Computational Statistics and Data Analysis 192 (2024) 107909

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A Laplace-based model with flexible tail behavior

Cristina Tortora [a,*], Brian C. Franczak [b], Luca Bagnato [c], Antonio Punzo [d]

[a] Department of Mathematics and Statistics, San José State University, One Washington Square, San José, 95192, CA, USA
[b] Department of Mathematics and Statistics, MacEwan University, 10700 104 Avenue NW, Edmonton, T5J 4S2, Alberta, Canada
[c] Dipartimento di Scienze Economiche e Sociali, Università Cattolica del Sacro Cuore, Via Emilia Parmense, 84, Piacenza, 29122, Italy
[d] Dipartimento di Economia e Impresa, Università di Catania, Corso Italia, 55, Catania, 95129, Italy

ARTICLE INFO

Keywords:
Contaminated distributions
Directional outlier detection
Monte Carlo expectation-maximization algorithm
Multiple scaled distributions
Normal variance-mean mixtures

ABSTRACT

The proposed multiple scaled contaminated asymmetric Laplace (MSCAL) distribution is an extension of the multivariate asymmetric Laplace distribution to allow for a different excess kurtosis on each dimension and for more flexible shapes of the hyper-contours. These peculiarities are obtained by working on the principal component (PC) space. The structure of the MSCAL distribution has the further advantage of allowing for automatic PC-wise outlier detection – i.e., detection of outliers separately on each PC – when convenient constraints on the parameters are imposed. The MSCAL is fitted using a Monte Carlo expectation-maximization (MCEM) algorithm that uses a Monte Carlo method to estimate the orthogonal matrix of eigenvectors. A simulation study is used to assess the proposed MCEM in terms of computational efficiency and parameter recovery. In a real data application, the MSCAL is fitted to a real data set containing the anthropometric measurements of monozygotic/dizygotic twins. Both a skewed bivariate subset of the full data, perturbed by some outlying points, and the full data are considered.

1. Introduction

There are different ways to generalize the multivariate normal (MN) distribution (with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$) to account for skewness and leptokurtosis. One approach is through the multivariate normal variance-mean mixture (Barndorff-Nielsen et al., 1982); it is a finite/continuous mixture of MN distributions where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are weighted by a positive mixing random variable W, whose probability density/mass function depends on one or more parameters governing the leptokurtosis of the unconditional mixture. A member of this family of distributions is the multivariate asymmetric Laplace (MAL; Kotz et al., 2001) distribution; the peculiar peak and heavier than normal tails make it well suited in several disciplines, see Part III (Applications) in Kotz et al. (2001) for some examples. However, some empirical studies show that the MAL distribution is not appropriate because of two deficiencies: (*a*) the levels of excess kurtosis on each variate are limited; (*b*) the shape of the hyper-contours may be restrictive in some circumstances.

To handle deficiency (*a*), Morris et al. (2019) introduced a multivariate contaminated asymmetric Laplace (MCAL) distribution. This is a simple theoretical model that has two additional parameters compared to the MAL distribution. More specifically, it is a two-component MAL mixture in which one of the components represents the observations that most likely belong to a baseline, or reference, MAL distribution. The other component, typically with a smaller prior probability, the same mode/peak, and an inflated covariance matrix, represents the observations farther from the bulk of the data. Following Aitkin and Wilson (1980) and Davies

* Corresponding author.
E-mail address: cristina.tortora@sjsu.edu (C. Tortora).

https://doi.org/10.1016/j.csda.2023.107909
Received 28 December 2022; Received in revised form 6 December 2023; Accepted 6 December 2023

Available online 12 December 2023

and Gather (1993), one can also refer to the observations that belong to the reference MAL distribution as "good" and those that belong to the inflated MAL distribution as "bad" or outlying, herein we use bad and outlying interchangeably. Therefore, the MCAL distribution also allows for the automatic detection of global outlying points via a simple and natural procedure based on maximum *a posteriori* probabilities. Unfortunately, in contrast with deficiency (*b*), the shape of the hyper-contours of the MCAL distribution is the same as those of the reference MAL distribution. Further, the parameters of the MCAL that model the proportion of good points and degree of contamination are the same across every dimension of the data. This is restrictive because it implies that all dimensions have the same amount of leptokurtosis.

To handle deficiency (*b*), Franczak et al. (2015) developed a multiple scaled asymmetric Laplace (MSAL) distribution. Generally speaking, multiple scaling was proposed by Forbes and Wraith (2014) to generalize the multivariate normal variance-mean mixture to allow for a different amount of excess kurtosis on each principal component (PC) – and, as a by-product, on each dimension – and alternative shapes for the hyper-contours (Punzo and Bagnato, 2022a). Wraith and Forbes (2015) extend this concept to include a multivariate generalized hyperbolic distribution. A multiple scaled distribution can be considered an extension of the multivariate normal variance-mean mixture based on two key elements: 1. the decomposition of the scale matrix $\boldsymbol{\Sigma}$ by eigenvalues and eigenvectors matrices, $\boldsymbol{\Phi}$ and $\boldsymbol{\Gamma}$, respectively; 2. the use of the mixing random variable W separately for each dimension of the space spanned by the columns of $\boldsymbol{\Gamma}$ (Punzo and Bagnato, 2022a), i.e., separately for each PC. When applied to the MAL distribution, this approach leads to a multivariate peaked, asymmetric, and heavy-tailed distribution, with hypercube contours. Hence, the MSAL distribution can solve deficiency (*b*), but it does not consider deficiency (*a*).

Under the same trajectory of research, we propose to merge the models proposed in Franczak et al. (2015) and Morris et al. (2019). The result is the multiple scaled contaminated asymmetric Laplace (MSCAL) distribution, a multivariate peaked, asymmetric, and heavy-tailed distribution having marginal contaminated asymmetric Laplace distributions on each PC. The MSCAL distribution offers a remedy to deficiencies (*a*) and (*b*), in addition to having other benefits. Concerning deficiency (*a*), the excess kurtosis is free to vary on each PC and, as a by-product, on each dimension. Concerning deficiency (*b*), the hyper-contours have more flexible shapes because the MSCAL distribution estimates the number of good observations and the degree of contamination for each PC. As such, the proposed model accounts for different tail behaviors in each PC. In addition, the implicit procedure to detect outliers now works separately for each PC, such that a point may be detected as bad for some PCs only (see Punzo and Tortora, 2021, for examples).

The article is organized as follows. Section 2 contains the required background materials. The main contribution of this paper, i.e. the MSCAL distribution, is in Section 3. In Section 4, a parameter estimation scheme is developed. Section 5 discusses several other important considerations related to the implementation of the proposed model. In Sections 6 and 7 we conduct a simulation study and present a sensitivity analysis using real data, respectively. Finally, a discussion and suggestions for future work are provided in Section 8.

2. Required background

In this section, we present the key results used to develop the MSCAL distribution. Specifically, in Section 2.1 we review some properties of Laplace-based distributions and in Section 2.2 we discuss the MSAL distribution.

2.1. Laplace-based distributions

In this paper, we define $\mathbf{X} \sim \text{AL}_p(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$ to be a random vector following a MAL distribution with location parameter $\boldsymbol{\mu} \in \mathbb{R}$, skewness parameter $\boldsymbol{\alpha} \in \mathbb{R}$, and $p \times p$ non-negative definite matrix $\boldsymbol{\Sigma}$. It follows from Kotz et al. (2001) that the density of $\mathbf{X}$ can be expressed as

$$f_{\text{MAL}}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \frac{2\exp\{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \left(\frac{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2+\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right)^{\nu/2} K_\nu(u), \tag{1}$$

where $\nu = (2-p)/2$, $u = \sqrt{(2+\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$, $K_\nu(\cdot)$ is the modified Bessel function of the third kind with index ν, and all other terms are as previously defined (cf. Franczak et al., 2014). The random vector $\mathbf{X} \sim \text{AL}_p(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$ can be written using the following mixture representation

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\alpha} + \sqrt{W}\mathbf{N}, \tag{2}$$

where W follows an exponential distribution with rate 1, i.e., $W \sim \text{Exp}(1)$, and $\mathbf{N}$ follows a MN distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, i.e., $\mathbf{N} \sim \text{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. It follows from (2) that $\mathbf{X}$ belongs to the class of multivariate normal variance-mean mixtures (cf. Barndorff-Nielsen et al., 1982) and that the expected value and covariance of $\mathbf{X}$ are given by, respectively,

$$\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu} + \boldsymbol{\alpha} \quad \text{and} \quad \text{Cov}(\mathbf{X}) = \boldsymbol{\Sigma} + \boldsymbol{\alpha}\boldsymbol{\alpha}'. \tag{3}$$

If $p = 1$, then the characteristic function of $\mathbf{X}$ corresponds to a univariate asymmetric Laplace distribution, i.e., $X \sim \text{AL}_1(\mu, \alpha, \phi)$, where $\nu = 1/2$ and the Bessel function can be written as $K_{1/2}(u) = \sqrt{\frac{\pi}{2u}}\exp\{-u\}$. As a result, the density of $X \sim \text{AL}_1(\mu, \alpha, \phi)$ can be expressed as

$$f_{\text{AL}}(x \mid \mu, \alpha, \phi) = \frac{1}{\gamma}\exp\left\{-\frac{|x-\mu|}{\phi^2}\left[\gamma - \alpha\,\text{sign}(x-\mu)\right]\right\}, \tag{4}$$

where $\gamma = \sqrt{\alpha^2 + 2\phi^2}$ and all other terms are as previously defined (see Kotz et al., 2001, for details).

2.2. A multiple scaled asymmetric Laplace distribution

As $\mathbf{X} \sim \mathrm{AL}_p(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$ belongs to the class of multivariate normal variance-mean mixtures, the density of $\mathbf{X}$ can also be expressed as

$$f_{\mathrm{MAL}}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \int_0^\infty f_{\mathrm{MN}}(\mathbf{x} \mid \boldsymbol{\mu} + w\boldsymbol{\alpha}, w\boldsymbol{\Sigma})\, g(w) dw, \tag{5}$$

where $f_{\mathrm{MN}}(\mathbf{x} \mid \boldsymbol{\mu} + w\boldsymbol{\alpha}, w\boldsymbol{\Sigma})$ is the multivariate Gaussian density with mean $\boldsymbol{\mu} + w\boldsymbol{\alpha}$ and covariance matrix $w\boldsymbol{\Sigma}$ and $g(w) = \exp\{-w\}$.

If we let

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{\Phi}\boldsymbol{\Gamma}', \tag{6}$$

and

$$g_{\mathbf{W}}(w_1, \ldots, w_p) = \prod_{h=1}^{p} g(w_h),$$

then we can define the joint pdf of the multiple scaled asymmetric Laplace (MSAL) distribution as

$$f_{\mathrm{MSAL}}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}, \boldsymbol{\phi}) = \int_0^\infty \cdots \int_0^\infty f_{\mathrm{MN}}\left(\mathbf{x} \mid \boldsymbol{\mu} + \boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{\alpha}, \boldsymbol{\Gamma}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\right) g_{\mathbf{W}}(w_1, \ldots, w_p)\, dw_1 \ldots dw_p, \tag{7}$$

where $\boldsymbol{\Gamma}$ is a matrix of eigenvectors, $\boldsymbol{\Phi}$ is a diagonal matrix of eigenvalues with elements $\phi_1, \ldots, \phi_p$, and $\boldsymbol{\Delta}_{\mathbf{w}}$ is a diagonal matrix with elements $w_1, \ldots, w_p$. Since the exponential random variables are independent, we can also express the density of the MSAL distribution as

$$\begin{aligned} f_{\mathrm{MSAL}}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}, \boldsymbol{\phi}) &= \prod_{h=1}^{p} \int_0^\infty f_{\mathrm{N}}\left(\left[\boldsymbol{\Gamma}'\mathbf{x}\right]_h \mid \left[\boldsymbol{\Gamma}'\boldsymbol{\mu}\right]_h + \left[\boldsymbol{\Gamma}'\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{\alpha}\right]_h, w_h\phi_h\right) \exp\{-w_h\} dw_h \\ &= \prod_{j=1}^{p} f_{\mathrm{AL}}\left(\left[\boldsymbol{\Gamma}'\mathbf{x}\right]_h \mid \left[\boldsymbol{\Gamma}'\boldsymbol{\mu}\right]_h, \left[\boldsymbol{\Gamma}'\boldsymbol{\alpha}\right]_h, \phi_h\right), \end{aligned} \tag{8}$$

where $[\mathbf{z}]_h$ is the hth element of the vector $\mathbf{z}$ and all model parameters are as previously defined.

3. The proposed model

The idea of contaminating multivariate distributions dates back at least as far as Tukey (1960) who introduced a multivariate contaminated normal distribution. This is a two-component normal mixture in which one of the components typically represents the good observations with probability ρ and the other component represents the bad observations with probability $1 - \rho$. Both components share the same mean, but the component representing the bad observations has an inflated variance with respect to the contamination parameter $\eta > 1$ (Aitkin and Wilson, 1980). Formally, we can write the density of the multivariate contaminated normal distribution as

$$f_{\mathrm{MCN}}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \rho, \eta) = \rho f_{\mathrm{MN}}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) + (1 - \rho) f_{\mathrm{MN}}(\mathbf{x} \mid \boldsymbol{\mu}, \eta\boldsymbol{\Sigma}), \tag{9}$$

where all terms are as previously defined.

