Journal of Classification
https://doi.org/10.1007/s00357-023-09450-2

Missing Values and Directional Outlier Detection in Model-Based Clustering

Hung Tong[1] · **Cristina Tortora**[2]

Accepted: 6 September 2023

Abstract
Model-based clustering tackles the task of uncovering heterogeneity in a data set to extract valuable insights. Given the common presence of outliers in practice, robust methods for model-based clustering have been proposed. However, the use of many methods in this area becomes severely limited in applications where partially observed records are common since their existing frameworks often assume complete data only. Here, a mixture of multiple scaled contaminated normal (MSCN) distributions is extended using the expectation-conditional maximization (ECM) algorithm to accommodate data sets with values missing at random. The newly proposed extension preserves the mixture's capability in yielding robust parameter estimates and performing automatic outlier detection separately for each principal component. In this fitting framework, the MSCN marginal density is approximated using the inversion formula for the characteristic function. Extensive simulation studies involving incomplete data sets with outliers are conducted to evaluate parameter estimates and to compare clustering performance and outlier detection of our model to other mixtures.

Keywords Model-based clustering · Outliers · Missing data · Contaminated normal distribution · Multiple scaled distributions · EM algorithm

1 Introduction

Model-based clustering refers to the use of finite mixture models in cluster analysis. In these models, the population of interest is assumed to be a mixture of sub-populations, each of which is considered as a cluster and can be modeled by a probability distribution (McLachlan & Peel, 2000). The first choice for modeling each cluster has been the Gaussian distribution, which is appealing due to its computational and theoretical convenience. Dating back to 1965, many papers can be found on Gaussian mixture models (GMM); see, for example, Wolfe

✉ Cristina Tortora
cristina.tortora@sjsu.edu

Hung Tong
hmtong@crimson.ua.edu

[1] The University of Alabama, Tuscaloosa, AL 35487, USA

[2] San José State University, San José, CA 95192, USA

Published online: 31 October 2023

(1965), Banfield and Raftery (1993) and Celeux and Govaert (1995). Other distributions with greater flexibility have also been used in statistical literature, including but not limited to the skew normal distribution (Lin, 2009), the normal inverse Gaussian distribution (Karlis & Santourian, 2009), the shifted asymmetric Laplace distribution (SAL; Franczak et al. 2014), and the generalized hyperbolic distribution (GHD; Browne & McNicholas 2015).

The need for modeling each cluster with a more flexible distribution arises from the shape limitation of the Gaussian distribution. Specifically, being symmetric with lighter tails, models based on the Gaussian distribution lack robustness and often exhibit sensitivity to outlying observations. Even though various robust models that use heavy-tailed distributions have been proposed to overcome such a drawback in mixture modeling settings, choosing an appropriate model still largely depends on the structure of outliers at hand. Regarding that point, outliers may roughly be divided into two types: gross and mild (Ritter 2014, pp. 79-80). Outliers are gross when they do not appear to be sampled from a population. Consequently, they are unpredictable and incalculable, and no probability distribution can be used to sufficiently model them. When gross outliers are present, it is common to handle them by maximizing a trimmed likelihood of a partition model (Gallegos & Ritter, 2005, 2009) or by simultaneously clustering the non-noise observations and identifying a noise component; see, for example, Coretto and Hennig (2016) and Novi Inverardi and Taufer (2020).

In contrast, mild outliers, referred to as bad points in Aitkin and Wilson (1980), can be modeled using a weighted likelihood function or a more flexible distribution. A few approaches based on the weighted likelihood function have been proposed that show good performance in clustering and outlier detection, and represent a valid solution for data sets with no missing values; see for example, Greco and Agostinelli (2020), Sugasawa and Kobayashi (2022). Alternately, a more flexible distribution that is symmetric and endowed with heavy tails can be used, such as, the multivariate t (Mt) distribution (Peel & McLachlan, 2000; Andrews & McNicholas, 2012) or the multivariate contaminated normal (MCN) distribution (Punzo et al., 2018). Compared to the former, the MCN distribution can be more appealing due to its ability to automatically detect outliers. Furthermore, it has two additional parameters besides the mean vector and covariance matrix to characterize the proportion of good observations and the extra variability introduced by outliers, which greatly enhances interpretability. However, one important limitation that both distributions share is that they possess the exact same parameter(s) governing their tail behaviors in all dimensions. This limitation implies the following consequences.

1. All marginals are (M)t with the same degree of freedom or (M)CN distributions with the same proportion of good observations and degree of contamination, respectively, and thus, the same amount of tail weight.
2. Outliers are automatically down-weighted in the maximum likelihood estimation of each distribution's parameters but in the same way for each dimension.
3. Both distributions' outlier detection procedures could be defined as an omnibus in the sense that when a point is detected as bad, it is globally bad, and the specific dimension(s) to which it is outlying remains unknown.

To overcome the aforementioned limitation, multiple scaled distributions have been proposed, in which the idea is to introduce multidimensional weight random variables in a normal-scale mixture. Recall that a normal-scale mixture consists of different Gaussian distributions sharing the same mean vector, each of which has its covariance matrix determined by one realization of a univariate weight random variable. Many distributions can be derived from a normal-scale mixture; for example, choosing the weight random variable to follow a gamma distribution generates an Mt distribution, while choosing it to be Bernoulli results in an MCN distribution. To obtain a multiple-scaled distribution, the covariance matrix is first

decomposed using the eigen-decomposition, and then, a different weight random variable is assigned to each dimension spanned by the columns of the eigenvector matrix (the principal components in other words). The multiple scaled t distribution was proposed by Forbes and Wraith (2014), whereas the multiple scaled contaminated normal (MSCN) distribution was introduced by Punzo and Tortora (2021). Many other distributions have been transformed using the same approach to increase their flexibility; see, for example, Franczak et al. (2015), and Tortora et al. (2019).

Despite their extensive robustness and flexibility, all the mentioned models use only complete data sets without any missing values, which can be unreasonable given how ubiquitous missing values are in many real applications. Given that limitation, in this paper, we consider the problem of fitting mixtures of MSCN distributions with missing information. In dealing with data sets with missing values, determining the underlying missing data mechanism is critical to select an appropriate strategy. Hence, herein, we specifically focus on data missing at random (MAR; Little & Rubin, 2020), in which the probability for missingness to occur in some variates of a particular individual depends only on the values of other observed variates of such individual, but not on the values of the missing variates themselves.

When handling missing data, it might be tempting to apply case deletion, that is, excluding observations with partially observed information and proceeding with regular statistical methods. However, although convenient and simple, doing so comes at the cost of producing biased estimators which can lead to some invalid inferences. A more common approach is data imputation, in which missing values are filled in to result in a complete data set. Among the most popular imputation techniques are mean imputation (Wilks, 1932), regression imputation (Buck, 1960), and multiple imputations (Rubin, 1987, 1996). Alternately, there are also likelihood-based approaches where a statistical model is imposed so that inference and parameter estimation can be based on the likelihood under such a model. One well-known method under this category is the expectation-maximization (EM) algorithm (Dempster et al., 1977). In particular, the algorithm outlines an iterative procedure that alternates between an expectation (E) step and a maximization (M) step until convergence to obtain maximum likelihood parameter estimates in the presence of missing data and/or some latent variables. For a comprehensive survey of statistical methods for analyzing missing data, refer to Schafer and Graham (2002), Little and Rubin (2020), and Buuren (2021).

Parameter estimation in model-based clustering is also commonly obtained using the EM algorithm. Essentially, to fit in the EM framework, the clustering problem can be reformulated as an incomplete-data problem where the grouping information of every individual is unobserved. When data are not fully observed due to missingness, extending the EM algorithm to accommodate MAR values presents a sensible solution. This idea has been reflected by previous work in model-based clustering via various distributions: the multivariate normal distribution (Ghahramani & Jordan, 1994; Serafini et al., 2020), the Mt distribution (Wang et al., 2004; Goren & Maitra, 2022), the MCN distribution (Tong & Tortora, 2022b), the skew-t distribution (Wang & Lin, 2015), and the generalized hyperbolic distribution (Wei et al., 2019). Although other models have yet to be developed, this approach can be extended to models based on other distributions in which maximum likelihood parameters are estimated with the EM algorithm (or its variations). However, the existence and computability of the marginal and joint densities of the used distribution are necessary. The marginal distribution of the MSCN proposed by Punzo and Tortora (2021) is unknown, and thus, the extension of the MSCN to data sets with missing data is not trivial.

In this paper, we employ the expectation-conditional maximization (ECM) algorithm, a variant of the EM algorithm introduced by Meng and Rubin (1993), to propose a new framework for fitting mixtures of MSCN distributions in data sets whose values are MAR. The paper is organized as follows. In Sect. 2, we provide the necessary background of the MCN and

MSCN distributions, with a note that the former is not a special nor limiting case of the latter. In Sect. 3, we obtain the marginals of the MSCN distribution and some useful results related to them. From there, we then outline the ECM algorithm for parameter estimation in detail. In Sect. 4, we describe computational aspects when implementing the newly proposed model, namely initialization, convergence, cluster assignment, outlier detection, model selection, and other computational details regarding the ECM algorithm. In Sect. 5, we discuss results from extensive simulation studies where our model and other mixtures are benchmarked using incomplete data sets. In Sect. 6, we analyze the World Happiness Report data using our model. We conclude the paper in Sect. 7. Besides, Appendix A includes useful details on the characteristic functions and the inversion formula. Appendix B provides proofs for most propositions introduced in the paper. Supplementary material Section 1 covers mathematical details of the E-step. The tables summarizing results from the simulation studies are available in supplementary material Section 2.

2 Background

2.1 The Multivariate Contaminated Normal Distribution

The multivariate contaminated normal (MCN) distribution was first introduced by Tukey (1960) as a mixture of two multivariate normal (MN) distributions, one of which, despite sharing the same mean vector with the other, has an inflated covariance matrix to represent outliers. Mathematically, the probability density function (pdf) of a p-variate random vector $\boldsymbol{X} = \left(X_1, \ldots, X_p\right)^\top$ that follows an MCN distribution with mean vector $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, proportion of good observations $\alpha \in (0.5, 1)$, and degree of contamination $\eta > 1$ is given by

$$f_{\text{MCN}}\left(\boldsymbol{x}\,;\,\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \eta\right) = \alpha\, f_{\text{MN}}\left(\boldsymbol{x}\,;\,\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) + (1-\alpha)\, f_{\text{MN}}\left(\boldsymbol{x}\,;\,\boldsymbol{\mu}, \eta\boldsymbol{\Sigma}\right), \tag{1}$$

where $f_{\text{MN}}\left(\boldsymbol{x}\,;\,\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is the pdf of a p-variate normal random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Herein, the constraint of α to be within $(0.5, 1)$ is imposed to be consistent with the common assumption in robust statistics that at least half of the observations are good. In addition, the degree of contamination η multiplied by $\boldsymbol{\Sigma}$ captures the increase in variability due to the presence of outliers. It can be seen from Eq. 1 that as α and η tend to 1, we obtain the MN distribution as a limiting case of the MCN distribution. In general, decreasing α and/or increasing η has the effect of inflating the variability and kurtosis of the distribution, and, as a consequence, the tails. For more details on the variance and kurtosis of the MCN, see Appendix G of Bagnato et al. (2017).

Modeling data sets with outliers using the MCN distribution has several advantages. First, the two additional parameters α and η provide useful interpretations of the behavior of outlying observations. Second, robust estimation for the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ can be obtained, ensuring valid inference about the population. Third, the identifiability of the model is already shown in Punzo and McNicholas (2016). Last but not least, once all the parameters are estimated, a generic point $\boldsymbol{x}^*$ can be classified as outlying if its corresponding *a posteriori* probability does not exceed 0.5. This outlier detection procedure follows from the maximum *a posteriori* clustering procedure, i.e., once the cluster membership is defined, a point is flagged as an outlier or not based on the value of its *a posteriori* probability of belonging to the outlying component for that cluster. In addition, it is done automatically as a byproduct of parameter estimation without the need for a subjective threshold.

2.2 The Multiple Scaled Contaminated Normal Distribution

Punzo and Tortora (2021) introduced the multiple scaled contaminated normal (MSCN) distribution to address some drawbacks of the MCN distribution. Based on the idea of Forbes and Wraith (2014), the MSCN distribution is obtained by decomposing the covariance matrix $\boldsymbol{\Sigma}$ into eigenvalues and eigenvectors matrices, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$, and then introducing Bernoulli random variables V's each indicating whether a point is good or outlying separately for each principal component of the space spanned by the columns of $\boldsymbol{\Gamma}$. The parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^\top$ are now two vectors controlling the proportions of good points and degrees of contamination for all principal components. Formally, the pdf of the MSCN distribution can be written as

$$f_{\text{MSCN}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \prod_{h=1}^{p} f_{\text{CN}}\left(\left[\boldsymbol{\Gamma}^\top(\boldsymbol{x} - \boldsymbol{\mu})\right]_h; 0, \lambda_h, \alpha_h, \eta_h\right), \tag{2}$$

where p is the number of variables, $\boldsymbol{\Gamma}^\top(\boldsymbol{x} - \boldsymbol{\mu})$ is the principal-component transform of $\boldsymbol{x}$, or equivalently a rotation and a re-centering of $\boldsymbol{x}$, $\left[\boldsymbol{\Gamma}^\top(\boldsymbol{x} - \boldsymbol{\mu})\right]_h$ is the hth element of $\boldsymbol{\Gamma}^\top(\boldsymbol{x} - \boldsymbol{\mu})$, and λ_h is the hth eigenvalue of the matrix $\boldsymbol{\Lambda}$. It can be shown that when $\boldsymbol{X} \sim \mathcal{MSCN}_p(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \boldsymbol{\alpha}, \boldsymbol{\eta})$,

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}\boldsymbol{W}_{\boldsymbol{V}}^{1/2}\boldsymbol{Y}, \tag{3}$$

where $\boldsymbol{V} = \left(V_1, \ldots, V_p\right)^\top$, $\boldsymbol{Y} = \left(Y_1, \ldots, Y_p\right)^\top \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$, and

$$\boldsymbol{W}_{\boldsymbol{V}} = \text{diag}\left\{\left(V_1 + \frac{1 - V_1}{\eta_1}\right)^{-1}, \ldots, \left(V_p + \frac{1 - V_p}{\eta_p}\right)^{-1}\right\}. \tag{4}$$

The MSCN distribution has more flexibility in terms of symmetric shapes and tail behaviors in different principal components compared to the MCN distribution which is constrained to be elliptical. Tortora et al. (2019) showed that multiple scaled distributions are identifiable up to multiplication for negative one of $\boldsymbol{\Gamma}$, if the univariate distribution they are based on is identifiable. On the other hand, Punzo and McNicholas (2016) showed that mixtures of CN distributions are identifiable, and therefore mixtures of MSCN distributions are identifiable under the sign conditions discussed in Tortora et al. (2019).

3 Methodology

3.1 Marginals of the Multiple Scaled Contaminated Normal Distribution

Apart from the univariate case, the MCN distribution is not a special nor limiting case of the MSCN distribution; therefore, the marginal distribution of the MSCN is unknown and needs to be derived from the characteristic functions. Some concepts and formulas that will be used in this section can be found in Appendix A. From Eq. 3, we denote

$$\tilde{\boldsymbol{Y}} = \boldsymbol{W}_{\boldsymbol{V}}^{1/2}\boldsymbol{Y} = \left(\left(V_1 + \frac{1 - V_1}{\eta_1}\right)^{-1/2} Y_1, \ldots, \left(V_p + \frac{1 - V_p}{\eta_p}\right)^{-1/2} Y_p\right)^\top. \tag{5}$$

In this notation, $\tilde{\boldsymbol{Y}}$ is a vector of p independent univariate contaminated normal random variables with mean 0 and variance 1, each of which has its own proportion of good observations and degree of contamination. Marginals of the MSCN distribution turn out to be linear

combinations of the independent components of $\tilde{\boldsymbol{Y}}$ for which no closed-form expression is available in general. However, marginal densities can still be obtained by means of the inversion formula, in which the first step is to identify the characteristic function of a marginal variable. The following propositions introduce the characteristic functions of the MCN and MSCN distributions.

