Robust model-based clustering with mild and gross outliers

Alessio Farcomeni · Antonio Punzo

Abstract We propose a model-based clustering procedure where each component can take into account cluster-specific mild outliers through a flexible distributional assumption, and a proportion of observations is additionally trimmed. We propose a penalized likelihood approach for estimation and selection of the proportions of mild and gross outliers. A theoretically grounded penalty parameter is then obtained. Simulation studies illustrate the advantages of our procedure over flexible mixtures without trimming, and over trimmed normal mixture models (tclust). We conclude with an original real data example on the identification of the source from illicit drug shipments seized in Italy and Spain. The methodology proposed in this paper has been implemented in R functions which can be downloaded from https://github.com/afarcome/cntclust.

Keywords tclust · contaminated normal · penalized likelihood

1 Introduction

Resistance to contamination is a desirable property of statistical methods. This is in our opinion particularly true for model-based clustering methods, where just few outliers can disrupt the true underlying grouping structure and lead to severely biased estimates. A review of general ideas in robust model-based
clustering can be found in García-Escudero et al. (2010), Ritter (2015), and Farcomeni and Greco (2015). To be more precise, outliers may roughly be divided into two types (cf. Ritter, 2015, pp. 79–80): mild or gross. Mild outliers are points that deviate from the distribution within a cluster but they would fit well if the overall within-cluster distribution had heavy(er) tails or some (more) skewness. On the other hand, gross outliers are points far away from any of the components. Relatedly, there are two approaches to robust model-based clustering, which have been somehow separated so far. On the one hand, the reference normal mixture model has been made more flexible by relaxing the normality assumption for one, more, or even all components. The most commonly used solution, obtained by relaxing the normality assumption for one component, is the use of a uniform component with support in the convex hull of the data (Banfield and Raftery, 1993), to capture background noise. Approaches relaxing the normality assumption for all components, preserving elliptical contours of clusters, are mixtures of heavy-tailed distributions. Examples in this direction are mixtures of $t$ distributions (Peel and McLachlan, 2000; Andrews et al., 2018), mixtures of power exponential distributions (Zhang and Liang, 2010 and Dang et al., 2015), mixtures of leptokurtic-normal distributions (Bagnato et al., 2017), and mixtures of contaminated normal distributions (Punzo and McNicholas, 2016). These methods are well suited to work with mild outliers, and are sometimes labeled as weakly robust. In some cases they can handle even highly atypical observations, for instance mixtures of $t$ distributions can handle very large or small values thanks to the polynomial tails. All of these approaches give their explicit definition of mild outliers at the modeling stage (e.g., points in the tails of leptokurtic-normal distribution, points in the tails of a $t$ distribution, etc). In our work we explicitly define mild outliers as points in the tails of a contaminated normal, see (1).

A separate body of literature has instead worked with outliers in more general position, including gross outliers, and has usually proceeded by downweighting or completely discarding a proportion of the observations. A good example is tclust (García-Escudero et al., 2008), where a fixed proportion of observations is trimmed (i.e., receives zero weight) and the rest is assumed to follow a normal mixture model; but several other approaches exist including Gallegos and Ritter (2005) and Coretto and Hennig (2016); see also Dotto and Farcomeni (2019). As opposed to weakly robust procedures, the parametric distribution of gross outliers is seldom explicited and the procedures are designed to deal with contamination in general position, where each outlying observation is generated from a unit-specific unspecified density, whose only requirement is to generate points that do not live in a lower dimensional space. These procedures have often formal robustness properties, e.g., positive breakdown point asymptotically. In this work we propose a trimmed mixture of contaminated normal (CN) distributions. A CN distribution is a mixture of two normal distributions, where class-specific mild outliers arise from a normal component with the same centroid but inflated covariance matrix. The resulting CN mixture is estimated after trimming a fixed proportion of gross outliers. We essentially obtain a mixture of normal mixtures (e.g., Li, 2004 and
Di Zio et al., 2007), which is robustly estimated. Our model can be seen from two different perspectives. On the one hand, clusters having a distribution with slightly heavy tails might be desired in order to assign as many observations to clusters as possible. On the other hand, the trade off between mild and gross outliers is exploited in order to increase efficiency: some (mild) outliers are assigned to a cluster and contribute to centroid estimation, therefore decreasing the final mean squared error (MSE). In this work we tackle also an additional open problem with trimming procedures, that of selecting the trimming proportion. This tuning parameter is often selected using heuristic approaches, including ctlcurves (García-Escudero et al., 2011), G-statistic (Farcomeni, 2009 and Farcomeni and Greco, 2015), and simple descriptive methods comparing different choices. In several cases heuristic approaches work well, but they do not have theoretical properties and they have not been evaluated with respect to the actual probability of selecting a good trimming level; see also Dotto et al. (2018) for an iterative procedure which is tailored to avoid the need for this choice. Our proposal for selecting a trimming level is based on a penalized likelihood approach, where the trimming proportion is in practice substituted by a penalty parameter. The advantage is that we can identify a heuristic but theoretically justified way of choosing an optimal penalty level, and therefore an optimal trimming proportion. This is seen in our simulation study to work very well, leading to estimated trimming proportions that are very close and indeed just slightly larger than the true proportion of outliers. Our fixed-penalty approach in some sense solves the issue of selecting the trimming proportion both for our model and the special case of trimmed normal mixture models (tclust).

The rest of the paper is as follows: in the next section we discuss the general methodology. In Section 3 we introduce the specific algorithm for robust estimation. In Section 4 we discuss the penalized trimmed likelihood approach, and in Section 4.1 we discuss the theoretically optimal choice of the penalty parameter. Theoretical and robustness properties of the proposed procedure are briefly discussed in Section 4.2. In Section 5 we describe and report the results of a brief simulation study and in Section 6 we illustrate with an example on original data about the identification of the source from cocaine and heroin shipments seized in Italy and Spain. Finally, in Section 7 we provide some concluding remarks.

The methodology proposed in this paper has been implemented in R functions which can be downloaded from https://github.com/afarcome/cntclust.