Replacing the multivariate normal density functions in (9) can extend the contaminated framework to include other distributions. For example, Morris et al. (2019) developed an MCAL distribution, and Melnykov et al. (2021) proposed a multivariate contaminated normal mixture model that utilizes a transformation of the observed data; other examples are given by Mazza and Punzo (2019), Tomarchio and Punzo (2020), and Punzo and Bagnato (2021a), just to cite a few. We can easily develop a contaminated mixture of MSAL distributions by replacing the multivariate normal density functions in (9) with the density given in (8). However, as mentioned in Section 1, this model will not address deficiency (*a*). So, we consider a model that allows for the proportion of good points and degree of contamination to be modeled separately in each PC of the data. Formally, we write the density of the proposed MSCAL distribution as

$$\begin{aligned} f_{\mathrm{MSCAL}}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}, \boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\eta}) = \prod_{h=1}^{p} \Big[& \rho_h f_{\mathrm{AL}}\left(\left[\boldsymbol{\Gamma}'\mathbf{x}\right]_h \mid \left[\boldsymbol{\Gamma}'\boldsymbol{\mu}\right]_h, \left[\boldsymbol{\Gamma}'\boldsymbol{\alpha}\right]_h, \phi_h\right) \\ & + (1 - \rho_h) f_{\mathrm{AL}}\left(\left[\boldsymbol{\Gamma}'\mathbf{x}\right]_h \mid \left[\boldsymbol{\Gamma}'\boldsymbol{\mu}\right]_h, \sqrt{\eta_h}\left[\boldsymbol{\Gamma}'\boldsymbol{\alpha}\right]_h, \eta_h\phi_h\right)\Big], \end{aligned} \tag{10}$$

where $\rho_h \in (0,1)$ and $\eta_h > 1$ give, respectively, the proportion of good points and the degree of contamination in each PC, $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ are, respectively, the location and skewness parameters for the observed data, and all other terms and model parameters are as previously defined.

Fig. 1 displays contour plots obtained from the MCAL distribution (column 1), the MSAL distribution (column 2), and the MSCAL distribution (column 3). In every plot, the contours are centered at the origin. In column 1, $\rho = 0.75$ and $\eta = 5$. In column 3, $\boldsymbol{\rho} = (0.75, 0.75)'$ and $\boldsymbol{\eta} = (5, 10)'$. In row 1, $\boldsymbol{\alpha} = (0,0)'$ and $\text{vec}(\boldsymbol{\Sigma}) = (1,0,0,1)'$. In row 2, $\boldsymbol{\alpha} = (1,1)'$ and $\text{vec}(\boldsymbol{\Sigma}) = (1,0,0,1)'$. In row 3, $\boldsymbol{\alpha} = (1,1)'$ and $\text{vec}(\boldsymbol{\Sigma}) = (1,0.5,0.5,1)'$. Fig. 1 shows the effects that skewness and correlation have on the contours of the considered distributions. In columns 1 and 2, the contours take on the expected shapes following the discussions given in Morris et al. (2019) and Franczak et al. (2015). The most polarizing difference between the images in columns 1 and 2 is the rigid hypercube shapes of the MSAL distribution compared to the more "traditional" elliptical shape of the MCAL distribution. Comparing columns 2 and 3, we can see how the contours of the MSCAL distribution adapt to the influence of bad observations, becoming less rigid than the MSAL contours shown in column 2.

4. Parameter estimation

The expectation-maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Peel, 1998) is an iterative procedure that is commonly used to find maximum likelihood (ML) estimates in the presence of missing or incomplete data. The EM algorithm iterates between two steps: an E-step, where the expected value of the complete-data log-likelihood is computed, and an M-step, where the expected complete-data log-likelihood is maximized with respect to the model parameters. If one or more updates on either the E-step or M-step are analytically intractable, a Monte Carlo method can be used. The corresponding algorithm is called a Monte Carlo EM (MCEM). Commonly, MCEM refers to the use of a Monte Carlo method in the E-step; however, the Monte Carlo method can also be used in the M-step (see Section 6.2.1, p. 220–221, of McLachlan and Krishnan, 2007, for details). For the MSCAL we use an MCEM with a Monte Carlo method on the M-step since a closed-form estimate for $\boldsymbol{\Gamma}$ does not exist (see Section 4.3).

For the proposed model, we introduce a multi-dimensional indicator variable,

$$V_{ih} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ is good with respect to the } h\text{th PC} \\ 0 & \text{if } \mathbf{x}_i \text{ is bad with respect to the } h\text{th PC,} \end{cases}$$

for $i = 1, \ldots, n$ and $h = 1, \ldots, p$. It follows that the complete-data for the proposed MSCAL is comprised of the observed $\mathbf{x}_i, \ldots, \mathbf{x}_n$ and the missing $\mathbf{V}_1, \ldots, \mathbf{V}_n$, where $\mathbf{V}_i = (V_{i1}, \ldots, V_{ip})$. So, we can write the complete-data likelihood as

$$\begin{aligned} \mathcal{L}^c(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\rho}, \boldsymbol{\eta} \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) = &\prod_{i=1}^{n}\prod_{h=1}^{p} \left[\rho_h f_{\text{N}}\left(\left[\boldsymbol{\Gamma}'\mathbf{x}_i\right]_h \mid \left[\boldsymbol{\Gamma}'\boldsymbol{\mu}\right]_h + \left[\boldsymbol{\Gamma}'\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{\alpha}\right]_h, w_{ih}\phi_h\right) \exp\{-w_{ih}\}\right]^{v_{ih}} \\ &\times \prod_{i=1}^{n}\prod_{h=1}^{p} \left[(1-\rho_h) f_{\text{N}}\left(\left[\boldsymbol{\Gamma}'\mathbf{x}_i\right]_h \mid \left[\boldsymbol{\Gamma}'\boldsymbol{\mu}\right]_h + \sqrt{\eta}_h\left[\boldsymbol{\Gamma}'\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{\alpha}\right]_h, w_{ih}\eta_h\phi_h\right) \exp\{-w_{ih}\}\right]^{1-v_{ih}}, \end{aligned} \tag{11}$$

where all terms and model parameters are as previously defined.

For the proposed MSCAL we have two sources of missing data: the V_{ih} and the latent W_{ih}, for $i = 1, \ldots, n$ and $h = 1, \ldots, p$. From (11), we can write the complete-data log-likelihood of the MSCAL as

$$\ell_c(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\rho}, \boldsymbol{\eta}) = \ell_{c1}(\boldsymbol{\rho}) + \ell_{c2}(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\eta}), \tag{12}$$

where

$$\ell_{c1}(\boldsymbol{\rho}) = \sum_{i=1}^{n}\sum_{h=1}^{p}\left[v_{ih}\log\rho_h + (1-v_{ih})\log(1-\rho_h)\right] \tag{13}$$

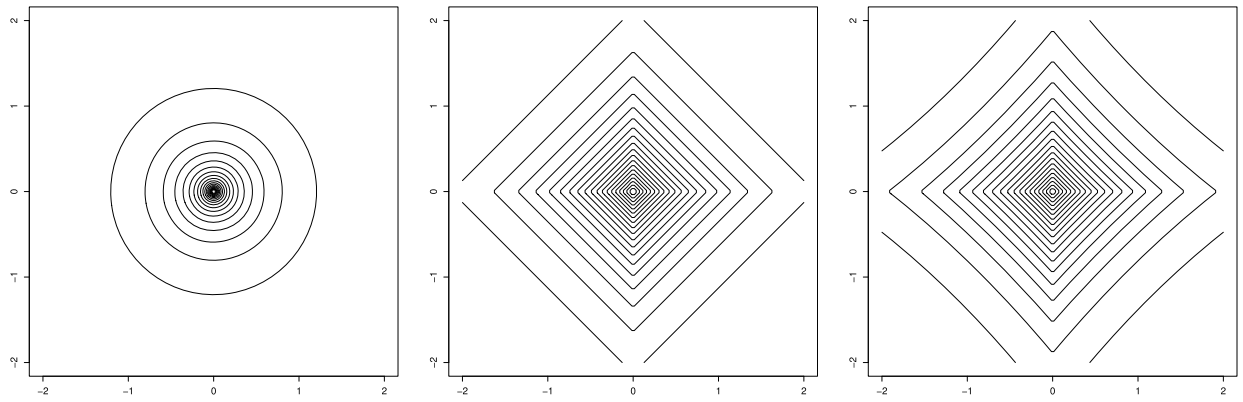
and

$$\ell_{c2}(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\eta}) = \sum_{i=1}^{n}\sum_{h=1}^{p} v_{ih}\log f_{\text{N}}\left(Y_{ih} \mid \mu_h^* + w_{ih}\alpha_h^*, w_{ih}\phi_h\right) + \sum_{i=1}^{n}\sum_{h=1}^{p}(1-v_{ih})\log f_{\text{N}}\left(Y_{ih} \mid \mu_h^* + w_{ih}\sqrt{\eta_h}\alpha_h^*, w_{ih}\eta_h\phi_h\right), \tag{14}$$

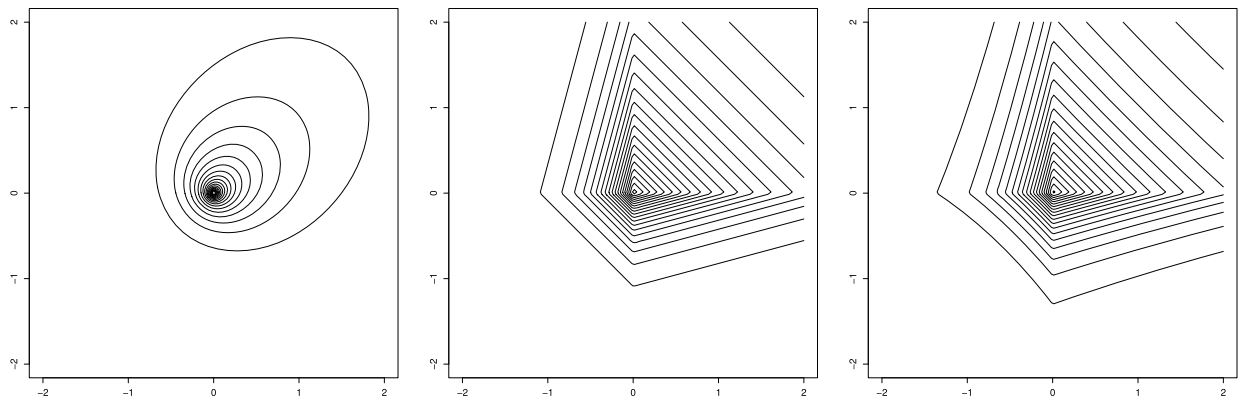
where $Y_{ih} = \left[\boldsymbol{\Gamma}'\mathbf{x}_i\right]_h$ is the hth element of the principal component transformation of $\mathbf{x}_i$, $\mu_h^* = \left[\boldsymbol{\Gamma}'\boldsymbol{\mu}\right]_h$ and $\alpha_h^* = \left[\boldsymbol{\Gamma}'\boldsymbol{\alpha}\right]_h$ represent, respectively, the location and skewness parameters for the hth PC, and all other terms are as previously defined.

4.1. E-step

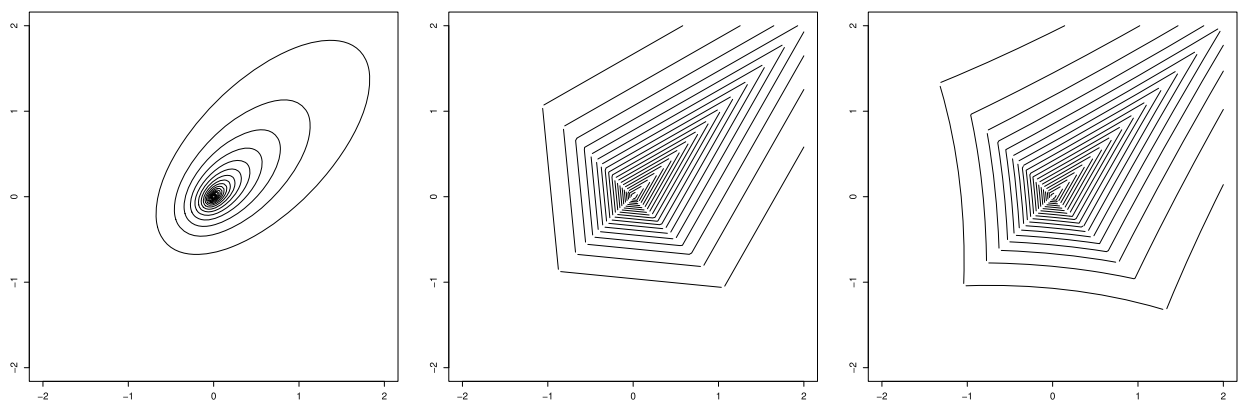
Recall that the latent W_{ih}, for $i = 1, \ldots, n$ and $h = 1, \ldots, p$, are exponential random variables with rate 1. Following Franczak et al. (2014), Franczak et al. (2015) and Morris et al. (2019), we can show that $W_{ih} \mid Y_{ij},\ v_{ih} = 1 \sim \text{GIG}\left(\chi_h, \omega_{ih}, 0.5\right)$, $W_{ih} \mid Y_{ij},\ v_{ih} = 0 \sim \text{GIG}\left(\chi_h, \omega_{ih}^b, 0.5\right)$, and $W_{ih} \mid \mathbf{x}_i, v_{ih} = 0 \sim \text{GIG}\left(\chi_h, \omega_{ih}^b, (2-p)/2\right)$, i.e., that each of these conditional random variables follow a generalized inverse Gaussian (GIG) distribution (Good, 1953). Given this result, we utilize the expected values of the GIG distribution laid down by Jørgensen (1982) to derive a portion of the expected values needed on the E-step of the proposed MCEM algorithm. In Appendix A, we give the density function and pertinent expected values of the GIG distribution.



