NONPARAMETRIC CONSISTENCY FOR MAXIMUM LIKELIHOOD ESTIMATION AND CLUSTERING BASED ON MIXTURES OF ELLIPTICALLY-SYMMETRIC DISTRIBUTIONS

Pietro Coretto* Christian Hennig†

Monday 29th April, 2024

Abstract

The consistency of the maximum likelihood estimator for mixtures of elliptically-symmetric distributions for estimating its population version is shown, where the underlying distribution P is nonparametric and does not necessarily belong to the class of mixtures on which the estimator is based. In a situation where P is a mixture of well enough separated but nonparametric distributions it is shown that the components of the population version of the estimator correspond to the well separated components of P. This provides some theoretical justification for the use of such estimators for cluster analysis in case that P has well separated subpopulations even if these subpopulations differ from what the mixture model assumes.

Keywords. Asymptotic analysis, finite mixture models, model-based clustering, canonical functional.

1. Introduction

Clustering methods are widely used in statistics, computer science, and many applied fields of science to discover group structures and to deal with the intrinsic inhomogeneity of complex data sets. While there is abundant work on methodology, algorithms, and applications, a smaller body of literature has investigated the relationship between the clusters found by a method and the underlying data-generating mechanism. Assuming that the observed data set is generated by independent and identical observations from a probability law P, consistency concerns the relationship between P and the outcome of a method for random samples of a size converging to infinity. In cluster analysis, the clustering itself and/or distributional parameters characterising the clustering may be of interest.

Here we will derive consistency results for model-based clustering, i.e., clustering based on probability mixture models. More precisely, the results will concern maximum likelihood (ML) estimators (MLE) of finite mixtures of distributions from elliptically symmetrical distribution (ESD) families such as the Gaussian distribution. Finite mixture models (FMM) are convex combinations of probability distributions suitable

*Departiment of Economics and Statistics; University of Salerno (Italy) – E-mail: pcoretto@unisa.it.
For this research the author was supported by Ministero dell'Università e della Ricerca (MUR, Italy), grant number 20223725WE.
†Department of Statistics; University of Bologna (Italy) – E-mail: christian.hennig@unibo.it

to represent inhomogeneous populations. FMMs are a versatile tool for modelling complex distributions, and are at the basis of a variety of data analysis, prediction, and inference tasks: density estimation, regression, clustering, and classification (see Frühwirth-Schnatter et al., 2019; Hennig et al., 2016).

Given K, the number of mixture components, FMMs are represented by parameters characterising the individual mixture components (such as Gaussian mean and variance) and the component proportions. MLEs can be obtained in practice using the Expectation-Maximisation (EM) algorithm of Dempster et al. (1977). Given the fitted parameters, a partition of the data can be obtained by use of the maximum posterior assignment rule, see Section 2.1.

In clustering, clusters are usually interpreted as corresponding to the estimated mixture components, although this is not always appropriate if different mixture components are not well separated, see Hennig (2010).

FMMs are parametric, and therefore, as a standard, statisticians are interested in whether the parameters can be consistently estimated if the true underlying P is indeed such a mixture. Consistency theory for ML estimation of mixture parameters is not as easy to obtain as for standard ML estimation, and standard conditions such as required by Wald (1949) are not fulfilled. Kiefer and Wolfowitz (1956) noted that the likelihood function of univariate Gaussian mixtures is unbounded if one of the mixture components' variances is allowed to converge to zero from above. This issue occurs for many classes of FMMs including those of ESDs as treated here. There have been several proposals to solve it. A popular strategy, originally due to Dennis (1981), is to impose an "eigen-ratio constraint" (ERC). This is a constant that bounds the ratio between the largest and smallest eigenvalue across all components' scatter matrices. The ERC allows for a non-compact parameter space; it is easier but also more restrictive to constrain the parameter space to be compact as done, e.g., by Redner (1981), who proved consistency for the MLE of a general class of FMMs. Hathaway (1985) expanded the work of Redner (1981) including the ERC in the case of a univariate Gaussian mixture. There is a long history of ML-theory for mixtures. For reviews see Redner and Walker (1984); Chen (2017).