2 Methodology

Let \( x_1, \ldots, x_i, \ldots, x_n \) be a sample of \( n \) observations in \( d \) dimensions. Let \( f_N(\cdot; \mu, \Sigma) \) denote the probability density function (pdf) of a \( d \)-variate normal (N) distribution with mean vector \( \mu \) and covariance matrix \( \Sigma \). Finally, let \( \alpha_0 \geq 0 \) denote a trimming proportion of units which shall not be used to estimate model parameters.
We propose to substitute the $k$ normally distributed clusters of the usual finite mixture model with clusters having a contaminated normal (CN) distribution (Tukey, 1960). The pdf of a $d$-variate CN distribution with mean vector $\mu$, scale matrix $\Sigma$, mixing probability $\alpha \in (0, 1)$, and inflation factor $\eta > 1$, is given by

$$f_{\text{CN}}(x; \mu, \Sigma, \alpha, \eta) = (1 - \alpha) f_N(x; \mu, \Sigma) + \alpha f_N(x; \mu, \eta \Sigma).$$

(1)

Although model (1) is identifiable only if $\alpha \in (0, 1)$, in our implementation we also allow $\alpha = 0$; in such a case, we do not estimate $\eta$, that is, we conventionally fix $f_{\text{CN}}(x; \mu, \Sigma, 0, \eta) = f_N(x; \mu, \Sigma)$.

We use the multivariate CN distribution as a component density since the component with weight $\alpha$ of this two-component normal mixture is well suited for capturing mild outliers due to an inflated covariance matrix $\eta \Sigma$ (Aitkin and Wilson, 1980). In detail, we assume data arise from the contaminated normal spurious outlier model

$$k \prod_{j=1}^k \prod_{i \in R_j} \pi_j f_{\text{CN}}(x_i; \mu_j, \Sigma_j, \alpha_j, \eta_j) \prod_{i \notin R} g_i(x_i),$$

(2)

where $R_j$ denotes the set of observations assigned to the $j$-th cluster, $R = \bigcup R_j$ is the set of non-trimmed observations of cardinality $\lceil (1 - \alpha_0) n \rceil$, and $g_i$ are pdfs generating the outliers in general position. In (2) $\alpha_0$ is a tuning parameter that is pre-specified by the user, and the optimization procedure leads to parameter estimation and assignment of $\lceil (1 - \alpha_0) n \rceil$ observations to the $k$ groups. The cluster weights $\pi_j$ can in principle be fixed as $\pi_j = 1/k$, promoting balance of the clusters in terms of cardinality. Estimation of those (as in our approach) corresponds to adding an ad-hoc entropy regularization term to the classification likelihood. See Bryant (1991) and references therein for further discussion. Note that we are assuming in (2) that the contamination rate by gross outliers is exactly $\alpha_0$. This is a working assumption used to set up inference but it will be relaxed when assessing robustness of the procedure below. Due to the fact that $g_i(\cdot)$ is observation-specific, any maximizer of the classification likelihood leads to $g_i(x_i) = 1$ (Farcomeni, 2014); hence we can remove these contributions and basically work with a profile log-likelihood

$$\ell(\theta) = \sum_{j=1}^k \sum_{i \in R_j} \ell_j(\theta) = \sum_{j=1}^k \sum_{i \in R_j} \left[ \ln \pi_j + \ln f_{\text{CN}}(x_i; \mu_j, \Sigma_j, \alpha_j, \eta_j) \right].$$

(3)

In order to make maximization of (3) a well defined problem, we adopt a scatter constraint which controls the eigenvalue ratio (ER; García-Escudero et al., 2008). Denoting with $\lambda_l(\Sigma_j)$ the $l$th eigenvalue of the $j$th scale matrix $\Sigma_j$, $l = 1, \ldots, d$ and $j = 1, \ldots, k$, the ER constraint is defined as

$$\frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \leq c,$$

(4)
Robust model-based clustering with mild and gross outliers

where $\lambda_{\text{max}} = \max_{j=1,\ldots,k} \max_{l=1,\ldots,d} \lambda_l(\Sigma_j)$ and $\lambda_{\text{min}} = \min_{j=1,\ldots,k} \min_{l=1,\ldots,d} \lambda_l(\Sigma_j)$ are the maximum and minimum eigenvalue, respectively, while $c \geq 1$ is a constant controlling the strength of the constraint (4). A careful selection of $c$ is not usually crucial, but there are ways to calibrate also this parameter (e.g., Cerioli et al., 2018). Note that the constraint guarantees the absence of spurious solutions also with the CN distribution, since $\eta_j > 1$.

3 Robust estimation

Maximization of the trimmed complete-data log-likelihood in (3), under the eigenvalues constraint (4), is not straightforward because of the combinatorial nature of the optimization problem. We adopt a classification expectation-conditional maximization (CECM) algorithm. The proposed algorithm can be seen either as a variant of the expectation-conditional maximization (ECM) algorithm (Meng and Rubin, 1993) where a classification (C) step is added after the expectation (E) step, or as a variant of the CEM algorithm where the maximization (M) step is replaced by two simpler conditional maximization (CM) steps. We have two levels of missing indicators: the cluster assignment (or trimming assignment) of each observation, and for observations assigned to a CN component, an indicator of being in the inflated variance subcluster.

We introduce a $n \times k$ indicator matrix $V$ with elements $v_{ij}$, where $v_{ij} = 1$ if $x_i$ in cluster $j$ arises from the inflated variance component and $v_{ij} = 0$ otherwise.

We can now define a trimmed doubly complete-data log-likelihood

$$
\ell_c(\vartheta) = \ell_{c1}(\pi) + \ell_{c2}(\alpha) + \ell_{c3}(\mu, \Sigma, \eta),
$$

where

$$
\ell_{c1}(\pi) = \sum_{j=1}^{k} \sum_{i \in R_j} \ln \pi_j,
$$

$$
\ell_{c2}(\alpha) = \sum_{j=1}^{k} \sum_{i \in R_j} [v_{ij} \ln \alpha_j + (1 - v_{ij}) \ln (1 - \alpha_j)],
$$

and

$$
\ell_{c3}(\mu, \Sigma, \eta) = -\frac{1}{2} \sum_{j=1}^{k} \sum_{i \in R_j} \left[ d \ln (2\pi) + \ln \det (\Sigma_j) + d v_{ij} \ln \eta_j + \right. \\
\left. + \left( 1 - v_{ij} + \frac{v_{ij}}{\eta_j} \right) \delta (x_i, \mu_j; \Sigma_j) \right].
$$

In (5)–(8), $\delta (x, \mu; \Sigma) = (x - \mu)' \Sigma^{-1} (x - \mu)$ denotes the squared Mahalanobis distance between $x$ and $\mu$, with covariance matrix $\Sigma$, $\pi = (\pi_1, \ldots, \pi_{k-1})'$, $\mu = (\mu'_1, \ldots, \mu'_k)'$, $\Sigma = (\text{vec} (\Sigma_1)' , \ldots , \text{vec} (\Sigma_k)')'$, $\alpha = (\alpha_1, \ldots, \alpha_k)'$, and $\eta = (\eta_1, \ldots, \eta_k)'$, with $\text{vec} (\Sigma)$ being the $d(d + 1)/2$-dimensional vector formed
by stacking the columns of the lower triangular portion of $\Sigma$ (see, e.g., Schott, 2016).