(a) Symmetric, uncorrelated MCAL (b) Symmetric, uncorrelated MSAL (c) Symmetric, uncorrelated MSCAL



(d) Skewed, uncorrelated MCAL (e) Skewed, uncorrelated MSAL (f) Skewed, uncorrelated MSCAL



(g) Skewed, correlated MCAL (h) Skewed, correlated MSAL (i) Skewed, correlated MSCAL

Fig. 1. Examples of contour plots from the MCAL distribution (column 1), the MSAL distribution (column 2), and the MSCAL distribution (column 3) for varying parameter sets.

Formally, on the E-step of the proposed parameter estimation scheme, we have the following expected values:

$$\ddot{E}_{Vih} := \mathbb{E}[v_{ih} \mid \mathbf{x}_i] = \frac{\dot{\rho}_h f_{\text{AL}}\left(\dot{Y}_{ih} \mid \dot{\mu}_h^*, \dot{\alpha}_h^*, \dot{\phi}_h\right)}{\dot{\rho}_h f_{\text{AL}}\left(\dot{Y}_{ih} \mid \dot{\mu}_h^*, \dot{\alpha}_h^*, \dot{\phi}_h\right) + (1-\dot{\rho}_h) f_{\text{AL}}\left(\dot{Y}_{ih} \mid \dot{\mu}_h^*, \sqrt{\dot{\eta}_h}\dot{\alpha}_h^*, \dot{\eta}_h\dot{\phi}_h\right)}, \tag{15}$$

$$\ddot{E}_{1ih} := \mathbb{E}\left[w_{ih} \mid Y_{ih}, v_{ih}=1\right] = \frac{\sqrt{\dot{\omega}_{ih}} K_{1.5}\left(\sqrt{\dot{\chi}_h \dot{\omega}_{ih}}\right)}{\sqrt{\dot{\chi}_h} K_{0.5}\left(\sqrt{\dot{\chi}_h \dot{\omega}_{ih}}\right)}, \tag{16}$$

$$\ddot{E}_{2ih} := \mathbb{E}\left[w_{ih}^{-1} \mid Y_{ih}, v_{ih}=1\right] = \frac{\sqrt{\dot{\chi}_h} K_{1.5}\left(\sqrt{\dot{\chi}_h \dot{\omega}_{ih}}\right)}{\sqrt{\dot{\omega}_{ih}} K_{0.5}\left(\sqrt{\dot{\chi}_h \dot{\omega}_{ih}}\right)} - \frac{1}{\dot{\omega}_{ih}}, \tag{17}$$

$$\ddot{E}_{1ih}^b := \mathbb{E}\left[w_{ih} \mid Y_{ih}, v_{ih}=0\right] = \frac{\sqrt{\dot{\omega}_{ih}^b} K_{1.5}\left(\sqrt{\dot{\chi}_h \dot{\omega}_{ih}^b}\right)}{\sqrt{\dot{\chi}_h} K_{0.5}\left(\sqrt{\dot{\chi}_h \dot{\omega}_{ih}^b}\right)}, \text{ and} \tag{18}$$

$$\ddot{E}_{2ih}^b := \mathbb{E}\left[w_{ih}^{-1} \mid Y_{ih}, v_{ih}=0\right] = \frac{\sqrt{\dot{\chi}_h} K_{1.5}\left(\sqrt{\dot{\chi}_h \dot{\omega}_{ih}^b}\right)}{\sqrt{\dot{\omega}_{ih}^b} K_{0.5}\left(\sqrt{\dot{\chi}_h \dot{\omega}_{ih}^b}\right)} - \frac{1}{\dot{\omega}_{ih}^b}, \tag{19}$$

where $\dot{\chi}_h = 2 + \left(\dot{\alpha}_h^*\right)^2 \dot{\phi}_h^{-1}$, $\dot{\omega}_{ih} = \left(\dot{Y}_{ih} - \dot{\mu}_h^*\right)^2 \dot{\phi}_h^{-1}$, $\dot{\omega}_{ih}^b = \left(\dot{Y}_{ih} - \dot{\mu}_h^*\right)^2 \left(\dot{\eta}_h \dot{\phi}_h\right)^{-1}$, one dot represents an update at the previous iteration, and two dots represent an update on the current iteration.

4.2. M-step

On the M-step of this MCEM algorithm, we have the following updates:

$$\ddot{\rho}_h = \sum_{i=1}^n \ddot{a}_{ih}, \tag{20}$$

$$\ddot{\mu}_h^* = \frac{\ddot{A}\sum_{i=1}^n \dot{Y}_{ih}\left(\ddot{b}_{ih} + \frac{\ddot{d}_{ih}}{\dot{\eta}_h}\right) - \ddot{C}\sum_{i=1}^n \dot{Y}_{ih}\left(\ddot{a}_{ih} + \frac{\ddot{c}_{ih}}{\sqrt{\dot{\eta}_h}}\right)}{\ddot{A}\ddot{B} - \ddot{C}^2}, \tag{21}$$

$$\ddot{\alpha}_h^* = \frac{\ddot{B}\sum_{i=1}^n \dot{Y}_{ih}\left(\ddot{a}_{ih} + \frac{\ddot{c}_{ih}}{\sqrt{\dot{\eta}_h}}\right) - \ddot{C}\sum_{i=1}^n \dot{Y}_{ih}\left(\ddot{b}_{ih} + \frac{\ddot{d}_{ih}}{\dot{\eta}_h}\right)}{\ddot{A}\ddot{B} - \ddot{C}^2}, \text{ and} \tag{22}$$

$$\ddot{\phi}_h = \frac{1}{n}\sum_{i=1}^n \left[\left(\ddot{b}_{ih} + \frac{\ddot{d}_{ih}}{\dot{\eta}_h}\right)\left(\dot{Y}_{ih} - \ddot{\mu}_h^*\right)^2 - 2\left(\ddot{a}_{ih} + \frac{\ddot{c}_{ih}}{\sqrt{\dot{\eta}_h}}\right)\left(\dot{Y}_{ih} - \ddot{\mu}_h^*\right)\ddot{\alpha}_h^* + \left(\ddot{\alpha}_h^*\right)^2 \ddot{A}\right], \tag{23}$$

where $\ddot{a}_{ih} = \ddot{E}_{Vih}$, $\ddot{b}_{ih} = \ddot{a}_{ih}\ddot{E}_{2ih}$, $\ddot{c}_{ih} = (1 - \ddot{E}_{Vih})$, $\ddot{d}_{ih} = \ddot{c}_{ih}\ddot{E}_{2ih}^b$, $\ddot{f}_{ih} = \ddot{a}_{ih}\ddot{E}_{1ih}$, $\ddot{h}_{ih} = \ddot{c}_{ih}\ddot{E}_{1ih}^b$, $\ddot{A} = \sum_{i=1}^n \left(\ddot{f}_{ih} + \ddot{h}_{ih}\right)$, $\ddot{B} = \sum_{i=1}^n \left(\ddot{b}_{ih} + \frac{\ddot{d}_{ih}}{\dot{\eta}_h}\right)$, and $\ddot{C} = \sum_{i=1}^n \left(\ddot{a}_{ih} + \frac{\ddot{c}_{ih}}{\sqrt{\dot{\eta}_h}}\right)$.

To update η_h, we solve the equation

$$\sum_{i=1}^n \eta_h \ddot{c}_{ih}\ddot{\Phi}_h + \sqrt{\eta_h}\ddot{c}_{ih}(\dot{Y}_{ih} - \ddot{\mu}_h^*)\ddot{\alpha}_h^* - \ddot{d}_{ih}(\dot{Y}_{ih} - \ddot{\mu}_h^*)^2 = 0.$$

Setting $\ddot{a}^* = \sum_{i=1}^n \ddot{c}_{ih}\ddot{\Phi}_h$, $\ddot{b}^* = \sum_{i=1}^n \ddot{c}_{ih}(\dot{Y}_{ih} - \ddot{\mu}_h^*)\ddot{\alpha}_h^*$, and $\ddot{c}^* = -\sum_{i=1}^n \ddot{d}_{ih}(\dot{Y}_{ih} - \ddot{\mu}_h^*)^2$, gives

$$\ddot{\eta}_h = \left(\frac{-\ddot{b}^* \pm \sqrt{\left(\ddot{b}^*\right)^2 - 4\ddot{a}^*\ddot{c}^*}}{2\ddot{a}^*}\right)^2. \tag{24}$$

Since (24) returns two solutions, we take the one greater than one.

4.3. Estimation of the orthogonal matrix

A closed-form estimate for $\boldsymbol{\Gamma}$ cannot be found either directly or through the EM algorithm. Therefore, we propose a two-step Monte Carlo optimization procedure that can be schematized as follows.

1. Compute the eigen-decomposition of the sample covariance matrix $\mathbf{S}$ and retain the resulting eigenvector matrix, say $\boldsymbol{\Gamma}_{\mathbf{S}}$.
2. Map $\boldsymbol{\Gamma}_{\mathbf{S}}$ to a $(p \times p)$ unit lower triangular matrix, say $\mathbf{L}$, having $p(p-1)/2$ unconstrained real-valued entries, via the PLR decomposition proposed by Bagnato and Punzo (2021).
3. At the first step of the procedure, for $r = 1, \ldots, R$:
   (a) generate a new $\mathbf{L}$-matrix, say $\dot{\mathbf{L}}_r$, by adding a uniform random number in $(-0.1, 0.1)$ to each element under the main diagonal of $\mathbf{L}$;
   (b) obtain the corresponding orthogonal matrix $\dot{\boldsymbol{\Gamma}}_r$ via the PLR back-decomposition of $\dot{\mathbf{L}}_r$;
   (c) run the EM algorithm with $\boldsymbol{\Gamma}$ fixed to $\dot{\boldsymbol{\Gamma}}_r$ and retain the observed data log-likelihood value at convergence, say $\dot{\ell}_r$.
4. Compute the maximum among the $\dot{\ell}_1, \ldots, \dot{\ell}_R$ values and retain the corresponding $\dot{\boldsymbol{\Gamma}}_r$ and $\dot{\mathbf{L}}_r$ matrices, say $\dot{\boldsymbol{\Gamma}}$ and $\dot{\mathbf{L}}$, respectively.
5. At the second, finer, step of the procedure, for $s = 1, \ldots, S$:
   (a) generate a new $\mathbf{L}$-matrix, say $\ddot{\mathbf{L}}_s$, by adding a uniform random number in $(-0.02, 0.02)$ to each element under the main diagonal of $\dot{\mathbf{L}}$;
   (b) obtain the corresponding orthogonal matrix $\ddot{\boldsymbol{\Gamma}}_s$ via the PLR back-decomposition of $\ddot{\mathbf{L}}_s$;
   (c) run the EM algorithm with $\boldsymbol{\Gamma}$ fixed to $\ddot{\boldsymbol{\Gamma}}_s$ and retain the observed data log-likelihood value at convergence, say $\ddot{\ell}_s$.
6. Compute the maximum among the $\ddot{\ell}_1, \ldots, \ddot{\ell}_S$ values and consider the corresponding $\ddot{\boldsymbol{\Gamma}}_s$ as the ML estimate of $\boldsymbol{\Gamma}$.

4.4. Computational details

The proposed MCEM algorithm is implemented in R (R Core Team, 2018). In all considered applications, $\boldsymbol{\mu}$ is initialized at the mean of the observed data, $\boldsymbol{\alpha}$ is set to $\mathbf{1}$, and in every PC, $\rho_h = 0.9$, $\eta_h = 2$, and $\phi_h = 1$, for $h = 1, \ldots, p$. In the two-step Monte Carlo optimization procedure discussed in Section 4.3, we fix $R = 100$ and $S = 40$. The MCEM algorithm is then run, starting with an E-step. The algorithm is iterated until convergence, or for a maximum of 100 iterations. Convergence is measured using a stopping criterion based on Aitken's acceleration (Aitken, 1926) with a tolerance of 0.1 (see McNicholas et al., 2010, for details).

