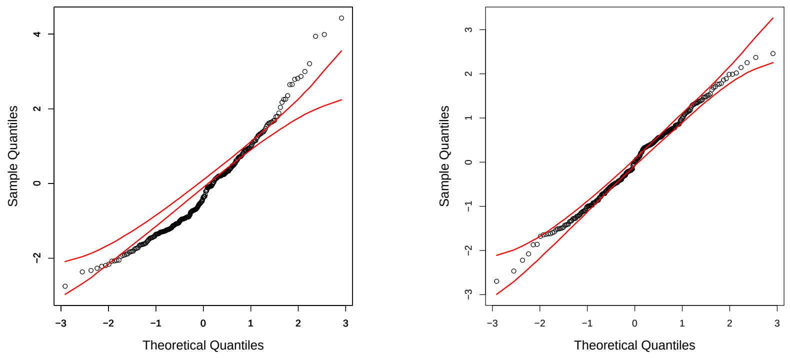
Variable	Location vector	Dispersion matrix			skewness
vs	25.792	262.290			1.073
gh	28.200	262.731	416.142		-4.974
kw	23.112	227.661	250.600	378.548	-0.282
log-likelihood: -3229.176					

Table 8 Skew-t fit: Wind speed dataset.

Variable	Location vector	Dispersion matrix			skewness
vs	27.486	175.507			0.633
gh	27.828	188.034	276.320		-4.548
kw	24.326	161.141	166.362	236.538	-0.124
log-likelihood: -3180.724, $\widehat{\nu} = 4.047$					

Table 9 Bootstrap confidence intervals, for $h_{\mathsf{SN}} = H_N(\boldsymbol{X})$, $h_{\mathsf{St}} = H_N(\boldsymbol{Y})$ and ν. Wind speed dataset ($n = 278, \widehat{h}_{\mathsf{SN}} = 0.246, \widehat{h}_{\mathsf{St}} = 0.439, \widehat{\nu} = 4.047$)

Bootstrap method	$CI_n(h_{\mathsf{SN}})$		$CI_n(h_{\mathsf{St}})$		$CI_n(\nu)$	
percentile	0.160	0.311	0.318	0.535	3.245	5.863
pivotal	0.181	0.331	0.343	0.559	2.230	4.848



(a) (b)

Fig. 5 Wind speed dataset: QQ-plot of transformed distances, for (a) skew-normal and (b) skew-t fits.

5.2.2 Transient sleep disorder

Svetnik et al. (2007) introduced a dataset of 82 patients who received a sleep-inducing drug. As part of a clinical study related to insomnia problems we sought to measure the latency to persistent sleep (LPS), i.e. which was recorded from the time the lights were turned off until 10 consecutive minutes of uninterrupted sleep. Originally, six measurement methods were used to study the sleep pattern. In this paper, we consider two of these methods: fully manual scoring and automated scoring by the Morpheus software (Manual and Automated, respectively).

This dataset has been previously analyzed by Feng et al. (2015) who considered a robust approach for the concordance between the manual and automated methods within a Bayesian framework, whereas Leal et al. (2019) developed methods to study the influence that outlying observations can exert on the concordance correlation coefficient and the probability of agreement. Both papers identify that observations 1, 30 and 79 are outliers. Indeed, these observations have a great impact on the estimation of the covariance matrix as well as on the concordance correlation coefficient. We can notice this aspect from the information displayed in Tables 11 and 12. In particular, the likelihood ratio statistic for the hypothesis given in Equation (10) is $LR = 70.217$. Thus, the normality assumption is rejected (see also QQ-plot presented in Figure 6 (b)). We should emphasize that the distribution for $\widehat{\eta}$ obtained by bootstrap is quite skewed and concentrated at high values. The empirical distribution for negentropy confirms non-normality for this dataset. Additional information is provided by the QQ-plot of transformed distances as well as from the influence plot considering the t-perturbation. Specifically, observations 30 and 79 are those that allow invalidating the normality assumption. These findings complement the results reported by Leal et al. (2019). We must highlight the great improvement on the fit produced by using the multivariate t-distribution (see, for instante, the QQ-plot displayed in Figure 7 (b)).

In addition, the estimation was performed using the skew-normal and skew-t models described in Appendix B. The estimation results are presented in Tables 14 and 15. It should be noted that the QQ-plot of transformed distances assuming skew-normal distribution (Figure 10 (a)) reveals the presence of heavy tails and that this feature has a strong impact on the estimation of the skewness parameter. This may be due to the fact that in the skewness coefficient proposed by Mardia (1970), there is a high interaction between ν and $\boldsymbol{\gamma}$ for small values of the degrees of freedom (see, for instance, Figure B1). It should also be noted that the estimation of $\widehat{\nu} = 1.749$ for the skew-t distribution, the covariance matrix does not exist, which prevents a direct comparison with the normal model or the model considering the t-distribution with the parameterization used in this work. Moreover, it is interesting to note that the negentropy is much more pronounced for the t-multivariate distribution than for its asymmetric counterpart.

Table 10 Descriptive statistics for the transient sleep disorder dataset.

Variable	Mean vector	Sample covariance matrix		Skewness	Kurtosis
manual	2.554	0.771		-0.230	2.569
automated	2.309	0.703	1.252	0.120	3.756
Mardia's skewness: 1.006, and kurtosis: 18.292					

Table 11 Gaussian fit: transient sleep disorder dataset.

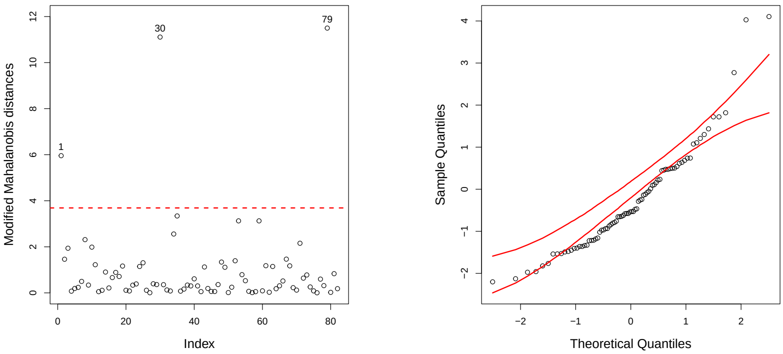
Variable	Mean vector	Covariance matrix	
manual	2.554	0.762	
automated	2.309	0.694	1.237
log-likelihood: -200.890			

Table 12 Multivariate t fit: transient sleep disorder dataset.