Here, however, we focus on nonparametric consistency, i.e., we ask what happens if P is general and not necessarily of the assumed mixture type. Note that we will treat the number of estimated mixture components K as fixed, so that the fitted mixture cannot approximate almost any P by selecting a sufficiently large K. The results will apply to a P that is indeed a mixture of the assumed family with K mixture components, however the underlying P could also have fewer or more than K mixture components of this kind (in clustering it may sometimes be desirable to fit a mixture with more than K not well separated components by a K-component mixture where the K components are better separated and justified to be interpreted as clusters, see, e.g., Biernacki et al. (2000)), or it may not be possible to represent it as a mixture of the assumed type at all.

In practice, arguably, parametric statistical model assumptions are never precisely fulfilled, yet often it is claimed that the application of data analytic methods based on parametric models require the parametric model assumptions to be fulfilled. In fact often standard theory relies on these model assumptions, and it may be suspected that if assumptions are violated, the method may not achieve the performance that is theoretically guaranteed assuming the model. Nonparametric theory as derived here, even if only asymptotic, will apply to and justify the use of such methods more generally.

Nonparametric consistency results for cluster analysis go back to the seminal works

of Pollard (1981) (for K-means clustering) and Hartigan (1981) (for single linkage). Further such results have been provided by Sriperumbudur and Steinwart (2012) for the DBSCAN algorithm and von Luxburg et al. (2008) for spectral clustering, among others. Multivariate Gaussian FMMs have been studied by García-Escudero et al. (2015) and Coretto and Hennig (2017). The latter paper also has a nonparametric consistency result regarding a robust ML-type method accommodating outliers by incorporating an improper uniform mixture component covering the whole data space, and an Expectation-Conditional Maximization (ECM) algorithm implementing the ERC.

There are two aspects of nonparametric consistency relevant in clustering (cp. von Luxburg et al. (2008)).

(i) Is the clustering method consistent at all, i.e., is there a limit clustering (or limit parameters characterising it) if the sample size grows larger and larger? Many clustering methods including MLE of FMMs are defined by parameters optimising an objective function. The asymptotic limit is then usually a functional defined on P by generalizing the objective function to general distributions ("population version" or "canonical functional") such as in Pollard (1981); García-Escudero et al. (2015); Coretto and Hennig (2017) and also in the present work.

(ii) Provided that a limit clustering exists, does this correspond to a reasonable clustering structure given P that is of interest when doing cluster analysis?

Addressing these aspects, the present paper provides two main contributions:

(i) We establish the existence and the consistency of the MLE for a general class of FMMs based on ESDs for a general nonparametric P, generalizing the results for Gaussian FMMs in García-Escudero et al. (2015); Coretto and Hennig (2017). This implies finite sample existence of the MLE under mild conditions.

(ii) Under the additional assumption that P is a mixture of K nonparametric components that put most of their probability mass on sufficiently well separated subsets of the data space, we show that the components of the canonical functional resulting from the MLE of FMMs of ESDs correspond to the well separated mixture components of P. This can be interpreted as follows. If P is a distribution generating well enough separated clusters of a very flexible kind, particularly not necessarily corresponding to the ESDs assumed to be the components of the FMM estimated by the MLE, the MLE will anyway for large enough data recover these clusters. We are not aware of any result of this kind in the literature regarding nonparametric consistency based on canonical functionals.

The paper is organized as follows. In Section 2, we define the class of ESDs and the corresponding FMMs. We review the connection between model-based clustering and FMMs. We also define the MLE and give an account of the issue of unbounded likelihood. Finally we introduce and motivate the regularization of the ML optimization problem based on the ERC. Section 3 is devoted to the finite sample analysis of the ML procedure. In Section 4, we establish the MLE's existence and consistency, assuming that P is some general nonparametric distribution. Section 5 investigates the canonical functional corresponding to the MLE for FMMS of ESDs for distributions P that can be written as mixtures of sufficiently well separated nonparametric components. In Section 6 we present numerical experiments that illustrate the results. Section 7 concludes the paper.