The CECM algorithm iterates four steps: one E-step, one C-step, and two CM-steps; until convergence. The two CM-steps arise from the partition $\vartheta = \{\vartheta_1, \vartheta_2\}$, where $\vartheta_1 = \{\pi, \mu, \Sigma, \alpha\}$ and $\vartheta_2 = \eta$. Conditionally on the trimming proportion $\alpha_0$ and initial value $\vartheta^{(0)}$ for the parameters, the CECM algorithm for the generic $(r + 1)$th iteration can be outlined as follows:

**E-step.** Based on the current parameters $\vartheta^{(r)}$, for each observation $x_i$ and cluster $j$ we compute

$$D_j \left( x_i; \vartheta_j^{(r)} \right) = \pi_j^{(r)} f_{CN} \left( x_i; \mu_j^{(r)}, \Sigma_j^{(r)}, \alpha_j^{(r)}, \eta_j^{(r)} \right),$$

which quantifies the distance between $x_i$ and $\mu_j^{(r)}$, with $\vartheta_j = (\pi_j, \mu_j, \text{vec}(\Sigma_j), \alpha_j, \eta_j)$. Then, we use these distances for the computation of the posterior probabilities

$$D_j \left( x_i; \vartheta_j^{(r)} \right),$$

$$\sum_{h=1}^{k} D_h \left( x_i; \vartheta_h^{(r)} \right).$$

Moreover, we need to calculate $Q(\vartheta)$, the conditional expectation of $\ell_c(\vartheta)$ in (5) given the observed data $x_1, \ldots, x_n$, using $\vartheta^{(r)}$ for $\vartheta$. It can be seen from (7) and (8) that in order to do this, we need to substitute $v_{ij}$ with its expectation

$$v_{ij}^{(r)} = \frac{\alpha_j^{(r)} f_N \left( x_i; \mu_j^{(r)}, \eta_j^{(r)} \Sigma_j \right)}{f_{CN} \left( x_i; \mu_j^{(r)}, \Sigma_j^{(r)}, \alpha_j^{(r)}, \eta_j^{(r)} \right)}.$$

which corresponds to the posterior probability that $x_i$ in cluster $j$ is a mild outlier, $i = 1, \ldots, n$ and $j = 1, \ldots, k$. Thus, we obtain $Q(\vartheta) = Q_1(\pi) + Q_2(\alpha) + Q_3(\mu, \Sigma, \eta)$, where the three terms on the right-hand side are ordered as the three terms on the right-hand side of (5).

**C-step.** We first trim the $\lceil \alpha_0 n \rceil$ observations with smallest values of $D \left( x_i; \vartheta^{(r)} \right) =$

$$\max \left\{ D_1 \left( x_i; \vartheta_1^{(r)} \right), \ldots, D_k \left( x_i; \vartheta_k^{(r)} \right) \right\}$$

by letting

$$R_0^{(r)} = \left\{ i : D \left( x_i; \vartheta^{(r)} \right) \leq D \left( x_{\left\lceil \alpha_0 n \right\rceil}; \vartheta^{(r)} \right) \right\},$$

where $D \left( x_{(i)}; \vartheta^{(r)} \right)$, with a slight abuse of notation, is the $i$-th ordered (from the smallest to the largest) value of $D \left( x_i; \vartheta^{(r)} \right)$. Then, we assign each remaining observation $x_i$ to the cluster $j$ such that $D_j \left( x_i; \vartheta_j^{(r)} \right) =$
CM-step 1. Based on the untrimmed observations $\mathbf{x}$, and their assignment to the $k$ clusters, we calculate $\vartheta^{(r)}$ as the value of $\vartheta_1$ that maximizes $Q(\vartheta)$ with $\vartheta_2$ fixed at $\vartheta^{(r)}_2$. As $Q_1(\pi)$, $Q_2(\alpha)$, and $Q_3(\mu, \Sigma, \eta)$ have zero cross-derivatives, they can be maximized separately. Maximizing $Q_1(\pi)$ with respect to $\pi$, subject to the constraints on these parameters, yields $\pi_i^{(r+1)} = \frac{n_i^{(r)}}{\pi_i^{(r)} \alpha}$, where $n_i^{(r)} = \#R_i^{(r)}$, $j = 1, \ldots, k$. Maximizing $Q_2(\alpha)$ with respect to the generic element $\alpha$, $j = 1, \ldots, k$, of $\alpha$, subject to the constraints on this parameter, yields $\alpha_j^{(r+1)} = \min \left\{ \frac{1}{n_j^{(r)}} \sum_{i \in R_j^{(r)}} v_{ij}^{(r)}, \alpha_{\text{max}} \right\}$, where $\alpha_{\text{max}}$ is an optional maximum value for $\alpha$, to be considered if we require that in the $j$th cluster, $j = 1, \ldots, k$, the proportion of mild outliers is at most equal to a predetermined value $\alpha_{\text{max}}$. In the analyses herein, we fix $\alpha_{\text{max}} = 0.5$ (see also Punzo and McNicholas, 2016, 2017). Finally, maximizing $Q_3(\mu, \Sigma, \eta)$ with respect to $\mu$ and $\Sigma$, with $\eta$ fixed at $\eta^{(r)}$, yields the updates

$$\mu_j^{(r+1)} = \frac{1}{s_j^{(r)}} \sum_{i \in R_j^{(r)}} \left( 1 - v_{ij}^{(r)} + \frac{v_{ij}^{(r)}}{\eta_j^{(r)}} \right) \mathbf{x}_i,$$  

and

$$\Sigma_j^{(r+1)} = \frac{1}{n_j^{(r)}} \sum_{i \in R_j^{(r)}} \left( 1 - v_{ij}^{(r)} + \frac{v_{ij}^{(r)}}{\eta_j^{(r)}} \right) (\mathbf{x}_i - \mu_j^{(r+1)})(\mathbf{x}_i - \mu_j^{(r+1)})^T,$$

where

$$s_j^{(r)} = \sum_{i \in R_j^{(r)}} \left( 1 - v_{ij}^{(r)} + \frac{v_{ij}^{(r)}}{\eta_j^{(r)}} \right).$$