Following the proposal in Punzo and Bagnato (2021b), in the data analyses discussed in Sections 6 and 7, three approaches to compute the ML estimates were considered. For each data set, a direct approach with the Nelder-Mead algorithm, a direct approach with the BFGS algorithm, and the MCEM algorithm discussed above were used to estimate the parameters. In the end, the solution providing the best-observed data log-likelihood value was retained. In Punzo and Bagnato (2021b), this approach has been shown to provide a higher likelihood value than when any one of the three methods is used alone, i.e. the highest log-likelihood value is not always obtained with the same method.

The direct approaches were implemented via the general-purpose optimizer `optim()` for R (R Core Team, 2018), included in the **stats** package.

5. Other important aspects

5.1. Identifiability

An important point in dealing with the proposed MSCAL model is establishing its identifiability. Without identifiability, the parameters might not be estimated and interpreted, and, more generally, the inference might be meaningless (Wang et al., 2014). Following the arguments below, we show that the model proposed in (10) is identifiable.

Tortora et al. (2019) prove that the identifiability of a multiple scaled distribution is ensured if the p univariate distributions on the PCs are identifiable. On the generic hth PC of model (10), for $h = 1, \ldots, p$, we have the (univariate) CAL distribution with pdf

$$f_{\text{CAL}}\left(y|\mu_h^*, \alpha_h^*, \phi_h, \rho_h, \eta_h\right) = \rho_h f_{\text{AL}}\left(y|\mu_h^*, \alpha_h^*, \phi_h\right) + \left(1-\rho_h\right) f_{\text{AL}}\left(y|\mu_h^*, \sqrt{\eta_h}\alpha_h^*, \eta_h\phi_h\right), \tag{25}$$

which is a mixture of two (univariate) AL distributions that only differ in terms of asymmetry (by $\sqrt{\eta_h}$) and scale (by η_h) parameters (cf. Punzo and Bagnato, 2021a, 2022b). Now, the AL distribution is a special case of the generalized hyperbolic (GH) distribution (see Section 2.2 of Browne and McNicholas, 2015). Browne and McNicholas (2015) proved, in the multivariate context, the identifiability (up to label switching) for finite mixtures of GH distributions. Therefore, it follows that the MSCAL distribution is identifiable (up to label switching).

As for the label-switching issue, it is overcome by using the constraint $\eta_h > 1$, $h = 1, \ldots, p$, as explained below. Suppose we relax this assumption on η_h so that $\eta_h > 0$, for $h = 1, \ldots, p$. Under this assumption, model (25) is nonidentifiable due to label-switching because, if $\tilde{\mu}_h^* = \mu_h^*$, $\tilde{\alpha}_h^* = \sqrt{\eta_h}\alpha_h^*$, $\tilde{\phi}_h = \eta_h\phi_h$, $\tilde{\rho}_h = 1-\rho_h$, and $\tilde{\eta}_h = 1/\eta_h$, then $f_{\text{CAL}}\left(y|\mu_h^*, \alpha_h^*, \phi_h, \rho_h, \eta_h\right) = f_{\text{CAL}}\left(y|\tilde{\mu}_h^*, \tilde{\alpha}_h^*, \tilde{\phi}_h, \tilde{\rho}_h, \tilde{\eta}_h\right)$. This tricky label-switching case, which is the only one possible, can be avoided by adding at least one of the following constraints:

Table 1
Values of $\boldsymbol{\alpha}$, $\boldsymbol{\rho}$, and $\boldsymbol{\eta}$ used in the eight different scenarios considered in this simulation study.

	$\boldsymbol{\alpha}$		$\boldsymbol{\rho}$		$\boldsymbol{\eta}$	
Scenario	Dim. 1	Dim. 2	PC 1	PC 2	PC 1	PC 2
S1	−2	2	0.9	0.9	3	3
S2	−2	2	0.9	0.9	6	10
S3	−2	2	0.7	0.9	3	3
S4	−2	2	0.7	0.9	6	10
S5	−2	6	0.9	0.9	3	3
S6	−2	6	0.9	0.9	6	10
S7	−2	6	0.7	0.9	3	3
S8	−2	6	0.7	0.9	6	10

1. $\rho_h > 0.5$ (as $\rho_h = 1 - \tilde{\rho}_h$, it follows that $\tilde{\rho}_h < 0.5$ and we obtain a contradiction);
2. $\eta_h > 1$ (as $\eta_h = 1/\tilde{\eta}_h$, it follows that $\tilde{\eta}_h \in (0, 1)$ and we obtain a contradiction).

Note that adding both constraints (as we do in Section 5.3) is not necessary but it has an interpretative advantage from a robust statistics perspective. The simultaneous use of both constraints ($\rho_h > 0.5$ and $\eta_h > 1$) allows us to label the observations as either 'good' or 'bad' and it forces the number of bad observations to be less than the good ones, while having a larger variability, allowing the user to perform directional outlier detection. Nevertheless, considering both constraints simultaneously would introduce an extra restriction on the parameter space, reducing the number of members/models of the MSCAL family. This is the reason why, in defining our model, we only use the constraint $\eta_h > 1$, $h = 1, \ldots, p$.

To complete the discussion on identifiability, it is important to realize that the two (trivial but important) conditions $\eta \neq 1$ and $\rho_h > 0$ also prevent overfitting (a potential problem for identifiability first noted by Crawford, 1994). Indeed, identifiability problems may occur due to empty AL components (i.e., when either $\rho_h = 0$ or $\rho_h = 1$), where their parameters cannot be uniquely determined and due to components with equal component parameter vectors (i.e., when either $\eta_h = 1$) where different values for ρ_h are possible (see Frühwirth-Schnatter, 2006, Chapter 1.3, for details).

5.2. Existence of a global maximum

The traditional EM algorithm monotonically increases the observed data (log-)likelihood function and returns the ML parameter estimates. However, a well-known issue in mixture models is the potential nonexistence of the global maximizer for ML estimates. With unrestricted covariance matrices, mixtures of normal distributions do not have a global maximizer (Melnykov, 2013). Since the AL itself is an infinite mixture of normal distributions, we expect to inherit the same issue.

In our MCEM algorithm, for every MC estimate of $\boldsymbol{\Gamma}$, we run the EM algorithm to estimate all the other parameters until convergence is reached. Therefore, the monotonicity is maintained within each EM step of the MCEM algorithm but not between the MC and EM steps. Anyway, it is important to point out that, in some cases, the MCEM algorithm gets closer to a maximizer with high probability (Booth and Hobert, 1999).

5.3. Automatic directional detection of outliers

At convergence, adding suitable constraints (cf. Section 5.1), the MSCAL model can be used for directional outlier detection on the PC space. Following Punzo and Tortora (2021), we additionally require that at least half of the observations are good points on the hth PC, i.e., $\rho_h > 0.5$. Under this additional constraint, $(1 - \rho_h)$ and η_h represent the proportion of outlying observations and degree of contamination, respectively, and E_{Vih} gives the *a posteriori* probability that the observation $\mathbf{x}_i$ is good with respect to the hth PC. Therefore, we label $\mathbf{x}_i$ in the hth PC, what we refer to as y_{ih}, as good if $E_{Vih} > 0.5$, for $i = 1, \ldots, n$ and $h = 1, \ldots, p$.

6. Simulation study

We use a simulation study to measure the computational time required to run the proposed MCEM algorithm and to evaluate parameter recovery. In total, we consider eight different parameter sets to generate the data, herein referred to as scenarios (see Table 1). For all scenarios, $\boldsymbol{\mu} = (0, 0)'$ and $\text{vec}(\boldsymbol{\Sigma}) = (4, -0.8, -0.8, 1)'$. The parameters $\boldsymbol{\alpha}$, $\boldsymbol{\rho}$, and $\boldsymbol{\eta}$ were allowed to take on one of two possible sets of values. Specifically, $\boldsymbol{\alpha}$ was either $(-2, 2)'$ or $(-2, 6)'$, $\boldsymbol{\rho}$ was either $(0.9, 0.9)'$ or $(0.7, 0.9)'$, and $\boldsymbol{\eta}$ was either $(3, 3)'$ or $(6, 10)'$. For each scenario, we simulated 100 data sets of size $n = 500$, 100 data sets of size $n = 1000$, and 100 data sets of size $n = 2000$. On each data set, the parameters are estimated using the approach described in Section 4.4. The MCEM algorithm is run until convergence with a tolerance of 0.1 or for a maximum of 100 iterations. The MCEM reached convergence before 100 iterations in 94% of the cases.

Table 2 shows the average computational time (in seconds) of the MCEM algorithm for each scenario and considered sample size. The overall average time is 317 seconds. As the value of n increases the average elapsed time also increases. For $n = 500$ the overall average is 148 seconds, for $n = 1000$ and $n = 2000$ the overall averages are 273 and 529 seconds, respectively.

Table 2
Average computational time to run the MCEM algorithm per scenario and values of n.

Scenario	500	1000	2000
S1	126.55	245.03	503.79
S2	126.34	249.14	508.72
S3	122.43	242.58	502.74
S4	127.83	237.03	500.68
S5	164.64	311.10	559.54
S6	177.97	311.09	557.75
S7	174.36	297.66	551.86
S8	168.71	296.36	549.12

Table 3
For each scenario, the bias and variance of the estimates of μ_j and α_j in each dimension j of the observed space and of ρ_h and η_h for the hth PC, $j, h = 1, 2$.

		Bias $[\hat{\boldsymbol{\mu}}^*]$			Var $[\hat{\boldsymbol{\mu}}^*]$			Bias $[\hat{\boldsymbol{\alpha}}^*]$			Var $[\hat{\boldsymbol{\alpha}}^*]$		
	n	500	1000	2000	500	1000	2000	500	1000	2000	500	1000	2000
S1	Dim. 1	-0.00	0.01	-0.01	0.01	0.00	0.01	0.11	0.07	0.10	0.10	0.07	0.05
	Dim. 2	0.00	-0.00	-0.00	0.00	0.00	0.00	-0.08	-0.05	-0.09	0.05	0.03	0.02
S2	Dim. 1	0.00	0.00	0.00	0.00	0.02	0.00	0.15	0.07	0.03	0.12	0.07	0.03
	Dim. 2	0.00	0.00	0.00	0.00	0.00	0.00	-0.08	-0.02	-0.02	0.04	0.02	0.01
S3	Dim. 1	0.01	-0.02	-0.01	0.00	0.01	0.02	-0.04	-0.04	-0.08	0.16	0.09	0.07
	Dim. 2	-0.00	0.00	0.02	0.00	0.00	0.01	-0.08	-0.03	-0.06	0.04	0.03	0.02
S4	Dim. 1	-0.02	-0.00	-0.00	0.01	0.00	0.01	0.03	0.05	-0.01	0.16	0.09	0.07
	Dim. 2	0.00	-0.00	0.00	0.00	0.00	0.00	-0.04	-0.05	-0.01	0.02	0.02	0.01
S5	Dim. 1	0.00	0.01	0.01	0.01	0.02	0.01	0.06	0.05	0.05	0.22	0.14	0.07
	Dim. 2	-0.00	0.00	0.01	0.00	0.01	0.01	-0.12	-0.21	-0.16	0.39	0.35	0.18
S6	Dim. 1	0.00	0.01	0.01	0.00	0.01	0.02	0.17	0.11	0.05	0.22	0.12	0.06
	Dim. 2	-0.00	-0.00	-0.00	0.00	0.00	0.00	-0.17	-0.07	-0.03	0.37	0.18	0.07
S7	Dim. 1	0.00	-0.00	-0.01	0.02	0.00	0.01	-0.13	-0.04	-0.09	0.38	0.22	0.16
	Dim. 2	0.00	0.00	0.01	0.00	0.00	0.00	-0.11	-0.12	-0.06	0.38	0.32	0.17
S8	Dim. 1	-0.00	-0.01	-0.00	0.01	0.02	0.00	-0.08	-0.07	-0.00	0.29	0.21	0.10
	Dim. 2	-0.00	-0.01	0.00	0.00	0.01	0.00	-0.18	-0.11	-0.02	0.43	0.14	0.09
		Bias $[\hat{\boldsymbol{\rho}}]$			Var $[\hat{\boldsymbol{\rho}}]$			Bias $[\hat{\boldsymbol{\eta}}]$			Var $[\hat{\boldsymbol{\eta}}]$		
	n	500	1000	2000	500	1000	2000	500	1000	2000	500	1000	2000
S1	PC 1	-0.11	-0.09	-0.14	0.05	0.04	0.04	-1.16	-1.07	-1.10	0.82	1.07	0.50
	PC 2	-0.06	-0.08	-0.12	0.04	0.04	0.04	-1.01	-0.91	-1.03	1.31	2.80	0.58
S2	PC 1	-0.13	-0.06	-0.04	0.04	0.02	0.01	-3.45	-3.32	-3.47	0.95	1.18	0.34
	PC 2	-0.06	-0.02	-0.02	0.02	0.01	0.00	-6.70	-6.73	-6.81	0.84	0.49	0.27
S3	PC 1	0.03	0.01	0.03	0.04	0.03	0.03	-0.80	-1.14	-1.04	0.91	0.21	0.24
	PC 2	-0.10	-0.09	-0.14	0.04	0.04	0.04	-1.20	-1.08	-1.04	0.70	0.88	0.59
S4	PC 1	-0.02	-0.02	0.00	0.02	0.02	0.01	-3.30	-3.43	-3.49	0.33	0.13	0.05
	PC 2	-0.03	-0.03	-0.02	0.01	0.01	0.00	-6.55	-6.68	-6.78	1.23	0.81	0.32
S5	PC 1	-0.06	-0.07	-0.10	0.04	0.03	0.03	-1.21	-1.14	-1.19	1.58	0.69	0.50
	PC 2	-0.03	-0.09	-0.08	0.03	0.04	0.03	-1.04	-1.02	-0.99	1.51	0.99	1.71
S6	PC 1	-0.10	-0.07	-0.04	0.03	0.02	0.01	-3.31	-3.36	-3.48	1.33	1.29	0.41
	PC 2	-0.04	-0.01	-0.01	0.01	0.00	0.00	-6.70	-6.55	-6.81	1.14	0.83	0.24
S7	PC 1	0.03	-0.01	0.04	0.04	0.04	0.03	-0.99	-1.14	-1.02	3.06	0.31	0.21
	PC 2	-0.05	-0.07	-0.07	0.03	0.04	0.03	-1.27	-1.18	-1.08	1.11	1.06	0.66
S8	PC 1	0.01	0.02	0.00	0.02	0.02	0.01	-3.31	-3.27	-3.48	0.38	3.48	0.07
	PC 2	-0.05	-0.03	-0.01	0.02	0.01	0.00	-6.67	-6.87	-6.80	1.41	0.47	0.20

