Journal of Machine Learning Research 18 (2017) 1-39 Submitted 8/16; Revised 9/17; Published 11/17

Consistency, Breakdown Robustness, and Algorithms for Robust Improper Maximum Likelihood Clustering

Pietro Coretto PCORETTO@UNISA.IT
Department of Economics and Statistics
University of Salerno
Fisciano (SA), Italy

Christian Hennig C.HENNIG@UCL.AC.UK
Department of Statistical Science
University College London
London, United Kingdom

Editor: Ingo Steinwart

Abstract

The robust improper maximum likelihood estimator (RIMLE) is a new method for robust multivariate clustering finding approximately Gaussian clusters. It maximizes a pseudo-likelihood defined by adding a component with improper constant density for accommodating outliers to a Gaussian mixture. A special case of the RIMLE is MLE for multivariate finite Gaussian mixture models. In this paper we treat existence, consistency, and breakdown theory for the RIMLE comprehensively. RIMLE's existence is proved under non-smooth covariance matrix constraints. It is shown that these can be implemented via a computationally feasible Expectation-Conditional Maximization algorithm.

Keywords: Robustness, Improper density, Mixture models, Model-based clustering, Maximum likelihood, ECM-algorithm

1. Introduction

Maximum likelihood estimation (MLE) in a Gaussian mixture model with mixture components interpreted as clusters is a popular approach to cluster analysis (see, e.g., Fraley and Raftery (2002)). In many datasets not all observations can be assigned appropriately to clusters that can be properly modelled by a Gaussian distribution, and it is also well known that the MLE can be strongly affected by outliers (Hennig (2004)). In this paper we investigate the recently introduced "robust improper maximum likelihood estimator" (RIMLE, see Coretto and Hennig, 2016), a method for robust clustering with clusters that can be approximated by multivariate Gaussian distributions. The basic idea of RIMLE is to fit an improper density to the data that is made up by a Gaussian mixture density and a "pseudo mixture component" defined by a small constant density, which is meant to capture outliers and observations in low density areas of the data that cannot properly be assigned to a Gaussian mixture component (called "noise" in the following). This is inspired by the addition of a uniform "noise component" to a Gaussian mixture suggested by Banfield and Raftery (1993). Hennig (2004) showed that using an improper density improves the breakdown robustness of this approach for one-dimensional datasets. As in

many other statistical problems, violations of the model assumptions may cause problems in cluster analysis. Our general attitude to the use of statistical models in cluster analysis is that the models should not be understood as reflecting some underlying but in practice unobservable "truth", but rather as thought constructs implying a certain behaviour of methods derived from them (e.g., maximizing the likelihood), which may or may not be appropriate in a given application (more details on the general philosophy of clustering can be found in Hennig and Liao (2013); Hennig (2015b)). Using a model such as a mixture of multivariate Gaussian distributions, interpreting every mixture component as a "cluster", implies that we look for clusters that are approximately "Gaussian-shaped", but we do not want to rely on whether the data really were generated i.i.d. by a Gaussian mixture. We focus on situations in which the number of clusters G is fixed.

There is a number of proposals already in the literature for accounting for the presence of noise and outliers in model-based clustering problems. The contributions can be divided in two groups: methods based on mixture modelling, and methods based on fixed partition models. Within the first group Banfield and Raftery (1993) and Coretto and Hennig (2011) dealt with uniform distributions added as "noise components" to a finite Gaussian mixture. Peel and McLachlan (2000) proposed to model data based on Student t-distributions. Cuesta-Albertos et al. (1997) and García-Escudero and Gordaliza (1999) introduced and studied trimming in order to robustify the k-means partitioning method. Robust partitioning methods with homoscedastic clusters based on ML–type procedures where proposed in Gallegos (2002) and Gallegos and Ritter (2005). Heteroscedasticity in ML-type partitioning methods has been introduced with the TCLUST algorithm of García-Escudero et al. (2008) and the "k–parameters clustering" of Gallegos and Ritter (2013). More references and an in-depth overview are given in García-Escudero et al. (2015). Different from the methods based on fixed partition models, mixture models and RIMLE allow a smooth transition between different clusters and between clustered observations and noise, which improves parameter estimation in the presence of overlap between mixture components. The one-dimensional version of the RIMLE was introduced in Coretto and Hennig (2010) and was investigated based on Monte Carlo experiments. Extension of the methods to the multivariate setting is not straightforward. Existence and consistency of the MLE for the multivariate Gaussian mixtures is a long standing problem due to the ill-posedness of the likelihood function. Even for ML for plain multivariate Gaussian mixtures (i.e. the RIMLE with the improper constant density set to zero), consistency theory is limited to the situation in which the model is assumed to hold precisely, and restrictive conditions are required (e.g., Redner and Walker (1984)). Chen and Tan (2009) and Alexandrovich (2014) propose and study a penalized ML estimator. García-Escudero et al. (2014) studied a classification ML estimator for Gaussian mixture that is based on the TCLUST idea.

In this paper we study the theoretical properties of the RIMLE as well as its computation. A comprehensive treatment of existence, consistency and robustness is given. This treatment includes the case of ML for multivariate Gaussian mixture as special case. Particularly, the robustness properties of RIMLE are superior to those of the mixture-based methods proposed by Banfield and Raftery (1993) and Peel and McLachlan (2000), as demonstrated later in the paper. For fitting plain Gaussian mixtures, some issues that are treated here arise as well, particularly the need to constrain the covariance matrices in order to avoid degeneration of the likelihood. Some literature on this is cited in Section

3.2. The consistency results given here in Section 4 are of a nonparametric nature and show the consistency of the RIMLE for the RIMLE-functional defined for a general class of sampling distributions. Similar results have been shown for a partition likelihood model (Gallegos and Ritter (2013)) and for alternative, trimming-based approaches to robust clustering (e.g., García-Escudero et al. (2008); Gallegos and Ritter (2009)). Compared to these results, there is an additional difficulty for the RIMLE, namely that degeneration of the likelihood needs to be prevented also in the case that almost all observations are assigned to the noise component and the remaining observations are fitted arbitrarily well. This may look like a disadvantage, but in the literature cited above such problems are only avoided by fixing the trimming rate. An analysis like the one given here, and in Coretto and Hennig (2016), is required for understanding the case in which both the proportion of points considered as "noise" and the density level at which this happens are flexible. Coretto and Hennig (2016) introduce the OTRIMLE, a data-adaptive choice of the improper constant density, the method's tuning constant for achieving robustness. That paper also includes a comprehensive simulation study comparing the different approaches to robust clustering. In the study, every method turns out to be superior for one or more setups, but the OTRIMLE achieves the most satisfactory overall performance.

The paper is organized as follows. We first discuss in Section 2 an artificial dataset to illustrate the issues RIMLE is meant to deal with. The RIMLE is introduced and defined in Section 3. In Section 4 existence and consistency of the RIMLE are proved. Section 5 treats the computation of the RIMLE and the choice of input parameters for the algorithms. Section 6 studies the breakdown robustness of RIMLE. Numerical experiments are presented in Section 7. Section 8 concludes the paper.

2. Artificial data examples

Every clustering method is designed to recover certain types of clusters even when they are based on methods and algorithms that apply universally. For instance, the well known k-means method aims to discover spherical balanced clusters, although the algorithm will find a solution when this is not the case. In this section we introduce some issues in robust clustering by showing examples of data affected by noise that cause trouble to most clustering methods, including those that supposedly explicitly account for it. These examples will illustrate the kind of clustering problem that the method investigated in this paper aims to address. Two artificial data sets are generated in dimension $p = 20$ from two sampling designs, called AsyNoise and GEM respectively, also considered for the numerical experiments presented in Section 7. The two data sets are shown in Figure 1 and 3. A detailed description of the sampling designs is given in Section 7.

In AsyNoise (Figure 1) there are 500 observations in 5 moderately separated clusters from student-t distributions with varying degrees of freedom, 187 observations (37.4%) are background noise. We have symmetric and elliptical clusters that are not well separated along all directions, and they are of different size. Although there is a deviation from Gaussianity in the tails of the clusters' distribution, methods based on Gaussian shapes are candidates to reconstruct such groups. Plain Gaussian mixture clustering (without noise component) fixing the number of clusters at $G = 5$ using the popular R package mclust of Fraley et al. (2012) puts the clustered points into two big clusters and assigns the noise to the

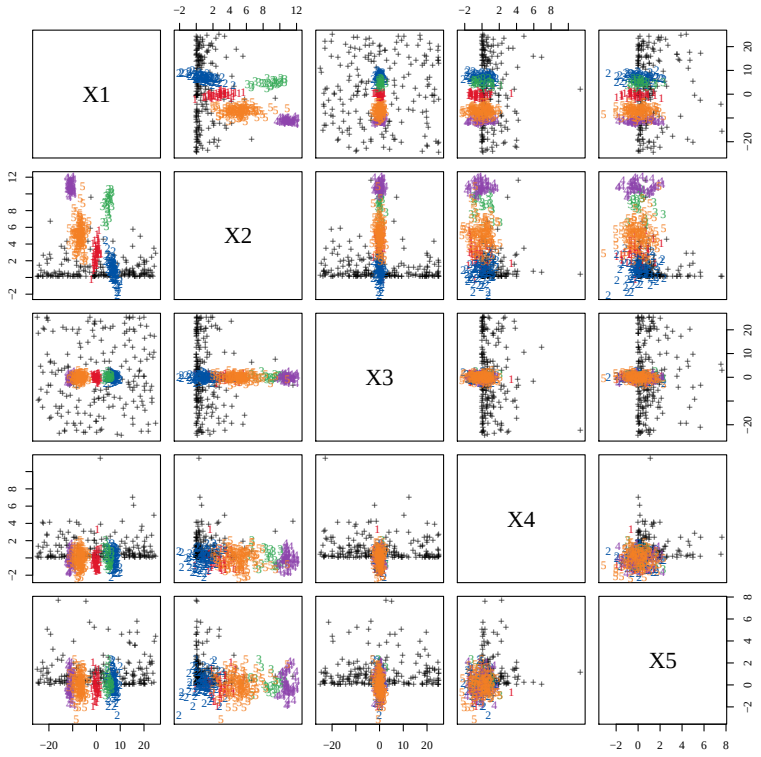

Figure 1: Scatter plots of $n = 500$ data points sampled from the AsyNoise design defined in Section 7. Marginals 1 to 5 are represented, further dimensions show a similar pattern. Colors denote the 5 clusters, while noise is represented by the "+" symbol.

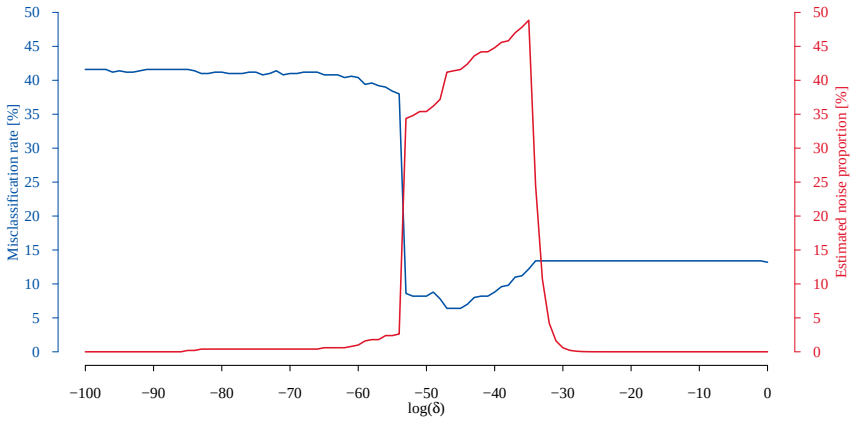

Figure 2: Misclassification rates (blue), and estimated noise proportions (red) by RIMLE over a grid of $\log(\delta)$ values for the data set in Figure 1.

remaining clusters achieving a misclassification rate of 61.4% (i.e., the best misclassification rate that can be achieved by permutation of the cluster labels so that no cluster is identified with the noise). Note that in this paper, mclust always automatically chooses an optimal covariance matrix parametrization by the BIC (see Fraley et al., 2012; Celeux and Govaert, 1995). One could wonder whether the data set may be an easy job for clustering methods that take into account outliers, but this is not the case as shown in Section 7. All robust methods require tuning that, directly or indirectly, controls the amount of noise present in the data set. Perhaps the only exception is the MLE for Gaussian mixtures with uniform noise of Banfield and Raftery (1993) implemented in the mclust package, but this can be led astray if the noise in fact behaves very differently from a uniform distribution. In real situations a priori information on the level of the noise is rarely available. The RIMLE method treated in this paper also requires tuning. The level of the noise is essentially controlled by the level of the improper noise density, called δ. For a given value of δ, the noise proportion then is estimated from the data.

For the data set in Figure 1 we computed the RIMLE for several values of $\log(\delta)$ (there are other constants required, chosen as $\gamma = 100$ and $\pi_{\max} = 0.5$, see Algorithm 2 and Sections 3). When choosing $\log(\delta)$ appropriately, namely $\log(\delta) \in [-53, -36]$, the RIMLE gets the structure of the data set right, and it stably produces a misclassification rate in the range [6.4%, 11%] with an estimated noise proportion in the range $[34.38\%, 45.8\%]$. This is clearly better than most other robust clustering methods we tried, see Section 7. Values of $\log(\delta)$ below -100 do not change the results. For large values of $\log(\delta)$ too much noise is found, hence the RIMLE's noise proportion constraint (see Section 3) becomes active and the resulting estimated noise proportion gets close to zero. The OTRIMLE criterion of Coretto and Hennig (2016) selects an optimal value $\log(\delta) = -40$, which is in the region where RIMLE shows its best performance. The RIMLE at $\log(\delta) = -40$ produces a misclassification rate of 8.8% with estimated noise proportion equal to 44.8%. Figure 2 shows how solutions change with changing values of $\log(\delta)$.

Another experimental situation considered in this paper is the GEM ("Gross Error Model") sampling design (Figure 3). In the GEM, 100 points are sampled from two normal populations with extremely different scatters, 2 points (2%) are outliers almost lying on a hyperplane. These outliers are not extremely separated from the regular points, and this can cause trouble to robust methods. Surprisingly, in a situation like this, some non-robust methods perform better than some robust alternatives. In fact, ML for plain Gaussian mixtures performed with mclust (without the noise component) assigns the two outliers to cluster 1 and achieves a misclassification rate of 2%, although the estimated mixture parameters are strongly biased. Robust methods may do worse if not well tuned (see Section 7). As for the previous data set the RIMLE has been computed for several values of $\log(\delta)$ maintaining all other parameters as before. The result can be seen in Figure 4. For any $-\infty < \log(\delta) \leq -46$ the RIMLE is 100% accurate and estimates a noise proportion of 2%. The OTRIMLE method for the data-driven choice of δ selects $\log(\delta) = -200$, which is in the region where the RIMLE achieves the best results.

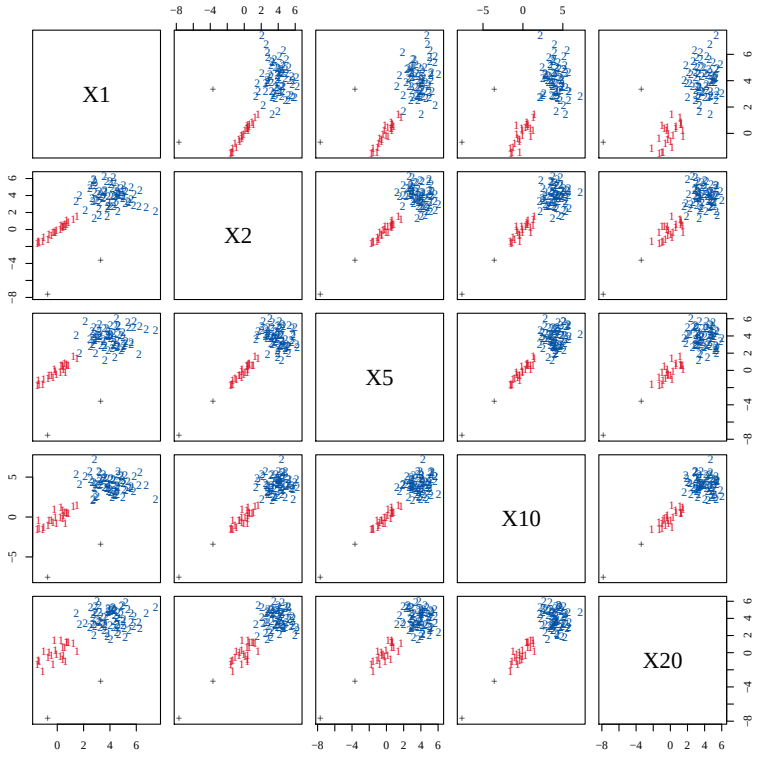

Figure 3: Scatter plots of $n = 100$ points sampled from the GEM sampling design defined in Section 7. Marginals 1,2,5,10, and 20 are represented, further dimensions show a similar pattern. Colors denote the 2 clusters, while noise is represented by the "+" symbol.