CM-step 2. We calculate $\vartheta_2^{(r+1)}$ as the value of $\vartheta_2$ that maximizes $Q(\vartheta)$ with $\vartheta_1$ fixed at $\vartheta_1^{(r+1)}$. This is equivalent to maximize $Q_3(\mu = \mu^{(r+1)}, \Sigma = \Sigma^{(r+1)}, \eta)$ with respect to $\eta$. Such a maximization, with respect to the generic element $\eta$, $j = 1, \ldots, k$, of $\eta$, under the constraint $\eta_j^{(r+1)} > 1$, $j = 1, \ldots, k$, yields

$$\eta_j^{(r+1)} = \max \left\{ \eta, \frac{1}{d n_j^{(r)} \sum_{i \in R_j^{(r)}} v_{ij}^{(r)}} \delta(\mathbf{x}_i, \mu_j^{(r+1)}; \Sigma_j^{(r+1)}) \right\},$$
8 Alessio Farcomeni, Antonio Punzo

where \( n_{\text{out}, j}^{(r)} = \sum_{i \in R_j^{(r)}} v_{ij}^{(r)} \) and \( \eta^* \) is a number close to 1 from the right; for the analyses of this paper, we use \( \eta^* = 1.001 \). Further details can be found in Mazza and Punzo (2017) and Punzo et al. (2018, 2020).

As initial value \( \vartheta(0) \) we use the solution provided by the usual robust normal mixture model (tclust) with trimming level \( \alpha_0 \) and ER \( c \). For \( \alpha_j(0) \) and \( \eta_j(0) \) in \( \vartheta(0) \), we use \( \alpha_j(0) = 0.001 \) and \( \eta_j(0) = 1.001, j = 1, \ldots, k \); this choice is justified by the nesting relation between the normal and contaminated normal distributions.

### 4 Penalized estimation

Model (2) involves the difficult choice of \( k + 1 \) parameters \( \alpha_0, \alpha_1, \ldots, \alpha_k \), where \( \alpha_0 \) controls the proportion of trimmed observations and \( \alpha_j \) the proportion of observations in the \( j \)-th cluster belonging to the component with inflated variance. Note that the likelihood is monotone as a function of \( \alpha_0 \), while the MLE for \( \alpha_1, \ldots, \alpha_k \) is well defined. Penalization of the likelihood through a function of \( \alpha_1, \ldots, \alpha_k \) still makes sense in order to provide super-efficiency at the normal mixture model. These parameters are clearly intertwined, as obviously the optimal \( \alpha_0 \) for a classical trimmed normal mixture model (i.e. the case \( \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \)) is not smaller than the optimal \( \alpha_0 \) obtained under the case in which at least one \( \alpha_j > 0 \). Choice of the trimming proportion is an open and debated problem in the robust clustering literature (e.g., García-Escudero et al., 2011 and Dotto et al., 2018). A subjective choice is often needed at some level. We propose here a LASSO-type penalized likelihood.

Our penalized approach enforces a sparse model selection in which some values in the set \( (\alpha_0, \alpha_1, \ldots, \alpha_k) \) might be set to zero. The ultimate goal is to set \( \alpha_0 \) as close as possible to the true proportion of gross outliers, possibly larger rather than smaller; and similarly to set \( \alpha_j \) as small as possible without having to increase \( \alpha_0 \). If gross outliers are present, we expect \( \alpha_0 \) to be increased notwithstanding the penalty; while if only mild outliers are present, some of the \( \alpha_1, \ldots, \alpha_k \) will be larger than zero but there will be no hard trimming, thus increasing the efficiency of the estimates. Finally, in case data arise from a normal mixture model with no contamination, the finely tuned penalty shall enforce \( \hat{\alpha}_j = 0 \) for \( j = 0, 1, \ldots, k \). A general form of penalized log-likelihood is given by

\[
\ell(\vartheta) + P(\alpha_0, \alpha_1, \ldots, \alpha_k).
\]

We propose using \( P(\alpha_0, \ldots, \alpha_k) = -\log(n) \sum_{j=0}^{k} \nu_j \alpha_j \). In several cases it is desired to maximize \( \sum_{j=1}^{k} \alpha_j / \alpha_0 \), so that as many observations as possible are included in the estimation set. It is straightforward to check that this corresponds to setting \( \nu_j = 0 \) for \( j > 0 \). When superefficiency at the normal mixture model is desired, then it might be more convenient to penalize also \( \alpha_j \) for \( j > 0 \). In this case, in order to reduce the number of penalty parameters, we propose setting \( \nu_0 = n \nu \) and \( \nu_j = \nu \) for \( j > 0 \). In all cases the penalty parameter \( \nu \) can be chosen as described below. The fact that for \( \nu \) large enough

\[
\text{where } n_{\text{out}, j}^{(r)} = \sum_{i \in R_j^{(r)}} v_{ij}^{(r)} \text{ and } \eta^* \text{ is a number close to 1 from the right; for the analyses of this paper, we use } \eta^* = 1.001. \text{ Further details can be found in Mazza and Punzo (2017) and Punzo et al. (2018, 2020).}

As initial value \( \vartheta(0) \) we use the solution provided by the usual robust normal mixture model (tclust) with trimming level \( \alpha_0 \) and ER \( c \). For \( \alpha_j(0) \) and \( \eta_j(0) \) in \( \vartheta(0) \), we use \( \alpha_j(0) = 0.001 \) and \( \eta_j(0) = 1.001, j = 1, \ldots, k \); this choice is justified by the nesting relation between the normal and contaminated normal distributions.

