Journal of Classification
https://doi.org/10.1007/s00357-026-09542-9

ORIGINAL RESEARCH

Skew Dimension-Wise Scaled Mixtures of Normal Distributions and Their Applications in Model-Based Clustering

Abbas Mahdavi[1] · Anthony F. Desmond[2] · Antonio Punzo[3] · Luca Bagnato[4]

Received: 6 August 2024 / Accepted: 10 February 2026

Abstract
The main goal of this paper is to address scenarios in which the distribution of multivariate real-valued data exhibits skewness and diverse tail behavior across dimensions. The dimension-wise scaled mixtures of normal (DSMN) distributions have been shown to be effective in modeling data with varying degrees of tail heaviness by dimension. An extension of the DSMN distribution is introduced by incorporating a vector of shape parameters, leading to the skew dimension-wise scaled mixtures of normal (SDSMN) distributions. The SDSMN family offers flexibility in expressing a range of shapes by allowing control over tailedness and skewness in each dimension. This study examines the characteristics and probabilistic properties of SDSMN distributions, as well as explores their extension to finite mixtures thereof. An ECME algorithm is developed utilizing a selection mechanism to compute the maximum likelihood estimates of model parameters. Numerical experiments conducted on simulated data and four real datasets from various fields, including biometry and biomedicine, demonstrate the effectiveness and practicality of the proposed methodology.

Keywords ECME algorithm · Heavy-tailed distributions · Dimension-wise scaled distributions · Truncated normal distribution

✉ Antonio Punzo
antonio.punzo@unict.it

Abbas Mahdavi
a.mahdavi@vru.ac.ir

Anthony F. Desmond
tdesmond@uoguelph.ca

Luca Bagnato
luca.bagnato@unicatt.it

1 Department of Statistics, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran

2 Department of Mathematics and Statistics, University of Guelph, N1G 2W1, Guelph 10587, Ontario, Canada

3 Dipartimento di Economia e Impresa, Università di Catania, Catania, Italy

4 Dipartimento di Scienze Economiche e Sociali, Università Cattolica del Sacro Cuore, Piacenza, Italy

Published online: 26 February 2026

1 Introduction

For real-valued p-dimensional data, the multivariate normal (MN) distribution, characterized by its mean vector $\boldsymbol{\xi}$ and covariance matrix $\boldsymbol{\Sigma}$, plays a crucial role. However, numerous empirical studies indicate that the MN distribution is often unsuitable due to two main deficiencies: (a) the tails are lighter than required; (b) the distribution is symmetric with a particular type of symmetry, the elliptical one.

A classical approach to addressing deficiency (a) is through the use of a multivariate scale mixture of normal (SMN) distributions (Andrews & Mallows, 1974). They are defined as a finite/infinite mixture of MN distributions, with the same mean $\boldsymbol{\xi}$ and a covariance $\tau^{-1}\boldsymbol{\Sigma}$ which is scaled by a convenient positive mixing random variable τ whose probability mass/density function depends on a tailedness parameter $\boldsymbol{\nu}$ governing the leptokurtosis of the data. Most of the existing SMN distribution literature focuses on the mixing distribution selection. However, the resulting SMN distributions are still elliptically symmetric and have a single tailedness parameter τ, valid for all the p dimensions, which may be rather restrictive because it implies dimensions having the same excess kurtosis (Forbes & Wraith, 2014).

Multiple scaled mixtures of normal (MSMN; Forbes & Wraith, 2014) distributions have become increasingly popular for modeling complex tail behaviors. MSMN distributions generalize SMN distributions by employing the eigendecomposition $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top$ of the scale matrix, where $\boldsymbol{\Gamma}$ contains the eigenvectors of $\boldsymbol{\Sigma}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix of the corresponding eigenvalues. Additionally, they introduce a distinct mixing random variable τ for each principal component (PC) in the space spanned by the columns of $\boldsymbol{\Gamma}$ (Punzo & Bagnato, 2022b). As discussed in Punzo and Bagnato (2022a), this parameterization allows for different tail behaviors across PCs. The MSMN distributions are defined as

$$f_{\text{MSMN}}(\boldsymbol{y};\boldsymbol{\xi},\boldsymbol{\Gamma},\boldsymbol{\Lambda},\boldsymbol{\nu}) = \int_{\mathbf{0}}^{\infty} \phi_p(\boldsymbol{y}|\boldsymbol{\tau};\boldsymbol{\xi},\boldsymbol{\Gamma}\boldsymbol{\Delta}_{\boldsymbol{\tau}}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top) h_{\boldsymbol{\tau}}(\boldsymbol{\tau};\boldsymbol{\nu}) d\boldsymbol{\tau},$$

for $\boldsymbol{y} \in \mathbb{R}^p$, where

$$\int_{\mathbf{0}}^{\infty} g(\boldsymbol{x})\, d\boldsymbol{x} = \int_0^{\infty} \cdots \int_0^{\infty} g\left(x_1, \ldots, x_p\right) dx_p \cdots dx_1,$$

$\phi_p(\boldsymbol{y};\boldsymbol{\xi},\boldsymbol{\Sigma})$ is the pdf of $N_p(\boldsymbol{\xi},\boldsymbol{\Sigma})$, $h_{\boldsymbol{\tau}}$ is a p-variate pdf of the scale vector $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)^\top$, with independent components, indexed by a possibly multivariate parameter vector $\boldsymbol{\nu} = (\boldsymbol{\nu}_1^\top, \ldots, \boldsymbol{\nu}_p^\top)^\top$, and $\boldsymbol{\Delta}_{\boldsymbol{\tau}} = \text{diag}\{\tau_1^{-1}, \ldots, \tau_p^{-1}\}$ is a $p \times p$ diagonal matrix whose diagonal elements are the inverse weights. The result is a multivariate heavy-tailed distribution with a more general type of symmetry than the elliptical one. The primary advantage of the MSMN distribution lies in its ability to capture variations in kurtosis across its PCs, thereby accommodating non-identical leptokurtosis. Additionally, a notable feature is the inclusion of multidimensional weight variables, τ_j for $j = 1, \ldots, p$, which can adjust for outliers with varying magnitudes across the PCs. In this model, the single-tailedness parameter of the SMN distributions is replaced by p tailedness parameters, one for each PC. This allows the model to address deficiency (a) and partially address deficiency (b), specifically the type of symmetry. However, two issues arise. First, the tailedness parameters are not defined in the original space of interest, complicating their interpretation. Second, the model's definition involves the orthogonal matrix $\boldsymbol{\Gamma}$, whose estimation is challenging (Bagnato & Punzo, 2021).

To overcome the above-mentioned issues about the MSNMs, Punzo and Bagnato (2022a) introduced a new class of distributions called dimension-wise scaled mixtures of normal

(DSMN) distributions, a more natural way to alter the tail behavior separately for each dimension. The DSMN distribution is a flexible approach that allows for different excess kurtosis on each dimension. This means that the excess kurtosis in each dimension can be different, allowing for greater flexibility in modeling complex data structures. The pdf of the DSMN distribution is given by

$$f_{\text{DSMN}}(\boldsymbol{y};\boldsymbol{\xi},\boldsymbol{\Sigma},\boldsymbol{\nu}) = \int_{\mathbf{0}}^{\infty} \phi_p(\boldsymbol{y}|\boldsymbol{\tau};\boldsymbol{\xi},\boldsymbol{\Delta}_{\boldsymbol{\tau}}^{1/2}\boldsymbol{\Sigma}\boldsymbol{\Delta}_{\boldsymbol{\tau}}^{1/2}) h_{\boldsymbol{\tau}}(\boldsymbol{\tau};\boldsymbol{\nu}) d\boldsymbol{\tau}, \tag{1}$$

for $\boldsymbol{y} \in \mathbb{R}^p$. By breaking down the distribution into its constituent components, it becomes possible to see how the generic jth dimension is affected by the scaling parameters $\boldsymbol{\nu}_j$ and by the distribution of the mixing random variable τ_j. By using the variance-correlation decomposition in Eq. 1, a hierarchical representation for a random vector $\boldsymbol{Y}$ that follows a DSMN distribution can be constructed as

$$\boldsymbol{Y} \overset{d}{=} \boldsymbol{\xi} + \boldsymbol{\sigma}^{1/2}\boldsymbol{\Delta}_{\boldsymbol{\tau}}^{1/2}\boldsymbol{Z},$$

where $\boldsymbol{Z} \sim N_p(\mathbf{0}, \bar{\boldsymbol{\Sigma}})$, with $\bar{\boldsymbol{\Sigma}}$ being a correlation matrix. Furthermore, $\boldsymbol{\sigma}^{1/2}\bar{\boldsymbol{\Sigma}}\boldsymbol{\sigma}^{1/2}$ is the variance-correlation decomposition of $\boldsymbol{\Sigma}$ and $\boldsymbol{\sigma} = \text{diag}(\sigma_{11}, \ldots, \sigma_{pp})$ where σ_{ij} denotes the element at position (i, j) in $\boldsymbol{\Sigma}$.

However, the methods discussed so far do not handle empirical distributions that exhibit asymmetry, which is a part of deficiency (b). To overcome this limitation, a skewness parameter vector can be incorporated into the parametric model, allowing control over the degree of asymmetry. In recent decades, extensive research on skew distributions has led to a wide range of models to describe asymmetric empirical distributions. Different construction approaches yield various forms of skewness, leading to distinct distributional shapes. Several methodologies have been developed, including perturbation (Azzalini & Capitanio, 2003 and Azzalini & Dalla Valle, 1996), copula-based methods (Jajuga & Papla, 2005; Vrac et al., 2012, and Kosmidis & Karlis, 2016), and transformation-based techniques (Ley & Paindaveine, 2010; Zhu & Melnykov, 2018; Scrucca, 2019, and Melnykov et al., 2021), each driven by different theoretical and practical considerations. In this work, we focus on the perturbation approach and, as an illustration, we consider the multivariate skew-normal (MSN) distribution (Azzalini & Dalla Valle, 1996). In detail, a p-dimensional random vector $\boldsymbol{Y}$ follows a MSN distribution with location parameter $\boldsymbol{\xi} \in \mathbb{R}^p$, positive definite scale matrix $\boldsymbol{\Sigma}$, and skewness parameter $\boldsymbol{\lambda} \in \mathbb{R}^p$, if its pdf is given by

$$f(\boldsymbol{y};\boldsymbol{\xi},\boldsymbol{\Sigma},\boldsymbol{\lambda}) = 2\phi_p(\boldsymbol{y};\boldsymbol{\xi},\boldsymbol{\Sigma})\Phi\big(\boldsymbol{\lambda}^\top\boldsymbol{\sigma}^{-1/2}(\boldsymbol{y}-\boldsymbol{\xi})\big), \tag{2}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (cdf) of a standard normal random variable.

In addition to the deficiencies outlined earlier, real data often exhibit heterogeneity, making a single distribution insufficient to adequately represent their complexity. In such cases, model-based clustering, which assumes an underlying finite mixture for the data distribution, is a powerful approach (McLachlan & Peel, 2000). Researchers and practitioners continuously seek flexible distributions to use as mixture components to better describe each cluster, operating under the typical assumption that each mixture component represents a cluster (cf. McLachlan & Basford, 1988; Fraley & Raftery, 1998, and Böhning, 2000). For a comprehensive review of finite mixtures with flexible distributions, see Dávila et al. (2018), McNicholas (2016), and Lee and McLachlan (2013). In line with this research trajectory, we propose finite mixtures of skew DSMN (SDSMN) distributions as a powerful tool for clustering.

For estimating the parameters of the proposed models, we utilize the maximum likelihood (ML) approach with the expectation conditional maximization either (ECME) algorithm (Liu & Rubin, 1994), an extension of the EM algorithm (Dempster et al., 1977). The ECME algorithm has a faster convergence rate than the EM algorithm while retaining its appealing features. The EM algorithm and its extensions, such as the expectation conditional maximization (ECM) algorithm (Meng & Rubin, 1993) and the ECME algorithm, are popular due to their ease of implementation and numerical stability when estimating various multivariate skew distributions.

Previous studies by Lin et al. (2007a, b), Lin (2009), and Lin (2010) provided analytical ECM and ECME procedures for ML estimation of finite mixtures of skew-normal and skew-t distributions in both univariate and multivariate cases using the convolution-type representation of the models. However, the SDSMN model cannot be formulated in this way. Instead, we adopt the selection-type representation of the SDSMN model, inspired by recent work by Mahdavi et al. (2021a, b), to develop a feasible ECME procedure for computing ML estimates of the model parameters.

The paper is structured as follows. Section 2 introduces the general SDSMN model and describes the application of the ECME algorithm for ML estimation using a selection-type mechanism. Section 3 presents two examples of SDSMN distributions, applying different mixing distributions to the scale vector $\boldsymbol{\tau}$. Following that, a finite mixture of SDSMN distributions is defined, and the EM algorithm is implemented to estimate the model parameters. The effectiveness of the proposed techniques is demonstrated through three simulation studies in Sect. A of the Supplementary Material and the analysis of four real datasets in Sect. 5 and Sect. B of the Supplementary Material. The paper concludes with final remarks and suggestions for future research directions in Sect. 6.

2 Methodology

This study aims to extend existing DSMN distributions by incorporating skewness, thereby introducing a more diverse range of distributional forms. This approach allows for distinct tail and skewness behaviors in each dimension. By combining the advantages of skewness, as detailed in Eq. 2, with variable marginal amounts of tail weight, as described in Eq. 1, our proposed framework—referred to as skew dimension-wise scaled mixtures of normal (SDSMN) distributions—offers a broader range of distributional forms to better capture the complexity of real-world data. This section explores the key characteristics of the SDSMNs that are beneficial for inferential purposes, particularly in ML parameter estimation using the ECME algorithm.