2. Finite mixture modeling with elliptically symmetric distributions

Elliptically symmetric distributions can be obtained applying an affine transformation to a spherical distribution. Let the random vector $Y \in \mathbb{R}^p$ have a spherical distribution. With fixed $\boldsymbol{\mu} \in \mathbb{R}^p$, and $\boldsymbol{\Omega} \in \mathbb{R}^{p\times p}$, the random vector $X = \boldsymbol{\mu} + \boldsymbol{\Omega}Y$ is said to have an elliptically-symmetric distribution denoted by $X \sim \mathrm{ESD}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \coloneqq \boldsymbol{\Omega}\boldsymbol{\Omega}^\intercal$. With a fixed distribution of the generating Y, $\mathrm{ESD}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ forms a symmetric location–scatter class of models with location parameter $\boldsymbol{\mu}$ and a symmetric positive semi-definite scatter matrix parameter $\boldsymbol{\Sigma}$. If $\mathrm{E}[Y] = \mathbf{0}$ and $\mathrm{Var}[Y] \propto \mathbf{I}_p$, the $p \times p$-unit matrix, then $\boldsymbol{\mu} = \mathrm{E}[X]$ and $\boldsymbol{\Sigma} \propto \mathrm{Var}[X]$. If Y is absolutely continuous and $\boldsymbol{\Sigma}$ is non-singular, than X is absolutely continuous with density function

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \coloneqq \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} g\left((\boldsymbol{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right), \tag{1}$$

where $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is the so-called radial or density generator function. Here we assume that $g(\cdot)$ is monotonically non-increasing, which implies that $f(\cdot)$ is unimodal. $\lim_{t\to+\infty} g(t) = 0$ is required so that $f(\cdot)$ is a proper density.

The choice of $g(\cdot)$ (or, equivalently, the distribution of Y) defines a specific family of ESDs. A number of popular models can be obtained in this way, for example multivariate Gaussian distributions, multivariate Student-t distributions and multivariate logistic distributions. Some families of ESDs have parameters in addition to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, e.g. the degrees of freedom of the Student-t. We do not involve such additional parameters, so the results given here apply to MLEs for a family of multivariate Student-t distribution with fixed degrees of freedom, and generally to any location-scale family as defined in eq. (1) with g fulfilling the conditions above.

Kelker (1970) originally introduced elliptical distributions to generalize the multivariate Gaussian model. ESDs have elliptically shaped density contours. They can model joint linear dependence between the p components of a random vector and possibly heavy tails. For these reasons, ESDs occur frequently in the theory of statistics and applications (see Genton, 2004, for a comprehensive overview). Finite mixture densities based on eq. (1) are the convex combinations

$$\psi(\boldsymbol{x}; \boldsymbol{\theta}) \coloneqq \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2}$$

where $\pi_k \in [0, 1]$ so that $\sum_{k=1}^{K} \pi_k = 1$, and $\boldsymbol{\theta}$ is a parameter vector collecting all elements of $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$, $k = 1, 2, \ldots, K$. Given a family of ESDs, an FMM is defined by the set of distributions for all possible choices of $\boldsymbol{\theta}$. FMMs are popular tools for modeling multimodality, skewness and heterogenous populations, and for performing several supervised and unsupervised tasks such as semi-parametric density estimation, classification and clustering (McLachlan and Peel, 2000; Frühwirth-Schnatter et al., 2019).

2.1. Clustering and classification. FMMs used for model-based clustering and classification normally identify the classes or clusters with the mixture components. FMMs can be equivalently formulated involving component membership indicators for the observations.

Assume that there are K mixture components. Let $X_1, \ldots, X_n$ model a sequence of i.i.d. observations. These are accompanied by a sequence of 0-1 valued i.i.d. random

variables $\zeta_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iK})^\top$, $i = 1, \ldots, n$, with $\sum_{i=1}^K Z_{ik} = 1$, i.e., for given i one of the Z_{ik} is 1, all the others are 0. Assume that ζ_i has a categorical distribution with $\Pr[Z_{ik} = 1] = \mathrm{E}[Z_{ik} = 1] = \pi_k$ for $k = 1, 2, \ldots, K$. Let eq. (1) be the density function of the conditional distribution of $X_i \,|\, Z_{ik} = 1$, then eq. (2) is the density function of the unconditional distribution of X_i. Z_{ik} is the component (cluster) membership indicator for observation i and the kth mixture component. eq. (2) can be seen as a model generating an expected fraction π_k of points from its k-th component $f(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. If the K mixture components are reasonably separated, sampling from eq. (2) will generate clustered regions of data points. Cluster analysis is unsupervised, i.e., $\zeta_i, \ldots, \zeta_n$ are unobserved.

$\zeta_1, \ldots, \zeta_n$ represent a partition $\mathcal{G}_K := \{G_k;\ k = 1, 2, \ldots, K\}$ of $\{1, \ldots, n\}$, where $Z_{ik} := \mathbb{I}\{i \in G_k\}$, $\mathbb{I}\{\cdot\}$ being the usual indicator function, meaning that the mixture component (cluster) having generated X_i is k.