### 4 Penalized estimation

Model (2) involves the difficult choice of \( k + 1 \) parameters \( \alpha_0, \alpha_1, \ldots, \alpha_k \), where \( \alpha_0 \) controls the proportion of trimmed observations and \( \alpha_j \) the proportion of observations in the \( j \)-th cluster belonging to the component with inflated variance. Note that the likelihood is monotone as a function of \( \alpha_0 \), while the MLE for \( \alpha_1, \ldots, \alpha_k \) is well defined. Penalization of the likelihood through a function of \( \alpha_1, \ldots, \alpha_k \) still makes sense in order to provide super-efficiency at the normal mixture model. These parameters are clearly intertwined, as obviously the optimal \( \alpha_0 \) for a classical trimmed normal mixture model (i.e. the case \( \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \)) is not smaller than the optimal \( \alpha_0 \) obtained under the case in which at least one \( \alpha_j > 0 \). Choice of the trimming proportion is an open and debated problem in the robust clustering literature (e.g., García-Escudero et al., 2011 and Dotto et al., 2018). A subjective choice is often needed at some level. We propose here a LASSO-type penalized likelihood.

Our penalized approach enforces a sparse model selection in which some values in the set \( (\alpha_0, \alpha_1, \ldots, \alpha_k) \) might be set to zero. The ultimate goal is to set \( \alpha_0 \) as close as possible to the true proportion of gross outliers, possibly larger rather than smaller; and similarly to set \( \alpha_j \) as small as possible without having to increase \( \alpha_0 \). If gross outliers are present, we expect \( \alpha_0 \) to be increased notwithstanding the penalty; while if only mild outliers are present, some of the \( \alpha_1, \ldots, \alpha_k \) will be larger than zero but there will be no hard trimming, thus increasing the efficiency of the estimates. Finally, in case data arise from a normal mixture model with no contamination, the finely tuned penalty shall enforce \( \hat{\alpha}_j = 0 \) for \( j = 0, 1, \ldots, k \). A general form of penalized log-likelihood is given by

\[
\ell(\vartheta) + P(\alpha_0, \alpha_1, \ldots, \alpha_k).
\]

We propose using \( P(\alpha_0, \ldots, \alpha_k) = -\log(n) \sum_{j=0}^{k} \nu_j \alpha_j \). In several cases it is desired to maximize \( \sum_{j=1}^{k} \alpha_j / \alpha_0 \), so that as many observations as possible are included in the estimation set. It is straightforward to check that this corresponds to setting \( \nu_j = 0 \) for \( j > 0 \). When superefficiency at the normal mixture model is desired, then it might be more convenient to penalize also \( \alpha_j \) for \( j > 0 \). In this case, in order to reduce the number of penalty parameters, we propose setting \( \nu_0 = n \nu \) and \( \nu_j = \nu \) for \( j > 0 \). In all cases the penalty parameter \( \nu \) can be chosen as described below. The fact that for \( \nu \) large enough

some $\alpha$ parameters are set to zero is a direct consequence of the fact that $\ell(\theta)$ is bounded, while the penalty (as a function of $\nu$) is not.

Maximization of the penalized log-likelihood proceeds as in the previous section. The E-step is unchanged, and so is the update of centroids and scatter matrices at the first CM-step. On the other hand, in order to update $\alpha_1, \ldots, \alpha_k$, we proceed through numerical optimization of the expected complete log-likelihood $Q_2(\alpha)$ plus penalty. The log-likelihood is a step function of $\alpha_0$, with jumps at $j/n$ for $j = 1, 2, \ldots, n$. Consequently, it suffices to evaluate (12) at $\alpha_0 = 0/n, 1/n, \ldots, n_{0,\text{max}}/n$ and optimize $\ell(\theta) - n\nu \log(n) \sum_{j=0}^k \alpha_j$ as a function of $\alpha_1, \ldots, \alpha_k$ and the remaining parameters. This leads to a set of $n_{0,\text{max}} + 1$ possible solutions, which can be used to compute (12). Finally, the parameter set corresponding to the largest value of (12) is selected as the penalized maximum likelihood estimate (by definition). For illustration, in the case $k = 2$, Fig. 1(b) reports an example of trajectories of $\alpha_0, \alpha_1,$ and $\alpha_2$ as a function of $\nu$. The underlying data, displayed in Fig. 1(a), consist of $n = 500$ bivariate observations ($d = 2$) randomly generated by a balanced normal mixture with centroids $\mu_1 = (0, 0)'$ and $\mu_2 = (4, 0)'$, and identity covariance matrices. In the first dimension 2% of points are randomly selected and replaced by values from a uniform distribution with support $(-8, -6) \times (-5, -3)$. We then let $\alpha_0 \in [0, 0.2] — to be more precise $\alpha_0 \in \{0/500, 1/500, \ldots, 100/500\}$ — and include $\alpha_0, \alpha_1,$ and $\alpha_2$ in the penalty. Clearly, when $\nu$ is too small $\alpha_0$ is too large. As $\nu$ increases, $\alpha_0$ decreases as well and some outliers are now recognized as mild outliers (hence $\alpha_1$ increases). For large $\nu$, $\alpha_0$ goes to zero and all unusual observations are assigned to cluster 1 (whose centroid is the closest to the outliers on the left). The only remaining parameter $\alpha_1$ of course decreases as a function of $\nu$, but given that there indeed are outliers it decreases slowly and $\nu \geq 12.5$ is needed to obtain $\alpha_0 = \alpha_1 = \alpha_2 = 0$.

4.1 Selecting the penalty parameter

The choice of the penalty parameter $\nu$ has got direct consequences on the estimation of the trimming proportion $\alpha_0$. If also $\alpha_1, \ldots, \alpha_k$ are included in the penalty, it also affects their estimates. There are several approaches in the literature for selecting the trimming proportion which can be extended to the case of selecting $\nu$. The most popular approach is probably the use of control curves (García-Éscudero et al., 2011), which would involve plotting the penalized log-likelihood at convergence as a function of $\nu$, for different values of $k$. Given that the objective function is monotone in $\nu$, the optimal one can be chosen as the minimal value after which the curve becomes slightly flat. A similar reasoning applies for selecting $k$. Since in this work we are indirectly monitoring several parameters, we have found that a useful approach is also that of plotting the trajectories of $\hat{\alpha}_j, j = 0, 1, \ldots, k$, as a function of $\nu$, as in Fig. 1(b). This is in the spirit of Cerioli et al. (2018), Farcomeni and Dotto (2018), Riani et al. (2019). Among the other possible approaches we mention cross-validation, which is relatively new and unexplored in the robust
clustering context, having been explored to the best of our knowledge only in Dotto et al. (2017). We do not pursue this route further here as it would be computationally cumbersome.