2.1 The SDSMN Distribution

A p-dimensional random vector $\boldsymbol{Y}$ follows the SDSMN distribution with location parameter $\boldsymbol{\xi} \in \mathbb{R}^p$, positive definite scale matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, skewness parameter $\boldsymbol{\lambda}$, and flatness parameter $\boldsymbol{\nu}$, if its pdf is given by

$$f_{\text{SDSMN}}(\boldsymbol{y};\boldsymbol{\xi},\boldsymbol{\Sigma},\boldsymbol{\lambda},\boldsymbol{\nu}) = 2 f_{\text{DSMN}}(\boldsymbol{y};\boldsymbol{\xi},\boldsymbol{\Sigma},\boldsymbol{\nu})\Phi\big(\boldsymbol{\lambda}^\top \boldsymbol{\sigma}^{-1/2}(\boldsymbol{y}-\boldsymbol{\xi})\big). \tag{3}$$

The SDSMN family of distributions defined in Eq. 3 is quite vast and includes various subfamilies of asymmetric and dimension-wise scaled distributions. These subfamilies have not been previously explored in the literature. It is worth mentioning that the SDSMN distribution

includes the DSMN distribution as a special case when $\boldsymbol{\lambda} = \mathbf{0}_p$ where $\mathbf{0}_p$ denotes the column vector of length p with all entries equal to 0. The SDSMN distribution can be cast in the framework of selection distributions that arise under a selection mechanism (Arellano-Valle et al., 2006).

Lemma 1 *A random vector $\boldsymbol{Y}$ follows the MSN distribution, with the pdf given in Eq. 2, if it has the following convolution and selection representation, respectively:*

$$\boldsymbol{Y} \overset{d}{=} \boldsymbol{\xi} + \boldsymbol{\sigma}^{1/2}\Big(\boldsymbol{\delta} H + (\bar{\boldsymbol{\Sigma}} - \boldsymbol{\delta}\boldsymbol{\delta}^\top)^{1/2}\boldsymbol{U}_0\Big),$$
$$\boldsymbol{Y} \overset{d}{=} \boldsymbol{\xi} + \boldsymbol{\sigma}^{1/2}\Big[\boldsymbol{Z} \mid \big(Z_0 < \boldsymbol{\lambda}^\top \boldsymbol{U}_0\big)\Big],$$

where $\boldsymbol{Z} \sim N_p(\mathbf{0}, \bar{\boldsymbol{\Sigma}})$, $Z_0 \sim N(0,1)$, $\boldsymbol{U}_0 \sim N_p(\mathbf{0}, \boldsymbol{I}_p)$, $H \sim DTN(0,1)I_{(0,\infty)}$ are independent random variables and $\boldsymbol{\delta} = \bar{\boldsymbol{\Sigma}}\boldsymbol{\lambda}/\sqrt{1+\boldsymbol{\lambda}^\top\bar{\boldsymbol{\Sigma}}\boldsymbol{\lambda}}$. Here, $DTN(\mu,\sigma^2)I_{\mathbb{A}}$ represents a doubly truncated normal distribution with support $\mathbb{A} = \{a_1 < x < a_2\}$, and $I_{\mathbb{A}}$ is an indicator function of the set $\mathbb{A}$.

Lemma 1 plays a crucial role in uncovering the essential characteristics of MSN distributions, including aspects such as random number generation, the computation of moments, and the estimation of parameters (Azzalini & Capitanio, 2014). In a similar way, it is straightforward to demonstrate that a p-variate random vector $\boldsymbol{Y} \sim SDSMN_p(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ follows the SDSMN distribution if it has the following convolution and selection-type representation, respectively:

$$\boldsymbol{Y} \overset{d}{=} \boldsymbol{\xi} + \boldsymbol{\sigma}^{1/2}\Big(\boldsymbol{\Delta}_{\boldsymbol{\tau}}^{1/2}\boldsymbol{\delta}_{\boldsymbol{\tau}} H + \boldsymbol{\Delta}_{\boldsymbol{\tau}}^{1/2}(\bar{\boldsymbol{\Sigma}} - \boldsymbol{\delta}_{\boldsymbol{\tau}}\boldsymbol{\delta}_{\boldsymbol{\tau}}^\top)^{1/2}\boldsymbol{U}_0\Big), \tag{4}$$

$$\boldsymbol{Y} \overset{d}{=} \boldsymbol{V} \mid (U > 0), \tag{5}$$

where $\boldsymbol{\delta}_{\boldsymbol{\tau}} = \bar{\boldsymbol{\Sigma}}\boldsymbol{\Delta}_{\boldsymbol{\tau}}^{1/2}\boldsymbol{\lambda}/\sqrt{1+\boldsymbol{\lambda}^\top\boldsymbol{\Delta}_{\boldsymbol{\tau}}^{1/2}\bar{\boldsymbol{\Sigma}}\boldsymbol{\Delta}_{\boldsymbol{\tau}}^{1/2}\boldsymbol{\lambda}}$, $\boldsymbol{V} = \boldsymbol{\xi} + \boldsymbol{\sigma}^{1/2}\boldsymbol{\Delta}_{\boldsymbol{\tau}}^{1/2}\boldsymbol{Z}$ and $U = \boldsymbol{\lambda}^\top\boldsymbol{\Delta}_{\boldsymbol{\tau}}^{1/2}\boldsymbol{Z} - Z_0$. Obviously, the random variable U is a continuous random variable symmetric about zero. From Eq. 4, it is easy to show that

$$E(\boldsymbol{Y}) = \boldsymbol{\xi} + \sqrt{\frac{2}{\pi}}\boldsymbol{\sigma}^{1/2}E\big(\boldsymbol{\Delta}_{\boldsymbol{\tau}}^{1/2}\boldsymbol{\delta}_{\boldsymbol{\tau}}\big).$$

Consider p independent weights $\tau_1, \ldots, \tau_p$, each having an associated pdf $h_{\tau_j}(\tau_j; \boldsymbol{\nu}_j)$, where $\boldsymbol{\nu}_j$ is a parameter vector that characterizes the shape of τ_j. For fitting the SDSMN model within the complete-data framework via the EM algorithm, we introduce the latent variables $W \overset{d}{=} U \mid (U > 0)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top \overset{d}{=} \boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)^\top \mid (U > 0)$ based on Eq. 5. Then, $(\boldsymbol{Y}^\top, W, \boldsymbol{\gamma}^\top)^\top \overset{d}{=} (\boldsymbol{V}^\top, U, \boldsymbol{\tau}^\top)^\top | (U > 0)$ has the following joint pdf:

$$\begin{aligned} f_{\boldsymbol{Y},W,\boldsymbol{\gamma}}(\boldsymbol{y}, w, \boldsymbol{\gamma}) &= \frac{1}{\Pr(U>0)} f_{\boldsymbol{V},U,\boldsymbol{\tau}}(\boldsymbol{y}, w, \boldsymbol{\gamma}) \\ &= 2 f_{\boldsymbol{V}|\boldsymbol{\tau}}(\boldsymbol{y}) f_{U|\boldsymbol{V},\boldsymbol{\tau}}(w) f_{\boldsymbol{\tau}}(\boldsymbol{\gamma}) \\ &= 2\phi_p\big(\boldsymbol{y}; \boldsymbol{\xi}, \boldsymbol{\Delta}_{\boldsymbol{\gamma}}^{1/2}\boldsymbol{\Sigma}\boldsymbol{\Delta}_{\boldsymbol{\gamma}}^{1/2}\big)\phi\big(w - \boldsymbol{\alpha}^\top(\boldsymbol{y} - \boldsymbol{\xi})\big)\prod_{j=1}^{p} h_{\tau_j}(\gamma_j; \boldsymbol{\nu}_j), \end{aligned} \tag{6}$$

where $\boldsymbol{\Delta}_{\boldsymbol{\gamma}} = \mathrm{diag}(\gamma_1^{-1}, \ldots, \gamma_p^{-1})$, $\boldsymbol{\alpha}^\top = \boldsymbol{\lambda}^\top\boldsymbol{\sigma}^{-1/2}$ is a redefined parameter, and $h_{\tau_j}(\gamma_j; \boldsymbol{\nu}_j)$ is the pdf of τ_j evaluated at point γ_j.

Integrating out W from Eq. 6 gives the following joint pdf

$$f_{\boldsymbol{Y},\boldsymbol{\gamma}}(\boldsymbol{y},\boldsymbol{\gamma}) = 2\phi_p\big(\boldsymbol{y};\boldsymbol{\xi},\boldsymbol{\Delta}_{\boldsymbol{\gamma}}^{1/2}\boldsymbol{\Sigma}\boldsymbol{\Delta}_{\boldsymbol{\gamma}}^{1/2}\big)\Phi\big(\boldsymbol{\alpha}^\top(\boldsymbol{y}-\boldsymbol{\xi})\big)\prod_{j=1}^{p} h_{\tau_j}(\gamma_j;\boldsymbol{\nu}_j). \tag{7}$$

Dividing (6) by Eq. 7 yields the following relation

$$f(w \mid \boldsymbol{y},\boldsymbol{\gamma}) = \frac{\phi\big(w-\boldsymbol{\alpha}^\top(\boldsymbol{y}-\boldsymbol{\xi})\big)}{\Phi\big(\boldsymbol{\alpha}^\top(\boldsymbol{y}-\boldsymbol{\xi})\big)} \equiv f(w \mid \boldsymbol{y}),$$

where '$\equiv$' means that W and $\boldsymbol{\gamma}$ are conditionally independent. Moreover, it is straightforward to show that

$$W \mid \boldsymbol{y} \sim DTN\big(\boldsymbol{\alpha}^\top(\boldsymbol{y}-\boldsymbol{\xi}),1\big)I_{(0,\infty)}.$$

Using Lemma 2 of Lin et al. (2007b), we have the following conditional expectation:

$$E(W \mid \boldsymbol{y}) = \boldsymbol{\alpha}^\top(\boldsymbol{y}-\boldsymbol{\xi}) + \frac{\phi\big(\boldsymbol{\alpha}^\top(\boldsymbol{y}-\boldsymbol{\xi})\big)}{\Phi\big(\boldsymbol{\alpha}^\top(\boldsymbol{y}-\boldsymbol{\xi})\big)}. \tag{8}$$

By Bayes' rule, the conditional pdf of $\boldsymbol{\gamma}$ given $\boldsymbol{Y}=\boldsymbol{y}$ is

$$f(\boldsymbol{\gamma} \mid \boldsymbol{y}) = \frac{f(\boldsymbol{\gamma},\boldsymbol{y})}{f(\boldsymbol{y})} \propto \phi_p\big(\boldsymbol{y};\boldsymbol{\xi},\boldsymbol{\Delta}_{\boldsymbol{\gamma}}^{1/2}\boldsymbol{\Sigma}\boldsymbol{\Delta}_{\boldsymbol{\gamma}}^{1/2}\big)\prod_{j=1}^{p} h_{\tau_j}(\gamma_j;\boldsymbol{\nu}_j). \tag{9}$$

2.2 Parameter Estimation via the ECME Algorithm

Suppose that $\boldsymbol{y}_o = (\boldsymbol{y}_1^\top,\ldots,\boldsymbol{y}_n^\top)^\top$ constitutes a sample of size n arising from the SDSMN model. Under the EM framework, the latent variables $\boldsymbol{W} = (W_1,\ldots,W_n)^\top$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top,\ldots,\boldsymbol{\gamma}_n^\top)^\top$, introduced in Sect. 2.1, are treated as missing data. Then, the complete data is given by $\boldsymbol{y}_c = (\boldsymbol{y}_o^\top,\boldsymbol{W}^\top,\boldsymbol{\gamma}^\top)^\top$. Further, we let $\boldsymbol{\theta} = (\boldsymbol{\xi}^\top,\text{vech}(\boldsymbol{\Sigma})^\top,\boldsymbol{\lambda}^\top,\boldsymbol{\nu}^\top)^\top$ to denote the entire set of unknown parameters to be estimated in the SDSMN model, where vech$(\cdot)$ is the half-vectorization operator that stacks the lower triangular elements of a $p\times p$ symmetric matrix into a single $p(p+1)/2$ vector.

According to Eq. 6, the log-likelihood function of $\boldsymbol{\theta}$ corresponding to the complete-data $\boldsymbol{y}_c$, excluding additive constants and terms that do not involve parameters of the model, is given by

$$\begin{aligned}\ell_c(\boldsymbol{\theta} \mid \boldsymbol{y}_c) = -\frac{1}{2}\sum_{i=1}^{n}\Bigg\{ &\log\big(\det(\boldsymbol{\Sigma})\big) + (\boldsymbol{y}_i-\boldsymbol{\xi})^\top\boldsymbol{\Delta}_{\boldsymbol{\gamma}_i}^{-1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\gamma}_i}^{-1/2}(\boldsymbol{y}_i-\boldsymbol{\xi}) \\ &+\big(W_i-\boldsymbol{\alpha}^\top(\boldsymbol{y}_i-\boldsymbol{\xi})\big)^2 - 2\sum_{j=1}^{p}\log h_{\tau_j}(\gamma_{ij};\boldsymbol{\nu}_j)\Bigg\},\end{aligned} \tag{10}$$

where γ_{ij} denotes the jth component of $\boldsymbol{\gamma}_i$.

On the kth iteration, the E-step requires the calculation of the so-called Q-function, which is the conditional expectation of Eq. 10 given the observed data $\boldsymbol{y}$ and the current estimate

$\boldsymbol{\theta}^{(k)}$, where the superscript $^{(k)}$ denote the updated estimates at iteration k. To evaluate the Q-function, we require the conditional expectations

$$\boldsymbol{\gamma}_i^{(k)} = E(\boldsymbol{\gamma}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}) \quad \text{and} \quad w_i^{(k)} = E(W_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}), \tag{11}$$

which can be derived from Eqs. 8 and 9, along with

$$s_{ij}^{(k)}(\boldsymbol{\nu}_j) = E(\log h_{\tau_j}(\gamma_{ij}; \boldsymbol{\nu}_j) \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}), \tag{12}$$

which may not have a closed form for some subfamilies. Substituting (11) and (12) into (10) yields the following Q-function:

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = -\frac{1}{2}\sum_{i=1}^{n}\Bigg\{ & \log\big(\det(\boldsymbol{\Sigma})\big) + (\mathbf{y}_i - \boldsymbol{\xi})^\top \boldsymbol{\Delta}_{\boldsymbol{\gamma}_i^{(k)}}^{-1/2} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta}_{\boldsymbol{\gamma}_i^{(k)}}^{-1/2} (\mathbf{y}_i - \boldsymbol{\xi}) \\ & + [\boldsymbol{\alpha}^\top(\mathbf{y}_i - \boldsymbol{\xi})]^2 - 2w_i^{(k)} \boldsymbol{\alpha}^\top(\mathbf{y}_i - \boldsymbol{\xi}) - 2\sum_{j=1}^{p} s_{ij}^{(k)}(\boldsymbol{\nu}_j)\Bigg\}, \end{aligned} \tag{13}$$

where $\boldsymbol{\Delta}_{\boldsymbol{\gamma}_i^{(k)}} = \mathrm{diag}(1/\gamma_{i1}^{(k)}, \ldots, 1/\gamma_{ip}^{(k)})$.