In cluster analysis one wants to assign an object to one of the K groups of $\mathcal{G}_K$, which for $i = 1, \ldots, n$ amounts to predicting ζ_i from X_i.

This can be done based on the posterior probability

$$\tau_k(\boldsymbol{x}; \boldsymbol{\theta}) := \Pr[Z_{ik} = 1 \,|\, X_i = \boldsymbol{x}] = \frac{\pi_k f(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\psi(\boldsymbol{x}; \boldsymbol{\theta})}. \tag{3}$$

Given a parameter vector $\boldsymbol{\theta}$, an object $\boldsymbol{x}$ can be assigned to the cluster $\mathrm{cl}(\boldsymbol{x}, \boldsymbol{\theta}) \in \{1, 2, \ldots, K\}$ according to the "*maximum a posteriori probability*" (MAP) rule:

$$\mathrm{cl}(\boldsymbol{x}, \boldsymbol{\theta}) := \underset{k \in \{1,2,\ldots,K\}}{\arg\max}\ \tau_k(\boldsymbol{x}; \boldsymbol{\theta}). \tag{4}$$

The MAP rule implements the optimal Bayes classifier, i.e. the assignment that minimizes the expected 0-1 loss, also known as misclassification rate. Since $\boldsymbol{\theta}$ is not known, it has to be estimated from the data, and eq. (4) is then computed based on the estimator $\hat{\boldsymbol{\theta}}$.

Remark 1. The optimality of the MAP rule for the clustering problem requires that (*i*) data are generated from eq. (2), (*ii*) "true" clusters are defined in terms of $Z_{ik} = \mathbb{I}\{i \in G_k\}$, that is, $f(\cdot, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the "true" underlying k-th class-conditional density, and (*iii*) the posterior ratios in eq. (3) are computed at the "true" generating parameter $\boldsymbol{\theta}$ (the optimality would hold asymptotically with a consistent estimator for the "true" $\boldsymbol{\theta}$). In practice, these conditions are arguably never fulfilled ((ii) is a matter of definition, but relies on (i)).

Applying this approach to data generated from a general distribution P that does not necessarily have a density of type eq. (2) means that the method imposes a partition on P that is governed by "cluster prototype densities" of the form eq. (1). Here we investigate what happens then, at least asymptotically.

2.2. Maximum likelihood estimator. The unknown mixture parameter vector $\boldsymbol{\theta}$ is typically estimated by ML. Often, computations are performed based on the Expectation-Maximization (EM) algorithm (McLachlan and Peel, 2000). Let $\mathbb{X}_n = \{\boldsymbol{x}_i,\ i = 1, 2, \ldots, n\}$ be the observed sample. Define the sample likelihood and log-likelihood functions

$$\mathcal{L}_n(\boldsymbol{\theta}) := \prod_{i=1}^n \psi(\boldsymbol{x}_i; \boldsymbol{\theta}), \quad \ell_n(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^n \log(\psi(\boldsymbol{x}_i; \boldsymbol{\theta})). \tag{5}$$

Finding the maximum of eq. (5), the sample MLE, is not straightforward. In fact, eq. (5) does not have a maximum. Kiefer and Wolfowitz (1956) discovered the unboundedness of ℓ_n in the context of univariate Gaussian FMM. The issue extends to many classes of FMMs, including those studied here. To see why, let us fix additional notations also used throughout the rest of the paper. Let $\|\cdot\|$ be the Euclidean norm. If the parameter vectors $\boldsymbol{\theta}$ come with indexes and accents, its members $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ have the same indexes and accents, i.e., $\tilde{\boldsymbol{\theta}}_m$ contains $(\tilde{\pi}_{mk}, \tilde{\boldsymbol{\mu}}_{mk}, \tilde{\boldsymbol{\Sigma}}_{mk})$ for all $m \in \mathbb{N}$. For given $\boldsymbol{\theta}$, let $\boldsymbol{\kappa}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ (indexes and accents are applied to $\boldsymbol{\kappa}$ as above). Scalar parameters contained in $\boldsymbol{\theta}$'s subvectors are indexed so that the first two subscripts always denote the mixture component and the dimension in the feature space respectively, e.g. $\mu_{k,j} \in \mathbb{R}$ is the j-th coordinate of $\boldsymbol{\mu}_k$.