Surprisingly enough, mapping the problem of selecting mixing and trimming proportions to the scale of the likelihood also gives an asymptotically “optimal” fixed value, $\nu = \sqrt{2d}$. The rationale is as follows: suppose that an observation is not an outlier and that the $d$ dimensions are independent. Under the assumptions just stated $-2\ell_i(\hat{\theta})$, where $\ell_i(\hat{\theta})$ is the individual contribution to the log-likelihood, given in (3), as evaluated in $\hat{\theta}$, is the sum of $d$ weakly-dependent random variables. The latter in turn can be well approximated with a normal random variable once standardized. The approximation is clearly good only as $d$ increases, but it can be seen in simulations to work well also for small $d$. According to Embrechts et al. (2008), Section 3.5,

$$\frac{\max_i \{-2\ell_i(\hat{\theta})\} - d}{\log(n)\sqrt{2d}} \rightarrow 2,$$

implying that $\min \ell_i(\hat{\theta})/ \log(n)\sqrt{2d} \rightarrow 1$. The consequence is that the $\alpha_0n$ likelihood contributions below $\log(n)\sqrt{2d}$ might be considered as (asymptotically) not arising from the specified model (be it CN or normal mixture).

Of course, several refinements are possible, in the spirit for instance of Cerioli (2010). The most obvious is that of relaxing the independence working model. This would simply lead to a data-dependent optimal $\nu$ based on replacing $d$ and $\sqrt{2d}$ in (13) with the estimated mean and standard deviation of $-2\ell_i(\hat{\theta})$. Convergence properties would hold under slightly general assumptions also under dependence (Doukan, 1994 and Farcomeni, 2007). We leave this refinement as grounds for further work for the following reasons:

---

**Fig. 1** Simulated data set; scatter plot matrix (on the left) and trajectories of $\alpha_0$, $\alpha_1$ and $\alpha_2$ at convergence of (12) for different values of $\nu$ (on the right).
(i) $\nu = \sqrt{2d}$ choice seems to work well in several settings (see the simulation study of Section 5), (ii) estimation of the standard deviation of $-2\ell_i(\hat{\theta})$ seems to be cumbersome in practice, (iii) even after refinement we would obtain only an asymptotically optimal value, and (iv) it can be easily shown that optimal values for the common case in which dependence is positive (Esary et al., 1967) lead to larger asymptotically optimal $\nu$ (and hence lower trimming proportions).

4.2 Theoretical properties

In this section we discuss the theoretical properties of trimmed mixtures of contaminated normals. In a nutshell, basically all theoretical properties of trimmed mixtures of normal distributions (tclust) map to this case. We quickly recap the most important ones here and formally discuss how they extend to our setting.

First of all, in García-Escudero et al. (2008) it is shown that: (a) a sample solution exists for each $n > 0$, (b) a population solution exists and (c) the sample solution is consistent to the population solution. Conditions for results (a) and (b) are that the data generating distribution is strictly positive and is not concentrated on $k$ points after removing a probability mass equal to $\alpha_0$. These conditions of course go beyond the normal assumption for component density, and trivially hold for our CN (and even more general) assumptions. If additionally there exists a unique maximum of the objective function (which is not guaranteed even with normal mixture models), consistency holds.

Secondly, the breakdown point (BP) of tclust is studied in Ruwet et al. (2013) and Farcomeni (2014). The BP is defined as the maximum proportion of data than can be arbitrarily replaced without arbitrarily changing the final estimate (formally, such that the estimate lies within a compact set). It can be shown that arbitrarily replacing rows of the data matrix leads to a BP of $\alpha_0$ for $\Sigma_j$ as long as $\alpha_0 \leq 1/2 - k(d + 1)/2n$. For $\mu_j$, the BP is $\alpha_0$ as long as

$$\alpha_0 \leq \min \left( \frac{1}{2}, \frac{k(k - 1) + 1}{2n}, \frac{1}{k}, \frac{1}{n} \right),$$

and under the condition that data are generated from “well-clustered” data sets, that is, under conditions on the overlap (see Ruwet et al., 2013 for technical details). Finally, Farcomeni (2014) shows that arbitrarily replacing cells (instead of rows) of the data matrix leads to BPs $\alpha_0/d$ under the same conditions. Extension of these results to our case is not straightforward, but essentially follows the reasoning of Ruwet et al. (2013). We report here the formal statements and proofs.

**Theorem 1** Assume $\alpha_0 \leq 1/2 - k(d + 1)/2n$. The asymptotic breakdown point of the scatter matrices and of the inflation parameters $\eta_1, \ldots, \eta_k$ is $\alpha_0$. If additionally the data sets are “well-separated” as defined in Ruwet et al.
and
\[
\alpha_0 \leq \min \left( \frac{1}{2} - \frac{k(k-1) + 1}{2n}, \frac{1}{k} - \frac{1}{n} \right),
\]
then the asymptotic breakdown point of estimated cluster centers is \(\alpha_0\).

Proof
See Web Appendix

5 Simulation study

In this section, we investigate the behavior of the proposed methods through a simulation study. We consider three experimental factors: the dimension \(d \in \{2, 4\}\), the number of groups \(k \in \{2, 3\}\), and the data generating process (DGP). As concerns the DGP, we generate data from: i) a classical normal mixture (NM), ii) a contaminated normal mixture (CNM), iii) a normal mixture with replacement of 1% of the observations with gross outliers (NM with 1% of points randomly substituted by outliers generated from a uniform distribution over the interval \((99, 101)\) on each of the \(d\) dimensions); iv) a contaminated normal mixture with replacement of 1% with gross outliers (generated as in scenario iii); v) a normal mixture with 5% gross outliers as background noise (5% of points randomly substituted by noise generated from a uniform distribution over the interval \((-100, 100)\) on each of the \(d\) dimensions); and vi) a CNM with 5% background noise generated as in scenario v).

The precise parameters of the DGPs are given in the Web Appendix.

For each combination of the experimental factors, we generate 1000 data sets, each of size \(n = 200\), and compare the following approaches for cluster analysis: normal mixture (NM), \(t\) mixture (tM), contaminated normal mixture (CNM), \texttt{tclust} with \(\alpha_0\) chosen through our fixed penalty approach (NT-CLUST0), trimmed mixture of contaminated normals with \(\alpha_0\) chosen through our fixed penalty approach (CNTCLUST0), and a trimmed mixture of contaminated normals with fixed penalty on the entire vector \((\alpha_0, \alpha_1, \ldots, \alpha_k)\) (CNTCLUST). We initialize all algorithms using the results of a simple \(k\)-means approach. For selecting the trimming proportion we use a finite grid \(\{0/n, 1/n, \ldots, n_{0,\text{max}}/n\}\), with \(n_{0,\text{max}} = \lfloor 0.2n \rfloor = 40\) to pick \(\alpha_0\). In Table 1 we report results for \(k = 2\), while \(k = 3\) can be found in the Web Appendix.