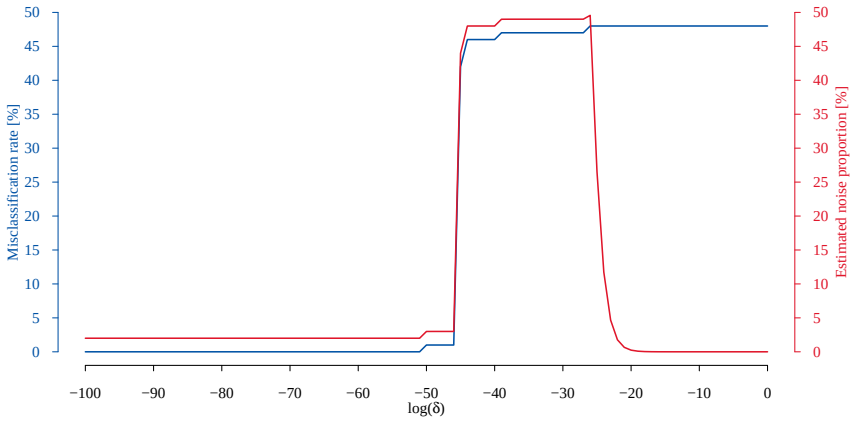

Figure 4: Misclassification rates (blue), and estimated noise proportions (red) by RIMLE over a grid of $\log(\delta)$ values for the data set in Figure 3.

3. Basic definitions

In this section we define the robust improper maximum likelihood estimator (RIMLE) along with its constrained parameter space.

3.1 RIMLE and clustering

The RIMLE is based on the "noise component" idea for robustification of the MLE based on the Gaussian mixture model. This models the noise by a uniform distribution, but in fact we are interested in more general patterns of noise or outliers. However, regions of high density are rather associated with clusters than with noise, so the noise regions should be those with the lowest density. This kind of distinction can be achieved by using the uniform density as in Banfield and Raftery (1993), but in the presence of gross outliers the dependence of the uniform distribution on the convex hull of the data causes a robustness problem (Hennig (2004)). The uniform distribution is not really used here as a model for the noise, but rather as a technical device to account for whatever goes on in low density regions. The RIMLE drives this idea further by using an improper uniform distribution the density value of which does not depend on how far awy extreme points in the data are from the main bulk. In the following, assume an observed sample $x_1, x_2, \ldots, x_n$, where x_i is the realization of a random variable $X_i \in \mathbb{R}^p$ with $p > 1$; $X_1, \ldots, X_n$ i.i.d. The goal is to cluster the sample points into G distinct groups. RIMLE then maximizes a pseudo-likelihood, which is based on the improper pseudo-density

$$\psi_\delta(x, \theta) = \pi_0 \delta + \sum_{j=1}^{G} \pi_j \phi(x; \mu_j, \Sigma_j), \tag{1}$$

where $\phi(\cdot, \mu, \Sigma)$ is the Gaussian density with mean μ and covariance matrix Σ, $\pi_0, \pi_j \in [0, 1]$ for $j = 1, 2, \ldots, G$, $\pi_0 + \sum_{i=1}^{G} \pi_j = 1$, while δ is the improper constant density. The parameter vector θ contains all Gaussian parameters plus all proportion parameters including π_0, ie. $\theta = (\mu_1, \ldots, \mu_G, \text{vect}(\Sigma_1), \ldots, \text{vect}(\Sigma_G), \pi_0, \ldots, \pi_G)$, where vect$(A)$ is the vectorized upper (or lower) triangle including the main diagonal of the symmetric square matrix A. δ and the number of Gaussian components G are considered fixed and known. Although this does not define a proper probability model, it yields a useful procedure for data modelled as a proportion of $(1 - \pi_0)$ of a mixture of Gaussian distributions, which have high enough density peaks to be interpreted as clusters plus a proportion π_0 times something unspecified with density smaller than or equal to δ (which may even contain further Gaussian components with so few points and/or so large within-component variation that they are not considered as "clusters"). The definition of the pseudo-model in (1) requires that the value of δ is fixed in advance. The choice of δ will be discussed in Section 5.2.

Given the sample improper pseudo-log-likelihood function

$$l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \psi_\delta(x_i, \theta), \tag{2}$$

the RIMLE is defined as

$$\theta_n(\delta) = \arg\max_{\theta \in \Theta_n} l_n(\theta), \tag{3}$$

where Θ_n is a constrained parameter space defined in Section 3.2. $\theta_n(\delta)$ is then used to cluster points using pseudo posterior probabilities for belonging to the Gaussian components or the improper uniform. These pseudo posterior probabilities are given by

$$\tau_j(x_i,\theta) := \begin{cases} \frac{\pi_0\delta}{\psi_\delta(x_i,\theta)} & \text{if } j=0 \\ \frac{\pi_j\phi(x_i,\mu_j,\Sigma_j)}{\psi_\delta(x_i,\theta)} & \text{if } j=1,2,\ldots,G; \end{cases} \quad \text{for } i=1,2,\ldots,n.$$

Points are assigned to the component for which the pseudo posterior probability is maximized. The assignment rule is then given by

$$J(x_i,\theta) := \underset{j\in\{0,1,2,\ldots,G\}}{\arg\max}\ \tau_j(x_i,\theta). \tag{4}$$

The assignment based on maximum posterior probabilities is common to all model-based clustering methods. Here, an improper density is involved, and so these are "pseudo posterior probabilities".

We also define a population version of the RIMLE for later deriving consistency results for the sequence $\{\theta_n(\delta)\}_{n\in\mathbb{N}}$. Let $\mathrm{E}_P f(x)$ be the expectation of $f(x)$ under P. The RIMLE population target function and the constrained parameter set can be obtained by replacing the empirical measure with P, and the population version of $l_n(\theta)$ is given by

$$L(\theta) = \mathrm{E}_P \log(\psi_\delta(x,\theta)).$$

Define $L_G = \sup_{\theta\in\Theta_G(P)} L(\theta)$, where $\Theta_G(P)$ is a constrained parameter space defined in Section 3.2.

3.2 The constrained parameter space

Some notation: the kth element of μ_j is denoted by $\mu_{j,k}$ for $k=1,2,\ldots,p$ and $j=1,2,\ldots,G$. Let $\lambda_{j,k}$ be the kth eigenvalue of Σ_j, define $\Lambda(\theta)=\{\lambda_{j,k}:\ j=1,2,\ldots,G;\ k=1,2,\ldots,p\}$, $\lambda_{\min}(\theta)=\min_{j,k}\{\Lambda(\theta)\}$, $\lambda_{\max}(\theta)=\max_{j,k}\{\Lambda(\theta)\}$.

Remark 1 *The p-dimensional Gaussian density can be written in terms of the eigen-decomposition of the covariance matrix:*

$$\phi(x;\mu,\Sigma) = (2\pi)^{-\frac{p}{2}}\left(\prod_{k=1}^{p}\lambda_k\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\sum_{k=1}^{p}\lambda_k^{-1}(x-\mu)'v_kv_k'(x-\mu)\right),$$

where λ_k is the k-th eigenvalue of Σ, and v_k is its associated eigenvector, for $k=1,2,\ldots,p$. Let $\lambda_0=\min\{\lambda_k;\ k=1,2,\ldots,p\}$. Then, $\lim_{\lambda_0\searrow 0}\phi(\mu;\mu,\Sigma)=\infty$, with $\phi(\mu;\mu,\Sigma)=O(\lambda_0^{-p/2})$ as $\lambda_0\searrow 0$. On the other hand $\lim_{\lambda_0\searrow 0}\phi(x;\mu,\Sigma)=0$ for all $x\neq\mu$, with $\phi(x;\mu,\Sigma)=o(\lambda_0^q)$ for any fixed q as $\lambda_0\searrow 0$. This implies that

$$\lim_{\lambda_0\searrow 0}\ \phi(\mu;\mu,\Sigma)\phi(x;\mu,\Sigma)\to 0 \quad \textit{for any } x\neq\mu$$

Furthermore, each of the density components in $\psi_\delta(\cdot)$ can be bounded above in terms of $\lambda_{\max}(\theta)$ and $\lambda_{\min}(\theta)$:

$$\phi(x;\mu_j,\Sigma_j)\leq(2\pi\lambda_{\min}(\theta))^{-\frac{p}{2}}\exp\left\{-\frac{1}{2}\lambda_{\max}(\theta)^{-1}\|x-\mu_j\|^2\right\}\leq(2\pi\lambda_{\min}(\theta))^{-\frac{p}{2}}. \tag{5}$$

The optimization problem in (3) requires that Θ_n is suitably defined, otherwise $\theta_n(\delta)$ may not exist. Consider a sequence $(\theta_m)_{m\in\mathbb{N}}$, as discovered by Kiefer and Wolfowitz (1956), the likelihood of a Gaussian mixtures degenerates if $\lambda_{1,1,m} \searrow 0$ if $\mu_{1,m} = x_1$, and this holds for (2), too. We here use the eigenratio constraint

$$\lambda_{\max}(\theta)/\lambda_{\min}(\theta) \leq \gamma < +\infty \tag{6}$$

with a constant $\gamma \geq 1$, where $\gamma = 1$ constrains all component covariance matrices to be spherical and equal, as in k-means clustering. This type of constraint has been proposed by Jr (1981), while Hathaway (1985) showed consistency of the scale-ratio constrained MLE for one-dimensional Gaussian mixtures. Ingrassia (2004) and Ingrassia and Rocci (2007) introduced EM algorithms for implementing these constraints for multivariate datasets. TCLUST by García-Escudero et al. (2008) and García-Escudero et al. (2014) also makes use of eigenratio constraints. Moreover there are a number of alternative constraints, see Ingrassia and Rocci (2011); Gallegos and Ritter (2009). It may be seen as a disadvantage of (6) that the resulting estimator will not be affine equivariant (this would require allowing $\lambda_{\max}(\theta)/\lambda_{\min}(\theta) \to \infty$ within any component). Affine equivariance can be achieved by defining a sphered version of the RIMLE as

$$\theta_n^*(\delta) = \theta_n^*(\delta, x_1, \ldots, x_n) = \theta_n(\delta, x_1^*, x_2^*, \ldots, x_n^*)$$

with $x_i^* = A(x_i - m),\ i = 1, \ldots, n$, where $S(x_1, \ldots, x_n)^{-1} = A'A$; $S(x_1, \ldots, x_n)$ could be the sample covariance matrix or another scale matrix and m the mean vector or another location estimator. This yields affine equivariance because the sphered versions of $\{x_1, \ldots, x_n\}$ and $\{Bx_1 + b, \ldots, Bx_n + b\}$ with some invertible $p \times p$-matrix B and $b \in \mathbb{R}^p$ are the same. Affine equivariance is not necessarily desirable though, see Hennig (2015a), Sec. 31.3.4.

This defines the parameter space

$$\tilde{\Theta} := \left\{ \theta :\ \pi_j \geq 0\ \forall j \geq 1,\ \pi_0 + \sum_{j=1}^{G} \pi_j = 1;\ \frac{\lambda_{\max}(\theta)}{\lambda_{\min}(\theta)} \leq \gamma \right\}. \tag{7}$$

Occasionally, later, the notation $\|\theta\|$ will refer to the Euclidean norm of a vector pieced together from all the parameters collected in θ, in which all covariance matrices are interpreted as subvectors of all the matrix entries.

Although (6) ensures the boundedness of the likelihood in standard mixture models and TCLUST, for RIMLE this is not enough. The Gaussian components could degenerate on a few points and all other points could be fitted by the improper uniform component. Therefore we impose an additional constraint:

$$\frac{1}{n} \sum_{i=1}^{n} \tau_0(x_i, \theta) \leq \pi_{\max}, \tag{8}$$

for fixed $0 < \pi_{\max} < 1$. The quantity $n^{-1} \sum_{i=1}^{n} \tau_0(x_i, \theta)$ can be interpreted as the estimated proportion of noise points. This constraint depends on the dataset. Unfortunately the similar looking constraint $\pi_0 \leq \pi_{\max}$ independent of the data will not do, because this could not stop more than a portion of π_0 points to be fitted by the improper uniform component.

There is therefore a constrained effective parameter space for RIMLE estimation depending on the dataset:

$$\Theta_n := \left\{ \theta \in \tilde{\Theta} : \ \pi_j \geq 0 \ \forall j \geq 1, \ \pi_0 + \sum_{j=1}^{G} \pi_j = 1; \ \frac{1}{n} \sum_{i=1}^{n} \tau_0(x_i, \theta) \leq \pi_{\max}; \ \frac{\lambda_{\max}(\theta)}{\lambda_{\min}(\theta)} \leq \gamma \right\}. \tag{9}$$

Analogously, existence and consistency of the RIMLE functional can only be showed on a parameter subset of $\tilde{\Theta}$ that depends on the underlying distribution and enforces that enough probability mass is fitted by Gaussian components rather than the improper uniform:

$$\Theta_G(P) := \left\{ \theta \in \tilde{\Theta} : \ \pi_j \geq 0 \ \forall j \geq 1; \ \pi_0 + \sum_{j=1}^{G} \pi_j = 1; \ \mathrm{E}_P \frac{\pi_0 \delta}{\psi_\delta(x, \theta)} \leq \pi_{\max}; \ \frac{\lambda_{\max}(\theta)}{\lambda_{\min}(\theta)} \leq \gamma \right\}. \tag{10}$$

4. RIMLE existence and consistency

We first show existence of the RIMLE for finite samples Let $\#(A)$ denote the cardinality of the set A. Let $\underline{x_n} = \{x_1, x_2, \ldots, x_n\}$. Lemma 2 concerns the important case of plain Gaussian mixtures ($\delta = 0$) and requires a weaker assumption A0(a) for existence than A0 required for the RIMLE with $\delta > 0$. Here are some assumptions:

A0(a) $\#(\underline{x_n}) > G$.

A0 $\#(\underline{x_n}) > G + \lceil n\pi_{\max} \rceil$.

Lemma 2 *Assume A0(a), $\delta = 0$. Let $(\theta_m)_{m\in\mathbb{N}}$ be a sequence such that $\lambda_{\max}(\theta_m)/\lambda_{\min}(\theta_m) \leq \gamma$. Assume also that for some $j = 1, 2, \ldots, G$ and $k = 1, 2, \ldots, p$, $\lambda_{k,j,m} \searrow 0$ as $m \to \infty$; then $\sup l_n(\theta_m) \to -\infty$.*

Proof $\lambda_{k,j,m} \searrow 0$ implies $\lambda_{\max}(\theta_m) \searrow 0$, $\lambda_{\min}(\theta_m) \searrow 0$ at the same speed because of (6). Assume w.l.o.g. (otherwise consider a suitable subsequence) that $(\theta_m)_{m\in\mathbb{N}}$ is such that $\mu_{j,m}$, $j = 1, 2, \ldots, G$ either leave every compact set for m large enough or converge, and assume w.l.o.g., that if their limits are in $\{x_1, \ldots, x_n\}$, they are in $\underline{x_G} = \{x_1, \ldots, x_G\}$. A0(a) implies that $\exists\, x_i \notin \underline{x_G}$, and $\exists \nu > 0$ such that for all such x_i, $j = 1, 2, \ldots, G$ and large enough $m : \|x_i - \mu_{j,m}\| \geq \nu$. Because the likelihood

$$\mathcal{L}_n(\theta_m) = \prod_{x_i \in \underline{x_G}} \Big\{ \sum_{j=1}^{G} \pi_j \phi(x_i; \mu_{j,m}, \Sigma_{j,m}) \Big\} \prod_{x_i \notin \underline{x_G}} \Big\{ \sum_{j=1}^{G} \pi_j \phi(x_i; \mu_{j,m}, \Sigma_{j,m}) \Big\}, \tag{11}$$

and Remark 1, the first product is of order $O(\lambda_{\min}(\theta_m)^{-p/2})^G$, and the second one of order $o(\lambda_{\min}(\theta_m)^q)$ for any fixed q, which implies that $\mathcal{L}_n(\theta_m) \to 0$ and $l_n(\theta_m) \to -\infty$. ■

Lemma 3 *Assume A0, $\delta > 0$. $(\theta_m)_{m\in N}$ is a sequence in Θ_n. Assume also that for some $j = 1, 2, \ldots, G$ and $k = 1, 2, \ldots, p$, $\lambda_{k,j,m} \searrow 0$ as $m \to \infty$. Then $l_n(\theta_m) \to -\infty$.*

