Multiple scaled contaminated normal distribution and its application in clustering

Antonio Punzo¹ and Cristina Tortora²

¹Department of Economics and Business, University of Catania, Catania, Italy.
²Department of Mathematics and Statistics, San José State University, San José, CA, USA.

Abstract: The multivariate contaminated normal (MCN) distribution represents a simple heavy-tailed generalization of the multivariate normal (MN) distribution to model elliptical contoured scatters in the presence of mild outliers (also referred to as 'bad' points herein) and automatically detect bad points. The price of these advantages is two additional parameters: proportion of good observations and degree of contamination. However, in a multivariate setting, only one proportion of good observations and only one degree of contamination may be limiting. To overcome this limitation, we propose a multiple scaled contaminated normal (MSCN) distribution. Among its parameters, we have an orthogonal matrix $\Gamma$. In the space spanned by the vectors (principal components) of $\Gamma$, there is a proportion of good observations and a degree of contamination for each component. Moreover, each observation has a posterior probability of being good with respect to each principal component. Thanks to this probability, the method provides directional robust estimates of the parameters of the nested MN and automatic directional detection of bad points. The term ‘directional’ is added to specify that the method works separately for each principal component. Mixtures of MSCN distributions are also proposed, and an expectation-maximization algorithm is used for parameter estimation. Real and simulated data are considered to show the usefulness of our mixture with respect to well-established mixtures of symmetric distributions with heavy tails.

Key words: contaminated normal distribution, heavy-tailed distributions, multiple scaled distributions, EM algorithm, mixture models, model-based clustering

Received December 2018; revised March 2019; accepted July 2019

1 Introduction

Statistical inference dealing with continuous multivariate data is commonly focused on the multivariate normal (MN) distribution, with mean $\mu$ and covariance matrix $\Sigma$, due to its computational and theoretical convenience. However, for many applied problems, the tails of this distribution are lighter than required. This is often due to the presence of outliers, that is, observations that do not comply with the model assumed and that affect the estimation of $\mu$ and $\Sigma$ (Maronna and Yohai, 2014). This has created a need for techniques that detect outliers and for which parameter estimates are robust in their presence (see, e.g., Devlin et al., 1981).
Outliers may roughly be divided into two types: gross and mild (Ritter, 2015). Outliers are ‘gross’ when they do not appear to be sampled from a population and, consequently, cannot be modelled by a distribution (Ritter, 2015, pp. 79–80), they are unpredictable and incalculable. In the presence of gross outliers, the statistician is recommended to choose a method for suppressing them. Outliers are ‘mild’ with respect to the MN distribution (reference distribution) when they do not deviate from the assumed MN model and are not strongly outlying; rather, they produce an overall distribution that is too heavy-tailed to be modelled by the MN. For a discussion of the concept of reference distribution see Davies and Gather (1993). Therefore, mild outliers (also referred to as bad points herein, in analogy with Aitkin and Wilson, 1980) can be modelled by means of more flexible distributions, usually symmetric and endowed with heavy tails (Ritter, 2015). To define them, the MN distribution is typically embedded in a larger symmetric model with one or more additional parameters denoting the deviation from normality in terms of tail weight. In this context, the multivariate t (Mt) distribution (see, e.g., Lange et al., 1989; Kotz and Nadarajah, 2004), the heavy-tailed versions of the multivariate power exponential (MPE) distribution (Gómez-Villegas et al., 2011) and the multivariate leptokurtic-normal (MLN) distribution (Bagnato et al., 2017) represent possible symmetric alternatives in the subclass of the elliptically contoured distributions.

The models/methods discussed above robustify the estimation of $\mu$ and $\Sigma$, when compared to the reference MN distribution, and detect outliers. However, the resulting detection rule is not ‘automatic’ because it requires a subjective threshold choice. To overcome this problem, we can consider the multivariate contaminated normal (MCN) distribution of Tukey (1960), a further common and simple elliptically contoured generalization of the MN distribution having heavier tails for the occurrence of bad points. It is a two-component normal mixture in which one of the components—with a large prior probability $\alpha$—represents the good observations (reference distribution), and the other with a small prior probability $1 - \alpha$, the same mean $\mu$ and an inflated (with respect to $\eta > 1$) covariance matrix $\eta\Sigma$—represents the bad observations (Aitkin and Wilson, 1980). Advantageously, once the MCN distribution is fitted to the observed data, by means of maximum a posteriori (MAP) probabilities, each observation, if desired (Berkane and Bentler, 1988), can be automatically (i.e., without the need of a subjective choice) classified as good or bad. Moreover, bad points are automatically down-weighted in the estimation of $\mu$ and $\Sigma$. Thus, the MCN distribution represents a model for the simultaneous robust estimation of $\mu$ and $\Sigma$, and the detection of mild outliers.

However, the MCN distribution has some drawbacks that are listed below.

1. When the scale matrix $\Sigma$ of the MCN distribution is diagonal, the variates are pair-wise uncorrelated, but can be statistically dependent (with strength of dependence affected by the values of the parameters $\alpha$ and $\eta$).
2. In relation to the previous point, the product of independent univariate contaminated normal (CN) distributions, with the same parameters $\alpha$ and $\eta$, is not an MCN distribution.
3. The MCN distribution, being a normal-scale mixture, belongs to the subclass of elliptically contoured distributions (see, e.g., Gómez et al., 2003, p. 347). Thus, its flexibility in terms of symmetric shapes is limited.

4. Another limitation of the MCN distribution is that all marginals are CN distributions with the same parameters $\alpha$ and $\eta$ and, hence the same amount of tail weight. Therefore, it is not possible to account for different tail behaviours.

5. In terms of robustness, bad points are automatically down-weighted in the maximum likelihood (ML) estimation of $\mu$ and $\Sigma$ but in the same way for each dimension. This does not take into consideration the fact that points may be bad in some dimensions but good in others (Alqallaf et al., 2009).

6. In relation to the previous point, the procedure to detect outliers induced by the MCN distribution could be defined as omnibus in the sense that when a point is detected as bad, it is globally bad. As a practical consequence, once the point is detected as bad, we do not know the dimension(s) yielding this decision.

To overcome these drawbacks, we introduce the multiple scaled contaminated normal (MSCN) distribution. The genesis of our model follows the idea developed by Forbes and Wraith (2014) to define the multiple scaled $t$ (MS$t$) distribution. The key elements of the approach are the decomposition of $\Sigma$ by eigenvalues and eigenvectors matrices $\Lambda$ and $\Gamma$ and the introduction of Bernoulli random variables indicating whether a point is good or bad separately for each dimension (principal component) of the space spanned by the columns of $\Gamma$. The result is a distribution in which the scalar parameters $\alpha$ and $\eta$ of the MCN distribution are replaced by two vectors, $\alpha$ and $\eta$, controlling the proportion of good points and the degree of contamination, respectively, separately for each principal component.

The MSCN distribution offers a remedy to the drawbacks of the MCN distribution discussed above in the following way. With respect to drawback 1, if the scale matrix $\Sigma$ of the MSCN distribution is diagonal, then the variates are independent; as a by-product of this property, the MSCN distribution contains the product of independent univariate CN distributions as a special case, thus providing a remedy to drawback 2. With respect to drawback 3, our distribution includes a greater variety of shapes and, in particular, contours that are symmetric but not necessarily elliptical. As concerns drawback 4, the MSCN distribution lets the parameters $\alpha$ and $\eta$ to be set or estimated differently in each principal component. It is then possible to account for very different tail behaviours across principal components. With respect to drawback 5, the down-weighting of the observations, the estimation of $\mu$ and $\Sigma$ are free to vary over principal components (directional robustness). Finally, with respect to drawback 6, the procedure to detect outliers induced by the MSCN distribution works separately for each principal component, such that a point may be detected as bad with respect to some principal components only (directional outlier detection).

The article is organized as follows. Section 2, after the summary of some results surrounding the MCN distribution, presents the main contribution of the work, namely the MSCN distribution. Section 3 illustrates a further proposal of the present article, the use of the MSCN distribution in robust model-based clustering. This
section also presents an expectation-maximization (EM) algorithm to fit mixtures of MSCN distributions. Further computational and operational aspects are discussed in Section 4. Section 5 investigates the performance of the proposed mixtures and compares them with mixtures of some well-established multivariate symmetric distributions with heavy tails, with regard to artificial and real data. Conclusions are given in Section 6 together with some avenues for further research.