Proposition 3.1 Let $\boldsymbol{X} \in \mathbb{R}^p$ be a p-variate random vector that follows a multivariate contaminated normal distribution with mean vector $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, proportion of good observations $\alpha \in (0.5, 1)$, and degree of contamination $\eta > 1$. The characteristic function of $\boldsymbol{X}$ is

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \alpha \exp\left(i\boldsymbol{t}^\top \boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^\top \boldsymbol{\Sigma}\boldsymbol{t}\right) + (1-\alpha)\exp\left(i\boldsymbol{t}^\top \boldsymbol{\mu} - \frac{1}{2}\eta\boldsymbol{t}^\top \boldsymbol{\Sigma}\boldsymbol{t}\right), \tag{6}$$

where $\boldsymbol{t} = \left(t_1, \ldots, t_p\right)^\top \in \mathbb{R}^p$ and i is an imaginary unit.

Proof See Appendix B □

Proposition 3.2 Let $\boldsymbol{X} \in \mathbb{R}^p$ be a p-variate multiple scaled contaminated normal random vector with mean vector $\boldsymbol{\mu}$, eigenvalues matrix $\boldsymbol{\Lambda}$, eigenvectors matrix $\boldsymbol{\Gamma}$, proportions of good observations $\boldsymbol{\alpha} = \left(\alpha_1, \ldots, \alpha_p\right)^\top$, and degrees of contamination $\boldsymbol{\eta} = \left(\eta_1, \ldots, \eta_p\right)^\top$. Consider a positive integer $q \leq p$ and partition $\boldsymbol{X}$ as

$$\boldsymbol{X} = \begin{bmatrix} \underset{q\times 1}{\boldsymbol{X}_1} \\ \underset{(p-q)\times 1}{\boldsymbol{X}_2} \end{bmatrix}.$$

Then, the characteristic function of the marginal variable $\boldsymbol{X}_1$ is given by

$$\phi_{\boldsymbol{X}_1}(\boldsymbol{t}) = \prod_{j=1}^{q} \exp(it_j\mu_j) \prod_{h=1}^{p} \phi_{\tilde{Y}_h}\left(\sum_{j=1}^{q} t_j [\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{jh}\right), \tag{7}$$

where

$$\phi_{\tilde{Y}_h}\left(\sum_{j=1}^{q} t_j [\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{jh}\right) = \alpha_h \exp\left[-\frac{1}{2}\left(\sum_{j=1}^{q} t_j [\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{jh}\right)^2\right]$$
$$+(1-\alpha_h)\exp\left[-\frac{1}{2}\eta_h\left(\sum_{j=1}^{q} t_j [\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{jh}\right)^2\right], \tag{8}$$

$\boldsymbol{t} = \left(t_1, \ldots, t_q\right)^\top \in \mathbb{R}^q$ and i is an imaginary unit.

Proof See Appendix B □

Using the inversion formula described in Theorem A.1, we can obtain the pdf of $\boldsymbol{X}_1$, which is the marginal density of $\boldsymbol{X}$. However, evaluating such marginal density involves a numerical procedure for multiple integrations such as the adaptive multivariate integration over hypercubes (see, for example, Dooren & Ridder 1976; Berntsen et al. 1991).

Here, we outline some useful propositions regarding an MSCN random vector and its marginals when conditioned on some directional good-observation indicator random variables. Let $\boldsymbol{X} \in \mathbb{R}^p$ be a p-variate MSCN random vector with mean vector $\boldsymbol{\mu}$, eigenvalues

matrix $\boldsymbol{\Lambda}$, eigenvectors matrix $\boldsymbol{\Gamma}$, proportions of good observations $\boldsymbol{\alpha} = \left(\alpha_1, \ldots, \alpha_p\right)^\top$, and degrees of contamination $\boldsymbol{\eta} = \left(\eta_1, \ldots, \eta_p\right)^\top$. Also let V_h be the indicator variable such that $V_h = 1$ if the hth element of the principal-component transformed random vector $\boldsymbol{\Gamma}^\top(\boldsymbol{X} - \boldsymbol{\mu})$ is good and $V_h = 0$ otherwise, for $h = 1, \ldots, p$. Consider a vector of 0/1 values $\boldsymbol{v} = \left(v_1, \ldots, v_p\right)^\top$ and the corresponding $p \times p$ diagonal matrix of inverse weights

$$\boldsymbol{W}_{\boldsymbol{v}} = \text{diag}\left\{\left(v_1 + \frac{1 - v_1}{\eta_1}\right)^{-1}, \ldots, \left(v_p + \frac{1 - v_p}{\eta_p}\right)^{-1}\right\}.$$

We have the following propositions regarding the relationship between the MSCN and MN distribution.

Proposition 3.3 (Punzo & Tortora, 2021) Given $V_1 = v_1, \ldots, V_p = v_p$, $\boldsymbol{X}$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{M} = \boldsymbol{\Gamma}\boldsymbol{W}_{\boldsymbol{v}}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top$.

Proposition 3.4 Let $q \leq p$ be a positive integer. If we partition $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{M}$ as

$$\boldsymbol{X} = \begin{bmatrix} \underset{q\times 1}{\boldsymbol{X}_1} \\ \underset{(p-q)\times 1}{\boldsymbol{X}_2} \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \underset{q\times 1}{\boldsymbol{\mu}_1} \\ \underset{(p-q)\times 1}{\boldsymbol{\mu}_2} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{M} = \begin{bmatrix} \underset{q\times q}{\boldsymbol{M}_{11}} & \underset{q\times(p-q)}{\boldsymbol{M}_{12}} \\ \underset{(p-q)\times q}{\boldsymbol{M}_{21}} & \underset{(p-q)\times(p-q)}{\boldsymbol{M}_{22}} \end{bmatrix},$$

then given $V_1 = v_1, \ldots, V_p = v_p$, $\boldsymbol{X}_1$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{M}_{11}$.

Proposition 3.5 With the same notations as in Proposition 3.4, the conditional distribution of $\boldsymbol{X}_2$, given $\boldsymbol{X}_1 = \boldsymbol{x}_1$ and $V_1 = v_1, \ldots, V_p = v_p$, is a multivariate normal distribution with mean vector and covariance matrix, respectively,

$$\boldsymbol{\mu}_2 + \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) \quad \text{and} \quad \boldsymbol{M}_{22} - \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}.$$

Given Proposition 3.3, the properties of the multivariate normal distribution can be applied to obtain Propositions 3.4 and 3.5. For more information on these properties, readers are invited to refer to Chapter 4 of Johnson and Wichern (2007).

Thus far, in this section, we have assumed the knowledge of all p indicator random variables $V_1, \ldots, V_p$ which reduces the MSCN distribution to the MN distribution with nice properties and closed-forms results. Now, in the following proposition, we adopt a more general view by assuming that we only know the values of a subset of the indicator random variables.

Proposition 3.6 Let $\mathcal{A}$ and $\mathcal{B}$ be two disjoint subsets of $\{1, \ldots, p\}$ such that $\mathcal{A} \cup \mathcal{B} = \{1, \ldots, p\}$. We define $V_{\mathcal{A}} = \{V_r, r \in \mathcal{A}\}$ and $V_{\mathcal{B}} = \{V_s, s \in \mathcal{B}\}$ in which we observe the value of every element in $V_{\mathcal{A}}$, that is, $v_r \in \{0, 1\}$, for $r \in \mathcal{A}$, but not any in $V_{\mathcal{B}}$. With the same notations as in Proposition 3.4, the characteristic function of the marginal variable $\boldsymbol{X}_1 \mid V_r = v_r, r \in \mathcal{A}$ is given by

$$\begin{aligned}
&\phi_{\boldsymbol{X}_1 \mid V_r = v_r, r\in\mathcal{A}}(\boldsymbol{t}) \\
&= \prod_{j=1}^{q} \exp(it_j\mu_j) \prod_{r\in\mathcal{A}} \left\{\exp\left[-\frac{1}{2}\left(\sum_{j=1}^{q} t_j[\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{jr}\right)^2\right]\right\}^{v_r} \left\{\exp\left[-\frac{1}{2}\eta_r\left(\sum_{j=1}^{q} t_j[\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{jr}\right)^2\right]\right\}^{1-v_r} \times \\
&\quad \times \prod_{s\in\mathcal{B}} \left\{\alpha_s \exp\left[-\frac{1}{2}\left(\sum_{j=1}^{q} t_j[\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{js}\right)^2\right] + (1 - \alpha_s)\exp\left[-\frac{1}{2}\eta_s\left(\sum_{j=1}^{q} t_j[\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{js}\right)^2\right]\right\}, \qquad (9)
\end{aligned}$$

where $\boldsymbol{t} = \left(t_1, \ldots, t_p\right)^\top \in \mathbb{R}^p$ and i is an imaginary unit.

Proof See Appendix B □

3.2 Model-Based Clustering via the Multiple Scaled Contaminated Normal Distribution

The MCN and MSCN distributions, as well as many other distributions like the MN distribution (Wolfe, 1965; Banfield & Raftery, 1993; Celeux & Govaert, 1995) or the multivariate t distribution (Peel & McLachlan, 2000; Andrews & McNicholas, 2012), are commonly used for model-based clustering, also known as cluster analysis on the notion of mixture models (McLachlan & Peel, 2000; Fraley & Raftery, 2002; Frühwirth-Schnatter, 2006; Melnykov & Maitra, 2010; McNicholas, 2016). Fundamentally, mixture models treat the population as a mixture of sub-populations, each modeled by a specified density function with different parameters. On model-based clustering via the MSCN distribution (Punzo & Tortora, 2021), a $p-$variate random vector $\boldsymbol{X}$ that arises from an MSCN mixture (MSCNM) with G components has its pdf given by

$$f_{\text{MSCNM}}(\boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{g=1}^{G} \pi_g f_{\text{MSCN}}\left(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Gamma}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\alpha}_g, \boldsymbol{\eta}_g\right), \tag{10}$$

where π_g is the mixing proportion of the gth component such that $\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$; the gth component is an MSCN distribution as defined in Eq. 2; and $\boldsymbol{\Psi} = \{\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Gamma}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\alpha}_g, \boldsymbol{\eta}_g\}_{g=1}^{G}$ contains all the parameters. The likelihood function of $\boldsymbol{\Psi}$ based on the observed data $\{\boldsymbol{x}_i\}_{i=1}^{n}$ is then given by

$$L(\boldsymbol{\Psi}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \prod_{i=1}^{n} \left[\sum_{g=1}^{G} \pi_g f_{\text{MSCN}}\left(\boldsymbol{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Gamma}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\alpha}_g, \boldsymbol{\eta}_g\right) \right]. \tag{11}$$

However, as mentioned by Melnykov and Maitra (2010), obtaining maximum likelihood estimates for $\boldsymbol{\Psi}$ using the observed likelihood function is typically challenging due to its complicated and multi-modal form, and thus, the EM algorithm (Dempster et al., 1977) has become a more common tool for fitting mixture models in general. In this algorithm, maximum likelihood estimation is carried out by maximizing the complete-data likelihood function that incorporates both the observed data $\{\boldsymbol{x}_i\}_{i=1}^{n}$ and some missing values and/or latent variables. The EM algorithm iteratively alternates between two steps, the expectation (E) step and the maximization (M) step. In the E-step, the conditional expectation of the complete data log-likelihood given the observed data and the current parameter estimates is computed. In the M-step, the parameters that maximize the expected log-likelihood from the E-step are computed. The framework for fitting the MSCNM using the EM algorithm is outlined in Punzo and Tortora (2021).

3.3 Mixtures of Multiple Scaled Contaminated Normal Distributions with Missing Values

Mixture models are generally formulated on complete data sets, i.e., there are no missing values. Some work has been done using the EM algorithm or its variants, such as the expectation-conditional maximization (ECM) algorithm (Meng & Rubin, 1993) or the

expectation-conditional maximization either (ECME) algorithm (Liu & Rubin, 1994), to extend mixtures of some well-known distributions to data sets with values missing at random; see, for example, Ghahramani and Jordan (1994) for the MN distribution, Wang et al. (2004), Lin (2014), and Goren and Maitra (2022) for the multivariate t distribution, and Tong and Tortora (2022b) for the MCN distribution. The MSCN distribution can also be extended in a similar fashion. Specifically, assuming a missing-at-random (MAR) mechanism, if we denote $\boldsymbol{x}_i^o$ and $\boldsymbol{x}_i^m$ as the observed and missing values per each observation $\boldsymbol{x}_i$, respectively, then $\boldsymbol{x}_i$ can be decomposed into $(\boldsymbol{x}_i^o, \boldsymbol{x}_i^m)$. Note that this is just a simplified notation where the superscripts o and m are used instead of o_i and m_i which represent how each observation can have a different number of missing values more accurately. In fact, our notation does not imply that the pattern of missingness is the same across all observations. Herein, we adopt the ECM algorithm for maximum likelihood estimation when fitting an MSCNM to incomplete data sets. This algorithm differs from the traditional EM algorithm in that the maximization steps are replaced by simpler conditional maximization (CM) steps where disjoint subsets of model parameters are updated. In light of the algorithm, we frame the problem as a maximum likelihood parameter estimation problem with three sources of missing data:

1. Cluster memberships for all observations: $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i=1}^n$, where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iG})^\top$. Herein, $z_{ig} = 1$ if observation $\boldsymbol{x}_i$ belongs to cluster g and $z_{ig} = 0$ otherwise for $g = 1, \ldots, G$;
2. Within cluster g, whether the hth variate of the transformed observation $\boldsymbol{\Gamma}_g^\top(\boldsymbol{x}_i - \boldsymbol{\mu}_g)$ is good or bad (outlier): $\boldsymbol{V} = \{\boldsymbol{v}_{i1}, \ldots, \boldsymbol{v}_{iG}\}_{i=1}^n$, where $\boldsymbol{v}_{ig} = \left(v_{i1g}, \ldots, v_{ipg}\right)^\top$ for $g = 1, \ldots G$. Herein, $v_{ihg} = 1$ if $\left[\boldsymbol{\Gamma}_g^\top(\boldsymbol{x}_i - \boldsymbol{\mu}_g)\right]_h$ is good and $v_{ihg} = 0$ otherwise for $h = 1, \ldots, p$;
3. Missing values of each observation: $\boldsymbol{X}^m = \left\{\boldsymbol{x}_i^m\right\}_{i=1}^n$.