The CM-steps are implemented to update estimates of $\boldsymbol{\theta}$ in the order of $\boldsymbol{\xi}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ by maximizing, one by one, the Q-function obtained in the E-step. After performing some algebraic manipulations, these procedures can be succinctly summarized by the following CMQ and CML steps. Here, the abbreviations CMQ and CML denote the function to be maximized in the CM step, specifically the Q-function for the CMQ steps, and the observed-data log-likelihood for the CML step.

CMQ-step 1 Fix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(k)}$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(k)}$, and update $\boldsymbol{\xi}^{(k)}$ by maximizing (13) with respect to $\boldsymbol{\xi}$, which leads to

$$\begin{aligned} \boldsymbol{\xi}^{(k+1)} = & \left[\sum_{i=1}^{n}\left(\boldsymbol{\Delta}_{\boldsymbol{\gamma}_i^{(k)}}^{-1/2} \boldsymbol{\Sigma}^{(k)-1} \boldsymbol{\Delta}_{\boldsymbol{\gamma}_i^{(k)}}^{-1/2} + \boldsymbol{\alpha}^{(k)} \boldsymbol{\alpha}^{\top(k)}\right)\right]^{-1} \left[\sum_{i=1}^{n} \boldsymbol{\Delta}_{\boldsymbol{\gamma}_i^{(k)}}^{-1/2} \boldsymbol{\Sigma}^{(k)-1} \boldsymbol{\Delta}_{\boldsymbol{\gamma}_i^{(k)}}^{-1/2} \mathbf{y}_i \right. \\ & \left. + \boldsymbol{\alpha}^{(k)} \boldsymbol{\alpha}^{\top(k)} \sum_{i=1}^{n} \mathbf{y}_i - \boldsymbol{\alpha}^{(k)} \sum_{i=1}^{n} w_i{}^{(k)}\right]. \end{aligned}$$

CMQ-step 2 Fix $\boldsymbol{\xi} = \boldsymbol{\xi}^{(k+1)}$ and update $\boldsymbol{\Sigma}^{(k)}$ by maximizing (13) over $\boldsymbol{\Sigma}$. This gives

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Delta}_{\boldsymbol{\gamma}_i^{(k)}}^{-1/2} (\mathbf{y}_i - \boldsymbol{\xi}^{(k+1)})(\mathbf{y}_i - \boldsymbol{\xi}^{(k+1)})^\top \boldsymbol{\Delta}_{\boldsymbol{\gamma}_i^{(k)}}^{-1/2}.$$

CMQ-step 3 Fix $\boldsymbol{\xi} = \boldsymbol{\xi}^{(k+1)}$ and obtain $\boldsymbol{\alpha}^{(k+1)}$ by maximizing (13) over $\boldsymbol{\alpha}$. This gives

$$\boldsymbol{\alpha}^{(k+1)} = \left[\sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\xi}^{(k+1)})(\mathbf{y}_i - \boldsymbol{\xi}^{(k+1)})^\top\right]^{-1} \sum_{i=1}^{n} w_i^{(k)} (\mathbf{y}_i - \boldsymbol{\xi}^{(k+1)}).$$

Accordingly, we can obtain $\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\sigma}^{1/2(k+1)} \boldsymbol{\alpha}^{(k+1)}$.

For some members of SDSMN, the calculation of $s_{ij}^{(k)}(\boldsymbol{\nu}_j)$ in Eq. 12 is not straightforward. An update of $\boldsymbol{\nu}^{(k)} = (\boldsymbol{\nu}_1^{\top(k)}, \ldots, \boldsymbol{\nu}_p^{(k)})^\top$ can be achieved by directly maximizing the constrained actual log-likelihood function. This gives rise to the following CML step.

CML step Update $\boldsymbol{\nu}^{(k)}$ by maximizing the constrained log-likelihood function

$$\boldsymbol{\nu}^{(k+1)} = \arg\max_{\boldsymbol{\nu}} \sum_{i=1}^{n} \log f_{\text{SDSMN}}(\boldsymbol{y}_i; \boldsymbol{\xi}^{(k+1)}, \boldsymbol{\Sigma}^{(k+1)}, \boldsymbol{\lambda}^{(k+1)}, \boldsymbol{\nu}).$$

3 Examples of SDSMN Distributions

3.1 The Skew Dimension-Wise Scaled Tail-Inflated Normal Distribution

Punzo and Bagnato (2022a) introduced the dimension-wise scaled tail-inflated normal distribution (DSTIN) by setting τ_j in Eq. 1 as a uniform distribution on $(1-\nu_j, 1)$, with $\nu_j \in (0, 1)$. In this case, $h_{\boldsymbol{\tau}}(\boldsymbol{\tau}; \boldsymbol{\nu}) = \prod_{j=1}^{p} \nu_j^{-1}$. The pdf of the DSTIN distribution is given by

$$f_{\text{DSTIN}}(\boldsymbol{y}; \boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\nu}) = \frac{2^p}{|\det(\boldsymbol{\Delta}_{\boldsymbol{y}-\boldsymbol{\xi}})| \prod_{j=1}^{p} \nu_j} F_2\left(\bar{\boldsymbol{\nu}}, \mathbf{1}_p; \mathbf{0}_p, \boldsymbol{\Delta}_{\boldsymbol{y}-\boldsymbol{\xi}}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Delta}_{\boldsymbol{y}-\boldsymbol{\xi}}^{-1}\right),$$

where $\bar{\boldsymbol{\nu}} = (\sqrt{1-\nu_1}, \ldots, \sqrt{1-\nu_p})$, $\mathbf{0}_p$ and $\mathbf{1}_p$ are p-dimensional vectors of zeros and ones, respectively, $\boldsymbol{\Delta}_{\boldsymbol{y}-\boldsymbol{\xi}} = \text{diag}(y_1 - \xi_1, \ldots, y_p - \xi_p)$, $\det(\boldsymbol{\Delta}_{\boldsymbol{y}-\boldsymbol{\xi}}) \neq 0$ and

$$F_k(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\xi}, \boldsymbol{\Sigma}) = \int_{\boldsymbol{a}}^{\boldsymbol{b}} \det(\boldsymbol{\Delta}_{\boldsymbol{t}})^k \phi_p(\boldsymbol{t}; \boldsymbol{\xi}, \boldsymbol{\Sigma}) d\boldsymbol{t}. \tag{14}$$

To see the extension to the case $\det(\boldsymbol{\Delta}_{\boldsymbol{y}-\boldsymbol{\xi}}) = 0$, see Punzo and Bagnato (2022a). The integral in Eq. 14 can be computed in an efficient and fast way for low-dimensional data by using a recursive relation whose details can be found in Kan and Robotti (2017) and Lee (1983) and Monte Carlo integration for high-dimensional data (Valeriano et al., 2023). Just as an example, R users can refer to the `recintab()` function of the **mnormt** package (Azzalini, 2016) and `mvtelliptical()` function of the **relliptical** package (Valeriano et al., 2024), respectively.

The skew dimension-wise scaled tail-inflated normal distribution (SDSTIN) distribution, denoted by $\boldsymbol{Y} \sim SDSTIN_p(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$, is obtained by taking $\tau_j \sim U(1-\nu_j, 1)$ in Eq. 5. The pdf of the SDSTIN distribution can be expressed as

$$f_{\text{SDSTIN}}(\boldsymbol{y}; \boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\nu}) = 2 f_{\text{DSTIN}}(\boldsymbol{y}; \boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\nu}) \Phi\left(\boldsymbol{\alpha}^\top (\boldsymbol{y} - \boldsymbol{\xi})\right).$$

According to Eq. 9, it can be shown that $\gamma_{ij} | \boldsymbol{Y}_i = \boldsymbol{y}_i \sim DTG_{(1-\nu_j,1)}(3/2, \eta_{ij})$, where DTG denotes the doubly truncated gamma distribution (Coffey & Muller, 2000) on the interval $(1-\nu_j, 1)$, with parameters $3/2$ and $\eta_{ij} = (y_{ij} - \xi_j)^2/\sigma_{jj}$. Here, σ_{ij} denotes the ith and jth component of $\boldsymbol{\Sigma}$. Therefore,

$$E(\gamma_{ij} \mid \boldsymbol{Y}_i = \boldsymbol{y}_i) = \frac{2}{\eta_{ij}} \frac{\Gamma(\frac{5}{2}, (1-\nu_j)\frac{\eta_{ij}}{2}) - \Gamma(\frac{5}{2}, \frac{\eta_{ij}}{2})}{\Gamma(\frac{3}{2}, (1-\nu_j)\frac{\eta_{ij}}{2}) - \Gamma(\frac{3}{2}, \frac{\eta_{ij}}{2})}, \tag{15}$$

where $\Gamma(a, x) = \int_x^\infty t^{a-1}e^{-t}dt$ denotes the incomplete gamma function. From Eq. 11, the E-step involves the calculation of $\gamma_{ij}^{(k)} = E(\gamma_{ij} \mid \boldsymbol{y}_i, \boldsymbol{\theta}^{(k)})$ and $w_i^{(k)} = E(W_i \mid \boldsymbol{y}_i, \boldsymbol{\theta}^{(k)})$, which can be easily evaluated via the results of Eqs. 8 and 15.

3.2 The Skew Dimension-Wise Scaled Shifted-Exponential Normal Distribution

The skew dimension-wise scaled shifted-exponential normal (SDSSEN) distribution is defined by setting τ_j in Eq. 5 as a shifted-exponential distribution, with the following pdf

$$h_{\tau_j}(\tau_j; \nu_j) = \nu_j e^{-\nu_j(\tau_j - 1)}, \quad \nu_j > 0 \text{ and } \tau_j > 1.$$

This yields the SDSSEN distribution with the following pdf

$$f_{\text{SDSSEN}}(\boldsymbol{y}; \boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\nu}) = 2 f_{\text{DSSEN}}(\boldsymbol{y}; \boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\nu}) \Phi\big(\boldsymbol{\alpha}^\top (\boldsymbol{y} - \boldsymbol{\xi})\big),$$

where f_{DSSEN} denotes the pdf of the dimension-wise scaled shifted-exponential normal (DSSEN; Punzo & Bagnato, 2022a) distribution given by

$$f_{\text{DSSEN}}(\boldsymbol{y}; \boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\nu}) = \frac{2^p \det(\boldsymbol{\Delta}_{\boldsymbol{\nu}}) e^{\boldsymbol{\nu}^\top \mathbf{1}_p}}{\sqrt{\det(\boldsymbol{\Sigma}) \det\left(\boldsymbol{\Delta}_{\boldsymbol{y}-\boldsymbol{\xi}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta}_{\boldsymbol{y}-\boldsymbol{\xi}} + 2\boldsymbol{\Delta}_{\boldsymbol{\nu}}\right)}} \times F_2\left(\mathbf{1}_p, \infty_p; \mathbf{0}_p, (\boldsymbol{\Delta}_{\boldsymbol{y}-\boldsymbol{\xi}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta}_{\boldsymbol{y}-\boldsymbol{\xi}} + 2\boldsymbol{\Delta}_{\boldsymbol{\nu}})^{-1}\right),$$

where ∞_p is a vector of length p of ∞s and $\det(\boldsymbol{\Delta}_{\boldsymbol{y}-\boldsymbol{\xi}}) \neq 0$. To see the extension to the case $\det(\boldsymbol{\Delta}_{\boldsymbol{y}-\boldsymbol{\xi}}) = 0$, see Punzo and Bagnato (2022a).

According to Eq. 9, the conditional distribution $\gamma_{ij} \mid \boldsymbol{Y}_i = \boldsymbol{y}_i$ is a left-truncated gamma distribution (see Dagpunar 1978, Philippe 1997 and Coffey and Muller 2000) on the interval $(1, \infty)$, with shape $3/2$ and rate $\nu_j + \eta_{ij}/2$, denoted by $\gamma_{ij} \mid \boldsymbol{Y}_i = \boldsymbol{y}_i \sim LTG_{(1,\infty)}(3/2, \nu_j + \eta_{ij}/2)$. In addition, according to Coffey and Muller (2000), it can be shown that

$$E(\gamma_{ij} \mid \boldsymbol{Y}_i = \boldsymbol{y}_i) = \frac{\psi_{\frac{3}{2}}\left(\nu_j + \frac{\eta_{ij}}{2}\right)}{\psi_{\frac{1}{2}}\left(\nu_j + \frac{\eta_{ij}}{2}\right)}, \tag{16}$$

where $\psi_g(x)$ is the Misra function (Misra, 1940), which is a further generalization of the generalized exponential integral function (Abramowitz & Stegun, 1968). To implement the ECME procedure for fitting the SDSSEN, the conditional expectations involved in Eq. 11 can be easily evaluated via the results of Eqs. 8 and 16. Besides, the parameters $\nu_j^{(k)}$ can be alternatively updated by maximizing the Q-function over ν_j, leading to the following CMQ step.