2 Methodology

2.1 Preliminaries: The multivariate contaminated normal

A \(d\)-variate random vector \(X = (X_1, \ldots, X_d)^\top\) is said to follow the MCN distribution with mean vector \(\mu\), scale matrix \(\Sigma\), proportion of good points \(\alpha \in (0, 1)\) and degree of contamination \(\eta > 1\) if its joint probability density function (pdf) is given by

\[
    f_{\text{MCN}}(x; \mu, \Sigma, \alpha, \eta) = \alpha f_{\text{MN}}(x; \mu, \Sigma) + (1 - \alpha) f_{\text{MN}}(x; \mu, \eta \Sigma),
\]

where \(f_{\text{MN}}(\cdot; \mu, \Sigma)\) denotes the pdf of a \(d\)-variate random vector having the MN distribution with mean vector \(\mu\) and covariance matrix \(\Sigma\). In the following, when \(d = 1\), we will substitute the subscripts MN and MCN with N and CN, respectively. If \(X\) follows the MCN distribution, we write \(X \sim \mathcal{CN}_d(\mu, \Sigma, \alpha, \eta)\). As a special case of (2.1), if \(\alpha\) and \(\eta\) tend to one (we do not consider the case of \(\alpha\) tending to 0; for details, see Section 4.4), we obtain the MN distribution with mean vector \(\mu\) and covariance matrix \(\Sigma\); in symbols, \(X \sim \mathcal{N}_d(\mu, \Sigma)\).

An advantage of (2.1) with respect to the multivariate t (Mt) distribution is that once the parameters in \(\theta = \{\mu, \Sigma, \alpha, \eta\}\) are estimated (say \(\hat{\theta} = \{\hat{\mu}, \hat{\Sigma}, \hat{\alpha}, \hat{\eta}\}\)), we can establish whether a generic point \(x^*\) is good via the \textit{a posteriori} probability

\[
    P\left( x^* \text{ is good} \mid \hat{\theta} \right) = \frac{\hat{\alpha} f_{\text{MN}}(x^*; \hat{\mu}, \hat{\Sigma})}{f_{\text{MCN}}(x^*; \hat{\theta})},
\]

and \(x^*\) will be considered good if \(P(x^* \text{ is good} \mid \hat{\theta}) > 1/2\), while it will be considered bad otherwise.

2.2 Proposal: The multiple scaled contaminated normal normal

In the same spirit of Forbes and Wraith (2014), we propose the extension of the MCN distribution to an MSCN distribution. It consists in using the classical eigen decomposition \(\Sigma = \Gamma \Lambda \Gamma^\top\) of the scale matrix, where \(\Lambda\) is the diagonal matrix of the eigenvalues of \(\Sigma\) and \(\Gamma\) is a \(d \times d\) orthogonal matrix whose columns are the normalized eigenvectors of \(\Sigma\), ordered according to their eigenvalues. Each element in the right-hand side of this decomposition has a different geometric interpretation: \(\Lambda\) determines the size and shape of the scatter, while \(\Gamma\) determines its orientation. Moreover, we introduce the indicator variable \(V_h\) to be good \((V_h = 1)\) or bad \((V_h = 0)\) with respect to the \(h\)th dimension, \(h = 1, \ldots, d\), and further define the \(d \times d\) diagonal
where the sum runs over all the \( 2^d \) possible 0/1 patterns \( \nu \), \( \alpha = (\alpha_1, \ldots, \alpha_d)^\top \), \( \eta = (\eta_1, \ldots, \eta_d)^\top \), and

\[
p(\nu; \alpha) = \prod_{b=1}^{d} p(\nu_b; \alpha_b),
\]

with \( p(\nu_b; \alpha_b) = \alpha_b^{\nu_b} (1 - \alpha_b)^{1-\nu_b} \) and \( \sum_{\nu} p(\nu; \alpha) = 1 \). So, model (2.3) can be seen as a mixture of \( 2^d \) MN distributions \( \mathcal{N}_d(\mu, \Gamma W_{\nu} \Lambda \Gamma^\top) \) with weights \( p(\nu; \alpha) \). As an example, in the case \( d = 2 \), model (2.3) reduces to a mixture of 4 MN distributions with the same mean \( \mu \), eigenvalues matrix \( \Lambda \) and eigenvectors matrix \( \Gamma \), but with different patterns \( \nu \), diagonal matrices of inverse weights \( W_{\nu} \) and components probabilities \( p(\nu; \alpha) \); see Table 1. If \( X \) follows the MSCN distribution, we write \( X \sim \mathcal{SCN}_d(\mu, \Gamma, \Lambda, \alpha, \eta) \). The pdf in (2.3) can be equivalently written as

\[
f_{\text{MSCN}}(x; \mu, \Gamma, \Lambda, \alpha, \eta) = \prod_{b=1}^{d} f_{\text{CN}} \left( \left[ \Gamma^\top (x - \mu) \right]_b; 0, \lambda_b, \alpha_b, \eta_b \right),
\]

where \( \Gamma^\top (x - \mu) \) is the principal components transform of \( x \) and can be thought of as a rotation and a recentring of \( x \), \( \left[ \Gamma^\top (x - \mu) \right]_b \) is the \( b \)th element of \( \Gamma^\top (x - \mu) \) and \( \lambda_b \) the \( b \)th diagonal element of \( \Lambda \) (or, equivalently, the \( b \)th eigenvalue of \( \Sigma \)).

Table 1  Model (2.3). Values of \( \nu, W_{\nu} \) and \( p(\nu;\alpha) \), for each mixture component, in the case \( d = 2 \)

| Component 1 | Component 2 | Component 3 | Component 4 |
|-------------|-------------|-------------|-------------|
| \( \begin{pmatrix} 0 \\ 0 \end{pmatrix} \) | \( \begin{pmatrix} 0 \\ 1 \end{pmatrix} \) | \( \begin{pmatrix} 1 \\ 0 \end{pmatrix} \) | \( \begin{pmatrix} 1 \\ 1 \end{pmatrix} \) |
| \( \begin{pmatrix} \eta_1 & 0 \\ 0 & \eta_2 \end{pmatrix} \) | \( \begin{pmatrix} \eta_1 & 0 \\ 0 & 1 \end{pmatrix} \) | \( \begin{pmatrix} 1 & 0 \\ 0 & \eta_2 \end{pmatrix} \) | \( \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \) |
| \( (1 - \alpha_1)(1 - \alpha_2) \) | \( (1 - \alpha_1)\alpha_2 \) | \( \alpha_1(1 - \alpha_2) \) | \( \alpha_1\alpha_2 \) |
Figure 1 Examples of contour plots of bivariate SCN distributions with $\mu = 0$ and $\Lambda = 0.75I$.

Note that apart from the univariate case $d = 1$, the MCN distribution is not a special or limiting case of the MSCN. The number of free parameters of the MSCN distribution is $d + d(d - 1)/2 + d + d + d$, where the addends refer to $\mu$, $\Gamma$, $\Lambda$, $\alpha$ and $\eta$, respectively.

In the bivariate case ($d = 2$), Figure 1 shows, via isodensities, some possible shapes of the MSCN distribution by varying $\Gamma$, $\alpha$ and $\eta$, with the mean vector and the eigenvalue matrix fixed, respectively, to $\mu = 0$ and $\Lambda = 0.75I$, where $I$ denotes the identity matrix. The orientation matrix $\Gamma$ is seen as a rotation matrix of angle $\theta$, that is,

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

Figure 1 clearly shows how the shape of the MSCN distribution is not constrained as elliptical, although the symmetry is preserved. In particular, the choices made for $\alpha$ and $\eta$ produce, among others, ‘smoothed’ rhomboidal (Figures 1(a) and 1(b)) and starred (Figures 1(c) and 1(d)) contours.
Finally, it is easy to show that if $Y \sim \mathcal{N}_d(0, I)$, then
\[
X = \mu + \Gamma \Lambda^{1/2} W^{1/2}_0 Y \sim \text{SCN}_d(\mu, \Gamma, \Lambda, \alpha, \eta).
\] (2.4)

This alternative way to see the MSCN distribution may be useful for random generation. Moreover, Equation (2.4) makes it easier to see that univariate marginal distributions are linear combinations of CN distributions for which, in general, no closed-form expression is available, although it is possible to show that symmetry is preserved. Therefore, univariate marginal distributions are not in general CN distributions.

3 Mixtures of MSCN distributions

Finite mixtures of distributions are commonly used in statistical modelling as a powerful device for clustering and classification by often assuming that each mixture component represents a cluster (or group) into the original data (see McLachlan and Basford, 1988; Fraley and Raftery, 1998; Böhning, 2000; McNicholas, 2016).