Proof Using the definitions of the proof of Lemma 2, instead of (11) now

$$\mathcal{L}_n(\theta_m) = \prod_{x_i \in \underline{x_G}} \left\{ \pi_{0,m}\delta + \sum_{j=1}^{G} \pi_j \phi(x_i; \mu_{j,m}, \Sigma_{j,m}) \right\} \prod_{x_i \notin \underline{x_G}} \left\{ \pi_{0,m}\delta + \sum_{j=1}^{G} \pi_j \phi(x_i; \mu_{j,m}, \Sigma_{j,m}) \right\} \tag{12}$$

has to be considered, so that the limit behaviour of $(\pi_{0,m})_{m\in\mathbb{N}}$ is relevant. (8) implies

$$\frac{1}{n} \sum_{x_i \in \underline{x_G}} \left(1 + \sum_{j=1}^{G} \frac{\pi_j \phi(x_i, \mu_{j,m}, \Sigma_{j,m})}{\pi_{0,m}\delta} \right)^{-1} + \frac{1}{n} \sum_{x_i \notin \underline{x_G}} \left(1 + \sum_{j=1}^{G} \frac{\pi_j \phi(x_i, \mu_{j,m}, \Sigma_{j,m})}{\pi_{0,m}\delta} \right)^{-1} \leq \pi_{\max}. \tag{13}$$

Suppose that $(\pi_{0,m})_{m\in\mathbb{N}}$ does not converge to zero as $O(\phi(x_i, \mu_{j,m}, \Sigma_{j,m}))$ for at least one $x_i \notin \underline{x_G}$. For $m \to \infty$, the left term of (13) is ≥ 0, and the right term (at least a subsequence) converges to $\frac{\#(\underline{x_n} \setminus \underline{x_G})}{n}$, which A0 requires to be $> \pi_{\max}$ with contradiction, thus $\pi_{0,m} = O(\phi(x_i, x_j, \Sigma_{j,m}))$. Therefore, by the same argument as in the proof of Lemma 2, the right product in (12) vanishes fast enough so that $l_n(\theta_m) \to -\infty$. ■

From these Lemmas:

Theorem 4 (Finite Sample Existence) *Assume A0. Then $\theta_n(\delta)$ exists for all $\delta \geq 0$.*

Proof Θ_n depends on δ via (8). Θ_n is not empty for any δ, because for any fixed values of the other parameters, small enough π_0 will fulfil (8). Next show that there exists a compact set $K_n \subset \Theta_n$ such that $\sup_{\theta\in K_n} l_n(\theta) = \sup_{\theta\in\Theta_n} l_n(\theta)$.
Step A: consider θ such that $\pi_1 = 1$, $\mu_1 = x_1$, $\Sigma_j = I_p$ for all $j = 1, 2, \ldots, G$, arbitrary μ_j and $\pi_j = 0$ for all $j \neq 1$. For this, $l_n(\theta) = \sum_{i=1}^{n} \log \phi(x_i; x_1, \Sigma_1) > -\infty$, thus $\sup_{\theta\in\Theta_n} l_n(\theta) > -\infty$.
Step B: consider a sequence $(\dot{\theta}_m)_{m\in N}$. It needs to be proved that if $(\dot{\theta}_m)_{m\in N}$ leaves a suitably chosen compact set K_n, it cannot achieve as large values of l_n as one could find within K_n. Lemma 3 (Lemma 2 for $\delta = 0$) rules out the possibility of any $\lambda_{k,j,m} \searrow 0$.
Step C: (5) implies that $l_n(\theta)$ can be bounded from above in terms of $\pi_0, \lambda_{\min}$ and δ:

$$l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \left(\pi_0\delta + \sum_{j=1}^{G} \pi_j \phi(x_i; \mu_j, \Sigma_j) \right) \leq \log \left(\pi_0\delta + (1-\pi_0)(2\pi\lambda_{\min}(\theta))^{-\frac{p}{2}} \right).$$

Consider $\dot{\theta} \in \Theta_n$ such that $\dot{\lambda}_{k,j} < +\infty$ for all $k = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, G$ (using the obvious notation of components of "dotted" parameter vectors). Also consider a sequence $(\ddot{\theta}_m)_{m\in\mathbb{N}}$ such that $\ddot{\theta}_m \to \ddot{\theta}$ where $\ddot{\theta}$ is equal to $\dot{\theta}$ except that $\ddot{\lambda}_{k,j,m} \to +\infty$ for some $k \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, G\}$. By (6), $\lambda_{\min}(\ddot{\theta}_m) \to +\infty$ and thus $l_n(\ddot{\theta}_m) \to \log(\ddot{\pi}_0\delta)$. Clearly $\ddot{\pi}_0 < 1$ because otherwise $n^{-1}\sum_{i=1}^{n} \tau_0(x_i, \ddot{\theta}) = 1$, violating (8). Therefore $\lim_{m\to\infty} l_n(\ddot{\theta}_m) \leq l_n(\dot{\theta})$.
Step D: now consider $||\dot{\mu}_{j,m}|| \to +\infty$, for $j = 1$, w.l.o.g. Choose $\ddot{\theta}_m$ equal to $\dot{\theta}_m$ except now $\ddot{\mu}_{1,m} = 0$ for all m. Note that $\phi(x_i; \dot{\mu}_{1,m}, \dot{\Sigma}_1) \to 0$ for all $i = 1, 2, \ldots, n$, which implies that $\psi_\delta(x_i, \dot{\theta}_m) < \psi_\delta(x_i, \ddot{\theta}_m)$ for large enough m, for which then $l_n(\dot{\theta}_m) < l_n(\ddot{\theta}_m)$. Applying

this argument to all j with $||\dot{\mu}_{j,m}|| \to +\infty$ shows that better l_n can be achieved inside a compact set.

Continuity of l_n now guarantees existence of $\theta_n(\delta)$. ■

We now derive consistency for the sequence $\{\theta_n(\delta)\}_{n\in\mathbb{N}}$ as estimator of L_G. Consistency of the RIMLE can be achieved only if L_G exists. In order to ease the notation we define $\eta(x,\theta) = \sum_{j=1}^{G} \pi_j \phi(x;\mu_j,\Sigma_j)$. Consider the following assumptions on P:

A1 For every $x_1,\ldots,x_G \in \mathbb{R}^p : P\{x_1,\ldots,x_G\} < 1-\pi_{max}$.

A2 $L_G > L_{G-1}$, where for $G \in \mathbb{N}$ let $L_G = \sup_{\theta\in\Theta_G(P)} L(\theta)$.

A3 There exist $\epsilon_1,\epsilon_2 > 0$ so that for every θ with $\pi_0 < \epsilon_1 : L(\theta) \leq L_G - \epsilon_2$.

Remark 5 *Assumption A1 requires that no set of G points carries probability $1-\pi_{\max}$ or more. Otherwise the log-likelihood can be driven to ∞ by fitting G mixture components to G points with all covariance matrix eigenvalues converging to zero. The improper noise component could take care of all other points.*

Note that assumptions A2 and A3 are not both required, but only any single one of them. A2 states that G mixture components fit the data better than $G-1$ components. If this is not the case, there is at least one redundant component, and one cannot make sure that $L(\theta)$ is bounded away from L_G for large n in some distance from the "true" RIMLE-functional as the redundant component can be moved around, see Theorem 11. In case that A2 is not fulfilled, a weaker result can still be achieved, namely the existence of a not necessarily unique consistent sequence of local maximizers of l_n. This requires A3, which states that a noise proportion bounded away from zero is required for maximizing L. If neither A2 nor A3 are fulfilled, P can be fitted perfectly by fewer than G mixture components and no noise. In this case one cannot stop the remaining mixture components from leaving every compact set, and therefore one cannot expect consistency of all components for any method; as long as there is still noise bounded away from zero, those mixture components still can contribute to fitting what otherwise would be noise, and fits become worse if these degenerate.

Note that this is less often the case than one might expect; for example, a plain Gaussian mixture with $G-1$ components may still fulfill A3: if the density of one of the components is uniformly smaller than δ, a better pseudo-likelihood can obviously be achieved by assigning its proportion to the noise component than by choosing $\pi_0 = \pi_G = 0$ and otherwise the true parameters. A Gaussian component that "looks like noise" rather than like a "cluster" will be treated as noise.

Lemma 6 *For any probability measure P on $\mathbb{R}^p$, $L_G > -\infty$.*

Proof Choose compact $K \subset \mathbb{R}^p$ with $P(K) > 0$. Let $q = 1-P(K)$ and choose K big enough that $\pi_{\max} > q$. Choose $\mu_1 = \mathrm{E}_P(x|x\in K)$, $\Sigma_1 = \mathrm{Cov}_P(x|x\in K)$, $\pi_2 = \ldots = \pi_G = 0$. If all eigenvalues of Σ_1 are zero, choose $\Sigma_1 = I_p$. Let $\lambda_{\max,1}$ be the largest eigenvalue of Σ_1. If $\lambda_{\max,1}/\lambda_{i,1} > \gamma$ for any eigenvalue $\lambda_{i,1}$ of Σ_1, modify Σ_1 by replacing all eigenvalues smaller

than $\gamma\lambda_{\max,1}$ by $\gamma\lambda_{\max,1}$ in its spectral decomposition. Let $\phi_{\min} = \min_{x\in K}\phi(x,\mu_1,\Sigma_1) > 0$. Choose

$$\pi_0 = \frac{(\pi_{\max}-q)\phi_{\min}}{2((1-\pi_{\max})\delta + (\pi_{\max}-q)\phi_{\min})} > 0,\ \pi_1 = 1-\pi_0.$$

Observe that the resulting $\theta \in \Theta_G(P)$ (with all other parameters chosen arbitrarily) by

$$\begin{aligned}\mathrm{E}_P\frac{\pi_0\delta}{\psi_\delta(x,\theta)} &= \int_K \frac{\pi_0\delta}{\pi_0\delta+\eta(x,\theta)}dP(x) + \int_{K^c}\frac{\pi_0\delta}{\pi_0\delta+\eta(x,\theta)}dP(x) \leq \\ &\leq (1-q)\frac{\pi_0\delta}{\pi_0\delta+(1-\pi_0)\phi_{\min}} + q.\end{aligned}$$

This is smaller than $\pi_{\max}$ if $\pi_0 < \frac{(\pi_{\max}-q)\phi_{\min}}{(1-\pi_{\max})c+(\pi_{\max}-q)\phi_{\min}}$. Furthermore, $L(\theta) \geq (1-q)\log(\pi_0\delta+(1-\pi_0)\phi_{\min}) + q\log(\pi_0\delta) > -\infty$. ■

Lemma 7 *Assume A1. There are $\lambda^*_{\min} > 0,\ \lambda^*_{\max} < \infty,\ \epsilon > 0$, so that*

(a) *$L(\theta) \leq L_G - \epsilon$ for every θ with $\lambda_{\min}(\theta) < \lambda^*_{\min}$ or $\lambda_{\max}(\theta) > \lambda^*_{\max}$,*

(b) *for $x_1, x_2, \ldots$ i.i.d. with $\mathcal{L}(x_1) = P$, for sequences $(\theta_n)_{n\in\mathbb{N}}$ with $\lambda_{\min}(\theta_n) < \lambda^*_{\min}$ or $\lambda_{\max}(\theta_n) > \lambda^*_{\max}$ for large enough n: $l_n(\theta_n) \leq l_n(\theta_n(\delta)) - \epsilon$ a.s.*

Proof Start with part (a). First consider a sequence $\{\theta_m\}_{m\in\mathbb{N}} \in \Theta_G(P)^{\mathbb{N}}$ with $\lambda_{\max}(\theta_m) \to \infty$. The eigenvalue ratio constraint forces all covariance matrix eigenvalues to infinity, and therefore $\sup_x \phi(x,\mu_{j,m},\Sigma_{j,m}) \searrow 0$. But this means that $\mathrm{E}_P\frac{\pi_0\delta}{\psi_\delta(x,\theta)} \to 1 > \pi_{\max}$ and $\theta_m \notin \Theta_G(P)$ eventually, unless $\pi_{0,m} \searrow 0$, too. If the latter is the case, $\psi_\delta(x,\theta) \searrow 0$ uniformly over all x and $L(\theta_m) \searrow -\infty$, which together with Lemma 6 makes it impossible that $L(\theta_m)$ is close to L_G for m large enough and $\lambda_{\max}(\theta_m)$ too large, proving the existence of the upper bound $\lambda^*_{\max} < \infty$ as required.
Now consider a sequence $\{\theta_m\}_{m\in\mathbb{N}} \in \Theta_G(P)^{\mathbb{N}}$ with $\lambda_{\min}(\theta_m) \to 0$. Define

$$A_{m,\epsilon} = \left\{x : \min_{j=1,2,\ldots,G}\|x-\mu_{j,m}\| > \epsilon\right\}.$$

A1 ensures that for $0 < \epsilon_3$ there exists $\epsilon > 0$ so that for all $m \in \mathbb{N} : P(A_{m,\epsilon}) \geq \pi_{\max} + \epsilon_3$. Based on (5) derive an upper bound for $\pi_{0,m}$ from the constraint $\int \frac{\pi_{0,m}\delta}{\pi_{0,m}\delta+\eta(x,\theta_m)}dP(x) \leq \pi_{\max}$:

$$\begin{aligned}\int\frac{\pi_{0,m}\delta}{\pi_{0,m}\delta+\eta(x,\theta_m)}dP(x) &= \int_{A_{m,\epsilon}}\frac{\pi_{0,m}\delta}{\pi_{0,m}\delta+\eta(x,\theta_m)}dP(x) + \int_{A^c_{m,\epsilon}}\frac{\pi_{0,m}\delta}{\pi_{0,m}\delta+\eta(x,\theta_m)}dP(x) \geq \\ &\geq P(A_{m,\epsilon})\frac{\pi_{0,m}\delta}{\pi_{0,m}\delta+\max_{x\in A_{m,\epsilon}}\eta(x,\theta_m)},\end{aligned}$$

which by (5) implies

$$\pi_{0,m} \leq \frac{\pi_{\max}\max_{x\in A_{m,\epsilon}}\eta(x,\theta_m)}{\delta(P(A_{m,\epsilon})-\pi_{\max})} \leq \frac{\pi_{\max}\exp(-\frac{\epsilon^2}{2\gamma\lambda_{\min}(\theta_m)})}{\delta\epsilon_3(2\pi)^{p/2}\lambda_{\min}(\theta_m)^{p/2}}.$$

For the log-likelihood,

$$
\begin{aligned}
L(\theta_m) &= \int_{A_{m,\epsilon}} \log(\pi_{0,m}\delta + \eta(x,\theta_m))dP(x) + \int_{A^c_{m,\epsilon}} \log(\pi_{0,m}\delta + \eta(x,\theta_m))dP(x) \leq \\
&\leq \int_{A_{m,\epsilon}} \log\left(\frac{\delta\pi_{\max}\exp(-\frac{\epsilon^2}{2\gamma\lambda_{\min}(\theta_m)})}{\delta\epsilon_3(2\pi)^{p/2}\lambda_{\min}(\theta_m)^{p/2}} + \frac{\exp(-\frac{\epsilon^2}{2\gamma\lambda_{\min}(\theta_m)})}{(2\pi)^{p/2}\lambda_{\min}(\theta_m)^{p/2}}\right)dP(x) + \\
&\quad + \int_{A^c_{m,\epsilon}} \log\left(\frac{\delta\pi_{\max}\exp(-\frac{\epsilon^2}{2\gamma\lambda_{\min}(\theta_m)})}{\delta\epsilon_3(2\pi)^{p/2}\lambda_{\min}(\theta_m)^{p/2}} + \frac{1}{(2\pi)^{p/2}\lambda_{\min}(\theta_m)^{p/2}}\right)dP(x) \leq \\
&\leq P(A_{m,\epsilon})\left(-\frac{c_1}{\lambda_{\min}(\theta_m)} - c_2\log(\lambda_{\min}(\theta_m)) + c_3\right) + \\
&\quad + P(A^c_{m,\epsilon})\left(o(1) - c_4\log(\lambda_{\min}(\theta_m)) + c_5\right) = \\
&= -\frac{c_6}{\lambda_{\min}(\theta_m)} - c_7\log(\lambda_{\min}(\theta_m)) + c_8
\end{aligned}
$$

for positive constants c_1, c_2, c_4, c_6, c_7 and constants c_3, c_5, c_8, all independent of θ_m. If $\lambda_{\min}(\theta_m) \searrow 0$, this implies $L(\theta_m) \searrow -\infty$, proving together with Lemma 6 the existence of the lower bound $\lambda^*_{\min} > 0$.

