Journal of Computational and Applied Mathematics 437 (2024) 115433

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

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

Clustering asymmetrical data with outliers: Parsimonious mixtures of contaminated mean-mixture of normal distributions

Mehrdad Naderi [a], Mehdi Jabbari Nooghabi [a,b,*]

[a] Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran
[b] Department of Mathematical Sciences, University of Copenhagen, DK-2100 Copenhagen Ø, Denmark

ARTICLE INFO

Article history:
Received 6 February 2022
Received in revised form 28 February 2023

Keywords:
Finite mixture model
Contaminated mean-mixture of normal distributions
EM-type algorithm
Eigenvalue decomposition
Outliers detection

ABSTRACT

Mixture modeling has emerged as a statistical tool to perform unsupervised model-based clustering for heterogeneous data. A framework of using contaminated mean-mixture of normal distributions as the components of the mixture model is designed to accommodate asymmetric data with outliers. Fourteen parsimonious variants of the postulated model are introduced by employing an eigenvalue decomposition of the component scale matrices. Simultaneously clustering and outliers detection is an outstanding advantage of the proposed model in analyzing non-normally distributed data. A computationally feasible and flexible EM-type algorithm is outlined for obtaining maximum likelihood parameter estimates. Moreover, the score vector and empirical information matrix for calculating asymptotic standard errors of the parameter estimates are derived by offering an information-based approach. The applicability of the proposed method is demonstrated through the analysis of simulated and real datasets with varying proportions of outliers.

1. Introduction

Mixture model is a famous statistical model-based approach for supervised and unsupervised learning. The initial interest in the mixture models can be found in the work of Redner and Walker [1], who introduced the normal mixture (N-MIX) model as a promising exploratory analysis tool. Since then, cluster analysis aiming to identify the group structure of data, groping data such that the similarity within groups and the dissimilarity between them be maximized, using the N-MIX model has received considerable attention. It has extensively been considered in diverse areas of sciences such as pattern recognition, computer vision, signal and image processing, and machine learning, over the past three decades. However, the clustering task using the N-MIX has two challenges: (1) it is not robust to accommodate data that departure from normality (data with non-normal features such as asymmetry, heavy tails, and outliers), and (2) when the dimension of data is larger than the number of observations, the N-MIX model might produce misleading inferential results regarding singular or near-singular estimates of the component scale matrices. To address the first potential obstacle, various variants of the mixture model have recently flourished by exploiting the non-Gaussian (skew and heavy-tailed) distributions. The interested reader can see Punzo and McNicholas [2], Naderi et al. [3], Morris et al. [4], Farcomeni and Punzo [5], Naderi et al. [6] and references therein, as few recently published research articles. To overcome the second

* Corresponding author at: Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran.
E-mail addresses: mehrdad.naderi@ymail.com (M. Naderi), jabbarinm@um.ac.ir, mjn@math.ku.dk (M.J. Nooghabi).

https://doi.org/10.1016/j.cam.2023.115433
0377-0427/

issue, Celeux and Govaert [7] introduced the parsimonious variants of the N-MIX model by employing an eigenvalue decomposition of the component scale matrices. As an alternative estimation process, Benaglia et al. [8] exploited the EM-like algorithm for the semi-parametric mixture model and showed that it can be more flexible than other existing algorithms, and may be applicable even for the non-identifiable model.

The family of skew-normal (SN) distributions, proposed by Azzalini [9], is one of the baseline platforms which plays a substantial role in constructing an efficient statistical model for analyzing skewed data. For instance, Vilca et al. [10] considered the multivariate SN distribution to propose the family of skew-normal generalized hyperbolic distributions which provides more flexibility in modeling data with strong skewness and heavy-tailed features. Moreover, clustering using the univariate as well as multivariate mixture models by considering the class of scale-mixtures of skew-normal (SMSN-MIX) distributions was discussed in Basso et al. [11], Cabral et al. [12]. Although the SMSN-MIX model provides a flexible platform for analyzing skewed and heavy-tailed datasets, it might not detect and model "mild" and "gross" outliers simultaneously [5]. The term "mild outliers" refers to the data points that depart from the distribution within a cluster, but if the underlying mixture components are skew with heavier tails than normal they could be fitted appropriately. On the other hand, the term "gross outliers" is used for observations that are far away from any of the components. In order to simultaneously detect and model mild and gross outliers, Morris et al. [4] introduced a new multivariate mixture model by exploiting the contaminated shifted asymmetric Laplace distributions.

Recently, Negarestani et al. [13] extended the SN distribution to introduce a new class of skew models as an alternative to the baseline SN distribution. Calling the mean-mixture of normal (MMN) model, Negarestani et al. [13] showed that the MMN distribution not only inherits mathematical and computational aspects of the SN distribution but also it can provide a more flexible platform for statistical analysis. Hashemi et al. [14] designed a new factor analyzer model based on the class of MMN distributions and investigate its theoretical as well as practical properties. They showed through a simulation study that the class of MMN distributions outperforms the skew-t distribution whenever the degree of freedom increases. Numerical results of Hashemi et al. [14] confirmed that the calibration of MMN distribution to the data is much faster than some existing models. Moreover, Naderi et al. [15] postulated matrix variate MMN distribution and showed that the class of MMN distributions offers different orientations compared with the family of the mean–variance mixture of normal (MVMN) distributions [16].

Owing to the properties of the MMN class of distributions, we consider mixtures of contaminated mean-mixture of normal (CMMN) distributions. In short, the main advances introduced in this contribution are: (1) proposing a flexible mixture model by using the MMN class of distributions for analyzing skew-distributed datasets with mild and gross outliers; (2) introducing 14 parsimonious variants of the proposed model by employing an eigenvalue decomposition of the component scale matrices; (3) developing a unified framework for maximum likelihood (ML) estimation by implementing a feasible expectation–maximization (EM; Dempster et al. [17]) type algorithm with almost all closed-form expressions for the E- and M-steps; (4) calculating asymptotic standard errors of parameter estimates by offering a unified information-based approach.

The remainder of the paper is organized as follows. Section 2 will give a brief sketch of the MMN and CMMN distributions with a concentration on the model formulation and some preliminary theoretical properties. The mixture of CMMN (CMMN-MIX) distributions and its parameter estimation via an EM-type algorithm are developed in Section 3. Moreover, Section 3 discusses introducing 14 parsimonious variants of the proposed CMMN-MIX distributions and offers an information-based approach for calculating asymptotic standard errors. In Section 4, we address some computational issues regarding the EM starting points and stopping criteria, model selection and clustering performance measures, and procedure of outliers detection. Section 5 reports the results of simulation studies. The illustration of the proposed CMMN-MIX distribution on the real-world data is finally presented in Section 6.

2. Contaminated mean-mixture of normal distributions

This section mainly discusses the formulation of the MMN family of distributions. We also present the way of extension of this family to the contaminated framework for accommodating data in the presence of outliers. Firstly, we give some notations. Throughout this paper, let $\phi_p(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ stand for the probability density function (pdf) of a p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and scale (covariance) matrix $\boldsymbol{\Sigma}$, denoted by $\mathrm{N}_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Denote also the cumulative distribution function of a univariate standard normal by $\Phi(\cdot)$. The notation "$\sim$" represents "follows the distribution" or "follows the distribution with pdf (cdf)', and "$\overset{d}{=}$" is used for the "equality in distribution". Finally, we will use the notations $U\sim \mathrm{TN}(\mu,\sigma^2;(a,b))$ if U follows a truncated univariate normal distribution lying within a truncated interval (a,b), $U\sim\mathrm{Exp}(\nu)$ if U has an exponential distribution with shape parameter ν, and $U\sim\mathrm{Wei}(\tau,\nu)$ if U follows a Weibull distribution with scale and shape parameters τ and ν, respectively.

2.1. Formulation of the mean-mixture of normal distributions

A p-variate random vector $\boldsymbol{X}\in\mathbb{R}^p$ follows an MMN distribution if it admits the stochastic representation

$$\boldsymbol{X}\overset{d}{=}\boldsymbol{\mu}+U\boldsymbol{\lambda}+\boldsymbol{Z}, \tag{1}$$

where $\boldsymbol{\mu}, \boldsymbol{\lambda} \in \mathbb{R}^p$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ respectively are the location, shape and scale parameters. Here, $\boldsymbol{Z} \sim \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $U \sim H(\cdot; \boldsymbol{\tau})$ independently, where the arbitrary cdf $H(\cdot; \boldsymbol{\tau})$ is indexed by a vector (scalar) parameter $\boldsymbol{\tau}$. The random vector $\boldsymbol{X}$ in (1) can also be represented by the hierarchical formulation as

$$X|U = u \sim \mathrm{N}_p(\boldsymbol{\mu} + u\boldsymbol{\lambda}, \boldsymbol{\Sigma}), \qquad U \sim H(u; \boldsymbol{\tau}). \tag{2}$$

It then leads to obtain the pdf, mean and covariance matrix of $\boldsymbol{X}$ as

$$f_{\mathrm{MMN}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau}) = \int_{-\infty}^{\infty} \phi_p(\boldsymbol{x}; \boldsymbol{\mu} + u\boldsymbol{\lambda}, \boldsymbol{\Sigma})\, dH(u; \boldsymbol{\tau}),$$
$$\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu} + \mathrm{E}(U)\boldsymbol{\lambda} \qquad \text{and} \qquad Cov(\boldsymbol{X}) = \boldsymbol{\Sigma} + Var(U)\boldsymbol{\lambda}\boldsymbol{\lambda}^\top. \tag{3}$$

We will use the notation $\boldsymbol{X} \sim \mathrm{MMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$ if $\boldsymbol{X}$ has pdf (3). Notice that when $H(\cdot; \boldsymbol{\tau})$ degenerates with $u = 1$, or alternatively when $\boldsymbol{\lambda}$ tends to zero, the MMN distribution tends to $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. However, conditionally on $U = u$, one can see from (2) that the mean is not fixed for all members of the population as $\boldsymbol{\lambda}$ departs from zero. This desirable property helps researchers to construct suitable distributions for modeling data with stronger skewness and heavier tails than the normal distribution. As discussed in Naderi et al. [15], this property allows us to postulate models with different contour plots compared to the NMVM distributions. If U is distributed symmetrically and/or elliptically, the resulting distribution of $\boldsymbol{X}$ is also symmetric and/or elliptical. The class of MMN distributions contains symmetric models if $H(\cdot; \boldsymbol{\nu})$ be a symmetric density function. However, a more flexible and skew-type one can be obtained based on the assumption that U follows any asymmetric distribution, preferably a positive support model such as the truncated-normal, exponential, and Weibull distributions.

In what follows, three special cases of the general MMN class of distributions are discussed by considering different specifications on the distribution of U. For each case, the conditional distribution of U given $\boldsymbol{X} = \boldsymbol{x}$ along with the conditional expectations, needed for the implementation of the EM algorithm, is also derived. It should be mentioned that in developing a new family of distributions for modeling unimodal data, at last five parameters are commonly accepted for flexibility, and justification for more is extremely hard [18]. Since there are two parameters in the construction of contaminated distributions, discussed in the next sub-section, and we have location, shape, and scale parameters in the model (1), the indexing parameter $\boldsymbol{\tau}$ is considered to be a fixed value. In this regard, either we can focus on the well-known cases, e.g restricted skew-normal distribution, or we can formulate a new distribution having wide ranges of skewness and kurtosis (remark 2.1 of Negarestani et al. [13]).

- Convolution with truncated normal distribution. If the random variable U in (1) follows the truncated standard normal distribution lying within a truncated interval $(0, \infty)$, the pdf of restricted skew-normal (rSN) distribution obtained form (3) is

$$f_{\mathrm{rSN}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = 2\phi_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Omega})\Phi\left(\frac{\boldsymbol{\lambda}^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})}{\sqrt{1 - \boldsymbol{\lambda}^\top \boldsymbol{\Omega}^{-1}\boldsymbol{\lambda}}}\right), \quad \boldsymbol{x} \in \mathbb{R}^p, \tag{4}$$

  where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\lambda}\boldsymbol{\lambda}^\top$. The notations $\boldsymbol{X} \sim \mathrm{rSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ will be used if $\boldsymbol{X}$ has pdf (4). Since the truncated normal distribution belongs to the elliptical class of distributions, the rSN distribution is a skew elliptically-contoured model.

Lemma 1. *If* $U \sim TN\left(\mu, \sigma^2; (0, \infty)\right)$, *then*

$$E(U) = \mu + \sigma\frac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)}, \qquad E\left(U^r\right) = \mu E\left(U^{r-1}\right) + \sigma^2(r-1)E\left(U^{r-2}\right), \quad r = 2, 3, \ldots.$$

Proposition 1. *Let* $\boldsymbol{X} \sim rSN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ *and* $U \sim TN(0, 1; (0, \infty))$. *Then, U conditionally on* $\boldsymbol{X} = \boldsymbol{x}$, *denoted by* $U_{\boldsymbol{x}}$, *is distributed by* $TN\left(\xi, \omega^2; (0, \infty)\right)$, *where* $\xi = \boldsymbol{\lambda}^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ *and* $\omega^2 = 1 - \boldsymbol{\lambda}^\top \boldsymbol{\Omega}^{-1}\boldsymbol{\lambda}$. *Moreover, the first two moments of* $U_{\boldsymbol{x}}$ *obtained from Lemma 1 are*

$$E(U_{\boldsymbol{x}}) = \xi + \omega\frac{\phi(\xi/\omega)}{\Phi(\xi/\omega)} \qquad \text{and} \qquad E\left(U^r\right) = \xi E(U_{\boldsymbol{x}}) + \omega^2.$$

- Convolution with exponential distribution. The pdf of exponentiated MMN (MMNE) distribution, derived from (1) with $U \sim \mathrm{Exp}(1)$, takes the form

$$f_{\mathrm{MMNE}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = \frac{\sqrt{2\pi}}{\eta}\exp\left\{\frac{A_E^2}{2}\right\}\phi_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi(A_E), \quad \boldsymbol{x} \in \mathbb{R}^p, \tag{5}$$

  where $\eta^2 = \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}$, and $A_E = \eta^{-1}\left[\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) - 1\right]$. We will use the notation $\boldsymbol{X} \sim \mathrm{MMNE}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ in case the random vector $\boldsymbol{X}$ follows the MMNE distribution. It is also clear that the MMNE distribution is a skew non-elliptically contoured model since the exponential distribution is a non-elliptical model. Moreover, the pdf of MMNE distribution is a log-concave function.