The complete-data set of the MSCNM with missing values is thus given by $\mathcal{D} = \{\boldsymbol{X}^o, \boldsymbol{X}^m, \boldsymbol{Z}, \boldsymbol{V}\} = \left\{\boldsymbol{x}_i^o, \boldsymbol{x}_i^m, \boldsymbol{z}_i, \boldsymbol{v}_{i1}, \ldots, \boldsymbol{v}_{iG}\right\}_{i=1}^n$. If we let $\boldsymbol{\pi} = \{\pi_g\}_{g=1}^G, \boldsymbol{\alpha} = \{\boldsymbol{\alpha}_g\}_{g=1}^G$ and $\boldsymbol{\vartheta} = \{\boldsymbol{\mu}_g, \boldsymbol{\Gamma}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\eta}_g\}_{g=1}^G$, the complete-data likelihood is defined as

$$L_c(\boldsymbol{\Psi}; \mathcal{D}) = \prod_{i=1}^n \prod_{g=1}^G \left[\pi_g \prod_{h=1}^p \left[\alpha_{hg} f_{\mathrm{N}}\left(\left[\boldsymbol{\Gamma}_g^\top\left(\begin{bmatrix}\boldsymbol{x}_i^o \\ \boldsymbol{x}_i^m\end{bmatrix} - \boldsymbol{\mu}_g\right)\right]_h; 0, \lambda_{hg}\right)\right]^{v_{ihg}}\right.$$
$$\left. \times \left[(1-\alpha_{hg}) f_{\mathrm{N}}\left(\left[\boldsymbol{\Gamma}_g^\top\left(\begin{bmatrix}\boldsymbol{x}_i^o \\ \boldsymbol{x}_i^m\end{bmatrix} - \boldsymbol{\mu}_g\right)\right]_h; 0, \eta_{hg}\lambda_{hg}\right)\right]^{(1-v_{ihg})}\right]^{z_{ig}},$$

where the notation $[\ldots]_h$ refers to the hth coordinate of the vector within the brackets. Then, the complete-data log-likelihood function for $\boldsymbol{\Psi}$ can be written as $l(\boldsymbol{\Psi}; \mathcal{D}) = l_1(\boldsymbol{\pi}; \mathcal{D}) + l_2(\boldsymbol{\alpha}; \mathcal{D}) + l_3(\boldsymbol{\vartheta}; \mathcal{D})$ with

$$l_1(\boldsymbol{\pi}; \mathcal{D}) = \sum_{i=1}^n \sum_{g=1}^G z_{ig} \log \pi_g, \tag{12}$$

$$l_2(\boldsymbol{\alpha}; \mathcal{D}) = \sum_{i=1}^n \sum_{g=1}^G \sum_{h=1}^p z_{ig}\left[v_{ihg} \log \alpha_{hg} + (1 - v_{ihg}) \log(1 - \alpha_{hg})\right], \tag{13}$$

$$\text{and } l_3(\boldsymbol{\vartheta}; \mathcal{D}) \propto -\frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G z_{ig}\left[\log |\boldsymbol{W}_{\boldsymbol{v}_{ig}} \boldsymbol{\Lambda}_g| + \delta(\boldsymbol{x}_i^o, \boldsymbol{x}_i^m, \boldsymbol{\mu}_g; \boldsymbol{\Gamma}_g \boldsymbol{W}_{\boldsymbol{v}_{ig}} \boldsymbol{\Lambda}_g \boldsymbol{\Gamma}_g^\top)\right], \tag{14}$$

where

$$\delta(\boldsymbol{x}_i^o, \boldsymbol{x}_i^m, \boldsymbol{\mu}_g; \boldsymbol{\Gamma}_g \boldsymbol{W}_{\boldsymbol{v}_{ig}} \boldsymbol{\Lambda}_g \boldsymbol{\Gamma}_g^\top) = \left(\begin{bmatrix} \boldsymbol{x}_i^o \\ \boldsymbol{x}_i^m \end{bmatrix} - \boldsymbol{\mu}_g \right)^\top \boldsymbol{\Gamma}_g \boldsymbol{W}_{\boldsymbol{v}_{ig}}^{-1} \boldsymbol{\Lambda}_g^{-1} \boldsymbol{\Gamma}_g^\top \left(\begin{bmatrix} \boldsymbol{x}_i^o \\ \boldsymbol{x}_i^m \end{bmatrix} - \boldsymbol{\mu}_g \right) \quad (15)$$

is the squared Mahalanobis distance and

$$\boldsymbol{W}_{\boldsymbol{v}_{ig}} = \text{diag} \left\{ \left(v_{i1g} + \frac{1 - v_{i1g}}{\eta_{1g}} \right)^{-1}, \ldots, \left(v_{ipg} + \frac{1 - v_{ipg}}{\eta_{pg}} \right)^{-1} \right\}. \quad (16)$$

In our proposed framework, the ECM algorithm alternates between three steps, one E-step and two CM-steps, until convergence. Using maximum likelihood estimation, in the first CM-step, we update $\boldsymbol{\pi}$, $\boldsymbol{\alpha}$, and $\{\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\eta}_g\}_{g=1}^G$, while in the second CM-step, we update $\{\boldsymbol{\Gamma}_g\}_{g=1}^G$. The details for each step are outlined in the following subsections.

3.3.1 E-Step

In the E-step for the $(r+1)$th iteration, we need to compute the expectations of Z_{ig}, $Z_{ig} V_{ihg}$, $Z_{ig} \log |\boldsymbol{W}_{\boldsymbol{V}_{ig}} \boldsymbol{\Lambda}_g|$, and $Z_{ig}\, \delta(\boldsymbol{x}_i^o, \boldsymbol{X}_i^m, \boldsymbol{\mu}_g; \boldsymbol{\Gamma}_g \boldsymbol{W}_{\boldsymbol{V}_{ig}} \boldsymbol{\Lambda}_g \boldsymbol{\Gamma}_g^\top)$. Note that the last three expectations involve interactions between different random variables, so the law of iterated expectations can be employed. For ease of presentation, we show below only the most important results regarding the required expectations and leave additional details in supplementary material Section 1.

Conditional Expectations of Z_{ig} and $Z_{ig} V_{ihg}$ The first two expectations are obtained as the result of Bayes' rule. Specifically,

$$E_{\boldsymbol{\Psi}^{(r)}}(Z_{ig} \mid \boldsymbol{x}_i^o) = \frac{\pi_g^{(r)} f_{\boldsymbol{X}_i^o \mid Z_{ig}=1}\left(\boldsymbol{x}_i^o; \boldsymbol{\Psi}^{(r)}\right)}{\sum_{b=1}^G \pi_b^{(r)} f_{\boldsymbol{X}_i^o \mid Z_{ib}=1}\left(\boldsymbol{x}_i^o; \boldsymbol{\Psi}^{(r)}\right)} =: \tilde{z}_{ig}^{(r)}, \quad (17)$$

where $f_{\boldsymbol{X}_i^o \mid Z_{ig}=1}\left(\boldsymbol{x}_i^o; \boldsymbol{\Psi}^{(r)}\right)$ is the marginal density of the MSCN random vector $\boldsymbol{X}_i$ which can be approximated using the inversion formula as described in Sect. 3.1. In a similar manner,

$$E_{\boldsymbol{\Psi}^{(r)}}(Z_{ig} V_{ihg} \mid \boldsymbol{x}_i^o) = \tilde{z}_{ig}^{(r)} E_{\boldsymbol{\Psi}^{(r)}}(V_{ihg} \mid \boldsymbol{x}_i^o, Z_{ig} = 1) = \tilde{z}_{ig}^{(r)} \tilde{v}_{ihg}^{(r)}, \quad (18)$$

where

$$E_{\boldsymbol{\Psi}^{(r)}}(V_{ihg} \mid \boldsymbol{x}_i^o, Z_{ig} = 1) = \frac{\alpha_{hg}^{(r)} f_{\boldsymbol{X}_i^o \mid Z_{ig}=1, V_{ihg}=1}\left(\boldsymbol{x}_i^o; \boldsymbol{\Psi}^{(r)}\right)}{f_{\boldsymbol{X}_i^o \mid Z_{ig}=1}\left(\boldsymbol{x}_i^o; \boldsymbol{\Psi}^{(r)}\right)} =: \tilde{v}_{ihg}^{(r)}. \quad (19)$$

In the formula of $\tilde{v}_{ihg}^{(r)}$, $f_{\boldsymbol{X}_i^o \mid Z_{ig}=1, V_{ihg}=1}\left(\boldsymbol{x}_i^o; \boldsymbol{\Psi}^{(r)}\right)$ can be approximated using the inversion formula of the characteristic function under Proposition 3.6 with $\mathcal{A} = \{h\}$ and $\mathcal{B} = \{1, \ldots, p\} \setminus \{h\}$.

Conditional Expectation of $Z_{ig} \log |\boldsymbol{W}_{\boldsymbol{V}_{ig}} \boldsymbol{\Lambda}_g|$ First, recall that, $\boldsymbol{v}_{ig} = (v_{i1g}, \ldots, v_{ipg})^\top$ is a 0/1 vector with 2^p possible patterns. We now introduce the superscript k so that for $k = 1, \ldots, 2^p$,

$$\boldsymbol{v}_{ig}^k = \left(v_{i1g}^k, \ldots, v_{ipg}^k\right)^\top \quad (20)$$

refers to the kth 0/1 pattern and

$$\boldsymbol{W}^{(r)}_{\boldsymbol{v}^k_{ig}} = \text{diag}\left\{\left(v^k_{i1g} + \frac{1 - v^k_{i1g}}{\eta^{(r)}_{1g}}\right)^{-1}, \ldots, \left(v^k_{ipg} + \frac{1 - v^k_{ipg}}{\eta^{(r)}_{pg}}\right)^{-1}\right\} \tag{21}$$

refers to the inverse weight diagonal matrix built upon $\boldsymbol{v}^{k(r)}_{ig}$. Table 1 provides an example of $\boldsymbol{v}^{k(r)}_{ig}$ and $\boldsymbol{W}^{(r)}_{\boldsymbol{v}^k_{ig}}$ when $p = 3$ and k goes from 1 to $2^p = 2^3 = 8$. The new notations allow for a compact representation of the conditional expectation of $Z_{ig} \log |\boldsymbol{W}_{\boldsymbol{V}_{ig}} \boldsymbol{\Lambda}_g|$ given $\boldsymbol{x}^o_i$ and the current parameter estimates $\boldsymbol{\Psi}^{(r)}$, in which

$$\begin{aligned}
&E_{\boldsymbol{\Psi}^{(r)}}(Z_{ig} \log|\boldsymbol{W}_{\boldsymbol{V}_{ig}} \boldsymbol{\Lambda}_g| \mid \boldsymbol{x}^o_i) \\
&= \tilde{z}^{(r)}_{ig} E_{\boldsymbol{\Psi}^{(r)}}(\log|\boldsymbol{W}_{\boldsymbol{V}_{ig}} \boldsymbol{\Lambda}_g| \mid \boldsymbol{x}^o_i, Z_{ig} = 1) \\
&= \tilde{z}^{(r)}_{ig} \sum_{k=1}^{2^p} \hat{v}^{k(r)}_{ig} \log|\boldsymbol{W}^{(r)}_{\boldsymbol{v}^k_{ig}} \boldsymbol{\Lambda}^{(r)}_g|,
\end{aligned} \tag{22}$$

where

$$\hat{v}^{k(r)}_{ig} = \left[\prod_{h=1}^{p} \left(\alpha^{(r)}_{hg}\right)^{v^k_{ihg}} \left(1 - \alpha^{(r)}_{hg}\right)^{1 - v^k_{ihg}}\right] \frac{f_{\text{MN}}\left(\boldsymbol{x}^o_i; \boldsymbol{\mu}^{o(r)}_g, (\boldsymbol{\Gamma}^{(r)}_g \boldsymbol{W}^{(r)}_{\boldsymbol{v}^k_{ig}} \boldsymbol{\Lambda}^{(r)}_g \boldsymbol{\Gamma}^{(r)\top}_g)^{oo}\right)}{f_{\boldsymbol{X}^o_i \mid Z_{ig}=1}\left(\boldsymbol{x}^o_i; \boldsymbol{\Psi}^{(r)}\right)}. \tag{23}$$

In the above, the superscript oo denotes the $p^o \times p^o$ sub-matrix of $\boldsymbol{\Gamma}^{(r)}_g \boldsymbol{W}^{(r)}_{\boldsymbol{v}^k_{ig}} \boldsymbol{\Lambda}^{(r)}_g \boldsymbol{\Gamma}^{(r)\top}_g$, with p^o being the dimension of $\boldsymbol{x}^o_i$.

Table 1 Possible binary patterns of $V_{i1g}, \ldots, V_{ipg}$, corresponding vector $\boldsymbol{v}^k_{ig} = (v^k_{i1g}, \ldots, v^k_{ipg})^\top$ for $k = 1, \ldots, p$, and inverse weight matrix $\boldsymbol{W}^{(r)}_{\boldsymbol{v}^k_{ig}}$ when $p = 3$. A value of 1 indicates that the hth variate of the ith transformed observation $\boldsymbol{\Gamma}^\top_g(\boldsymbol{X}_i - \boldsymbol{\mu}_g)$ is good in cluster g; 0 otherwise

k	v^k_{i1g}	v^k_{i2g}	v^k_{i3g}	$\boldsymbol{v}^k_{ig}$	$\boldsymbol{W}^{(r)}_{\boldsymbol{v}^k_{ig}}$
1	0	0	0	$(0\ 0\ 0)^\top$	$\text{diag}\left\{\eta^{(r)}_{1g}, \eta^{(r)}_{2g}, \eta^{(r)}_{3g}\right\}$
2	0	0	1	$(0\ 0\ 1)^\top$	$\text{diag}\left\{\eta^{(r)}_{1g}, \eta^{(r)}_{2g}, 1\right\}$
3	0	1	0	$(0\ 1\ 0)^\top$	$\text{diag}\left\{\eta^{(r)}_{1g}, 1, \eta^{(r)}_{3g}\right\}$
4	0	1	1	$(0\ 1\ 1)^\top$	$\text{diag}\left\{\eta^{(r)}_{1g}, 1, 1\right\}$
5	1	0	0	$(1\ 0\ 0)^\top$	$\text{diag}\left\{1, \eta^{(r)}_{2g}, \eta^{(r)}_{3g}\right\}$
6	1	0	1	$(1\ 0\ 1)^\top$	$\text{diag}\left\{1, \eta^{(r)}_{2g}, 1\right\}$
7	1	1	0	$(1\ 1\ 0)^\top$	$\text{diag}\left\{1, 1, \eta^{(r)}_{3g}\right\}$
8	1	1	1	$(1\ 1\ 1)^\top$	$\text{diag}\{1, 1, 1\}$

Conditional Expectation of $Z_{ig}\,\delta(\boldsymbol{x}_i^o, \boldsymbol{X}_i^m, \boldsymbol{\mu}_g; \boldsymbol{\Gamma}_g \boldsymbol{W}_{V_{ig}} \boldsymbol{\Lambda}_g \boldsymbol{\Gamma}_g^\top)$ To conclude the E-step, we have

$$\begin{aligned}
&E_{\boldsymbol{\Psi}^{(r)}}(Z_{ig}\,\delta(\boldsymbol{x}_i^o, \boldsymbol{X}_i^m, \boldsymbol{\mu}_g; \boldsymbol{\Gamma}_g \boldsymbol{W}_{V_{ig}} \boldsymbol{\Lambda}_g \boldsymbol{\Gamma}_g^\top) \mid \boldsymbol{x}_i^o) \\
&= \tilde{z}_{ig}^{(r)} E_{\boldsymbol{\Psi}^{(r)}}(\,\delta(\boldsymbol{x}_i^o, \boldsymbol{X}_i^m, \boldsymbol{\mu}_g; \boldsymbol{\Gamma}_g \boldsymbol{W}_{V_{ig}} \boldsymbol{\Lambda}_g \boldsymbol{\Gamma}_g^\top) \mid \boldsymbol{x}_i^o, Z_{ig} = 1) \\
&= \tilde{z}_{ig}^{(r)} \sum_{k=1}^{2^p} \hat{v}_{ig}^{k(r)} \operatorname{trace}\left[\boldsymbol{M}_{ikg}^{(r)} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}_{ig}^k}^{(r)} \right],
\end{aligned} \tag{24}$$

where

$$\boldsymbol{M}_{ikg}^{(r)} = \left[\boldsymbol{\Gamma}_g^{(r)} \boldsymbol{W}_{\boldsymbol{v}_{ig}^k}^{(r)} \boldsymbol{\Lambda}_g^{(r)} \boldsymbol{\Gamma}_g^{(r)\top} \right]^{-1} = \boldsymbol{\Gamma}_g^{(r)} \boldsymbol{W}_{\boldsymbol{v}_{ig}^k}^{-1(r)} \boldsymbol{\Lambda}_g^{-1(r)} \boldsymbol{\Gamma}_g^{(r)\top} \tag{25}$$