Part (b) holds because if $(\theta_m)_{m\in\mathbb{N}}$ is chosen as above for $m = n \to \infty$ and P is replaced by the empirical distribution P_n, Glivenko-Cantelli enforces $P_n(A_{n,\epsilon}) - P(A_{n,\epsilon}) \to 0$ a.s. Glivenko-Cantelli applies because the class of all $A_{n,\epsilon}$ is a subset of the class of intersections of the complements of all closed balls, and therefore a Vapnik-Chervonenkis class, see van der Vaart and Wellner (1996). The argument carries over using all other integrals in the finite sample-form, i.e., w.r.t. P_n. Lemma 6 carries over because $l_n(\theta) \to L(\theta)$ a.s. by the strong law of large numbers for θ with $L(\theta) > -\infty$. ■

Remark 8 *Lemma 7 (a) and* (5) *imply that for all θ with $L(\theta) > L_G - \epsilon$, $j = 1, 2, \ldots, G$ and all x:*

$$\phi(x, \mu_j, \Sigma_j) \leq \phi_{\max} = \frac{1}{(2\pi)^{p/2}(\lambda^*_{\min})^{p/2}}.$$

This implies $L_G < \infty$.

The same holds because of Lemma 7 (b) for $x_1, x_2, \ldots$ i.i.d. with $\mathcal{L}(x_1) = P$ for large enough n a.s. for all θ with $l_n(\theta) > l_n(\theta_n(\delta)) - \epsilon$.

Lemma 9 *Assume A1 and A2. There is a compact set $K \subset \mathbb{R}^p$ so that*

(a) *L reaches its supremum for $\mu_1, \ldots, \mu_G \in K$ and is bounded away from the supremum if not all of $\mu_1, \ldots, \mu_G \in K$ (i.e., $\exists\epsilon_4 > 0$ so that L in this case is bounded from above by $\sup L - \epsilon_4$),*

(b) *for $x_1, x_2, \ldots$ i.i.d. with $\mathcal{L}(x_1) = P$ for large enough n, l_n reaches its supremum for $\mu_1, \ldots, \mu_G \in K$ and is bounded away from the supremum if not all of $\mu_1, \ldots, \mu_G \in K$, a.s.*

Proof Start with part (a). Consider a sequence $\{\theta_m\}_{m\in\mathbb{N}} \in \Theta_G(P)^{\mathbb{N}}$ with $\|\mu_{jm}\| \to \infty$ for $j = 1, \ldots, k,\ 1 \leq k < G$ and a compact set K with $\mu_{jm} \in K$ for $j > k$. Let

$$A_m = \left\{ x :\ \forall j \in \{1, \ldots, k\} : \phi(x, \mu_{j,m}, \Sigma_{j,m}) \leq \epsilon_m \sum_{l=k+1}^{G} \pi_{l,m} \phi(x, \mu_{l,m}, \Sigma_{l,m}) \right\},$$

where $\epsilon_m \searrow 0$ slowly enough that $P(A_m) \to 1$. Let $\pi^*_m = \sum_{j=1}^{k} \pi_{j,m}$. Let $\theta_{(G-k),m}$ for $m \in \mathbb{N}$ be defined by $\pi_{0,(G-k),m} = \pi_{0,m},\ \pi_{(j-k),(G-k),m} = \pi_{j,m}(1 - \pi_{0,m})(1 - \pi^*_m - \pi_{0,m})^{-1}$ for $j = k+1, \ldots, G$ accompanied by the μ_j, Σ_j-parameters belonging to the components $k+1, \ldots, G$ of θ_m. Observe, using Remark 8,

$$\begin{aligned} L(\theta_m) &= \int_{A_m} \log(\psi_\delta(x, \theta_m)) dP(x) + \int_{A_m^c} \log(\psi_\delta(x, \theta_m)) dP(x) \leq \\ &\leq \int_{A_m} \log((1+\epsilon_m)\psi_\delta(x, \theta_{(G-k)m})) dP(x) + P(A_m^c) \log(\delta + \phi_{max}) \end{aligned}$$

implying $L(\theta_{(G-k)m}) - L(\theta_m) \to 0$. $\mathrm{E}_P[\pi_{0,(G-k),m}\delta(\psi_\delta(x, \theta_{(G-k)m}))^{-1}] \leq \pi_{\max}$ will be fulfilled for m large enough because it is fulfilled for θ_m by definition and $\psi_\delta(x, \theta_{(G-k),m}) > \psi_\delta(x, \theta_m)$ on A_m with $P(A_m) \to 1$. $L(\theta_{(G-k)m}) < L_{G-1}$ implies that, because of A2, θ_m is bounded away from L_G.

Regarding existence of a maximum with $\mu_1, \ldots, \mu_G \in K$, observe that with Remark 8, $\psi_\delta(x, \theta)$ can be bounded by $\delta + \phi_{max}$ for all θ for which $L(\theta) > L_G - \epsilon$. Now consider a sequence $(\theta_m)_{m\in\mathbb{N}}$ so that $\forall m :\ \mu_{1m}, \ldots, \mu_{Gm} \in K$, with the notation of Lemma 7, $\lambda^*_{min} \leq \lambda_{min}(\theta_m) \leq \lambda_{max}(\theta_m) \leq \lambda^*_{max}$ and $L(\theta_m) \to L_G$. Because of compactness, w.l.o.g., $\theta_m \to \theta_+$ and, using Fatou's Lemma, $L_G = \lim_{m\to\infty} L(\theta_m) \leq E_P \limsup_m \psi_\delta(x, \theta_m) = L(\theta_+) \leq L_G$.

Part (b) holds because if $(\theta_m)_{m\in\mathbb{N}}$ is chosen as above for $m = n \to \infty$ and P is replaced by the empirical distribution P_n, Glivenko-Cantelli enforces $P_n(A_n) - P(A_n) \to 0$ a.s. Glivenko-Cantelli applies here because a sequence of closed balls $(B_n)_{n\in\mathbb{N}}$ can be constructed so that $B_n \subseteq A_n,\ P(B_n) \to 1$ a.s.; the closed balls are a Vapnik-Chervonenkis class, and $P_n(A_n) \geq P_n(B_n) \to 1$ a.s. Furthermore, for $\theta \in \Theta_G(P)$ with $L(\theta) > L_{G-1}$: $l_n(\theta) \to L(\theta)$ a.s. by the strong law of large numbers, so that for large enough n: $\sup_{\theta\in\Theta_G(P)} l_n(\theta) > L_{G-1}$ a.s. On the other hand, $\theta_{(G-k),n}$ can be chosen optimally in a compact set K because of Lemma 7, within which l_n converges uniformly to L a.s. (Theorem 2 in Jennrich (1969)), and therefore, $\limsup_{n\to\infty} l_n(\theta_{(G-k),n}) \leq L_{G-1}$. With these ingredients, the argument of part (a) carries over. ■

Lemma 10 *Assume A1 and A3. There is a compact set $K \subset \mathbb{R}^p$ so that*

(a) *L reaches its supremum for $\mu_1, \ldots, \mu_G \in K$,*

(b) *for $x_1, x_2, \ldots$ i.i.d. with $\mathcal{L}(x_1) = P$, there exists a sequence $(\tilde{\theta}_n)_{n\in\mathbb{N}}$ maximizing l_n locally for $\mu_1, \ldots, \mu_G \in K$ so that $l_n(\tilde{\theta}_n) \to L_G$ a.s., and a.s. there is no sequence $\theta_n \in \Theta_G(P)$ so that $\limsup_{n\to\infty} l_n(\theta_n) > L_G$.*

Proof Start with part (a). Consider a sequence $\{\theta_m\}_{m\in\mathbb{N}} \in \Theta_G(P)^{\mathbb{N}}$ with $\|\mu_{j,m}\| \to \infty$ for $j = 1,\ldots,k,\ 1 \leq k < G$ (the case $k = G$ is treated at the end), and a compact set K with $\mu_{j,m} \in K$ for $j > k$. By selecting a subsequence if necessary, assume that there exists $\mu_j = \lim_{m\to\infty} \mu_{j,m}$, $\Sigma_j = \lim_{m\to\infty} \Sigma_{j,m}$ for $k+1 \leq j \leq G$ and that $\pi_{j,m}$ converge for $j = 0,\ldots,G$. Let $j^* = \underset{k+1\leq j\leq G}{\arg\max}\, \mathrm{E}_P\phi(x,\mu_j,\Sigma_j)$. Suppose $L(\theta_m) \nearrow L_G$ monotonically and, by A3, assume $\pi_{0,m} > \epsilon_1$.

Consider first the case $\sum_{j=1}^k \pi_{j,m} \to \epsilon_3 > 0$. Construct another sequence $\{\theta_{*m}\}_{m\in\mathbb{N}} \in \Theta_G(P)^{\mathbb{N}}$ for which $\mu_{1,*m} = \ldots = \mu_{k,*m} = \mu_{j^*} \in K$, $\Sigma_{1,*m} = \ldots = \Sigma_{k,*m} = \Sigma_{j^*}$. All other parameters are the same as in θ_m. Let $A_m = \{x:\ \forall j \in \{1,\ldots,k\} : 2\phi(x,\mu_{j,m},\Sigma_{j,m}) \leq \phi(x,\mu_{j^*},\Sigma_{j^*})\}$. Observe $P(A_m) \to 1$. Now

$$L(\theta_{*m}) - L(\theta_m) = \int_{A_m} \log\frac{\psi_\delta(x,\theta_{*m})}{\psi_\delta(x,\theta_m)}dP(x) + \int_{A_m^c} \log\frac{\psi_\delta(x,\theta_{*m})}{\psi_\delta(x,\theta_m)}dP(x). \tag{14}$$

For large enough m,

$$\int_{A_m} \log\frac{\psi_\delta(x,\theta_{*m})}{\psi_\delta(x,\theta_m)}dP(x) \geq \epsilon_5 > 0,$$

whereas (using Remark 8)

$$\int_{A_m^c} \log\frac{\psi_\delta(x,\theta_{*m})}{\psi_\delta(x,\theta_m)}dP(x) \geq P(A_m^c)\log\frac{\epsilon_1\delta}{\delta+\phi_{\max}} \to 0. \tag{15}$$

Therefore $L(\theta_{*m}) - L(\theta_m) > 0$ for large enough m so that $L(\theta_m)$ is improved by a θ with $\mu_j \in K$ for $j = 1,2,\ldots,G$.

Consider now $\sum_{j=1}^k \pi_{jm} \to 0$. Construct another sequence $\{\theta_{*m}\}_{m\in\mathbb{N}} \in \Theta_G(P)^{\mathbb{N}}$ for which $\mu_{1,*m} = \ldots = \mu_{k,*m} = \mu_{j^*} \in K$, $\Sigma_{1,*m} = \ldots = \Sigma_{k,*m} = \Sigma_{j^*}$, $\pi_{1*m} = \ldots = \pi_{k*m} = 0$, $\pi_{0*m} = \sum_{j=0}^k \pi_{jm}$, all other parameters taken from θ_m. Set $A_m = \{x:\ \forall j \in \{1,\ldots,k\} : \phi(x,\mu_{jm},\Sigma_{jm}) < \delta\}$. Again $P(A_m) \to 1$. With this, (14) holds again. This time

$$\int_{A_m} \log\frac{\psi_\delta(x,\theta_{*m})}{\psi_\delta(x,\theta_m)}dP(x) > 0$$

and again (15).

Let $\theta_* = \lim_{m\to\infty}\theta_{*m}$ (this exists by construction). Continuity of L implies that $L(\theta_*) = L_G$ and therefore for all $m : L(\theta_*) \geq L(\theta_m)$. $\sum_{j=1}^k \pi_{j,m} \to 0$ is required here because $\pi_{0,*m} \geq \pi_{0,m}$ does not necessarily fulfill $\mathrm{E}_P\frac{\pi_{0,*m}c}{\psi_\delta(x,\theta_{*m})} \leq \pi_{\max}$, but $\lim_{m\to\infty}\pi_{0,*m} = \lim_{m\to\infty}\pi_{0,m}$ does.

Finally, consider $k = G$. With $A_{m,\epsilon} = \{x:\ \forall j \in \{1,\ldots,k\} : \phi(x,\mu_{j,m},\Sigma_{j,m}) < \delta\epsilon\}$, observe

$$\mathrm{E}_P\frac{\pi_{0,m}\delta}{\psi_\delta(x,\theta_m)} \geq P(A_{m,\epsilon})\frac{\pi_{0,m}\delta}{\pi_{0,m}(c+\epsilon)} > \pi_{\max},$$

for small enough ϵ and large enough m, violating for large m the corresponding constraint in $\Theta_G(P)$ as long as $\pi_{0,m}$ is bounded from below, as was assumed. Existence follows in the same way as in the proof of Lemma 9.

For part (b) let θ^* have $\mu_1^*,\ldots,\mu_G^* \in K$ and $L(\theta^*) = L_G$, which exists because of part (a) and Lemma 7, which ensures further that θ^* is in a compact K^*. Then the strong law

of large numbers yields $l_n(\theta^*) \to L_G$ a.s., and Theorem 2 of Jennrich (1969) implies that for all sequences $(\theta_n)_{n\in\mathbb{N}} \in (K^*)^{\mathbb{N}}$: $\limsup_{n\to\infty} l_n(\theta_n) \leq L_G$. This also holds for sequences $(\theta_n)_{n\in\mathbb{N}}$ that are eventually outside K^* because of part (a) of Lemma 7 and the proof of part (a) above, because if $(\theta_m)_{m\in\mathbb{N}}$ is chosen as above for $m = n \to \infty$ and P is replaced by the empirical distribution P_n, Glivenko-Cantelli (which applies by the same argument as used in the proof of Lemma 9) enforces $P_n(A_n) - P(A_n) \to 0$ a.s., which means that as in part (a), a.s., eventually $l_n(\theta_n)$ cannot converge to anything larger than L_G. ■

Theorem 11 (RIMLE existence) *Assume A1 and any one of A2 or A3. There is a compact subset $K \subset \Theta_G(P)$ so that there exists $\theta \in K$: $\infty > L(\theta) = L_G > -\infty$. Assuming A2, for $\theta \notin K$, $L(\theta)$ is bounded away from L_G.*

Proof Pieced together from Lemmas 6-9 parts (a) and Remark 8. ■

Theorem 11 establishes existence of the RIMLE functional

$$\theta^\star(\delta) = \underset{\theta\in\Theta_G(P)}{\arg\max}\, L(\theta). \tag{16}$$

Unfortunately neither $L(\theta)$ nor $l_n(\theta)$ can be expected to have a unique maximum. If we take the vector θ and we permute some of the triples (π_j, μ_j, Σ_j) we still obtain the same value for $L(\theta)$ and $l_n(\theta)$. This known as "label switching" in the mixture literature. There could be other causes for multiple maxima. Without strong restrictions on P, we cannot identify any specific source of multiple optima in the target function. Instead we show that asymptotically the sequence of estimators is close to some maximum of the pseudo-loglikelihood, which amounts to consistency of the RIMLE with respect to a quotient space topology identifying all loglikelihood maxima, as done in Redner (1981). By $\theta^\star(\delta)$ in (16) we mean any of the maximizer of $L(\theta)$. Define the sets

$$S(\dot\theta) = \left\{\theta \in \Theta_G(P) : \int \log\psi_\delta(x;\theta)dP(x) = \int \log\psi_\delta(x;\dot\theta)dP(x)\right\},$$

$$\mathcal{K}(\dot\theta,\varepsilon) = \left\{\theta \in \Theta_G(P) : \|\theta - \ddot\theta\| < \varepsilon \;\forall\; \ddot\theta \in S(\dot\theta)\right\}, \quad \text{for any} \quad \varepsilon > 0.$$

The following theorem makes a stronger statement assuming A2 than A3, because if A2 does not hold, the Gth mixture component is asymptotically not needed and cannot be controlled for finite n outside a compact set.

Theorem 12 (Consistency) *Assume A1 and A2. Then for every $\varepsilon > 0$ and every sequence of maximizers $\theta_n(\delta)$ of l_n:*

$$\lim_{n\to\infty} P\left\{\theta_n(\delta) \in \mathcal{K}(\theta^\star(\delta),\varepsilon)\right\} = 1.$$

Assuming A3 instead of A2, for every compact $K \supset \mathcal{K}(\theta^\star(\delta),\varepsilon)$ there exists a sequence of θ_n that maximize l_n locally in K so that

$$\lim_{n\to\infty} P\left\{\theta_n \in \mathcal{K}(\theta^\star(\delta),\varepsilon)\right\} = 1.$$

Proof Under A2, because of the parts (b) of the Lemmas 7 and 9 it can be assumed that there is a compact set K so that all $\theta_n(\delta) \in K$ for large enough n a.s. Under A3, considerations are restricted to K anyway.