Table 3 gives bias and variance of the estimates of $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, $\boldsymbol{\eta}$, and $\boldsymbol{\rho}$. For $\boldsymbol{\Sigma}$, the bias and variance of each unique element are given in Table B.11 of Appendix B. The parameters $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\rho}$ have small biases and variances that do not seem to change in the different scenarios, the variances slightly reduce as n increases. The parameter $\boldsymbol{\eta}$ shows the highest bias (in absolute value) and variance, as expected. The parameters $\boldsymbol{\eta}$ and $\boldsymbol{\rho}$ are the most difficult to estimate. Like the degrees of freedom in the t distribution (Thompson et al., 2020), or the parameters impacting the tails of a heavy-tailed distribution in general (Punzo and Bagnato, 2021b), $\boldsymbol{\eta}$ and $\boldsymbol{\rho}$ need more data to have good convergence properties for the estimator. This is, even more, emphasized when the percentage of bad points is small and the tailedness parameter values are high. Looking at the simulation results, the bias increases significantly in S2, S4, S6, and S8, where $\boldsymbol{\eta} = (6, 10)'$. The values of $\boldsymbol{\eta}$ are extremely difficult to estimate because only a small percentage of data points belong to the bad component. On PC2, $\rho_2 = 0.9$ for all the scenarios, on PC1, $\rho_1 = 0.9$ in S2 and S6, and $\rho_1 = 0.7$ in S4 and S8. The results suggest that the percentage of outlying points affects the variance of $\hat{\boldsymbol{\eta}}$ since the values of Var $[\hat{\boldsymbol{\eta}}]$ are smaller for S4 and S8 than they are for S2 and S6.

Table 4
ML-estimates for a subset of the parameters for the MSCAL distribution when fitted to the entire `f.twins` data set.

	STA1	HIP1	CHE1	STA2	HIP2	CHE2
$\hat{\boldsymbol{\mu}}$	158.79	29.10	75.20	160.13	28.91	75.23
$\hat{\boldsymbol{\alpha}}$	-8.28	-1.85	-1.90	-9.15	-1.71	-2.09
	PC1	PC2	PC3	PC4	PC5	PC6
$\hat{\boldsymbol{\rho}}$	0.99	1.00	1.00	1.00	1.00	1.00
$\hat{\boldsymbol{\eta}}$	1.04	1.00	1.00	1.00	1.00	1.00

7. Analysis on anthropometric measurements of female twins

In this Section, we illustrate the proposed model using real data. Information about the data set is given in Section 7.1. In Section 7.2, we focus on two variables for the sake of illustration and graphical representation. In Section 7.2.1, we compare the ability of several well-established parametric models to reproduce the joint distribution of the two variables. Finally, in Section 7.2.2, we investigate how the MSAL distribution – the reference model for the good data behind our approach – reacts to the inclusion of *ad-hoc* located outliers and how these points are instead handled by the proposed MSCAL distribution.

7.1. The data

We consider the `f.twins` data set accompanying the **Flury** package (Flury, 1997) for R available at https://CRAN.R-project.org/package=Flury. The data set contains $p = 6$ anthropometric measurements for $n = 79$ pairs of monozygotic/dizygotic female twins measured in the 1950s at the University of Hamburg, Germany. The available measurements are the stature of the first twin (STA1), the hip circumference of the first twin (HIP1), the chest circumference of the first twin (CHE1), and the same measurements for the second twin, namely STA2, HIP2, and CHE2. The variables are expressed in centimeters. For further details on these data, see Flury (2013).

We fitted the MSCAL to the `f.twins` data set. Table 4 gives the corresponding parameter estimates. The estimate of $\hat{\boldsymbol{\Sigma}}$ is reported in Appendix C, Table C.12.

The parameters estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ are measured in the observed data space, while $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{\eta}}$ are measured in the PC space. All the dimensions are negatively skewed, with the statures having the highest values of skewness. The hip circumferences are the least skewed for both twins. The values of $\hat{\boldsymbol{\mu}}$ do not seem to differ among twins. The values of $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\rho}}$ indicate that the are no outliers in the data set.

7.2. Stature and chest circumference of the second twin

In this section, we focus on $p = 2$ of the available measurements, STA2 and CHE2. The scatter plot of the data is displayed in Fig. 2.

The scatter plot in Fig. 2 looks negatively skewed in both dimensions and this is corroborated by the Mardia test of symmetry (p-value = 0.002), as implemented by the `mvn()` function of the **MVN** package (Korkmaz et al., 2019). A visual inspection of Fig. 2 does not give any evidence of outlying points.

7.2.1. Model comparison

The first aim of the analysis is to evaluate the best parametric model for the bivariate distribution of the data. Table 5 presents the model comparison. The first column provides the thirty models we consider. The first group of sixteen models includes the multivariate generalized hyperbolic (MGH) distribution (see, e.g., McNeil et al., 2005) and all of its special or limiting cases. For a useful hierarchical representation of the taxonomy among these models, see Bagnato et al. (2023). This group contains both elliptical-symmetric (the first nine models) and asymmetric models. Then we have: 1) four other elliptical-symmetric models, the multivariate tail-inflated normal (MTIN; Punzo and Bagnato, 2021b), multivariate shifted-exponential normal (MSEN; Punzo and Bagnato, 2020), multivariate leptokurtic normal (MLN; Bagnato et al., 2017), and multivariate contaminated normal (MCN; Punzo and McNicholas, 2016); 2) four other well-known asymmetric models, the multivariate skew-normal, multivariate skew-t, multivariate skew contaminated normal, as formulated by Cabral et al. (2012), and the multivariate contaminated asymmetric Laplace (MCAL; Morris et al., 2019); and 3) six multiple scaled distributions, the multiple scaled shifted-exponential normal (MSSEN; Punzo and Bagnato, 2022a)), multiple scaled t (MSt; Forbes and Wraith, 2014), and multiple scaled contaminated normal (MSCN; Punzo and Tortora, 2021) multiple scaled asymmetric Laplace (MSAL; Franczak et al., 2015), multiple scaled contaminated asymmetric Laplace (MSCAL), and multiple scaled GH (MSGH; Wraith and Forbes, 2015; Tortora et al., 2019).

While the differences between some of the reported AIC values are small and may not represent a significant difference, Table 5 shows that the best model is the MSAL, with the proposed MSCAL being ranked tenth. Generally speaking, the models allowing for skewness behave better, and this corroborates the results from the Mardia test given at the end of Section 7.2. Another interesting result is that there are only two multiple scaled distributions in the first ten positions (the MSAL and MSCAL). When compared to

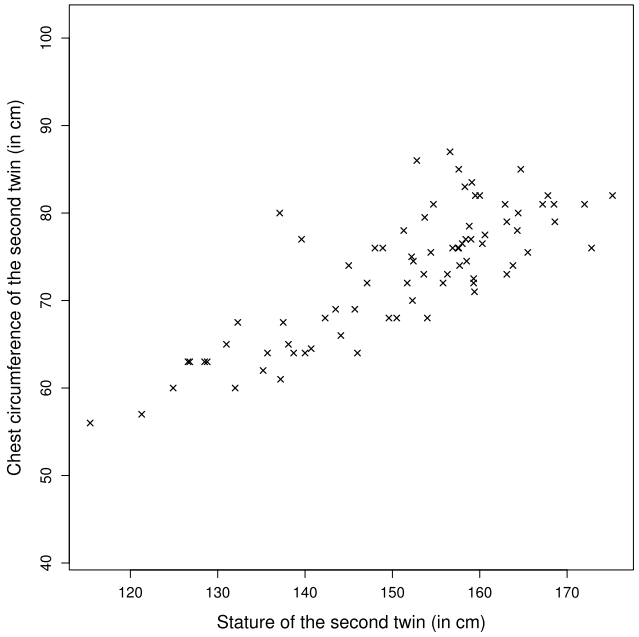


Fig. 2. Scatter plot of stature and chest circumference of the second twin from the `f.twins` data.

the other multiple scaled models, this implies that the peculiar peak of the Laplace distribution is useful when trying to model the joint distribution of the considered data set.

For the sake of coherence, we estimate the parameters of all the models by the ML approach. We use the `fit.ghypmv()` function of the **ghyp** package (Weibel et al., 2022) to fit all the members of the MGH-family. We adopt the `WML.MLN()` function from the code available at http://docenti.unict.it/punzo/Rcode.htm to fit the MLN, the `mtin.ML.ECME()` function of the **mtin** package (available at http://docenti.unict.it/punzo/Rpackages.htm) to fit the MTIN, the `msen.ML.EM()` function of the **msen** package (always available at http://docenti.unict.it/punzo/Rpackages.htm) to fit the MSEN, and the `CNmixt()` function of the **ContaminatedMixt** package (Punzo et al., 2018a,b) to fit the MCN. To fit the models proposed by Cabral et al. (2012), we use `smsn.mmix()` function of the **mixsmsn** package (Weibel et al., 2022). As for the multiple scaled models, we use the function `msdist.ML.EM()` available at http://docenti.unict.it/punzo/Rcode.htm to fit MSSEN model, the `mst()` and `mscn()` functions of the **MSclust** package to fit the MS*t* and MSCN (Tortora et al., 2023), and the `MSGHD()` function of the **MixGHD** package (Tortora et al., 2022, 2021) to fit the MSGH. The MSAL, MCAL, and MSCAL models are fitted using R code that we wrote for this project. The code to fit the proposed MSCAL distribution is available at https://github.com/cristinatortora/MSCAL.

In Table 5 we report the number of parameters and the maximum log-likelihood value for each model. Since the considered models have a different number of parameters, we compare their goodness-of-fit using the Akaike information criterion (AIC; Akaike, 1974) that, in our formulation, needs to be maximized and multiplied by -1 in accordance with the returned log-likelihood value. Under certain assumptions, the AIC has been shown to be appropriate for detecting the best approximating model (Punzo and Bagnato, 2021b). Therefore, we use the AIC because the true underlying model is unknown and it is highly unlikely that it is exactly one of the considered models.

The ML estimates for the MSAL model selected by the AIC are

$$\widehat{\boldsymbol{\mu}} = \begin{pmatrix} 161.609 \\ 76.811 \end{pmatrix}, \ \widehat{\boldsymbol{\Phi}} = \begin{pmatrix} 155.109 & 0.000 \\ 0.000 & 15.668 \end{pmatrix}, \ \widehat{\boldsymbol{\Gamma}} = \begin{pmatrix} -0.898 & -0.441 \\ -0.441 & 0.898 \end{pmatrix}, \text{ and } \widehat{\boldsymbol{\alpha}} = \begin{pmatrix} -10.620 \\ -3.667 \end{pmatrix}. \tag{26}$$

Notably, the returned estimates for $\boldsymbol{\alpha}$ in (26), show that skewness is negative in both dimensions (stature and chest circumference). Fig. 3 displays a scatterplot with isodensities from the MSAL superimposed.

The isodensities appear to be coherent with the scatter's shape. The MSAL model will be used as the benchmark (or "reference") to judge the results of the sensitivity analysis in Section 7.2.2.

Finally, it is important to note that, the ML-estimates $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{\Phi}}$, $\widehat{\boldsymbol{\Gamma}}$, and $\widehat{\boldsymbol{\alpha}}$ for the MSCAL model are practically the same as those given in (26); however, for this model, we also have the additional estimates

$$\widehat{\boldsymbol{\rho}} \approx \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \widehat{\boldsymbol{\eta}} \approx \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{27}$$

Notably, the estimates given in (27) indicate there are no outlying points or contamination. Therefore, we can conclude that the results in Table 5 suggest that a distribution with more flexible tail behavior in each principal component is needed for the considered data set.