Consider a sequence $(\boldsymbol{\theta}_m)_{m\in\mathbb{N}}$ where $\boldsymbol{\mu}_{1,m} = \boldsymbol{x}_1 \in \mathbb{X}_n$ and the smallest eigenvalue of $\boldsymbol{\Sigma}_{1,m}$ converges to 0 as $m \longrightarrow +\infty$, then $\ell_n(\boldsymbol{\theta}_m) \longrightarrow +\infty$. The likelihood degeneracy is caused by the fact that the density peak of $f(\cdot)$ is controlled by the smallest eigenvalue of the scatter matrix. In fact, eq. (1) can be parameterized in terms of the eigenvalue decomposition of $\boldsymbol{\Sigma}$. That is

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(\prod_{j=1}^{p} \lambda_j(\boldsymbol{\Sigma}) \right)^{-\frac{1}{2}} g\left(\sum_{j=1}^{p} \lambda_j(\boldsymbol{\Sigma})^{-1} (\boldsymbol{x} - \boldsymbol{\mu})^{\intercal} V_j(\boldsymbol{\Sigma}) V_j(\boldsymbol{\Sigma})^{\intercal} (\boldsymbol{x} - \boldsymbol{\mu}) \right), \tag{6}$$

where $\lambda_j(\boldsymbol{\Sigma})$ is the j-th eigenvalue of $\boldsymbol{\Sigma}$, and $V_j(\boldsymbol{\Sigma})$ is its corresponding normalized eigenvector, i.e. $\|V_j(\boldsymbol{\Sigma})\| = 1$ for all $j = 1, 2, \ldots, p$. Define $\lambda^*_{\min}(\boldsymbol{\Sigma}) = \min\{\lambda_j(\boldsymbol{\Sigma});\ j = 1, 2, \ldots, p\}$, and $\lambda^*_{\max}(\boldsymbol{\Sigma}) = \max\{\lambda_j(\boldsymbol{\Sigma});\ j = 1, 2, \ldots, p\}$. The density $f(\cdot)$ can be bounded as

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq (\lambda^*_{\min}(\boldsymbol{\Sigma}))^{-\frac{p}{2}} g\left(\lambda^*_{\max}(\boldsymbol{\Sigma})^{-1} \|\boldsymbol{x} - \boldsymbol{\mu}\|^2 \right) \leq g(0) (\lambda^*_{\min}(\boldsymbol{\Sigma}))^{-\frac{p}{2}}. \tag{7}$$

Furthermore, $f(\cdot) \in O(\lambda^*_{\min}(\boldsymbol{\Sigma})^{-\frac{p}{2}})$, meaning that that $f(\boldsymbol{\mu}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \longrightarrow +\infty$ as $\lambda^*_{\min}(\boldsymbol{\Sigma}) \searrow 0$ unless also $\lambda^*_{\max}(\boldsymbol{\Sigma}) \searrow 0$. The ML problem can therefore not be solved in plain uncostrained form, and requires either constraints on the parameter space or a penalty. Let $\Lambda(\boldsymbol{\theta}) = \{\lambda_j(\boldsymbol{\Sigma}_k),\ j = 1, \ldots, p,\ k = 1, \ldots, k\}$, $\lambda_{\min}(\boldsymbol{\theta}) = \min\{\Lambda(\boldsymbol{\theta})\}$, $\lambda_{\max}(\boldsymbol{\theta}) = \max\{\Lambda(\boldsymbol{\theta})\}$.

A possible constraint to define a proper ML problem is the eigenratio constraint (ERC)

$$\frac{\lambda_{\max}(\boldsymbol{\theta})}{\lambda_{\min}(\boldsymbol{\theta})} \leq \gamma, \tag{8}$$

with a constant $\gamma \geq 1$. $\gamma = 1$ constrains all component scatter matrices to be spherical and equal. Increasing γ continuously allows for more flexible scatter shapes. The ERC has been introduced by Dennis (1981) and Hathaway (1985) for the Gaussian case and brought back to the attention of the literature by Ingrassia (2004). For a recent review see García-Escudero et al. (2018). Although the resulting MLE will not be affine equivariant on the original feature space, affine equivariance can be achieved by sphering the data. The ERC captures the discrepancy between scatter shapes across components. Therefore, at least in clustering applications, these constraints have a data-analytic interpretation. Other proposals of fully affine equivariant constraints exist in the literature (see Gallegos and Ritter, 2009; Ritter, 2014); however, there are no algorithms for their exact implementation. Another approach to solving the MLE existence problem is to rely on penalized ML methods. Ciuperca et al. (2003) treated the case of the univariate

Gaussian mixture. A recent comprehensive review is found in Chen (2017).