Based on Theorem 11 and related Lemmas $|\log \psi_\delta(x,\theta)| \leq C$ for some finite constant C for all $\theta \in K$. Sufficient conditions for Theorem 2 in Jennrich (1969) are satisfied, which implies uniform convergence of $l_n(\theta)$, that is $\sup_{\theta\in K} |l_n(\theta) - L(\theta)| \to 0$ P-a.s. Based on the latter, and applying the same argument as in proof of Theorem 5.7 in van der Vaart (2000), it holds true that $L(\theta_n(\delta)) \to L(\theta^\star(\delta))$ P-a.s. By continuity of $L(\theta)$ and Theorem 11 we have that for every $\varepsilon > 0$ there exists a $\beta > 0$ such that $L(\theta^\star(\delta)) - L(\theta) > \beta$ for all $\theta \in K \setminus \mathcal{K}(\theta^\star(\delta), \varepsilon)$. Denote $(\Omega, \mathcal{A}, P)$ the probability space where the sample random variables are defined and consider the following events

$$A_n = \{\omega \in \Omega : \ \theta_n(\delta) \in K \setminus \mathcal{K}(\theta^\star(\delta), \varepsilon)\},$$

and

$$B_n = \{\omega \in \Omega : \ L(\theta^\star(\delta)) - L(\theta_n(\delta)) > \beta\}.$$

Clearly $A_n \subseteq B_n$ for all n. $P(B_n) \to 0$ for $n \to \infty$ implies $P(A_n) \to 0$. The latter proves the result. ■

5. Algorithms and practical issues

Following the presentation of the proposed algorithm to compute the RIMLE, we discuss its initialization and tuning.

5.1 RIMLE computing

In this section we develop Expectation–Maximization type algorithms (EM) to compute the RIMLE (for a fixed δ). Let $s = 0, 1, \ldots$ be the iteration index. Let $a^{(s+1)}$ be the quantity a computed at the sth step of the EM algorithm. Define

$$\begin{aligned} Q(\theta, \theta^{(s)}) = &\sum_{i=1}^{n}\sum_{j=0}^{G} \tau_j(x_i, \theta^{(s)}) \log \pi_j + \sum_{i=1}^{n} \tau_0(x_i, \theta^{(s)}) \log \delta + \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{G} \tau_j(x_i, \theta^{(s)}) \log \phi(x_i; \mu_j, \Sigma_j). \end{aligned} \tag{17}$$

Increasing (17) by an appropriate choice of θ increases $l_n(\cdot)$. An approximate candidate maximum of $l_n(\theta)$ can be found by the EM Algorithm 1.

Proposition 13 *Assume A0. The sequence $\{\theta^{(s)}\}_{s\in\mathbb{N}}$ produced by Algorithm 1 converges to a point $\theta_n^{em} \in \Theta$, and $l_n(\theta^{(s)})$ is increased in every step.*

Proof Find a set $A_n \subset \mathbb{R}^p$ that contains all points in $\underline{x_n}$ with Lebesgue measure $M(A_n) = 1/\delta$. δ is then a proper uniform density function on $\overline{A_n}$. Hence, for a given dataset the pseudo-density $\psi_\delta(\cdot)$ can be written as proper density function. Therefore the convergence

Algorithm 1: EM-algorithm

input : $\{x_1, x_2, \ldots, x_n\}$, δ, $\pi_{\max}$, $\gamma, \theta^{(0)}$, `tol¿0`
output : θ^{em}

while $|l_n(\theta^{(s+1)}) - l_n(\theta^{(s)})| >$`tol` **do**
  E–step: compute $\tau_j(x_i, \theta^{(s)})$, for all $i = 1, 2, \ldots, n$ and $j = 0, 1, \ldots, G$
  M–step: $\theta^{(s+1)} \leftarrow \arg\max_{\theta \in \Theta} Q(\theta, \theta^{(s)})$
end
$\theta^{\text{em}} \leftarrow \theta^{(s+1)}$

Theorem 4.1 in Redner and Walker (1984) holds, with $Q(\theta, \theta^{(s)})$ playing the role of their $Q(\cdot)$ function. ■

Algorithm 1 is the analog of the EM algorithm for plain Gaussian mixtures (see Redner and Walker, 1984) except that now the M-step is a constrained optimization. Wu (1983) showed that the EM algorithm converges to the global maximum if the likelihood function is unimodal and certain differentiability conditions are satisfied. In general the limit of the EM algorithm is not guaranteed to coincide with a global maximum of likelihood function. However, Proposition 13 guarantees that θ_n^{em} is a stationary point of $l_n(\cdot)$. Running the EM algorithm for a large number of starting values increases the chances of finding the optimal solution. For finite Gaussian mixtures models it is well known that the likelihood surface is difficult to explore even when p is not too large, and the main advantage of the EM algorithm is that the M-step can be divided in a number of simpler optimization problems each of which has a closed form solution. However, for the RIMLE the constraints add some complexity, and in particular the noise proportion constraint does not allow to separate the M-step in simpler subprograms. One possibility is to perform the M-step using numerical optimization packages, but the eigenratio constraints requires to parameterize each Σ_j terms of its spectral components. The latter has the drawback to add $G \times p(1-p)/2$ parameters. Furthermore, the eigenratio constraint has a non-smooth nature that would make numerical techniques hard to adapt.

In Coretto and Hennig (2016) computations are based on Algorithm 1 where the M-step is performed as if the problem would be unconstrained, and breaking the iteration when updates drive the parameters outside the constrained parameter space. Coretto and Hennig (2016) also propose a heuristic method to enforce the constraints at the end of the iterations if necessary. Of course in such situations there would be no guarantee that the delivered solution is a stationary point of $l_n(\cdot)$. Here, we propose Algorithm 2 where constraints are applied exactly in each iteration. The M-step in Algorithm 1 is replaced with two conditional maximization (CM) steps. This transforms Algorithm 1 into an Expectation-Conditional Maximization algorithm (ECM) as introduced by Meng and Rubin (1993). For ease of notation, for $j = 0, 1, \ldots, G$ define $\tau_{i,j}^{(s)} = \tau_j(x_i, \theta^{(s)})$ and $T_j^{(s)} = \sum_{i=1}^n \tau_j(x_i, \theta^{(s)})$. Rewrite (17), using $\theta_1 = (\mu_1, \ldots, \mu_G, \text{vect}(\Sigma_1), \ldots, \text{vect}(\Sigma_G))$, $\theta_2 = (\pi_0, \pi_1, \ldots, \pi_G)'$, and $\theta = (\theta_1, \theta_2)'$, as

$$Q(\theta_1, \theta_2, \theta^{(s)}) = Q_1(\theta_1, \theta^{(s)}) + Q_2(\theta_2, \theta^{(s)}) + \text{const}, \tag{18}$$

where

$$Q_1(\theta_1, \theta^{(s)}) = \sum_{i=1}^{n}\sum_{j=1}^{G} \tau_{i,j}^{(s)} \log \phi(x_i; \mu_j, \Sigma_j), \quad Q_2(\theta_2, \theta^{(s)}) = \sum_{j=0}^{G} T_j^{(s)} \log \pi_j,$$

and const $= T_0^{(s)} \log(\delta)$ which does not depend on θ. Consider the following programs:

$$\underset{\theta_1}{\text{maximize}}\ Q_1(\theta_1, \theta^{(s)}) \quad \text{subject to} \quad \frac{\lambda_{\max}(\theta_1)}{\lambda_{\min}(\theta_1)} \leq \gamma, \tag{\textbf{CM1}}$$

and

$$\underset{\theta_2}{\text{maximize}}\ Q_2(\theta_2, \theta^{(s)}) \quad \text{subject to} \quad \sum_{j=0}^{G} \pi_j = 1,\ \sum_{i=1}^{n} \tau_0(x_i, \theta_1^{(s+1)}, \theta_2) \leq n\pi_{\max}. \tag{\textbf{CM2}}$$

The ECM algorithm consists of solving (**CM1**) and then (**CM2**). The sequence of optimizations replaces the M-step in Algorithm 1. Notice that in (**CM2**), $\pi_j = 0$ for some $j = 0, 1, \ldots, G$ would drive the objective function toward $-\infty$, so we do not need to restrict the π_j as > 0. $T_j^{(s)} = 0$ will not happen, see Remark 17. Also notice that for δ=0 the noise proportion constraint is automatically fulfilled, and for more analysis on these cases see Remark 17.

Before presenting the ECM Algorithm 2 we introduce additional notations. For $a \in R^d$ let diag(a) be the $d \times d$ diagonal matrix with elements of a on the main diagonal. For a matrix A let Spec$(A) = \Gamma\Lambda\Gamma'$ be the spectral decomposition of A, that is, Γ contains the normalized eigenvectors of A corresponding to the eigenvalues contained in the diagonal matrix Λ. Moreover for $a, m \in \mathbb{R}$ define the shrinkage operator $\ell_\gamma(a, m) = \min\{\max\{m, a\}, \gamma m\}$.
In each step of Algorithm 2 closed form expressions are computed except that for computing m_* in (CM1), and ω_* in (CM2). m_* is the solution of a one-dimensional convex problem. The resulting updates for the eigenvalues almost coincide with those of TCLUST. For TCLUST, Fritz et al. (2013) show that their analogue of m_* can be computed by $2pG + 1$ evaluations of the objective function. A similar result may hold here, however we do not consider it because in numerical experiments we found that the simple golden section search algorithm of Kiefer (1953) requires on average few objective function evaluations independently of p and G. Computation of ω_* can be performed by a one-dimensional root finder algorithm. Both are simple problems that do not require much computational effort.

Some additional results are given to show how the CM1–step and the CM2–step solve (**CM1**) and (**CM2**) respectively.

Lemma 14 *Assume Algorithm 2 has been run for s iterations. The vector $\theta_1^{(s+1)} = (\mu_1^{(s+1)}, \ldots, \mu_G^{(s+1)}, \text{vect}(\Sigma_1^{(s+1)}), \ldots, \text{vect}(\Sigma_G^{(s+1)}))'$ computed in the CM1–step is the global optimal solution to* (**CM1**). *Moreover, m_* exists and it is unique.*

Proof Based on standard normal likelihood theory, one can see that the unique maximum of $Q_1(\theta_1, \theta^{(s)})$ with respect to mean parameters is $\mu_j^{(s+1)}$ for all $j = 1, 2, \ldots, G$. Substituting $\mu_j^{(s+1)}$ into $Q_1(\cdot)$, and rearranging the exponent of the Gaussian density by using the cyclic

Algorithm 2: ECM

input : $\{x_1, x_2, \ldots, x_n\}$, δ, $\pi_{\max}$, $\gamma, \theta^{(0)} \in \Theta$, tol¿0
output : θ^{ecm}

while $|l_n(\theta^{(s+1)}) - l_n(\theta^{(s)})| >$tol **do**

E–step

compute $\tau_{i,j}^{(s)}$ for all $i = 1, 2, \ldots, n$ and $j = 0, 1, \ldots, G$

CM1–step

for $j = 1, 2, \ldots, G$ **do**

$\mu_j^{(s+1)} \leftarrow \frac{1}{T_j^{(s)}} \sum_{i=1}^n \tau_{i,j}^{(s)} x_i$

$S_j^{(s+1)} \leftarrow \frac{1}{T_j^{(s)}} \sum_{i=1}^n \tau_{i,j}^{(s)} \ (x_i - \mu_j^{(s+1)})(x_i - \mu_j^{(s+1)})'$,

$V_j^{(s+1)} \operatorname{diag}(e_{j,1}, \ldots, e_{j,p}) V_j^{(s+1)\prime} \quad \leftarrow \quad \operatorname{Spec}(S_j^{(s+1)})$

end

if $\max\{e_{j,k}/e_{t,k}; \quad t, j = 1, 2, \ldots, G, \ j \neq t, \ k = 1, 2, \ldots, p\} \leq \gamma$ **then**

$\Sigma_j^{(s+1)} \leftarrow S_j^{(s+1)}$

else

$$m_* \leftarrow \arg\min_{m>0} \quad \sum_{j=1}^{G} T_j^{(s)} \sum_{k=1}^{p} \left(\log(\ell_\gamma(e_{j,k}, m)) + \frac{e_{j,k}}{\ell_\gamma(e_{j,k}, m)} \right)$$

$\Sigma_j^{(s+1)} \leftarrow V_j^{(s+1)} \operatorname{diag}(\ell_\gamma(e_{j,1}, m_*), \ldots, \ell_\gamma(e_{j,p}, m_*)) \ V_j^{(s+1)\prime}$ for all $j = 1, 2, \ldots, G$

end

CM2–step

if $\sum_{i=1}^n \tau_0(x_i, (\theta_1^{(s+1)}, \dot{\theta}_2)) \leq n\pi_{\max}$, *where* $\dot{\theta}_2 = \left(T_0^{(s)}/n, \ldots, T_G^{(s)}/n \right)'$ **then**

$\pi_j^{(s+1)} \leftarrow T_j^{(s)}/n$ for all $j = 0, 1, \ldots, G$

else

$$\text{compute } \omega_* : \quad \left(\sum_{i=1}^{n} \frac{\omega_* \delta}{\omega_* \delta + \frac{1-\omega_*}{n - T_0^{(s)}} \sum_{j=1}^{G} T_j^{(s)} \phi(x_i; \mu_j^{(s+1)}, \Sigma_j^{(s+1)})} \right) - n\pi_{\max} = 0$$

$\pi_0^{(s+1)} \leftarrow \quad \omega_*$

$\pi_j^{(s+1)} \leftarrow \quad \frac{1-\omega_*}{n - T_0^{(s)}} T_j^{(s)}$

end

end

$\theta^{\text{ecm}} \leftarrow \theta^{(s+1)}$

property of the matrix trace (see Anderson and Olkin, 1985), program (**CM1**) is completed by choosing $\Sigma_1, \ldots, \Sigma_G$ maximizing

$$Q_1(\dot{\theta}_1, \theta^{(s)}) = \text{const} - \frac{1}{2}\sum_{j=1}^{G} T_j^{(s)} \left(\text{tr}(\Sigma_j^{-1} S_j^{(s+1)}) - \log\det(\Sigma_j^{-1}) \right) \tag{19}$$

under the eigenratio constraint, where $\dot{\theta}_1 = (\mu_1^{(s+1)}, \ldots, \mu_j^{(s+1)}, \text{vect}(\Sigma_1), \ldots, \text{vect}(\Sigma_G))$. Consider the spectral decompositions

$$\text{Spec}(\Sigma_j) = \Gamma_j \Lambda_j \Gamma_j', \quad \text{and} \quad \text{Spec}(S_j^{(s+1)}) = V_j^{(s+1)} E_j V_j^{(s+1)\prime},$$

where $\Lambda_j = \text{diag}(\lambda_{j,1}, \ldots, \lambda_{j,p})$ and $E_j = \text{diag}(e_{j,1}, \ldots, e_{j,p})$. Theorem 1 and Corollary 1 in Theobald (1975) imply that

$$\text{tr}\left(\Lambda_j^{-1} E_j\right) - \log\det(\Lambda_j^{-1}) \leq \text{tr}\left(\Sigma_j^{-1} S_j^{(s+1)}\right) - \log\det(\Sigma_j^{-1}), \tag{20}$$

with the previous holding with equality if and only if $\Gamma_j = V_j^{(s+1)}$. Therefore, $\Sigma_j^{(s+1)} = V_j^{(s+1)} \Lambda_j V_j^{(s+1)\prime}$ is plugged into (19), and (**CM1**) reduces to

$$\begin{aligned} &\underset{\lambda_{1,1},\ldots,\lambda_{G,p}}{\text{minimize}} \quad \sum_{j=1}^{G} T_j^{(s)} \sum_{k=1}^{p} \left(\log(\lambda_{j,k}) + \frac{e_{j,k}}{\lambda_{j,k}} \right), \\ &\text{subject to} \quad 0 < m \leq \lambda_{j,1}, \ldots, \leq \lambda_{j,p} \leq m\gamma \quad \forall j = 1, 2, \ldots, G. \end{aligned} \tag{21}$$

Program (21) is separable in the optimization variables, and therefore the summands of (21) can be minimized separately for a given m. Fix $m > 0$, then $\ell_\gamma(e_{j,k}, m)$ is the unique optimal solution to the minimization of $\log(\lambda_{j,k}) + e_{j,k}\, \lambda_{j,k}^{-1}$. Notice that $e_{j,k} \leq e_{j,t} \Longrightarrow \ell_\gamma(e_{j,k}, m) \leq \ell_\gamma(e_{j,t}, m)$ for any $m > 0$ and $t = 1, 2 \ldots, p$. This means that the relative ordering of the elements on the diagonal of E_j remains unchanged after having applied the shrinkage operator $\ell(\cdot)$. Replace $\lambda_{j,k}$ with $\ell_\gamma(e_{j,k}, m)$ and (21) is transformed into

$$\begin{aligned} &\underset{m}{\text{minimize}} \quad \sum_{j=1}^{G} T_j^{(s)} \sum_{k=1}^{p} \left(\log(\ell_\gamma(e_{j,k}, m)) + \frac{e_{j,k}}{\ell_\gamma(e_{j,k}, m)} \right), \\ &\text{subject to} \quad m > 0 \end{aligned} \tag{22}$$

(22) is now a convex program in m. Therefore (**CM1**) is solved by the unique m_* that solves (22). This implies that the CM1–step is solved by taking $\Sigma_j^{(s+1)} = V_j^{(s+1)} E_j^* V_j^{(s+1)\prime}$ where $E_j^* = \text{diag}(\ell_\gamma(e_{j,1}, m_*), \ldots, \ell_\gamma(e_{j,p}, m_*))$. Notice that uniqueness of m_* implies the uniqueness of the solution to **CM1**. Observe that when the eigenvalues of $S_1^{(s+1)}, \ldots, S_G^{(s+1)}$ fulfill the eigenratio constraint, then $E_j^* = E_j$ and $\Sigma_j^{(s+1)} = S_j^{(s+1)}$ for all $j = 1, \ldots, G$. The latter completes the proof. (The result is connected to Lemma 1 in Won et al. (2013), by which the last part of the proof is inspired.)