Variable	Mean vector	Covariance matrix	
manual	2.615	4.826	
automated	2.530	4.750	5.218
log-likelihood: -165.782, $\widehat{\eta} = 0.453$			

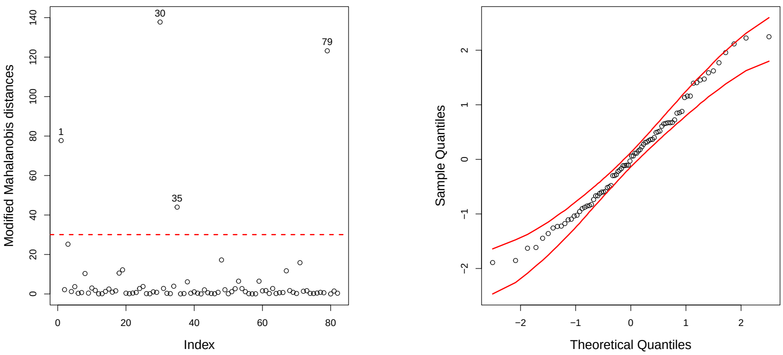
Table 13 Asymptotic and Bootstrap confidence intervals, for η and $h(\eta)$. Transient sleep disorder dataset $(n = 82, \widehat{\eta} = 0.453, h(\widehat{\eta}) = 1.454)$

Method	$CI_n(\eta)$		$CI_n(h(\eta))$	
normal approximation	0.367	0.538	0.159	2.749
bootstrap percentile	0.320	0.490	0.382	2.936
bootstrap pivotal	0.416	0.586	-0.028	2.526



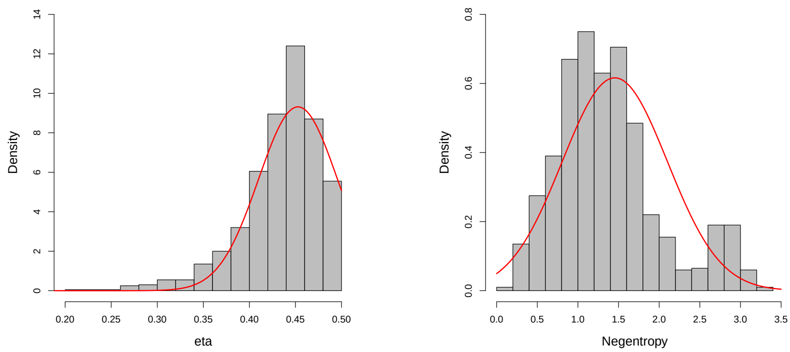
(a) (b)

Fig. 6 Transient sleep disorder dataset: (a) Mahalanobis distances and (b) QQ-plot of transformed distances, fitted model under the normality assumption.



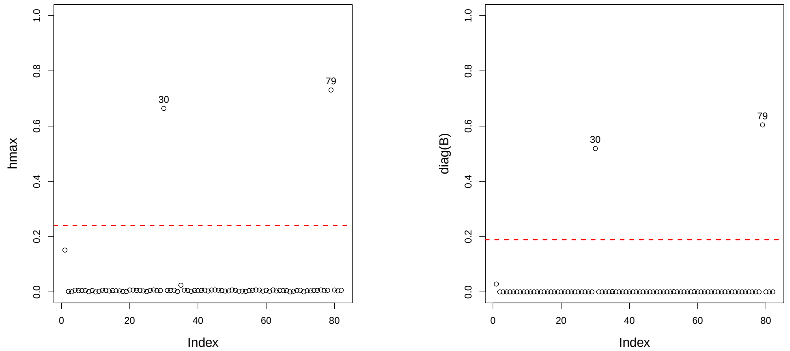
(a) (b)

Fig. 7 Transient sleep disorder dataset: (a) Mahalanobis distances and (b) QQ-plot of transformed distances, fitted model under the multivariate t distribution.



(a) (b)

Fig. 8 Transient sleep disorder dataset: Bootstrap distributions for (a) $\widehat{\eta}$ and (b) $h(\widehat{\eta})$.



(a) (b)

Fig. 9 Transient sleep disorder dataset: Indices plot for the t-perturbation, (a) $|\boldsymbol{l}_{\max}|$ and (b) B_i.

Table 14 Skew normal fit: transient sleep disorder dataset.

Variable	Location vector	Dispersion matrix		skewness
manual	2.247	0.856		-0.110
automated	1.697	0.882	1.612	0.846
log-likelihood: -200.782				

Table 15 Skew-t fit: transient sleep disorder dataset.

Variable	Location vector	Dispersion matrix		skewness
manual	2.822	0.465		8.437
automated	2.930	0.485	0.584	-10.424
log-likelihood: -151.476, $\widehat{\nu} = 1.749$				

Table 16 Bootstrap confidence intervals, for $h_{\mathsf{SN}} = H_N(\boldsymbol{X})$, $h_{\mathsf{St}} = H_N(\boldsymbol{Y})$ and ν. Transient sleep disorder dataset ($n = 82, \widehat{h}_{\mathsf{SN}} = 0.065, \widehat{h}_{\mathsf{St}} = 0.599, \widehat{\nu} = 1.749$)

Bootstrap method	$CI_n(h_{\mathsf{SN}})$		$CI_n(h_{\mathsf{St}})$		$CI_n(\nu)$	
percentile	0.023	0.508	0.335	0.850	1.182	4.601
pivotal	-0.379	0.106	0.348	0.864	-1.103	2.316