and

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}_{ig}^k}^{(r)} = \begin{bmatrix}
\left(\boldsymbol{x}_i^o - \boldsymbol{\mu}_g^{o(r)}\right)\left(\boldsymbol{x}_i^o - \boldsymbol{\mu}_g^{o(r)}\right)^\top & \left(\boldsymbol{x}_i^o - \boldsymbol{\mu}_g^{o(r)}\right)\left(\tilde{\boldsymbol{x}}_{ig}^{k(r)} - \boldsymbol{\mu}_g^{m(r)}\right)^\top \\
\left(\tilde{\boldsymbol{x}}_{ig}^{k(r)} - \boldsymbol{\mu}_g^{m(r)}\right)\left(\boldsymbol{x}_i^o - \boldsymbol{\mu}_g^{o(r)}\right)^\top & \left(\tilde{\boldsymbol{x}}_{ig}^{k(r)} - \boldsymbol{\mu}_g^{m(r)}\right)\left(\tilde{\boldsymbol{x}}_{ig}^{k(r)} - \boldsymbol{\mu}_g^{m(r)}\right)^\top + \\
& \tilde{\tilde{\boldsymbol{x}}}_{ig}^{k(r)} - \tilde{\boldsymbol{x}}_{ig}^{k(r)} \tilde{\boldsymbol{x}}_{ig}^{k(r)\top}
\end{bmatrix}, \tag{26}$$

with

$$\begin{aligned}
&E_{\boldsymbol{\Psi}^{(r)}}(\boldsymbol{X}_i^m \mid \boldsymbol{x}_i^o, Z_{ig} = 1, V_{i1g} = v_{i1g}^k, \ldots, V_{ipg} = v_{ipg}^k) \\
&= \boldsymbol{\mu}_g^{m(r)} + \left(\boldsymbol{M}_{ikg}^{(r)}\right)^{mo} \left(\boldsymbol{M}_{ikg}^{(r)}\right)^{oo-1} \left(\boldsymbol{x}_i^o - \boldsymbol{\mu}_g^{o(r)}\right), \\
&=: \tilde{\boldsymbol{x}}_{ig}^{k(r)},
\end{aligned} \tag{27}$$

$$\begin{aligned}
&E_{\boldsymbol{\Psi}^{(r)}}(\boldsymbol{X}_i^m \boldsymbol{X}_i^{m\top} \mid \boldsymbol{x}_i^o, Z_{ig} = 1, V_{i1g} = v_{i1g}^k, \ldots, V_{ipg} = v_{ipg}^k) \\
&= \left(\boldsymbol{M}_{ikg}^{(r)}\right)^{mm} - \left(\boldsymbol{M}_{ikg}^{(r)}\right)^{mo} \left(\boldsymbol{M}_{ikg}^{(r)}\right)^{oo-1} \left(\boldsymbol{M}_{ikg}^{(r)}\right)^{om} + \tilde{\boldsymbol{x}}_{ig}^{k(r)} \tilde{\boldsymbol{x}}_{ig}^{k(r)\top} \\
&=: \tilde{\tilde{\boldsymbol{x}}}_{ig}^{k(r)}.
\end{aligned} \tag{28}$$

In the above, $\tilde{\boldsymbol{x}}_{ig}^{k(r)}$ and $\tilde{\tilde{\boldsymbol{x}}}_{ig}^{k(r)}$ are obtained by the conditional distribution properties of an MSCN distribution as described in Sect. 3.1. Moreover, the superscript *oo* denotes the $p^o \times p^o$ sub-matrix of $\boldsymbol{M}_{ikg}^{(r)}$, with p^o being the dimension of $\boldsymbol{x}_i^o$. The superscript *mm* denotes the $p^m \times p^m$ sub-matrix of $\boldsymbol{M}_{ikg}^{(r)}$, with p^m being the dimension of $\boldsymbol{x}_i^m$. On the other hand, the superscripts *mo* and *om* denote the $p^m \times p^o$ and $p^o \times p^m$ sub-matrices of $\boldsymbol{M}_{ikg}^{(r)}$. We also note that the terms

$$\left(\boldsymbol{M}_{ikg}^{(r)}\right)^{mo} \left(\boldsymbol{M}_{ikg}^{(r)}\right)^{oo-1} \left(\boldsymbol{x}_i^o - \boldsymbol{\mu}_g^{o(r)}\right) \text{ in } \tilde{\boldsymbol{x}}_{ig}^{k(r)} \tag{29}$$

and

$$-\left(\boldsymbol{M}_{ikg}^{(r)}\right)^{mo} \left(\boldsymbol{M}_{ikg}^{(r)}\right)^{oo-1} \left(\boldsymbol{M}_{ikg}^{(r)}\right)^{om} \text{ in } \tilde{\tilde{\boldsymbol{x}}}_{ig}^{k(r)} \tag{30}$$

can be regarded as the adjustment for imputing the conditions in the expectation computation according to the kth 0/1 pattern.

3.3.2 CM-Steps

In the first CM-step for the $(r+1)$th iteration, we fix $\boldsymbol{\Gamma}_g$ at $\boldsymbol{\Gamma}_g^{(r)}$ and update $\pi_g^{(r)}$, $\boldsymbol{\mu}_g^{(r)}$, $\boldsymbol{\Lambda}_g^{(r)}$, and $\boldsymbol{\eta}_g^{(r)}$ with $\pi_g^{(r+1)}$, $\boldsymbol{\mu}_g^{(r+1)}$, $\boldsymbol{\Lambda}_g^{(r+1)}$, and $\boldsymbol{\eta}_g^{(r+1)}$ respectively. We will also update $\boldsymbol{\alpha}_g^{(r)}$, but unlike the mentioned parameters, we update it through each of its hth component, α_{hg}, for $h = 1, \ldots, p$. Specifically, the mixing proportion and the proportion of good observations per cluster are updated first

$$\pi_g^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n} \tilde{z}_{ig}^{(r)}, \tag{31}$$

$$\alpha_{hg}^{(r+1)} = \frac{\sum_{i=1}^{n} \tilde{z}_{ig}^{(r)} \tilde{v}_{ihg}^{(r)}}{\sum_{i=1}^{n} \tilde{z}_{ig}^{(r)}}, \tag{32}$$

where $\tilde{z}_{ig}^{(r)}$ and $\tilde{v}_{ig}^{(r)}$ depends on the observed values. The update of the cluster means is then computed, in which the expected value $\tilde{\boldsymbol{x}}_{ig}^{k(r)}$ is used.

$$\boldsymbol{\mu}_g^{(r+1)} = \boldsymbol{\Gamma}_g^{(r)} \left[\sum_{i=1}^{n} \tilde{z}_{ig}^{(r)} \sum_{k=1}^{2^p} \hat{v}_{ig}^{k(r)} \boldsymbol{W}_{\boldsymbol{v}_{ig}^k}^{-1(r)} \right]^{-1} \left\{ \sum_{i=1}^{n} \tilde{z}_{ig}^{(r)} \sum_{k=1}^{2^p} \hat{v}_{ig}^{k(r)} \boldsymbol{W}_{\boldsymbol{v}_{ig}^k}^{-1(r)} \boldsymbol{\Gamma}_g^{(r)\top} \begin{bmatrix} \boldsymbol{x}_i^o \\ \tilde{\boldsymbol{x}}_{ig}^{k(r)} \end{bmatrix} \right\}. \tag{33}$$

The update of $\boldsymbol{\Lambda}_g$ is

$$\boldsymbol{\Lambda}_g^{(r+1)} = \left[\sum_{i=1}^{n} \tilde{z}_{ig}^{(r)} \right]^{-1} \left\{ \boldsymbol{I}_p \odot \boldsymbol{\Gamma}_g^{(r)\top} \left[\sum_{i=1}^{n} \tilde{z}_{ig}^{(r)} \sum_{k=1}^{2^p} \hat{v}_{ig}^{k(r)} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}_{ig}^k}^{(r)} \boldsymbol{W}_{\boldsymbol{v}_{ig}^k}^{-1(r)} \right] \boldsymbol{\Gamma}_g^{(r)} \right\}, \tag{34}$$

where $\odot$ is the Hadamard product and $\boldsymbol{I}_p$ is a $p \times p$ identity matrix, and the computation of $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}_{ig}^k}^{(r)}$ involves $\tilde{\boldsymbol{x}}_{ig}^{k(r)}$ and $\tilde{\tilde{\boldsymbol{x}}}_{ig}^{k(r)}$. The update $\boldsymbol{\eta}_g^{(r+1)}$ of $\boldsymbol{\eta}_g^{(r)}$ is the solution of the following equation

$$\sum_{i=1}^{n} \tilde{z}_{ig}^{(r)} \sum_{k=1}^{2^p} \hat{v}_{ig}^{k(r)} \left[\boldsymbol{W}_{\boldsymbol{v}_{ig}^k}^{-1(r)} \frac{\partial}{\partial \boldsymbol{\eta}_g^{(r)}} \boldsymbol{W}_{\boldsymbol{v}_{ig}^k}^{(r)} - \boldsymbol{I}_p \odot \boldsymbol{W}_{\boldsymbol{v}_{ig}^k}^{-1(r)} \boldsymbol{\Lambda}_g^{-1(r)} \boldsymbol{\Gamma}_g^{(r)\top} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}_{ig}^k}^{(r)} \boldsymbol{\Gamma}_g^{(r)} \boldsymbol{W}_{\boldsymbol{v}_{ig}^k}^{-1(r)} \frac{\partial}{\partial \boldsymbol{\eta}_g^{(r)}} \boldsymbol{W}_{\boldsymbol{v}_{ig}^k}^{(r)} \right] = \boldsymbol{0}_p, \tag{35}$$

where $\boldsymbol{0}_p$ is a vector whose p entries are all zeros and

$$\frac{\partial}{\partial \boldsymbol{\eta}_g^{(r)}} \boldsymbol{W}_{\boldsymbol{v}_{ig}^k}^{(r)} = \operatorname{diag}\left\{ \left(v_{ihg}^k + \frac{1 - v_{ihg}^k}{\eta_{hg}^{(r)}} \right)^{-2} \left(\frac{1 - v_{ihg}^k}{\left(\eta_{hg}^{(r)}\right)^2} \right) \right\} \quad \text{for } h = 1, \ldots, p. \tag{36}$$

It is worth noting that for the hth diagonal entry of $\frac{\partial}{\partial \boldsymbol{\eta}_g^{(r)}} \boldsymbol{W}_{\boldsymbol{v}_{ig}^k}^{(r)}$, its value is 0 if $v_{ihg}^k = 1$ or 1 if $v_{ihg}^k = 0$, so as a whole, this partial derivative matrix actually does not contain anything relevant to $\boldsymbol{\eta}_g^{(r)}$.

In the second CM-step for the $(r+1)$th iteration, the update $\boldsymbol{\Gamma}_g^{(r+1)}$ of $\boldsymbol{\Gamma}_g^{(r)}$ would be the solution of the following optimization problem

$$\boldsymbol{\Gamma}_g^{(r+1)} = \underset{\boldsymbol{\Gamma}_g^{(r)}}{\operatorname{argmin}} \sum_{i=1}^{n} \tilde{z}_{ig}^{(r)} \sum_{k=1}^{2^p} \hat{v}_{ig}^{k(r)} \operatorname{trace}\left[\boldsymbol{\Gamma}_g^{(r)} \boldsymbol{W}_{\boldsymbol{v}_{ig}^k}^{-1(r)} \boldsymbol{\Lambda}_g^{-1(r)} \boldsymbol{\Gamma}_g^{(r)\top} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}_{ig}^k}^{(r)}\right], \tag{37}$$

with $\boldsymbol{\Gamma}_g^{(r+1)}$ constrained to be an orthogonal matrix. To satisfy such constraint, the PLR decomposition for orthogonal matrices proposed by Bagnato and Punzo (2021) is first applied on $\boldsymbol{\Gamma}_g^{(r)}$ before solving the optimization problem.

4 Notes on Implementation

4.1 Initial Values

Despite its popularity and effectiveness in incomplete data problems, the EM algorithm, as well as its variants, are known to be highly dependent on the choice of initial values, which in turn can affect clustering performance and convergence speed (Biernacki et al., 2000; Karlis & Xekalaki, 2003; Shireman et al., 2017). Some recent literature such as Michael and Melnykov (2016) and You et al. (2023) focused on improving the initialization of the EM algorithm. When the data are characterized by outliers, the problem becomes more complex; see for example, Cuesta-Albertos et al. (2008). A study on the impact of the initialization technique is beyond the scope of this work and deserves further investigation. In this paper, we use the following standard initialization technique:

- From the full data set, obtain observations without any missing variates.
- Set mixing proportions equal across G clusters, that is, $\pi_g^{(0)} = \dfrac{1}{G}$.
- Perform k-medoids clustering (Kaufman & Rousseeuw, 1990) and assign component mean vector $\boldsymbol{\mu}_g^{(0)}$ and covariance matrix $\boldsymbol{\Sigma}_g^{(0)}$ according to the resulting solution.
- Apply an eigen-decomposition on component covariance matrix $\boldsymbol{\Sigma}_g^{(0)}$ to obtain $\boldsymbol{\Gamma}_g^{(0)}$ and $\boldsymbol{\Lambda}_g^{(0)}$.
- Set proportions of good observations $\boldsymbol{\alpha}_g^{(0)} = (0.999, \ldots, 0.999)^\top$ and degrees of contamination $\boldsymbol{\eta}_g^{(0)} = (1.001, \ldots, 1.001)^\top$.

4.2 Convergence, Cluster Assignment, and Directional Outlier Detection

In our framework, convergence is determined using the Aitken stopping criterion (Aitken, 1926). At convergence of the ECM algorithm, $\hat{z}_{ig}$ and $\hat{v}_{ihg}$ are computed as the values of $\tilde{z}_{ig}^{(r)}$ and $\tilde{v}_{ihg}^{(r)}$, respectively. Then, cluster memberships can be assigned to all observations by means of the maximum *a posteriori* probabilities (MAP). More information on the Aitken stopping criterion and the MAP in the context of model-based clustering can be found in McNicholas et al. (2010). Within cluster g, observation $\boldsymbol{x}_i$ is considered good with respect to the hth principal component if MAP $\left(\hat{z}_{ig}\right) = 1$ and $\hat{v}_{ihg} > 0.5$. As a byproduct of the ECM algorithm, this outlier detection procedure is done automatically and requires no additional distributional assumptions or subjective thresholds. Not only that, it is said to be directional, meaning whether observation $\boldsymbol{x}_i$ resembles an outlier or not is evaluated separately for each principal component, and thus, complex behaviors of outliers in a multivariate setting can be

effectively captured. This provides an important advantage compared to the outlier detection procedure of the MCN mixture where outlying observations can be identified, but without the knowledge of which dimension(s) yielding such decision (see, for example, Punzo et al. 2018; Tong & Tortora 2022b).

4.3 Model Selection

In practice, the number of mixture components G is often not given in advance, thus resulting in the model selection problem. One common strategy to address this problem is to fit the assumed model multiple times over a range of values for G and select the model that optimizes an information criterion such as Akaike information criterion (AIC; Akaike 1998) or Bayesian information criterion (BIC; Schwarz 1978). Their formulas are AIC $= -2l(\hat{\boldsymbol{\Psi}}) + 2\rho$ and BIC $= -2l(\hat{\boldsymbol{\Psi}}) + \rho \log n$, where $\hat{\boldsymbol{\Psi}}$ the estimated parameters at convergence of the ECM algorithm, $l(\hat{\boldsymbol{\Psi}})$ is the associated observed log-likelihood, and ρ is the number of free parameters in the model. In our model, $\rho = (G-1) + 4Gp + Gp(1-p)/2$. Many other criteria have also been proposed; commonly used criteria are integrated classification likelihood (ICL; Biernacki et al. 2000), Kullback information criterion (KIC; Cavanaugh 1999), corrected Kullback information criterion (KICc; Seghouane & Bekara 2004), approximate weight of evidence criterion (AWE; Banfield & Raftery 1993), modified Akaike information criterion (AIC3; Bozdogan 1993), consistent Akaike information criterion (CAIC; Bozdogan 1987), corrected Akaike information criterion (AICc; Hurvich & Tsai 1989), and classification likelihood criterion (CLC; Biernacki & Govaert 1997). Using different simulated and real data sets, some studies have been conducted to compare these information criteria in model-based clustering of complete data sets; see for example, Tran and Tortora (2021) for the MSCN distribution and Akogul and Erisoglu (2016) for the MN distribution. The mentioned papers also contain a unified summary of the formulas of the information criteria.