CMQ-step 4

$$\nu_j^{(k+1)} = \frac{n}{\sum_{i=1}^{n}(\gamma_{ij}^{(k)} - 1)}, \quad j = 1, \ldots, p.$$

In Fig. 1, we can see visual representations of bivariate SDSTIN and SDSSEN distributions. We consider $\boldsymbol{\xi} = (0, 0)^\top$, $\boldsymbol{\Sigma} = \text{diag}(1, 1)$, and different values for the skewness

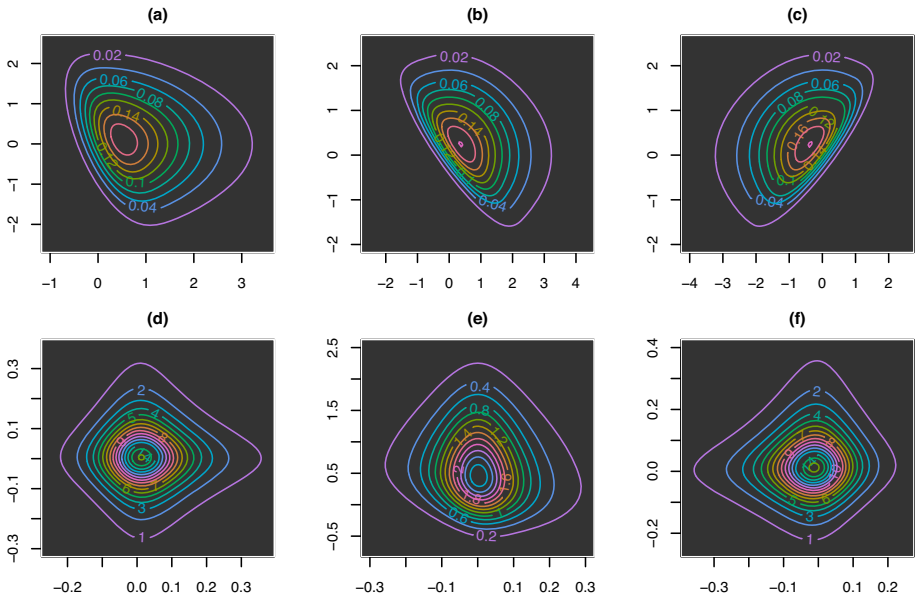

Fig. 1 Contour plots of the bivariate SDSMN models. The **a**–**c** correspond to the SDSTIN distribution with **a** $\boldsymbol{\lambda} = (3, 1)^\top$, **b** $\boldsymbol{\lambda} = (3, 3)^\top$, and **c** $\boldsymbol{\lambda} = (-3, 3)^\top$. The **d**–**f** correspond to the SDSSEN distribution with **d** $\boldsymbol{\lambda} = (3, 1)^\top$, **e** $\boldsymbol{\lambda} = (3, 3)^\top$, and **f** $\boldsymbol{\lambda} = (-3, 3)^\top$

parameters; as for the tailedness parameters, we use $\boldsymbol{\nu} = (0.9, 0.1)^\top$ for the SDSTIN and $\boldsymbol{\nu} = (0.008, 0.008)^\top$ for the SDSSEN. The plots demonstrate that the SDSMN distributions offer a wide range of shapes depending on the choice of shape parameters and can exhibit different levels of directional skewness and tail characteristics. It highlights the versatility of the SDSMN distributions and their ability to represent a wide range of shapes and characteristics.

4 Finite Mixtures of SDSMN (FM-SDSMN) Distributions

4.1 ML Estimation via the ECME Algorithm

We consider n independent random variables $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ which are taken from a G-component mixture of SDSMN distributions, given by

$$f(\boldsymbol{y}_i; \boldsymbol{\Theta}) = \sum_{g=1}^{G} \omega_g f_{\text{SDSMN}}(\boldsymbol{y}_i; \boldsymbol{\xi}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\lambda}_g, \boldsymbol{\nu}_g),$$

where $\omega_g > 0$, $\sum_{g=1}^{G} \omega_g = 1$, and $\boldsymbol{\Theta} = (\omega_1, \ldots, \omega_{G-1}, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_G^\top)^\top$ denotes all parameters, with $\boldsymbol{\theta}_g = (\boldsymbol{\xi}_g^\top, \text{vech}(\boldsymbol{\Sigma}_g)^\top, \boldsymbol{\lambda}_g^\top, \boldsymbol{\nu}_g^\top)$ and $\boldsymbol{\nu}_g = (\boldsymbol{\nu}_{1g}^\top, \ldots, \boldsymbol{\nu}_{pg}^\top)^\top$ denoting the parameters of component g. Let us introduce a G-dimensional binary random variable $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iG})^T$ where a particular element Z_{ig} is equal to 1 if $\boldsymbol{Y}_i$ belongs to group g and is equal to zero, otherwise. Note that, $\boldsymbol{Z}_i$ follows a multinomial random vector with 1

trial and cell probabilities $\omega_1, \ldots, \omega_G$ denoted by $\boldsymbol{Z}_i \sim Mult(1; \omega_1, \ldots, \omega_G)$. Obviously

$$\boldsymbol{Y}_i | Z_{ig} = 1 \sim SDSMN_p\Big(\boldsymbol{\xi}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\lambda}_g, \nu_g\Big),$$

$$W_i | (\boldsymbol{Y}_i = \boldsymbol{y}_i, Z_{ig} = 1) \sim DTN\Big(\boldsymbol{\alpha}_g^\top (\boldsymbol{y}_i - \boldsymbol{\xi}_g), 1\Big) \boldsymbol{I}_{(0,\infty)},$$

where $\boldsymbol{\alpha}_g^\top = \boldsymbol{\lambda}_g^\top \boldsymbol{\sigma}_g^{-1/2}$.

In addition, according to Eq. 9 the conditional distribution of $\boldsymbol{\gamma}_i$ can be derived by

$$f(\boldsymbol{\gamma}_i \mid \boldsymbol{y}_i, z_{ig} = 1) \propto \phi_p\big(\boldsymbol{y}_i; \boldsymbol{\xi}_g, \boldsymbol{\Delta}_{\boldsymbol{\gamma}_i}^{1/2} \boldsymbol{\Sigma}_g \boldsymbol{\Delta}_{\boldsymbol{\gamma}_i}^{1/2}\big) \prod_{j=1}^{p} h_{\tau_j}(\gamma_j; \nu_{gj}).$$

Letting $\boldsymbol{y}_o = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_n^\top)^\top$, $\boldsymbol{W} = (W_1, \ldots, W_n)^\top$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_n^\top)^\top$ and $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)^\top$, excluding additive constants, the complete data log-likelihood function of $\boldsymbol{\Theta}$ for the complete data $\boldsymbol{y}_c = (\boldsymbol{y}_o^\top, \boldsymbol{W}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{Z})^\top$ is given by

$$\begin{aligned}\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{y}_c) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} Z_{ig} \Bigg\{ & \log\big(\det(\boldsymbol{\Sigma}_g)\big) + (\boldsymbol{y}_i - \boldsymbol{\xi}_g)^\top \boldsymbol{\Delta}_{\boldsymbol{\gamma}_i}^{-1/2} \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\Delta}_{\boldsymbol{\gamma}_i}^{-1/2} (\boldsymbol{y}_i - \boldsymbol{\xi}_g) \\ & + \big(W_i - \boldsymbol{\alpha}_g^\top (\boldsymbol{y}_i - \boldsymbol{\xi}_g)\big)^2 - 2 \sum_{j=1}^{p} \log h_{\tau_j}(\gamma_{ij}; \nu_{gj}) - 2 \log \omega_g \Bigg\}. \end{aligned} \tag{17}$$

The expected value of Eq. 17 to start the E-step given the current parameter $\boldsymbol{\Theta}^{(k)}$ requires some necessary conditional expectations, including

$$z_{ig}^{(k)} = E(Z_{ig} | \boldsymbol{y}_i, \boldsymbol{\Theta}^{(k)}) = \frac{\omega_g^{(k)} f_{\text{SDSMN}}(\boldsymbol{y}_i; \boldsymbol{\theta}_g^{(k)})}{f(\boldsymbol{y}_i; \boldsymbol{\Theta}^{(k)})}, \tag{18}$$

$$\boldsymbol{\gamma}_{ig}^{(k)} = E(\boldsymbol{\gamma}_i \mid \boldsymbol{y}_i, Z_{ig} = 1, \boldsymbol{\Theta}^{(k)}) = E(\boldsymbol{\gamma}_i \mid \boldsymbol{y}_i, \boldsymbol{\theta}_g^{(k)}), \tag{19}$$

$$w_{ig}^{(k)} = E(W_i \mid \boldsymbol{y}_i, Z_{ig} = 1, \boldsymbol{\Theta}^{(k)}) = E(W_i \mid \boldsymbol{y}_i, \boldsymbol{\theta}_g^{(k)}), \tag{20}$$

and

$$\begin{aligned} s_{ijg}^{(k)}(\nu_{jg}) &= E(\log h_{\tau_j}(\gamma_{ij}; \nu_{jg}) \mid \boldsymbol{y}_i, Z_{ig} = 1, \boldsymbol{\Theta}^{(k)}) \\ &= E(\log h_{\tau_j}(\gamma_{ij}; \nu_{jg}) \mid \boldsymbol{y}_i, \boldsymbol{\theta}_g^{(k)}). \end{aligned} \tag{21}$$

Therefore, the conditional expectation of the complete data log-likelihood yields

$$\begin{aligned} Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(k)}) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig}^{(k)} \Bigg\{ & \log\big(\det(\boldsymbol{\Sigma}_g)\big) + (\boldsymbol{y}_i - \boldsymbol{\xi}_g)^\top \boldsymbol{\Delta}_{\boldsymbol{\gamma}_{ig}^{(k)}}^{-1/2} \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\Delta}_{\boldsymbol{\gamma}_{ig}^{(k)}}^{-1/2} (\boldsymbol{y}_i - \boldsymbol{\xi}_g) \\ & + [\boldsymbol{\alpha}_g^\top (\boldsymbol{y}_i - \boldsymbol{\xi}_g)]^2 - 2 w_{ig}^{(k)} \boldsymbol{\alpha}_g^\top (\boldsymbol{y}_i - \boldsymbol{\xi}_g) - 2 \sum_{j=1}^{p} s_{ijg}^{(k)}(\nu_{jg}) - 2 \log \omega_g \Bigg\}. \end{aligned}$$

In summary, the ECME algorithm iterates one E-step, four CMQ steps, and one CML step until convergence. These steps are summarized below.

E-step Given $\boldsymbol{\Theta} = \boldsymbol{\Theta}^{(k)}$, compute $z_{ig}^{(k)}$, $\boldsymbol{\gamma}_{ig}^{(k)}$, $w_{ig}^{(k)}$, and $s_{ijg}^{(k)}$ for $i = 1, \ldots, n$, $j = 1, \ldots, p$, and $g = 1, \ldots, G$ using Eqs. 18–21.

CMQ-step 1 Calculate

$$\omega_g^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} z_{ig}^{(k)}.$$

CMQ-step 2 Update $\boldsymbol{\xi}_g^{(k)}$ by

$$\boldsymbol{\xi}_g^{(k+1)} = \left[\sum_{i=1}^{n} z_{ig}^{(k)}\left(\boldsymbol{\Delta}_{\boldsymbol{\gamma}_{ig}^{(k)}}^{-1/2}\boldsymbol{\Sigma}_g^{(k)-1}\boldsymbol{\Delta}_{\boldsymbol{\gamma}_{ig}^{(k)}}^{-1/2} + \boldsymbol{\alpha}_g^{(k)}\boldsymbol{\alpha}_g^{\top(k)}\right)\right]^{-1}\left[\sum_{i=1}^{n} z_{ig}^{(k)}\boldsymbol{\Delta}_{\boldsymbol{\gamma}_{ig}^{(k)}}^{-1/2}\boldsymbol{\Sigma}_g^{(k)-1}\boldsymbol{\Delta}_{\boldsymbol{\gamma}_{ig}^{(k)}}^{-1/2}\boldsymbol{y}_i + \boldsymbol{\alpha}_g^{(k)}\boldsymbol{\alpha}_g^{\top(k)}\sum_{i=1}^{n} z_{ig}^{(k)}\boldsymbol{y}_i - \boldsymbol{\alpha}_g^{(k)}\sum_{i=1}^{n} z_{ig}^{(k)} w_{ig}^{(k)}\right],$$

where $\boldsymbol{\Delta}_{\boldsymbol{\gamma}_{ig}^{(k)}} = \text{diag}(1/\gamma_{i1g}^{(k)}, \ldots, 1/\gamma_{ipg}^{(k)})$.

CMQ-step 3 Update $\boldsymbol{\Sigma}_g^{(k)}$by

$$\boldsymbol{\Sigma}_g^{(k+1)} = \frac{1}{\sum_{i=1}^{n} z_{ig}^{(k)}}\sum_{i=1}^{n} z_{ig}^{(k)}\boldsymbol{\Delta}_{\boldsymbol{\gamma}_{ig}^{(k)}}^{-1/2}(\boldsymbol{y}_i - \boldsymbol{\xi}_g^{(k+1)})(\boldsymbol{y}_i - \boldsymbol{\xi}_g^{(k+1)})^{\top}\boldsymbol{\Delta}_{\boldsymbol{\gamma}_{ig}^{(k)}}^{-1/2}.$$

CMQ-step 4 Update $\boldsymbol{\alpha}_g^{(k)}$ by

$$\boldsymbol{\alpha}_g^{(k+1)} = \left[\sum_{i=1}^{n} z_{ig}^{(k)}(\boldsymbol{y}_i - \boldsymbol{\xi}_g^{(k+1)})(\boldsymbol{y}_i - \boldsymbol{\xi}_g^{(k+1)})^{\top}\right]^{-1}\sum_{i=1}^{n} z_{ig}^{(k)} w_{ig}^{(k)}(\boldsymbol{y}_i - \boldsymbol{\xi}_g^{(k+1)}).$$

CML step Update $\boldsymbol{\nu}^{(k)} = (\boldsymbol{\nu}_1^{\top}, \ldots, \boldsymbol{\nu}_G^{\top})^{\top}$ by optimizing the following constrained log-likelihood function:

$$\boldsymbol{\nu}^{(k+1)} = \arg\max_{\boldsymbol{\nu}} \sum_{i=1}^{n} \log\left(\sum_{g=1}^{G} \omega_g^{(k+1)} f_{\text{SDSMN}}(\boldsymbol{y}_i; \boldsymbol{\xi}_g^{(k+1)}, \boldsymbol{\Sigma}_g^{(k+1)}, \boldsymbol{\lambda}_g^{(k+1)}, \boldsymbol{\nu}_g)\right).$$

For the SDSSEN distribution the update of ν_{gj} can be alternatively obtained by maximizing the Q-function over ν_{gj}, leading to the following CMQ step.