For continuous multivariate random variables, attention is commonly focused on mixtures of MN distributions. However, in clustering applications, data are often contaminated by mild outliers (see, e.g., Bock, 2002; Gallegos and Ritter, 2009; Ritter, 2015), affecting the estimation of the component means and covariance matrices and the recovery of the underlying clustering structure. For MN mixtures (MNMs), one of the possible solutions used to deal with mild outliers is the ‘component-wise’ approach: the component MN distributions are separately protected against mild outliers by embedding them in more general heavy-tailed, usually symmetric, distributions. Examples are $Mt$ mixtures ($MtMs$; Peel and McLachlan, 2000), MPE mixtures (MPEMs; Zhang and Liang, 2010; Dang et al., 2015), MLN mixtures (Bagnato et al., 2017) and $MSt$ mixtures (MStMs; Forbes and Wraith, 2014). These methods robustify, by down-weighting, the estimation of the component means and covariance matrices with respect to mixtures of MN distributions, but they do not automatically detect bad points, although an a posteriori procedure (i.e., a procedure taking place once the model is fitted) to detect bad points with $MtMs$ is illustrated by McLachlan and Peel (2000). To overcome this problem, Punzo and McNicholas (2016) introduced MCN mixtures (MCNMs); for further recent uses of the MCN distribution in model-based clustering, see Punzo et al. (2020), Punzo and McNicholas (2017), Punzo and Maruotti (2016), Maruotti and Punzo (2017) and Farcomeni and Punzo (2019).

3.1 The model

For a $d$-variate random vector $X$, the pdf of a MSCN mixture (MSCNM) with $k$ components can be written as
The corresponding two-level complete-data log-likelihood can be so written as

$$f_{MSCNM}(x; \Psi) = \sum_{j=1}^{k} \pi_j f_{MSCN}(x; \mu_j, \Gamma_j, \Lambda_j, \alpha_j, \eta_j), \quad (3.1)$$

where $\pi_j$ is the mixing proportion of the $j$th component, with $\pi_j > 0$ and $\sum_{j=1}^{k} \pi_j = 1$, $\mu_j = (\mu_{1j}, \ldots, \mu_{dj})^T$, $\Lambda_j = \text{diag}(\lambda_{1j}, \ldots, \lambda_{dj})$, $\alpha_j = (\alpha_{1j}, \ldots, \alpha_{dj})^T$, $\eta_j = (\eta_{1j}, \ldots, \eta_{dj})^T$ and $\Psi$, of dimension $(k - 1) + k[d + d(d - 1)/2 + d + d + d]$, contains all of the parameters of the mixture.

### 3.2 ML estimation via the EM algorithm

Let $x_1, \ldots, x_n$ be a random sample from model (3.1). To find the ML estimates for its parameters $\Psi$, we adopt the classical EM algorithm (Dempster et al., 1977), which is a natural approach for ML estimation when the data are incomplete.

In the implementation of the EM algorithm for our model, we have a two-level source of incompleteness. The first-level source of incompleteness arises from the fact that for each observation we do not know its component membership; to govern this source, we use an indicator vector $z_i = (z_{i1}, \ldots, z_{ik})$, where $z_{ij} = 1$ if $x_i$ comes from component $j$ and $z_{ij} = 0$ otherwise. This source of incompleteness is common in the use of mixture models. The second-level source of incompleteness arises from the fact that, within the $j$th cluster, in the space spanned by the principal components $\Gamma_j^T (X - \mu_j)$ of that cluster, we do not know if the $h$th variate of the $i$th transformed observation $\Gamma_j^T (x_i - \mu_j)$ is good or bad, $i = 1, \ldots, n$ and $h = 1, \ldots, d$; this source of incompleteness is governed by a $n \times d \times k$ indicator array with elements $v_{ihj}$, $i = 1, \ldots, n$, $h = 1, \ldots, d$, and $j = 1, \ldots, k$, where $v_{ihj} = 1$ if $\left[\Gamma_j^T (x_i - \mu_j)\right]_h$ is good and $v_{ihj} = 0$ otherwise. The values of $z_{ij}$ and $v_{ihj}$ are used for the definition of the following two-level complete-data likelihood

$$L_c(\Psi) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left\{ \pi_j \prod_{h=1}^{d} [\alpha_{ihj} f_N(\left[\Gamma_j^T (x_i - \mu_j)\right]_h; 0, \lambda_{ihj})]^{v_{ihj}} \times \left[ (1 - \alpha_{ihj}) f_N(\left[\Gamma_j^T (x_i - \mu_j)\right]_h; 0, \eta_{ihj} \lambda_{ihj}) \right]^{1-v_{ihj}} \right\}^{z_{ij}}. \quad (3.2)$$

The corresponding two-level complete-data log-likelihood can be so written as

$$l_c(\Psi) = l_{c,1}(\pi) + \sum_{j=1}^{k} \left[ l_{c,2j}(\alpha_j) + l_{c,3j}(\psi_j) \right], \quad (3.3)$$
where \( \pi = (\pi_1, \ldots, \pi_k) \) and

\[
I_{c,1}(\pi) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \ln \pi_j, \tag{3.4}
\]

\[
I_{c,2j}(\alpha_j) = \sum_{i=1}^{n} \sum_{h=1}^{d} z_{ih} \left[ v_{ihj} \ln \alpha_{hj} + (1 - v_{ihj}) \ln \left(1 - \alpha_{hj}\right) \right], \tag{3.5}
\]

\[
I_{c,3j}(\psi_j) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{h=1}^{d} z_{ih} \left\{ \ln \lambda_{hj} + (1 - v_{ihj}) \ln \eta_{hj} + \left( \frac{1 - v_{ihj}}{\eta_{hj}} \right) \frac{\left( \Gamma_j^T (x_i - \mu_j) \right)_h^2}{\lambda_{hj}} \right\}, \tag{3.6}
\]

with \( \psi_j = \{ \mu_j, \Gamma_j, \Lambda_j, \eta_j \}, j = 1, \ldots, k \). Equation (3.6) only shows the terms that depend on \( \psi_j \).

The algorithm iterates between two steps, one E-step and one M-step, until convergence. These two steps, for the generic \((r + 1)\)th iteration, \(r = 0, 1, \ldots\), are detailed below.

### 3.2.1 E-step

The E-step requires the calculation of

\[
E_{\Psi^{(r)}}(Z_{ij} | x_i) = \frac{\pi_j^{(r)} f_{M^{(r)}}(x_i; \mu_j^{(r)}, \Gamma_j^{(r)}, \Lambda_j^{(r)}, \alpha_j^{(r)}, \eta_j^{(r)})}{f_{M^{(r)}}(x_i; \Psi^{(r)})} := z_{ij}^{(r)},
\]

which is the posterior probability that \( x_i \) belongs to the \( j \)th component of the mixture, using the current fit \( \Psi^{(r)} \) for \( \Psi \), and

\[
E_{\vartheta^{(r)}}(V_{ihj} | x_i, Z_{ij} = 1) = \frac{\alpha_{hj}^{(r)} f_{N}(\left[ \Gamma_j^{(r)^T} (x_i - \mu_j^{(r)}) \right]_h^T ; 0, \lambda_{hj}^{(r)})}{f_{N}\left( \left[ \Gamma_j^{(r)^T} (x_i - \mu_j^{(r)}) \right]_h^T ; 0, \lambda_{hj}^{(r)^T}, \alpha_{hj}^{(r)}, \eta_{hj}^{(r)} \right)} := v_{ijh}^{(r)}, \tag{3.7}
\]

which is the posterior probability that \( \left[ \Gamma_j^{(r)^T} (x_i - \mu_j^{(r)}) \right]_h \) is good using the current fit \( \vartheta^{(r)} \) for \( \vartheta_j = \{ \mu_j, \Gamma_j, \Lambda_j, \alpha_j, \eta_j \}, i = 1, \ldots, n, h = 1, \ldots, d \) and \( j = 1, \ldots, k \). Then, by substituting \( z_{ij} \) with \( z_{ij}^{(r)} \) and \( v_{ijh} \) with \( v_{ijh}^{(r)} \) in (3.4)–(3.6), we obtain the functions to be maximized in the M-step, at the \((r + 1)\)th iteration, to obtain the updates for the parameters of the model.
3.2.2 M-step
The M-step requires the maximization, with respect to \( \Psi \), of the expected counterpart of \( l_c(\Psi) \) in (3.3). The maximization yields the following closed-form updates of the parameters

\[
\pi_j^{(r+1)} = \frac{n_j^{(r)}}{n},
\]

\[
\alpha_{bj}^{(r+1)} = \frac{1}{n_j^{(r)}} \sum_{i=1}^{n} z_{ij}^{(r)} v_{ibj}^{(r)},
\]

\[
\mu_{bj}^{(r+1)} = \sum_{i=1}^{n} z_{ij}^{(r)} \left[ \Gamma_j W_{\tilde{\eta}}^{(r)} \Gamma_j^{\top} x_i \right]_{b}
\]