4.4 Other Computational Details

Like many mixture models including normal mixtures, the likelihood function of the proposed model presents spurious local maxima and is unbounded. Therefore, the existence of the global maximizer is not guaranteed. Some possible solutions to this issue are discussed in Melnykov (2013). The proposed model is estimated using the ECM algorithm, whose convergence to a stationary point is obtained under the same conditions as the classification EM algorithm (McLachlan & Krishnan 2008, Chapter 5). However, the expectations $\tilde{z}_{ig}^{(r)}$ and $\tilde{v}_{ihg}^{(r)}$ in the E-step involve marginal densities obtained from the inversion formula, these densities, in turn, require the numerical calculation of multiple integrals; moreover, two parameter updates do not have closed-form solutions, implying additional conditions for monotonicity and convergence; for more details, see McLachlan and Krishnan (2008, Chapter3).

To numerically calculate multiple integrals required to compute the expectations $\tilde{z}_{ig}^{(r)}$ and $\tilde{v}_{ihg}^{(r)}$ in the E-step, we employ the adaptive multivariate integration over hypercubes introduced by Berntsen et al. (1991). It is similar to the idea suggested by Dooren and Ridder (1976) for single integrands. This method is based on a globally adaptive subdivision strategy, which carries out numerical integration over hyperrectangular regions, and is readily implemented in the R package **cubature** (Narasimhan et al., 2022).

In the CM-steps, we also need some numerical procedures for two parameters. The first is $\boldsymbol{\eta}_g^{(r)}$, where its update amounts to solving p non-linear equations that can be achieved

using the Newton–Raphson method implemented in the `multiroot()` function of the R package **rootSolve** (Soetaert, 2009; Soetaert & Herman, 2009). The second is $\boldsymbol{\Gamma}_g^{(r)}$, which involves an optimization problem subject to the constraint that $\boldsymbol{\Gamma}_g^{(r)}$ is orthogonal. To turn this problem into an unconstrained problem, we first apply the PLR decomposition introduced by Bagnato & Punzo (2021) to $\boldsymbol{\Gamma}_g^{(r)}$ and obtain the objective function described in Eq. 37. This decomposition factorizes an $p \times p$ invertible orthogonal matrix $\boldsymbol{Q}$ into the product of a $p \times p$ permutation matrix $\boldsymbol{P}$, a $p \times p$ unit lower-triangular matrix $\boldsymbol{L}$, and $\boldsymbol{R}^{-1}$, where $\boldsymbol{R}$ a $p \times p$ upper-triangular matrix with a positive diagonal and obtained from the QR decomposition of the matrix $\boldsymbol{PL}$. Bagnato and Punzo (2021) showed that there exists a one-to-one correspondence between $\boldsymbol{Q}$ and the lower-triangular unconstrained real values in $\boldsymbol{L}$, which allows the optimization problem to be instead formulated in terms of $\boldsymbol{L}$. Advantageously, the new formulation can be solved with unconstrained optimization methods implemented in the function `optim()` of base R. In this paper, our optimization method of choice is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970).

5 Simulation Study

In this section, we evaluate the clustering and outlier detection performance of our newly proposed extension in Sect. 3.3 through a series of four simulation studies involving data sets with outliers and missing-at-random values. In the first three studies, the number of clusters is fixed to the real known number of clusters. In Sect. 5.1, we evaluate the performance of our model in recovering the true underlying parameters. In Sect. 5.2, we focus on directional outliers which refer to outlying observations along particular principal components. In Sect. 5.3, we consider a higher number of overlapping clusters sampled from heavy-tailed distributions. Lastly, in Sect. 5.4, we revisit the simulated scenarios in studies 2 and 3 and examine model selection using the information criteria introduced in Sect. 4.3. The detailed results of all simulation studies can be found in supplementary material Section 2.

Herein, clustering performance is measured using the adjusted Rand index (ARI; Hubert & Arabie 1985) which corrects the Rand index (Rand, 1971) for chance and has an expected value of 0 under random partitions and an expected value of 1 under perfect agreement (for more information, see Steinley 2004). Since all the clustering techniques used in the simulation study also assign cluster labels to the outliers, the outliers are included in the computation of the ARI. To assess outlier detection, we rely on the true positive rate (TPR), measuring the proportion of outliers correctly detected, and the false positive rate (FPR), measuring the proportion of good observations incorrectly detected as outliers. Our competitors include mixture models via different distributions: MCN, multivariate t (Mt), and MN. However, we do not consider the multivariate normal mixture (MNM) with regard to outlier detection due to it not being a robust model-based clustering method. The outlier detection rule of the multivariate contaminated normal mixture (MCNM) is similar to the MSCNM, except that it is limited to global detection only, i.e., an outlier is identified without knowing in which specific dimension(s) it is outlying. Such detection rule of the MCNM is described in greater detail by Punzo and McNicholas (2016) and Tong and Tortora (2022b). On the other hand, in the multivariate t mixture (MtM), an observation $\boldsymbol{x}_i$ is considered bad based on the quantity

$$\sum_{g=1}^{G} \text{MAP}(\hat{z}_{ig})\delta(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_g; \hat{\boldsymbol{\Sigma}}_g), \tag{38}$$

where $\hat{\boldsymbol{\mu}}_g$ and $\hat{\boldsymbol{\Sigma}}_g$ are the values of $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ at convergence of the MtM's fitting framework, respectively, and the corresponding squared Mahalanobis distance random variable $\delta(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)$ is approximately χ^2_p. According to Peel and McLachlan (2000), if the quantity is greater the 95th percentile of the χ^2_p distribution, then $\boldsymbol{x}_i$ is treated as a bad point. A different percentile could be used, but this would become a tuning problem that undermines the automatic outlier detection benefit.

All the algorithms are initialized by setting the means equal to the medoids obtained using k-medoids clustering, equal proportions across clusters, and covariance matrices as identity matrices. The analysis was conducted using the software R (R Core Team, 2021), the code for the mixtures of MCN, multivariate t (Mt), and MN distributions with missing data can be found in the package **MixtureMissing** (Tong & Tortora, 2022a).

5.1 Study 1: True Parameter Recovery

In this study, we examine the performance of our model in recovering true parameters when the number of components is correctly specified. For data simulation, we consider the setting of two bivariate multiple scaled contaminated normal clusters ($G = 2$) with the following parameters

$$n_1 = 420, \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & 1.5 \\ 1.5 & 2 \end{bmatrix}, \quad \boldsymbol{\alpha}_1 = (0.95, 0.95)^\top, \quad \boldsymbol{\eta}_1 = (10, 10)^\top,$$

$$n_2 = 180, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 0 \\ 8 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & -1.5 \\ -1.5 & 2 \end{bmatrix}, \quad \boldsymbol{\alpha}_2 = (0.7, 0.7)^\top, \quad \boldsymbol{\eta}_2 = (3, 10)^\top.$$

The number of observations in each cluster implies that the true mixing proportions are $\pi_1 = 0.7$ and $\pi_2 = 0.3$. Then, using the function `rmscn()` in the package **MSclust** (Tortora et al., 2023), we generate 20 complete data sets based on the above setting. For each of these complete data sets, we hide one of the two variates of 10%, 30%, 50%, and 70% observations under the missing-at-random mechanism in a way such that there are 10 general MAR missing-data patterns per percentage using the `ampute()` function. Essentially, in doing so, with each missing percentage, we replicate the complete data set 10 times and introduce a different general MAR missing data pattern to each replicate, yielding 10 different data sets with the same number of observations missing one variate. In total, we have $(20)(4)(10) = 800$ data sets for this study. Figure 1 provides an example of data sets generated for this study.

Herein, we compare our model to the complete case analysis using the multiple scaled contaminated normal mixture proposed by Punzo and Tortora (2021). Recall that the model of Punzo and Tortora (2021) assumes no missing values so it is fitted to complete observations only in our simulation, whereas our model is fitted to the full data set. The number of clusters for each model is fixed to the ground truth $G = 2$. After fitting both models, their parameter estimates are compared to the true parameters used for the data simulation above. Tables 1, 2, 3 and 4 in supplementary material Section 2 summarize the results of the simulation study for each missing percentage. The estimates for the mixing proportions and the component mean vectors of our model have lower bias and standard deviation than in the complete case analysis for almost every missing percentage. This is expected as the complete case analysis only uses information from complete observations and thus can yield biased estimates overall. Also with the complete case analysis, as more observations are subject to missingness, the amount of bias and standard deviation rises up. For component covariance matrices, proportions of good observations, and degrees of contamination, the estimates of our model have higher

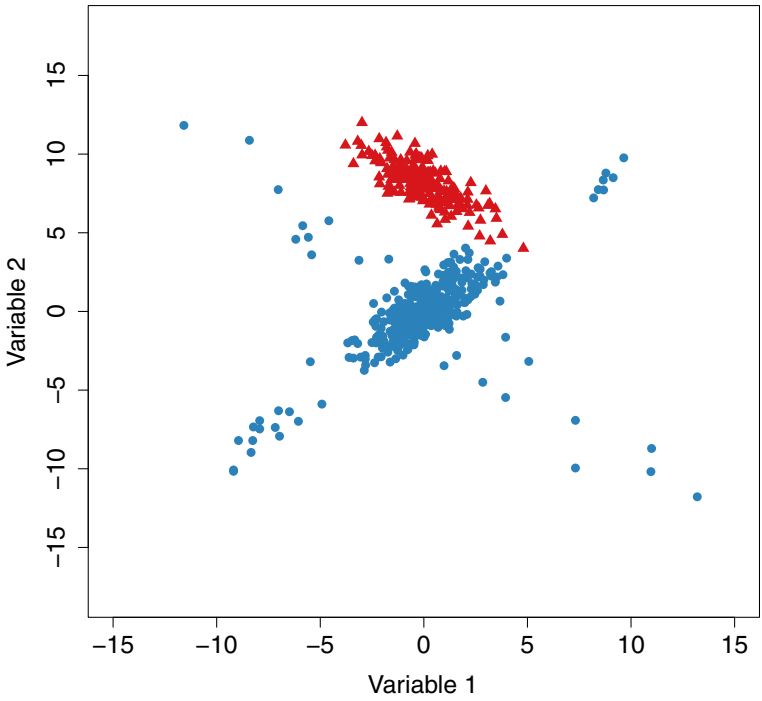


Fig. 1 An example of data set generated in the simulation study 1

bias and standard deviation than in the complete case analysis when there are 10% and 30% observations with missing values. However, when the missing percentage goes to 50% and 70%, our model yields estimates with lower bias and standard deviation for those parameters compared to the complete case analysis.

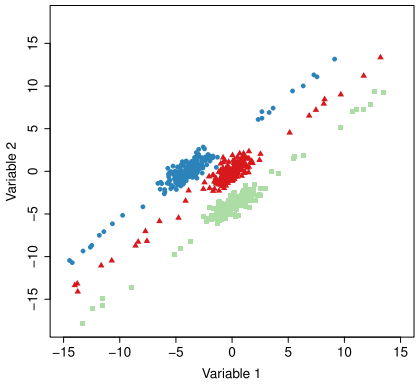
5.2 Study 2: Clusters with Directional Outliers

Herein, we consider the following four scenarios of directional outliers, all of which are bivariate ($p = 2$) and contain 600 observations ($n = 600$).

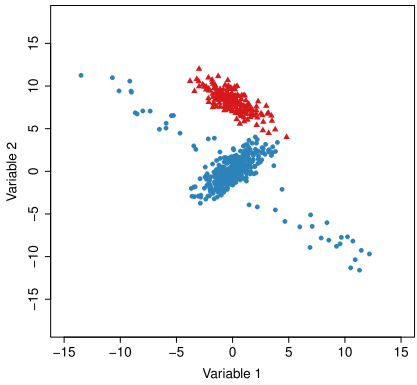
(a) **Three parallel elongated clusters with outliers in the first principal component.** Example in Fig. 2a. The setup is inspired by the supplementary materials of Forbes and Wraith (2014). We first generate three Gaussian clusters ($G = 3$), 200 observations each ($n_1 = n_2 = n_3 = 200$), with mean vectors $\boldsymbol{\mu}_1 = (-4, 0)^\top$, $\boldsymbol{\mu}_2 = (0, 0)^\top$, and $\boldsymbol{\mu}_3 = (0, -4)^\top$, respectively, and the common covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix}.$$

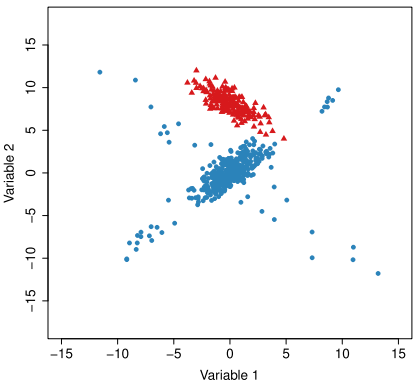
For each cluster, we introduce directional outliers by applying a principal component analysis, in which we denote $\bar{x}^{(j)}$ and $s^{(j)}$ as the sample mean and standard deviation of principal component scores associated with the jth principal component, for $j = 1, 2$. Then, in the space spanned by the two principal components, we replace some observations by (x^*_{i1}, x^*_{i2}), where



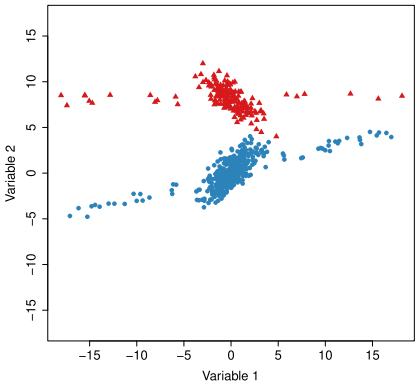
(a) Three parallel elongated clusters with outliers in the first principal component



(b) One cluster with outliers in the second principal component and one elliptical cluster



(c) One cluster with outliers in both principal components and one elliptical cluster



(d) Two clusters with outliers forming an angle with their first principal component

Fig. 2 Examples of data sets generated according to the four scenarios considered in the simulation study 2

- x_{i1}^* can be drawn with an equal chance from either a uniform random variable over the interval

$$\left(\bar{x}^{(1)} - 15s^{(1)}, \bar{x}^{(1)} - 4s^{(1)}\right)$$

or a uniform random variable over the interval

$$\left(\bar{x}^{(1)} + 4s^{(1)}, \bar{x}^{(1)} + 15s^{(1)}\right).$$

- x_{i2}^* is a realization of a uniform random variable over the interval

$$\left(\bar{x}^{(2)} - s^{(2)}, \bar{x}^{(2)} + s^{(2)}\right).$$

Finally, we transform back to the original data space to obtain the simulated data set.

(b) **One cluster with outliers in the second principal component and 1 elliptical cluster.** Example in Fig. 2b. We first generate two Gaussian clusters ($G = 2$) with the following

parameters

$$n_1 = 420, \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & 1.5 \\ 1.5 & 2 \end{bmatrix}, \quad n_2 = 180, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 0 \\ 8 \end{bmatrix},$$

$$\text{and} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & -1.5 \\ -1.5 & 2 \end{bmatrix}.$$

Then, in cluster 1, we apply a principal component analysis, in which we denote $\bar{x}^{(j)}$ and $s^{(j)}$ as the sample mean and standard deviation of principal component scores associated with the jth principal component, for $j = 1, 2$. Now, in the space spanned by the two principal components, we replace some observations by (x_{i1}^*, x_{i2}^*), where

- x_{i1}^* is a realization of a uniform random variable over the interval

$$\left(\bar{x}^{(1)} - s^{(1)}, \bar{x}^{(1)} + s^{(1)}\right).$$

- x_{i2}^* can be drawn with an equal chance from either a uniform random variable over the interval

$$\left(\bar{x}^{(2)} - 25s^{(2)}, \bar{x}^{(2)} - 5s^{(2)}\right)$$

or a uniform random variable over the interval

$$\left(\bar{x}^{(2)} + 5s^{(2)}, \bar{x}^{(2)} + 25s^{(2)}\right).$$

Finally, we transform cluster 1 back to the original data space and combine it with cluster 2, which remains unchanged, to obtain the simulated data set.