CMQ-step 5

$$\nu_{gj}^{(k+1)} = \frac{\sum_{i=1}^{n} z_{ig}^{(k)}}{\sum_{i=1}^{n} z_{ig}^{(k)}(\gamma_{ijg}^{(k)} - 1)}, \quad j = 1, \ldots, p \text{ and } g = 1, \ldots, G.$$

In the fitting of finite mixture models, one common type of non-identifiability is induced by an overfitting of the number of components (McLachlan & Peel, 2000; Frühwirth-Schnatter, 2006). This kind of problem generally occurs when the resulting mixture model either has a

nearly empty component or two components with identical parameter values. Another type of non-identifiability arises from the well-known label switching problem (Celeux et al., 2000; Stephens, 2000), namely two resulting mixture models are equivalent apart from a permutation of component labels. A thorough theoretical investigation of identifiability would be worthwhile but is extremely challenging. This is the reason why we assess identifiability empirically through simulations, demonstrating that the proposed estimation procedure can accurately recover the true parameter values. The simulation study, described in Sect. A.1 of the Supplementary Material, provides a valuable complement to the theoretical considerations.

4.2 Initialization and Convergence Criterion

Before initiating iterative procedures, it is crucial to define a sensible set of initial values to facilitate faster convergence. To begin the ECME algorithm for fitting the FM-SDSMN models, a practical approach is to create an initial data partition $\{\boldsymbol{y}_g^{(0)}\}_{g=1}^G$ using the K-means method or its modified extensions (Cuesta-Albertos et al., 1997; Pelleg et al.,2000, and Kaufman & Rousseeuw, 2009) for getting a reliable estimate of $z_{ig}^{(0)}$. This leads directly to the calculation of $\boldsymbol{\omega}_g^{(0)} = \sum_{i=1}^n z_{ig}^{(0)}/n$. Next, we compute the sample means and covariance of $\boldsymbol{y}_g^{(0)}$ as initial estimates for $\boldsymbol{\xi}_g^{(0)}$ and $\boldsymbol{\Sigma}_g^{(0)}$. The initial skewness vectors for each component, $\boldsymbol{\lambda}_g^{(0)}$, are determined by the sample skewness of $\boldsymbol{y}_g^{(0)}$. Finally, we initialize $\hat{\boldsymbol{\nu}}_g^{(0)}$ with moderately small values such as 5 or 10.

Regarding the stopping criterion, the algorithm is considered to have converged when the relative change between two successive observed-data log-likelihood values is less than a specified tolerance, expressed as $\ell(\boldsymbol{\Theta}^{(k+1)})/\ell(\boldsymbol{\Theta}^{(k)}) - 1 < \varepsilon$, where $\varepsilon = 10^{-6}$ is employed.

5 Real Data Analyses

In this section, we first demonstrate real-world scenarios where our SDSMN family of distributions outperforms well-established multivariate distributions available in the literature (refer to Sect. 1). We focus on the field of allometry (Sect. 5.1). Next, we demonstrate the importance of using flexible distributions to accommodate complex cluster shapes in model-based clustering through our FM-SDSMN models, with a particular focus on the biomedical and customer segmentation fields (Sect. 5.2 and Sect. B in the Supplementary Material). The entire analysis is conducted in R.

5.1 No Clusters: Allometry

5.1.1 Preliminary Statistical Tests

As a preliminary exploratory step of the analysis, along with providing summary statistics and graphical representations of the data, we test the null hypotheses of no correlation, elliptical symmetry, and central symmetry. Following Punzo and Bagnato (2022a), we compute a two-sided test of no correlation using the classical test statistic based on the Pearson product-moment correlation coefficient. This test is performed with the `cor.test()` function from the **stats** package. For testing elliptical symmetry, we use the MPQ test by Manzotti

et al. (2002), implemented through the `MPQ()` function in the **ellipticalsymmetry** package (Babić et al., 2020). This test is selected because it maintains the claimed nominal significance level (Babić et al., 2011). Lastly, to assess the more challenging null hypothesis of central symmetry—the type of symmetry involved by DSMN distributions (Punzo & Bagnato, 2022a)—we apply the simple union-intersection multiple testing procedure proposed by Punzo and Bagnato (2022a), and we refer to that paper for further details.

5.1.2 Model Fitting

As a second step in the analysis, we fit 31 different well-known models to the data, including our SDSMN distributions (i.e., the SDSSEN and SDSTIN). The first group of sixteen models includes the multivariate generalized hyperbolic (MGH) distribution and its most well-known special or limiting cases, as described by McNeil et al. (2005). For a hierarchical representation of the relationships among these models, see Bagnato et al. (2024). This group includes nine elliptical-symmetric and seven asymmetric models. In addition, we consider four other elliptical-symmetric models: the multivariate tail-inflated normal (MTIN) by Punzo and Bagnato (2021), the multivariate shifted-exponential normal (MSEN) by Punzo and Bagnato (2020), the multivariate leptokurtic normal (MLN) by Bagnato et al. (2023, see also Browne et al. 2017), and the multivariate contaminated normal (MCN) as parametrized by Punzo et al. (2018a). We also include three well-known asymmetric models: the multivariate skew-normal, multivariate skew-t, and multivariate skew contaminated normal, as formulated by Cabral et al. (2012). For the sake of completeness, we also fit the multivariate Manly-transformed normal (MMTN) by Zhu and Melnykov (2018), a well-established multivariate asymmetric distribution constructed using the transformation-based approach (cf. Sect. 1). Lastly, we examine four multiple scaled distributions: the multiple scaled shifted-exponential normal (MSSEN) by Punzo and Bagnato (2022b), multiple scaled t (MSt) by Forbes and Wraith (2014), multiple scaled contaminated normal (MSCN) by Punzo and Tortora (2021), and two DSMN distributions, the DSSEN and DSTIN (Punzo & Bagnato, 2022a).

5.1.3 Computational and Operational Details

For consistency, we estimate the parameters of all models using the ML approach. We use the ECME algorithm described in Sect. 2.2 to obtain ML estimates for our SDSMN distributions. The code for density evaluation, random number generation, and fitting for the SDSMN models is available at https://github.com/a-mahdavi/EM.mixSDSMN. For the competing models, we use the `fit.ghypmv()` function from the **ghyp** package (Weibel et al., 2022) to fit all members of the MGH family. We fit the MLN using the `WML.MLN()` function available at http://docenti.unict.it/punzo/Rcode.htm. The MTIN is fitted using the `mtin.ML.ECME()` function from the **mtin** package, available at http://docenti.unict.it/punzo/Rpackages.htm. The MSEN is fitted using the `msen.ML.EM()` function from the **msen** package, also available at http://docenti.unict.it/punzo/Rpackages.htm. We use the `CNmixt()` function from the **ContaminatedMixt** package (Punzo et al., 2018a, b) to fit the MCN. For the models proposed by Cabral et al. (2012), we use the `smsn.mmix()` function from the **mixsmsn** package (Weibel et al., 2022). We use the **ManlyMix** package (Zhu & Melnykov, 2023) to fit the MMTN distribution. To fit the multiple scaled models, we use the `msdist.ML.EM()` function available at http://docenti.unict.it/punzo/Rcode.htm for the MSSEN, and the `mst()` function from the **MSclust** package to fit the MSt (Tortora et al., 2023). The code to fit the DSMN distributions is available at http://docenti.unict.it/punzo/Rcode.

5.1.4 Model Selection

In the set of 31 candidate models used to describe the data distribution, models often differ in the number of parameters. In such cases, model selection is typically required to identify the most suitable model among them. When adopting the maximum likelihood paradigm, a standard approach for model selection is to compute an appropriate likelihood-based criterion. Two widely used criteria, which in our formulation need to be maximized and are employed in this work, are the Akaike information criterion (AIC; Akaike, 1974)

$$\text{AIC} = 2\ell(\hat{\boldsymbol{\Theta}}) - 2 \times \#\text{par}, \tag{22}$$

and the Bayesian information criterion (BIC; Schwarz, 1978)

$$\text{BIC} = 2\ell(\hat{\boldsymbol{\Theta}}) - \log n \times \#\text{par}, \tag{23}$$

where #par represents the total number of free parameters in the model, and $\hat{\boldsymbol{\Theta}}$ denotes the ML estimate of $\boldsymbol{\Theta}$ obtained upon convergence of the ECME algorithm, as described in Sect. 4.1; see Burnham and Anderson (2013) for further details on these criteria. As evident from Eqs. 22 and 23, AIC and BIC share a similar structure but differ in their penalty for model complexity. Specifically, BIC penalizes free parameters more strongly than AIC when $n > 7$. Unlike the likelihood ratio (LR) test, which requires models to be nested, AIC and BIC can be used to compare non-nested models (Schwarz, 1978). A comparative analysis of AIC and BIC is provided by Burnham and Anderson (2013, Sects. 6.3–6.4), with additional insights from Burnham and Anderson (2004). A recent simulation-based study by Punzo and Bagnato (2021), conducted to select among a sufficiently large set of candidate elliptical models—similar to the context of this paper—demonstrated that AIC and BIC perform comparably. Theoretically, BIC is often recommended for selecting the "true model"—that is, the data-generating process—if it is present in the candidate set. More precisely, when the "true model" is included in the set, BIC will identify it with probability 1 as $n \to \infty$. In contrast, the probability of AIC selecting the "true model" may be lower than 1 (Burnham & Anderson, 2013, Sects. 6.3–6.4, Vrieze, 2012, and Aho et al., 2014). However, proponents of AIC argue that this distinction is largely theoretical, as the "true model" is almost never part of the candidate set. In cases where the "true model" is not included, the best possible outcome is selecting the model that best approximates it. Under certain assumptions, AIC is well-suited for identifying the best approximating model.

Another important factor to consider when choosing between model selection criteria is the type of analysis. While in this section, we focus on density estimation, where the goal is to model the data distribution as accurately as possible, in Sect. 5.2, the objective shifts to clustering, aiming to uncover the true underlying group structure within the data. Studies have shown that the AIC often overestimates the number of clusters due to its lighter penalty on model complexity, leading to models that fit the data well but may introduce unnecessary complexity (Biernacki et al., 2000). In contrast, the BIC applies a stronger penalty, typically selecting models with fewer clusters and aligning more closely with the true underlying structure, particularly in large samples (Biernacki et al., 2000). In summary, while both AIC and BIC are valuable model selection tools, their differing penalty structures make AIC more suitable for density estimation, where capturing the full data distribution is the priority. Conversely, BIC's emphasis on model parsimony and its consistency properties make it more appropriate for clustering applications, where determining the correct number of distinct groups is essential (Li & Maitra, 2021). All these considerations explain why we rely solely on the AIC for the density estimation analysis in this section and why we use

only the BIC for the model-based clustering analyses presented in Sect. 5.2 and Sect. B of the Supplementary Material.

Additionally, we use the LR test to compare the multivariate normal (MN) distribution (null model) against each of the competing distributions as alternative models, effectively conducting 30 different LR tests of normality. For all the tests, we compare the resulting p-values to the standard 5% significance level.

5.1.5 Motivation

In allometric studies, one of the most important issues is modeling the joint distribution of morphometric variables related to various organisms (Klingenberg, 1996). Therefore, the statistical models that are commonly considered are multivariate (almost always bivariate).

As a habit in this literature, measurements are log-transformed. The main motivation behind this choice is to remove marginal skewness and leptokurtosis such that new variables are more likely to be jointly MN-distributed (for other practical and theoretical reasons why it is often useful to transform data to logarithms, see, e.g., Pimentel1979, Reyment 1991, and Bookstein 1997). However, the log-transformation may fail to do what it is purported to do (Feng et al., 2013 and Curran-Everett, 2018), so that the log-transformed measurements are still inconsistent with the MN distribution. Moreover, the departures from normality may not only be due to marginal behavior but also to the joint one, especially in terms of symmetric/asymmetric shapes for the underlying isodensities. All these findings are corroborated by many empirical studies (see, e.g., Punzo and Bagnato 2020, 2022a, b). Distributions defined under a multiple-scaled approach should allow us to cope with all these issues. However, as we have emphasized in Sect. 1, for the sake of interpretation of the obtained results, the need arises to have skewness and tailedness parameters directly related to the skewness and leptokurtosis on each dimension. All these considerations motivate the use of our SDSMN family.

5.1.6 Data and Results

The allometric data analysis considers the anthropometric survey of U.S. Army Soldiers (ANSUR II) conducted by the Natick Soldier Research, Development and Engineering Center (NSRDEC) from October 2010 to April 2012 and comprised of personnel representing the total U.S. Army force. The dataset[1] contains 93 anthropometric measurements, along with 15 demographic/administrative variables, of U.S. male and female soldiers; see Gordon et al. (2014) for details and summary statistics on these data. The initial goal of the survey was to acquire a large body of data from comparably measured males and females to serve the Army's design and engineering needs. Nowadays, these data have many commercial, industrial, and academic applications.

We analyze the ANSUR II Male data set related to a sample of $n = 4,082$ male soldiers. We focus on the logarithm of two available anthropometric measurements on the head: head circumference (HC) and tragion to top-of-head (TTOH); see Park et al. (2021) for a clear explanation and an example of the use of these measurements. All measurements are expressed in millimeters.

Table 1 provides descriptive statistics for the logarithms of HC and TTOH separately considered.

[1] Available at http://mreed.umtri.umich.edu/mreed/downloads.html#ansur2

Table 1 Descriptive statistics for the logarithms of HC and TTOH

	Mean	Std. Dev.	Skewness	Excess kurtosis
log-HC	6.353	0.028	0.455	0.076
log-TTOH	4.875	0.047	−1.517	0.103

While the distribution of log-HC shows a positive skewness coefficient (0.455), for log-TTOH we have a negative skewness coefficient which is larger in magnitude (−1.517). As for the kurtosis, there is a positive excess kurtosis for both variables.

Figure 2 shows the scatterplot of the data.