Table 5
For the `f.twins` data, a comparison of thirty models in terms of number of parameters (#par), log-likelihood value (Log-Lik), and AIC. The ranking based on the AIC is given in the final column.

Model	# par.	Log-Lik.	AIC	Ranking
Normal	5	-542.982	-1095.964	12
Cauchy	5	-554.261	-1118.522	30
Laplace	5	-551.239	-1112.477	28
t	6	-542.981	-1097.962	18
Hyperbolic Univariate Marginals	6	-542.992	-1097.985	21
Symmetric Variance Gamma	6	-542.992	-1097.984	19
Symmetric Hyperbolic	6	-542.992	-1097.985	21
Symmetric Normal Inverse Gaussian	6	-542.992	-1097.985	21
Symmetric Generalized Hyperbolic	7	-542.992	-1099.984	25
Asymmetric Cauchy	7	-550.881	-1115.761	29
Asymmetric Laplace	7	-541.774	-1097.547	14
Asymmetric *t*	8	-538.755	-1093.510	9
Normal Inverse Gaussian	8	-538.387	-1092.773	8
Variance Gamma	8	-536.940	-1089.879	4
Hyperbolic	8	-538.263	-1092.526	7
Generalized Hyperbolic	9	-536.940	-1091.879	6
Tail-Inflated Normal	6	-542.976	-1097.952	16
Shifted-Exponential Normal	6	-542.976	-1097.952	16
Leptokurtic Normal	6	-542.976	-1097.952	16
Contaminated Normal	7	-542.976	-1099.952	24
Skew-normal	7	-536.615	-1087.230	2
Skew-*t*	8	-536.426	-1088.852	3
Contaminated asymmetric Laplace	9	-541.818	-1101.636	27
Skew Contaminated Normal	9	-536.573	-1091.146	5
Multiple scaled Shifted-Exponential Normal	7	-541.522	-1097.044	13
Multiple scaled *t*	7	-542.424	-1098.847	23
Multiple scaled Contaminated Normal	9	-541.415	-1100.829	26
Multiple scaled Asymmetric Laplace	7	-536.397	-1086.793	1
Multiple scaled Contaminated Asymmetric Laplace	11	-536.396	-1094.792	10
Multiple scaled Generalized Hyperbolic	11	-536.821	-1095.643	11

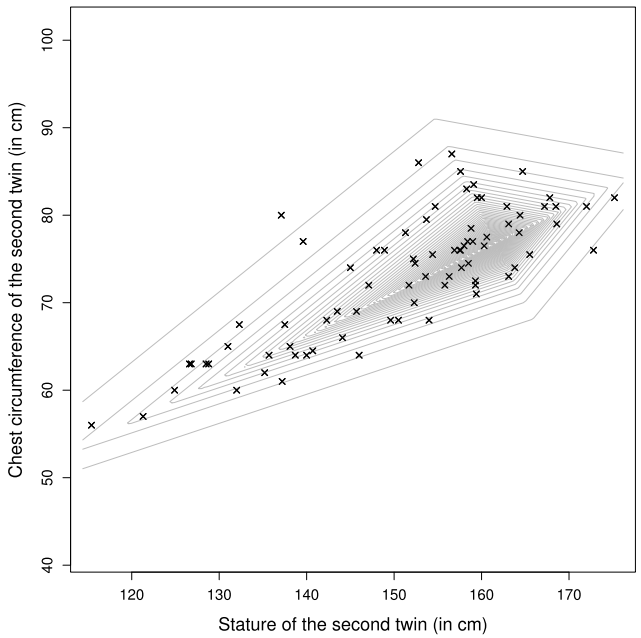


Fig. 3. Scatter plot of stature and chest circumference of the second twin from the `f.twins` data with superimposed isodensities from the AIC-selected MSAL model.

Table 6
Coordinates of the artificially added outliers for the sensitivity analysis.

Outliers	Stature (cm)	Chest circumference (cm)
P1	130	140
P2	120	160
P3	270	130
P4	200	180

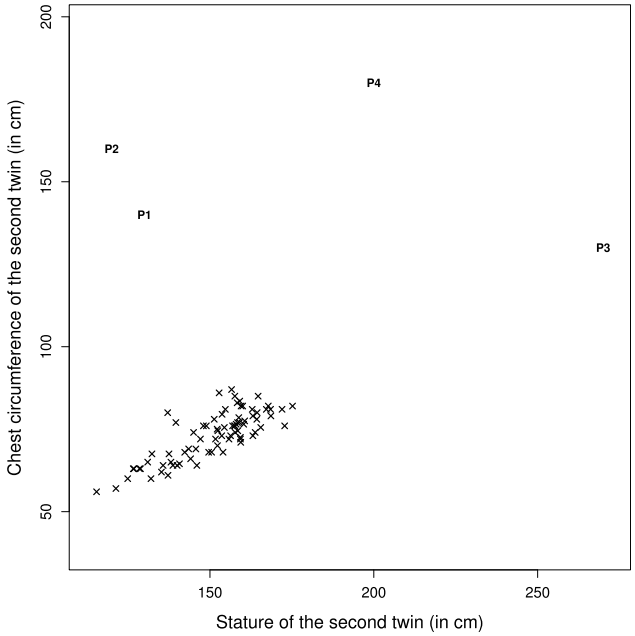


Fig. 4. Scatter plot of stature and chest circumference of the second twin, from the `f.twins` data, including artificially outlying points labeled with an initial "P".

Table 7
Comparison between MSAL and MSCAL distributions on each perturbed version of the `f.twins` data set. Considered scenarios are listed in the first column. The comparison is in terms of log-likelihood, AIC, and *p*-values from the LR test of MSALity.

	Log-Lik		AIC		LR test
Scenario	MSAL	MSCAL	MSAL	MSCAL	*p*-value
P1	-556.803	-550.742	-1127.606	-1123.484	0.195
P2	-560.889	-550.690	-1135.779	-1123.379	0.037
P1+P2	-576.865	-563.215	-1167.730	-1148.429	0.008
P3	-557.566	-550.574	-1129.133	-1123.149	0.136
P1+P2+P3	-618.967	-577.430	-1251.934	-1176.859	0.000
P4	-570.678	-559.237	-1155.355	-1140.475	0.022
P1+P2+P3+P4	-623.231	-597.059	-1260.461	-1216.118	0.000

7.2.2. Sensitivity analysis

The second aim of the first analysis is to investigate how the MSAL and MSCAL distributions react to the inclusion of *ad-hoc* located outliers. These points, whose coordinates are given in Table 6, are labeled with an initial "P" and are displayed in Fig. 4. The points are purposefully placed in such a way that either one or two of the PCs from the MSAL distribution will be impacted; refer to $\widehat{\mathbf{\Gamma}}$ in (26). Furthermore, it is important to note that using less extreme points would not show the benefits of the MSCAL model as clearly. In detail, P1 and P2 are roughly in the direction of the second PC, P3 is roughly in the direction of the first PC, while P4 is in between. By considering one or more of these points, we define seven "perturbed" versions of the original data.

For each perturbed data set, we fit the MSAL and MSCAL distributions. Table 7 presents the model comparison. The first column provides the considered scenarios; for example, the scenario identified by P1+P2+P3 refers to the perturbed data set which includes the points P1, P2, and P3. As in Section 7.2.1, we compare the models using the log-likelihood and AIC values. In addition, we also report the results from a likelihood-ratio (LR) test to compare the MSAL distribution (null model) with the MSCAL distribution

Table 8
ML-estimates for the parameters of the MSAL distribution when fitted to each perturbed version of the considered subset of variables from `f.twins` data set. The eigenvectors matrix $\mathbf{\Gamma}$ is expressed as a rotation matrix of angle θ.

Scenario	$\hat{\theta}$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\alpha}_1$	$\hat{\alpha}_2$
P1	0.423	162.412	74.096	154.453	12.864	-11.778	-0.170
P2	0.462	162.387	75.150	153.196	16.728	-11.702	-0.899
P1+P2	0.400	162.361	73.392	154.973	14.431	-12.021	1.647
P3	0.456	158.049	75.060	303.842	15.287	-5.544	-1.196
P1+P2+P3	0.329	167.003	71.160	288.538	0.004	-13.334	5.844
P4	0.467	161.072	74.660	264.555	16.128	-10.475	-0.701
P1+P2+P3+P4	0.414	159.203	71.961	349.142	15.734	-6.832	4.969

Table 9
ML-estimates for the parameters for the MSCAL distribution when fitted to each perturbed version of the considered subset of variables from `f.twins` data set. The eigenvectors matrix $\mathbf{\Gamma}$ is expressed as a rotation matrix of angle θ.

Scenario	$\hat{\theta}$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\eta}_1$	$\hat{\eta}_2$
P1	0.468	161.486	77.312	150.477	15.900	-10.467	-4.117	1.000	0.984	1.000	16.284
P2	0.459	161.393	76.789	153.738	15.630	-10.257	-3.549	1.000	0.985	1.000	21.660
P1+P2	0.454	161.425	76.601	150.780	15.277	-10.203	-3.293	1.000	0.969	1.000	18.222
P3	0.456	161.400	76.686	159.409	15.264	-10.108	-3.417	0.985	1.000	19.756	1.000
P1+P2+P3	0.456	161.382	76.681	152.418	15.093	-9.837	-3.230	0.985	0.969	20.008	18.710
P4	0.454	161.625	76.695	157.501	15.473	-10.311	-3.360	0.982	0.984	11.819	16.550
P1+P2+P3+P4	0.456	161.310	76.639	154.626	14.847	-9.605	-3.091	0.968	0.954	15.459	18.571

(alternative model). Hereafter, we will refer to this test as the "LR test of MSALity". The p-values from this test are reported in the last column of Table 7. Regardless of the considered scenario, the AIC always selects the MSCAL model. The p-values from the LR test provide additional insights. First, all of the p-values, apart from the ones on the P1 scenario (0.195) and P3 scenario (0.136), are lower than the commonly used 5% significance level, leading us to conclude that the MSCAL model is required for the corresponding data set; moreover, the p-values decrease as the contamination increases, where the term "contamination" describes the number of outliers and/or the distance of the added points from the bulk of the data. For example, if we consider the first two scenarios, labeled as P1 and P2, we note that, the further the added outlier in the direction of the second PC, the lower the p-value. In addition, when these points are considered together in the scenario P1+P2, the p-value is very low.

Tables 8 and 9 give the parameter estimates for the MSAL and MSCAL distributions, respectively, for each scenario. To simplify the interpretation and presentation of the estimates, in analogy with Greselin et al. (2011), Greselin and Punzo (2013), Bagnato et al. (2014), and Punzo et al. (2016), we see the eigenvectors matrix $\mathbf{\Gamma}$ as a rotation matrix of angle θ, with $\theta \in (-\pi/2, \pi/2)$, and we report the estimate of this parameter in the corresponding tables.

As a benchmark, the value of $\hat{\theta}$ related to $\hat{\mathbf{\Gamma}}$ in (26) is 0.457. Regardless of the considered scenario, the estimates of the common parameters $\hat{\theta}$, $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\lambda}_1$, $\hat{\lambda}_2$, $\hat{\alpha}_1$, and $\hat{\alpha}_2$, are less affected by the presence of the outliers for the MSCAL distribution. Moreover, the estimate for the additional parameters of the MSCAL distribution, i.e., the proportion of good points and the degree of contamination in each PC, are as expected when assessing the location of the outliers in Fig. 4. For example, in the first three scenarios, where the outlying points are only in the direction of the second PC, no contamination is detected in the first PC ($\hat{\rho}_1 \approx \hat{\eta}_1 \approx 1$); instead, on the second PC, the proportion of good points decreases as the number of added outliers increases and the degree of contamination increases in line with the magnitude of the outlier(s). In the fourth scenario involving P3, we observe a similar result, but in the direction of the first PC rather than the second. In all the remaining scenarios, the contamination parameters "activate" for both the PCs because there is at least one outlier on each of them. In other words, P1 and P2 activate $\hat{\rho}_2$ and $\hat{\eta}_2$ only, P3 has an impact on $\hat{\rho}_1$ and $\hat{\eta}_1$ only, while P4 has an impact on all of these contamination parameters. For completeness, Figs. 5 and 6 show the contour plots for each scenario from the fitted MSAL and MSCAL distributions, respectively.

We also superimpose the PC-axes from $\hat{\mathbf{\Gamma}}$. As we can see, the presence of outliers affects the fitting of the MSAL distribution if we compare the results with those in Fig. 3. However, the PC-wise tails of the MSCAL distribution can adapt to the presence of the outlying points, and the fit of the model is not compromised by their presence. This result is also observed by evaluating the stability of the PC-axes.