(c) **One cluster with outliers in both principal components and 1 elliptical cluster.** Example in Fig. 2c. We first generate two Gaussian clusters ($G = 2$), apply a principal component analysis on cluster 1, and introduce some sample statistics as in scenario (b). However, in the space spanned by the two principal components, we replace some observations with an equal chance by (x_{i1}^*, x_{i2}^*) or $(x_{i1}^{**}, x_{i2}^{**})$. Herein, (x_{i1}^*, x_{i2}^*) is defined as in scenario (b), while $(x_{i1}^{**}, x_{i2}^{**})$ is made as follows

- x_{i1}^{**} can be drawn with an equal chance from either a uniform random variable over the interval

$$\left(\bar{x}^{(1)} - 8s^{(1)}, \bar{x}^{(1)} - 4s^{(1)}\right)$$

or a uniform random variable over the interval

$$\left(\bar{x}^{(1)} + 4s^{(1)}, \bar{x}^{(1)} + 8s^{(1)}\right).$$

- x_{i2}^{**} is a realization of a uniform random variable over the interval

$$\left(\bar{x}^{(2)} - s^{(2)}, \bar{x}^{(2)} + s^{(2)}\right).$$

Finally, we transform cluster 1 back to the original data space and combine it with cluster 2, which remains unchanged, to obtain the simulated data set.

(d) **Two clusters with outliers forming an angle with their first principal component.** Example in Fig. 2d. We first generate two Gaussian clusters ($G = 2$) as in scenario (b). Let $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ be the sample covariance matrices of clusters 1 and 2, respectively. We then perform the following spectral decompositions on the following two matrices

$$\boldsymbol{\Gamma}_1 \boldsymbol{S}_1 \boldsymbol{\Gamma}_1^\top \quad \text{and} \quad \boldsymbol{\Gamma}_2 \boldsymbol{S}_2 \boldsymbol{\Gamma}_2^\top,$$

where $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ are rotation matrices of angles $\theta_1 = \frac{5\pi}{6}$ and $\theta_2 = \frac{\pi}{4}$, respectively. Note that a rotation matrix of angle θ is defined as follows

$$\mathbf{\Gamma} \equiv \mathbf{\Gamma}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

We transform the data using the eigenvectors and eigenvalues obtained from each decomposition. This is equivalent to identifying the two principal components of each cluster and rotating them by an angle as specified. For cluster g, we denote $\bar{x}_g^{(j)}$ and $s_g^{(j)}$ as the sample mean and standard deviation of principal component scores associated with the jth rotated principal component, for $j = 1, 2$ and $g = 1, 2$. Now, in the same fashion as in scenario (b), in the space spanned by the two rotated principal components, we replace some observations in cluster g by (x^*_{ig1}, x^*_{ig2}), where

- x^*_{ig1} is a realization of a uniform random variable over the interval

$$\left(\bar{x}_g^{(1)} - s_g^{(1)}, \bar{x}_g^{(1)} + s_g^{(1)}\right).$$

- x^*_{ig2} can be drawn with an equal chance from either a uniform random variable over the interval

$$\left(\bar{x}_g^{(2)} - 10s_g^{(2)}, \bar{x}_g^{(2)} - 3s_g^{(2)}\right)$$

or a uniform random variable over the interval

$$\left(\bar{x}_g^{(2)} + 3s_g^{(2)}, \bar{x}_g^{(2)} + 10s_g^{(2)}\right).$$

Finally, we transform both clusters back to the original data space to obtain the simulated data set.

Note that to generate each Gaussian cluster, we use the `rmvnorm()` function of the **mvtnorm** package (Genz et al., 2021). Under each scenario, we generate 20 complete data sets, in which the number of outliers can account for 10% and 20% of 600 observations. For each of these complete data sets, we hide one of the two variates of 10%, 30%, 50%, and 70% observations under the missing-at-random mechanism per percentage using the `ampute()` function of the R package **mice** (Buuren & Groothuis-Oudshoorn, 2011). In doing so, with each missing percentage, we replicate the complete data set 10 times and introduce a different general MAR missing data pattern to each replicate, yielding 10 different data sets with the same number of observations missing one variate. For ease of presentation, we denote $\kappa \in \{10\%, 30\%, 50\%, 70\%\}$ to refer to the percentage of observations missing one variate. In total, we have $(4)(20)(2)(4)(10) = 6,400$ data sets for this study.

The detailed results of this study containing the average ARIs, TPRs, and FPRs, with associated standard deviations, are available in supplementary material Section 2. Figure 3 shows the box plots of resulting ARIs over different percentages of observations with missing values (κ's), scenarios, and percentages of outliers. Recall that the ARI measures clustering performance of the methods. Here, in every scenario and for every mixture model, the ARI decreases as the percentage of observations with missing values increases. In scenario (a), the MSCNM tends to outperform the other methods in most cases and perform equally in the rest. All the methods perform better with 10% outliers when the percentage of observations with missing values is 50% or below. However, with 10% observations with missing values and 20% outliers, the variability is higher. In scenario (b), the MSCNM clearly outperforms the others, with the MNM being the second best. Moreover, in this scenario, there is less

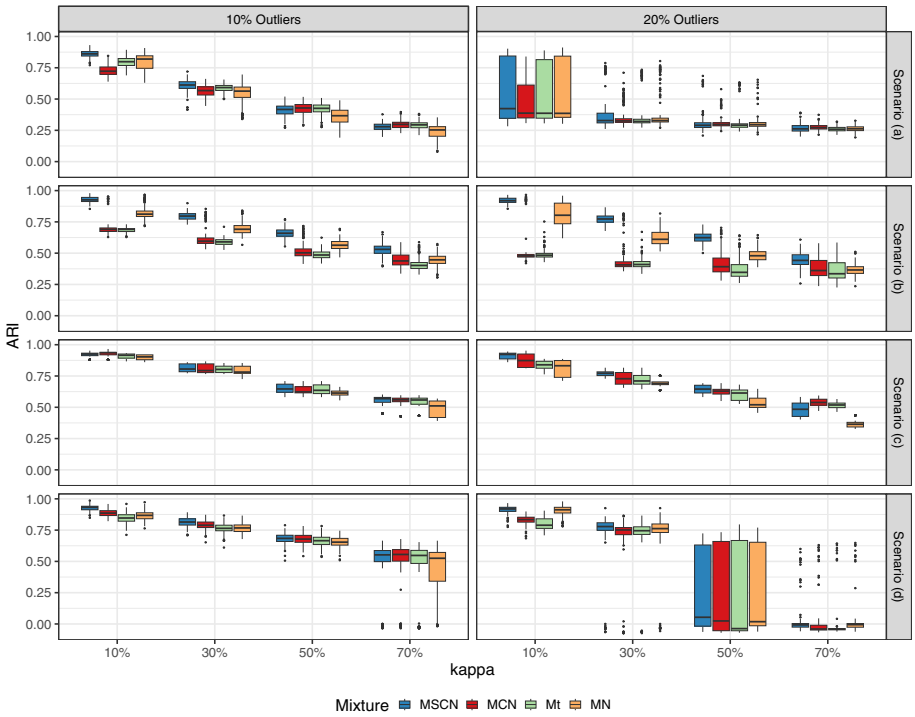


Fig. 3 Box plots of resulting ARIs over different percentages of observations with missing values (denoted by κ) and colors representing the mixture. Each row is a different scenario and each column is a percentage of outliers

difference between the cases with 10% and 20% outliers. In scenario (c), all the methods have similar clustering performance without any particular differences between 10% and 20% outliers. In scenario (d), the MSCNM still has better performance than other competitors. The 10%-outliers case is similar to the corresponding one of scenario (c), except that there are more outlying results, especially when 70% observations have missing values. Outlying results appear more in the 20%-outliers case, and we can see high variability when 50% observations have missing values.

Figures 4 and 5 show the TPRs and FPRs, respectively, over different percentages of observations with missing values (κ's), scenarios, and percentages of outliers. In scenario (a), the MSCNM's TPRs are lower and exhibit greater variability than the competitors for 50% and 70% observations of missing values. The performances of all the methods are similar in the other cases, with the exception of 70% missing data and all missing percentages in the 20%-outliers case of scenario (d) where the MSCNM has a lower TPR. When looking at the FPRs in Fig. 5, the MtM always has the highest values, i.e., the worst performance. The MSCNM and MCNM have similar performance, with the exception of scenario (A) - 20% outliers where the MCNM always has higher FPRs. Overall, although the MtM performs fairly well in detecting true outliers, it also labels many points that are not outliers as outliers. In contrast, the MSCNM and MCNM have better performance, although it is not clear which one is better in terms of outlier detection, unlike in terms of clustering performance where the MSCNM excels.

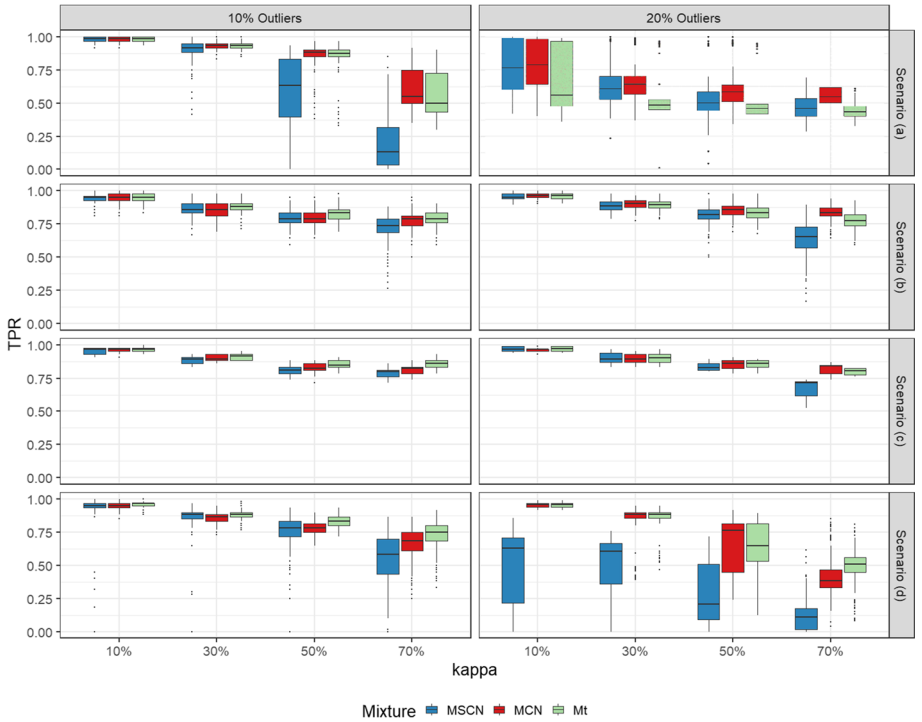


Fig. 4 Box plots of resulting TPRs over different percentages of observations with missing values (denoted by κ) and colors representing the mixture. Each row is a different scenario and each column is a percentage of outliers

5.3 Study 3: Four Overlapping Heavy-Tailed Clusters

Herein, we evaluate our proposed model and its competitors on data sets with overlapping clusters simulated from heavy-tailed distributions. Using the R package **MixSim** (Melnykov et al., 2012), with a desired average pairwise overlap of 0.05, we first generate 20 parameter sets each containing four mean vectors and four covariance matrices. Note that in **MixSim**, the pairwise overlap is defined as a sum of two misclassification probabilities (Maitra & Melnykov, 2010). Then, for each parameter set, we use the corresponding mean vectors and covariance matrices to generate four clusters ($G = 4$), each containing 150 observations ($n = 600$ in total) from a multivariate Student's t whose degrees of freedom are randomly chosen between 5 and 20. With the exact same parameter set, we generate another four clusters but from multivariate contaminated normal distributions whose proportions of good observations and degrees of contamination are randomly chosen between 0.6 and 0.9 and between 2 and 10, respectively. For each of the $(2)(20) = 40$ complete data sets obtained so far, just like in the previous study, we hide one of the two variates of 10%, 30%, 50%, and 70% observations under the missing-at-random mechanism in a way such that there are 10 general MAR missing-data patterns per percentage using the `ampute()` function. In total, we have $(2)(20)(4)(10) = 1,600$ data sets for this study. Figure 6 provides examples of data sets generated for this study.

Figure 7 shows the ARIs of our model and its competitors over different percentages of observations with missing values (κ's) and percentages of outliers for the two distributions

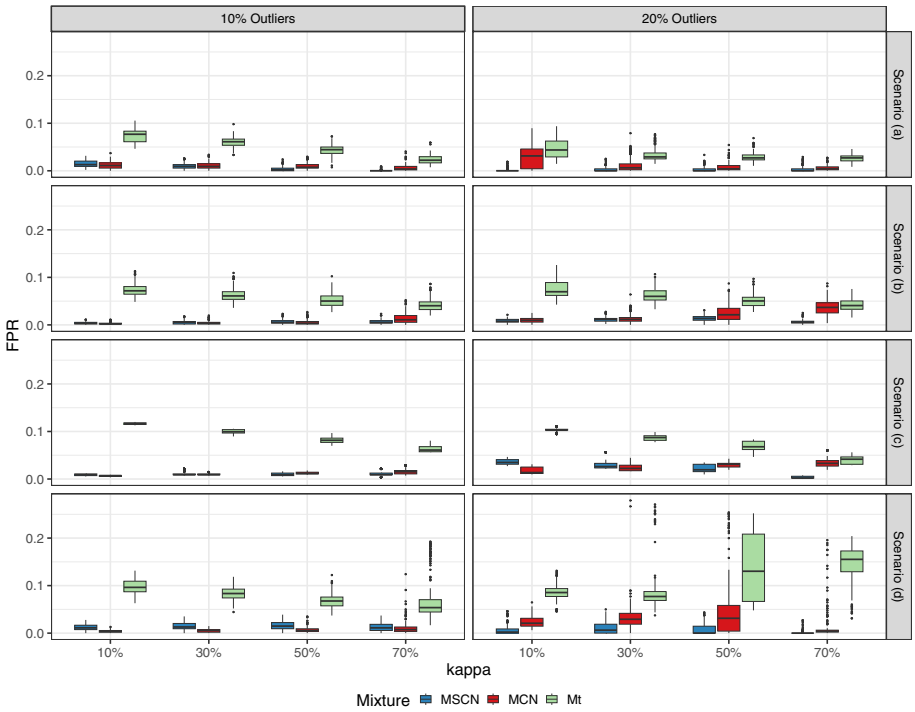
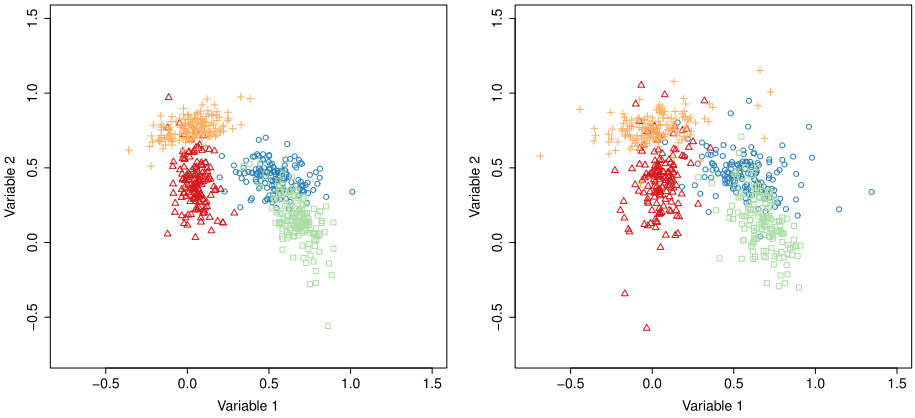


Fig. 5 Box plots of resulting FPRs over different percentages of observations with missing values (denoted by κ) and colors representing the mixture. Each row is a different scenario and each column is a percentage of outliers



(a) Student's t (b) Contaminated Normal

Fig. 6 Examples of data sets generated in the simulation study 3. The mean vectors and covariance matrices here are just one of the 20 parameter sets generated

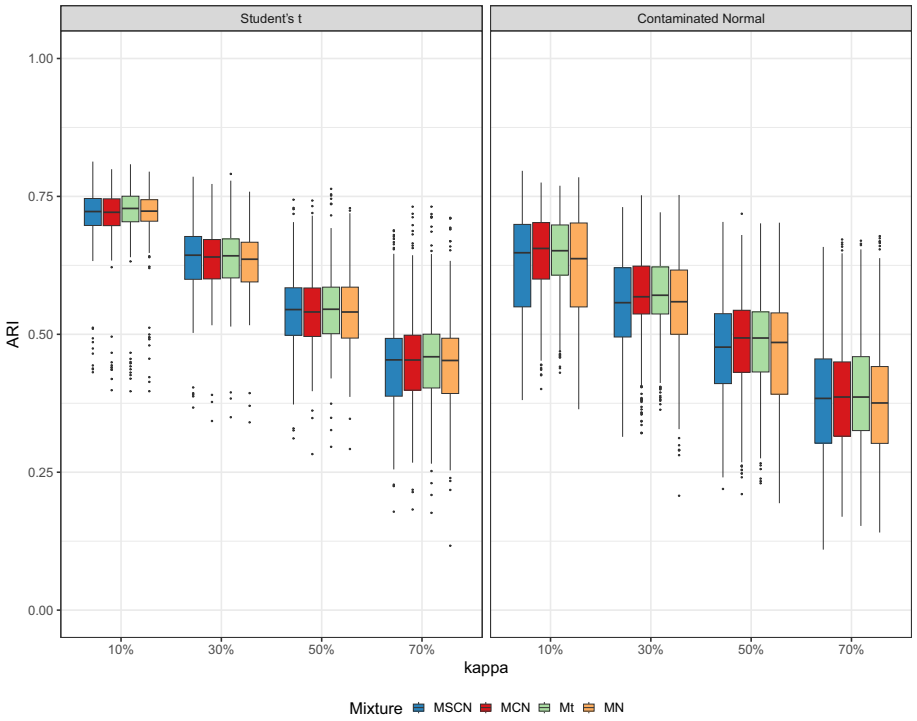


Fig. 7 Box plots of resulting ARIs over different percentages of observations with missing values (denoted by κ) and colors representing the mixture models. Each column is a distribution of the four clusters

mentioned. We can observe a decreasing linear trend as more observations with missing values appear. The stronger overlap between clusters results in much lower ARIs for cases with 10% observations with missing values compared to the corresponding results in study 1, but for higher missing percentages, the results are generally comparable and stable.