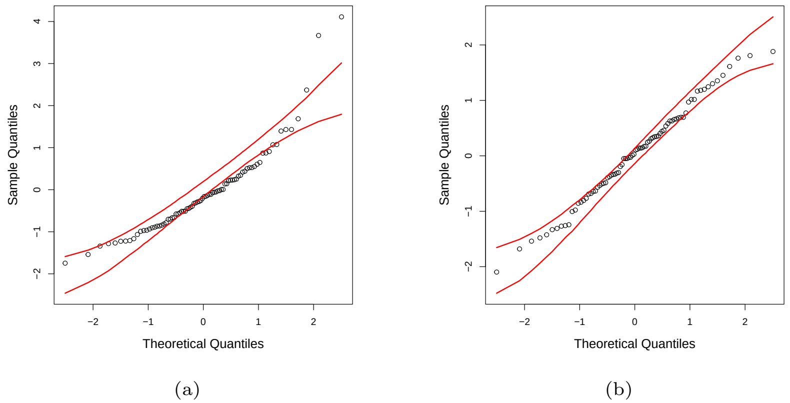


Fig. 10 Transient sleep disorder dataset: QQ-plot of transformed distances, for (a) skew-normal and (b) skew-t fits.

6 Discussion

In this work, we have discussed the problem of statistical inference on the mean vector and the covariance matrix when there is a sample of observations from a continuous population following a multivariate t distribution. Specifically, we assumed the existence of the second moment, and we have used a reparameterized version of the multivariate t distribution proposed by Sutradhar (1993) that allows for a more direct comparison with the normal distribution. Measures to determine non-normality were proposed using different procedures. Namely, measures based on the Kullback-Liebler divergence, the local influence procedure and a graphical tool to assess goodness-of-fit. We must highlight the good performance of each of these procedures in our numerical experiments.

A point that is often raised, for instance, associated with data from financial contexts is the use of distributions suitable for modelling skewness and heavy tails. To deal with this pitfalls and based on the suggestion of a referee, we have considered the class of skew-t distributions (Azzalini and Genton, 2008; Gupta et al., 2003). Appendix B presents the negentropy for skew-normal and skew-t distributions, as well as a brief comment on the skewness coefficient proposed by Mardia (1970) for both distributions. Among the procedures for the detection of non-normality outlined in this work, the use of the local influence technique is very relevant, which also reveals those observations that exert a strong influence on deviations from normality. The extension of these results represents an interesting field that the authors plan to address in future research.

Another interesting aspect that one of the referees pointed out to us corresponds to the ill-conditioning that can occur due to the covariance structure. It is illustrative to note that, for the equicorrelation matrix $\mathbf{\Sigma} = (1-\rho)\boldsymbol{I}+\rho\mathbf{1}_p\mathbf{1}_p^\top$, used in the simulation study reported in Section 5.1, we have that its condition number $\kappa(\mathbf{\Sigma})$, takes the form

$$\kappa(\mathbf{\Sigma}) = \sqrt{\lambda_1/\lambda_p} = \sqrt{1 + p\rho/(1-\rho)},$$

where λ_1 and λ_p are the largest and smallest eigenvalue of $\mathbf{\Sigma}$, respectively. It is easy to notice that $\kappa(\mathbf{\Sigma})$ is an increasing function in p. Thus, we should expect ill-conditioning problems in situations whose dimension is large. An alternative to overcome this may be to consider the work of Ledoit and Wolf (2004), Bodnar et al. (2014) or more recently Ledoit and Wolf (2020). Moreover, it is well known that under normality, i.e., for $\eta = 0$, the condition for the sample covariance matrix to be positive definite is guaranteed when $n - 1 \geq p$ (see Dykstra, 1970). However, for $\eta \in (0, 1/2]$ the positive definiteness of the ML estimator for $\mathbf{\Sigma}$ described in the Supplementary Material depends not only on n and p, but also on the weights obtained by the estimation algorithm, $\widehat{v}_i = (1/\widehat{\eta} + p)/(1/c(\widehat{\eta} + \widehat{\delta}_i^2))$ for $i = 1, \dots, n$. Details about this topic and the distribution of the estimated weights under the t-multivariate distribution is being developed by the authors and will be the subject of an incoming paper.

Supplementary information. This material is subdivided into two sections. First, we present basic properties of the multivariate t-distribution introduced by Sutradhar (1993). Then, a detailed description of the maximum likelihood estimation procedure considering an EM algorithm is provided.

Acknowledgments. This work was supported by Comisión Nacional de Investigación Científica y Tecnológica, FONDECYT grants 1140580 and 1150325. Authors are grateful for the valuable comments and suggestions made by the associate editor and anonymous reviewers who, as well as Carla Leal and Ronny Vallejos, who careful read the initial version of the manuscript, allowed improvement to the paper.

Conflict of Interest. The authors have declared no conflict of interest.

Orcid. Felipe Osorio, https://orcid.org/0000-0002-4675-5201
Manuel Galea, https://orcid.org/0000-0001-9819-5843
Reinaldo Arellano-Valle, https://orcid.org/0000-0002-5121-9702

Appendix A The expected information matrix

We use the results shown in the Supplementary Material and note that the Fisher information matrix for $\boldsymbol{\theta}$ can be written as

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\{\boldsymbol{U}_i(\boldsymbol{\theta})\boldsymbol{U}_i^\top(\boldsymbol{\theta})\},$$

where the score function $\boldsymbol{U}_i(\boldsymbol{\theta}) = (\boldsymbol{U}_i^\top(\boldsymbol{\mu}), \boldsymbol{U}_i^\top(\boldsymbol{\phi}), U_i(\eta))^\top$ associated to the ith component of the log-likelihood, with $i = 1, \dots, n$, is defined in Equations (4)-(6), and the expected value $\mathrm{E}(\cdot)$ is taken with respect to the density function in (1). Next, we obtain each of the blocks of the Fisher information matrix reported in Equation (7).