■

Based on the previous Lemma, the constrained eigenvalues can be found by simply solving a convex one-dimensional problem. The optimal choice of the covariances is a form of Steinian-type nonlinear shrinkage (see Gavish and Donoho, 2017).

Lemma 15 *Assume Algorithm 2 has been run for s iterations. The vector* $\theta_2^{(s+1)} = (\pi_0^{(s+1)}, \ldots, \pi_G^{(s+1)})'$ *computed in the CM2–step is the global optimal solution to* (**CM2**). *Moreover,* ω_* *exists and it is unique.*

Proof The objective function in (**CM2**) is strictly concave and the equality constraint is linear. Take $\theta_2', \theta_2'' \in \{\theta_2 : \ \pi_0 + \ldots, +\pi_G = 1, \ \pi_j \in [0,1] \ \forall j = 0, 1, \ldots, G\}$, if $\pi_0' \leq \pi_0''$ then $\tau_0(\cdot, \cdot, \theta_2') \leq \tau_0(\cdot, \cdot, \theta_2'')$. Therefore, for $\beta \in (0,1)$

$$\sum_{i=1}^{n} \tau_0(x_i, \theta_1^{(s+1)}, \ \beta\theta_2' + (1-\beta)\theta_2'') \leq \max\left\{ \sum_{i=1}^{n} \tau_0(x_i, \theta_1^{(s+1)}, \theta_2'), \sum_{i=1}^{n} \tau_0(x_i, \theta_1^{(s+1)}, \theta_2'') \right\},$$

which implies that $\sum_{i=1}^{n} \tau_0(x_i, \theta_1^{(s+1)}, \theta_2)$ is quasiconvex in the optimization variable θ_2. It is concluded that Karush–Kuhn–Tucker (KKT) conditions are necessary for a global optimal solution (see Bertsekas, 1999). Such a solution will be a stationary point of the Langrangean function

$$H(\theta_2, h_1, h_2) := Q_2(\theta_2, \theta^{(s)}) + h_1 \left(1 - \sum_{j=0}^{G} \pi_j \right) + h_2 \left(n\pi_{\max} - \sum_{i=1}^{n} \tau_0(x_i, \theta_1^{(s+1)}, \theta_2) \right),$$

where h_1 and h_2 are the dual variables. Let ∇_j denote derivatives with respect to the j component of θ_2. Let $\theta_2^\star$ the optimal solution, then based on KKT conditions there exists $h_1^\star, h_2^\star$ such that the following hold

$$\nabla_j Q_2(\theta_2^\star, \theta^{(s)}) - h_1^\star - h_2^\star \nabla_j \sum_{i=1}^{n} \tau_0(x_i, \theta_1^{(s+1)}, \theta_2^\star) = 0 \qquad \text{for all} \quad j = 0, 1, \ldots, G, \tag{23}$$

$$h_2^\star \left(\sum_{i=1}^{n} \tau_0(x_i, \theta_1^{(s+1)}, \theta_2^\star) - n\pi_{\max} \right) = 0, \quad h_2^\star \geq 0. \tag{24}$$

First consider the case when the noise proportion constraint does not bind, that is $h_2^\star = 0$. Than (23) becomes $T_j^{(s)}/\pi_j^\star - h_1^\star = 0$ for all $j = 0, 1, \ldots, G$. Solving the latter for $\pi_j^\star$, using the equality constraints and that $\sum_{j=0}^{G} T_j^{(s)} = n$, it results that $h_1^\star = n$ and $\pi_j^\star = T_j^{(s)}/n$ for all $j = 0, 1, \ldots, G$.
Now assume that the noise proportion constraints binds, hence $h_2^\star > 0$. Let $\pi_0^\star = \omega$ and rewrite the equality constraint as $\pi_1^\star +, \ldots, +\pi_G^\star = 1 - \omega$. Stationary points of $H(\cdot)$ satisfy $T_j^{(s)}/\pi_j^\star - h_1^\star = 0$. Solving the latter for $\pi_j^\star$ and using the equality constraints it results that $h_1^\star = \sum_{j=1}^{G} T_j^{(s)}/(1-\omega)$. Since $\sum_{j=1}^{G} T_j^{(s)} = n - T_0^{(s)}$, then

$$\pi_j^\star = \frac{1-\omega}{n - T_0^{(s)}} T_j^{(s)} \quad \text{for all} \quad j = 1, 2, \ldots, G.$$

Now the solution for $j = 1, 2, \ldots, G$ is a function of ω, which can be determined by using the fact that the inequality constraints binds. Define

$$g(\omega) = \left(\sum_{i=1}^{n} \frac{\omega\delta}{\omega\delta + \frac{1-\omega}{n-T_0^{(s)}} \sum_{j=1}^{G} T_j^{(s)} \phi(x_i; \mu_j^{(s+1)}, \Sigma_j^{(s+1)})} \right) - n\pi_{\max}.$$

$g(\omega)$ is bracketed on the interval $[0, 1]$, in fact $g(0) = -n\pi_{\max} < 0$ and $g(1) = n(1 - \pi_{\max}) > 0$. Moreover $g(\omega)$ is continuous, and it can be easily verified that it's derivative is continuous and positive at any $\omega \in (0, 1)$. This implies that there exists a unique ω_* such that $g(\omega_*) = 0$. Setting $\pi_0^\star = \omega_*$ and replacing ω_* into $\pi_j^\star$ gives the optimal solution. We now compare the two solutions in terms of objective function, and we show that there is hierarchy between them. Define

$$\dot{\theta}_2 = \left(\frac{T_0^{(s)}}{n}, \frac{T_1^{(s)}}{n}, \ldots, \frac{T_1^{(s)}}{n} \right)', \qquad \ddot{\theta}_2 = \left(\omega_*, \frac{1-\omega_*}{n-T_0^{(s)}} T_1^{(s)}, \ldots, \frac{1-\omega_*}{n-T_0^{(s)}} T_G^{(s)} \right)'.$$

Using Wald's information inequality it can be shown that

$$\frac{T_0^{(s)}}{n} \log(\omega_*) + \sum_{j=1}^{G} \frac{T_j^{(s)}}{n} \log\left(\frac{1-\omega_*}{n-T_0^{(s)}} T_j^{(s)} \right) \leq \sum_{j=0}^{G} \frac{T_j^{(s)}}{n} \log\left(\frac{T_j^{(s)}}{n} \right),$$

with the previous holding with equality if and only if $\omega_* = T_0^{(s)}/n$. The latter implies that $Q_2(\ddot{\theta}_2, \theta^{(s)}) < Q_2(\dot{\theta}_2, \theta^{(s)})$ whenever $\ddot{\theta}_2 \neq \dot{\theta}_2$. Hence $\dot{\theta}_2$ is the global optimal solution whenever it is feasible, otherwise the global optimal solution is $\ddot{\theta}_2$. The latter proves that the updating in CM2-step selects the global optimal solution to (**CM2**). ■

Theorem 16 *Assume A0. The $\{\theta^{(s)}\} s \in \mathbb{N}$ produced by Algorithm 2 converges to a point $\theta_n^{ecm} \in \Theta$, and $l_n(\theta^{(s)})$ is increased in every step.*

Proof As consequence of Lemma 14 and 15, $Q(\theta, \theta^{(s)})$ is never decreased, in fact for all $s = 0, 1, \ldots$

$$Q(\theta_1^{(s+1)}, \theta_2^{(s+1)}, \theta^{(s)}) \geq Q(\theta_1^{(s+1)}, \theta_2^{(s)}, \theta^{(s)}) \geq Q(\theta_1^{(s)}, \theta_2^{(s)}, \theta^{(s)}),$$

A0 ensures existence of $\theta_n(\delta)$, and the convergence Theorem 4.1 in Redner and Walker (1984) holds with $Q(\theta, \theta^{(s)})$ playing the role of their $Q(\cdot)$ function. ■

Remark 17 *The eigenratio constraint together with the noise proportion constraint rule out the possibility that at some point along the iteration $T_j^{(s)} = 0$ for some $j = 1, 2, \ldots, G$ and updates in CM1-step are guaranteed to exist. In fact $T_j^{(s)} = 0$ means that according to $\theta^{(s-1)}$ none of points contributes to the jth Gaussian component. In theory this can only happen if the jth component has an infinite dispersion according to $\theta^{(s-1)}$. However, in*

that case the eigenratio constraint would force all eigenvalues in $\theta^{(s-1)}$ to diverge to $+\infty$ at the same rate so that $T_j^{(s)} \searrow 0$ for all $j = 1, 2, \ldots, G$, which is not possible because of the noise proportion constraint. Although in theory an appropriate choice of $\theta^{(0)} \in \Theta$ should not produce such a degeneracy, it may well be that in practice this is caused because of limited numerical resolution. Notice also that for δ=0 the noise proportion constraint is automatically fulfilled, and this would take the problem back to the EM algorithm for the MLE of a finite Gaussian mixture model with the additional eigenratio constraint. Therefore Algorithm 2 would become the EM Algorithm 1 where the M-step would coincide with CM1-step of Algorithm 2 plus the usual updating for the proportion parameters: $\pi_j^{(s+1)} \leftarrow T_j^{(s)}/n$ for all $j = 0, 1, \ldots, G$.

Remark 18 *There are substantial differences between Algorithm 2 and the one proposed in Coretto and Hennig (2016). Algorithm 2 implements the constraints of the RIMLE exactly and it is shown to converge under mild conditions on the data set. The algorithm presented in Coretto and Hennig (2016) handles the constraint heuristically as follows: (i) if the EM iteration converges, the eigenratio constraint is checked at the end of the iteration, and it is enforced simply increasing the eigenvalues smaller than $\lambda_{\max}(\theta)/\gamma$; (ii) if at some point along the EM iteration the estimated π_0 hits $\pi_{\max}$, the iteration is stopped and the solution discarded. Of course, this approximate algorithm is on average faster than Algorithm 2 for low values of γ and $\pi_{\max}$. In fact, in such a situation it is likely that Algorithm 2 needs to go through the additional computation of m_* and ω_* for most of the iterations. In order to give an idea to the reader, we estimated the relative speed for the AsyNoise data presented in Section 2 with $\gamma = 1, \pi_{\max} = 1\%$, and we found that on average the algorithm of Coretto and Hennig (2016) runs ten times faster. However, setting $\gamma = 100, \pi_{\max} = 50\%$ the difference becomes negligible. Furthermore, for low settings of $\pi_{\max}$ the algorithm of Coretto and Hennig (2016) is likely to record premature stops not leading to a solution (this happened in the previous experiment with $\pi_{\max} = 1\%$).*

5.2 Choice of initial values and input parameters

Algorithms 1 and 2 require the initial value $\theta^{(0)}$, and the input parameters $\pi_{\max}, \gamma$ and δ. The initial value $\theta^{(0)}$ can be set by randomly assigning points to G clusters and then computing cluster parameters. Initialization like this needs to be performed a number of times so that the solution with the largest pseudo-likelihood is selected. Implementation of the RIMLE given in the otrimle software of Coretto and Hennig (2017) relies on a more refined initialization strategy which consist in the following steps.

Initial denoising: for each data point compute its kth-nearest neighbors distance (k-NND), for some k. All points with k-NND larger than the $(1 - \pi_{\max})$-quantile of the k-NND are initialized as noise. The interpretation of k is that $(k-1)$, but not k, points close together may still be interpreted as noise/outliers, whereas k such points would constitute a cluster. The default value in the otrimle package is $k = 3$.

Initial clusters: agglomerative hierarchical clustering based on ML criteria for Gaussian mixture models as in Fraley (1998) is performed on the remaining $\lfloor n(1 - \pi_{\max}) \rfloor$ regular points to find the initial clusters. The sample mean and covariance matrix of

points belonging to each cluster are computed to define $\theta^{(0)}$. This step is performed based on the `hc()` function from the mclust package.

The constraint defining quantities $\pi_{\max}$ and γ are regularization parameters that allow solving an otherwise ill-posed optimization problem. $\pi_{\max}$ also controls robustness because it specifies the maximum proportion of points assignable to the noise component. In order to be as robust as possible $\pi_{\max} = 1/2$ is a convenient choice that guarantees maximum protection. This implements a familiar condition in robust statistics that at most half of the data should be classified as "outliers/noise". A choice of $\pi_{\max}$ lower than the actual noise/outlier proportion will enforce some outliers to be assigned to clusters with potentially problematic implications. Hence, unless one has prior knowledge about the contamination process, we suggested to stick to $\pi_{\max} = 1/2$.

The role of the eigenratio can be twofold. If γ is set to a low-value, strong restrictions on clusters' shape are imposed. In this respect, the eigenratio constraint acts as a model selector. Unless one knows precisely the implications of a low choice of γ, it is suggested to use the eigenratio constraint as a regularization parameter. In fact, a large value of γ will regularize the covariance matrices without affecting clusters' shape too much. For example, a large γ would allow discovering an elongated concentrated cluster along with clusters having widespread spherical scatters. Ritter (2014) contains an in-depth analysis of constraints in model-based clustering. In Section 7 we present Monte Carlo experiments where the effect of different γ values is investigated.

Although through the presence of the product $\pi_0\delta$ in (1) the parameters π_0 and δ may seem confounded, they actually play a very different role in the RIMLE. δ is not treated as a model parameter to be estimated, but rather as a tuning device to enable a good robust clustering. The interpretation is that δ is the density value below which groups of observations should rather be treated as "noise" than as "cluster". This means that a larger value of δ will normally yield a larger estimate of π_0 because more observations will be classified as noise, as opposed to the intuition suggested by having the product $\pi_0\delta$ in (1). Whether small groups of observations of a certain size and with a certain density peak should rather count as "cluster" or rather as "group of outliers" cannot be identified from the data alone, but is rather a matter of interpretation. RIMLE may be sensitive to the choice of δ, and a good choice of δ is therefore important in practice. For instance, in the example of Figure 2 it has been shown that outside a certain interval of δ values the RIMLE does not perform well. Occasionally, subject matter knowledge may be available aiding the choice of a fixed value of δ, but often such knowledge may not exist. The OTRIMLE, a data dependent method ("optimally tuned RIMLE") to choose δ is presented in Coretto and Hennig (2016). The basic idea is to find a δ that optimizes a weighted Kolmogorov-type distance measure between the Mahalanobis distances of all objects to their corresponding cluster centers and the χ^2-distribution, which the Mahalanobis distances should follow if the clusters were indeed Gaussian. The current implementation of the OTRIMLE in the otrimle package selects the best RIMLE solution computed with algorithm 2 on a selected grid of 50 values of $\log(\delta)$. The default grid includes $\log(\delta) = -\infty$ so that a pure Gaussian mixture is always included in the competition (see the otrimle manual for more details).

6. Breakdown robustness of the RIMLE

Although robustness results for some clustering methods can be found in the literature, robustness theory in cluster analysis remains a tricky issue. Some work exists on breakdown points (García-Escudero and Gordaliza, 1999; Hennig, 2004; Gallegos and Ritter, 2005), addressing whether parameters can diverge to infinity (or zero, for covariance eigenvalues and mixture proportions) under small modifications of the data. An addition breakdown point of $r/(n+r)$ means that r, but not $r-1$, points can be added to a data set of size n so that at least one of the parameters "breaks down" in the above sense.