\[
\lambda_{bj}^{(r+1)} = \frac{1}{n_j^{(r)}} \sum_{i=1}^{n} z_{ij}^{(r)} \left( v_{ibj}^{(r)} + \frac{1 - v_{ibj}^{(r)}}{n_j^{(r)}} \right) \left[ (\Gamma_j^{(r)})^{\top} (x_i - \mu_j^{(r+1)}) \right]_b^2,
\]

\[
\eta_{bj}^{(r+1)} = \max \left\{ \eta^*, \sum_{i=1}^{n} \frac{z_{ij}^{(r)} (1 - v_{ibj}^{(r)})}{\lambda_{bj}^{(r+1)}} \left[ (\Gamma_j^{(r)})^{\top} (x_i - \mu_j^{(r+1)}) \right]_b^2 \right\},
\]

where \( n_j^{(r)} = \sum_{i=1}^{n} z_{ij}^{(r)} \),

\[
W_{\tilde{\eta}}^{(r)} = \text{diag} \left\{ \left( v_{i1j}^{(r)} + \frac{1 - v_{i1j}^{(r)}}{\eta_{1j}^{(r)}} \right), \ldots, \left( v_{idj}^{(r)} + \frac{1 - v_{idj}^{(r)}}{\eta_{dj}^{(r)}} \right) \right\},
\]

and \( \eta^* \) is a number close to 1 from the right, \( j = 1, \ldots, k \) and \( b = 1, \ldots, d \); for the analyses in Section 5, we use \( \eta^* = 1.001 \).
The update of $\Gamma_j, j = 1, \ldots, k$, is obtained as

$$\Gamma_j^{(r+1)} = \arg \min_{\Gamma_j} \sum_{i=1}^{n} \text{trace} \left[ \Gamma_j \left( W_{ij}^{(r+1)} \Lambda_j^{(r+1)} \right)^{-1} \left( x_i - \mu_j^{(r+1)} \right) \left( x_i - \mu_j^{(r+1)} \right)^\top \right],$$

where

$$W_{ij}^{(r+1)} = \text{diag} \left\{ \left( v_{i1j}^{(r)} + \frac{1}{\eta_{i1j}^{(r+1)}} \right)^{-1}, \ldots, \left( v_{idj}^{(r)} + \frac{1}{\eta_{idj}^{(r+1)}} \right)^{-1} \right\}.$$

We consider the PLR decomposition for orthogonal matrices, proposed in Bagnato and Punzo (2019), in order to make the minimization unconstrained. Operationally, the minimization is done via the `optim()` function for $\mathbb{R}$. The BFGS (Broyden, Fletcher, Goldfarb and Shanno) algorithm, passed to `optim()` via the argument `method`, is used for minimization.

### 4 Further computational and operational aspects

#### 4.1 Initialization

As is well-documented in literature, the starting values impact the results of the EM algorithm; therefore, their choice constitutes a very important issue (see, e.g., Biernacki et al., 2003; Karlis and Xekalaki, 2003; Bagnato and Punzo, 2013). We decided to start our EM algorithm by the M-step. This implies the need of initial quantities $z_{ij}^{(0)}, v_{ihj}^{(0)}, \Gamma_j^{(0)}$ and $\eta_{ij}^{(0)}$. For $z_{ij}^{(0)}$, we tried several options: $k$-means, $k$-medoids and MNM. The best one, used in the data analyses of Section 5, was the partition arising from a preliminary run of the $k$-medoids method (Kaufman and Rousseeuw, 1990). Finally, we fix $v_{ihj}^{(0)} = 0.999$, $\eta_{ij}^{(0)} = 1.001$ and define $\Gamma_j^{(0)}$ as the eigenvectors matrix of the $j$th cluster covariance matrix.

#### 4.2 Convergence criterion

A stopping criterion based on Aitken’s acceleration (Aitken, 1926) is used to determine convergence of the EM algorithm illustrated in Section 3.2. The commonly used stopping rules can yield convergence earlier than the Aitken stopping criterion, resulting in estimates that might not be close to the ML estimates. The Aitken acceleration at iteration $r$ is

$$a^{(r)} = \frac{l_{\text{new}}^{(r)} - l^{(r)}}{l^{(r)} - l^{(r-1)}}.$$
where \( l^{(r)} \) is the (observed-data) log-likelihood value from iteration \( r \). An asymptotic (with respect to the iteration number) estimate of the log likelihood at iteration \( r + 1 \) can be computed via

\[
l^{new}_A = l^{(r)} + \frac{1}{1 - d^{(r)}} \left( l^{new} - l^{(r)} \right);
\]

cf. Böhning et al. (1994). Convergence is assumed to have been reached when \( l^{new}_A - l^{(r)} < \epsilon \), provided that this difference is positive (cf. Lindsay, 1995; McNicholas et al., 2010; Subedi et al., 2013, 2015). We use \( \epsilon = 0.001 \) in the analyses herein and set the maximum number of iterations to 200.

### 4.3 Some notes on directional robustness

The MSCNM model provides improved directional estimates (robust directional estimates) of \( \mu_j \) and \( \Lambda_j \), \( j = 1, \ldots, k \), in the presence of mild outliers. This is made possible because in the \( j \)th cluster, the influence of the \( h \) dimension of the \( i \)th transformed observation, i.e. \( \left[ \Gamma_j^T x_i \right]_h \) is reduced (down-weighted) as the squared Mahalanobis distance \( \delta_{ihj} = \left[ \Gamma_j^T (x_i - \mu_j) \right]_h^2 / \lambda_{bhj} \) increases. This is the underlying idea of \( M \) estimation (Maronna, 1976), which uses a decreasing weighting function \( w(\delta_{ihj}) : (0, \infty) \rightarrow (0, \infty) \) to down-weight the transformed observations \( \left[ \Gamma_j^T x_i \right]_h \) with large \( \delta_{ihj} \) values. To be more precise, according to (3.9), \( \mu_{bhj}^{(r+1)} \) can be viewed, because \( \alpha_{bhj} \) and \( \eta_{bhj} \) are estimated from the data by ML, as an adaptively weighted sample mean, in the sense used by Hogg (1974), with weights

\[
\nu_{ihj}^{(r)} + \frac{1 - \nu_{ihj}^{(r)}}{\eta_{ihj}^{(r)}}.
\]  

(4.1)

Analogous considerations can be applied to the update \( \lambda_{bhj}^{(r+1)} \), given in (3.10), for \( \lambda_{bhj} \). This approach, in addition to be a type of \( M \) estimation, in each cluster and each principal component follows Box (1980) and Box and Tiao (2011) in embedding the normal model in a larger model with one or more parameters (here \( \alpha_{bhj} \) and \( \eta_{bhj} \)) that afford protection against non-normality. For a discussion on down-weighting for the contaminated normal distribution, see also Little (1988), Punzo and McNicholas (2016) and Punzo et al. (2019).

Below, we make explicit the formulation of our weighting function and demonstrate its decreasing behaviour with respect to \( \delta_{ihj} \). If we substitute \( \nu_{ihj}^{(r)} \) in (4.1) with its explicit formulation given in (3.7), avoid the use of the iteration superscript and use the simplified notation \( \delta_{ihj} \) for the squared Mahalanobis distance, then the
The weighting function of our approach results in:

\[ w(\delta_{ihj}; \alpha_{hj}, \eta_{hj}) = 1 + \frac{(1 - \alpha_{hj})(\eta_{hj} - 1)e^{\frac{\delta_{ihj}}{2}}}{(\alpha_{hj} - 1)\eta_{hj}e^{\frac{\delta_{ihj}}{2}} - \alpha_{hj}\sqrt{\eta_{hj}^3}e^{\frac{3\delta_{ihj}}{2}}}. \]  \hspace{1cm} (4.2)

The first order derivative of \( w(\delta_{ihj}; \alpha_{hj}, \eta_{hj}) \) is:

\[ w'(\delta_{ihj}; \alpha_{hj}, \eta_{hj}) = -\frac{\alpha_{hj}(1 - \alpha_{hj})(\eta_{hj} - 1)^2 e^{\frac{\delta_{ihj}(\eta_{hj} + 1)}{2\eta_{hj}}}}{2\sqrt{\eta_{hj}^3}\left[(\alpha_{hj} - 1)e^{\frac{\delta_{ihj}}{2}} - \alpha_{hj}\sqrt{\eta_{hj}^3}e^{\frac{3\delta_{ihj}}{2}}\right]^2}. \] \hspace{1cm} (4.3)

Due to the constraints \( \alpha_{hj} \in (0, 1) \) and \( \eta_{hj} > 0 \), it is straightforward to realize that \( w'(\delta_{ihj}; \alpha_{hj}, \eta_{hj}) \) is always negative, and this implies that \( w(\delta_{ihj}; \alpha_{hj}, \eta_{hj}) \) is a decreasing function of \( \delta_{ihj} \). For further details about down-weighting with mixture models based on the contaminated normal distribution, see Punzo and McNicholas (2016) and Mazza and Punzo (2017).