From the score functions (4) and (5), it follows that

$$
\begin{aligned}
\mathrm{E}\{\boldsymbol{U}_i(\boldsymbol{\mu})\boldsymbol{U}_i^\top(\boldsymbol{\mu})\} &= c_\mu(\eta)\boldsymbol{\Sigma}^{-1}, \\
\mathrm{E}\{\boldsymbol{U}_i(\boldsymbol{\phi})\boldsymbol{U}_i^\top(\boldsymbol{\phi})\} &= \frac{1}{4}\boldsymbol{D}_p^\top\Big\{2c_\phi(\eta)(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\boldsymbol{N}_p \\
&\qquad + (c_\phi(\eta)-1)(\operatorname{vec}\boldsymbol{\Sigma}^{-1})(\operatorname{vec}\boldsymbol{\Sigma}^{-1})^\top\Big\}\boldsymbol{D}_p,
\end{aligned}
$$

where $c_\mu(\eta) = c_\phi(\eta)/(1-2\eta)$, $c_\phi(\eta) = (1+p\eta)/(1+(p+2)\eta)$. Note that $c_\mu(\eta)$ and $c_\phi(\eta) \to 1$ when $\eta \to 0$. On the other hand, we have that $\boldsymbol{N}_p\boldsymbol{D}_p = \boldsymbol{D}_p$ (see Magnus and Neudecker, 1999), which produces the expressions corresponding to the normal case.

Therefore, it is clear that

$$
\frac{\partial v_i}{\partial \eta} = (1-2\eta)^{-3}\Big\{(p+2)(1-2\eta)q_i^{-1} - (1+\eta p)q_i^{-2}\delta_i^2\Big\},
$$

with $q_i = 1 + c(\eta)\delta_i^2$. Thus, we obtain

$$
\begin{aligned}
\frac{\partial \boldsymbol{U}_i(\boldsymbol{\mu})}{\partial \eta} &= (1-2\eta)^{-3}\boldsymbol{\Sigma}^{-1}\Big\{(p+2)(1-2\eta)q_i^{-1}\boldsymbol{Z}_i - (1+\eta p)q_i^{-2}\delta_i^2\boldsymbol{Z}_i\Big\}, \\
\frac{\partial \boldsymbol{U}_i(\boldsymbol{\phi})}{\partial \eta} &= \frac{1}{2}(1-2\eta)^{-3}\boldsymbol{D}_p^\top\operatorname{vec}\Big\{\boldsymbol{\Sigma}^{-1}\Big((p+2)(1-2\eta)q_i^{-1}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top \\
&\quad - (1+\eta p)q_i^{-2}\delta_i^2\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\Big)\boldsymbol{\Sigma}^{-1}\Big\}.
\end{aligned}
$$

By applying Lemmas 3 and 5 from Supplementary Material, it follows that

$$
\begin{aligned}
\mathrm{E}\Big\{\frac{\partial \boldsymbol{U}_i(\boldsymbol{\mu})}{\partial \eta}\Big\} &= \boldsymbol{0}, \\
\mathrm{E}\Big\{\frac{\partial \boldsymbol{U}_i(\boldsymbol{\phi})}{\partial \eta}\Big\} &= \frac{c(\eta)(p+2)}{(1+\eta p)(1+(p+2)\eta)}\boldsymbol{D}_p^\top\operatorname{vec}(\boldsymbol{\Sigma}^{-1}),
\end{aligned}
$$

for $i = 1, \dots, n$. The score function for η can be written as,

$$
\begin{aligned}
U_i(\eta) &= \frac{1}{2\eta^2}\Big\{pc(\eta) + \psi\Big(\frac{1}{2\eta}\Big) - \psi\Big(\frac{1+p\eta}{2\eta}\Big) - \frac{1+p\eta}{1-2\eta}\frac{c(\eta)\delta_i^2}{1+c(\eta)\delta_i^2} + \log(1+c(\eta)\delta_i^2)\Big\}, \\
&= \frac{1}{2\eta^2}\Big\{\log(1+Q_{i\eta}) - \Big(\psi\Big(\frac{1+p\eta}{2\eta}\Big) - \psi\Big(\frac{1}{2\eta}\Big)\Big) - \Big(\frac{1+p\eta}{1-2\eta}\frac{Q_{i\eta}}{1+Q_{i\eta}} - pc(\eta)\Big)\Big\},
\end{aligned}
$$

where $Q_{i\eta} = c(\eta)\delta_i^2 \sim \chi_p^2/\chi_{1/\eta}^2$, $\mathrm{E}\{Q_{i\eta}(1+Q_{i\eta})^{-1}\} = \frac{p\eta}{1+p\eta}$ and $\mathrm{E}\{\log(1+Q_{i\eta})\} = \psi\Big(\frac{1+p\eta}{2\eta}\Big) - \psi\Big(\frac{1}{2\eta}\Big)$. Let $U_1 = \log(1+Q_{i\eta})$, $U_2 = \frac{1+p\eta}{1-2\eta}\frac{Q_{i\eta}}{1+Q_{i\eta}}$, $\overline{U}_1 = U_1 - \mathrm{E}(U_1)$ and $\overline{U}_2 = U_2 - \mathrm{E}(U_2)$. Then,

$$\begin{aligned}
\mathrm{E}\{U_i(\eta)\} &= \mathrm{E}\{(U_1 - \mathrm{E}(U_1)) - (U_2 - \mathrm{E}(U_2))\} = 0, \\
\mathrm{var}\{U_i(\eta)\} &= \frac{1}{4\eta^4}\{\mathrm{E}(\overline{U}_1^2) - 2\,\mathrm{E}(\overline{U}_1\overline{U}_2) + \mathrm{E}(\overline{U}_2^2)\} \\
&= \frac{1}{4\eta^4}\{\mathrm{var}(\overline{U}_1) - 2\,\mathrm{Cov}(\overline{U}_1,\overline{U}_2) + \mathrm{var}(\bar{U}_2)\} \\
&= \frac{1}{4\eta^4}\{\mathrm{E}(U_1^2) - \mathrm{E}^2(U_1) - 2(\mathrm{E}(U_1U_2) - \mathrm{E}(U_1)\,\mathrm{E}(U_2)) + \mathrm{E}(U_2^2) - \mathrm{E}^2(U_2)\}.
\end{aligned}$$