Proposition 2. *If* $\boldsymbol{X} \sim MMNE_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ *and* $U \sim Exp(1)$, *then* $U_{\boldsymbol{x}} \sim TN\left(A_E \eta^{-1}, \eta^{-2}; (0, \infty)\right)$. *Furthermore, we have*

$$E(U_{\boldsymbol{x}}) = A_E \eta^{-1} + \eta^{-1} \frac{\phi(A_E)}{\Phi(A_E)}, \qquad E(U^r) = A_E \eta^{-1} E(U_{\boldsymbol{x}}) + \eta^{-2}.$$

- Convolution with Weibull distribution. The mixed-Weibull MMN (MMNW) distribution, denoted by $\boldsymbol{X} \sim \text{MMNW}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, arises from (1) when $U \sim \text{Wei}(1, 2)$ with $\text{E}(U) = \sqrt{\pi}/2$. The pdf of $\boldsymbol{X}$ is given by

$$f_{\text{MMNW}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = \frac{2\sqrt{2\pi}}{\eta_{WE}^2} \exp\left\{\frac{A_{WE}^2}{2}\right\} \phi_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \left(A_{WE} \Phi(A_{WE}) + \phi(A_{WE})\right), \quad \boldsymbol{x} \in \mathbb{R}^p, \tag{6}$$

  where $\eta_{WE}^2 = \eta^2 + 2$, and $A_{WE} = \eta_{WE}^{-1}\left[\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right]$. The MMNW model is a non-elliptically contoured distribution with a log-concave pdf (since the Weibull distribution with shape parameter greater than 1 has a log-concave pdf).

Proposition 3. *Let* $\boldsymbol{X} \sim MMNW_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ *and* $U \sim Wei(1, 2)$. *Then,* $U_{\boldsymbol{x}}$ *has pdf*

$$f_{U_{\boldsymbol{x}}}(u) = \frac{u \eta_{WE}}{A_{WE} \Phi(A_{WE}) + \phi(A_{WE})} \phi\left(u; A_{WE} \eta_{WE}^{-1}, \eta_{WE}^{-2}\right), \qquad u > 0.$$

For any $\boldsymbol{x} \in \mathbb{R}^p$ *and* $r = 1, 2, \ldots$, *we have also*

$$E(U_{\boldsymbol{x}}^r) = \frac{\eta_{WE} \Phi(A_{WE})}{A_{WE} \Phi(A_{WE}) + \phi(A_{WE})} E(V^{r+1}),$$

where $V \sim TN\left(A_{WE} \eta_{WE}^{-1}, \eta_{WE}^{-2}; (0, \infty)\right)$.

2.2. Contaminated mean-mixture of normal distributions

We say the random vector $\boldsymbol{X}$ has the contaminated mean-mixture of normal (CMMN) distribution if it is generated by the stochastic representation

$$\boldsymbol{X} \overset{d}{=} \boldsymbol{\mu} + \sqrt{W}\left(U\boldsymbol{\lambda} + \boldsymbol{Z}\right),$$

where three independent variables are $\boldsymbol{Z} \sim \text{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, $U \sim H(u)$, and W takes 1 with probability γ and states ν with probability $1 - \gamma$. Therefore, the pdf of $\boldsymbol{X}$ can be obtained as

$$f_{\text{CMMN}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu, \gamma) = \gamma f_{\text{MMN}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) + (1 - \gamma) f_{\text{MMN}}(\boldsymbol{x}; \boldsymbol{\mu}, \nu\boldsymbol{\Sigma}, \sqrt{\nu}\boldsymbol{\lambda}), \qquad \gamma \in (0.5, 1), \nu > 1. \tag{7}$$

In the rest of the paper, the notation $\boldsymbol{X} \sim \text{CMMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu, \gamma)$ shall be used if the random vector $\boldsymbol{X}$ has pdf (7). The pdf of three special cases of the CMMN distribution, namely contaminated rSN (CrSN), contaminated MMNE (CMMNE), and MMNW (CMMNW) distributions can respectively be obtained by plugging the pdfs (4)–(6) into (7). It is interesting to highlight that the CMMN distribution contains two parameters, ν and γ, more than the MNN distribution. It is clear from (7) that the pdf of CMMN distribution is a mixture of two pdfs of the MMN model with equal location parameters and proportional covariance matrix and shape parameters to the contamination factor ν. The pdf with lower dispersion fits the "ordinary" or "good" observations whereas the other one accommodates "mild" and/or "gross" outliers ("bad data"). Therefore, contamination factor ν controls the increase in variability due to the presence of outliers. Note also the additional parameter γ in (7) denotes the rate of good points. Plugging the scale matrix and shape parameters of the contaminated part into Eq. (3), we have

$$\text{E}(\boldsymbol{X}) = \boldsymbol{\mu} + \sqrt{\nu}\text{E}(U)\boldsymbol{\lambda} \qquad \text{and} \qquad Cov(\boldsymbol{X}) = \nu\left(\boldsymbol{\Sigma} + Var(U)\boldsymbol{\lambda}\boldsymbol{\lambda}^\top\right).$$

which shows that the covariance matrix of the bad data points is inflated by the contamination factor ν as compared to the covariance matrix in (3) for good data. It indicates that the formulation of CMMN distribution is analogous to the works of Punzo and McNicholas [2], Cabral et al. [12]. Moreover, similar to Cabral et al. [12], the component for good and bad points in the CMMN distribution do not share an identical mode. It should also be note that $\gamma \in (0, 1)$ in model (7) is also acceptable. However, the constrain $\gamma \in (0.5, 1)$ is considered to avoid switching issue between good and bad data, and to have a robust inference.

Theorem 1. *If* $\boldsymbol{X} \sim CMMN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu, \gamma)$, *then the mean, covariance matrix and moment generating function of the* $\boldsymbol{X}$ *are*

$$\text{E}(\boldsymbol{X}) = \boldsymbol{\mu} + \boldsymbol{\lambda}E(U)(\gamma + \sqrt{\nu}(1 - \gamma)), \qquad \text{Cov}(\boldsymbol{X}) = (\boldsymbol{\Sigma} + Var(U)\boldsymbol{\lambda}\boldsymbol{\lambda}^\top)(\gamma + \nu(1 - \gamma)),$$
$$\text{M}_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left\{\boldsymbol{t}^\top\boldsymbol{\mu}\right\}\left(\gamma \exp\left\{\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}\right\}\text{M}_U(\boldsymbol{t}^\top\boldsymbol{\lambda}; \boldsymbol{\tau}) + (1 - \gamma)\exp\left\{\nu\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}\right\}\text{M}_U(\sqrt{\nu}\boldsymbol{t}^\top\boldsymbol{\lambda})\right),$$

where $\text{M}_U(\cdot)$ *denotes the moment generating function of* U.

Theorem 2. *The CMMN distribution is closed under linear transformation, i.e. if* $\boldsymbol{X} \sim CMMN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu, \gamma)$, *then for any full rank matrix* $\boldsymbol{L} \in \mathbb{R}^{q \times p}$, $1 \le q \le p$, *the random vector* $\boldsymbol{LX}$ *is distributed by* $\boldsymbol{X} \sim CMMN_p(\boldsymbol{L\mu}, \boldsymbol{L\Sigma L}^\top, \boldsymbol{L\lambda}, \nu, \gamma)$.

3. Mixtures of contaminated mean-mixture of normal distributions

3.1. Model specification

Let $\mathcal{X} = \left(\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top\right)$ be a matrix of n independent random vectors identically come from a CMNN-MIX distributions with pdf

$$f(\boldsymbol{x}_i;\, \boldsymbol{\Theta}) = \sum_{j=1}^{G} \pi_j f_{\text{CMMN}}(\boldsymbol{x}_i;\, \boldsymbol{\mu}_j,\, \boldsymbol{\Sigma}_j,\, \boldsymbol{\lambda}_j,\, \nu_j,\, \gamma_j), \qquad i = 1, \ldots, n, \tag{8}$$

where π_j denotes the mixing proportions subjected to $\pi_1 + \cdots + \pi_G = 1$, and for $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j,\, \boldsymbol{\Sigma}_j,\, \boldsymbol{\lambda}_j,\, \nu_j,\, \gamma_j)$ the set of entire free parameters is $\boldsymbol{\Theta} = \left\{\pi_j, \boldsymbol{\theta}_j;\ j = 1, \ldots, G\right\}$. We provide a discussion about the identifiability issue here since it represents a regularity condition for the asymptotic theory to hold for the ML estimator. As is common in the mixture models, model (8) is affected by two sources of non-identifiability. Specifically, model (8) is invariant under label switching (relabeling the distributions of the mixture) and under potential overfitting due to empty or equal classes of the mixture. To cope with this difficulty, some constraints have been imposed on the parameter space. Recommended by Richardson and Green [19], we assume that $\pi_j > 0$ and $\pi_1 < \pi_2 < \cdots < \pi_G$.

For each row of $\mathcal{X}$, it is convenient to introduce a missing allocation random vector $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iG})^\top$ to identify the group membership of $\boldsymbol{X}_i$. The allocation indicators $\boldsymbol{z}_i$ is a binary-valued vector whose element z_{ij} equals to one if the individual $\boldsymbol{X}_i$ belongs to the jth component and zero otherwise. Therefore, it is assumed that $\boldsymbol{z}_i$ has a one-trail multinomial distribution with prior probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$, denoted by $\boldsymbol{z}_i \sim \mathcal{M}(1; \pi_1, \ldots, \pi_G)$. For outlier detection purposes, one can also introduce a binary random variable b_{ij} as the outlying label that takes 1 if $\boldsymbol{X}_i$ in group j comes from the good component and $b_{ij} = 0$ if $\boldsymbol{X}_i$ in group j comes from the contaminated component. So, conditionally on $z_{ij} = 1$, the random variable b_{ij} has the Bernoulli distribution with success probability γ_j, denoted by $b_{ij}|z_{ij} = 1 \sim B(1, \gamma_j)$. These assumptions lead to the four-level hierarchical representation of the CMMN-MIX model (8) as

$$\begin{aligned}
\boldsymbol{X}_i|(u_i, z_{ij} = 1, b_{ij}) &\sim \begin{cases} \mathrm{N}_p(\boldsymbol{\mu}_j + u_i\boldsymbol{\lambda}_j,\, \boldsymbol{\Sigma}_j) & \text{if} \quad b_{ij} = 1; \\ \mathrm{N}_p(\boldsymbol{\mu}_j + u_i\sqrt{\nu_i}\boldsymbol{\lambda}_j,\, \nu_i\boldsymbol{\Sigma}_j) & \text{if} \quad b_{ij} = 0, \end{cases} \\
U_i|(z_{ij} = 1, b_{ij}) &\sim h(u_i), \\
b_{ij}|z_{ij} = 1 &\sim B(1, \gamma_j), \\
Z_i &\sim \mathcal{M}(1, \pi_1, \ldots, \pi_G),
\end{aligned}$$

where $h(u_i)$ denotes the pdf of U_i.

Let regard the observed data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ as being incomplete, where $\boldsymbol{u} = (u_1, \ldots, u_n)$ refer to the unobservable latent points, and the component allocation indicators $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ and the matrix of outlying labels $\boldsymbol{b} = \left(b_{ij}\right)$ are the missing data. Consequently, the complete-data log-likelihood function for the unknown parameter $\boldsymbol{\Theta}$ associated with the complete data $\boldsymbol{x}_c = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{u}, \boldsymbol{z}, \boldsymbol{b}\}$ is given by

$$\begin{aligned}
\ell(\boldsymbol{\Theta}|\boldsymbol{x}_c) = &\sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij} \log \pi_j + \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij}\left[b_{ij}\log\gamma_j + (1 - b_{ij})\log(1 - \gamma_j)\right] \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij}b_{ij}\log\left(\phi_p(\boldsymbol{x}_i;\, \boldsymbol{\mu}_j + u_i\boldsymbol{\lambda}_j,\, \boldsymbol{\Sigma}_j)h(u_i)\right) + \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij}(1 - b_{ij})\log\left(\phi_p(\boldsymbol{x}_i;\, \boldsymbol{\mu}_j + u_i\sqrt{\nu_j}\boldsymbol{\lambda}_j,\, \nu_j\boldsymbol{\Sigma}_j)h(u_i)\right).
\end{aligned} \tag{9}$$

3.2. Parameter estimation through an EM-type algorithm

For fitting the CMMN-MIX model to the observed data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, we need to compute the ML estimate of the parameters by directly maximizing the log-likelihood function

$$\ell(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{G} \pi_j f_{\text{CMMN}}(\boldsymbol{x}_i;\, \boldsymbol{\mu}_j,\, \boldsymbol{\Sigma}_j,\, \boldsymbol{\lambda}_j,\, \nu_j,\, \gamma_j)\right). \tag{10}$$

However, the optimization of the log-likelihood function (10) is complicated and a numerical strategy should be developed. The EM algorithm is a well-known approach for carrying out ML estimation in a variety of incomplete data problems. The iterations of the EM algorithm proceeds with two steps: an E-step in which the missing information is estimated by their conditional expectations, and an M-step which updates parameter through maximizing the conditionally expected complete-data log-likelihood function, computed in the E-step and called Q-function, with respect to all parameters. Among various variants of the EM algorithm, the parameter estimation of CMMN-MIX distributions is carried out herein by offering an expectation conditional maximization either (ECME; Liu and Rubin [20]) algorithm. The ECME algorithm is a generalization of the EM and ECM [21] algorithms with a faster convergence rate in terms of the number of iterations

or actual computer time. The ECME algorithm is implemented by replacing the M-step of the ECM algorithm with a sequence of computationally simpler conditional maximization (CM) steps where, as explained below, the Q-function or the corresponding constrained actual likelihood function is maximized. The ECME algorithm for obtaining the ML estimate of the CMMN-MIX distributions proceeds as follows.