Table 10 gives the *a posteriori* probability that an observation is labelled as good (denoted as $\hat{v}_{ih}$, for $h = 1, 2$) in each PC for each of the considered scenarios. Regardless of the considered scenario, the probability P1 and P2 are labeled good in the second PC is effectively zero, the probability P3 is labeled as good in the first PC is effectively zero, while the probability P4 is labeled as good is effectively zero on both the PCs. It is interesting to note how these are the only points that are detected as PC-wise outliers; this means that we have a perfect (null) PC-wise false positive rate as well as a perfect (unitary) PC-wise true positive rate.

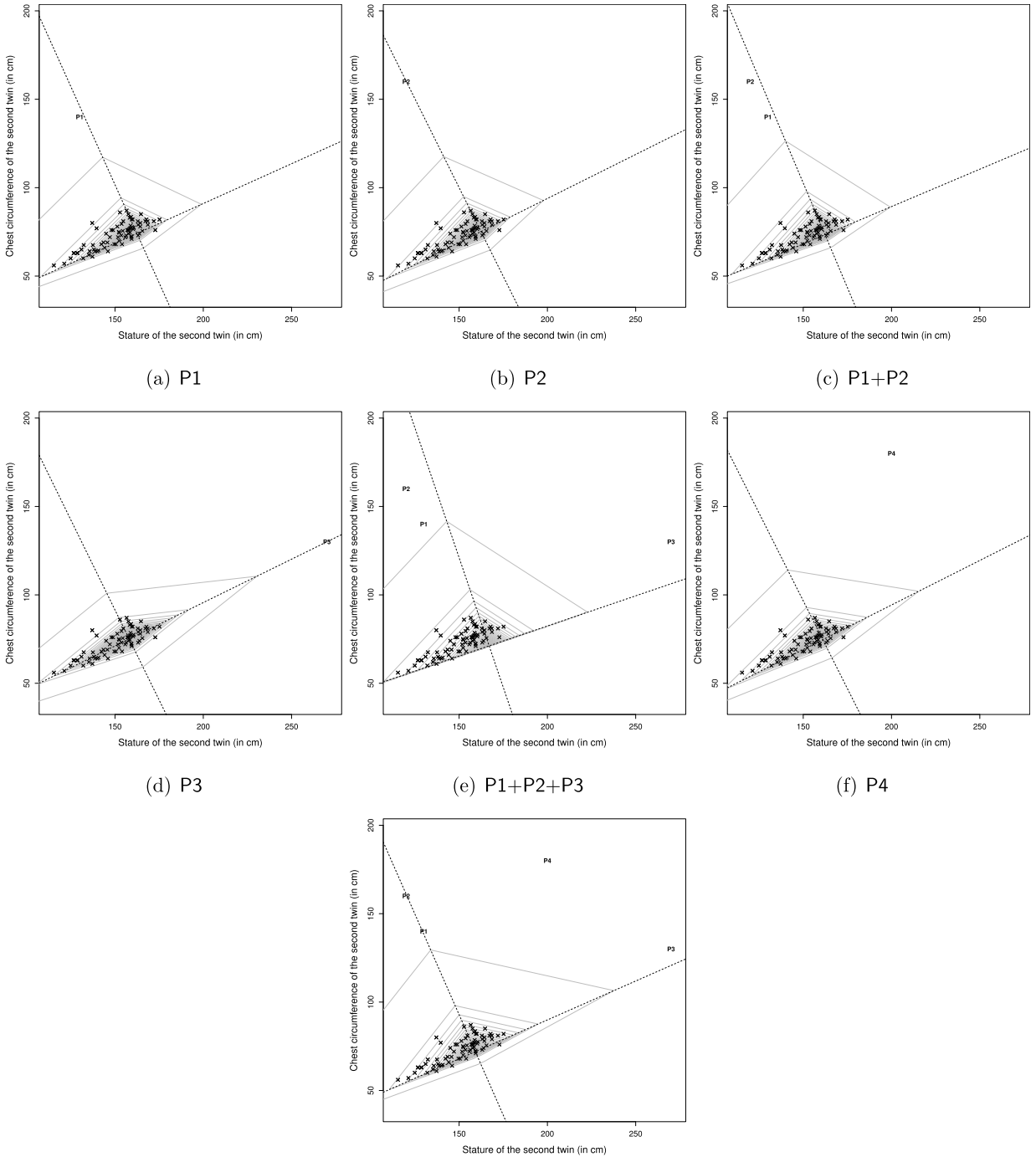


Fig. 5. Scatter plot of stature and chest circumference of the second twin from the `f.twins` data under the different scenarios of the sensitivity analysis. Contours and PC-axes from the fitted MSAL distribution are also superimposed.

8. Conclusions

This paper introduces a multiple scaled contaminated asymmetric Laplace (MSCAL) distribution for outlier detection. Compared to a multiple scaled asymmetric Laplace (MSAL) distribution, the levels of excess kurtosis on each variate are free to vary and it emits hyper-contours with less restrictive shapes. The MSCAL distribution is fitted using an MCEM algorithm that considers a two-

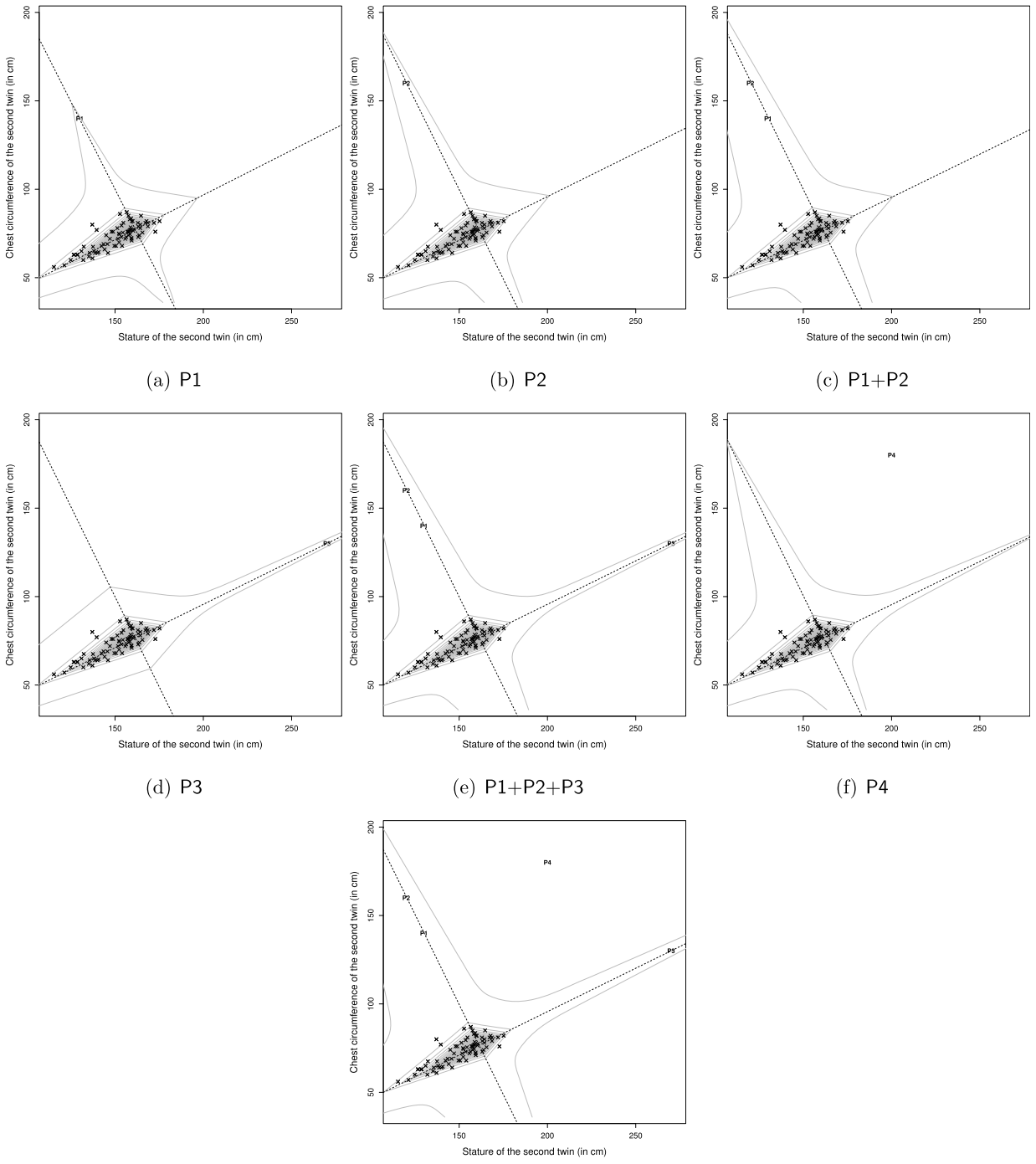


Fig. 6. Scatter plot of stature and chest circumference of the second twin from the `f.twins` data under the different scenarios of the sensitivity analysis. Contours and PC-axes from the fitted MSCAL distribution are also superimposed.

step Monte Carlo optimization procedure to estimate the matrix of eigenvectors, $\mathbf{\Gamma}$. A simulation study was used to evaluate the proposed MCEM algorithm in terms of parameter recovery and demonstrated the efficiency of this algorithm, in terms of elapsed run time. An analysis on the `f.twins` data showed how a multiple scaled asymmetric Laplace (MSAL) distribution can be a good fit for real data even when compared to other flexible distributions. However, a sensitivity analysis demonstrated the need for an MSCAL when outliers are included in the data.

Table 10
A posteriori probability to be good (denoted as $\widehat{v}_{ih}$, $j = 1, 2$) for each added point in each scenario, and separately for each PC.

Scenario	Point (i)	$\widehat{v}_{i1}$ (PC1)	$\widehat{v}_{i2}$ (PC2)
P1	P1	1.00000	0.00000
P2	P2	1.00000	0.00000
P1+P2	P1	1.00000	0.00001
	P2	1.00000	0.00000
P3	P3	0.00000	1.00000
P1+P2+P3	P1	0.99922	0.00000
	P2	0.99922	0.00000
	P3	0.00000	0.99828
P4	P4	0.00030	0.00000
P1+P2+P3+P4	P1	0.99785	0.00000
	P2	0.99783	0.00000
	P3	0.00000	0.99742
	P4	0.00020	0.00000

In terms of future work, the MSCAL can be extended to include a complete finite mixture modeling framework that can account for data composed of multiple sub-populations and PC-wise bad points simultaneously. Other variants of the proposed model that allow for combinations of either global or PC-wise outlier detection and either global or PC-wise parameterization of skewness can be considered. Moreover, alternative parameter estimation schemes for updating the matrix $\mathbf{\Gamma}$ should be explored. Finally, the proposed method could be extended to accommodate data sets with values missing at random (Tong and Tortora, 2022, 2023).

Acknowledgements

This work was supported by a grant from the National Science Foundation No. 2209974 (Tortora), by a discovery grant from the Natural Sciences and Engineering Research Council of Canada grant No. RGPIN-2017-04676 (Franczak), and by MUR, grant number 2022XRHT8R - *The SMILE project: Statistical Modelling and Inference to Live the Environment* (Punzo).

Appendix A. The generalized inverse Gaussian distribution

The generalized inverse Gaussian (GIG) distribution has been utilized and studied on several occasions (see Barndorff-Nielsen, 1977, 1978; Blæsild, 1978; Halgreen, 1979; Jørgensen, 1982 for examples). Let $Z \sim \text{GIG}(\chi, \omega, \nu)$ mean that Z follows a GIG distribution with parameters $\chi, \omega \in \mathbb{R}_+$ and $\nu \in \mathbb{R}$. The density of Z can be written as

$$f_{\text{GIG}}(z \mid \chi, \omega, \nu) = \frac{(\chi/\omega)^{\nu/2} z^{\nu-1}}{2K_\nu(\sqrt{\chi\omega})} \exp\left\{-\frac{\chi z + \omega/z}{2}\right\}, \tag{A.1}$$

for $z > 0$, where $K_\nu(\cdot)$ is the modified Bessel function of the third kind with index ν. From Jørgensen (1982), we utilize the expected values of Z and $1/Z$ in the E-step of the parameter estimation scheme proposed in Section 4. Formally, the expected values of interest can be written as

$$\mathbb{E}[Z] = \sqrt{\frac{\omega}{\chi}} R_\nu\left(\sqrt{\chi\omega}\right) \text{ and } \mathbb{E}\left[1/Z\right] = \sqrt{\frac{\chi}{\omega}} R_\nu\left(\sqrt{\chi\omega}\right) - \frac{2\nu}{\omega}, \tag{A.2}$$

respectively, where $R_\nu(\cdot) := K_{\nu+1}(\cdot)/K_\nu(\cdot)$ and all other terms as previously defined.

Appendix B. Simulation study results

Table B.11 gives the bias and variance of the estimates of $\boldsymbol{\Sigma}$ for the simulation study in Section 6.

Table B.11
For each scenario, the bias and variance of the estimates of Σ_{jh} for $j, h = 1, 2$.