It is well known (García-Escudero and Gordaliza, 1999; Hennig, 2008), assuming the fitted number of clusters to be fixed, that robustness in cluster analysis has to be data dependent, for the following reasons:

- If there are two not well separated clusters in the data set, a very small amount of "contamination" can merge them, freeing up a cluster to fit outliers converging to infinity.
- Very small clusters cannot be robust because a group of outlying points can legitimately be seen as a "cluster" and will compete for fit with non-outlying clusters of the same size. Noise component-based and trimming methods are prone to trimming whole clusters if they are small enough.

Therefore, all nontrivial breakdown results (i.e., with breakdown point larger than the minimum $1/(n+1)$) in clustering require a condition that makes sure that the clusters in the data set are strongly clustered in some sense, which usually means that the clusters are homogeneous and strongly separated.

The theory for the RIMLE given here generalizes the argument given in Hennig (2004), Theorem 4.11, to the multivariate setup. We consider fixed datasets $\underline{x_n} = (x_1, x_2, \ldots, x_n)$ and sequences of estimators $(E_n)_{n\in\mathbb{N}}$ mapping observations from $(\mathbb{R}^p)^n$ to Θ. Denote the components of $E_n(\underline{x_n})$ by
$(\pi_{En0}, \pi_{En1}, \ldots, \pi_{EnG}, \mu_{En1}, \ldots, \mu_{EnG}, \Sigma_{En1}, \ldots, \Sigma_{EnG})$, G being the number of mixture components as usually.

The following assumption in the definition of the breakdown point makes sure that $E_n(\underline{x_n})$ indeed parametrizes G different mixture components; if there was a mixture component with proportion zero or two equal ones, one mixture component would be free to be driven to breakdown.

A4 For $j = 1, \ldots, G : \pi_{Enj} > 0$, and all $(\mu_{Enj}, \Sigma_{Enj})$ are pairwise different.

Definition 19 *Assume that $(E_n)_{n\in\mathbb{N}}$ and $\underline{x_n}$ fulfil A4. Then,*

$$
\begin{aligned}
B(E_n, \underline{x_n}) \quad = \quad & \min_r \Bigg\{ \frac{r}{n+r} : \ \exists 1 \le j \le G \\
& \forall D = [\pi_{min}, 1] \times C \text{ for which } \pi_{min} > 0, C \subset \mathbb{R}^p \times \mathcal{S}_p \text{ compact} \\
& \exists \underline{x_{n+r}} = (x_1, x_2, \ldots, x_n, x_{n+1}, \ldots x_{n+r}) \text{ so that for} \\
& E_{n+g}(\underline{x_{n+g}}) : \ (\pi_{E(n+g)j}, \mu_{E(n+g)j}, \Sigma_{E(n+g)j}) \notin D \Bigg\},
\end{aligned}
$$

where $\mathcal{S}_p$ is the set of all positive definite real valued $p \times p$-matrices, is called the **breakdown point** *of E_n at dataset $\underline{x_n}$.*

Denote the sequence of RIMLE estimators defined in (3) as $(\theta_{mH})_{m\in\mathbb{N}}$, write $l_{mH}(\underline{x_m}, \theta)$ for $l_m(\theta)$ with any $m \in \mathbb{N}$ and number of components H in (2), $l^o_{mH} = l_{mH}(\underline{x_m}, \theta_{mH}(\underline{x_m}))$. Let $\theta^* = \theta_{nG}(\underline{x_n})$ for the specific $\underline{x_n}$ and G considered here. Components of θ^* and later θ^+ are denoted with upper index "$*$" and "$+$", respectively. For $j = 1, \ldots, G$, let $\phi^*_j(x) = \phi(x, \mu^*_j, \Sigma^*_j)$, same with upper index "$+$". Assume $\delta > 0$ fixed throughout this section. We start with a straightforward extension of Lemma 3.

Lemma 20 *Assume A0 for $\underline{x_n}$. If $(\theta_m)_{m\in N}$ is any sequence in Θ so that for some $j = 1, 2, \ldots, G$ and $k = 1, 2, \ldots, p$, $\lambda_{k,j,m} \searrow 0$ as $m \to \infty$. For $\underline{x_{n+r}} = (x_1, x_2, \ldots, x_n, x_{n+1}, \ldots x_{n+r})$: $\sup_{(x_{n+1},\ldots,x_{n+r})\in(\mathbb{R}^p)^r} l_{(n+r)G}(\underline{x_{n+r}}, \theta_m) \to -\infty$.*

Proof The proof of Lemma 3 still applies because adding further observations only adds further positive terms to the sum in (13). ■

Corollary 21 *Assume that $\underline{x_n}$ is a fixed dataset fulfilling A0 and A4 for $E_n = \theta_{nG}$. Then there is a $\lambda_0 > 0$ bounding from below all λ_{min} for $\theta = \theta_{(n+r)G}(\underline{x_{n+r}})$ where $\underline{x_{n+r}} = (x_1, x_2, \ldots, x_n, x_{n+1}, \ldots x_{n+r})$ for any $(x_{n+1}, \ldots, x_{n+r}) \in (\mathbb{R}^p)^r$. Consequently $\phi_{max} = (2\pi)^{-\frac{p}{2}}\lambda_0^{-\frac{p}{2}}$ is an upper bound for all $\phi(x; \mu, \Sigma)$ with (μ, Σ) occurring as component parameters in any such θ.*

Proof Observe

$$l_{(n+r)G}(\underline{x_{n+r}}, \theta^*) \geq \frac{1}{n+r}\left(\sum_{i=1}^{n} \log \psi_\delta(x_i, \theta^*) + r \log(\pi^*_0 \delta)\right) > -\infty. \tag{25}$$

If the Corollary was wrong, it would be possible to construct a sequence $(\theta_m)_{m\in N}$ with $\lambda_{k,j,m} \searrow 0$ for some $j = 1, 2, \ldots, G$ and $k = 1, 2, \ldots, p$ so that each $\theta_m = \theta_{(n+r)G}(\underline{x_{n+r}})$ for an admissible $\underline{x_{n+r}}$. But (25) implies that there is a lower bound for $l_{(n+r)G}(\underline{x_{n+r}}, \theta_m)$, contradicting Lemma 20. ■

The following theorem gives conditions under which the RIMLE estimator is breakdown robust against adding r observation to $\underline{x_n}$. (26) states that the dataset needs to be fitted by G Gaussian components considerably better than by $G-1$ components, because otherwise the remaining mixture component would be available for fitting the added observations without doing much damage to the original fit. (27) makes sure that the noise proportion in $\underline{x_n}$ is low enough that the added observations can still be fitted by the noise component without exceeding π_{max}.

Theorem 22 *Assume that $\underline{x_n}$ fulfils A0 and A4 for $E_n = \theta_{nG}$. If*

$$\begin{aligned} l^o_{n(G-1)} \quad < \quad & \sum_{i=1}^{n} \log\left(\sum_{j=1}^{G} \pi^*_j \phi^*_j(x_i) + \left(\pi^*_0 + \frac{r}{n}\right)\delta\right) + \\ & r \log\left(\left(\pi^*_0 + \frac{r}{n}\right)\delta\right) + (n+r)\log\frac{n}{n+r} - r\log\phi_{max}, \end{aligned} \tag{26}$$

ϕ_{max} defined in Corollary 21, and

$$\frac{1}{n+r}\left(\sum_{i=1}^{n}\frac{(n\pi_0^*+r)\delta}{(n+r)\psi_\delta(x_i,\theta^*)}+r\right)<\pi_{max}, \tag{27}$$

then $B(\theta_{nG},\underline{x_n})>\frac{r}{n+r}$.

Proof For $\underline{x_{n+r}}=(x_1,\dots x_{n+r})$, let $\theta^+=\theta_{(n+r)G}(\underline{x_{n+r}})$. Let $H<G$. Then,

$$l^o_{(n+r)G}\leq\sum_{i=1}^{n}\log\left(\sum_{j=1}^{H}\pi_j^+\phi_j^+(x_i)+\sum_{j=H+1}^{G}\pi_j^+\phi_j^+(x_i)+\pi_0^+\delta\right)+r\log\phi_{max}.$$

Assume w.l.o.g. that the parameter estimators of the mixture components $H+1,\dots,G$ leave a compact set D of the form $D=[\pi_{min},1]\times C,\ C\subset\mathbb{R}^p\times\mathcal{S}_p$ compact, $\pi_{min}>0$. Then there exists ϕ_{min} bounding $\phi_j^+(x_i)$ from below for $j=1,\dots,H$ and $i=1,\dots,n$, so $\sum_{j=1}^{H}\pi_j^+\phi_j^+(x_i)\geq H\pi_{min}\phi_{min}$.

Consider sequences $(\theta_m)_{m\in\mathbb{N}}\in\Theta$ with $l_{(n+r)G}(\underline{x_{n+r}},\theta_m)\to l^o_{(n+r)G}$ and leaving any D for $j=H+1\dots,G$, i.e., $\|\mu_{mj}\|\to\infty$ or $\lambda_{k,j,m}\to\infty$ or $\pi_{mj}\to0$, but with all $\lambda_{k,j,m}\geq\lambda_0$ as established in Corollary 21. Observe that for such sequences $\sum_{j=H+1}^{G}\pi_{mj}\phi_{mj}(x_i)$ becomes arbitrarily small for $i=1,\dots,n$. Thus, for arbitrary $\epsilon>0$ and D large enough:

$$\begin{aligned}l^o_{(n+r)G}&\leq\sum_{i=1}^{n}\log\left(\sum_{j=1}^{H}\pi_j^+\phi_j^+(x_i)+\pi_0^+\delta\right)+r\log\phi_{max}+\epsilon\\&\leq\max_{H<G}l^o_{nH}+r\log\phi_{max}+\epsilon\leq l^o_{n(G-1)}+r\log\phi_{max}+\epsilon.\end{aligned}$$

But a potential estimator $\hat\theta$ could be defined by $\hat\pi_0=\frac{n\pi_0^*+r}{n+r}$, $\hat\pi_j=\frac{n}{n+r}\pi_j^*$, $\hat\mu_j=\mu_j^*$, $\hat\Sigma_j=\Sigma_j^*$, $j=1,\dots,G$. Note that $\hat\theta\in\Theta$ because of (27). Therefore,

$$\begin{aligned}l^o_{(n+r)G}&\geq\sum_{i=1}^{n}\log\left(\sum_{j=1}^{G}\pi_j^*\phi_j^*(x_i)+\left(\pi_0^*+\frac{r}{n}\right)\delta\right)\\&\quad+r\log\left[\left(\pi_0^*+\frac{r}{n}\right)\delta\right]+(n+r)\log\frac{n}{n+r}\\\Rightarrow l^o_{n(G-1)}&\geq\sum_{i=1}^{n}\log\left(\sum_{j=1}^{G}\pi_j^*\phi_j^*(x_i)+\left(\pi_0^*+\frac{r}{n}\right)\delta\right)\\&\quad+r\log\left[\left(\pi_0^*+\frac{r}{n}\right)\delta\right]+(n+r)\log\frac{n}{n+r}-r\log\phi_{max}-\epsilon.\end{aligned}$$

This contradicts (26) by $\epsilon\to0$. ■

Table 1: Parameters of the AsyNoise sampling design. Let π and ν be the expected proportion and the degrees of freedom. m_1, v_1 and m_2, v_2 are the mean parameters (m) and variance parameters (v) along dimensions 1 and 2 respectively. $c_{1,2}$ denotes the covariance between marginals 1 and 2. All remaining variances are set equal to 1, while all remaining mean and covariance parameters are set equal to 0.

Parameter	Cluster				
	1	2	3	4	5
π	10.05%	20.10%	6.70%	10.05%	20.10%
ν	10	11	12	13	14
m_1	0	7	5	-11	-7
m_2	3	1	9	11	5
v_1	1	2	2	0.5	2.5
v_2	1	2	2	0.5	2.5
c_{12}	0.5	-1.5	1.3	0	0

7. Numerical experiments

In this section, we perform Monte Carlo experiments to compare robust clustering methods on the two sampling designs introduced in Section 2. There is already a comprehensive simulation study involving OTRIMLE and competitors in Coretto and Hennig (2016), so here we use different setups. Below, apart from involving competing methods from the literature, the two OTRIMLE algorithms are compared (see Remark 18).

The AsyNoise design of Figure 1 generates $G = 5$ clusters in $p = 20$ dimensions and an expected noise proportion of 33%. The five clusters are generated from a mixture of t-distributions with parameters given in Table 1. The five clusters show a combination of structures that are often difficult to handle together. Some of them are not well separated, they are of different size, and although they are all elliptically shaped, there are strong differences in cluster scatters, and deviations from normality. The noise originates from a distribution obtained as the product of two independent one-dimensional uniform distributions with support on the interval $[-25, 25]$, and 18 independent one-dimensional χ^2-distributions with 1 degree of freedom. The first and the third marginal are distributed uniformly, producing background noise on both clustered and non-clustered dimensions, and the χ^2-distribution adds a strong dose of asymmetry.

The second sampling design is called GEM (see Figure 3). In this case, the sampling design is a mixture of two Gaussian distributions in $p = 20$ dimensions, with the addition of a few potential outliers. In this design, the first cluster has strongly correlated marginals, whereas the second one is spherical, and this produces a large discrepancy between the clusters' shapes. Define the $p \times p$ correlation matrix $C(\rho) := (\rho^{|l-k|})_{l,k}$ for $l, k = 1, \ldots, p$ (also called AR(1) correlation model). The parameters of the GEM design are specified in Table 2. An expected 2% of points are generated from a 20-dimensional t-distribution with 3 degrees of freedom, centered at $(0, 0, -7, \ldots, -7)'$, with unit variances and correlation matrix $C(0.9999)$. This produces a few points far from both clusters, although these outliers

Table 2: Parameters of the GEM sampling design. Let π be the expected proportion. m is the mean parameter constant across all marginals for the same cluster. Each clusters has unit variance across all marginals and correlation matrix given by $C(\rho)$.

Parameter	Cluster	
	1	2
π	29.4%	68.6%
m	0	4
ρ	0.99	0

are not extremely separated from the regular data. While non-robust methods can cope with weakly separated outliers at the expense of large estimation bias, some robust methods capable of handling extreme outliers might get in trouble if the separation gap between regular and nonregular points is modest.

In this experiment, ML for Gaussian mixtures with uniform noise and TCLUST are compared with the RIMLE optimally tuned according to the OTRIMLE method introduced in Coretto and Hennig (2016). Methods under comparison are set up as follow:

OtrimleECM: RIMLE is computed based on the ECM algorithm 2 on a grid of 50 $\log(\delta))$ values as described in Section 5.2. The OTRIMLE criterion proposed in Coretto and Hennig (2016) selects the best solution. The input parameter $\pi_{\max}$ is always set to the conventional 50%. The eigenratio constraint is varied between the strongest restriction ($\gamma = 0$), and no restriction at all ($\gamma = +\infty$). In particular, $\log_{10}(\gamma) = \{0, 0.5, 1, 2, 3, 6, +\infty\}$. The initial partition is computed as described in Section 5.2. OtrimleECM is computed using the otrimle package of Coretto and Hennig (2017).

OtrimleAEM: RIMLE is computed using the approximate EM-algorithm introduced in Coretto and Hennig (2016). Both $\pi_{\max}$ and γ are set for OtrimleECM as well as initial values. Software for OtrimleAEM is available as part of the supplementary materials in Coretto and Hennig (2016).

TclustOracle: TCLUST with trimming rate set to the true underlying noise proportion. The eigenratio constraint is treated as for OtrimleECM. TclustOracle is computed using the tclust package of Fritz et al. (2012) which does not allow the user to choose an initial partition. TCLUST initialization is random, and we increased the default number of random starts to the sample size. The default maximum number of iterations is also increased to 500 because several convergence problems were recorded.

TclustFix: same as TclustOracle but with trimming rate fixed to a low 5% for the GEM design, and a high 50% for the AsyNoise design. The package tclust also includes the "ctlcurves"-tool for guiding the user toward the choice of a suitable trimming rate. Unfortunately, this graphical tool does not give any clear indication for the data sets generated in this experiment.

MCLUSTn: ML for Gaussian mixtures with uniform noise as implemented in the mclust package of Fraley et al. (2012). Regularization of the covariance matrices is done by choosing an appropriate covariance parameterization based on the BIC (Bayesian Information Criterion). Mclust requires noise initialization, and this is initialized as for the OtrimleECM. Note that the otrimle package uses mclust initialization for the regular points, hence OtrimleECM and MCLUSTn both start from the same partition.

There exist other methods capable of handling noise not considered here. The true clusters can be characterized by having a considerably higher density than the noise region, so density based clustering would seem to be another promising approach, but it suffers from the high dimensionality of the data, too. The DBSCAN algorithm of Ester et al. (1996) can handle noise, however its performance strongly depends on a pair of tunings that need to be carefully selected based on the dataset, and this makes it hardly comparable in a Monte Carlo experiment like the present one. ML based on t-mixtures of Peel and McLachlan (2000) would also be appropriate, however it requires discretionary decisions on how to define noise in terms of the tails of the estimated student-t distributions.