- Parameter initialization: Set $k = 0$ and choose a relative starting point. Due to the multimodality of the mixture models' log-likelihood function, this primary step might have a significant effect on the search for finding global parameter estimates. If the initial points depart too far from the real values, the EM-type algorithm might converge to any local maximizers. Therefore, the choice of starting points constitutes a fundamental issue in EM-based algorithms. The initialization strategy exploited in this paper is discussed in the next section.
- E-step: At iteration $k \in \{1, 2, \ldots\}$, the so-called Q-function, defined as the expected value of complete-data log-likelihood (9) with $\boldsymbol{\Theta}$ valued at $\boldsymbol{\Theta}^{(k-1)}$, is calculated as

$$Q(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k-1)}) = \sum_{i=1}^{n}\sum_{j=1}^{G}\hat{z}_{ij}^{(k)}\log\pi_j + \sum_{i=1}^{n}\sum_{j=1}^{G}\hat{z}_{ij}^{(k)}\left[\hat{b}_{ij}^{(k)}\log\gamma_j + (1-\hat{b}_{ij}^{(k)})\log(1-\gamma_j)\right] + Q^{good}\left(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k-1)}\right) + Q^{bad}\left(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k-1)}\right),$$

where for $r = 1, 2$, $\hat{u}_{rij}^{(k)} = \mathrm{E}\left(U_i^r \middle| \boldsymbol{X} = \boldsymbol{x}_i, z_{ij} = 1, b_{ij} = 1, \hat{\boldsymbol{\theta}}_j^{(k-1)}\right)$ and $\tilde{u}_{rij}^{(k)} = \mathrm{E}\left(U_i^r \middle| \boldsymbol{X} = \boldsymbol{x}_i, z_{ij} = 1, b_{ij} = 0, \hat{\boldsymbol{\theta}}_j^{(k-1)}\right)$,

$$\hat{z}_{ij}^{(k)} = \frac{\hat{\pi}_j^{(k-1)} f_{\mathrm{CMMN}}\left(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_j^{(k-1)}, \hat{\boldsymbol{\Sigma}}_j^{(k-1)}, \hat{\boldsymbol{\lambda}}_j^{(k-1)}, \hat{\nu}_j^{(k-1)}, \hat{\gamma}_j^{(k-1)}\right)}{\sum_{j=1}^{G}\hat{\pi}_j^{(k-1)} f_{\mathrm{CMMN}}\left(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_j^{(k-1)}, \hat{\boldsymbol{\Sigma}}_j^{(k-1)}, \hat{\boldsymbol{\lambda}}_j^{(k-1)}, \hat{\nu}_j^{(k-1)}, \hat{\gamma}_j^{(k-1)}\right)}, \quad \hat{b}_{ij}^{(k)} = \frac{\hat{\gamma}_j^{(k-1)} f_{\mathrm{MMN}}\left(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_j^{(k-1)}, \hat{\boldsymbol{\Sigma}}_j^{(k-1)}, \hat{\boldsymbol{\lambda}}_j^{(k-1)}\right)}{f_{\mathrm{CMMN}}\left(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_j^{(k-1)}, \hat{\boldsymbol{\Sigma}}_j^{(k-1)}, \hat{\boldsymbol{\lambda}}_j^{(k-1)}, \hat{\nu}_j^{(k-1)}, \hat{\gamma}_j^{(k-1)}\right)},$$

we have

$$Q^{good}(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G}\hat{z}_{ij}^{(k)}\hat{b}_{ij}^{(k)}\log|\boldsymbol{\Sigma}_j| - \frac{1}{2}\mathtt{tr}\left(\sum_{j=1}^{G}\boldsymbol{\Sigma}_j^{-1}\sum_{i=1}^{n}\hat{z}_{ij}^{(k)}\hat{b}_{ij}^{(k)}\left[\boldsymbol{e}_{ij}\boldsymbol{e}_{ij}^{\top} - \hat{u}_{1ij}^{(k)}\boldsymbol{e}_{ij}\boldsymbol{\lambda}_j^{\top} - \hat{u}_{1ij}^{(k)}\boldsymbol{\lambda}_j\boldsymbol{e}_{ij}^{\top} + \hat{u}_{2ij}^{(k)}\boldsymbol{\lambda}_j\boldsymbol{\lambda}_j^{\top}\right]\right),$$

$$\begin{aligned} Q^{bad}(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)}) = &-\frac{p}{2}\sum_{i=1}^{n}\sum_{j=1}^{G}\hat{z}_{ij}^{(k)}(1-\hat{b}_{ij}^{(k)})\log\nu_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G}\hat{z}_{ij}^{(k)}(1-\hat{b}_{ij}^{(k)})\log|\boldsymbol{\Sigma}_j| \\ &- \frac{1}{2}\mathtt{tr}\left(\sum_{j=1}^{G}\boldsymbol{\Sigma}_j^{-1}\sum_{i=1}^{n}\hat{z}_{ij}^{(k)}(1-\hat{b}_{ij}^{(k)})\left[\frac{\boldsymbol{e}_{ij}\boldsymbol{e}_{ij}^{\top}}{\nu_j} - \frac{\tilde{u}_{1ij}^{(k)}}{\sqrt{\nu_j}}\boldsymbol{e}_{ij}\boldsymbol{\lambda}_j^{\top} - \frac{\tilde{u}_{1ij}^{(k)}}{\sqrt{\nu_j}}\boldsymbol{\lambda}_j\boldsymbol{e}_{ij}^{\top} + \tilde{u}_{2ij}^{(k)}\boldsymbol{\lambda}_j\boldsymbol{\lambda}_j^{\top}\right]\right), \end{aligned}$$

such that $\boldsymbol{e}_{ij} = (\boldsymbol{x}_i - \boldsymbol{\mu}_j)$, and $\mathtt{tr}(\cdot)$ is the trace operator of a matrix. Notice that the conditional expectations $\hat{u}_{rij}^{(k)}$ and $\tilde{u}_{rij}^{(k)}$ for the CrSN, CMMNE, and CMMNW distributions can be computed by the closed forms provided in Propositions 1 to 3, respectively.

- CM-step 1: Let $n_j^{(k)} = \sum_{i=1}^{n}\hat{z}_{ij}^{(k)}$, $S(b, u_1, u_2, \nu) = \left(u_1 b + \nu u_2(1-b)\right)$, and $S(b, \nu) = \left(b + \nu(1-b)\right)$. Update π_j, γ_j, $\boldsymbol{\mu}_j$, $\boldsymbol{\lambda}_j$ and $\boldsymbol{\Sigma}_j$ by maximizing Q-function as

$$\begin{aligned} \hat{\pi}_j^{(k)} &= \frac{n_j^{(k)}}{n}, \qquad \hat{\gamma}_j^{(k)} = \frac{1}{n_j^{(k)}}\sum_{i=1}^{n}\hat{z}_{ij}^{(k)}\hat{b}_{ij}^{(k)}, \\ \hat{\boldsymbol{\mu}}_j^{(k)} &= \frac{1}{\sum_{i=1}^{n}\hat{z}_{ij}^{(k)}S\left(\hat{b}_{ij}^{(k)}, (\hat{\nu}_j^{(k)})^{-1}\right)}\sum_{i=1}^{n}\hat{z}_{ij}^{(k)}\left\{\boldsymbol{x}_i S\left(\hat{b}_{ij}^{(k)}, (\hat{\nu}_j^{(k)})^{-1}\right) - \boldsymbol{\lambda}_j^{(k-1)} S\left(\hat{b}_{ij}^{(k)}, \hat{u}_{1ij}^{(k)}, \tilde{u}_{1ij}^{(k)}, (\hat{\nu}_j^{(k)})^{-\frac{1}{2}}\right)\right\}, \\ \hat{\boldsymbol{\lambda}}_j^{(k)} &= \frac{1}{\sum_{i=1}^{n}\hat{z}_{ij}^{(k)}S\left(\hat{b}_{ij}^{(k)}, \hat{u}_{2ij}^{(k)}, \tilde{u}_{2ij}^{(k)}, 1\right)}\sum_{i=1}^{n}\hat{z}_{ij}^{(k)}S\left(\hat{b}_{ij}^{(k)}, \hat{u}_{1ij}^{(k)}, \tilde{u}_{1ij}^{(k)}, (\hat{\nu}_j^{(k)})^{-\frac{1}{2}}\right)\boldsymbol{e}_{ij}^{(k)}, \\ \hat{\boldsymbol{\Sigma}}_j^{(k)} &= \frac{1}{n_j^{(k)}}\sum_{i=1}^{n}\hat{z}_{ij}^{(k)}\left\{\boldsymbol{e}_{ij}^{(k)}\left(\boldsymbol{e}_{ij}^{(k)}\right)^{\top} S\left(\hat{b}_{ij}^{(k)}, (\hat{\nu}_j^{(k)})^{-1}\right) + \hat{\boldsymbol{\lambda}}_j^{(k)}\left(\hat{\boldsymbol{\lambda}}_j^{(k)}\right)^{\top} S\left(\hat{b}_{ij}^{(k)}, \hat{u}_{2ij}^{(k)}, \tilde{u}_{2ij}^{(k)}, 1\right)\right. \\ &\quad \left. - \left(\boldsymbol{e}_{ij}^{(k)}\left(\hat{\boldsymbol{\lambda}}_j^{(k)}\right)^{\top} + \hat{\boldsymbol{\lambda}}_j^{(k)}\left(\boldsymbol{e}_{ij}^{(k)}\right)^{\top}\right) S\left(\hat{b}_{ij}^{(k)}, \hat{u}_{1ij}^{(k)}, \tilde{u}_{1ij}^{(k)}, (\hat{\nu}_j^{(k)})^{-\frac{1}{2}}\right)\right\}, \end{aligned} \tag{11}$$

where $\boldsymbol{e}_{ij}^{(k)} = (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_j^{(k)})$.

- CM-step 2: For the fix $\boldsymbol{\mu}_j = \hat{\boldsymbol{\mu}}_j^{(k)}$, $\boldsymbol{\lambda}_j = \hat{\boldsymbol{\lambda}}_j^{(k)}$, $\boldsymbol{\Sigma}_j = \hat{\boldsymbol{\Sigma}}_j^{(k)}$, and $\gamma_j = \hat{\gamma}_j^{(k)}$, $\hat{\boldsymbol{\nu}}^{(k)} = (\hat{\nu}_1^{(k)}, \ldots, \hat{\nu}_G^{(k)})$ is calculated by maximizing either the Q-function over $\boldsymbol{\nu}$ with the constraint $\nu_j > 1$, or the constrained actual observed log-likelihood function corresponding to (10). In the former approach, $\hat{\boldsymbol{\nu}}^{(k)}$ is obtained by maximizing $Q^{bad}(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)})$ over $\boldsymbol{\nu}$, while in the latter

Table 1
The nomenclature, $\boldsymbol{\Sigma}_j$ structure, and number of free parameters in $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G$. Here, $a^* = p(p+1)/2$, $b^* = (G-1)(p-1)$, and $\boldsymbol{I}_p$ denotes the Identity matrix of size p.

Model	$\omega_j = \omega$	$\boldsymbol{\Delta}_j = \boldsymbol{\Delta}$	$\boldsymbol{\Gamma}_j = \boldsymbol{\Gamma}$	$\boldsymbol{\Sigma}_j$	#Free par.	Model	$\omega_j = \omega$	$\boldsymbol{\Delta}_j = \boldsymbol{\Delta}$	$\boldsymbol{\Gamma}_j = \boldsymbol{\Gamma}$	$\boldsymbol{\Sigma}_j$	#Free par.
EII	Equal	Identity	Identity	$\omega \boldsymbol{I}_p$	1	VEE	Variable	Equal	Equal	$\omega_j \boldsymbol{\Gamma} \boldsymbol{\Delta} \boldsymbol{\Gamma}^\top$	$a^* + G - 1$
VII	Variable	Identity	Identity	$\omega_j \boldsymbol{I}_p$	G	EVE	Equal	Variable	Equal	$\omega \boldsymbol{\Gamma} \boldsymbol{\Delta}_j \boldsymbol{\Gamma}^\top$	$a^* - b^*$
EEI	Equal	Equal	Identity	$\omega \boldsymbol{\Delta}$	p	VVE	Variable	Variable	Equal	$\omega_j \boldsymbol{\Gamma} \boldsymbol{\Delta}_j \boldsymbol{\Gamma}^\top$	$a^* + p(G-1)$
VEI	Variable	Equal	Identity	$\omega_j \boldsymbol{\Delta}$	$G + p - 1$	EEV	Equal	Equal	Variable	$\omega \boldsymbol{\Gamma}_j \boldsymbol{\Delta} \boldsymbol{\Gamma}_j^\top$	$Ga^* - p(G-1)$
EVI	Equal	Variable	Identity	$\omega \boldsymbol{\Delta}_j$	$G(p-1)+1$	VEV	Variable	Equal	Variable	$\omega_j \boldsymbol{\Gamma}_j \boldsymbol{\Delta} \boldsymbol{\Gamma}_j^\top$	$Ga^* - b^*$
VVI	Variable	Variable	Identity	$\omega_j \boldsymbol{\Delta}_j$	Gp	EVV	Equal	Variable	Variable	$\omega \boldsymbol{\Gamma}_j \boldsymbol{\Delta}_j \boldsymbol{\Gamma}_j^\top$	$Ga^* - G + 1$
EEE	Equal	Equal	Equal	$\omega \boldsymbol{\Gamma} \boldsymbol{\Delta} \boldsymbol{\Gamma}^\top$	a^*	VVV	Variable	Variable	Variable	$\omega_j \boldsymbol{\Gamma}_j \boldsymbol{\Delta}_j \boldsymbol{\Gamma}_j^\top$	Ga^*

method, the CM-step switches to the CML-step that obtains $\hat{\boldsymbol{\nu}}^{(k)}$ by searching the solution of following function

$$\hat{\nu}_j^{(k)} = \arg\max_{\nu > 1} \sum_{i=1}^{n} \hat{z}_{ij}^{(k)} \log f_{\text{CMMN}}\left(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_j^{(k)}, \hat{\boldsymbol{\Sigma}}_j^{(k)}, \hat{\boldsymbol{\lambda}}_j^{(k)}, \nu, \hat{\gamma}_j^{(k)}\right).$$

Herein, we adopt the CML approach by exploiting the R function optim(·) to perform a numerical search of maximization.

3.3. Parsimonious models

For the N-MIX model, it is known that there are $p(p+1)/2$ free parameters for each scale matrix $\boldsymbol{\Sigma}_j$. By increasing the number of components, we will face a potentially over-parameterization issue in model fitting. To overcome this obstacle, the parsimony into the Gaussian and non-Gaussian mixture models was introduced by Banfield and Raftery [22]. They considered the eigenvalue decomposition parameterization (EPD) on the component scale matrices

$$\boldsymbol{\Sigma}_j = \omega_j \boldsymbol{\Gamma}_j \boldsymbol{\Delta}_j \boldsymbol{\Gamma}_j^\top, \tag{12}$$

where $\omega_j = |\boldsymbol{\Sigma}_j|^{1/p}$ is the constant of proportionality, $\boldsymbol{\Delta}_j$ is a diagonal matrix of the eigenvalues of $\boldsymbol{\Sigma}_j$ sorted in decreasing order with $|\boldsymbol{\Delta}_j| = 1$, and $\boldsymbol{\Gamma}_j$ is the orthogonal matrix of eigenvectors of $\boldsymbol{\Sigma}_j$ ordered according to their eigenvalues. By applying various constraints to the EDP (12), Celeux and Govaert [7] introduced fourteen normally parsimonious clustering models summarized in Table 1. Bear in mind that the elements in parameterization (12) have a different geometric interpretation in the normally parsimonious clustering, i.e. ω_j represents the volume of the jth cluster while $\boldsymbol{\Delta}_j$ and $\boldsymbol{\Gamma}_j$ determine the shape and orientation of the jth cluster, respectively.