		Bias			Variance		
Scenario	n	500	1000	2000	500	1000	2000
S1	Σ_{11}	0.38	0.29	0.33	0.86	0.55	0.52
	Σ_{22}	0.06	0.04	0.08	0.05	0.02	0.02
	Σ_{12}	-0.08	-0.06	-0.07	0.06	0.04	0.03
S2	Σ_{11}	0.49	0.24	0.15	0.93	0.68	0.30
	Σ_{22}	0.08	0.02	0.03	0.03	0.02	0.01
	Σ_{12}	-0.10	-0.06	-0.03	0.06	0.04	0.02
S3	Σ_{11}	-0.18	-0.14	-0.14	1.66	0.96	0.74
	Σ_{22}	0.06	0.02	0.06	0.05	0.02	0.02
	Σ_{12}	0.07	0.04	0.05	0.11	0.06	0.05
S4	Σ_{11}	0.13	0.08	-0.06	1.57	0.99	0.69
	Σ_{22}	0.03	0.03	0.03	0.04	0.02	0.01
	Σ_{12}	-0.02	-0.01	0.03	0.11	0.07	0.04
S5	Σ_{11}	0.12	0.27	0.20	1.40	0.71	0.38
	Σ_{22}	0.08	0.11	0.08	0.22	0.07	0.04
	Σ_{12}	-0.04	-0.04	-0.03	0.08	0.05	0.02
S6	Σ_{11}	0.38	0.24	0.16	0.99	0.64	0.33
	Σ_{22}	0.14	0.09	0.02	0.23	0.06	0.03
	Σ_{12}	-0.06	-0.04	-0.04	0.08	0.05	0.02
S7	Σ_{11}	-0.47	-0.06	-0.28	2.08	1.27	1.82
	Σ_{22}	-0.06	0.08	-0.02	0.20	0.08	0.26
	Σ_{12}	0.12	0.04	0.05	0.16	0.08	0.06
S8	Σ_{11}	-0.02	-0.30	-0.02	1.36	2.07	0.68
	Σ_{22}	0.13	0.02	-0.01	0.14	0.12	0.07
	Σ_{12}	0.04	0.09	0.00	0.10	0.13	0.05

Appendix C. Female twins data

Table C.12
Estimates of $\boldsymbol{\Sigma}$ obtained using the MSCAL algorithm on the female twins data.

	STA1	HIP1	CHE1	STA2	HIP2	CHE2
STA1	149.26	38.25	69.31	150.41	34.71	63.28
HIP1	38.25	13.85	22.19	39.00	12.13	19.34
CHE1	69.31	22.19	50.53	69.54	19.42	40.42
STA2	150.41	39.00	69.54	168.94	38.61	70.26
HIP2	34.71	12.13	19.42	38.61	13.22	21.61
CHE2	63.28	19.34	40.42	70.26	21.61	46.57

References

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Aitken, A., 1926. On Bernoulli's numerical solution of algebraic equations. Proc. R. Soc. Edinb. 46, 289–305.
Aitkin, M., Wilson, G.T., 1980. Mixture models, outliers, and the EM algorithm. Technometrics 22 (3), 325–331.
Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19 (6), 716–723.
Bagnato, L., Punzo, A., 2021. Unconstrained representation of orthogonal matrices with application to common principal components. Comput. Stat. 36 (2), 1177–1195.
Bagnato, L., Greselin, F., Punzo, A., 2014. On the spectral decomposition in normal discriminant analysis. Commun. Stat., Simul. Comput. 43 (6), 1471–1489.
Bagnato, L., Punzo, A., Zoia, M.G., 2017. The multivariate leptokurtic-normal distribution and its application in model-based clustering. Can. J. Stat. 45 (1), 95–119.
Bagnato, L., Farcomeni, A., Punzo, A., 2023. The generalized hyperbolic family and automatic model selection through the multiple-choice LASSO. Stat. Anal. Data Min., 1–13.
Barndorff-Nielsen, O., 1977. Exponentially decreasing distributions for the logarithm of particle size. Proc. R. Soc. Lond. Ser. A, Math. Phys. Sci. 353 (1674), 401–419.
Barndorff-Nielsen, O., 1978. Hyperbolic distributions and distributions on hyperbolae. Scand. J. Stat. 5 (3), 151–157.
Barndorff-Nielsen, O., Kent, J., Sørensen, M., 1982. Normal variance-mean mixtures and z distributions. Int. Stat. Rev. / Rev. Int. Stat. 50 (2), 145–159.
Blæsild, P., 1978. The shape of the generalized inverse Gaussian and hyperbolic distributions. Research Report 37. Department of Theoretical Statistics, Aarhus University, Denmark.
Booth, J.G., Hobert, J.P., 1999. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. J. R. Stat. Soc., Ser. B, Stat. Methodol. 61 (1), 265–285.
Browne, R.P., McNicholas, P.D., 2015. A mixture of generalized hyperbolic distributions. Can. J. Stat. 43 (2), 176–198.
Cabral, C.S.B., Lachos, V.H., Prates, M.O., 2012. Multivariate mixture modelling using skew-normal independent distributions. Comput. Stat. Data Anal. 56, 126–142.
Crawford, S.L., 1994. An application of the Laplace method to finite mixture distributions. J. Am. Stat. Assoc. 89 (425), 259–267.

Davies, L., Gather, U., 1993. The identification of multiple outliers. J. Am. Stat. Assoc. 88 (423), 782–792.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B 39 (1), 1–38.
Flury, B., 1997. Flury: data sets from flury. R package version 0.1-3 https://CRAN.R-project.org/package=Flury.
Flury, B., 2013. A First Course in Multivariate Statistics. Springer Texts in Statistics. Springer, New York. https://books.google.it/books?id=lGXTBwAAQBAJ.
Forbes, F., Wraith, D., 2014. A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights: application to robust clustering. Stat. Comput. 24 (6), 971–984.
Franczak, B.C., Browne, R.P., McNicholas, P.D., 2014. Mixtures of shifted asymmetric Laplace distributions. IEEE Trans. Pattern Anal. Mach. Intell. 36 (6), 1149–1157.
Franczak, B.C., Tortora, C., Browne, R.P., McNicholas, P.D., 2015. Unsupervised learning via mixtures of skewed distributions with hypercube contours. Pattern Recognit. Lett. 58, 69–76.
Frühwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer, New York.
Good, I.J., 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40, 237–260.
Greselin, F., Punzo, A., 2013. Closed likelihood ratio testing procedures to assess similarity of covariance matrices. Am. Stat. 67 (3), 117–128.
Greselin, F., Ingrassia, S., Punzo, A., 2011. Assessing the pattern of covariance matrices via an augmentation multiple testing procedure. Stat. Methods Appl. 20, 141–170.
Halgreen, C., 1979. Self-decomposibility of the generalized inverse Gaussian and hyperbolic distributions. Z. Wahrscheinlichkeitstheor. Verw. Geb. 47, 13–18.
Jørgensen, B., 1982. Statistical Properties of the Generalized Inverse Gaussian Distribution. Springer-Verlag, New York.
Korkmaz, S., Goksuluk, D., Zararsiz, G., 2019. MVN: multivariate normality tests. R package version 5.6 https://CRAN.R-project.org/package=MVN.
Kotz, S., Kozubowski, T.J., Podgorski, K., 2001. The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance, 1st edition. Birkhäuser, Boston.
Mazza, A., Punzo, A., 2019. Modeling household income with contaminated unimodal distributions. In: Petrucci, A., Racioppi, F., Verde, R. (Eds.), New Statistical Developments in Data Science. In: Springer Proceedings in Mathematics & Statistics, vol. 88. Springer, Cham, Switzerland, pp. 373–391.
McLachlan, G.J., Krishnan, T., 2007. The EM Algorithm and Extensions. John Wiley & Sons.
McLachlan, G.J., Peel, D., 1998. Robust Cluster Analysis via Mixtures of Multivariate T-Distributions. Lecture Notes in Computer Science, vol. 1451. Springer-Verlag, Berlin, pp. 658–666.
McNeil, A., Frey, R., Embrechts, P., 2005. Quantitative Risk Management: Concepts, Techniques and Tools. Princeton Series in Finance. Princeton University Press.
McNicholas, P.D., Murphy, T.B., McDaid, A.F., Frost, D., 2010. Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. Comput. Stat. Data Anal. 54 (3), 711–723.
Melnykov, V., 2013. Challenges in model-based clustering. Wiley Interdiscip. Rev.: Comput. Stat. 5 (2), 135–148.
Melnykov, Y., Zhu, X., Melnykov, V., 2021. Transformation mixture modeling for skewed data groups with heavy tails and scatter. Comput. Stat. 36, 61–78.
Morris, K., Punzo, A., Blostein, M., McNicholas, P.D., 2019. Asymmetric clusters and outliers: mixtures of multivariate contaminated shifted asymmetric Laplace distributions. Comput. Stat. Data Anal. 132, 145–166.
Punzo, A., Bagnato, L., 2020. Allometric analysis using the multivariate shifted exponential normal distribution. Biom. J. 62 (6), 1525–1543.
Punzo, A., Bagnato, L., 2021a. Modeling the cryptocurrency return distribution via Laplace scale mixtures. Phys. A, Stat. Mech. Appl. 563, 125–354.
Punzo, A., Bagnato, L., 2021b. The multivariate tail-inflated normal distribution and its application in finance. J. Stat. Comput. Simul. 91 (1), 1–36.
Punzo, A., Bagnato, L., 2022a. Multiple scaled symmetric distributions in allometric studies. Int. J. Biostat. 18 (1), 219–242.
Punzo, A., Bagnato, L., 2022b. Asymmetric Laplace scale mixtures for the distribution of cryptocurrency returns. arXiv.org e-print arXiv:2209.12848. available at http://arxiv.org/abs/2209.12848, 2022.
Punzo, A., McNicholas, P.D., 2016. Parsimonious mixtures of multivariate contaminated normal distributions. Biom. J. 58 (6), 1506–1537.
Punzo, A., Tortora, C., 2021. Multiple scaled contaminated normal distribution and its application in clustering. Stat. Model. 21 (4), 332–358.
Punzo, A., Browne, R.P., McNicholas, P.D., 2016. Hypothesis testing for mixture model selection. J. Stat. Comput. Simul. 86 (14), 2797–2818.
Punzo, A., Mazza, A., McNicholas, P.D., 2018a. ContaminatedMixt: model-based clustering and classification with the multivariate contaminated normal distribution. R package Version 1.3 (2018-01-29). https://CRAN.R-project.org/package=ContaminatedMixt.
Punzo, A., Mazza, A., McNicholas, P.D., 2018b. ContaminatedMixt: an R package for fitting parsimonious mixtures of multivariate contaminated normal distributions. J. Stat. Softw. 85 (10), 1–25.
Thompson, G.Z., Maitra, R., Meeker, W.Q., Bastawros, A.F., 2020. Classification with the matrix-variate-t distribution. J. Comput. Graph. Stat. 29 (3), 668–674.
Tomarchio, S.D., Punzo, A., 2020. Dichotomous unimodal compound models: application to the distribution of insurance losses. J. Appl. Stat. 47 (13–15), 2328–2353.
Tong, H., Tortora, C., 2022. Model-based clustering and outlier detection with missing data. Adv. Data Anal. Classif. 16 (1), 5–30.
Tong, H., Tortora, C., 2023. Missing values and directional outlier detection in model-based clustering. J. Classif. online, 1–34.
Tortora, C., Franczak, B.C., Browne, R.P., McNicholas, P.D., 2019. A mixture of coalesced generalized hyperbolic distributions. J. Classif. 36, 26–57.
Tortora, C., Browne, R.P., ElSherbiny, A., Franczak, B.C., McNicholas, P.D., 2021. Model-based clustering, classification, and discriminant analysis using the generalized hyperbolic distribution: MixGHD R package. J. Stat. Softw. 98 (3), 1–24. https://doi.org/10.18637/jss.v098.i03. https://www.jstatsoft.org/index.php/jss/article/view/v098i03.
Tortora, C., ElSherbiny, A., Browne, R.P., Franczak, B.C., McNicholas, P.D., Amos, D.D., 2022. MixGHD: model based clustering, classification and discriminant analysis using the mixture of generalized hyperbolic distributions. R package version 2.3.7. https://CRAN.R-project.org/package=MixGHD.
Tortora, C., Punzo, A., Tran, L., 2023. MSclust: multiple-scaled clustering. R package version 1.0.3. https://cran.r-project.org/web/packages/MSclust/index.html.
Tukey, J.W., 1960. A survey of sampling from contaminated distributions. In: Olkin, I. (Ed.), Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. In: Stanford Studies in Mathematics and Statistics. Stanford University Press, California, pp. 448–485. Ch. 39.
Wang, S., Yao, W., Huang, M., 2014. A note on the identifiability of nonparametric and semiparametric mixtures of GLMs. Stat. Probab. Lett. 93, 41–45.
Weibel, M., Luethi, D., Breymann, W., 2022. ghyp: generalized hyperbolic distribution and its special cases. R package version 1.6.3. https://CRAN.R-project.org/package=ghyp.
Wraith, D., Forbes, F., 2015. Location and scale mixtures of Gaussians with flexible tail behaviour: properties, inference and application to multivariate clustering. Comput. Stat. Data Anal. 90, 61–73.