5.4 Study 4: Information Criteria for Model Selection

In this study, we revisit all the scenarios mentioned in simulation studies 2 and 3 in the case where 70% of observations have one variate missing. Furthermore, for those in study 2, we focus specifically on data sets with 20% outliers. On these $(6)(20)(10) = 1,200$ data sets of consideration, we fit our model and the other three competitors with the number of mixture components (clusters) to be $G = 2, 3$, and 4. For each fitted model, we record its AIC, BIC, ICL, KIC, KICc, AWE, AIC3, CAIC, and CLC. After $(3)(1,200) = 3,600$ simulations, we obtain the percentage of time each criterion correctly specifies the number of clusters in a particular scenario for every model.

Figure 8 visualizes percentages of times the information criteria correctly specify the number of clusters. Numerical results can be found in Table 12 in supplementary material Section 2. In scenario (d), none of the criteria is able to detect the correct number of clusters for any of the methods. BIC, CAIC, and ICL work best for the MSCN in scenarios (a), (b), and (c) in study 2, while AWE only excels in scenarios (b) and (c). Interestingly, those indices have the worst performance in study 3. Similar behavior of BIC, CAIC, ICL,

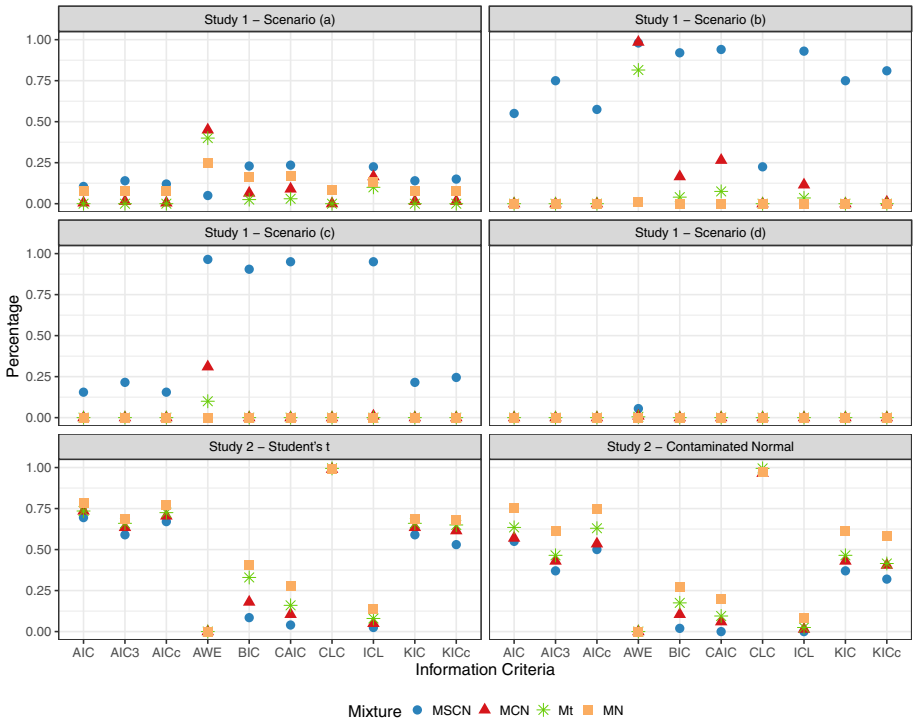


Fig. 8 Percentage of times the information criteria correctly specify the number of clusters in different scenarios of simulation studies 2 and 3 with colors and shapes representing the mixture. For both studies, only the case of 70% observations with missing values is considered. For study 1 specifically, only the case of 20% outliers is considered

and AWE can be seen for the other distributions; however, for the MCN, Mt, and MN those indices also have low success in study 2. To summarize, BIC, CAIC, and ICL tend to detect the correct number of clusters for the MSCN when directional outliers in the direction of the principal components are present, but, for overlapping clusters, the other indices are better. When the outliers are not in the direction of the principal components, all the indices fail. For the MCN, Mt, and MN, CLC works best in detecting the correct number of clusters when there is overlap, but none of the indices works particularly well in the other situations.

6 World Happiness Report data

World Happiness Report by the United Nations Sustainable Development Solutions Network (Sachs et al., 2018) contains six main measurements of happiness in 142 countries. The data used are from 2016. The first two variables are *Log GDP per capita* (per capita Gross Domestic Product on the log scale) and *Healthy life expectancy at birth*, while the last three variables are obtained as averages of binary responses to Gallup World Poll (GWP) questions. Social support, is the response to the question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?" *Freedom to make life choices* answers the question "Are you satisfied or not with your freedom to choose what you do with

your life?" *Perception of corruption* is the average of two questions "Is corruption widespread throughout the government or not?" and "Is corruption widespread within businesses or not?" The last variable, *Generosity*, is measured by the residual of regressing the national average of GWP responses to the question "Have you donated money to a charity in the past month?" on GDP per capita. Out of the 142 countries, 14 have missing values, contributing to a total of 21 missing values in the data set. Moreover, the data is characterized by outliers, i.e., countries with anomalous traits.

Herein, the goal is to find homogeneous groups of countries. The number of clusters is unknown, and the algorithm is thus run with $G = 2, 3, 4$, and 5. The corresponding BIC values are 2128.104, 2127.574, 2202.824, and 2390.527, where the lowest value is obtained by setting $G = 3$. Figure 9 shows the countries colored based on cluster memberships. Cluster 1 has Rwanda as a single outlier, while cluster 3 has 12 outliers, namely China, France, Hong Kong, Iceland, Indonesia, Malta, Myanmar, Norway, Somalia, Thailand, the UK, and Uzbekistan. With respect to cluster 3, China is an outlier for two principal components, while all the others are outliers for one of the components.

The means per each cluster are represented in Fig. 10. In the figure, each line represents a cluster mean vector, and the overall mean vector is set to 0. Countries in cluster 1, red, have the lowest averages for all the variables but the Perception of corruption whose value is equal to the global average. Countries in cluster 2, blue, are average, with a high perception of corruption. Cluster 3, green, contains the happiest countries, with high values for all the variables and a low perception of corruption.

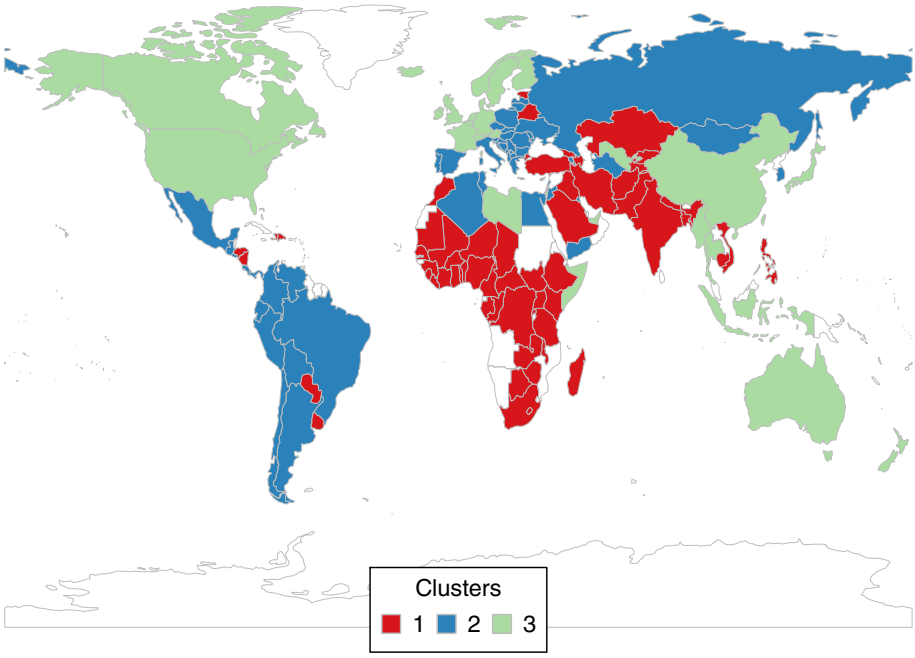


Fig. 9 Countries colored according to cluster membership, white countries have no data

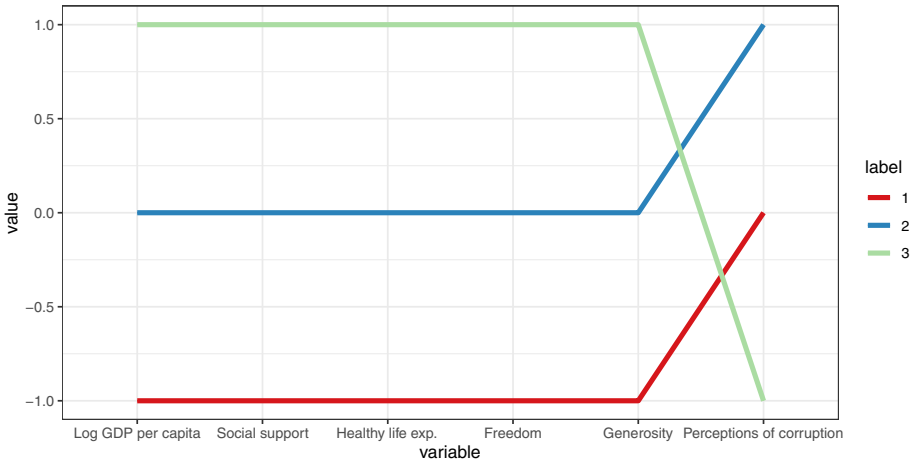


Fig. 10 Parallel coordinate plots, each line represents a cluster mean where the overall mean is set to 0

7 Conclusion

The literature on cluster analysis and model-based clustering is extensive. Many distributions have been used within the model-based clustering framework to offer greater flexibility to the shape of the clusters. Among them, the multivariate contaminated normal (MCN) distribution has the advantage of performing clustering and outlier detection simultaneously. The multiple-scaled contaminated normal distribution (MSCN) solves some limitations of the MCN distribution because it has flexible tails that allow for directional outlier detection and different down-weighting of outlying observations per principal component. This paper extends the MSCN mixture to accommodate data sets with missing values at random (MAR). The marginals of the MSCN distribution are first obtained and then used within the ECM algorithm's framework to perform cluster analysis and directional outlier detection on data sets with values missing at random. The advantages and limitations of the MSCN mixture, compared to its main competitors, are illustrated on simulated data sets. For all the methods, as the percentage of missing values increases, clustering performance measured using the ARI decreases. The impact on outlier detection is less evident; the true positive rate (TPR) slightly decreases as the percentage of missing observation increases, while the false positive rate (FPR) seems stable. Overall, the proposed MSCN mixture has good clustering performance, great FPR, and variable TPR. Moreover, when there is a high percentage of observations with missing values, it yields parameter estimates with lower bias and standard deviations. A study on model selection finds that BIC, CAIC, and ICL are the best indices to select the number of clusters when there is not much overlap among clusters.

Despite the interesting results, the proposed model has three main limitations. First, the MSCN and all the other models used in this paper assume symmetric clusters, which can be a limitation in some real-world applications. Some recent work focused on outlier detection using asymmetric distributions (see for example, Morris et al. 2019). The extension of methods that include asymmetric clusters to data sets with missing values can be very valuable in many applied fields. Second, all the results shown are obtained on a small number of variables, six maximum, because the computational cost of the algorithm does not make it usable on large data sets, and this is an avenue for future research. Third, the initialization

method plays a key role in EM-based algorithms, and this is, even more, emphasized when data sets have missing values and outliers. Future work can focus on a study on the effect of the initialization techniques on the performance of the algorithm.

Appendix A: Characteristic Functions and the Inversion Formula

In statistics, characteristic functions provide a powerful tool for deriving probability density functions by means of Fourier transformations. One major advantage of this approach is that there always exists a unique characteristic function for every probability distribution.

Definition A.1 Let $\boldsymbol{X} = \left(X_1, \ldots, X_p\right)^\top \in \mathbb{R}^p$ be a p-variate random vector, $\boldsymbol{t} = \left(t_1, \ldots, t_p\right)^\top \in \mathbb{R}^p$, and i be an imaginary unit. The function

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = E\left(\exp(i\boldsymbol{t}^\top \boldsymbol{X})\right) \tag{39}$$

is called the characteristic function of $\boldsymbol{X}$.

From a characteristic function, the associated probability density function can be obtained using the inversion formula.

Theorem A.1 (Inversion Formula) Let $\boldsymbol{X} = \left(X_1, \ldots, X_p\right)^\top \in \mathbb{R}^p$ be a p-variate random vector, $\phi_{\boldsymbol{X}}(\boldsymbol{t})$ be the characteristic function of $\boldsymbol{X}$ with $\boldsymbol{t} = \left(t_1, \ldots, t_p\right)^\top \in \mathbb{R}^p$, and i be an imaginary unit. The probability density function of $\boldsymbol{X}$ can be obtained by

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= (2\pi)^{-p} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-i\boldsymbol{t}^\top \boldsymbol{x}\right) \phi_{\boldsymbol{X}}(\boldsymbol{t})\, d\boldsymbol{t} \\ &= (2\pi)^{-p} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-i \sum_{j=1}^{p} t_j x_j\right) \phi_{\boldsymbol{X}}(t_1, \ldots, t_p)\, dt_1 \ldots dt_p. \end{aligned} \tag{40}$$

To obtain the marginals of the MSCN distribution, the propositions describing the characteristic functions of the MN and MCN distributions are needed. The marginals of the MCN and MSCN distribution are outlined in the methodology under Sect. 3.

Proposition A.1 The characteristic function of a p-variate random vector $\boldsymbol{X} = \left(X_1, \ldots, X_p\right)^\top \in \mathbb{R}^p$ that follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(i\boldsymbol{t}^\top \boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t}\right), \tag{41}$$

where $\boldsymbol{t} = \left(t_1, \ldots, t_p\right)^\top \in \mathbb{R}^p$ and i is an imaginary unit.