Stemming from Celeux and Govaert [7], we impose constraints on the eigenvalue decomposition on scale matrix structure (12) to obtain a family of fourteen parsimonious CMMN-MIX distributions. It is obvious that the estimate of $\boldsymbol{\Sigma}_j$ for the case VVV can be obtained by the aforementioned ECME algorithm (Eq. (11)). The update of ω_j, $\boldsymbol{\Delta}_j$, and $\boldsymbol{\Gamma}_j$ for the other thirteen constrained models can be analogously obtained by using an identical procedure of Celeux and Govaert [7] with considering $\boldsymbol{U}_j^{(k)} = \sum_{i=1}^{n} \hat{z}_{ij}^{(k)} \boldsymbol{U}_{ij}^{(k)}$ and $\boldsymbol{U} = \sum_{j=1}^{G} \boldsymbol{U}_j^{(k)}$, where

$$\begin{aligned}\boldsymbol{U}_{ij}^{(k)} = {} & \boldsymbol{e}_{ij}^{(k)} \left(\boldsymbol{e}_{ij}^{(k)}\right)^\top S\left(\hat{b}_{ij}^{(k)}, (\hat{\nu}_j^{(k)})^{-1}\right) + \hat{\lambda}_j^{(k)} \left(\hat{\lambda}_j^{(k)}\right)^\top S\left(\hat{b}_{ij}^{(k)}, \hat{u}_{2ij}^{(k)}, \tilde{u}_{2ij}^{(k)}, 1\right) \\ & - \left(\boldsymbol{e}_{ij}^{(k)} \left(\hat{\lambda}_j^{(k)}\right)^\top + \hat{\lambda}_j^{(k)} \left(\boldsymbol{e}_{ij}^{(k)}\right)^\top\right) S\left(\hat{b}_{ij}^{(k)}, \hat{u}_{1ij}^{(k)}, \tilde{u}_{1ij}^{(k)}, (\hat{\nu}_j^{(k)})^{-\frac{1}{2}}\right).\end{aligned}$$

Despite the geometric interpretation of the normally parsimonious clustering model in terms of volume, shape, and orientation, notice that it is impossible to make any inference on the component volume, shape, and orientation for the CMMN-MIX distributions due to its covariance structure $Cov(\boldsymbol{X}) = (\boldsymbol{\Sigma}_j + Var(W)\boldsymbol{\lambda}_j \boldsymbol{\lambda}_j^\top)(\gamma_j + \nu_j(1-\gamma_j))$. Form $Cov(\boldsymbol{X})$, one can observe the influence of $\boldsymbol{\lambda}_j$, γ_j, and ν_j on the covariance between the variables in $\boldsymbol{X}$. For instance, if the CMMN-MIX model with EII is selected to fit the data ($\boldsymbol{\lambda} \neq \boldsymbol{0}$, $\nu_j > 1$), the resulting model does not have a diagonal covariance structure for each component.

3.4. Provision of standard error estimates

To compute the asymptotic covariance of the ML estimates, the information-based method recommended by Basford et al. [23] is offered. Based on Meilijson [24]'s formula, the empirical information matrix is obtained as

$$I_e(\boldsymbol{\Theta} \mid \boldsymbol{x}) = \sum_{i=1}^{n} \boldsymbol{s}(\boldsymbol{x}_i \mid \boldsymbol{\Theta}) \boldsymbol{s}^\top(\boldsymbol{x}_i \mid \boldsymbol{\Theta}) - \frac{1}{n} \boldsymbol{S}(\boldsymbol{x} \mid \boldsymbol{\Theta}) \boldsymbol{S}^\top(\boldsymbol{x} \mid \boldsymbol{\Theta}), \tag{13}$$

where $\boldsymbol{S}(\boldsymbol{x} \mid \boldsymbol{\Theta}) = \sum_{i=1}^{n} \boldsymbol{s}(\boldsymbol{x}_i \mid \boldsymbol{\Theta})$. Inspired by Louis [25], the individual score $\boldsymbol{s}(\boldsymbol{x}|\boldsymbol{\Theta})$ can be determined as

$$\boldsymbol{s}(\boldsymbol{x}_i \mid \boldsymbol{\Theta}) = \frac{\partial f(\boldsymbol{x}_i \mid \boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = \mathrm{E}\left[\frac{\partial \ell_{ci}(\boldsymbol{\Theta} \mid \boldsymbol{x}_{ci})}{\partial \boldsymbol{\Theta}} \;\middle|\; \boldsymbol{x}_i, \boldsymbol{\Theta}\right],$$

where

$$\begin{aligned}\ell_{c_i}(\boldsymbol{\Theta} \mid \boldsymbol{x}_{ci}) = &\sum_{i=1}^{n} z_{ij}\left[\log \pi_j + b_{ij}\log\gamma_j + (1-b_{ij})\log(1-\gamma_j)\right] \\ &+ \sum_{i=1}^{n} z_{ij} b_{ij} \log\left(\phi_p(\boldsymbol{\mu}_j + u_i\boldsymbol{\lambda}_j, \boldsymbol{\Sigma}_j)h(u_i)\right) + \sum_{i=1}^{n} z_{ij}(1-b_{ij})\log\left(\phi_p(\boldsymbol{\mu}_j + u_i\sqrt{v_j}\boldsymbol{\lambda}_j, v_j\boldsymbol{\Sigma}_j)h(u_i)\right) \end{aligned} \tag{14}$$

is the complete-data log-likelihood formed with the ith $\boldsymbol{x}_c$. Substituting the ML estimates ($\hat{\boldsymbol{\Theta}}$) into (13), we obtain

$$I_e(\hat{\boldsymbol{\Theta}} \mid \boldsymbol{x}) = \sum_{i=1}^{n} \hat{\mathbf{s}}_j \hat{\mathbf{s}}_j^{\top}, \tag{15}$$

where $\hat{\mathbf{s}}_i = \boldsymbol{s}(\boldsymbol{x}_i \mid \hat{\boldsymbol{\Theta}})$ is a $m \times 1$ vector comprising elements of

$$(\hat{\mathbf{s}}_{i,\pi_1}, \ldots, \hat{\mathbf{s}}_{i,\pi_{G-1}}, \hat{\mathbf{s}}_{i,\boldsymbol{\mu}_1}^{\top}, \ldots, \hat{\mathbf{s}}_{i,\boldsymbol{\mu}_G}^{\top}, \hat{\mathbf{s}}_{i,\boldsymbol{\lambda}_1}^{\top}, \ldots, \hat{\mathbf{s}}_{i,\boldsymbol{\lambda}_G}^{\top}, \hat{\mathbf{s}}_{i,\boldsymbol{\sigma}_1}^{\top}, \ldots, \hat{\mathbf{s}}_{i,\boldsymbol{\sigma}_G}^{\top}, \hat{\mathbf{s}}_{i,\nu_1}, \ldots, \hat{\mathbf{s}}_{i,\nu_G}, \hat{\mathbf{s}}_{i,\gamma_1}, \ldots, \hat{\mathbf{s}}_{i,\boldsymbol{\gamma}_G})^{\top},$$

with $\boldsymbol{\sigma}_j = \mathrm{vech}(\boldsymbol{\Sigma}_j)$ being a $p(p+1)/2$ vector containing the lower triangular elements of $\boldsymbol{\Sigma}_j$. The explicit expressions for the elements of $\hat{\mathbf{s}}_i$, obtained by taking expectation on the partial derivatives of (14), are summarized as follows:

$$\begin{aligned}
\hat{\mathbf{s}}_{i,\pi_j} &= \mathrm{E}\left[\frac{\partial \ell_{ci}(\boldsymbol{\Theta} \mid \boldsymbol{x}_{ci})}{\partial \pi_j} \;\middle|\; \boldsymbol{x}_i, \hat{\boldsymbol{\Theta}}\right] = \frac{\hat{z}_{ij}}{\hat{\pi}_j} - \frac{\hat{z}_{iG}}{\hat{\pi}_G}, \qquad \hat{\mathbf{s}}_{i,\gamma_j} = \mathrm{E}\left[\frac{\partial \ell_{ci}(\boldsymbol{\Theta} \mid \boldsymbol{x}_{ci})}{\partial \gamma_r} \;\middle|\; \boldsymbol{x}_i, \hat{\boldsymbol{\Theta}}\right] = \hat{z}_{ij}\frac{\hat{b}_{ij} - \hat{\gamma}_j}{\hat{\gamma}_j(1-\hat{\gamma}_j)}, \\
\hat{\mathbf{s}}_{i,\boldsymbol{\mu}_j} &= \mathrm{E}\left[\frac{\partial \ell_{ci}(\boldsymbol{\Theta} \mid \boldsymbol{x}_{ci})}{\partial \boldsymbol{\mu}_j} \;\middle|\; \boldsymbol{x}_i, \hat{\boldsymbol{\Theta}}\right] = \hat{z}_{ij}\hat{\boldsymbol{\Sigma}}_j^{-1}\left(\hat{\boldsymbol{e}}_{ij} S(\hat{b}_{ij}, \hat{v}_j^{-1}) - \hat{\boldsymbol{\lambda}}_j S(\hat{b}_{ij}, \hat{u}_{1ij}, \tilde{u}_{1ij}, \hat{v}_j^{-1/2})\right), \\
\hat{\mathbf{s}}_{i,\boldsymbol{\lambda}_j} &= \mathrm{E}\left[\frac{\partial \ell_{ci}(\boldsymbol{\Theta} \mid \boldsymbol{x}_{ci})}{\partial \boldsymbol{\lambda}_j} \;\middle|\; \boldsymbol{x}_i, \hat{\boldsymbol{\Theta}}\right] = \hat{z}_{ij}\hat{\boldsymbol{\Sigma}}_j^{-1}\left(\hat{\boldsymbol{e}}_{ij} S\left(\hat{b}_{ij}, \hat{u}_{1ij}, \tilde{u}_{1ij}, \hat{v}_j^{-1/2}\right) - \hat{\boldsymbol{\lambda}}_j S\left(\hat{b}_{ij}, \hat{u}_{2ij}, \tilde{u}_{2ij}, 1\right)\right), \\
\hat{\mathbf{s}}_{i,\nu_j} &= \mathrm{E}\left[\frac{\partial \ell_{ci}(\boldsymbol{\Theta} \mid \boldsymbol{x}_{ci})}{\partial \nu_j} \;\middle|\; \boldsymbol{x}_i, \hat{\boldsymbol{\Theta}}\right] = \frac{1}{2\hat{v}_j^2}\hat{z}_{ij}(1-\hat{b}_{ij})\,\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}_j^{-1}\left[\hat{\boldsymbol{e}}_{ij}\hat{\boldsymbol{e}}_{ij}^{\top} - \frac{\tilde{u}_{1ij}\sqrt{\hat{v}_j}}{2}(\hat{\boldsymbol{\lambda}}_j\hat{\boldsymbol{e}}_{ij}^{\top} + \hat{\boldsymbol{e}}_{ij}\hat{\boldsymbol{\lambda}}_j^{\top})\right]\right),
\end{aligned}$$

where $\hat{\boldsymbol{e}}_{ij} = (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_j)$. Let σ_{js} also be the sth entry of $\boldsymbol{\sigma}_j$. Making use the formulae of

$$\frac{\partial \ln|\boldsymbol{\Sigma}_j|}{\partial \sigma_{js}} = \mathrm{tr}\left(\boldsymbol{\Sigma}_j^{-1}\frac{\partial \boldsymbol{\Sigma}_j}{\partial \sigma_{js}}\right) \quad \text{and} \quad \frac{\partial \boldsymbol{\Sigma}_j^{-1}}{\partial \sigma_{js}} = -\boldsymbol{\Sigma}_j^{-1}\frac{\partial \boldsymbol{\Sigma}_j}{\partial \sigma_{js}}\boldsymbol{\Sigma}_j^{-1},$$

we have

$$\begin{aligned}
\hat{\mathbf{s}}_{i,\sigma_{js}} &= \mathrm{E}\left[\frac{\partial \ell_{ci}(\boldsymbol{\Theta} \mid \boldsymbol{x}_{ci})}{\partial \sigma_{js}} \;\middle|\; \boldsymbol{x}_i, \hat{\boldsymbol{\Theta}}\right] \\
&= -\frac{\hat{z}_{ij}}{2}\left\{\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}_j^{-1}\frac{\partial \boldsymbol{\Sigma}_r}{\partial \sigma_{js}}\Big|_{\hat{\sigma}_{js}}\right) + \hat{\boldsymbol{e}}_{ij}^{\top}\hat{\boldsymbol{A}}_{js}\hat{\boldsymbol{e}}_{ij} S\left(\hat{b}_{ij}, \hat{v}_j^{-1}\right) + \hat{\boldsymbol{\lambda}}_j^{\top}\hat{\boldsymbol{A}}_{js}\hat{\boldsymbol{\lambda}}_j S\left(\hat{b}_{ij}, \hat{u}_{2ij}, \tilde{u}_{2ij}, 1\right)\right. \\
&\quad \left. - S\left(\hat{b}_{ij}, \hat{u}_{1ij}, \tilde{u}_{1ij}, \hat{v}_j^{-\frac{1}{2}}\right)\left(\hat{\boldsymbol{e}}_{ij}^{\top}\hat{\boldsymbol{A}}_{js}\hat{\boldsymbol{\lambda}}_j + \hat{\boldsymbol{\lambda}}_j^{\top}\hat{\boldsymbol{A}}_{js}\hat{\boldsymbol{e}}_{ij}\right)\right\},
\end{aligned}$$

where $\hat{\boldsymbol{A}}_{js} = \hat{\boldsymbol{\Sigma}}_j^{-1}\partial \boldsymbol{\Sigma}_j/\partial \sigma_{js}\hat{\boldsymbol{\Sigma}}_j^{-1}\big|_{\hat{\sigma}_{js}}$. The standard errors of ML estimates are calculated by taking the square root of the diagonal elements of (15).

4. Computational and operational aspects

4.1. Starting values specification

As discussed in Section 3, the MMN class of distributions contains the normal model as a special case. Consequently, the CMMN distribution (and CMMN-MIX model) degenerates to the contaminated normal when $H(\cdot)$ degenerates with $u = 1$, or alternatively when $\boldsymbol{\lambda}$ tends to zero. Therefore, to implement the ECME algorithm described in Section 3, the parameter initial values, $\hat{\boldsymbol{\Theta}}^{(0)}$, can be adopted by fitting a mixture of contaminated normal (CN-MMN, Punzo and McNicholas [2]) distributions. In this regard, one can set $\hat{\boldsymbol{\lambda}}_1^{(0)} = \cdots = \hat{\boldsymbol{\lambda}}_G^{(0)} = \mathbf{0}$, and obtain the estimate of parameters $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G$, $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G$, $\nu_1, \ldots, \nu_G$, and $\gamma_1, \ldots, \gamma_G$ by calibrating CN-MIX distributions to the data, as implemented by the CNmixt(·)

function of the R package **ContaminatedMixt**. The returned parameter estimates are considered as the starting values of the first E-step of the ECME algorithm to fit the CMMN-MIX distributions. In the numerical parts of this paper, we initialize the parameters via explained strategy. However, as an alternative approach, the starting points can be obtained by setting $\hat{\boldsymbol{\lambda}}_j^{(0)} = \mathbf{0}$, $\hat{\nu}_j^{(0)} = 1.001$, and $\hat{\gamma}_j^{(0)} = 0.999$ for $j = 1, \ldots, G$, and fitting the N-MIX distributions to the data by using the gpcm(·) function of the R package **mixture**.