In the following sections we study the sequence of constrained MLEs

$$\boldsymbol{\theta}_n \in \underset{\boldsymbol{\theta}\in\tilde{\Theta}_K}{\arg\max}\ \ell_n(\boldsymbol{\theta}), \tag{9}$$

where the constrained parameter space is

$$\tilde{\Theta}_K := \left\{\boldsymbol{\theta}:\ \ \pi_k \geq 0\ \forall k \geq 1,\ \sum_{k=1}^{K}\pi_k = 1;\ \frac{\lambda_{\max}(\boldsymbol{\theta})}{\lambda_{\min}(\boldsymbol{\theta})} \leq \gamma\right\}. \tag{10}$$

The ERC implies that $\lambda_{\max}(\boldsymbol{\theta}) \in O(\lambda_{\min}(\boldsymbol{\theta}))$ for all $\boldsymbol{\theta} \in \tilde{\Theta}_K$. For a sequence $(\boldsymbol{\theta}_m)_{m\in\mathbb{N}}$ such that $\boldsymbol{\theta}_m \in \tilde{\Theta}_K$ and $\lambda_j(\boldsymbol{\Sigma}_{k,m}) \searrow 0$ as $m \longrightarrow +\infty$ for some $j \in \{1,2,\ldots,p\}$ and $k \in \{1, 2,\ldots,K\}$ the ERC enforces $\lambda_{\max}(\boldsymbol{\theta}_m) \searrow 0$. The ERC comes with two major technical challenges: (i) the parameter space $\tilde{\Theta}_K$ is not compact; (ii) the resulting ML problem cannot be solved by the standard constrained optimization methods.

3. Finite sample existence

In this section we give precise non-asymptotic conditions involving n, p, K and $g(\cdot)$ that guarantee the existence of $\boldsymbol{\theta}_n$ given the input data set $\mathbb{X}_n$. The existence of the sequence $(\boldsymbol{\theta}_n)_{n\in\mathbb{N}}$ is the prerequisite for the study of its asymptotic behavior investigated in Section 4 and 5.

Assumption 1. For fixed $p, n, K, \mathbb{X}_n$:

(a) $n > K$, and $\mathbb{X}_n$ contains at least $K+1$ distinct points;

(b) for all $\beta \in \mathbb{R}_{>0}$ and $\alpha \in \{0, 1, \ldots, K\}$, $g(\beta y^{-1})^{n-\alpha} \in o(y^{\frac{p}{2}n})$ as $y \searrow 0$.

Assumption 1-(b) plays the central role in dealing with the sample likelihood function's unboundedness. It guarantees that for $\boldsymbol{\theta} \in \tilde{\Theta}_K$, such that $\lambda_{\min}(\boldsymbol{\theta}) \searrow 0$, the densities in the product $\mathcal{L}_n(\boldsymbol{\theta})$ vanish sufficiently fast at all data points $\boldsymbol{x}_i \in \mathbb{X}_n$ that do not coincide with mixture components' centers $\boldsymbol{\mu}_k$ for all $k \in \{1, 2, \ldots, K\}$. The following Lemma is the key result to obtain the compactification of the parameter space.

Lemma 1. *Let Assumption 1 hold. Consider a sequence $(\boldsymbol{\theta}_m)_{m\in\mathbb{N}} \in \tilde{\Theta}_K$ such that $\lambda_j(\boldsymbol{\Sigma}_{k,m}) \searrow 0$ as $m \longrightarrow \infty$ for some $k \in \{1, 2, \ldots, K\}$ and $j \in \{1, 2, \ldots, p\}$. Then, $\sup_{\tilde{\Theta}_K} \ell_n(\boldsymbol{\theta}_m) \longrightarrow -\infty$ as $m \longrightarrow \infty$.*