### 4.4 Constraints for directional detection of bad points

When our MSCNM is used for the directional detection of bad points, \( (1 - \alpha_{hj}) \) and \( \eta_{hj} \) represent the proportion of bad points and the degree of contamination, respectively, in the \( h \)th principal component and \( j \)th group. Then, for the former parameter, one could require that the proportion of good data in the \( h \)th principal component is at least equal to a predetermined value, say \( \alpha^*_h \). Note that it does not make sense for \( \alpha^*_h \) to be cluster dependent due to the well-known label switching issue (Stephens, 2000), but it makes sense to be principal component dependent if, for example, we order the eigenvalues from largest to smallest. Under the considered constraint, it is easy to show (Punzo et al., 2018b) that the update for \( \alpha_{hj} \) in (3.8) becomes:

\[ \alpha_{hj}^{(r+1)} = \max\left\{ \alpha^*_h, \frac{1}{\eta_{hj}}\sum_{i=1}^{n} z_{ij}^{(r)} v_{ij}^{(r)} \right\}. \]

In the data analyses of Section 5, we use this approach to update \( \alpha_{hj} \), and we take \( \alpha^*_h = 0.5, h = 1, \ldots, d \). The value 0.5 is justified because in robust statistics, it is usually assumed that at least half of the points are good (Punzo and McNicholas, 2016).
4.5 Automatic directional detection of outliers

Here, we illustrate how the automatic directional detection of bad points works for the MSCNM model introduced in Section 3.1. In detail, the classification of a generic observation \( x_i \), according to model (3.1), means to determine its cluster membership and establish whether its generic \( h \)th principal component, \( h = 1, \ldots, d \), is good or bad in that cluster. Let \( \tilde{z}_{ij} \) and \( \tilde{v}_{ihj} \) be the values of \( z_{ij}^{(r)} \) and \( v_{ihj}^{(r)} \), respectively, at convergence of the EM algorithm. To evaluate the cluster membership of \( x_i \), we use, as is typical in model-based clustering applications, the MAP classification, that is,

\[
\text{MAP}(\tilde{z}_{ij}) = \begin{cases} 
1 & \text{if } \max_g \{\tilde{z}_{ig}\} \text{ occurs in cluster } j, \\
0 & \text{otherwise.}
\end{cases}
\]

We then consider \( \tilde{v}_{ihg} \), where \( g \) is selected such that \( \text{MAP}(\tilde{z}_{ig}) = 1 \) and \( x_i \) is considered good with respect to the \( h \)th principal component if \( \tilde{v}_{ihg} > 0.5 \) and bad otherwise, \( i = 1, \ldots, n \) and \( h = 1, \ldots, d \). This is somehow related, in the space spanned by the principal components, to the concept of snipping, complementing that of trimming, introduced in robust cluster analysis by Farcomeni (2014b) and studied in model-based clustering by Farcomeni (2014a). Roughly speaking, an observation is snipped when some of its dimensions are discarded but the remaining are used for clustering and estimation. For further details about snipping refer to Farcomeni and Greco (2016, Ch. 8 and 9).

5 Data analyses

In this section, we evaluate the performance of the proposed MSCNM on synthetic and real data. Particular attention is devoted to the problem of detecting bad points. We also compare the performance of the proposed method with some (unconstrained) finite mixtures of well-established multivariate symmetric distributions. In detail, we compare MNMs, MtMs (Peel and McLachlan, 2000), MCNMs (Punzo and McNicholas, 2016) and MS\( t \)Ms (Forbes and Wraith, 2014). Apart from MNMs, all the other mixture components of the models above have one (in the case of MtMs) or more (in the case of MCNMs and MS\( t \)Ms) additional parameters governing the tail weight. We also compare the proposed method with trimmed mixture models (Garcia-Escudero et al., 2008). We picked two different values of the trimming parameter (i.e., the proportion of data to be trimmed): 0.05 (TRI05) and 0.02 (TRI02).

The whole analysis is conducted in R version 3.6.0 64-bit, using a MacOS Mojave 10.14.5 PC, with 4.2 GHz Intel Core i7 CPU and 16 GB 2400 MHz DDR RAM. All the fitting algorithms are EM or EM variants. We use the \texttt{gpcm()} function of the \texttt{mixture} package (Browne et al., 2018), with the option \texttt{mnames = "VVV"}, to fit the MNMs, the \texttt{teigen()} function of the \texttt{teigen} package (Andrews et al.,...
2018), specifying the argument models = "UUUU", to fit the MtMs, the CNmixt() function of the ContaminatedMixt package (Punzo et al., 2018b), with the option model = "VVV", to fit the MCNMs, the tclust() function of the tclust package (Fritz et al., 2012) to fit TRI05 and TRI02, while we have implemented a specific R code, available at http://www.statmod.org/smiJ/archive.html to fit MSTMs and MSCNMs. For a fair comparison, the updates of \( \Gamma_j \), \( j = 1, \ldots, k \), for the MSTM are not computed with the approach discussed in Forbes and Wraith (2014) but with arguments analogous to the those discussed in Section 3.2.2. To allow for a direct comparison of the competing models, all of these algorithms are initialized by providing the initial quantities \( z_i^{(0)} \), \( i = 1, \ldots, n \), using the partition provided by a preliminary run of the \( k \)-medoids method, as implemented by the pam() function of the cluster package (Maechler et al., 2018). For the competing mixture models based on the \( t \) distribution, the degrees of freedom are initialized to 20. The stopping criterion is based on Aitken’s acceleration (refer to Section 4.2).

We use the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978) for model selection, under a ‘greater is better’ formulation. To compare the classification results, when the true partition is available, we use the error rate (ER) and the adjusted Rand index (ARI; Hubert and Arabie, 1985). The ARI corrects the Rand index (Rand, 1971) for chance; its expected value under random classification is 0, and it takes a value of 1 when there is perfect class agreement.

The comparison is also based on the ability of the models to detect outliers. In this regard, the MCNM can be used to detect outliers using an analogous procedure like the one described in Section 4.5; see Punzo and McNicholas (2016, Section 5.6) for details. An \textit{a posteriori} procedure (i.e., a procedure taking place once the model is fitted) to detect bad points with MtMs is illustrated by McLachlan and Peel (2000, p. 232): An observation \( x_i \) is treated as a bad point in the \( j \)th cluster if

\[
\sum_{j=1}^{k} \text{MAP} \left( \bar{z}_j \right) \delta \left( x_i, \bar{\mu}_j; \bar{\Sigma}_j \right)
\]

is sufficiently large, where \( \delta(x_i, \bar{\mu}_j; \bar{\Sigma}_j) \) is the squared Mahalanobis distance between \( x_i \) and \( \bar{\mu}_j \) with covariance matrix \( \bar{\Sigma}_j \), \( i = 1, \ldots, n \) and \( j = 1, \ldots, k \). To decide how large the statistic (5.1) must be in order for \( x_i \) to be classified as a bad point, McLachlan and Peel (2000, p. 232) compare it to the 95th percentile of the chi-squared distribution with \( d \) degrees of freedom, where the chi-squared result is used to approximate the distribution of \( \delta(X_i, \bar{\mu}_j; \bar{\Sigma}_j) \). This procedure can be easily extended to the MSTM to define a strategy for the directional detection of bad points by considering the statistic

\[
\sum_{j=1}^{k} \text{MAP} \left( \bar{z}_j \right) \delta \left( \bar{\Gamma}_j^T \left( x_i - \bar{\mu}_j \right), 0; \bar{\lambda}_{ij} \right).
\]
Table 2 Synthetic data: AIC, BIC and number of free parameters (# Par) for the competing (untrimmed) mixture models with \( k = 2 \) components

|      | MNM     | MtM     | MSTm    | MCNM    | MSCNM   |
|------|---------|---------|---------|---------|---------|
| AIC  | -4 551.956 | -4 338.507 | -4 355.454 | -4 303.575 | -4 330.291 |
| BIC  | -4 600.323 | -4 395.667 | -4 421.408 | -4 369.529 | -4 413.833 |
| # Par| 11      | 13      | 15      | 15      | 19      |

The quantity in (5.2) can be compared to the 95th percentile of the chi-squared distribution with one degree of freedom in order to classify \( \left[ \hat{\theta}_j \right] \) as good or bad in the \( j \)th cluster, \( i = 1, \ldots, n \), \( h = 1, \ldots, d \), and \( j = 1, \ldots, k \).