4.2. Stopping rule

A reasonable procedure to determine whether the proposed ECME algorithm has converged is Aitken's acceleration method [26]. Recommended by Lindsay [27], the algorithm can be considered to reach convergence at iteration k when

$$\ell_\infty(\hat{\boldsymbol{\Theta}}^{(k+1)}) - \ell(\hat{\boldsymbol{\Theta}}^{(k)}) < \epsilon,$$

where ϵ is the per-specified tolerance, $\ell(\hat{\boldsymbol{\Theta}}^{(k)})$ means the observed log-likelihood value evaluated at $\hat{\boldsymbol{\Theta}}^{(k)}$, and for the Aitken acceleration factor at iteration k, $a^{(k)} = \{\ell(\hat{\boldsymbol{\Theta}}^{(k+1)}) - \ell(\hat{\boldsymbol{\Theta}}^{(k)})\}/\{\ell(\hat{\boldsymbol{\Theta}}^{(k)}) - \ell(\hat{\boldsymbol{\Theta}}^{(k-1)})\}$, the asymptotic estimate of log-likelihood at iteration $k+1$, is

$$\ell_\infty(\hat{\boldsymbol{\Theta}}^{(k+1)}) = \ell(\hat{\boldsymbol{\Theta}}^{(k+1)}) + \frac{1}{1-a^{(k)}}\{\ell(\hat{\boldsymbol{\Theta}}^{(k+1)}) - \ell(\hat{\boldsymbol{\Theta}}^{(k)})\}.$$

In our numerical work, we set $\epsilon = 10^{-5}$.

4.3. Automatic outliers detection

By achieving the convergence criterion in the ECME algorithm, the fitted CMMN-MIX model allows us to determine cluster membership of observations as well as to detect outliers points in each cluster. To do so, we record the values of $\hat{z}_{ij}$ and $\hat{b}_{ij}$ at the convergence of the ECME algorithm. For clustering purposes, the maximum a posteriori (MAP) classification for each observed data $\boldsymbol{x}_i$ is computed as

$$\text{MAP}(\hat{z}_{ij}) = \begin{cases} 1 & \text{if } \arg\max_g(\hat{z}_{ig}) = j; \\ 0 & \text{otherwise.} \end{cases}$$

Then, the observation $\boldsymbol{x}_i$ is classified into the gth cluster if $\text{MAP}(\hat{z}_{ig}) = 1$. To detect outliers, we consider $\hat{b}_{ig}$ if $\boldsymbol{x}_i$ is determined to belong to the gth group. Therefore, if $\hat{b}_{ig} \geq 0.5$, $\boldsymbol{x}_i$ is a good point and if $\hat{b}_{ig} < 0.5$, it is considered a bad observation.

4.4. Model comparison and classification performance

As is typical in mixture models, we assume that the number of mixing components G is unknown. The key for the statistical inference based on model (8) is to determine (estimate) G. In the majority of the mixture literature, a natural idea to answer this issue is to use some information criteria. In our numerical examples, we use the Bayesian information criterion (BIC; Schwarz [28]) defined as

$$\text{BIC} = -2\ell_{\max} + m\log(n),$$

where $\ell_{\max}$ is the maximized log-likelihood value and m denotes the number of free parameters in $\boldsymbol{\Theta}$. For model comparison purposes, the smaller score of BIC, the better fit of the model to the data.

We also regard the identification of optimal classification of the data. To measure the ability of clustering agreement, we employ the misclassification rate (MCR) and adjusted Rand index (ARI, Hubert and Arabie [29]). The model with the highest ARI (smallest MCR) score is considered to provide the most reliable classification accuracy.

5. Simulated data analysis

Herein, three Monte-Carlo simulation studies are conducted to examine the performance and effectiveness of the proposed model. We mainly focus on the impact of outliers and noise points respectively on the parameter estimation and data clustering.

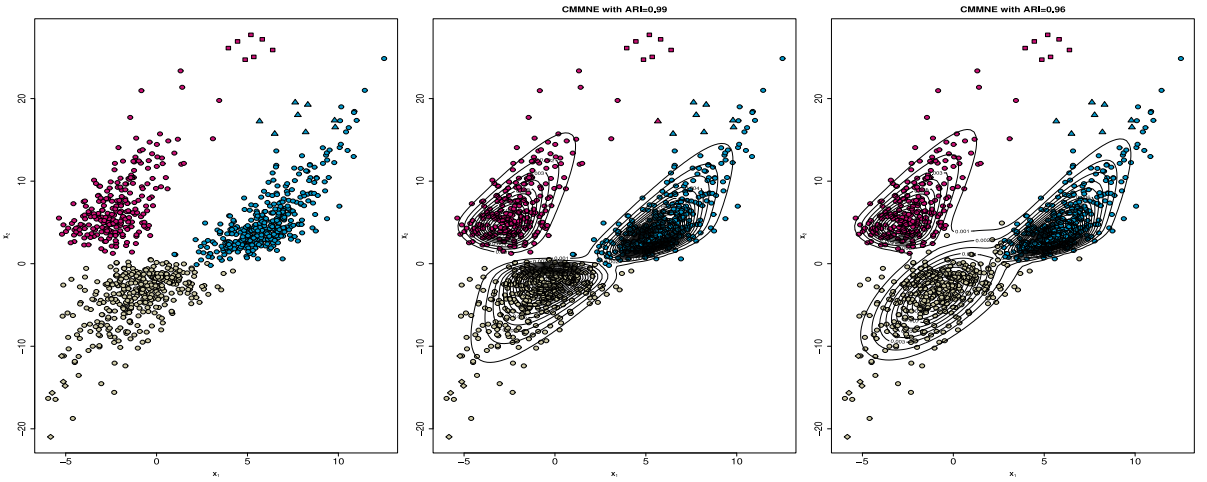

Fig. 1. Scatter plots of a sample generated from a three-components MMNE distributions containing 2% outliers with and without overlaid contours of the fitted CMMNE-MIX and MMNE-MIX models.

5.1. Assessing the impact of outliers on the parameter estimates

This section aims at examining the sensitivity of the ML estimates in presence of outliers. Data are simulated to illustrate how much the common parameter estimates of the CMMN-MIX and MMN-MIX models and their clustering performances are affected by outliers. As an illustration for this purpose, bivariate data (x_{1i}, x_{2i}) are firstly generated from a three-component MMNE-MIX distributions with true parameters, $\boldsymbol{\pi} = (0.2, 0.35, 0.45)$, $\boldsymbol{\mu}_1 = (-3, 3)$, $\boldsymbol{\mu}_2 = (0, -1)$, $\boldsymbol{\mu}_3 = (5, 2)$ $\boldsymbol{\lambda}_1 = (1, 4)$, $\boldsymbol{\lambda}_2 = (-1, -3)$, $\boldsymbol{\lambda}_3 = (1, 3)$, and

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & -0.18 \\ -0.18 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2.40 & 0 \\ 0 & 0.88 \end{pmatrix}, \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1.60 & 0.85 \\ 0.85 & 0.90 \end{pmatrix}.$$

We then randomly add some outliers generated from the uniform distribution over an interval on each dimension. Let $\max_{X_{lj}}$ and $\min_{X_{lj}}$ for $j = l = 1, 2$, respectively denote the maximum and minimum of the generated x_{li} in the jth group. Then, the intervals $(\max_{X_{11}}, \max_{X_{11}} +3)$ and $(\max_{X_{21}}, \max_{X_{21}} +5)$, $(\min_{X_{12}}, -5)$ and $(\min_{X_{22}} -5, \min_{X_{22}})$, and (5, 10) and (15, 20) are considered for the outliers generation of the first, second, and third group, respectively. Under this scenario, we simulate 500 samples of various sizes ($n = 100, 200, 500$ and 1000) with 2% added outliers. The scatter plot of a generated sample is shown in Fig. 1 where different colors represent the distinct true groups of individuals.

In each replication of 500 trails, we fit the unconstrained variant (VVV) of the CMMNE-MIX and MMNE-MIX models to the data with the fixed number of components $G = 3$ and record the parameter estimates and clustering criteria (ARI and MCR). For sake of comparison, Fig. 2 reports the bias and root mean squared error (RMSE) of the common parameters calculated as

$$\text{Bias} = \frac{1}{500}\sum_{r=1}^{500}(\hat{\theta}_r - \theta_{\text{true}}) \quad \text{and} \quad \text{RMSE} = \sqrt{\frac{1}{500}\sum_{r=1}^{500}(\hat{\theta}_r - \theta_{\text{true}})^2},$$

where $\hat{\theta}_r$ denotes the ML estimate of a specific parameter at the rth replication and θ_{true} is its corresponding true value.

Looking at Fig. 2, the following results are striking:

- The ML estimate of mixing probabilities $\hat{\pi}_j$ for both MMNE-MIX and CMMNE-MIX models are robust against outliers.
- It is notable that the outliers have significant effect on the estimation process of the common parameters $\hat{\boldsymbol{\mu}}_j$, $\hat{\boldsymbol{\lambda}}_j$, and $\hat{\boldsymbol{\Sigma}}_j$ for both MMNE-MIX and CMMNE-MIX models. Although almost all of these parameters are bias with large MSE, the ML estimates for the CMMNE-MIX model is more robust in presence of outliers since they have smaller values of RMSE.
- Regardless of the considered MMNE-MIX and CMMNE-MIX models, it can be seen that the bias and RMSE for $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\lambda}}_j$ tend to decrease when the sample size growth. However, it is evidence that the outliers significantly affect on $\hat{\boldsymbol{\Sigma}}_j$ for the MMNE-MIX model, in the sense that its estimate gets far from the true value.

To check the influence of the outliers on the classification performance, Table 2 reports the average values of ARI and MCR for both fitted models (CMMNE-MIX and MMNE-MIX) to the generated data. It can clearly be seen that the clustering performance of the CMMNE-MIX is improved with the increase of n, however, a reverse pattern is observed for

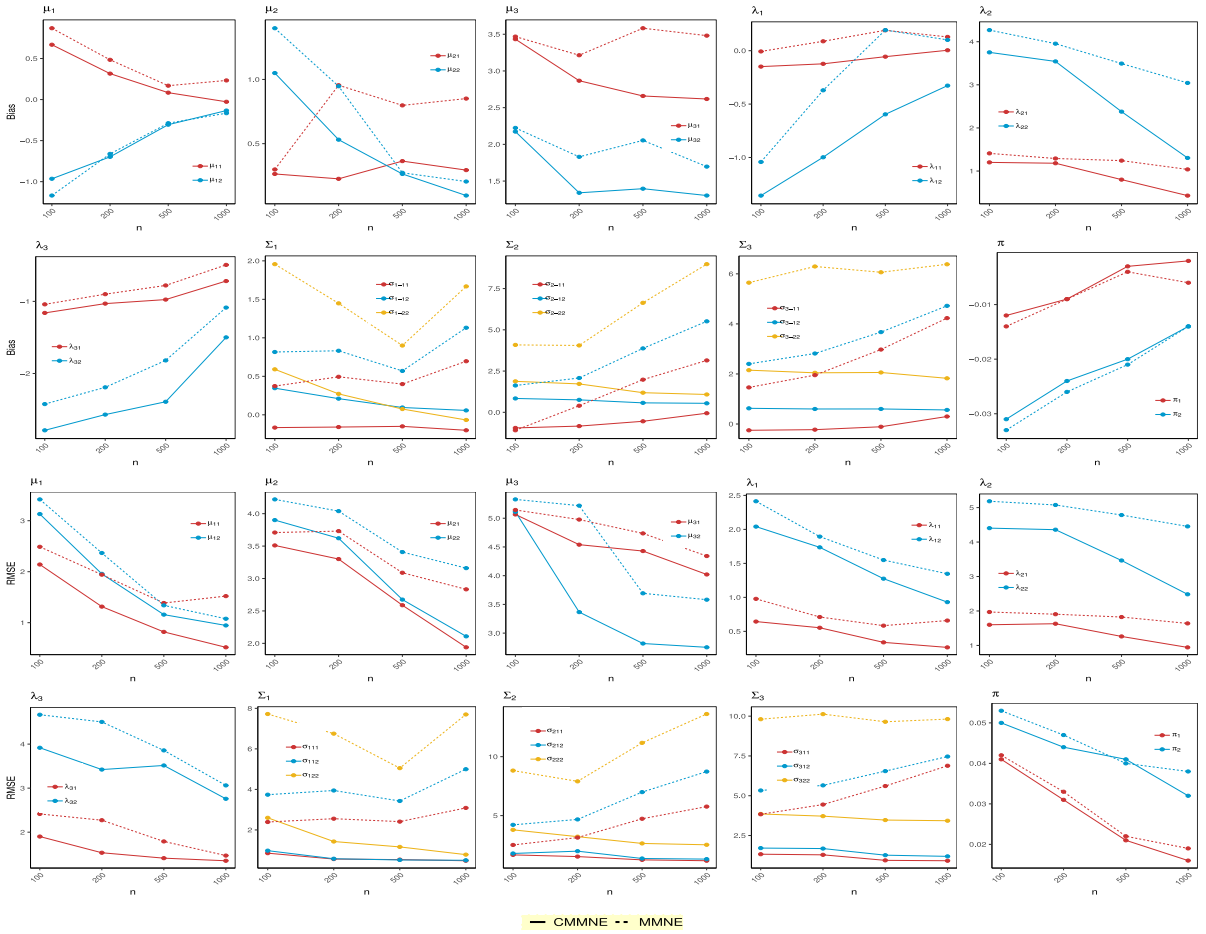

Fig. 2. Simulation results for assessing the consistency of parameter estimates for 500 samples from the MMNE-MIX model with various sample sizes and 2% outliers.

Table 2
The mean of classification criteria (with standard deviation in parenthesis) for the CMMNE-MIX and MMNE-MIX models over 500 samples generated in simulation study 5.1.

	CMMNE-MIX				MMNE-MIX			
	100	200	500	1000	100	200	500	1000
MCR	0.0696 (0.0682)	0.0538 (0.0518)	0.0346 (0.0549)	0.0329 (0.0769)	0.0749 (0.0880)	0.0588 (0.0648)	0.0908 (0.1209)	0.1136 (0.1387)
ARI	0.8153 (0.1644)	0.8381 (0.1471)	0.9053 (0.1317)	0.9137 (0.1523)	0.8144 (0.1865)	0.8215 (0.1571)	0.8030 (0.2271)	0.7681 (0.2577)

the MMNE-MIX model. It is also striking that the standard deviations of ARI and MCR values for the CMMNE-MIX model are smaller than those for the CMMNE-MIX model, showing its flexibility for classification purposes. The scatter plots superimposed on the fitted CMMNE-MIX and MMNE-MIX contours shown in the middle and right panels of Fig. 1 give an illustration of one simulated case where different colors represent the distinct groups of individuals detected by the fitted models. Clearly, the fitted CMMNE-MIX model provides better classification and its contours adapt the scattering patterns more closely than the MMNE-MIX model.