Proof. First, rearrange $\mathcal{L}_n(\cdot)$. Consider a vector of indexes $\boldsymbol{k} := (k_1, k_2 \ldots, k_n)^\intercal$ where $k_r \in \{1, 2, \ldots, K\}$ for $r \in \{1, 2, \ldots, n\}$. For a fixed such $\boldsymbol{k}$ define the products

$$p^f_{\boldsymbol{k}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}) := \prod_{r=1}^{n} f(\boldsymbol{x}_r; \boldsymbol{\mu}_{k_r}, \boldsymbol{\Sigma}_{k_r}), \quad \text{and} \quad p^\pi_{\boldsymbol{k}}(\boldsymbol{\theta}) := \prod_{r=1}^{n} \pi_{k_r}. \tag{11}$$

There exists K^n of such possible vectors of indexes $\boldsymbol{k}$, say $\boldsymbol{k}_1, \boldsymbol{k}_2, \ldots, \boldsymbol{k}_{K^n}$. Based on these vectors it is possible to write the sample likelihood function as a mixture of K^n components like eq. (11), that is

$$\mathcal{L}_n(\boldsymbol{\theta}) = \sum_{h=1}^{K^n} p^\pi_{\boldsymbol{k}_h}(\boldsymbol{\theta})\, p^f_{\boldsymbol{k}_h}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}). \tag{12}$$

Since $\boldsymbol{\theta}_m \in \tilde{\Theta}_K$, $\lambda_j(\boldsymbol{\Sigma}_{km}) \searrow 0$ implies that $\lambda_{\max}(\boldsymbol{\theta}_m) \searrow 0$ as $m \longrightarrow +\infty$ Assume w.l.o.g. (otherwise consider a suitable subsequence) that for sufficiently large m, the sequence $(\boldsymbol{\theta}_m)_{m\in\mathbb{N}}$ is such that, for $j \in \{1, 2, \ldots, p\}$, and $k \in \{1, 2, \ldots, K\}$, the centrality parameter μ_{kjm} either converges, or it leaves any compact set. If the limits of $(\boldsymbol{\mu}_{km})_{m\in N}$ belong to $\mathbb{X}_n$, there are at most K of them. Therefore, according to Assumption 1-(a), there exists $i \in \{1, 2, \ldots, n\}$ and $\varepsilon > 0$ such that $\boldsymbol{x}_i \in \mathbb{X}_n \backslash \mathbb{X}'$ and $\|\boldsymbol{x}_i - \boldsymbol{\mu}_{km}\| > \varepsilon$ for large enough m. Consider a factorization such as the one in eq. (11) for some vector of indexes $\boldsymbol{k}$. Define $\mathbb{X}'_{\boldsymbol{k}} \subseteq \mathbb{X}'$ where

$$\mathbb{X}'_{\boldsymbol{k}} := \left\{ \boldsymbol{x}_r \in \mathbb{X}' : \lim_{m\to\infty} \boldsymbol{\mu}_{k_r,m} = \boldsymbol{x}_r, \text{ for any } k_r \in \{k_1, k_2, \ldots, k_n\} \right\}.$$

Let $\#(\mathbb{X}'_{\boldsymbol{k}}) = q_{\boldsymbol{k}} \leq K$. Therefore, $p^f_{\boldsymbol{k}}(\cdot)$ in eq. (11) can be factorized as

$$p^f_{\boldsymbol{k}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}_m) = \prod_{r:\, \boldsymbol{x}_r \in \mathbb{X}'_{\boldsymbol{k}}} f(\boldsymbol{x}_r; \boldsymbol{\mu}_{k_r m}, \boldsymbol{\Sigma}_{k_r m}) \prod_{r:\, \boldsymbol{x}_r \in \mathbb{X}\backslash\mathbb{X}'_{\boldsymbol{k}}} f(\boldsymbol{x}_r; \boldsymbol{\mu}_{k_r m}, \boldsymbol{\Sigma}_{k_r m}). \tag{13}$$

As $m \longrightarrow +\infty$, $f(\boldsymbol{x}_r; \boldsymbol{\mu}_{k_r m}, \boldsymbol{\Sigma}_{k_r m}) \in O(\lambda_{\min}(\boldsymbol{\theta}_m)^{-\frac{p}{2}})$ for all r such that $\boldsymbol{x}_r \in \mathbb{X}'_{\boldsymbol{k}}$, therefore

$$\sup_{\tilde{\Theta}_K} \prod_{r:\, \boldsymbol{x}_r \in \mathbb{X}'_{\boldsymbol{k}}} f(\boldsymbol{x}_r; \boldsymbol{\mu}_{k_r m}, \boldsymbol{\Sigma}_{k_r m}) \in O\left(\lambda_{\min}(\boldsymbol{\theta}_m)^{-\frac{p}{2} q_{\boldsymbol{k}}}\right).$$