The sum of the mean square errors (MSEs) are computed over three groups of parameters: the weights \(\pi_1, \ldots, \pi_k\), the means \(\mu_1, \ldots, \mu_k\), and the scale matrices \(\Sigma_1, \ldots, \Sigma_k\), notationally compacted as \(\pi\), \(\mu\), and \(\Sigma\) hereafter. Results for \(k = 3\) are in Web Appendix. To tackle label switching issues we assigned labels by minimizing the distance between the estimated and true values of \(\pi\) and \(\mu\) (Stephens, 2000).

Additionally, in Web Appendix we report the estimated values of \(\alpha_0, \alpha_1, \ldots, \alpha_k\) in all scenarios for the three penalized estimation methods considered.

The following considerations can be put forward. When the data arise from a classical normal mixture with no contamination, all methods perform comparably well to the NM method; with in some cases a slight increase in MSE.
which is due to the loss of efficiency that must be paid in order to guarantee robustness in the presence of contamination. Our penalty approach seems to work well in that $\hat{\alpha}_0$ is very close to zero (even always zero in some settings) and this holds also for $\hat{\alpha}_j$, $j = 1, \ldots, k$ when also these parameters are penalized. In the presence of only mild outliers, that is, when data arise from a CNM, $tM$, CNM, TCLUST0, CNTCLUST0, CNTCLUST are clearly better than the traditional NM. Under the remaining scenarios, where gross outliers are always present, NM, $tM$, and CNM break down. In these cases trimming is mandatory, and the other three methods work well. Notably, our penalized approach used in conjunction with tclust works very well in selecting the trimming proportion, which is either corresponding to the true one or slightly larger. Note that a trimming proportion slightly larger than the true proportion of gross outliers is only leading to a mild loss of efficiency, while the reverse can lead to break down of the estimates. This phenomenon is particularly evident with trimmed mixtures of contaminated normals when all parameters are penalized, as some mild outliers are hard trimmed.

Summarizing, robust penalized methods work globally better than NM, $tM$, and CNM. Among NTCLUST0, CNTCLUST0, and CNTCLUST, we think that the best compromise to be CNTCLUST0, which is able to robustify the classical NM in the presence of mild and/or gross outliers, and does not trim too many mild outliers. On the other hand, NTCLUST0 (tclust with automatically selected trimming proportion) works generally better when only

---

Table 1 Simulation study. MSE of parameter groups when $k = 2$. The results are averaged over 1000 replicates.

<table>
<thead>
<tr>
<th></th>
<th>NM</th>
<th>$tM$</th>
<th>CNM</th>
<th>TCLUST0</th>
<th>CNTCLUST0</th>
<th>CNTCLUST</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d = 2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\pi$</td>
<td>0.001</td>
<td>0.002</td>
<td>0.001</td>
<td>0.003</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>$\mu$</td>
<td>0.050</td>
<td>0.050</td>
<td>0.050</td>
<td>0.050</td>
<td>0.050</td>
<td>0.050</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>0.135</td>
<td>0.138</td>
<td>0.222</td>
<td>0.135</td>
<td>0.219</td>
<td>0.139</td>
</tr>
<tr>
<td>$d = 4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\pi$</td>
<td>0.001</td>
<td>0.002</td>
<td>0.001</td>
<td>0.003</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>$\mu$</td>
<td>0.095</td>
<td>0.095</td>
<td>0.095</td>
<td>0.095</td>
<td>0.095</td>
<td>0.095</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>0.344</td>
<td>0.347</td>
<td>0.417</td>
<td>0.344</td>
<td>0.417</td>
<td>0.417</td>
</tr>
</tbody>
</table>

...
gross outliers are expected to be present. CNTCLUST, as expected, performs well for the identification of a classical normal mixture model.

6 Example about clustering illicit drug shipments

In this section we analyze data about 151 seizures of shipments of cocaine and heroin in Italy and Spain. The data are a subset of the database of transnational controlled deliveries in drug trafficking investigations, elaborated within the EU funded Project TDT JUST/2011/ISEC/DRUGS/AG/3670 during 2013, 2014 and 2015. Seizures were performed by the governmental police officers of each country, and the substance confiscated ranged from having an aspect of flakes, to powder, to solid bricks. Weights ranged from less than 100 grams to several kilograms. At each event, a sample was sent to the forensic laboratories for checking the nature of the substance and quantifying the absolute and relative contents of each of several chemical compounds. In modern forensics it is believed that the contents of certain solvents might be useful for identifying the source, that is, clustering packages with respect to the illicit laboratory where the drug was processed. We verify this assumption by focusing on three compounds: hexane, acetone, and 2-propanol. Our data matrix is therefore made of the concentration of these \( d = 3 \) compounds for \( n = 151 \) seizures. Since data are severely skewed, and some measurements are exactly zero, we add one unit and take the logarithm of the original measurements. We fix \( k = 2 \) and estimate a classical normal mixture model and a contaminated normal mixture model without trimming first. Then we use robust clustering methods: \texttt{tclust} and the contaminated normal mixture model with trimming.

In Table 2 we report, for several values of the trimming level, the adjusted Rand-index (ARI; Hubert and Arabie, 1985) showing the agreement between the class labels and the true underlying Italy/Spain location of seizure.

In Table 2 it can be seen that with no or insufficient trimming one might conclude that there is no relationship between solvent contents and seizure location. On the other hand, after trimming the agreement becomes fairly high. We shall add that, of course, true labels can not be used for choosing the most appropriate technique and trimming level, hence in practice one should select among NM, \( t \text{M} \), CNM, \texttt{tclust}(0.066), CNTCLUST0(0.053) and CNTCLUST(0.053), where the trimming levels are chosen using our fixed-penalty approach. First of all, as expected we note that the optimal trimming level using \texttt{tclust} is slightly larger than those using CNTCLUST0. While in our low sample size example this might not have strong consequences in terms of MSE, \( |151(0.066 - 0.053)| = 2 \) seizures will not be attributed to a location using \texttt{tclust}, which can have forensic consequences.