5.1 Synthetic data

The synthetic data analysis considers \( n = 600 \) observations, subdivided in \( k = 2 \) groups of sizes \( n_1 = 180 \) and \( n_2 = 420 \), randomly generated by bivariate \((d = 2)\) normal distributions with parameters

\[
\mu_1 = (0, 0)^\top, \quad \mu_2 = (2, 4)^\top, \quad \Sigma_1 = \begin{pmatrix} 2 & 1.5 \\ 1.5 & 2 \end{pmatrix}, \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}.
\]

Moreover, 18 observations have been transformed in outliers in the following three ways:

1. by substituting the first dimension \((X_1)\) of 12 randomly selected points of the first cluster with values randomly generated from a uniform distribution with support \((-8, -5)\);
2. by substituting the second dimension \((X_2)\) of 3 randomly selected points of the second cluster with values randomly generated from a uniform distribution with support \((7, 8)\) and
3. by substituting both dimensions \((X_1, X_2)\) of 3 randomly selected points of the first cluster with values randomly generated from a uniform distribution with support \((2, 4) \times (-8, -6)\).

Figure 2 shows the scatter plot of the generated data, with labels and colours representing the different clusters and with bullets denoting the outliers. As we can note, the clusters are separated sufficiently, and the outliers fall outside them. Thus, we would expect the competing robust mixtures, directly fitted with \( k = 2 \) components, to be able to easily recognize the underlying clusters and to detect the outliers.

Table 2 presents a comparison between the ‘untrimmed’ models in terms of AIC, BIC and number of estimated parameters (# Par). As we can see from Table 2, AIC and BIC indicate the MCNM as the best one. The MSCNM model is the second best for the AIC and the third one for the BIC.
Statistical Modelling

and this must be taken into account when evaluating the obtained classification approaches always in terms of classification. For a fair comparison, we compute by the MCNM. Table 5 shows a comparison between trimmed and untrimmed MSCNM gives the best performance with an ARI of 0.874 and ER of 0.032, followed

5.1.1 Classification

update the orthogonal matrices require more time to be fitted, and this is due to the procedure used in the M-step to

Table 3 reports the elapsed time (in seconds) required by the algorithms to fit all the models. This time is computed via the proc.time() function of the base package. As we can note, the mixture models based on multiple scaled distributions require more time to be fitted, and this is due to the procedure used in the M-step to update the orthogonal matrices $\Gamma_1$ and $\Gamma_2$ in the two latent groups (cf. Section 3.2.2).

Table 3 Synthetic data: Elapsed time in seconds for the competing mixture models with $k = 2$ components

| Model   | MrM  | MSrM | TRI05 | TRI02 | MCNM | MSCNM |
|---------|------|------|-------|-------|------|-------|
| MNM     | 0.011| 0.033| 78.180| 0.056 | 0.041| 0.016 |
| MrM     |      |      |       |       |      | 74.683|

Table 3 reports the elapsed time (in seconds) required by the algorithms to fit all the models. This time is computed via the proc.time() function of the base package. As we can note, the mixture models based on multiple scaled distributions require more time to be fitted, and this is due to the procedure used in the M-step to update the orthogonal matrices $\Gamma_1$ and $\Gamma_2$ in the two latent groups (cf. Section 3.2.2).

5.1.1 Classification

Table 4 shows the obtained ER and ARI values for the untrimmed methods. The MSCNM gives the best performance with an ARI of 0.874 and ER of 0.032, followed by the MCNM. Table 5 shows a comparison between trimmed and untrimmed approaches always in terms of classification. For a fair comparison, we compute the ARI and ER only using units that are not trimmed by any of the trimmed methods involved; this data selection technique naturally benefits trimming methods, and this must be taken into account when evaluating the obtained classification...
Multiple scaled contaminated normal distribution

Table 4  Synthetic data: ER and ARI values for the competing (untrimmed) mixture models with \( k = 2 \) components

|      | MNM | MrM | MS\(\text{tM} \) | MCNM | MSCNM |
|------|-----|-----|-----------------|------|-------|
| ER   | 0.093 | 0.073 | 0.057          | 0.040 | 0.032 |
| ARI  | 0.659 | 0.725 | 0.783          | 0.843 | 0.874 |

Table 5  Synthetic data: ER and ARI values, for the competing mixture models with \( k = 2 \) components, computed using units that are not trimmed by any of the trimmed procedures involved

|      | MNM | MrM | MS\(\text{tM} \) | TRI05 | TRI02 | MCNM | MSCNM |
|------|-----|-----|-----------------|------|------|------|-------|
| ER   | 0.098 | 0.075 | 0.060          | 0.032 | 0.070 | 0.042 | 0.033 |
| ARI  | 0.643 | 0.718 | 0.773          | 0.875 | 0.736 | 0.836 | 0.868 |

Table 6  Synthetic data: TPP and FPN values for the competing robust methods

|      | MrM | MS\(\text{tM} \) | TRI05 | TRI02 | MCNM | MSCNM |
|------|-----|-----------------|------|------|------|-------|
| TPP(\%) | 100 | 100            | 100  | 67  | 100  | 100  |
| FPN   | 31  | 29             | 12   | 0   | 0    | 0    |

5.1.2  Outlier detection

Table 6 reports the true positive percentage (TPP), that is, the percentage of outliers correctly classified as outliers, and the false positive number (FPN), that is, the number of points incorrectly detected as outliers, related to the outlier detection rule of the competing robust methods. We can note from the TPPs how, apart from TRI02 (having TPP = 67\%), all the competing robust methods are able to detect the perturbed observations (TPP = 100\%). However, by looking at the FPN values, MS\(\text{tM} \), MrM and TRI05 detect many more outliers than there should be. These results confirm the open problem with trimming procedures, that of selecting the trimming proportion. Figure 3 shows classification and outlier detection results for each robust competing method. So, in terms of outlier detection, and according to the adopted measures, the best performers are MCNM and MSCNM. However, the estimates of the contamination parameters for the MSCNM, shown in Table 7, give us the opportunity to appreciate the usefulness of this model. In particular, the estimates indicate a different proportion of outliers and a different degree of contamination

Table 7  Synthetic data: Estimates of the contamination parameters for the MSCNM

| \( \hat{\alpha}_1 \) | \( \hat{\eta}_1 \) | \( \hat{\alpha}_2 \) | \( \hat{\eta}_2 \) |
|---------------------|---------------------|---------------------|---------------------|
| 0.974               | 1.882               | 0.911               | 7.442               |
| 0.937               | 34.004              | 0.999               | 1.001               |
across the principal components. In fact in cluster 2 only one of the two principal components presents outliers, while in cluster 1 both principal components have outliers in different proportions and degrees of contamination.

Statistical Modelling 2021; 21(4): 332–358
Table 8  Wholesale data: AIC, BIC and number of free parameters (# Par) for the competing (untrimmed) mixture models with $k = 2$ components

|       | MNM    | MtM    | MSTM   | MCNM   | MSCNM   |
|-------|--------|--------|--------|--------|---------|
| AIC   | -4 523.908 | -2 865.809 | -3 038.972 | -3 135.962 | -3 119.700 |
| BIC   | -4 511.550 | -3 098.756 | -3 312.786 | -3 377.082 | -3 442.556 |
| # Par | 89     | 91     | 105    | 93     | 121     |

Summarizing, both MSCNM and MCNM can find the true outliers. However, MSCNM performs better in terms of ARI and ER (refer to Table 4). This can be explained by the fact that the dataset has different outliers per dimension and the estimates of the mean vectors and scale matrices for the MSCNM are directionally robust.

5.2 Wholesale data

The real data analysis considers the wholesale dataset, which is freely available on the University of California Irvine (UCI) machine learning repository (https://archive.ics.uci.edu/ml/datasets/wholesale+customers). The dataset originates from a larger database (Abreu, 2011) and contains information about the annual spending, in monetary units, on $d = 6$ products for $n = 440$ customers of a wholesale merchant in Portugal. The product categories are fresh, milk, grocery, frozen, detergents paper (DP) and delicatessen. The dataset also contains two nominal variables: region (Lisboa, Porto or other) and channel (hotel/restaurant/caf´e or retail). There is no distinguishable difference in consumption among the regions, but there is a distinguishable difference between channels. The objective of this analysis is to segment the customers based on their spending and to compare these segments to the channel.

Figure 4 shows the scatter plot matrix of the standardized data, with each colour and symbol representing a different channel. There is a high level of overlap between groups; many outliers are located along the principal axes of these groups, and their number and distance from the bulk of the groups varies across principal axes.

The competing models are fitted with $k = 2$ components. Table 8 shows a model comparison in terms of AIC and BIC. The MSCNM model is not the best one also in this data analysis, with the MtM model working better regardless of the adopted criterion.

Table 9 displays the computational elapsed times (in seconds). The additional computational burden required by the multiple scaled models is clear also in this case.