5.2. Assessing the impact of background noise on fitting and clustering performance

In the second simulation study, we focus on the effect of background noises on the fitting and clustering performance of the CMMN-MIX submodels. Streaming from Morris et al. [4], we generate "good" data from a three-component t-slice mixture distributions (a three-component mixture model in which each component comes from a t-slice distribution). Specifically, for each group $j = 1, 2, 3$, we generate $(x_{1i}, x_{2i})^\top$ data from a bivariate Student-t distribution with parameters $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j$, and ν_j and record it as a good point if $x_{2i} + a_j x_{1i} + b_j \geq 0$. This algorithm is iterated until the required sample size of the group has been reached. By obtaining good data points with total size $n = 500$, we generate background noises

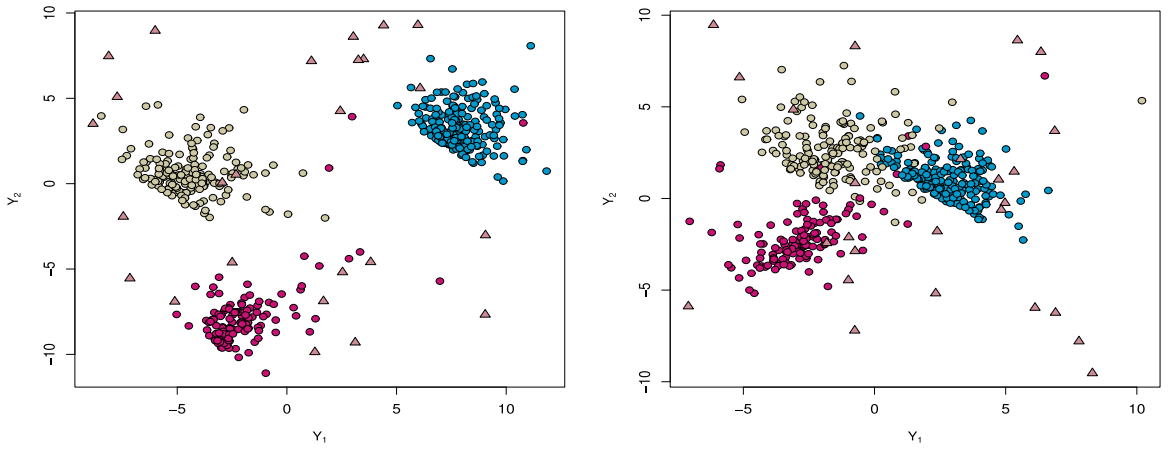

Fig. 3. Scatter plots of samples generated under well-separated (left panel) and poor-separated (right panel) scenarios from the t.slice mixtures.

of size 25 from the uniformly distributed two-dimensional plate over the interval $(-5, 10)$. We then have an overall 525 observations in each replication. The true parameter values are set to $\boldsymbol{\pi} = (0.25, 0.35, 0.4)$, $\boldsymbol{\nu} = (\nu_1, \nu_2, \nu_3) = (2, 6, 15)$, and

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2.35 & 0.5 \\ 0.5 & 2 \end{pmatrix}, \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1.6 & 0.5 \\ 0.5 & 2 \end{pmatrix}.$$

To check the effect of the degree of overlapping on the fitting and clustering performance, we consider two scenarios: well-separated and poor-separated groups. In this regard, we set $a_j = 0.8;\ j = 1, 2, 3$,

$$\text{well-separated:} \quad \boldsymbol{\mu}_1 = \begin{pmatrix} -7 \\ 7 \end{pmatrix}, \ \boldsymbol{\mu}_2 = \begin{pmatrix} -5 \\ 0 \end{pmatrix}, \ \boldsymbol{\mu}_3 = \begin{pmatrix} 7 \\ 2 \end{pmatrix}, \quad \boldsymbol{b} = (b_1, b_2, b_3) = (12, 5, -8)$$
$$\text{poor-separated:} \quad \boldsymbol{\mu}_1 = \begin{pmatrix} -3 \\ 3 \end{pmatrix}, \ \boldsymbol{\mu}_2 = \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \ \boldsymbol{\mu}_3 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad \boldsymbol{b} = (b_1, b_2, b_3) = (9, 1, -2).$$

Fig. 3 shows the scatter plots of samples generated from the well-separated and poor-separated scenarios. The colored bullets represent the good data points where as background noises are shown with triangles (▲).

We simulate 100 samples for each scenario and fit the CMMNE-MIX, CMMNW-MIX, and CrSN-MIX models, with the constrained and unconstrained variants of eigen-decomposition in Table 1, to the data. For the assumed unknown number of mixture components G, forty two competing models are fitted to the data, where G ranges from 1 to 5. Tables 3 and 4 list the average of BIC values (with standard deviation in parentheses) of the parsimonious variants over 100 trials respectively for the well- and poor-separated scenarios. For choosing the most plausible model, these Tables also report: the frequencies of the chosen model with $G = 3$ supported by BIC (FrEq. (1)) and the frequencies of the chosen parsimonious variant supported by BIC (FrEq. (2)). For the sake of clustering comparison, Tables 3 and 4 report the average value of ARI and MCR criteria for $G = 3$. Moreover, we give the proportion of bad points that are correctly identified as bad points (TPR) and the proportion of good points that are miss-specified as bad points (FPR) for evaluating outliers detection performance.

Form FrEq. (1) in Table 3, it can be seen that the constrained variants of the CrSN-MIX, CMMNE-MIX, and CMMNW-MIX models with the fixed matrix of eigenvalues ($\Gamma_j = \Gamma$) outperform others, respectively with rates 0.92, 0.92, and 0.93. The results of FrEq. (1) depicted in Table 3 show that the fitted models with assumption $\Gamma_j = \Gamma$ in their covariance matrices successfully detect the true value of G based on the BIC criterion. Focusing on the averaged BIC of the forty two mixture models, it is supported that the VVE version of the CrSN-MIX, CMMNE-MIX, and CMMNW-MIX models significantly perform better than other covariance structure, whereas the CrSN-MIX distributions is the best in terms of model fitting in this simulation scenario. As can be expected from the considered simulation scenario, we have great clustering results (MCR close to zero and ARI near to one) in all fitted models. Although the constrained and unconstrained CMMNW-MIX models are the best with respect to TPR value, the CMMNE-MIX performs much better in the discrimination of good data points in this scenario.

When we turn our attention to the poor-separated scenario, the results significantly support the performance of the CMMNE-MIX model. Form Table 4, We first notice that the rate of selecting the correct number of components supported by BIC for all CMMNE-MIX, CMMNW-MIX, and CrSN-MIX models is greater than 60% (FrEq. (1)), showing the ability of the proposed contaminated mixture models for the model selection purposes. It should be also highlighted that for $G = 3$ all CMMNE-MIX, CMMNW-MIX, and CrSN-MIX models suggest constrained variants with a fixed constant of proportionality $\omega_j = \omega$ in the poor-separated scenario, respectively with the rate 0.88, 0.92, and 0.76. Overall, the BIC values in Table 4 show that the EVE is the best among covariance structures and the CMMNE-MIX model is best among CMMN-MIX submodels in this scenario. Comparing the ARI, MCR, TPR, and FPR values in Tables 3 and 4, we secondly note

Table 3
Simulation results for the well-separated mixture of t-slice distributions.

Mixture	Meas.	Parsimonious variant													
model		EII	VII	EEI	VEI	EVI	VVI	EEE	VEE	EVE	VVE	EEV	VEV	EVV	VVV
CrSN	BIC	4743.157	4713.862	4743.792	4718.813	4753.002	4725.126	4714.261	4667.750	4690.830	4660.406	4719.282	4682.202	4697.432	4669.675
		(64.584)	(65.375)	(62.788)	(62.432)	(65.313)	(65.811)	(63.929)	(63.607)	(69.698)	(84.489)	(66.561)	(78.805)	(68.340)	(82.198)
	Freq.1	46	78	45	73	66	82	60	78	78	90	61	81	82	90
	Freq.2	0	0	0	0	0	0	0	30	4	59	0	2	0	5
	ARI	0.959	0.965	0.961	0.965	0.958	0.969	0.942	0.942	0.936	0.949	0.950	0.955	0.936	0.947
	MCR	0.015	0.013	0.014	0.013	0.015	0.011	0.021	0.021	0.024	0.019	0.018	0.016	0.024	0.019
	TPR	0.864	0.816	0.865	0.818	0.847	0.816	0.860	0.824	0.845	0.816	0.849	0.819	0.840	0.813
	FPR	0.188	0.111	0.190	0.108	0.178	0.108	0.165	0.111	0.151	0.110	0.144	0.115	0.126	0.108
CMMNE	BIC	4744.402	4726.982	4748.399	4730.094	4758.227	4742.641	4715.763	4689.337	4697.642	4668.201	4725.331	4699.099	4705.458	4687.227
		(64.998)	(68.751)	(63.865)	(63.516)	(65.034)	(66.072)	(71.801)	(75.956)	(71.182)	(72.843)	(71.878)	(73.934)	(75.449)	(73.785)
	Freq.1	51	80	60	80	74	87	69	79	80	90	66	67	82	89
	Freq.2	0	2	1	0	0	1	5	26	11	40	1	5	1	7
	ARI	0.965	0.958	0.968	0.961	0.971	0.966	0.954	0.954	0.951	0.957	0.963	0.959	0.950	0.954
	MCR	0.013	0.016	0.012	0.015	0.011	0.012	0.017	0.017	0.018	0.016	0.013	0.015	0.018	0.017
	TPR	0.871	0.815	0.882	0.817	0.856	0.813	0.864	0.802	0.841	0.812	0.831	0.793	0.800	0.780
	FPR	0.169	0.081	0.170	0.082	0.167	0.075	0.146	0.059	0.143	0.067	0.133	0.073	0.126	0.067
CMMNW	BIC	4740.651	4717.801	4737.247	4720.284	4741.710	4730.299	4720.810	4687.830	4710.420	4691.778	4726.093	4702.495	4718.224	4702.536
		(78.326)	(88.046)	(79.950)	(86.757)	(82.784)	(86.597)	(89.084)	(93.393)	(85.224)	(95.311)	(78.470)	(93.487)	(83.362)	(99.795)
	Freq.1	46	76	55	76	70	85	67	79	68	89	62	82	76	93
	Freq.2	1	0	0	0	0	0	4	50	4	35	2	0	1	3
	ARI	0.966	0.973	0.962	0.973	0.958	0.971	0.957	0.963	0.951	0.966	0.960	0.966	0.953	0.967
	MCR	0.012	0.010	0.014	0.010	0.015	0.010	0.015	0.013	0.018	0.012	0.014	0.012	0.017	0.012
	TPR	0.891	0.822	0.889	0.821	0.861	0.819	0.903	0.839	0.856	0.836	0.862	0.830	0.851	0.834
	FPR	0.226	0.130	0.234	0.129	0.188	0.123	0.221	0.148	0.144	0.142	0.175	0.146	0.134	0.140

Table 4
Simulation results for the poor-separated mixture of t-slice distributions.

Mixture	Meas.	Parsimonious variant													
model		EII	VII	EEI	VEI	EVI	VVI	EEE	VEE	EVE	VVE	EEV	VEV	EVV	VVV
CrSN	BIC	4571.555	4579.788	4575.000	4586.249	4581.713	4591.093	4560.880	4570.632	4522.579	4545.454	4531.079	4549.720	4529.378	4559.886
		(64.584)	(65.375)	(62.788)	(62.432)	(65.313)	(65.811)	(63.929)	(63.607)	(69.698)	(84.489)	(66.561)	(78.805)	(68.340)	(82.198)
	Freq.1	64	67	69	68	77	78	81	89	95	95	96	94	97	94
	Freq.2	0	0	0	1	0	0	5	4	58	5	14	2	11	0
	ARI	0.738	0.733	0.733	0.724	0.734	0.733	0.752	0.764	0.785	0.794	0.802	0.794	0.790	0.789
	MCR	0.093	0.096	0.097	0.103	0.094	0.094	0.092	0.089	0.080	0.072	0.071	0.074	0.078	0.076
	TPR	0.790	0.702	0.787	0.711	0.756	0.702	0.788	0.702	0.813	0.682	0.774	0.700	0.803	0.780
	FPR	0.257	0.094	0.262	0.098	0.188	0.099	0.288	0.092	0.264	0.069	0.173	0.089	0.239	0.068
CMMNE	BIC	4577.817	4589.910	4582.160	4598.480	4590.310	4603.901	4571.968	4589.285	4521.774	4556.441	4529.723	4553.988	4528.702	4564.609
		(64.998)	(64.751)	(63.865)	(63.516)	(65.034)	(64.072)	(60.801)	(61.956)	(65.182)	(72.843)	(62.878)	(73.934)	(66.449)	(73.785)
	Freq.1	72	80	79	75	80	83	91	83	95	82	98	92	97	83
	Freq.2	0	0	0	0	0	0	2	0	59	5	19	1	12	2
	ARI	0.744	0.737	0.734	0.736	0.735	0.732	0.765	0.761	0.799	0.787	0.798	0.797	0.798	0.791
	MCR	0.092	0.094	0.099	0.098	0.095	0.097	0.086	0.091	0.070	0.079	0.074	0.075	0.075	0.077
	TPR	0.781	0.675	0.785	0.669	0.749	0.641	0.767	0.649	0.727	0.695	0.724	0.585	0.709	0.692
	FPR	0.240	0.087	0.263	0.084	0.190	0.080	0.247	0.078	0.245	0.040	0.171	0.063	0.235	0.041
CMMNW	BIC	4572.226	4578.083	4575.125	4585.381	4580.830	4589.518	4559.996	4568.983	4537.030	4559.428	4544.170	4560.742	4546.906	4573.013
		(78.326)	(88.046)	(79.950)	(86.757)	(82.784)	(86.597)	(89.084)	(93.393)	(85.224)	(95.311)	(78.470)	(93.487)	(83.362)	(99.795)
	Freq.1	61	61	64	63	70	74	81	79	88	97	89	92	90	94
	Freq.2	1	1	0	1	0	0	11	11	50	5	13	4	3	0
	ARI	0.727	0.730	0.730	0.716	0.728	0.731	0.723	0.753	0.776	0.795	0.780	0.773	0.784	0.789
	MCR	0.098	0.098	0.097	0.109	0.096	0.095	0.102	0.093	0.083	0.072	0.080	0.084	0.081	0.076
	TPR	0.799	0.716	0.793	0.724	0.761	0.721	0.803	0.733	0.818	0.710	0.786	0.737	0.802	0.706
	FPR	0.269	0.120	0.281	0.126	0.195	0.124	0.312	0.140	0.273	0.093	0.191	0.146	0.235	0.095

that the clustering and outliers detection performances of the fitted mixtures are suffered from the group overlapping. However, almost all models can detect more than 70% of outliers in the data.