Using the fact that $\psi(x+1) = \psi(x) + 1/x$ we have

$$\begin{aligned}
\mathrm{E}(U_1^2) &= \mathrm{E}\{(\log(1+Q_{i\eta}))^2\} = \mathrm{E}^2(U_1) - \psi'\Big(\frac{1+p\eta}{2\eta}\Big) + \psi'\Big(\frac{1}{2\eta}\Big), \\
\mathrm{E}\{U_1U_2) &= \frac{1+p\eta}{1-2\eta}\,\mathrm{E}\{Q_{i\eta}(1+Q_{i\eta})^{-1}\log(1+Q_{i\eta})\} \\
&= \Big\{\mathrm{E}(U_1) + \frac{2\eta}{1+p\eta}\Big\}\,\mathrm{E}(U_2), \\
\mathrm{E}(U_2^2) &= \Big(\frac{1+p\eta}{1-2\eta}\Big)^2\,\mathrm{E}\{Q_{i\eta}^2(1+Q_{i\eta})^{-2}\} \\
&= \frac{p+2}{p}\frac{1+p\eta}{1+(p+2)\eta}\,\mathrm{E}(U_2)^2.
\end{aligned}$$

Finally, we have that the expected information in relation to η is given by,

$$\begin{aligned}
\mathrm{var}\{U_i(\eta)\} &= \frac{1}{4\eta^4}\Big\{-\psi'\Big(\frac{1+p\eta}{2\eta}\Big) + \psi'\Big(\frac{1}{2\eta}\Big) - \frac{4\eta}{1+p\eta}\,\mathrm{E}(U_2) \\
&\quad + \frac{p+2}{p}\frac{1+p\eta}{1+(p+2)\eta}\,\mathrm{E}(U_2)^2 - \mathrm{E}(U_2)^2\Big\} \\
&= \frac{1}{4\eta^4}\Big\{\psi'\Big(\frac{1}{2\eta}\Big) - \psi'\Big(\frac{1+p\eta}{2\eta}\Big) + 2pc(\eta)^2\Big(\frac{4(p+2)\eta^2 - p\eta - 1}{(1+p\eta)(1+(p+2)\eta)}\Big)\Big\}.
\end{aligned}$$

From the expansion (Abramowitz and Stegun, 1970, Sec. 6.4.12),

$$\begin{aligned}
\psi'(x) &= \frac{1}{x} + \frac{1}{2x^2} + \frac{1}{6x^3} + O\Big(\frac{1}{x^5}\Big) \quad \text{as} \quad x \to \infty, \\
(1+ax)^{-k} &= 1 - kax + \frac{k(k+1)}{2}a^2x^2 - \frac{k(k+1)(k+2)}{6}a^3x^3 + O(x^4) \quad \text{as} \quad x \to 0,
\end{aligned}$$

we find as $\eta \to 0$ that,

$$\begin{aligned}\psi'\Big(\frac{1}{2\eta}\Big) - \psi'\Big(\frac{1+p\eta}{2\eta}\Big) &= 2\eta + 2\eta^2 + \frac{4}{3}\eta^3 - 2\eta(1+p\eta)^{-1} - 2\eta^2(1+p\eta)^{-2} \\ &\quad - \frac{4}{3}\eta^3(1+p\eta)^{-3} + O(\eta^5) \\ &= 2p\eta^2 - (2p^2 - 4p)\eta^3 + (2p^3 - 6p^2 + 4p)\eta^4 + O(\eta^5).\end{aligned}$$

Similarly,

$$\begin{aligned}2pc(\eta)^2\Big(\frac{4(p+2)\eta^2 - p\eta - 1}{(1+p\eta)(1+(p+2)\eta)}\Big) &= \frac{2p\eta^2(4(p+2)\eta^2 - p\eta - 1)}{(1-2\eta)^2(1+p\eta)(1+(p+2)\eta)} \\ &= (8p(p+2)\eta^4 - 2p^2\eta^3 - 2p\eta^2)(1 + 4\eta + 12\eta^2 + O(\eta^3)) \\ &\quad \times (1 - p\eta + p^2\eta^2 + O(\eta^3))(1 - (p+2)\eta - (p+2)^2\eta^2 O(\eta^3)) \\ &= -2p\eta^2 + (2p^2 - 4p)\eta^3 + (2p^3 + 24p^2 + 16p)\eta^4 + O(\eta^5).\end{aligned}$$

Hence,

$$\operatorname{var}\{U_i(\eta)\} = \frac{1}{4}(4p^3 + 18p^2 + 20p) + O(\eta) = \frac{p(p+2)(2p+5)}{2} + O(\eta).$$

Appendix B Non-normality due to asymmetry

Another source of non-normality is the possible asymmetry present in the observations. Shannon entropy, Kullback-Leibler divergence and mutual information for multivariate skew-elliptical distributions have been considered in the literature, see for instance Arellano-Valle et al. (2012) and Contreras-Reyes and Arellano-Valle (2012). We summarize some of these results for the multivariate skew normal and skew t distributions below.