Table 9  Wholesale data: Elapsed time in seconds for the competing mixture models with $k = 2$ components

|       | MNM     | MtM     | MSTM    | TRI05   | TRI02   | MCNM    | MSCNM   |
|-------|---------|---------|---------|---------|---------|---------|---------|
|       | 0.009   | 0.115   | 614.534 | 0.058   | 0.056   | 0.038   | 631.604 |

Statistical Modelling 2021; 21(4): 332–358
Figure 4 Wholesale data: Scatter plot matrix with classification of the customers based on the channel (× and + denote retail and horeca channels, respectively).

Table 10 Wholesale data: ER and ARI values for the competing mixture models with \( k = 2 \) components

|       | MNM | MrM | MStM | MCNM | MSCNM |
|-------|-----|-----|------|------|-------|
| ER    | 0.311 | 0.220 | 0.207 | 0.248 | 0.177 |
| ARI   | 0.100 | 0.311 | 0.343 | 0.252 | 0.415 |

Table 10 shows ER and ARI for each model. The MSCNM gives the best performance with ER = 0.177 and ARI = 0.415. To compare the results with the trimming techniques, the ER and ARI are calculated, as for Section 5.1, only using units that are not trimmed by any of the trimming procedures involved. Table 11 shows the results. MSCNM gives the best ER (0.167) and ARI (0.441) also in this case.

Some of the estimated parameters from the fitted MSCNM with \( k = 2 \) components can help in the interpretation of the results. Table 12 shows the estimates of \( \mu_j \), while Table 13 reports the estimates of \( \alpha_j \) and \( \eta_j, j = 1, 2 \).
Table 11 Wholesale data: ER and ARI values for the competing mixture models with \( k = 2 \) components computed using units that are not trimmed by any of the procedures involved

|           | MNM | MrM | MS\(\text{tM}\) | TRI05 | TRI02 | MCNM | MSCN\(\text{M}\) |
|-----------|-----|-----|----------------|-------|-------|------|-----------------|
| ER        | 0.301 | 0.225 | 0.208 | 0.335 | 0.342 | 0.251 | 0.167          |
| ARI       | 0.069 | 0.301 | 0.339 | 0.229 | 0.251 | 0.246 | 0.441          |

Table 12 Wholesale data: Estimated mean vectors from the MSCN mixture with \( k = 2 \) components

|       | Fresh | Milk | Grocery | Frozen | DP | Delicatessen |
|-------|-------|------|---------|--------|----|--------------|
| \( \mu_1 \) | -0.035 | -0.561 | -0.613 | -0.214 | -0.545 | -0.273 |
| \( \mu_2 \) | -0.421 | 0.185 | 0.334 | -0.405 | 0.400 | -0.045 |

Table 13 Wholesale data: Estimated contamination parameters from the MSCN mixture with \( k = 2 \) components

|       | \( \alpha_1 \) | \( \eta_1 \) | \( \alpha_2 \) | \( \eta_2 \) |
|-------|-----------------|---------------|-----------------|---------------|
| \( \alpha_1 \) | 0.726 | 68.655 | 0.885 | 26.244 |
| \( \eta_1 \) | 0.810 | 12.670 | 0.853 | 112.044 |
| \( \alpha_2 \) | 0.856 | 13.239 | 0.811 | 23.280 |
| \( \eta_2 \) | 0.626 | 8.488 | 0.839 | 23.830 |

The customers in cluster 1 spend more for fresh and frozen products. The customers in cluster 2 are those spending more for milk, grocery, detergent paper and delicatessen categories. The outliers in this cluster are generally farther away from the centres when compared to cluster 1 (refer to the values of \( \eta_1 \) and \( \eta_2 \)), but in cluster 1, the proportion of outliers in each principal component is often higher (compare \( \alpha_1 \) with \( \alpha_2 \)), with the fifth component being the strongest example of that having good and bad points in the same proportion.

Summarizing, both MSCNM and MCNM can find the true outliers. However, MSCNM performs better in terms of ARI and ER (refer to Table 4). This can be explained by the fact that the dataset has different outliers per dimension and the estimates of the mean vectors and scale matrices for the MSCNM are directionally robust.

6 Conclusions

When compared to the classical MN, the MCN distribution has two additional parameters, \( \alpha \) and \( \eta \), denoting the proportion of good data and the degree of contamination, respectively.

In this article, following the strategy of Forbes and Wraith (2014), we derived the MSCN distribution. In our setting, such a strategy was based on two key elements. The first is the eigen decomposition of the scale matrix \( \Sigma \) of the MCN distribution. The eigenvectors matrix arising from this decomposition defines a new space that we refer to as the space of the principal components. The second element is the introduction of \( d \) Bernoulli random variables indicating whether a
point is good or bad separately for each principal component, being \( d \) the number of dimensions. The MSCN distribution has a closed-form representation of the pdf and depends on \( 2d \) additional parameters, with respect to the MN distribution, representing the proportion of good data and the degree of contamination on each principal component. Advantageously, the MSCN distribution enables directional (with respect to the \( d \) principal components) robust estimation of the mean vector and covariance matrix of the MN distribution and also gives automatic directional detection of bad points.

The MSCN distribution was applied to robust model-based clustering by introducing mixtures of MSCN distributions; an EM algorithm was also described to obtain ML estimates for the mixture parameters. On the real and artificial data analyses of Section 5, we demonstrated the good behaviour of our directional contaminated approach when compared to mixtures of the following distributions: MN, \( Mt \), MCN and multiple MS\( Mt \).

Future work will focus on the following avenues.

- Although AIC and BIC are commonly used in model-based clustering, Tortora et al. (2019) illustrated that they are not well suited when using mixture models based on multiple scaled components; this may be the reason why these criteria never selected a model of this class in the data analyses of Section 5. A development of an _ad hoc_ model selection criterion for this kind of models could be the subject of future research.

- Our mixture model uses symmetric distributions to model each cluster which could be rather limiting under some empirical conditions (Franczak et al., 2014). However, the use of symmetric distributions can be justified considering that a mixture of several basic symmetric distributions can approximate non-symmetric distributions. While this can be very helpful for modelling purposes, it can be misleading when dealing with clustering and classification applications because one cluster may be represented by more than one mixture component simply because it has, in fact, a skewed distribution. To overcome this issue, we could extend our MSCN distribution with the aim of introducing skewness; the resulting model could be used to define the components of a mixture. Examples of competing approaches in this directions are given in Franczak et al. (2015) and Tortora et al. (2019). Equivalently, but from a different perspective, we could extend the multivariate contaminated shifted asymmetric Laplace distribution (Morris et al., 2019), or the multivariate skew-contaminated normal distribution (Cabral et al., 2012), to account for a different contamination on each principal component and then use these models as mixture components; in the univariate case, see also Punzo et al. (2018a).

- In the fashion of McLachlan and Peel (2000), McLachlan et al. (2003) and McNicholas and Murphy (2008), for mixtures of MN distributions, McLachlan et al. (2007) and Andrews and McNicholas (2011) for mixtures of Mt distributions, and Punzo et al. (2020) for mixtures of MCN distributions,
parsimony and dimension reduction could be obtained by exploiting local (eventually common) factor analysers.

**Declaration of conflicting interests**

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

**Funding**

The authors received no financial support for the research, authorship and/or publication of this article.

**References**

Abreu NG (2011) Analise do perfil do cliente Recheio e desenvolvimento de um sistema promocional [Customer profile analysis. Filling and development of a promotional system]. PhD thesis, Mestrado em Marketing, ISCTE-IUL, Lisbon.

Aitken AC (1926) On Bernoullis numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh*, 46, 289–305.

Aitkin M and Wilson GT (1980). Mixture models, outliers, and the EM algorithm. *Technometrics*, 22, 325–31.

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, edited by BN Petrov and F Csaki, pages 267–81. New York, NY: Springer-Verlag.

Alqallaf F, Van Aelst S, Yohai VJ and Zamar RH (2009) Propagation of outliers in multivariate data. *The Annals of Statistics*, 37, 311–31.

Andrews JL and McNicholas PD (2011) Extending mixtures of multivariate t-factor analysers. *Statistics and Computing*, 21, 361–73.

Andrews J, Wickins J, Boers N and McNicholas P (2018) teigen: An R package for model-based clustering and classification via the multivariate t distribution. *Journal of Statistical Software*, 83, 1–32.

Bagnato L and Punzo A (2013) Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm. *Computational Statistics*, 28, 1571–97.

——— (2019) Unconstrained representation of orthogonal matrices with application to common principle components. ArXiv.org e-print 1906.00587. URL http://arxiv.org/abs/1906.00587.

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

Berkane M and Bentler PM (1988) Estimation of contamination parameters and identification of outliers in multivariate data. *Sociological Methods & Research*, 17, 55–64.