The scatter appears to be skewed, based on visual inspection. This observation is supported by the outcomes of the statistical tests conducted. Indeed, the p-value from the MPQ test of elliptical symmetry is lower than 0.05 (0.011), and the union-intersection test of central symmetry rejects the null hypothesis (minimum adjusted p-value lower than 0.001). As for the correlation structure, the sample correlation coefficient is 0.385, with a p-value lower

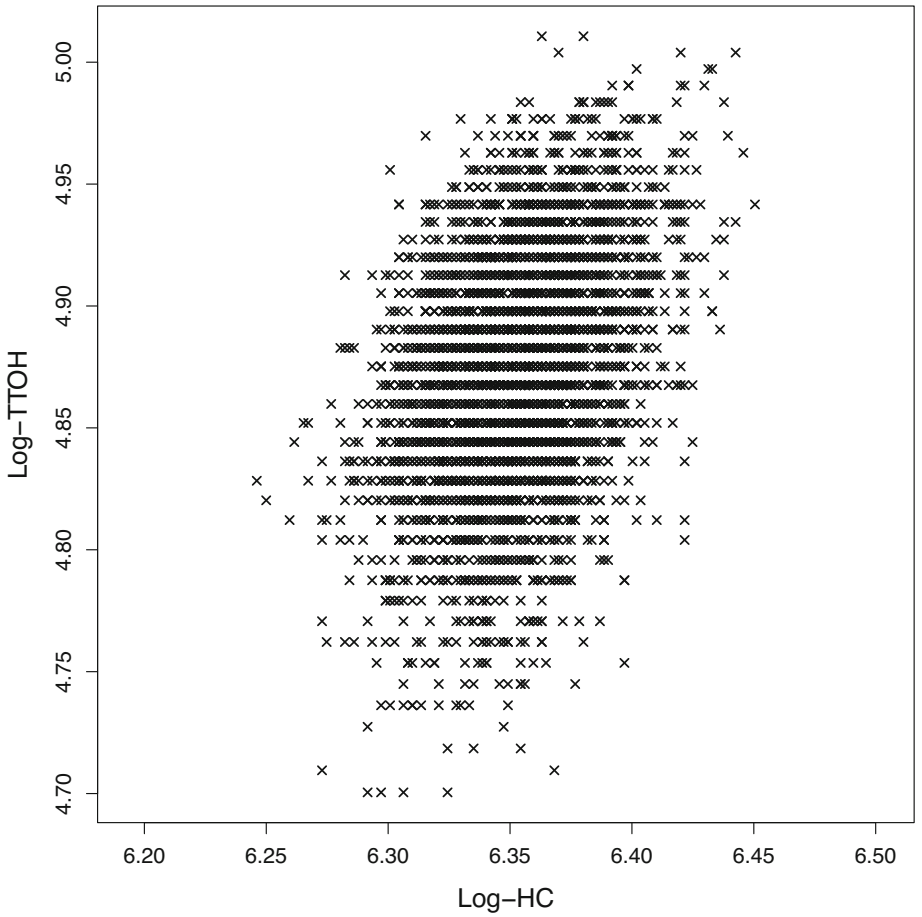

Fig. 2 Scatterplot of the logarithms of HC and TTOH

than 0.001 from the test of no correlation. Summarizing the insights from Table 1 and Fig. 2, it seems we would need a bivariate model allowing for correlation between variables, bivariate skewness, and excess kurtosis on both the logarithmic-transformed measurements.

Table 2 presents the model comparison results. First, regardless of the criterion used—AIC or LR test of bivariate normality—our SDSMNs, along with the Manly-transformed normal distribution, consistently rank among the top three models with the SDSSEN being always the best-performing model. Notably, only models that incorporate skewness show a significant improvement over the normal model. Moreover, the LR test suggests that if skewed models are excluded from the candidate set, the bivariate normal distribution would provide the best fit for the data. This finding supports the earlier visual assessments in Table 1 and Fig. 2, which highlighted the importance of accounting for skewness. On the other hand, symmetric multiple scaled distributions are consistently ranked among the worst models according to the considered criteria. Notably, based on the AIC and the LR test, only models that incorporate skewness show a significant improvement over the normal model. Moreover, the LR test suggests that if skewed models are excluded from the candidate set, the bivariate normal distribution would provide the best fit for the data. This finding supports the earlier visual assessments in Table 1 and Fig. 2, which highlighted the importance of accounting for skewness. On the other hand, symmetric multiple scaled distributions are consistently ranked among the worst models.

Interestingly, the set of candidate models also includes two asymmetric distributions—the asymmetric Cauchy and asymmetric Laplace—but they are positioned at the opposite end of the ranking, regardless of the criterion used. To understand why these asymmetric distributions perform poorly despite their skewness, it is worth noting that their symmetric counterparts (Cauchy and Laplace) also exhibit poor performance. In fact, all four models (symmetric/asymmetric Cauchy and symmetric/asymmetric Laplace) rank as the worst-performing models overall, regardless of the selection criterion. This result can be explained from two different perspectives—one common to both distribution families (Cauchy and Laplace), regardless of symmetry, and the other specific to each family. The common limitation is the absence of a dedicated parameter for controlling leptokurtosis, which may be too restrictive for this dataset. From a family-specific standpoint, Cauchy-based distributions have undefined marginal higher-order moments beyond the first absolute moment, potentially leading to excessively heavy tails relative to the empirical marginal excess kurtoses (see Table 1). Meanwhile, Laplace-based distributions exhibit a distinctive cusp at the mode, which may not align well with the shape of the empirical data distribution in its bulk.

The estimated parameters for the SDSSEN distribution, the model selected by the AIC, are

$$\widehat{\boldsymbol{\xi}} = \begin{bmatrix} 6.353 \\ 4.911 \end{bmatrix}, \quad \widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.0009 & 0.0006 \\ 0.0006 & 0.0042 \end{bmatrix}, \quad \widehat{\boldsymbol{\lambda}} = \begin{bmatrix} 0.035 \\ -0.241 \end{bmatrix}, \quad \text{and} \quad \widehat{\boldsymbol{\nu}} = \begin{bmatrix} 6.630 \\ 4.000 \end{bmatrix}.$$

We can observe that the signs of the estimated values of λ_1 and λ_2 are consistent with the skewness coefficients shown in Table 1. Similarly, the magnitude order of the estimated values of ν_1 and ν_2 aligns with expectations, considering that for the SEN distribution, a lower ν corresponds to higher excess kurtosis (refer to Punzo and Bagnato 2020).

Figure 3 shows the scatterplot of log-HC and log-TTOH with superimposed contour lines from the SDSSEN.

The isodensities highlight a complex type of asymmetry and, at the same time, visually corroborate a good fit of the model to the data.

Table 2 Model comparison, for the joint distribution of log-HC and log-TTOH, in terms of number of parameters (#par), log-likelihood, AIC (multiplied by −1), and p-values from the LR tests of bivariate normality. Rankings induced by the considered criteria are also displayed

Model	#par	Log-lik.	AIC	Ranking	Normality LR p-value	Ranking
Normal	5	15,814.66	31,619.32	12	–	–
Cauchy	5	15,222.91	30,435.81	31	1.000	27
Laplace	5	15,338.46	30,666.93	28	1.000	27
t	6	15,814.74	31,617.49	21	0.686	21
Hyperbolic univariate marginals	6	15,814.76	31,617.52	19	0.657	19
Symmetric variance gamma	6	15,814.76	31,617.53	17	0.650	17
Symmetric hyperbolic	6	15,814.76	31,617.52	20	0.658	20
Symmetric normal inverse Gaussian	6	15,814.76	31,617.52	18	0.657	18
Symmetric generalized hyperbolic	7	15,814.76	31,615.52	23	0.906	22
Asymmetric Cauchy	7	15,227.64	30,441.28	30	1.000	27
Asymmetric Laplace	7	15,338.62	30,663.24	29	1.000	27
Asymmetric t	8	15,837.99	31,659.99	4	<0.001	4
Normal inverse Gaussian	8	15,836.68	31,657.37	6	<0.001	5
Variance Gamma	8	15,835.70	31,655.39	8	<0.001	8
Hyperbolic	8	15,836.65	31,657.30	7	<0.001	6
Generalized hyperbolic	9	15,837.76	31,657.53	5	<0.001	7
Tail-inflated normal	6	15,814.91	31,617.82	14	0.482	15

Table 2 continued

Model	#par	Log-lik.	AIC	Ranking	Normality LR p-value	Ranking
Shifted-exponential normal	6	15,814.81	31,617.61	16	0.591	16
Leptokurtic normal	6	15,814.92	31,617.84	13	0.470	14
Contaminated normal	7	15,814.66	31,615.32	24	1.000	23
Skew-normal	7	15,827.68	31,641.37	9	<0.001	9
Skew-t	8	15,823.56	31,631.12	11	<0.001	11
Skew contaminated normal	9	15,827.87	31,637.74	10	<0.001	10
Manly-transformed normal	7	15,840.98	31,667.97	3	<0.001	2
Multiple-scaled shifted-exponential normal	7	15,813.43	31,612.86	26	1.000	27
Multiple-scaled t	7	15,814.51	31,615.02	25	1.000	27
Multiple-scaled contaminated normal	9	15,814.66	31,611.32	27	1.000	24
Dimension-wise tail-inflated normal	7	15,815.84	31,617.69	15	0.307	12
Dimension-wise shifted-exponential normal	7	15,815.67	31,617.34	22	0.364	13
Skew dimension-wise tail-inflated normal	9	15,844.34	31,670.68	2	<0.001	3
Skew dimension-wise shifted-exponential normal	9	15,844.59	31,671.17	1	<0.001	1

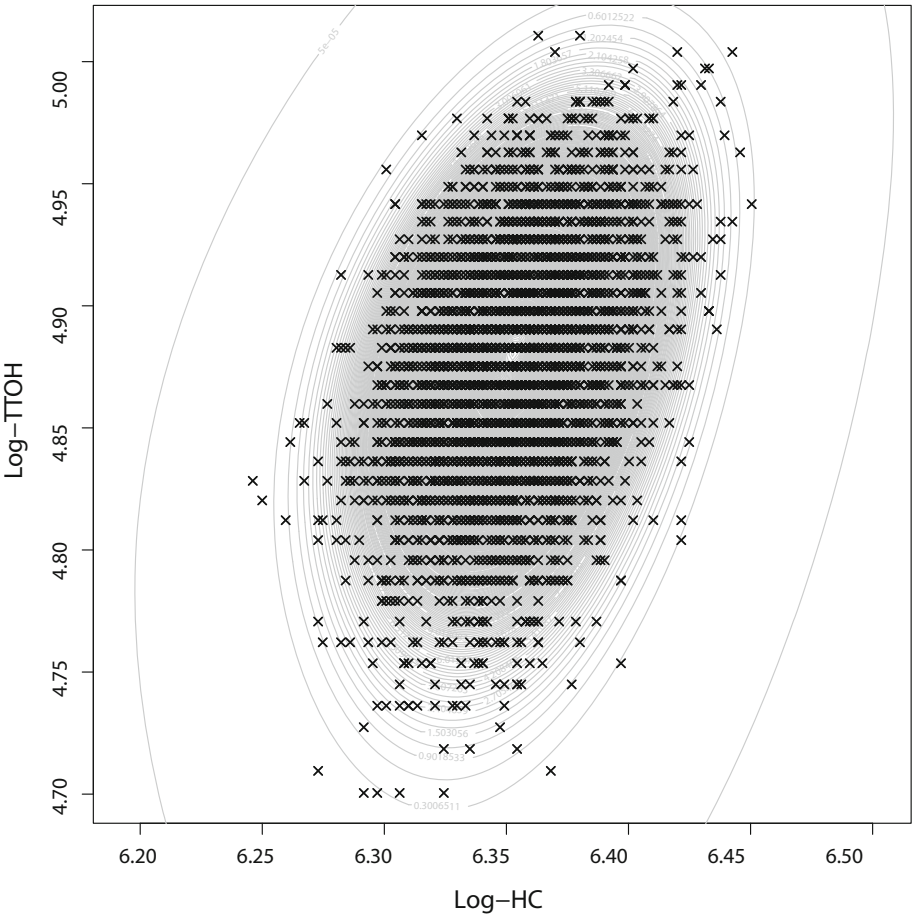

Fig. 3 Scatterplot for log-BSA and log-IB. Grey lines from the ML-fitted SDSSEN distribution are superimposed on the scatterplot

In conclusion, the SDSSEN distribution successfully captures all the characteristics of the data: the positive correlation between variables, the type of asymmetry with differing skewness signs for the two measurements, and the higher excess kurtosis for log-TTOH.

5.2 Model-Based Clustering

In this section, we examine three benchmark real-world datasets commonly used in model-based clustering literature. The first two datasets come from the biomedical field (Sects. 5.2.1 and B.1 of the Supplementary Material), while the third pertains to customer segmentation (Sect. B.2 of the Supplementary Material). For insights into the significance of clustering in these areas, refer to Wedel and Kamakura (2012) for market/customer segmentation and Schlattmann (2010) for medical applications.

On each dataset, we fit eight finite mixtures, including our FM-SDSMN models described in Sect. 4. The first group of competing models includes finite mixtures with multivariate

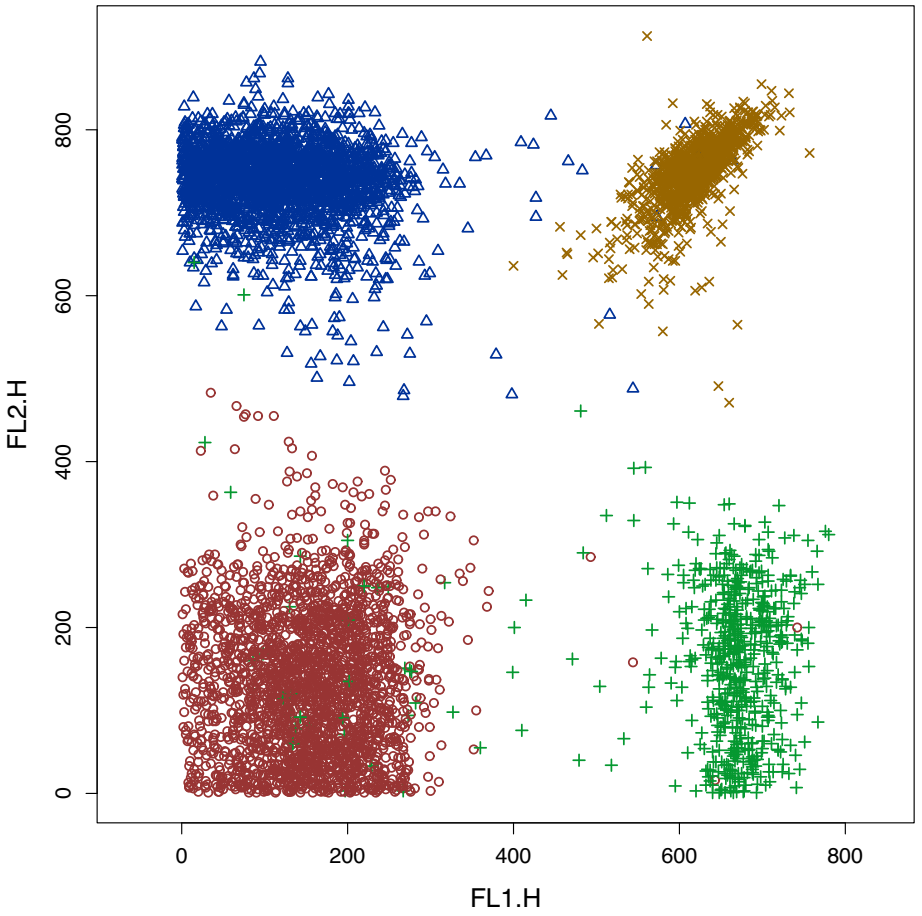

Fig. 4 HSCT data. Scatterplot of FL1.H and FL2.H

skew-normal (MSN) and multivariate skew-*t* (MST) components, both of which belong to the family of skew-normal independent (SNI) distributions (Cabral et al., 2012). These two models are fitted using the **mixsmsn** package. The second group involves finite mixtures of MTIN and MSEN distributions (Tomarchio et al., 2022), fitted using the **SenTinMixt** package (Tomarchio et al., 2021). Furthermore, we also consider finite mixtures of multiple-scaled generalized hyperbolic (FM-MSGH) distributions and convex mixtures of multiple-scaled generalized hyperbolic (FM-CMSGH) distributions; these are fitted using the **MixGHD** package (Tortora et al., 2021, 2019). Finally, we fit the finite mixtures of multivariate Manly-transformed normal (FM-MMTN) model introduced by Zhu and Melnykov (2018) using **ManlyMix** package.