We conclude this section by reporting in Table 3 cluster centroids for some methods, where it can be seen that with NM and CNM clusters are too separated due to the presence of outliers. On the other hand, with \( t \text{M} \) clusters do not seem to separate sufficiently, even if a slightly better performance than NM and CNM is seen in terms of ARI. Surprisingly, also with \texttt{tclust} the
Table 2 Real data example. Adjusted Rand-index for assessing agreement of cluster and true location labels. In parentheses, the trimming level $\alpha_0$ of the trimmed approaches. NM: normal mixture, $tM$: $t$ mixture, CNM: contaminated normal mixture, tclust: trimmed NM, CNTCLUST0: trimmed CNM, CNTCLUST: penalized trimmed CNM with $\nu = \sqrt{2d}$ and fixed trimming level. The trimming level selected with our fixed-penalty approach is indicated with $\hat{\alpha}_0$.

<table>
<thead>
<tr>
<th>Method</th>
<th>ARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM</td>
<td>-0.076</td>
</tr>
<tr>
<td>$tM$</td>
<td>-0.032</td>
</tr>
<tr>
<td>CNM</td>
<td>-0.069</td>
</tr>
<tr>
<td>tclust(0.025)</td>
<td>-0.076</td>
</tr>
<tr>
<td>tclust(0.05)</td>
<td>0.435</td>
</tr>
<tr>
<td>tclust($\alpha_0 = 0.066$)</td>
<td>0.657</td>
</tr>
<tr>
<td>tclust(0.075)</td>
<td>0.653</td>
</tr>
<tr>
<td>CNTCLUST(0.025)</td>
<td>-0.076</td>
</tr>
<tr>
<td>CNTCLUST(0.05)</td>
<td>0.660</td>
</tr>
<tr>
<td>CNTCLUST($\alpha_0 = 0.053$)</td>
<td>0.660</td>
</tr>
<tr>
<td>CNTCLUST(0.075)</td>
<td>0.628</td>
</tr>
<tr>
<td>CNTCLUST(0.025)</td>
<td>0.076</td>
</tr>
<tr>
<td>CNTCLUST(0.05)</td>
<td>0.660</td>
</tr>
<tr>
<td>CNTCLUST($\alpha_0 = 0.053$)</td>
<td>0.658</td>
</tr>
<tr>
<td>CNTCLUST(0.075)</td>
<td>0.628</td>
</tr>
</tbody>
</table>

Table 3 Real data example. Cluster centers obtained with different methods.

<table>
<thead>
<tr>
<th></th>
<th>Hexane Cluster 1</th>
<th>Hexane Cluster 2</th>
<th>Acetone Cluster 1</th>
<th>Acetone Cluster 2</th>
<th>2-Propanol Cluster 1</th>
<th>2-Propanol Cluster 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM</td>
<td>0.057</td>
<td>0.255</td>
<td>0.006</td>
<td>0.010</td>
<td>0.029</td>
<td>0.145</td>
</tr>
<tr>
<td>$tM$</td>
<td>0.001</td>
<td>0.091</td>
<td>0.002</td>
<td>0.006</td>
<td>0.000</td>
<td>0.047</td>
</tr>
<tr>
<td>CNM</td>
<td>0.055</td>
<td>0.328</td>
<td>0.006</td>
<td>0.029</td>
<td>0.029</td>
<td>0.130</td>
</tr>
<tr>
<td>tclust(0.066)</td>
<td>0.046</td>
<td>0.240</td>
<td>0.006</td>
<td>0.017</td>
<td>0.022</td>
<td>0.137</td>
</tr>
<tr>
<td>CNTCLUST(0.053)</td>
<td>0.002</td>
<td>0.094</td>
<td>0.004</td>
<td>0.007</td>
<td>0.004</td>
<td>0.047</td>
</tr>
<tr>
<td>CNTCLUST(0.053)</td>
<td>0.002</td>
<td>0.095</td>
<td>0.004</td>
<td>0.007</td>
<td>0.004</td>
<td>0.048</td>
</tr>
</tbody>
</table>

7 Conclusions

In this work we have given two contributions: on the one hand, we have proposed a trimmed mixture of contaminated normal distributions, therefore somehow bridging the gap between use of flexible mixtures and trimming methods for model-based clustering. Our approach provides fine tuning for
the trade-off between robustness and efficiency, as some mild outliers can be included into the estimation set and contribute to unbiased parameter estimation. Secondly, we have proposed a novel penalized likelihood approach for selecting the trimming parameter(s), which seems to work well especially with trimmed normal mixtures (tclust) when a theoretically grounded penalty parameter is used.

In real data applications it is particularly interesting also to monitor $\nu$ to check whether $\alpha_0$ is set to zero before any $\alpha_j$, $j > 0$. This will give a clear message to the practitioner about the nature of outliers (e.g., whether they indeed are mild or gross outliers). Beyond monitoring, a route for further work is represented by providing a high breakdown procedure which is initialized by a very low tuning parameter, which can then be increased. We speculate a close relationship might exist between low initial penalty parameters and wild trimming (Cerioli et al., 2019). See also Atkinson et al. (2018).

The proposed trimmed CN mixture is provably more efficient than mixtures of normal distributions when there are mild outliers. Indeed, when data arise from the CN mixture, tclust needs to trim a proportion $\alpha_0 + \sum_j \pi_j \alpha_j$ of observations, while our proposed CNTCLUST only $\alpha_0$. For moderately large $n$, this translates into a substantial difference in terms of MSE. Our trimmed contaminated normal mixture model implies clusters with regular observations and mild outliers being both elliptically-shaped. Under specific empirical settings this could be rather restrictive, and to overcome this issue we could extend the untrimmed part of our model to account for possible component-wise skewness. This extension may be achieved by first considering a mixture of skewed distributions for the regular data, such as mixtures of skew-normal distributions (Lin, 2009) or mixtures of shifted asymmetric Laplace (SAL) distributions (Franczak et al., 2014), and then contaminating each component pdf via mixtures of skew-contaminated normal distributions (Cabral et al., 2012), or mixtures of contaminated SAL distributions (Morris et al., 2019), respectively. The use of a contaminated approach to make heavier the tails of the component distributions allows us to have a proportion of mild outliers in each cluster, aiding in the interpretation and simplifying the use of the penalized approach as all trimming parameters are defined on the same scale.

Acknowledgements The authors are grateful to two referees for constructive and helpful suggestions.

References