Following Arellano-Valle et al. (2012) we say that a random vector $\boldsymbol{Z} \in \mathbb{R}^p$ has a skew-normal distribution with location vector $\boldsymbol{\xi} \in \mathbb{R}^p$, dispersion matrix $\boldsymbol{\Omega} > 0$ and shape/skewness parameter $\boldsymbol{\gamma} \in \mathbb{R}^p$, denoted by $\boldsymbol{Z} \sim \mathsf{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\gamma})$, if its probability density function is

$$f(\boldsymbol{z}) = 2\phi_p(\boldsymbol{z};\ \boldsymbol{\xi}, \boldsymbol{\Omega})\,\Phi\{\boldsymbol{\gamma}^\top(\boldsymbol{z} - \boldsymbol{\xi})\}, \qquad \boldsymbol{z} \in \mathbb{R}^p, \tag{B1}$$

where

$$\phi_p(\boldsymbol{z};\ \boldsymbol{\xi}, \boldsymbol{\Omega}) = (2\pi)^{-p/2}|\boldsymbol{\Omega}|^{-1/2}\exp(-\delta^2_{\mathsf{skew}}/2),$$

is the probability density function of the p-variate $\mathsf{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$ distribution, $\Phi(\cdot)$ is the univariate $\mathsf{N}(0,1)$ cumulative distribution function and $\delta^2_{\mathsf{skew}} = (\boldsymbol{z} - \boldsymbol{\xi})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{z} - \boldsymbol{\xi}) \sim \chi^2(p)$. The vector of means and the covariance matrix of $\boldsymbol{Z}$ are given, respectively by

$$\boldsymbol{\mu}_{\mathsf{SN}} = \boldsymbol{\xi} + \sqrt{\frac{2}{\pi}}\boldsymbol{\delta}, \qquad \text{and} \qquad \boldsymbol{\Sigma}_{\mathsf{SN}} = \boldsymbol{\Omega} - \frac{2}{\pi}\boldsymbol{\delta}\boldsymbol{\delta}^\top,$$

where, $\boldsymbol{\delta} = \boldsymbol{\Omega}\boldsymbol{\gamma}/\sqrt{1+\tau^2}$ and $\tau^2 = \boldsymbol{\gamma}^\top\boldsymbol{\Omega}\boldsymbol{\gamma}$.

We say that a random vector $\boldsymbol{Z} \in \mathbb{R}^p$ has a skew-t distribution with location vector $\boldsymbol{\xi} \in \mathbb{R}^p$, dispersion matrix $\boldsymbol{\Omega} \in \mathbb{R}^{p\times p}$, shape/skewness parameter $\boldsymbol{\gamma} \in \mathbb{R}^p$ and $\nu > 0$ degrees of freedom, denoted by $\boldsymbol{Z} \sim \mathsf{St}_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\gamma}, \nu)$, if its probability density function is given by

$$f(\boldsymbol{z}) = 2t_p(\boldsymbol{z};\ \boldsymbol{\xi}, \boldsymbol{\Omega}, \nu)\, T\Big(\sqrt{\frac{\nu+p}{\nu+\delta^2_{\mathsf{skew}}}}\,\boldsymbol{\gamma}^\top(\boldsymbol{z}-\boldsymbol{\xi});\ \nu+p\Big), \tag{B2}$$

where

$$t_p(\boldsymbol{z};\ \boldsymbol{\xi}, \boldsymbol{\Omega}, \nu) = \frac{\Gamma(\frac{\nu+p}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{p/2}}|\boldsymbol{\Omega}|^{-1/2}\Big(1+\frac{1}{\nu}\delta^2_{\mathsf{skew}}\Big)^{-(\nu+p)/2}, \qquad \boldsymbol{z} \in \mathbb{R}^p,$$

is the probability density function of the p-variate $t_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \nu)$ distribution, $\delta^2_{\mathsf{skew}} = (\boldsymbol{z}-\boldsymbol{\xi})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{z}-\boldsymbol{\xi})/p \sim F(p,\nu)$ and $T(x;\ \nu+p)$ is the $T_1(0,1,\nu+p)$ cumulative distribution function (see, for instance Arellano-Valle et al., 2012; Azzalini and Genton, 2008, for details).

If $\boldsymbol{Z} \sim \mathsf{St}_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\gamma}, \nu)$ then the vector of means and the covariance matrix of $\boldsymbol{Z}$ is given by

$$\begin{aligned} \boldsymbol{\mu}_{\mathsf{St}} &= \boldsymbol{\xi} + \alpha(\nu)\boldsymbol{\delta}, \quad \nu > 1 \\ \boldsymbol{\Sigma}_{\mathsf{St}} &= \frac{\nu}{\nu-2}\boldsymbol{\Omega} - \{\alpha(\nu)\}^2\boldsymbol{\delta}\boldsymbol{\delta}^\top, \quad \nu > 2, \end{aligned}$$

where $\alpha(\nu) = \{\Gamma((\nu-1)/2)/\Gamma(\nu/2)\}\sqrt{\nu/\pi}$. Note that $\alpha(\nu) \to \sqrt{2/\pi}$ as $\nu \to \infty$, and we obtain the results for the skew-normal distribution given above.

Arellano-Valle et al. (2012), show that for the skew-normal and skew-t distributions the Shannon entropy has explicit form and given in the following lemmas.

Lemma 4 *If* $\boldsymbol{X} \sim \mathsf{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\gamma})$ *and* $\boldsymbol{Y} \sim \mathsf{St}_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\gamma}, \nu)$, *then the Shannon entropy is given by*

(i) $H(\boldsymbol{X}) = \frac{1}{2}\log|\boldsymbol{\Omega}| + \frac{p}{2}(1+\log 2\pi) - \mathrm{E}[\log\{2\Phi(\tau W)\}]$,

(ii) $H(\boldsymbol{Y}) = \frac{1}{2}\log|\boldsymbol{\Omega}| - \log\Gamma\big(\frac{\nu+p}{2}\big) + \log\Gamma\big(\frac{\nu}{2}\big) + \frac{p}{2}\log(\nu\pi) + \frac{\nu+p}{2}\big\{\psi\big(\frac{\nu+p}{2}\big) - \psi\big(\frac{\nu}{2}\big)\big\} - \mathrm{E}[\log\{2T(\tau W^*;\ \nu+p)\}]$,

with $W \sim \mathsf{SN}(0,1,\tau)$, $\tau^2 = \boldsymbol{\gamma}^\top\boldsymbol{\Omega}\boldsymbol{\gamma}$; $W^* = \sqrt{\nu+p}\,W_{\mathsf{St}}/\sqrt{\nu+p-1+W^2_{\mathsf{St}}}$ *where* $W_{\mathsf{St}} \sim \mathsf{St}(0,1,\tau,\nu+p-1)$.