Appendix B: Proofs

Proposition 3.1

Proof For data generation purposes, the MCN random variable $\boldsymbol{X}$ can be represented as

$$\boldsymbol{X} = \left(V + \frac{1-V}{\eta}\right)^{-1/2} \boldsymbol{Y},$$

where V follows a Bernoulli distribution such that $V = 1$ with probability $\alpha \in (0.5, 1)$ and $V = 0$ with probability $1 - \alpha$; and $\boldsymbol{Y}$ follows an MN distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. By Definition A.1 and the law of total expectation, we can establish the following

$$\begin{aligned}\phi_X(t) = E(\exp(i\boldsymbol{t}^\top \boldsymbol{X})) &= \sum_{v=0}^{1} E(\exp(i\boldsymbol{t}^\top \boldsymbol{X}) \mid V = v) P(V = v) \\ &= \alpha E(\exp(i\boldsymbol{t}^\top \boldsymbol{Y})) + (1-\alpha) E(\exp(i\eta^{1/2}\boldsymbol{t}^\top \boldsymbol{Y})) \\ &= \alpha \phi_{\boldsymbol{Y}}(\boldsymbol{t}) + (1-\alpha)\phi_{\boldsymbol{Y}}(\eta^{1/2}\boldsymbol{t}) \\ &= \alpha \exp\left(i\boldsymbol{t}^\top \boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t}\right) + (1-\alpha)\exp\left(i\boldsymbol{t}^\top \boldsymbol{\mu} - \frac{1}{2}\eta \boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t}\right).\end{aligned}$$

□

Proposition 3.2

Proof From Definition A.1 and the fact that $\tilde{\boldsymbol{Y}}$ contains p independent univariate contaminated normal random variables, the characteristic function of the marginal variable $\boldsymbol{X}_1$ is given by

$$\phi_{\boldsymbol{X}_1}(\boldsymbol{t}) = E\left(\exp(i\boldsymbol{t}^\top \boldsymbol{X}_1)\right) = \prod_{j=1}^{q} \exp(it_j\mu_j) \prod_{h=1}^{p} \phi_{\tilde{Y}_h}\left(\sum_{j=1}^{q} t_j [\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{jh}\right),$$

where from Proposition 3.1, for $h = 1, \ldots p$, we know that

$$\begin{aligned}\phi_{\tilde{Y}_h}\left(\sum_{j=1}^{q} t_j [\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{jh}\right) &= \alpha_h \exp\left[-\frac{1}{2}\left(\sum_{j=1}^{q} t_j [\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{jh}\right)^2\right] \\ &+ (1-\alpha_h)\exp\left[-\frac{1}{2}\eta_h\left(\sum_{j=1}^{q} t_j [\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{jh}\right)^2\right].\end{aligned}$$

□

Proposition 3.6

Proof From Definition A.1 and the fact that we are dealing with linear combinations of independent random variables, we have the characteristic function of $\boldsymbol{X}_1 \mid V_r = v_r, r \in \mathcal{A}$ to be

$$\phi_{\boldsymbol{X}_1 \mid V_r, r\in\mathcal{A}}(\boldsymbol{t}) = \prod_{j=1}^{q} \exp(it_j\mu_j) \prod_{r\in\mathcal{A}} \phi_{\tilde{Y}_r}\left(\sum_{j=1}^{q} t_j [\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{jr}\right) \prod_{s\in\mathcal{B}} \phi_{\tilde{Y}_s}\left(\sum_{j=1}^{q} t_j [\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{js}\right).$$

Herein, for $r \in \mathcal{A}$, $\tilde{Y}_r$ follows a univariate normal distribution with mean 0 and variance 1 if $v_r = 1$ or variance η_r if $v_r = 0$. Thus,

$$\begin{aligned}&\phi_{\tilde{Y}_r}\left(\sum_{j=1}^{q} t_j [\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{jr}\right) \\ &= \left\{\exp\left[-\frac{1}{2}\left(\sum_{j=1}^{q} t_j [\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{jr}\right)^2\right]\right\}^{v_r} \left\{\exp\left[-\frac{1}{2}\eta_r\left(\sum_{j=1}^{q} t_j [\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}]_{jr}\right)^2\right]\right\}^{1-v_r}.\end{aligned}$$

On the other hand, for $s \in \mathcal{B}$, $\tilde{Y}_s$ follows a univariate contaminated normal distribution with mean 0, variance 1, proportion of good observation α_s, and degree of contamination η_s. As the result,

$$\phi_{\tilde{Y}_s}\left(\sum_{j=1}^{q} t_j[\mathbf{\Gamma}\mathbf{\Lambda}^{1/2}]_{js}\right) = \left\{\alpha_s \exp\left[-\frac{1}{2}\left(\sum_{j=1}^{q} t_j[\mathbf{\Gamma}\mathbf{\Lambda}^{1/2}]_{js}\right)^2\right] + (1-\alpha_s)\exp\left[-\frac{1}{2}\eta_s\left(\sum_{j=1}^{q} t_j[\mathbf{\Gamma}\mathbf{\Lambda}^{1/2}]_{js}\right)^2\right]\right\}.$$

□

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00357-023-09450-2.

Funding This material is based upon work supported by the National Science Foundation under Grant No. 2209974

Data Availability The data that support the findings of this study are available from the corresponding author upon request.

Code Availability The code can be found on github at https://github.com/cristinatortora/MSCN_missing.

Declarations

Ethical Approval The authors agree to follow Springer ethical conduct.

Conflict of Interest The authors declare no competing interest.

References

Aitken, A. (1926). A series formula for the roots of algebraic and transcendental equations. *Proceedings of the Royal Society of Edinburgh, 45*(1), 14–22.

Aitkin, M., & Wilson, G. T. (1980). Mixture models, outliers, and the EM algorithm. *Technometrics, 22*(3), 325–331.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In: E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 199–213). Springer New York, New York, NY

Akogul, S., & Erisoglu, M. (2016). A comparison of information criteria in clustering based on mixture of multivariate normal distributions. *Mathematical and Computational Applications, 21*(3), 34.

Andrews, J. L., & McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Statistics and Computing, 22*(5), 1021–1029.

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

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

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics, 49*(3), 803.

Berntsen, J., Espelid, T. O., & Genz, A. (1991). An adaptive algorithm for the approximate calculation of multiple integrals. *ACM Transactions on Mathematical Software, 17*(4), 437–451.

Biernacki, C., & Govaert, G. (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics,* (pp. 451–457)

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(7), 719–725.
Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix. In: O. Opitz, B. Lausen, & R. Klar (Eds.), *Information and Classification* (pp. 40–54). Berlin, Heidelberg. Springer Berlin Heidelberg
Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370.
Browne, R. P., & McNicholas, P. D. (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics, 43*(2), 176–198.
Broyden, C. (1970). The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and its Applications, 6*(2), 76–90.
Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society. Series B (Methodological), 22*(2), 302–306.
Buuren, S. v. (2021). *Flexible imputation of missing data*. Chapman & Hall/CRC interdisciplinary statistics series. Chapman & Hall/CRC, Boca Raton, 2nd ed.
Buuren, S. v., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software,45*(3), 1–67
Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics & Probability Letters, 42*(4), 333–343.
Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition, 28*(5), 781–793.
Coretto, P., & Hennig, C. (2016). Robust improper maximum likelihood: Tuning, computation, and a comparison with other methods for robust Gaussian clustering. *Journal of the American Statistical Association, 111*(516), 1648–1659.
Cuesta-Albertos, J., Matrán, C., & Mayo-Iscar, A. (2008). Robust estimation in the normal mixture model based on robust clustering. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 70*(4), 779–802.
Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*(1), 1–22.
Dooren, P. V., & Ridder, L. D. (1976). An adaptive algorithm for numerical integration over an n-dimensional cube. *Journal of Computational and Applied Mathematics, 2*(3), 207–217.
Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal, 13*(3), 317–322.
Forbes, F., & Wraith, D. (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: Application to robust clustering. *Statistics and Computing, 24*(6), 971–984.
Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*(458), 611–631.
Franczak, B. C., Browne, R. P., & McNicholas, P. D. (2014). Mixtures of shifted asymmetric Laplace distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(6), 1149–1157.
Franczak, B. C., Tortora, C., Browne, R. P., & McNicholas, P. D. (2015). Unsupervised learning via mixtures of skewed distributions with hypercube contours. *Pattern Recognition Letters, 58*, 69–76.
Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Series in Statistics. Springer New York
Gallegos, M. T., & Ritter, G. (2005). A robust method for cluster analysis. *The Annals of Statistics, 33*(1), 347–380.
Gallegos, M. T., & Ritter, G. (2009). Trimmed ML estimation of contaminated mixtures. *Sankhyā: The Indian Journal of Statistics, Series A (2008-), 71*(2), 164–220.
Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2021). mvtnorm: Multivariate normal and *t* distributions. R package version 1.1-3.
Ghahramani, Z., & Jordan, M. I. (1994). Learning from incomplete data. Technical report, Defense Technical Information Center, Fort Belvoir, VA
Goldfarb, D. (1970). A family of variable metric methods derived by variational means. *Mathematics of Computation, 24*(109), 23–26.
Goren, E. M., & Maitra, R. (2022). Fast model-based clustering of partial records. *Stat,11*(1), e416. Publisher: John Wiley & Sons, Ltd.
Greco, L., & Agostinelli, C. (2020). Weighted likelihood mixture modeling and model-based clustering. *Statistics and Computing, 30*(2), 255–277.
Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218.
Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*(2), 297–307.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis*. Pearson Prentice Hall, Upper Saddle River, N.J, 6th ed. edition. OCLC: ocm70867129.

Karlis, D., & Santourian, A. (2009). Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing, 19*(1), 73–83.

Karlis, D., & Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis, 41*, 577–590.

Kaufman, L., & Rousseeuw, P. J. (Eds.). (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, USA.

Lin, T. I. (2009). Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis, 100*(2), 257–265.

Lin, T.-I. (2014). Learning from incomplete data via parameterized t mixture models through eigenvalue decomposition. *Computational Statistics & Data Analysis, 71*, 183–195.

Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 3rd ed.

Liu, C., & Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika, 81*(4), 633–648.

Maitra, R., & Melnykov, V. (2010). Simulating data to study performance of finite mixture modeling and clustering algorithms. *Journal of Computational and Graphical Statistics*, *19*(2), 354–376. Publisher: Taylor & Francis

McLachlan, G. J., & Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, N.J.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, USA.

McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification, 33*(3), 331–373.

McNicholas, P., Murphy, T., McDaid, A., & Frost, D. (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Second Special Issue on Statistical Algorithms and Software, 54*(3), 711–723.

Melnykov, V., Chen, W.-C., & Maitra, R. (2012). MixSim : An R package for simulating data to study performance of clustering algorithms. *Journal of Statistical Software*, *51*(12)

Melnykov, V. (2013). Challenges in model-based clustering. *Wiley interdisciplinary reviews: computational statistics, 5*(2), 135–148.

Melnykov, V., & Maitra, R. (2010). Finite mixture models and model-based clustering. *Statistics Surveys, 4*, 80–116.

Meng, X.-L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika, 80*(2), 267–278.

Michael, S., & Melnykov, V. (2016). An effective strategy for initializing the em algorithm in finite mixture models. *Advances in Data Analysis and Classification, 10*, 563–583.

Morris, K., Punzo, A., Blostein, M., & McNicholas, P. D. (2019). Asymmetric clusters and outliers: Mixtures of multivariate contaminated shifted asymmetric laplace distributions. *Computational Statistics and Data Analysis, 132*, 145–166.

Narasimhan, B., Johnson, S. G., Hahn, T., Bouvier, A., & Kiêu, K. (2022). cubature: Adaptive multivariate integration over hypercubes.

Novi Inverardi, P. L., & Taufer, E. (2020). Outlier detection through mixtures with an improper component. *Electronic Journal of Applied Statistical Analysis, 13*(1), 146–163.

Peel, D., & McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing, 10*(4), 339–348.

Punzo, A., Mazza, A., & McNicholas, P. D. (2018). ContaminatedMixt: An R package for fitting parsimonious mixtures of multivariate contaminated normal distributions. *Journal of Statistical Software, 85*(10), 1–25.

Punzo, A., & McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal, 58*(6), 1506–1537.

Punzo, A., & Tortora, C. (2021). Multiple scaled contaminated normal distribution and its application in clustering. *Statistical Modelling, 21*(4), 332–358.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846–850.

Ritter, G. (2014). *Robust cluster analysis and variable selection*. Chapman and Hall/CRC, 1st ed.

Rubin, D. B. (Ed.). (1987). *Multiple imputation for nonresponse in surveys*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., Hoboken, NJ, USA

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*(434), 473–489.
Sachs, J. D., Layard, R., Helliwell, J. F., et al. (2018). World happiness report 2018. Technical report.
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics,6*(2)
Seghouane, A., & Bekara, M. (2004). A small sample model selection criterion based on Kullback's symmetric divergence. *IEEE Transactions on Signal Processing, 52*(12), 3314–3323.
Serafini, A., Murphy, T. B., & Scrucca, L. (2020). Handling missing data in model-based clustering. arXiv preprint arXiv:2006.02954
Shanno, D. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation, 24*(111), 647–656.
Shireman, E., Steinley, D., & Brusco, M. J. (2017). Examining the effect of initialization strategies on the performance of Gaussian mixture modeling. *Behavior Research Methods, 49*(1), 282–293.
Soetaert, K. (2009). *rootSolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations*. R package 1.6.
Soetaert, K., & Herman, P. M. (2009). *A practical guide to ecological modelling. Using R as a Simulation Platform*. Springer. ISBN 978-1-4020-8623-6
Steinley, D. (2004). Properties of the Hubert-Arabie adjusted Rand index. *Psychological Methods, 9*(3), 386–396.
Sugasawa, S., & Kobayashi, G. (2022). Robust fitting of mixture models using weighted complete estimating equations. *Computational Statistics & Data Analysis, 174*, 107526.
Tong, H., & Tortora, C. (2022). MixtureMissing: Robust model-based clustering for data sets with missing values at random. R package version 1.0.2.
Tong, H., & Tortora, C. (2022). Model-based clustering and outlier detection with missing data. *Advances in Data Analysis and Classification, 16*(1), 5–30.
Tortora, C., Punzo, A., & Tran, L. (2023). *MSclust: Multiple-scaled clustering*. R package version 1.0.3.
Tortora, C., Franczak, B. C., Browne, R. P., & McNicholas, P. D. (2019). A mixture of coalesced generalized hyperbolic distributions. *Journal of Classification, 36*(1), 26–57.
Tran, L., & Tortora, C. (2021). How many clusters are best? Investigating model selection in robust clustering. In *JSM Proceedings, Statistical Learning and Data Science Section.* Alexandria, VA: American Statistical Association. 1159–1180 2021.
Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In: I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics: Essays in Honor of Harold Hotelling* (pp. 448–485). Stanford University Press, Stanford, CA
Wang, W.-L., & Lin, T.-I. (2015). Robust model-based clustering via mixtures of skew-t distributions with missing information. *Advances in Data Analysis and Classification, 9*(4), 423–445.
Wang, H., Zhang, Q., Luo, B., & Wei, S. (2004). Robust mixture modelling using multivariate t-distribution with missing information. *Pattern Recognition Letters, 25*(6), 701–710.
Wei, Y., Tang, Y., & McNicholas, P. D. (2019). Mixtures of generalized hyperbolic distributions and mixtures of skew-t distributions for model-based clustering with incomplete data. *Computational Statistics & Data Analysis, 130*, 18–41.
Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *Annals of Mathematical Statistics, 3*, 163–195.
Wolfe, J. H. (1965). A computer program for the maximum likelihood analysis of types. USNPRA Technical Bulletin 65-15, U.S. Naval Personnel Research Activity, San Diego, USA.
You, J., Li, Z., & Du, J. (2023). A new iterative initialization of em algorithm for gaussian mixture models. *Plos one, 18*(4), e0284114.