5.3. Model comparison when data are came form a mixture of heavy-tailed distributions

Due to the similarity in their distributional forms, the last simulation study is conducted to examine the performance of the CMMN-MIX submodels when data are generated from a strongly-skewed and heavy-tailed model. To proceed with the experimental study, we use the mean–variance mixture of multivariate normal (MVMN) scheme for data generation. Specifically, in each replication of 100 trials, bivariate data are simulated from a three-component MVMN as

$$\boldsymbol{x}_i = \boldsymbol{\mu}_j + u_i\boldsymbol{\lambda}_j + \sqrt{u_i}\boldsymbol{z}_i, \qquad \text{with probability} \quad \pi_j,$$

where u_i is generated from the Birnbaum–Saunders distribution (Birnbaum and Saunders [30]; BS(ν_j, 1)) and $\boldsymbol{z}_i$ comes from $\mathrm{N}_2(\boldsymbol{0}, \boldsymbol{\Sigma}_j)$. As pointed out in Section 1, Naderi et al. [15] showed that the class of MVMN and MMN offer different contours and orientation. The considered true parameters are: $\boldsymbol{\pi} = (0.25, 0.35, 0.4)$, $\boldsymbol{\nu} = (\nu_1, \nu_2, \nu_3) = (0.5, 1, 1.5)$,

Table 5
Simulation results for fitting performance comparison of four contaminated mixtures under simulation scenario 5.3.

Mixture		Parsimonious variant													
model		EII	VII	EEI	VEI	EVI	VVI	EEE	VEE	EVE	VVE	EEV	VEV	EVV	VVV
CN	BIC	2333.535	2409.875	2324.507	2417.445	2340.644	2413.138	2276.366	2283.364	2387.038	2392.598	2309.242	2326.026	2334.350	2347.794
	Freq.1	42	26	56	23	42	31	65	67	21	7	45	36	43	25
	Freq.2	0	0	0	0	0	0	7	10	16	2	26	17	11	11
CrSN	BIC	2209.216	2227.595	2210.696	2232.441	2212.956	2241.419	2216.611	2229.701	2209.682	2224.497	2207.229	2217.803	2213.651	2227.505
	Freq.1	98	86	97	86	100	80	98	95	98	94	100	97	100	97
	Freq.2	22	1	6	0	6	0	1	0	12	2	38	7	4	1
CMMNE	BIC	2209.249	2233.647	2212.708	2239.579	2217.901	2245.928	2213.851	2229.864	2210.406	2233.145	2201.822	2221.552	2214.175	2231.412
	Freq.1	95	89	96	85	97	84	97	92	94	93	98	96	95	96
	Freq.2	28	0	4	0	3	0	3	0	13	2	37	4	5	1
CMMNW	BIC	2212.708	2226.654	2214.111	2230.506	2213.888	2238.804	2221.623	2228.152	2210.215	2224.154	2209.234	2217.174	2215.784	2229.544
	Freq.1	98	91	99	88	100	87	98	97	99	98	99	99	100	98
	Freq.2	15	4	1	0	14	0	1	0	21	0	35	6	3	0

Table 6
Average of ARI, MCR, and TPR values over 100 replication for clustering and outliers detection performances of four contaminated mixtures studied in simulation scenario 5.3.

Mixture		Parsimonious variant													
model		EII	VII	EEI	VEI	EVI	VVI	EEE	VEE	EVE	VVE	EEV	VEV	EVV	VVV
CN	ARI	0.827	0.629	0.839	0.616	0.822	0.607	0.876	0.832	0.637	0.514	0.824	0.721	0.748	0.638
	MCR	0.059	0.193	0.058	0.199	0.060	0.215	0.043	0.075	0.199	0.307	0.067	0.155	0.130	0.209
	TPR	0.818	0.523	0.843	0.513	0.771	0.516	0.867	0.846	0.486	0.223	0.722	0.572	0.575	0.418
CrSN	ARI	0.899	0.898	0.895	0.896	0.896	0.902	0.893	0.879	0.895	0.889	0.894	0.893	0.894	0.891
	MCR	0.034	0.034	0.035	0.035	0.034	0.033	0.036	0.042	0.036	0.039	0.036	0.037	0.036	0.038
	TPR	0.903	0.916	0.906	0.911	0.907	0.889	0.894	0.911	0.942	0.888	0.925	0.883	0.924	0.872
CMMNE	ARI	0.914	0.878	0.901	0.885	0.914	0.890	0.914	0.888	0.902	0.887	0.914	0.897	0.902	0.894
	MCR	0.032	0.041	0.033	0.039	0.032	0.037	0.032	0.038	0.033	0.039	0.032	0.036	0.033	0.037
	TPR	0.845	0.745	0.844	0.721	0.827	0.705	0.844	0.828	0.844	0.662	0.920	0.769	0.818	0.738
CMMNW	ARI	0.883	0.890	0.881	0.888	0.884	0.888	0.868	0.857	0.885	0.877	0.881	0.886	0.882	0.880
	MCR	0.040	0.037	0.040	0.038	0.039	0.038	0.045	0.051	0.040	0.044	0.040	0.040	0.041	0.042
	TPR	0.926	0.914	0.92	0.907	0.904	0.896	0.911	0.905	0.951	0.881	0.958	0.893	0.943	0.862

$\boldsymbol{\mu}_1^\top = (9, 6)$ $\boldsymbol{\lambda}_1^\top = (1, 1)$, $\boldsymbol{\mu}_2^\top = (-5, -5)$, $\boldsymbol{\lambda}_2^\top = (2, 2)$, $\boldsymbol{\mu}_3^\top = (3, 13)$, $\boldsymbol{\lambda}_3^\top = (-0.5, -0.5)$, and

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix}, \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 2 & -0.6 \\ -0.6 & 1.5 \end{pmatrix}.$$

For each simulated dataset of size 200, we also add some noise points of size 10 uniformly generated in the range of the data. The intervals $(\max_{X_{11}}, \max_{X_{11}} +3)$ and $(\max_{X_{21}}, \max_{X_{21}} +5)$, $(\min_{X_{12}}, -5)$ and $(\min_{X_{22}} -5, \min_{X_{22}})$, and $(-12,-7)$ and $(15, 10)$ are considered for the noise generation of the first, second, and third group, respectively. We, therefore, have 200 good points (possibly with some outliers which are unknown) and 10 (5%) prespecified bad points in the data.

A comparison with respect to the fourteen constrained and unconstrained variants of the four contaminated mixture models is provided. These mixtures are CrSN-MIX, CMMNE-MIX, CMMNW-MIX, and the contaminated normal mixture (CN-MIX). In this simulation study, it is assumed that the number of mixture components G is unknown. Therefore, all models are fitted to the generated data with G ranging from 1 to 5 in each replication. Table 5 lists the average of BIC values over 100 trails for the parsimonious variants summarized in Table 1. For choosing the most plausible model, Table 5 also reports FrEq. (1) and FrEq. (2). It can be seen from FrEq. (1) that the skew contaminated mixture models, CrSN-MIX, CMMNE-MIX, and CMMNW-MIX, significantly perform better than the CN-MIX model for estimating the true value of G. Focusing on the results FrEq. (2) in Table 1, one can conclude that the CrSN-MIX, CMMNE-MIX, and CMMNW-MIX models suggest constrained versions of the covariance matrix with $\omega_j = \omega$, respectively with rate 0.89, 0.93, and 0.90. Overall, the average of BIC values supports the outperformance of the CMMNE-MIX model with the EEV covariance structure in terms of model selection.

For the sake of clustering comparison, Table 6 reports the average value of ARI and MCR criteria for $G = 3$. Moreover, we report the TPR for evaluating outliers detection performance. Results depicted in Table 6 suggest that the skew mixtures with condition $\omega_j = \omega$ in covariance matrices significantly improved the clustering and outliers detection performances. Subsequently, it could clearly be observed from Tables 5 and 6 that the CMMNE-MIX model with EEV covariance structure is the best for analyzing data generated under the simulation scenario considered in this section.

Table 7
Mardia's multivariate skewness and Kurtosis and GSW test with p-values in parentheses for the wine data.

	Name	Size	Skewness	Kurtosis	GSW	Overall	
Class 1	Barolo	59	47.2702 (0.3648)	181.3741 (0.0080)	0.96231 (3.935e−06)	Skewness:	42.7782 (5.4482e−78)
Class 2	Grignolino	71	57.6820 (2.3277e−11)	205.2270 (0.0291)	0.95379 (3.648e−13)	Kurtosis:	222.5919 (1.1602e−20)
Class 3	Barbera	48	50.444 (0.9601)	178.3657 (0.0035)	0.97209 (0.156)	GSW:	0.9607 (<2.2321e−16)

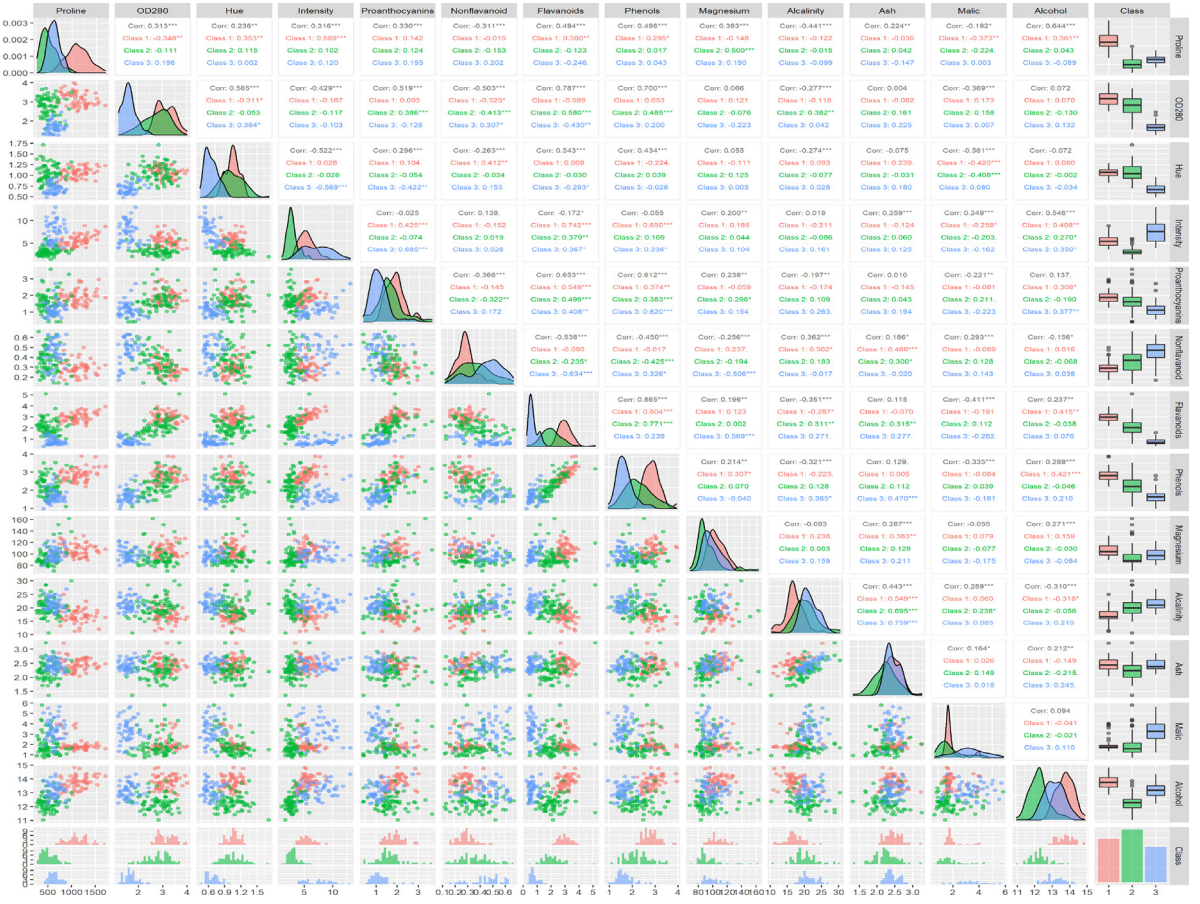

Fig. 4. The pairwise scatter plots of the 13 attributes of the wine data with the density curves in the diagonals.

6. Applications to Italian wine recognition data

The real data analysis focuses on the Italian wine recognition dataset available at UCI Machine Learning Repository (http://archive.ics.uci.edu/ml) and in the "**pgmm**" package of R. The dataset comprises 13-dimensional chemical measurements of $n = 178$ wines grown in three different cultivars (classes), Barolo, Grignolino, and Barbera. Table 7 presents some initial descriptive statistics of the data, including Mardia's multivariate skewness and Kurtosis, the generalized Shapiro–Wilk test for multivariate normality (GSW; Villasenor Alva and Estrada [31]) with their corresponding p-values for each class and overall data samples. The Mardia's multivariate skewness in Table 7 strongly recommends that Grignolino class is skew while the Mardia's multivariate Kurtosis highlights the heavy-tail behavior for all classes. Moreover, the GSW test suggests that the multivariate normality assumption is rejected in favor of the multivariate skew and heavy-tailed distribution. Fig. 4 shows the scatter plots of the 13 attributes of the wine data with the density curves in the diagonals. It can be seen that most of the attributes do not have well-separated scatter.

We fit our proposed mixture models, CrSN-MIX, CMMNE-MIX and CMMNW-MIX, and CN-MIX distributions to the data with G ranging from 1 to 5. We find that all models recommend the best fit to the data, in terms of BIC, for $G = 3$. The detailed numerical results, including the BIC, ARI, and MCR scores for $G = 3$ are reported in Table 8. The bold value of BIC for each model represents the best fit over all parsimonious structures. Moreover, The bold values of ARI and MCR for each model represent the clustering performance (with the smallest BIC if they are equal) over all parsimonious structures. It is observed from Table 8 that parsimonious models have smaller BIC values than those with the VVV structure. Furthermore,

Table 8
Fitting and clustering results for each of the four competing contaminated mixtures for the wine dataset.