Lemma 5 *Let* $\boldsymbol{Z} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *If* $\boldsymbol{X} \sim \mathsf{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\gamma})$, $\boldsymbol{Y} \sim \mathsf{St}_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\gamma}, \nu)$ *then the negentropy of* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are given, respectively by,*

(i) $H_N(\boldsymbol{X}) = \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\log|\boldsymbol{\Omega}| + \mathrm{E}[\log\{2\Phi(\tau W)\}]$, *and*

(ii) $H_N(\boldsymbol{Y}) = \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{p}{2}(1+\log 2\pi) - \frac{1}{2}\log|\boldsymbol{\Omega}| + \log\Gamma\left(\frac{\nu+p}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) - \frac{p}{2}\log(\nu\pi) - \frac{\nu+p}{2}\left\{\psi\left(\frac{\nu+p}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right\} + \mathrm{E}[\log\{2T(\tau W^*;\ \nu+p)\}]$.

Mardia (1970) introduced one of the popular and commonly used measures of multivariate skewness of an arbitrary p-dimensional random vector $\boldsymbol{Z}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Mardia's skewness coefficient is defined as,

$$\beta_{1,p} = \mathrm{E}[\{(\boldsymbol{Z}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu})\}^3],$$

which can be expressed as $\beta_{1,p} = \mathrm{tr}\{\boldsymbol{S}^\top(\boldsymbol{Y})\boldsymbol{S}(\boldsymbol{Y})\}$, where $\boldsymbol{S}(\boldsymbol{Y}) = \mathrm{E}(\boldsymbol{Y}\otimes\boldsymbol{Y}^\top\otimes\boldsymbol{Y})$, with $\boldsymbol{Y} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{Z}-\boldsymbol{\mu})$ and $\otimes$ denotes the Kronecker product. The following lemmas, extracted from Kim and Mallick (2003), allow us to obtain explicit formulas for $\boldsymbol{S}(\boldsymbol{Y})$. In particular, Figure B1 leads us to note the interaction between the degrees of freedom and the skewness parameter on the coefficient $\beta_{1,p}$ proposed by Mardia (1970). In fact, as ν grows, it has less impact on $\beta_{1,p}$.

Lemma 6 *If* $\boldsymbol{Y} \sim \mathsf{SN}_p(\mathbf{0},\boldsymbol{\Omega},\boldsymbol{\gamma})$, *then*

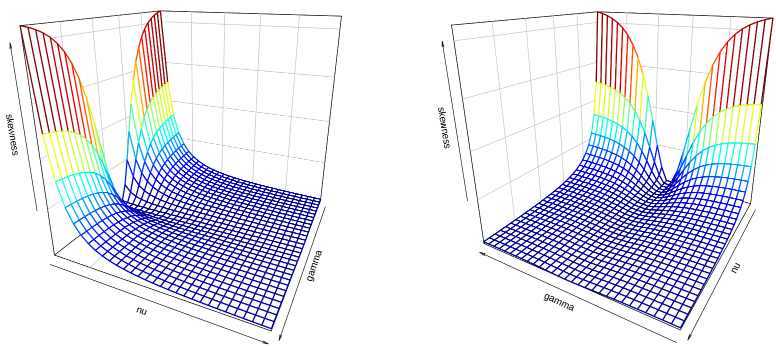
$$\boldsymbol{S}(\boldsymbol{Y}) = \sqrt{2/\pi}[\boldsymbol{\delta}\otimes\boldsymbol{\Omega} + \mathrm{vec}(\boldsymbol{\Omega})\boldsymbol{\delta}^\top + (\boldsymbol{I}_p\otimes\boldsymbol{\delta})\boldsymbol{\Omega} - \boldsymbol{\delta}\otimes\boldsymbol{\delta}\boldsymbol{\delta}^\top],$$

where $\boldsymbol{\delta} = \boldsymbol{\Omega}\boldsymbol{\gamma}/\sqrt{1+\tau^2}$. *In addition, if* $\boldsymbol{\gamma} = \mathbf{0}$, *that is* $\boldsymbol{Y} \sim \mathsf{N}_p(\mathbf{0},\boldsymbol{\Omega})$, *then* $\beta_{1,p} = 0$.

Lemma 7 *If* $\boldsymbol{Y} \sim \mathsf{St}_p(\mathbf{0},\boldsymbol{\Omega},\boldsymbol{\gamma},\nu)$, *then*

$$\boldsymbol{S}(\boldsymbol{Y}) = \frac{\alpha(\nu)\nu}{\nu-3}[\boldsymbol{\delta}\otimes\boldsymbol{\Omega} + \mathrm{vec}(\boldsymbol{\Omega})\boldsymbol{\delta}^\top + (\boldsymbol{I}_p\otimes\boldsymbol{\delta})\boldsymbol{\Omega} - \boldsymbol{\delta}\otimes\boldsymbol{\delta}\boldsymbol{\delta}^\top],$$

where $\alpha(\nu) = \sqrt{\nu/\pi}\,\Gamma((\nu-1)/2)/\Gamma(\nu/2)$ *and* $\boldsymbol{\delta} = \boldsymbol{\Omega}\boldsymbol{\gamma}/\sqrt{1+\tau^2}$. *In addition, if* $\boldsymbol{\gamma} = \mathbf{0}$, *that is* $\boldsymbol{Y} \sim \mathsf{St}_p(\mathbf{0},\boldsymbol{\Omega},\nu)$, *then* $\beta_{1,p} = 0$.



(a) (b)

Fig. B1 Plot of Mardia's skewness coefficient $\beta_{1,p}$ for the univariate skew t-distribution with $\xi = 0$ and $\Omega = 1$.

References

Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions. Dover, New York (1970).

Anderson, T.W.: An Introduction to Multivariate Statistical Analysis. Wiley, New York (2003).