Biernacki C,Celeux G and Govaert G (2003) Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41, 561–75.

Bock HH (2002) Clustering methods: From classical models to new approaches. *Statistics in Transition*, 5, 725–58.
Böhning D (2000) Computer-assisted Analysis of Mixtures and Applications: Meta-analysis, Disease Mapping and Others (Monographs on Statistics and Applied Probability 81). London: Chapman & Hall/CRC.

Böhning D, Dietz E, Schaub R, Schlattmann P and Lindsay B (1994) The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46, 373–88.

Box GEP (1980) Sampling and Bayes inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 143, 383–430.

Box GEP and Tiao GC (2011) *Bayesian Inference in Statistical Analysis*. New York: Wiley Classics Library.

Browne RP, ElSherbiny A and McNicholas PD (2018) mixture: Finite Gaussian Mixture Models for Clustering and Classification. R package Version 1.5. URL http://CRAN.R-project.org/package=mixture

Cabral CSB, Lachos VH and Prates MO (2012) Multivariate mixture modelling using skew-normal independent distributions. *Computational Statistics & Data Analysis*, 56, 126–42.

Dang UJ, Browne RP and McNicholas PD (2015) Mixtures of multivariate power exponential distributions. *Biometrics*, 71, 1081–89.

Davies L and Gather U (1993). The identification of multiple outliers. *Journal of the American Statistical Association*, 88, 782–92.

Dempster A, Laird N and Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39, 1–38.

Devin SJ, Gnanadesikan R and Kettenring JR (1981) Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76, 354–62.

Farcomeni A (2014a) Robust constrained clustering in presence of entry-wise outliers. *Technometrics*, 56, 102–11.

——— (2014b) Snipping for robust k-means clustering under component-wise contamination. *Statistics and Computing*, 24, 907–19.

Farcomeni A and Greco L (2016) Robust Methods for Data Reduction. Boca Raton, FL: CRC Press.

Farcomeni A and Punzo A (2019) Robust model-based clustering with mild and gross outliers. *TEST*, URL: https://doi.org/10.1007/s11749-019-00693-z

Forbes F and 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, 971–84.

Fraley C and Raftery AE (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal*, 41, 578–88.

Franczak BC, Browne RP and McNicholas PD (2014) Mixtures of shifted asymmetric Laplace distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36, 1149–57.

Franczak BC, Tortora C, Browne RP and McNicholas PD (2015) Unsupervised learning via mixtures of skewed distributions with hypercube contours. *Pattern Recognition Letters*, 58, 69–76.

Fritz H, Garcia-Escudero LA and Mayo-Iscar A (2012) tclust: An R package for a trimming approach to cluster analysis. *Journal of Statistical Software*, 47, 1–26.

Gallegos MT and Ritter G (2009) Trimmed ML estimation of contaminated mixtures. *Sankhy?: The Indian Journal of Statistics A*, 71, 164–220.

Garcia-Escudero LA, Gordaliza A, Matran C and Mayo-Iscar A (2008) A general trimming approach to robust cluster analysis. *Annals of Statistics*, 36, 1324–45.

Gómez E, Gómez-Villegas MA and Marín JM (2003) A survey on continuous elliptical vector distributions. *Revista Matemática Complutense*, 16, 345–61.

Gómez-Villegas MA, Gómez-Sánchez-Manzano E, Main P and Navarro H (2011). The effect of non-normality in the power exponential distributions. In *Modern Mathematical Tools and Techniques in Capturing Com-*
plexity, Understanding Complex Systems, edited by L Pardo, N Balakrishnan and MA Gil, pages 119–29. Berlin and Heidelberg: Springer-Verlag.

Hogg RV (1974) Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association, 69*, 909–23.

Hubert L and Arabie P (1985) Comparing partitions. *Journal of Classification, 2*, 193–218.

Karlis D and Xekalaki E (2003) Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis, 41*, 577–90.

Kaufman L and Rousseeuw PJ (Eds) (1990) Partitioning around medoids (program PAM). In Finding groups in data: An introduction to cluster analysis, pages 68–125. Hoboken, New Jersey: Wiley.

Kotz S and Nadarajah S (2004) *Multivariate t-Distributions and Their Applications*. Cambridge: Cambridge University Press.

Lange KL, Little RJA and Taylor JMG (1989) Robust statistical modeling using the t distribution. *Journal of the American Statistical Association, 84*, 881–96.

Lindsay B (1995) *Mixture Models: Theory, Geometry and Applications* (NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 5). Hayward, CA: Institute of Mathematical Statistics.

Little RJA (1988) Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics, 37*, 23–8.

Maechler M, Rousseeuw P, Struyf A and Hubert M (2018) ‘Finding groups in data’—Cluster analysis extended Rousseeuw et al. R package Version 2.0.7-1. URL: https://CRAN.R-project.org/package=cluster

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

Maronna RA and Yohai VJ (2014). *Robust Estimation of Multivariate Location and Scatter*. John Wiley & Sons.

Maruotti A and Punzo A (2017) Model-based time-varying clustering of multivariate longitudinal data with covariates and outliers. *Computational Statistics & Data Analysis, 113*, 475–96.

Mazza A and Punzo A (2017) Mixtures of multivariate contaminated normal regression models. *Statistical Papers*. URL: https://doi.org/10.1007/s00362-017-0964-y

McLachlan GJ and Basford KE (1988) *Mixture models: Inference and Applications to clustering*. New York, NY: Marcel Dekker.

McLachlan GJ, Bean RW and Ben-Tovim Jones L (2007) Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. *Computational Statistics & Data Analysis, 51*, 5327–38.

McLachlan GJ and Peel D (2000) *Finite Mixture Models*. New York, NY: John Wiley & Sons.

McLachlan GJ, Peel D and Bean RW (2003) Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis, 41*, 379–88.

McNicholas PD (2016) *Mixture Model-Based Classification*. Boca Raton, FL: Chapman and Hall/CRC Press.

McNicholas PD and Murphy TB (2008) Parsimonious Gaussian mixture models. *Statistics and Computing, 18*, 285–96.

McNicholas PD, Murphy TB, McDaid AF and Frost D (2010) Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics & Data Analysis, 54*, 711–23.

Morris, K., Punzo, A., McNicholas, P. D., and Browne, R. P. (2019). Asymmetric clusters and outliers: Mixtures of multivariate contaminated shifted asymmetric Laplace distributions. *Computational Statistics & Data Analysis, 132*, 145–66.

Peel D and McLachlan GJ (2000) Robust mixture modelling using the t distribution. *Statistics and Computing, 10*, 339–48.

Punzo A, Blostein M and McNicholas PD (2020) High-dimensional unsupervised classification via parsimonious contaminated mixtures. *Pattern Recognition, 98*: 107031.
Punzo A, Ingrassia S and Maruotti A (2019) Multivariate hidden Markov regression models: Random covariates and heavy-tailed distributions. Statistical Papers, to appear. URL: https://doi.org/10.1007/s00362-019-01146-3.

Punzo A and Maruotti A (2016) Clustering multivariate longitudinal observations: The contaminated Gaussian hidden Markov model. Journal of Computational and Graphical Statistics, 25, 1097–116.

Punzo A, Mazza A and Maruotti A (2018a) Fitting insurance and economic data with outliers: A flexible approach based on finite mixtures of contaminated gamma distributions. Journal of Applied Statistics, 45, 2563–84.

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

Punzo A and McNicholas PD (2016) Parsimonious mixtures of multivariate contaminated normal distributions. Biometrical Journal, 58, 1506–37.

——— (2017) Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. Journal of Classification, 34, 249–93.

Rand WM (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66, 846–50.

Ritter G (2015) Robust Cluster Analysis and Variable Selection (Chapman & Hall/CRC Monographs on Statistics & Applied Probability Volume 137). Boca Raton, FL: CRC Press.

Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics, 6, 461–64.

Stephens M (2000) Dealing with label switching in mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62, 795–809.

Subedi S, Punzo A, Ingrassia S and McNicholas PD (2013) Clustering and classification via cluster-weighted factor analyzers. Advances in Data Analysis and Classification, 7, 5–40.

——— (2015) Cluster-weighted t-factor analyzers for robust model-based clustering and dimension reduction. Statistical Methods & Applications, 24, 623–49.

Tortora C, Franczak B, Browne R and McNicholas P (2019). A mixture of coalesced generalized hyperbolic distributions. Journal of Classification, 36, 26–57.

Tukey JW (1960) A survey of sampling from contaminated distributions. In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling (Stanford Studies in Mathematics and Statistics), edited by I Olkin. Chapter 39, pages 448–85. Stanford, CA: Stanford University Press.

Zhang J and Liang F (2010). Robust clustering using exponential power mixtures. Biometrics, 66, 1078–86.