To compare the performance of the candidate models, besides calculating the BIC as motivated in Sect. 5.1.4, we will also evaluate the Adjusted Rand Index (ARI; Hubert & Arabie, 1985) and the number of misclassified units using external information provided as a categorical variable for each dataset. The ARI is a measure of classification accuracy that takes a value of 1 when two partitions perfectly match. A negative ARI value can occur in

Table 3 Model comparison, for the joint distribution of FL1.H and FL2.H, in terms of the number of parameters (#par), log-likelihood, and BIC (multiplied by −1) computed over the set $G \in \{3, 4, 5\}$. Bold values indicate the best (lowest) BIC for each model

Model	G	Log-lik.	#par	BIC
FM-MSN	3	−79,801.82	23	−159,804.80
	4	−76,783.39	31	−153,837.90
	5	−76,728.71	39	**−153,798.40**
FM-MST	3	−78,070.91	24	−156,351.70
	4	−76,634.49	32	**−153,548.80**
	5	−76,605.38	40	−153,560.50
FM-MTIN	3	−78,938.83	20	−158,052.50
	4	−76,684.51	27	**−153,605.00**
	5	−76,659.21	34	−153,615.70
FM-MSEN	3	−78,988.02	17	−158,124.70
	4	−76,738.22	24	**−153,686.30**
	5	−76,712.89	31	−153,696.80
FM-MSGH	3	−78,016.30	29	−156,286.20
	4	−76,918.02	39	**−154,177.10**
	5	−76,893.06	49	−154,214.60
FM-CMSGH	3	−78,821.83	29	−157,897.20
	4	−76,990.95	39	**−154,322.90**
	5	−76,961.25	49	−154,351.00
FM-MMTN	3	−77,920.75	23	−156,042.60
	4	−76,469.97	31	−153,211.00
	5	−76,335.22	39	**−153,011.50**
FM-SDSTIN	3	−78,437.47	29	−157,128.50
	4	−76,519.34	39	**−153,379.70**
	5	−76,489.03	49	−153,406.50
FM-SDSSEN	3	−78,452.36	29	−157158.30
	4	−76,541.81	39	**−153,424.60**
	5	−76,504.73	49	−153,437.90

cases of highly discordant clusterings, indicating that the level of agreement is lower than what would be expected from random clustering.

5.2.1 Flow Cytometric Data

The first application originates from a hematopoietic stem cell transplant (HSCT) experiment, collected by the British Columbia Cancer Agency (Aghaeepour et al., 2013 and Spidlen et al., 2012). Lee and McLachlan (2013) consider a cluster analysis of the HSCT dataset. The dataset contains close to 10,000 observations, each stained with four fluorescent markers. In this example, we focus on a subset of the data in $p = 2$ dimensions, FL1.H and FL2.H, after removing 3551 dead cells with values equal to 0, for a total of $n = 6274$ observations.

Figure 4 presents the scatterplot of the data, with colors and symbols corresponding to the expert's clusters.

As we can see, the scenario is quite challenging because there are several points that lie between clusters, and some points belong to one cluster but appear in others. In these latter

cases, no clustering method can correctly classify them. This should be considered when evaluating the classification results from the fitted models.

In the first clustering phase of our analysis, we fit each candidate model to the data, allowing the number of components, G, to vary from 1 to 5. For brevity, Table 3 focuses on $G \in \{3, 4, 5\}$, as the best BIC value is always found within this range, regardless of the model considered. Overall, the best-performing model according to the BIC is the FM-MMTN with $G = 5$ components, although this result contradicts the true number of clusters. For all the other models, apart from FM-MSN where $G = 5$ is still favored, the BIC instead favors a solution with $G = 4$ mixture components. To understand why the BIC selects $G = 5$ components for the FM-MMTN, it is sufficient to inspect the scatterplot in Fig. 5, which displays the classification obtained under the best-BIC FM-MMTN model.

Roughly speaking, the BIC favors splitting one of the expert's clusters—the one located in the bottom-left region—into two subclusters.

In the second classification phase of the analysis, following Lee and McLachlan (2013), we compare the same models using a fixed number of components, $G = 4$, to assess their classification performance. Table 4 presents the results of this comparison.

The FM-SDSMN distributions follow the FM-MMTN model as the best fit according to the BIC measure, and they rank among the top two models based on the ARI, with the FM-SDSTIN being the best classifier. Notably, finite mixtures of multiple scaled distributions, namely FM-MSGH and FM-CMSGH, perform the worst according to all the criteria considered, despite their flexibility in component-wise skewness and tail behavior. Figure 6 illustrates separate contour lines (for each mixture component) from fitting mixtures of (a) FM-MTIN vs. FM-SDSTIN and (b) FM-MSEN vs. FM-SDSSEN models. The goal is to evaluate whether adding skewness and flexible dimension-wise tail behavior to the mixture components of the FM-MTIN and FM-MSEN models enhances their fitting performance. As expected, the fitted FM-SDSMN models not only result in a lower misclassification rate (refer

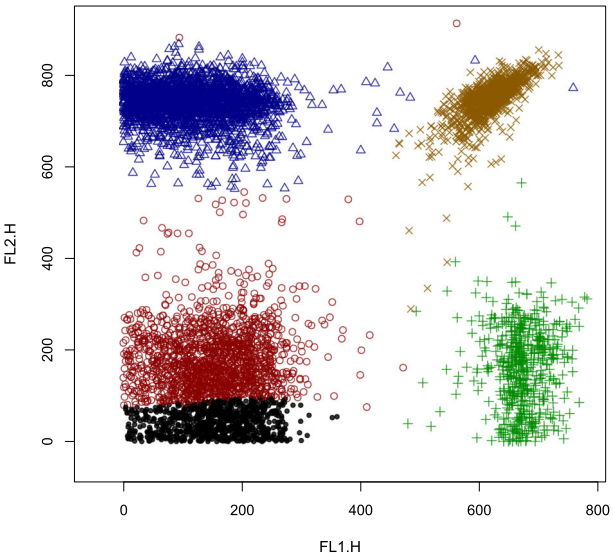

Fig. 5 HSCT data. Scatterplot of FL1.H and FL2.H colored according to the classification of the FM-MMTN model with 5 components

Table 4 HSCT data. Model comparison, for the joint distribution of FL1.H and FL2.H, in terms of the number of parameters (#par), log-likelihood, BIC (multiplied by −1), ARI, and number of misclassified units (#mis) with respect to the expert's clusters. Rankings induced by the considered criteria are also displayed

Model	G	Log-lik.	#par	BIC	Ranking	ARI	Ranking	#mis	Ranking
FM-MSN	4	−76,783.39	31	−153,837.90	7	0.974	5	71	5
FM-MST	4	−76,634.49	32	−153,548.80	4	0.974	5	69	4
FM-MTIN	4	−76,684.51	27	−153,605.00	5	0.975	4	71	5
FM-MSEN	4	−76,738.22	24	−153,686.30	6	0.977	3	64	3
FM-MSGH	4	−76,918.02	39	−154,177.10	8	0.966	8	84	9
FM-CMSGH	4	−76,990.95	39	−154,322.90	9	0.970	7	78	7
FM-MMTN	4	−76,469.97	31	−153,211.00	1	0.969	8	79	8
FM-SDSTIN	4	−76,519.34	39	−153,379.70	2	0.979	1	59	1
FM-SDSSEN	4	−76,541.81	39	−153,424.60	3	0.978	2	61	2

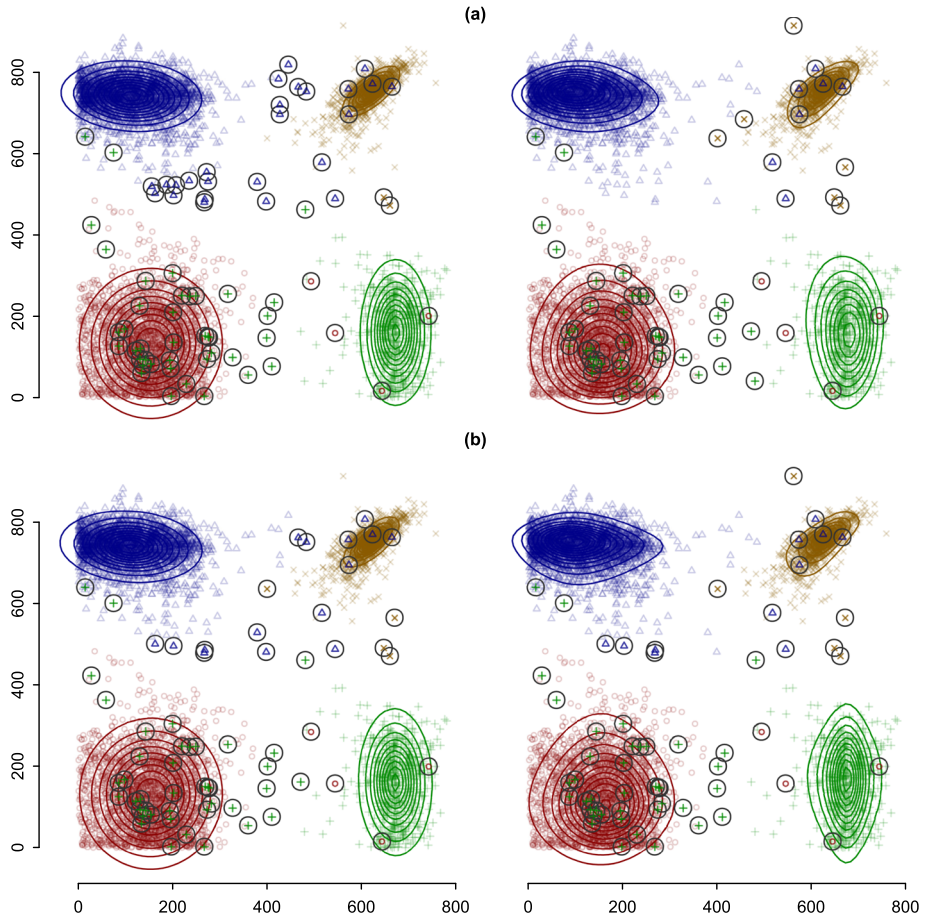

Fig. 6 HSCT data. Scatterplot of FL1.H and FL2.H and fitted contours for **a** FM-MTIN vs. FM-SDSTIN and **b** FM-MSEN vs. FM-SDSSEN models. The black circles indicate misclassified units

to Table 4) but also more effectively capture component-wise asymmetry and tail behavior; in other words, they better fit the data compared to the considered candidate models.

6 Discussion and Future Work

In this paper, we first introduced the novel family of SDSMN distributions. We then explored these models as mixture components within a model-based clustering framework. Two members of the SDSMN family were presented: the skew dimension-wise scaled tail-inflated normal (SDSTIN) and the skew dimension-wise scaled shifted-exponential normal (SDSSEN) distributions, both offering closed-form expressions for the joint density function. The flexibility of these distributions allows them to effectively manage skewness and diverse tail behaviors across each dimension.

We also developed an ECME algorithm for parameter estimation using the ML approach within a complete data framework. Through numerical analyses on four real datasets, we

demonstrated scenarios where the proposed models, both independently and as components in a finite mixture, outperformed well-established models commonly used in the literature.

This approach sets a promising foundation for future developments, potentially integrating a multivariate regression framework by assuming the error term distribution originates from the SDSMN family. Additionally, our methodology has potential applications in various domains, including clusterwise regression and cluster-weighted modeling (Danget al., 2017; Punzo & McNicholas, 2017, and Gallaugher et al.,2022), and linear mixed modeling (Ferreira et al., 2022 and Schumacher et al., 2021). The matrix-variate paradigm is also of interest and could follow a similar path to the work conducted by Gallaugher and McNicholas (2018, 2019).

Moreover, this approach could be extended to handle "asymmetric" missing values, as shown in previous research on finite mixtures of other skew distributions like the multivariate skew normal and multivariate skew-*t* (Lin et al., 2009; Lin & Lin,2011), as well as for finite mixtures of common factor analyzers dealing with incomplete or missing data (Wang, 2013).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00357-026-09542-9.

Acknowledgements This study (for Antonio Punzo) was funded by the European Union - NextGenerationEU, Mission 4, Component 2, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP E63C22002120006). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

Funding Open access funding provided by Università degli Studi di Catania within the CRUI-CARE Agreement.

Data Availability The real datasets used in this manuscript are publicly available as described in the manuscript.

Declarations

Ethical Approval Not applicable.

Conflict of Interest The authors declare no competing interests.