Arellano-Valle, R.B., Contreras-Reyes, J. Genton, M.: Shannon entropy and mutual information for multivariate skew-elliptical distributions. Scandinavian Journal of Statistics **40**, 42-62 (2012).

Arellano-Valle, R.B., Ferreira, C.S., Genton, M.G.: Scale and shape mixtures of multivariate skew-normal distributions. Journal of Multivariate Analysis **166**, 98-110 (2018).

Azzalini, A., Genton, M.G.: Robust likelihood methods based on the skew-t and related distributions. International Statistical Review **76**, 106-129 (2008).

Bolfarine, H., Galea, M.: On structural comparative calibration under a t-model. Computational Statistics **11**, 63-85 (1996).

Bodnar, T., Gupta, A.K., Parolya, N.: On the strong convergence of the optimal linear shrinkage estimator for large dimensional covariance matrix. Journal of Multivariate Analysis **132**, 215-228 (2014).

Contreras-Reyes, J., Arellano-Valle, R.: Kullback-Leibler divergence measure for multivariate skew-normal distributions. Entropy **14**, 1606-1626 (2012).

Cook, R.D.: Assessment of local influence (with discussion). Journal of the Royal Statistical Society B **48**, 133-169 (1986).

Dykstra, R.L.: Establishing the positive definiteness of the sample covariance matrix. The Annals of Mathematical Statistics **41**, 2153-2154 (1970).

Fang, K.T., Zhang, Y.T.: Generalized Multivariate Analysis. Springer, Berlin (1990).

Feng, D., Baumgartner, R., Svetnik, V.: A robust bayesian estimate of the concordance correlation coefficient. Journal of Biopharmaceutical Statistics **25**, 490-507 (2015).

Fiorentini, G., Sentana, E., Calzolari, G.: Maximum likelihood estimation and inference in multivariate conditionally heteroscedastic dynamic regression models with Student t innovations. Journal of Business & Economic Statistics **21**, 532-546 (2003).

Galea, M., Cademartori, D., Curci, R., Molina, A.: Robust inference in the capital asset pricing model using the multivariate t-distribution. Journal of Risk and Financial Management **13**, 123 (2020).

Gao, J., Zhang, B.: Estimation of seismic wavelets based on the multivariate scale mixture of gaussian model. Entropy **12**, 14-33 (2010).

Gómez-Villegas, M.A., Gómez-Sánchez-Manzano, E., Maín, P., Navarro, H.: The effect of non-normality in the Power Exponential distribution. In: Pardo, L., Balakrihnan, N., Gil, M.A. (eds.) Modern Mathematical Tools and Techniques in Capturing Complexity, pp. 119-129, Springer-Verlag, Berlin.

Gupta, A.K.: Multivariate skew t-distribution. Statistics **37**, 359-363 (2003).

Gupta, A.K., Varga, T., Bodnar, T.: Elliptically Contoured Models in Statistics and Portfolio Theory (2nd Ed.). Springer, New York (2013).

Härdle, W.K., Simar, L.: Applied Multivariate Statistical Analysis (3rd Ed.). Springer, New York (2012).

Kent, J.T., Tyler, D.E., Vardi, Y.: A curious likelihood identity for the multivariate t-distribution. Communications in Statistics - Simulation and Computation **23**, 441-453 (1994).

Kent, J.T., Tyler, D.E.: Redescending M-estimates of multivariate location and scatter. The Annals of Statistics **19**, 2102-2119 (1991).

Kim, H.M., Mallick, B.K.: Moments of random vectors with skew t distribution and their quadratic forms. Statistics & Probability Letters **63**, 417-423 (2003).

Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Mathematical Statistics **22**, 79-86 (1951).

Lange, K., Little, R.J.A., Taylor, J.M.G.: Robust statistical modeling using the t distribution. Journal of the American Statistical Association **84**, 881-896 (1989).

Leal, C., Galea, M., Osorio, F.: Assessment of local influence for the analysis of agreement. Biometrical Journal **61**, 955-972 (2019).

Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis **88**, 365-411 (2004).

Ledoit, O., Wolf, M.: Analytical nonlinear shrinkage of large-dimensional covariance matrices. The Annals of Statistics **48**, 3043-3065 (2020).

Magnus, J.R., Neudecker, H.: Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley, Chichester (1999).

Mardia, K.V.: Measures of multivariate skewness and kurtosis with applications. Biometrika **36**, 519-530 (1970),

Mardia, K.V.: Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. Sankhyā, Series B **36**, 115-128 (1974).

Maronna, R.A.: Robust M-estimators of multivariate location and scatter. The Annals of Statistics **4**, 51-67 (1976).

Poon, W., Poon, Y.S.: Conformal normal curvature and assessment of local influence. Journal of the Royal Statistical Society B **61**, 51-61 (1999).

Serfling, R.J.: Approximation Theorems of Mathematical Statistics. Wiley, New York (2009).

Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27**, 379-423 (1948).

Song, P.X.K., Zhang, P., Qu, A.: Maximum likelihood inference in robust linear mixed-effects models using the multivariate t distributions. Statistica Sinica **17**, 929-943 (2007).

Sutradhar, B.C.: Score test for the covariance matrix of elliptical t-distribution. Journal of Multivariate Analysis **46**, 1-12 (1993).

Svetnik, V., Ma, J., Soper, K.A., Doran, S., Renger, J.J., Deacon, S., Koblan, K.S.: Evaluation of automated and semi-automated scoring of polysomnographic recordings from a clinical trial using zolpidem in the treatment of insomnia. SLEEP **30**, 1562-1574 (2007).

Wilson, E.B., Hilferty, M.M.: The distribution of chi-square. Proceedings of the National Academy of Sciences of the United States of America **17**, 684-688 (1931).

Zhu H.T., Lee S.Y.: Local influence for incomplete-data models. Journal of the Royal Statistical Society B **63**, 111-126 (2001).

Zhu, H., Ibrahim, J.G., Lee, S. Zhang, H.: Perturbation selection and influence measures in local influence analysis. The Annals of Statistics **35**, 2565-2588 (2007).