The sample size is set to $n = 500$ for AsyNoise, and $n = 100$ for GEM. With these relatively low sample sizes, the regularization of the covariance matrices becomes crucial because often small clusters are found compared to the dimensionality p. For both data sets 1000 Monte Carlo replicates have been considered. The true cluster label of a point is defined based on the component of the sampling distribution that generates it. Misclassification rates are computed with respect to the minimizing permutation of clusters' indexes not involving the estimated noise, which is always matched to the true noise. The underlying eigenratio behavior of these designs is largely varying. The true γ is 7 for AsyNoise, and it is 3704.7 for GEM. However, if one computes the eigenratio of sample clusters' covariances based on true labels the figure can be completely different. In fact, we computed the (5%, 95%)-quantiles of the Monte Carlo distribution of these quantities, and we obtained (44.5, 273.4) for AsyNoise, and (19899.3, 246826.8) for GEM. In the examples given in Section 2 we fixed $\gamma = 100$, because in real world applications one typically does not have information on it, and we used the central value adopted in these experiments.

Results are summarized in Tables 3 and 4, and Figures 5 and 6. Since MCLUSTn does not enforce an eigenratio constraint, results are recorded at $\gamma = +\infty$, although MCLUSTn has its own covariance regularization. MCLUSTn is seriously affected by contamination in both designs. Its performance is better for the AsyNoise design for which the boxplot of Figure 5 shows that in some replica it can produce misclassification rates below 10%. Regarding the Otrimle and Tclust versions, the performance depends on the setting of the eigenratio constraint. However, OtrimleECM offers the most stable performance in both designs. OtrimleECM achieves the best misclassification performance in all situations except for few cases where TclustOracle does better, but in fact, TclustOracle is run with the assumption that one knows the expected amount of noise exactly, which is never true in reality. Note that TclustOracle seems to not tolerate large values of γ in both designs. This is counterintuitive at least for GEM, where the true $\log_{10}(\gamma)$ is between 3 and 6, but in this range both TclustOracle and TclustFix have serious problems. OtrimleAEM is the second best overall, although it shows a large positive skewness in the distribution of the misclassification rates for AsyNoise, and in both designs it is not able to enforce low γ values appropriately. This is because in OtrimleAEM an approximate eigenratio constraint

Table 3: Monte Carlo averages, with standard errors in parenthesis, of misclassification rates (%) for the AsyNoise sampling design. "na" is reported if the software did not produce a valid answer in more than 50% of the replicates.

$\log_{10}(\gamma)$	OtrimleECM	OtrimleAEM	TclustOracle	TclustFix	MCLUSTn
0	15.31(0.00)	31.96(0.03)	4.33(0.00)	18.63(0.00)	—
0.5	11.25(0.01)	25.57(0.03)	5.31(0.00)	18.65(0.00)	—
1	9.46(0.00)	22.14(0.02)	16.52(0.00)	23.30(0.00)	—
2	11.48(0.00)	19.40(0.01)	50.26(0.00)	48.66(0.00)	—
3	12.37(0.01)	19.71(0.01)	57.88(0.00)	56.72(0.00)	—
6	12.05(0.01)	19.90(0.01)	57.26(0.00)	56.89(0.00)	—
$+\infty$	12.08(0.00)	19.92(0.01)	na	na	27.05(0.02)

Table 4: Monte Carlo averages, with standard errors in parenthesis, of misclassification rates (%) for the GEM sampling design. "na" is reported if the software did not produce a valid answer in more than 50% of the replicates.

$\log_{10}(\gamma)$	OtrimleECM	OtrimleAEM	TclustOracle	TclustFix	MCLUSTn
0	1.10(0.00)	63.97(0.04)	1.78(0.00)	3.32(0.00)	—
0.5	0.87(0.00)	11.96(0.03)	1.50(0.00)	3.17(0.00)	—
1	1.57(0.01)	3.49(0.01)	1.25(0.00)	3.11(0.00)	—
2	0.52(0.00)	2.25(0.00)	8.10(0.01)	9.68(0.01)	—
3	0.50(0.00)	0.71(0.00)	16.41(0.00)	18.08(0.00)	—
6	3.82(0.01)	1.18(0.01)	14.07(0.00)	16.80(0.00)	—
$+\infty$	4.66(0.01)	1.17(0.01)	na	na	41.60(0.01)

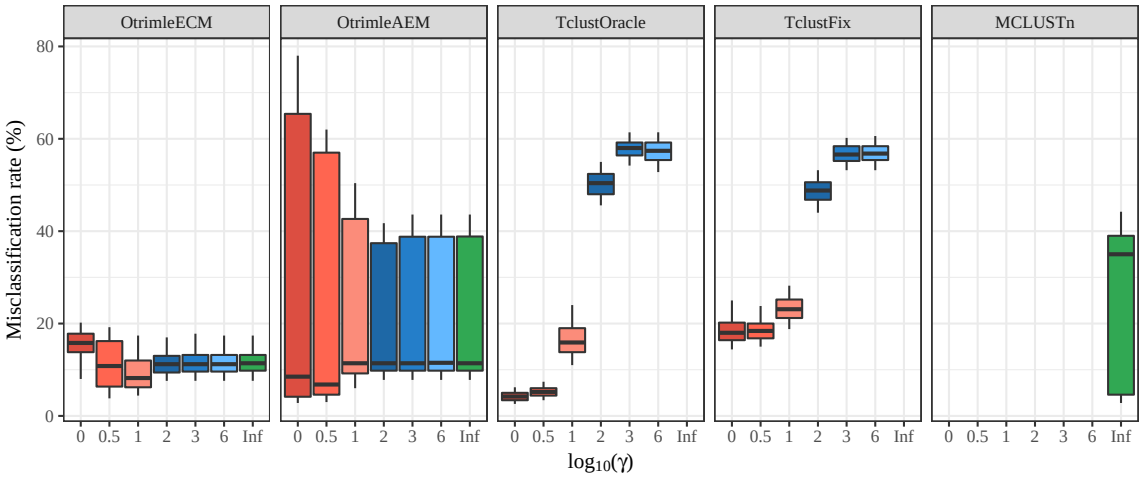

Figure 5: Modified boxplots of the Monte Carlo distribution of misclassification rates for the AsyNoise design: whiskers coincide with (5%, 95%)–quantiles of the distribution.

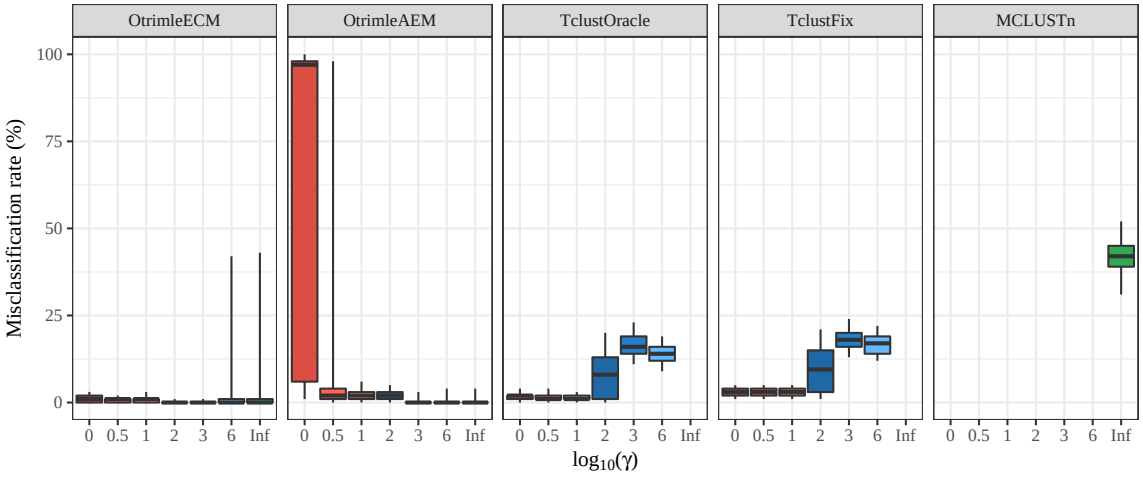

Figure 6: Modified boxplots of the Monte Carlo distribution of misclassification rates for the GEM design: whiskers coincide with (5%, 95%)–quantiles of the distribution.

is applied at the end of the EM iteration (see Remark 18). These results show a remarkable improvement of the ECM algorithm 2 over the approximate solution proposed in Coretto and Hennig (2016).

In practice the user has to specify γ. According to the results shown here, OtrimleECM is not very sensitive to this choice. Also the results show that good misclassification rates can be achieved in GEM, with a true $\gamma > 3000$, using a much lower γ for OTRIMLE; actually for TCLUST a much lower γ is even required to achieve good results. Choosing a lower γ in such situations may provide some welcome regularization. $\gamma = 100$ often seems to be a sensible choice. However, the user needs to have in mind that a straight interpretation of γ requires that a variance of 1 (say) along a one-dimensional projection has the same meaning in all directions in data space, which is particularly doubtful if variables have different measurement units or variable-wise variations are not meaningfully comparable. In such cases sphering or at least variable standardization may be advisable.

8. Concluding Remarks

The RIMLE robustifies the MLE in the Gaussian mixture model by adding an improper constant mixture component to catch outliers and points that cannot appropriately assigned to any cluster. Characteristics of the method compared to other robust clustering methods aiming for approximately Gaussian clusters are a smooth mixture-type transition between clusters and noise, and the fact that noise and outliers are not modelled by a specific and usually misspecified distribution, but rather as anything where the estimated mixture density is so low that the observation is rather classified to the constant noise than to any mixture component. If needed, the density value of the improper constant noise component can be chosen in a data-adaptive based on the OTRIMLE criterion developed in (Coretto and Hennig, 2016). The RIMLE/OTRIMLE has shown competitive performance when compared with state of the art methods for robust model-based clustering methods. In this paper we investigated theoretical properties of the RIMLE, and it is shown existence, consistency, breakdown behaviour, and convergence of algorithms. Since the RIMLE coincides with the MLE for Gaussian finite mixture models (when $\delta = 0$), the present paper also gives a comprehensive treatment for it which was missing in the literature.

References

Grigory Alexandrovich. A note on the article 'Inference for multivariate normal mixtures' by J. Chen and X. Tan. *Journal of Multivariate Analysis*, 129:245–248, 2014.

Theodore Wilbur Anderson and Ingram Olkin. Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear Algebra and its Applications*, 70: 147–171, 1985.

Jeffrey D. Banfield and Adrian E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49(3):803, sep 1993. doi: 10.2307/2532201.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

Gilles Celeux and Gérard Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.

Jiahua Chen and Xianming Tan. Inference for multivariate normal mixtures. *J. Multivariate Anal.*, 100(7):1367–1383, 2009.

Pietro Coretto and Christian Hennig. A simulation study to compare robust clustering methods based on mixtures. *Advances in Data Analysis and Classification*, 4(2):111–135, 2010.

Pietro Coretto and Christian Hennig. Maximum likelihood estimation of heterogeneous mixtures of gaussian and uniform distributions. *Journal of Statistical Planning and Inference*, 141(1):462–473, 2011.

Pietro Coretto and Christian Hennig. Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust Gaussian clustering. *Journal of the American Statistical Association*, 111:1648–1659, 2016.

Pietro Coretto and Christian Hennig. otrimle: Robust model-based clustering, 2017. R package version 1.1. Available at: https://CRAN.R-project.org/package=otrimle.

Juan Antonio Cuesta-Albertos, Alfonso Gordaliza, and Carlos Matrán. Trimmed k-means: an attempt to robustify quantizers. *The Annals of Statistics*, 25(2):553–576, 1997.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. Institute for Computer Science, University of Munich, 1996.

Chris Fraley. Algorithms for model-based gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20(1):270–281, 1998.

Chris Fraley and Adrian E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

Chris Fraley, Adrian E. Raftery, Thomas Brendan Murphy, and Luca Scrucca. mclust version 4 for r: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report 597, University of Washington, Department of Statistics, 2012.

Heinrich Fritz, Luis A. García-Escudero, and Agustín Mayo-Iscar. tclust: An R package for a trimming approach to cluster analysis. *Journal of Statistical Software*, 47(12):1–26, 2012.

Heinrich Fritz, Luis A. García-Escudero, and Agustín Mayo-Iscar. A fast algorithm for robust constrained clustering. *Comput. Statist. Data Anal.*, 61:124–136, 2013. ISSN 0167-9473.

María Teresa Gallegos and Gunter Ritter. A robust method for cluster analysis. *The Annals of Statistics*, 33(1):347–380, 2005.

María Teresa Gallegos and Gunter Ritter. Trimmed ML estimation of contaminated mixtures. *Sankhyā. The Indian Journal of Statistics*, 71(2, Ser. A):164–220, 2009. ISSN 0972-7671.

María Teresa Gallegos. Maximum likelihood clustering with outliers. In *Classification, Clustering, and Data Analysis*, pages 247–255. Springer, 2002.

María Teresa Gallegos Gallegos and Gunter Ritter. Strong consistency of k-parameters clustering. *Journal of Multivariate Analysis*, 117:14–31, 2013.

Luis Angel García-Escudero and Alfonso Gordaliza. Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association*, 94(447):956–969, 1999.

Luis Angel García-Escudero, Alfonso Gordaliza, Carlos Matrán, and Agustin Mayo-Iscar. A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 36(3): 1324–1345, 2008.

Luis Angel García-Escudero, Alfonso Gordaliza, Carlos Matrán, and Agustin Mayo-Iscar. Avoiding spurious local maximizers in mixture modeling. *Statistics and Computing*, 25: 1–15, 2014.

Luis Angel García-Escudero, Alfonso Gordaliza, Carlos Matrán, Agustin Mayo-Iscar, and Christian Hennig. Robustness and outliers. In Christian Hennig, Marina Meila, Fionn Murtagh, and Roberto Rocci, editors, *Handbook of Cluster Analysis*, pages 653–678. CRC Press, Boca Raton FL, 2015.

Matan Gavish and David L. Donoho. Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, 63(4):2137–2152, 2017.

Richard J. Hathaway. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13(2):795–800, 1985.

Christian Hennig. Breakdown points for maximum likelihood estimators of location?scale mixtures. *The Annals of Statistics*, 32(4):1313–1340, 2004.

Christian Hennig. Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods. *Journal of Multivariate Analysis*, 99(6):1154–1176, 2008.

Christian Hennig. Clustering strategy and method selection. In C. Hennig, M. Meila, F. Murtagh, and R. Rocci, editors, *Handbook of Cluster Analysis*, chapter 31, pages 703–730. Chapman & Hall/CRC, Boca Raton FL, 2015a.

Christian Hennig. What are the true clusters? *Pattern Recognition Letters*, 64:53–62, 2015b.

Christian Hennig and Tim F. Liao. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3):309–369, 2013.

Salvatore Ingrassia. A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods & Applications*, 13(2):151–166, 2004.

Salvatore Ingrassia and Roberto Rocci. Constrained monotone EM algorithms for finite mixture of multivariate gaussians. *Computational Statistics & Data Analysis*, 51(11): 5339–5351, 2007.

Salvatore Ingrassia and Roberto Rocci. Degeneracy of the EM algorithm for the MLE of multivariate gaussian mixtures and dynamic constraints. *Computational Statistics & Data Analysis*, 55(4):1715–1725, 2011.

Robert I. Jennrich. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643, 1969.

John E. Dennis Jr. Algorithms for nonlinear fitting. In M. J. D. Powell, editor, *Proceedings of the NATO Advanced Research Institute on "Nonlinear Optimization", held at Trinity Hall, Cambridge*, NATO Conference Series. Series II: Systems Science, London, 1981. Academic Press in cooperation with NATO Scientific Affairs Division.

Jack Kiefer. Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society*, 4(3):502–502, 1953.

Jack Kiefer and Jacob Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27(4):887–906, 1956.

Xiao-Li Meng and Donald B. Rubin. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80(2):267–278, 1993.

David Peel and Geoffrey John McLachlan. Robust mixture modelling using the t–distribution. *Statistics and Computing*, 10(4):339–348, 2000.

Richard Redner. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics*, 9(1):225–228, 1981.

Richard Redner and Homer F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

Gunter Ritter. *Robust Cluster Analysis and Variable Selection*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, 2014.

Chris M. Theobald. An inequality with application to multivariate analysis. *Biometrika*, 62(2):461–466, 1975.

Aad van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

Aad van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

Joong-Ho Won, Johan Lim, Seung-Jean Kim, and Bala Rajaratnam. Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society, Series B*, 75 (3):427–450, 2013.

Chien-Fu Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.