References

Abramowitz, M., & Stegun, I. A. (1968). *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (Vol. 55). US: US Government printing office.

Aghaeepour, N., Finak, G., Consortium, F., Consortium, D., Hoos, H., Mosmann, T. R., Brinkman, R., Gottardo, R., & Scheuermann, R. H. (2013). Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, *10*(3), 228–238.

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology, 95*(3), 631–636.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Andrews, D. F., & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological), 36*(1), 99–102.

Arellano-Valle, R. B., Branco, M. D., & Genton, M. G. (2006). A unified view on skewed distributions arising from selections. *Canadian Journal of Statistics, 34*(4), 581–601.

Azzalini, A. (2016). *'Package 'mnormt"*. The multivariate normal and t distributions (version 1.5-5).

Azzalini, A., & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*(2), 367–389.

Azzalini, A., & Capitanio, A. (2014). *The skew-normal and related families*. US: Cambridge University Press.

Azzalini, A., & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika, 83*(4), 715–726.

Babić, S., Ley, C., & Palangetić, M. (2011), Elliptical symmetry tests in R. arXiv:2011.12560v1.

Babić, S., Palangetić, M., & Ley, C. (2020). Ellipticalsymmetry: Elliptical symmetry tests. R package version 0.1. https://CRAN.R-project.org/package=ellipticalsymmetry.

Bagnato, L., Farcomeni, A., & Punzo, A. (2024). The generalized hyperbolic family and automatic model selection through the multiple-choice lasso. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 17*(1), e11652.

Bagnato, L., & Punzo, A. (2021). Unconstrained representation of orthogonal matrices with application to common principal components. *Computational Statistics, 36*(2), 1177–1195.

Bagnato, L., Punzo, A., & Zoia, M. G. (2017). The multivariate leptokurtic-normal distribution and its application in model-based clustering. *Canadian Journal of Statistics, 45*(1), 95–119.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(7), 719–725.

Böhning, D. (2000), *Computer-assisted analysis of mixtures and applications: Meta-analysis, disease mapping and others, Vol. 81 of monographs on statistics and applied probability*, Chapman & Hall/CRC, London.

Bookstein, F. L. (1997). *Morphometric tools for landmark data*. Geometry and Biology: Cambridge University Press, UK.

Browne, R. P., Bagnato, L., & Punzo, A. (2023). Parsimony and parameter estimation for mixtures of multivariate leptokurtic-normal distributions. *Advances in Data Analysis and Classification*, 1–29.

Burnham, K. P., & Anderson, D. R. (2013), Model selection and inference: A practical information-theoretic approach, Springer, New York. https://books.google.it/books?id=W63hBwAAQBAJ.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261–304.

Cabral, C. R. B., Lachos, V. H., & Prates, M. O. (2012). Multivariate mixture modeling using skew-normal independent distributions. *Computational Statistics & Data Analysis, 56*(1), 126–142.

Celeux, G., Hurn, M., & Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association, 95*(451), 957–970.

Coffey, C. S., & Muller, K. E. (2000). Properties of doubly-truncated gamma variables. *Communications in Statistics-Theory and Methods, 29*(4), 851–857.

Cuesta-Albertos, J. A., Gordaliza, A., & Matrán, C. (1997). Trimmed k-means: An attempt to robustify quantizers. *The Annals of Statistics, 25*(2), 553–576.

Curran-Everett, D. (2018). Explorations in statistics: The log transformation. *Advances in Physiology Education, 42*(2), 343–347.

Dagpunar, J. S. (1978). Sampling of variates from a truncated gamma distribution. *Journal of Statistical Computation and Simulation, 8*(1), 59–64.

Dang, U. J., Punzo, A., McNicholas, P. D., Ingrassia, S., & Browne, R. P. (2017). Multivariate response and parsimony for Gaussian cluster-weighted models. *Journal of Classification, 34*, 4–34.

Dávila, V., Cabral, C., & Zeller, C. (2018). *Finite mixture of skewed distributions*. Springer Briefs in Statistics, Springer International Publishing, Germany. https://books.google.com/books?id=vox5DwAAQBAJ.

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

Feng, C., Wang, H., Lu, N., & Tu, X. M. (2013). Log transformation: Application and interpretation in biomedical research. *Statistics in Medicine, 32*(2), 230–239.

Ferreira, C. S., Bolfarine, H., & Lachos, V. H. (2022). Linear mixed models based on skew scale mixtures of normal distributions. *Communications in Statistics-Simulation and Computation, 51*(12), 7194–7214.

Forbes, F., & Wraith, D. (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights: Application to robust clustering. *Statistics and Computing, 24*(6), 971–984.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal, 41*(8), 578–588.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer.
Gallaugher, M. P., Tomarchio, S. D., McNicholas, P. D., & Punzo, A. (2022). Multivariate cluster weighted models using skewed distributions. *Advances in Data Analysis and Classification*, 1–32.
Gallaugher, M. P., & McNicholas, P. D. (2018). Finite mixtures of skewed matrix variate distributions. *Pattern Recognition, 80*, 83–93.
Gallaugher, M. P., & McNicholas, P. D. (2019). Three skewed matrix variate distributions. *Statistics & Probability Letters, 145*, 103–109.
Gordon, C. C., Blackwell, C. L., Bradtmiller, B., Parham, J. L., Barrientos, P., Paquette, S. P., Corner, B. D., Carson, J. M., Venezia, J. C., Rockwell, B. M., Mucher, M., & Kristensen, S. (2014). *2012 anthropometric survey of U.S. army personnel: Methods and summary statistics*, Technical Report NATICK/TR-15/007, U.S. Army Natick Soldier Research, Development and Engineering Center, Natick, Massachusetts.
Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218.
Jajuga, K., & Papla, D. (2006). Copula functions in model based clustering. In *From Data and Information Analysis to Knowledge Engineering: Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation eV*, University of Magdeburg, March 9–11, 2005', Springer, (pp. 606–613).
Kan, R., & Robotti, C. (2017). On moments of folded and truncated multivariate normal distributions. *Journal of Computational and Graphical Statistics, 26*(4), 930–934.
Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.
Klingenberg, C. P. (1996). Multivariate allometry. In L. E. Marcus, M. Corti, A. Loy, G. J. P. Naylor, & D. E. Slice (Eds.), *Advances in morphometrics* (pp. 23–49). US: Springer.
Kosmidis, I., & Karlis, D. (2016). Model-based clustering using copulas with applications. *Statistics and Computing, 26*, 1079–1099.
Lee, L.-F. (1983). The determination of moments of the doubly truncated multivariate normal Tobit model. *Economics Letters, 11*(3), 245–250.
Lee, S. X., & McLachlan, G. J. (2013). Model-based clustering and classification with non-normal mixture distributions. *Statistical Methods & Applications, 22*(4), 427–454.
Ley, C., & Paindaveine, D. (2010). Multivariate skewing mechanisms: A unified perspective based on the transformation approach. *Statistics & Probability Letters, 80*(23–24), 1685–1694.
Li, J., & Maitra, R. (2021). Model selection strategies for determining the number of clusters in cluster analysis. *Journal of Classification, 38*, 552–580.
Lin, T. I., Lee, J. C., & Yen, S. Y. (2007). Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, 909–927.
Lin, T. I. (2009). Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis, 100*(2), 257–265.
Lin, T. I. (2010). Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing, 20*(3), 343–356.
Lin, T. I., Ho, H. J., & Chen, C. L. (2009). Analysis of multivariate skew normal models with incomplete data. *Journal of Multivariate Analysis, 100*(10), 2337–2351.
Lin, T. I., Lee, J. C., & Hsieh, W. J. (2007). Robust mixture modeling using the skew t distribution. *Statistics and Computing, 17*(2), 81–92.
Lin, T. I., & Lin, T.-C. (2011). Robust statistical modelling using the multivariate skew t distribution with complete and incomplete data. *Statistical Modelling, 11*(3), 253–277.
Liu, C., & Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika, 81*(4), 633–648.
Mahdavi, A., Amirzadeh, V., Jamalizadeh, A., & Lin, T. I. (2021). Maximum likelihood estimation for scale-shape mixtures of flexible generalized skew normal distributions via selection representation. *Computational Statistics, 36*, 2201–2230.
Mahdavi, A., Amirzadeh, V., Jamalizadeh, A., & Lin, T. I. (2021). A multivariate flexible skew-symmetric-normal distribution: Scale-shape mixtures and parameter estimation via selection representation. *Symmetry, 13*(8), 1343.
Manzotti, A., Pérez, F. J., & Quiroz, A. J. (2002). A statistic for testing the null hypothesis of elliptical symmetry. *Journal of Multivariate Analysis, 81*(2), 274–285.
McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker.
McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: John Wiley & Sons.
McNeil, A., Frey, R., & Embrechts, P. (2005). *Quantitative risk management: Concepts*. Techniques and Tools: Princeton Series in Finance, Princeton University Press, US.
McNicholas, P. D. (2016). *Mixture model-based classification*. Boca Raton: Chapman & Hall/CRC Press.

Melnykov, Y., Zhu, X., & Melnykov, V. (2021). Transformation mixture modeling for skewed data groups with heavy tails and scatter. *Computational Statistics, 36*, 61–78.
Meng, X.-L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika, 80*(2), 267–278.
Misra, R. D. (1940). On the stability of crystal lattices. ii. *Mathematical Proceedings of the Cambridge Philosophical Society, 36*(2), 173–182.
Park, B.-K.D., Corner, B. D., Hudson, J. A., Whitestone, J., Mullenger, C. R., & Reed, M. P. (2021). A three-dimensional parametric adult head model with representation of scalp shape variability under hair. *Applied Ergonomics, 90*, 103239.
Pelleg, D., Moore, A. et al. (2000). X-means: Extending K-means with efficient estimation of the number of clusters. In *ICML'00'*. Citeseer, (pp. 727–734).
Philippe, A. (1997). Simulation of right and left truncated gamma distributions by mixtures. *Statistics and Computing, 7*, 173–181.
Pimentel, R. A. (1979). *Morphometrics, the multivariate analysis of biological data*. New York: Kendall/Hunt Pub. Co.
Punzo, A., Mazza, A., & McNicholas, P. D. (2018b). *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., & Bagnato, L. (2020). Allometric analysis using the multivariate shifted exponential normal distribution. *Biometrical Journal, 62*(6), 1525–1543.
Punzo, A., & Bagnato, L. (2021). The multivariate tail-inflated normal distribution and its application in finance. *Journal of Statistical Computation and Simulation, 91*(1), 1–36.
Punzo, A., & Bagnato, L. (2022). Dimension-wise scaled normal mixtures with application to finance and biometry. *Journal of Multivariate Analysis, 191*, 105020.
Punzo, A., & Bagnato, L. (2022). Multiple scaled symmetric distributions in allometric studies. *The International Journal of Biostatistics, 1*(18), 219–242.
Punzo, A., Mazza, A., & McNicholas, P. D. (2018). ContaminatedMixt: An R package for fitting parsimonious mixtures of multivariate contaminated normal distributions. *Journal of Statistical Software, 85*(10), 1–25.
Punzo, A., & McNicholas, P. D. (2017). Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. *Journal of Classification, 34*, 249–293.
Punzo, A., & Tortora, C. (2021). Multiple scaled contaminated normal distribution and its application in clustering. *Statistical Modelling, 21*(4), 332–358.
Reyment, R. A. (1991). *Multidimensional palaeobiology*. New York: Pergamon Press.
Schlattmann, P. (2010). *Medical applications of finite mixture models, statistics for biology and health*, Springer, Berlin Heidelberg, Berlin. https://books.google.com/books?id=S6JBYgEACAAJ.
Schumacher, F. L., Lachos, V. H., & Matos, L. A. (2021). Scale mixture of skew-normal linear mixed models with within-subject serial dependence. *Statistics in Medicine, 40*(7), 1790–1810.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.
Scrucca, L. (2019). A transformation-based approach to gaussian mixture density estimation for bounded data. *Biometrical Journal, 61*(4), 873–888.
Spidlen, J., Breuer, K., Rosenberg, C., Kotecha, N., & Brinkman, R. R. (2012). Flowrepository: A resource of annotated flow cytometry datasets associated with peer-reviewed publications. *Cytometry Part A, 81*(9), 727–731.
Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62*(4), 795–809.
Tomarchio, S. D., Bagnato, L., & Punzo, A. (2021). *SenTinMixt: Parsimonious mixtures of MSEN and MTIN distributions*. R package version 1.0.0. https://cran.r-project.org/web/packages/MSclust/index.html.
Tomarchio, S. D., Bagnato, L., & Punzo, A. (2022). Model-based clustering via new parsimonious mixtures of heavy-tailed distributions. *AStA Advances in Statistical Analysis, 106*(2), 315–347.
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.
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. *Journal of Statistical Software, 98*, 1–24.
Tortora, C., Franczak, B. C., Browne, R. P., & McNicholas, P. D. (2019). A mixture of coalesced generalized hyperbolic distributions. *Journal of Classification, 36*(1), 26–57.
Valeriano, K. A., Matos, L. A., Morales, C. G., & Valeriano, M. K. A. (2024). *Relliptical: The truncated elliptical family of distributions*. R package Version 1.3.0 (2024–02-07). https://CRAN.R-project.org/package=relliptical.

Valeriano, K. A., Galarza, C. E., & Matos, L. A. (2023). Moments and random number generation for the truncated elliptical family of distributions. *Statistics and Computing, 33*(1), 32.

Vrac, M., Billard, L., Diday, E., & Chédin, A. (2012). Copula analysis of mixture models. *Computational Statistics, 27*, 427–457.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*(2), 228.

Wang, W.-L. (2013). Mixtures of common factor analyzers for high-dimensional data with missing information. *Journal of Multivariate Analysis, 117*, 120–133.

Wedel, M., & Kamakura, W. (2012). Market segmentation: Conceptual and methodological foundations. In *International Series in Quantitative Marketing*, Springer, US. https://books.google.com/books?id=XxLaBwAAQBAJ.

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.

Zhu, X., & Melnykov, V. (2023). ManlyMix: An R package for model-based clustering with manly mixture models. R package version 0.1.15.1.

Zhu, X., & Melnykov, V. (2018). Manly transformation in finite mixture modeling. *Computational Statistics & Data Analysis, 121*, 190–208.