Mixture		Parsimonious variant													
model		EII	VII	EEI	VEI	EVI	VVI	EEE	VEE	EVE	VVE	EEV	VEV	EVV	VVV
CN	BIC	21709.040	21432.443	7011.999	7008.105	6942.448	6943.741	6936.314	7058.691	6835.764	**6812.246**	7331.313	7315.496	7258.388	7183.994
	MCR	0.2753	0.2809	0.0449	0.0393	0.0562	0.0506	**0.0000**	0.2247	0.1011	0.0505	0.3820	0.4719	0.2191	0.2471
	ARI	0.4007	0.4070	0.8649	0.8819	0.8309	0.8483	**1.0000**	0.5793	0.7133	0.8469	0.3115	0.2501	0.4996	0.4702
CrSN	BIC	22737.471	23955.570	**6776.933**	6788.711	6786.626	6788.390	6878.386	6900.626	7027.805	7030.439	7259.906	7233.911	7336.685	7324.245
	MCR	0.2753	0.3820	0.0169	0.0112	0.0225	0.0112	**0.0000**	0.0000	0.0112	0.0000	0.0000	0.0000	0.0112	0.0169
	ARI	0.3718	0.1864	0.9487	0.9649	0.9295	0.9637	**1.0000**	1.0000	0.9667	1.0000	1.0000	1.0000	0.9651	0.9472
CMMNE	BIC	22526.069	23931.786	**6793.569**	6797.597	6799.179	6800.820	6886.552	6918.624	6838.509	6886.820	7235.600	7226.009	7294.791	7307.154
	MCR	0.3258	0.3652	0.0225	0.0225	0.0281	0.0169	0.0112	0.0056	0.0112	**0.0056**	0.0056	0.0056	0.0393	0.0169
	ARI	0.2937	0.2564	0.9326	0.9326	0.9134	0.9472	0.9667	0.9833	0.9651	**0.9850**	0.9833	0.9833	0.8785	0.9472
CMMNW	BIC	22800.244	23969.412	**6779.027**	6789.491	6787.736	6794.072	6876.397	6892.284	7119.443	7171.141	7295.374	7286.731	7384.455	7370.332
	MCR	0.3089	0.3820	0.0112	0.0112	0.0169	0.0169	**0.0000**	0.0000	0.0112	0.0056	0.0056	0.0112	0.0056	0.0056
	ARI	0.3368	0.1766	0.9649	0.9649	0.9471	0.9471	**1.0000**	1.0000	0.9667	0.9832	0.9817	0.9637	0.9817	0.9817

it can be seen that the proposed contaminated mixture models with the EEI covariance structure significantly fit the data better than the CN-MIX distributions since they offer smaller BIC scores.

7. Conclusions and discussions

This paper developed a new family of finite mixture models where the components follow the class of contaminated mean-mixture distributions, called the CMMN-MIX model. It was shown that the pdf of the CMMN-MIX model is indeed a mixture of two MMN distributions, one with a high priori mixing proportion for describing ordinary ("good") data and the other with a low priori mixing proportion for describing outliers ("dad") points. The latter component had the same mean parameter as the former, but its shape and scale parameters were inflated by a contamination factor. This formulation of the CMMN-MIX model allows the user to have more options for model-based density estimation and to simultaneously cluster heterogeneous data and detect outliers. Parameter estimation was outlined within the expectation–maximization framework after presenting a convenient hierarchical representation for the CMMN-MIX distributions. In order to address the over-parameterization issue in the scale matrices, parsimonious variants of the CMMN-MIX model were also introduced based on the eigenvalue decomposition of $\boldsymbol{\Sigma}_j$. The performance and effectiveness of the model were examined by conducting some simulation studies. To illustrate the usefulness of the CMMN-MIX model, we also analyzed Italian wine recognition datasets. The simulation and real-data results suggested the CMMN-MIX model could be an alternative benchmarked for clustering and outliers detection purposes. All computational tasks are preformed in R language and the computer programs are available at https://github.com/MerNaderi/CMMN-MIX.git.

There are some open avenues for future research development. It is known that analyzing data with missing and/or censored values has found interest in a wide range of disciplines. A further development will be to consider the class of CMMN-MIX distributions to propose a flexible for analyzing skewed data with censored and/or missing values [32,33]. It would also be of interest to extend the current approach to formulate a new mixture of factor analyzer models simultaneously for clustering and outlier detection of high-dimensional data [34].

Data availability

Data will be made available on request.

Acknowledgments

The authors gratefully acknowledge the Editor, Fatih Tank, and an anonymous referee for their comments and suggestions that significantly improved the quality of this paper. Mehrdad Naderi was partially supported by a grant from Ferdowsi University of Mashhad (No. FUM-14002794075).

Appendix. Additional simulation study

This simulation study is interested in inspecting the consistency of the standard error estimates. Toward this end, we consider true parameters used in Section 5.1 with $\nu_1 = 2, \nu_1 = 3, \nu_1 = 1.5$, $\gamma_1 = 0.9, \gamma_2 = 0.85$, and $\gamma_3 = 0.95$ to generate data from a CMMNE-MIX distributions. In the replications of this study, the empirical standard deviations (STD) of the ML estimates across 500 simulated samples and the average standard errors (IMSE) obtained through the inverse of the outer product of gradient scores (15) are calculated as

$$\mathrm{STD}(\hat{\theta}_i) = \sqrt{\frac{1}{499}\left[\sum_{r=1}^{500}\left(\hat{\theta}_i^{(r)}\right)^2 - \frac{1}{500}\left(\sum_{r=1}^{500}\hat{\theta}_i^{(r)}\right)^2\right]} \quad \text{and} \quad \mathrm{IMSE}(\hat{\theta}_i) = \frac{1}{500}\sum_{r=1}^{500}\mathrm{SE}(\hat{\theta}_i^{(r)}),$$

Table A.1
Simulation results for assessing the consistency of standard errors.

n		$\boldsymbol{\pi}$		$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_3$	$\boldsymbol{\lambda}_1$	$\boldsymbol{\lambda}_2$	$\boldsymbol{\lambda}_3$	$\boldsymbol{\Sigma}_1$		$\boldsymbol{\Sigma}_2$		$\boldsymbol{\Sigma}_3$		$\boldsymbol{\nu}$		$\boldsymbol{\gamma}$	
		π_1	π_2	μ_{11}	μ_{21}	μ_{31}	λ_{11}	λ_{21}	λ_{31}	$\sigma_{1,11}$	–	$\sigma_{2,11}$	–	$\sigma_{3,11}$	–	ν_1	ν_2	γ_1	γ_2
				μ_{12}	μ_{22}	μ_{32}	λ_{12}	λ_{22}	λ_{32}	$\sigma_{1,21}$	$\sigma_{1,22}$	$\sigma_{2,21}$	$\sigma_{2,22}$	$\sigma_{3,21}$	$\sigma_{3,22}$	ν_3	–	γ_3	–
100	STD	0.105	0.098	0.657	0.391	1.239	0.531	0.515	0.510	0.612	–	1.115	–	0.732	–	1.450	1.329	0.230	0.197
				0.670	0.463	0.580	1.022	0.934	0.922	0.200	0.569	0.121	0.540	0.555	0.633	0.717	–	0.218	–
	IMSE	0.073	0.064	0.433	0.219	0.891	0.367	0.354	0.343	0.398	–	0.730	–	0.462	–	1.103	0.986	0.178	0.141
				0.451	0.275	0.477	0.729	0.594	0.585	0.122	0.373	0.089	0.362	0.380	0.411	0.531	–	0.169	–
500	STD	0.075	0.061	0.361	0.278	0.898	0.351	0.347	0.340	0.502	–	0.819	–	0.606	–	1.014	0.927	0.200	0.107
				0.374	0.339	0.356	0.822	0.791	0.780	0.114	0.408	0.096	0.388	0.401	0.437	0.396	–	0.147	–
	IMSE	0.052	0.049	0.212	0.186	0.569	0.283	0.266	0.262	0.345	–	0.624	–	0.388	–	0.839	0.652	0.165	0.088
				0.227	0.242	0.260	0.619	0.467	0.449	0.077	0.313	0.057	0.299	0.295	0.317	0.261	–	0.104	–
1000	STD	0.031	0.028	0.232	0.173	0.504	0.263	0.255	0.254	0.376	–	0.530	–	0.400	–	0.773	0.630	0.134	0.086
				0.256	0.203	0.239	0.501	0.473	0.465	0.081	0.259	0.067	0.214	0.274	0.251	0.308	–	0.092	–
	IMSE	0.027	0.025	0.183	0.125	0.368	0.209	0.193	0.190	0.287	–	0.464	–	0.411	–	0.839	0.509	0.114	0.073
				0.210	0.159	0.193	0.348	0.312	0.306	0.054	0.209	0.044	0.173	0.203	0.196	0.227	–	0.082	–

where $\mathrm{SE}(\hat{\theta}_i^{(r)})$ denotes the asymptotic standard errors of $\hat{\theta}_i$ at the rth replication (see Section 3.4). The detailed numerical results are presented in Table A.1. It can be seen that the empirical standard errors (STD) tend to get closer to IMSE, suggesting that the proposed asymptotic approximation through (15) for calculating the standard errors of the ML estimates are reliable.

References

[1] R.A. Redner, H.F. Walker, Mixture densities, maximum likelihood and the EM algorithm, SIAM Rev. 26 (2) (1984) 195–239.
[2] A. Punzo, P.D. McNicholas, Parsimonious mixtures of multivariate contaminated normal distributions, Biom. J. 58 (6) (2016) 1506–1537.
[3] M. Naderi, W.-L. Hung, T.-I. Lin, A. Jamalizadeh, A novel mixture model using the multivariate normal mean-variance mixture of Birnbaum–saunders distributions and its application to extrasolar planets, J. Multivariate Anal. 171 (2019) 126–138.
[4] K. Morris, A. Punzo, P.D. McNicholas, R.P. Browne, Asymmetric clusters and outliers: Mixtures of multivariate contaminated shifted asymmetric Laplace distributions, Comput. Stat. Data Anal. 132 (2019) 145–166.
[5] A. Farcomeni, A. Punzo, Robust model-based clustering with mild and gross outliers, TEST 29 (4) (2019) 989–1007.
[6] M. Naderi, F. Hashemi, A. Bekker, A. Jamalizadeh, Modeling right-skewed financial data streams: A likelihood inference based on the generalized birnbaum-saunders mixture model, Appl. Math. Comput. 376 (2020) 125109.
[7] G. Celeux, G. Govaert, Gaussian parsimonious clustering models, Pattern Recognit. 28 (5) (1995) 781–793.
[8] T. Benaglia, D. Chauveau, D.R. Hunter, An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures, J. Comput. Graph. Statist. 18 (2) (2009) 505–526.
[9] A. Azzalini, A class of distributions which includes the normal ones, Scand. J. Stat. (1985) 171–178.
[10] F. Vilca, N. Balakrishnan, C.B. Zeller, Multivariate skew-normal generalized hyperbolic distribution and its properties, J. Multivariate Anal. 128 (2014) 73–85.
[11] R.M. Basso, V.H. Lachos, C.R.B. Cabral, P. Ghosh, Robust mixture modeling based on scale mixtures of skew-normal distributions, Comput. Statist. Data Anal. 54 (12) (2010) 2926–2941.
[12] C.R.B. Cabral, V.H. Lachos, M.O. Prates, Multivariate mixture modeling using skew-normal independent distributions, Comput. Statist. Data Anal. 56 (1) (2012) 126–142.
[13] H. Negarestani, A. Jamalizadeh, S. Shafiei, N. Balakrishnan, Mean mixtures of normal distributions: properties, inference and application, Metrika 82 (4) (2019) 501–528.
[14] F. Hashemi, M. Naderi, A. Jamalizadeh, A. Bekker, A flexible factor analysis based on the class of mean-mixture of normal distributions, Comput. Stat. Data Anal. 157 (2021) 107162.
[15] M. Naderi, A. Bekker, M. Arashi, A. Jamalizadeh, A theoretical framework for landsat data modeling based on the matrix variate mean-mixture of normal model, in: D.C. Zanotta (Ed.), PLoS One 15 (4) (2020) e0230773.
[16] A.J. McNeil, R. Frey, P. Embrechts, Quantitative Risk Management: Concepts, Techniques and Tools-Revised Edition, Princeton University Press, 2015.
[17] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1) (1977) 1–22.
[18] M. Jones, On families of distributions with shape parameters, Internat. Statist. Rev. 83 (2) (2014) 175–192.
[19] S. Richardson, P.J. Green, On Bayesian analysis of mixtures with an unknown number of components (with discussion), J. R. Stat. Soc. Ser. B Stat. Methodol. 59 (4) (1997) 731–792.
[20] C. Liu, D.B. Rubin, The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence, Biometrika 81 (4) (1994) 633–648.
[21] X.-L. Meng, D.B. Rubin, Maximum likelihood estimation via the ECM algorithm: A general framework, Biometrika 80 (2) (1993) 267–278.
[22] J.D. Banfield, A.E. Raftery, Model-based Gaussian and non-Gaussian clustering, Biometrics (1993) 803–821.
[23] K. Basford, D. Greenway, G.J. McLachlan, D. Peel, Standard errors of fitted means under normal mixture models, Comput. Statist. 12 (1–17) (1997) 5If.
[24] I. Meilijson, A fast improvement to the EM algorithm on its own terms, J. R. Stat. Soc. Ser. B Stat. Methodol. 51 (1) (1989) 127–138.
[25] T.A. Louis, Finding the observed information matrix when using the EM algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. 44 (2) (1982) 226–233.
[26] A.C. Aitken, On Bernoulli's numerical solution of algebraic equations, Proc. Roy. Soc. Edinburgh 46 (1927) 289–305.
[27] B.G. Lindsay, Mixture models: theory, geometry and applications, in: NSF-CBMS Regional Conference Series in Probability and Statistics, JSTOR, 1995, pp. i–163.
[28] G. Schwarz, Estimating the dimension of a model, Ann. Statist. (1978) 461–464.

[29] L. Hubert, P. Arabie, Comparing partitions, J. Classification 2 (1) (1985) 193–218.
[30] Z.W. Birnbaum, S.C. Saunders, A new family of life distributions, J. Appl. Probab. 6 (2) (1969) 319–327.
[31] J.A. Villasenor Alva, E.G. Estrada, A generalization of Shapiro–Wilk's test for multivariate normality, Comm. Statist. Theory Methods 38 (11) (2009) 1870–1883.
[32] E. Mirfarah, M. Naderi, D.-G. Chen, Mixture of linear experts model for censored data: A novel approach with scale-mixture of normal distributions, Comput. Stat. Data Anal. 158 (2021) 107182.
[33] F.H.C. de Alencar, C.E. Galarza, L.A. Matos, V.H. Lachos, Finite mixture modeling of censored and missing data using the multivariate skew-normal distribution, Adv. Data Anal. Classif. 16 (3) (2021) 521–557.
[34] F. Hashemi, M. Naderi, M. Mashinchi, Clustering right-skewed data stream via Birnbaum-Saunders mixture models: A flexible approach based on fuzzy clustering algorithm, Appl. Soft Comput. 82 (2019) 105539.