On the other hand, for all r such that $\boldsymbol{x}_r \in \mathbb{X}\backslash\mathbb{X}'_{\boldsymbol{k}}$ and a positive constant c_{k_r}

$$f(\boldsymbol{x}_r; \boldsymbol{\mu}_{k_r m}, \boldsymbol{\Sigma}_{k_r m}) \leq O(\lambda_{\min}(\boldsymbol{\theta}_m))^{-\frac{p}{2}}\ g(\lambda_{\min}(\boldsymbol{\theta}_m)^{-1} c_{k_r}).$$

Assumption 1-(b) and $q_{\boldsymbol{k}} \leq K$ ensure that $g(\lambda_{\min}(\boldsymbol{\theta}_m)^{-1} c_{k_r})^{n-q_{\boldsymbol{k}}} \in o(\lambda_{\min}(\boldsymbol{\theta}_m)^{\frac{p}{2} n})$, therefore

$$\sup_{\tilde{\Theta}_K} \prod_{r:\, \boldsymbol{x}_r \in \mathbb{X}\backslash\mathbb{X}'_{\boldsymbol{k}}} f(\boldsymbol{x}_r; \boldsymbol{\mu}_{k_r m}, \boldsymbol{\Sigma}_{k_r m}) \in O\left(\lambda_{\min}(\boldsymbol{\theta}_m)^{-\frac{p}{2}(n-q_{\boldsymbol{k}})}\right)\ o\left(\lambda_{\min}(\boldsymbol{\theta}_m)^{\frac{p}{2} n}\right).$$

The latter implies that

$$\sup_{\tilde{\Theta}_K}\ p^f_{\boldsymbol{k}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}_m) \in O\left(\lambda_{\min}(\boldsymbol{\theta}_m)^{-\frac{p}{2} q_{\boldsymbol{k}}}\right) O\left(\lambda_{\min}(\boldsymbol{\theta}_m)^{-\frac{p}{2}(n-q_{\boldsymbol{k}})}\right) o\left(\lambda_{\min}(\boldsymbol{\theta}_m)^{\frac{p}{2} n}\right)$$

That is

$$\sup_{\tilde{\Theta}_K}\ p^f_{\boldsymbol{k}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}_m) \in o(1).$$

We conclude that $\sup_{\tilde{\Theta}_K} p^f_{\boldsymbol{k}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}_m) \longrightarrow 0$ for large enough m, whatever the vector of indexes $\boldsymbol{k}$. The latter implies that all factors of $\mathcal{L}_n(\boldsymbol{\theta}_m)$ in (12) vanish and therefore $\sup_{\tilde{\Theta}_K} \ell_n(\boldsymbol{\theta}_m) \longrightarrow -\infty$. □

Theorem 1 (finite sample existence)**.** *Under Assumption 1, $\boldsymbol{\theta}_n$ exists.*

Proof. First note that $\tilde{\Theta}_K$ is not empty because for any $\gamma \geq 1$ and any choice $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots \boldsymbol{\Sigma}_K$ such that the ERC is not fulfilled, for all $k = 1, 2, \ldots, K$ one can always replace $\boldsymbol{\Sigma}_k$ with $\dot{\boldsymbol{\Sigma}}_k = V(\boldsymbol{\Sigma}_k)\dot{\Lambda}(\boldsymbol{\Sigma}_k)V(\boldsymbol{\Sigma}_k)^{\intercal}$, where $V(\boldsymbol{\Sigma}_k) \in \mathbb{R}^{p\times p}$ is an orthogonal matrix whose columns are eigenvectors of $\boldsymbol{\Sigma}_k$, and $\dot{\Lambda}(\boldsymbol{\Sigma}_k) \in \mathbb{R}^{p\times p}$ is a diagonal matrix where $\dot{\Lambda}(\boldsymbol{\Sigma}_k)[j, j] = \min\{\lambda_j(\boldsymbol{\Sigma}_k), \gamma\lambda_{\min}(\boldsymbol{\theta})\}$. By construction any such choice $\dot{\boldsymbol{\Sigma}}_1, \dot{\boldsymbol{\Sigma}}_2, \ldots \dot{\boldsymbol{\Sigma}}_K$ will always satisfy the eigen-ratio constraint. With the following step we show that there

exists a compact set $T_K \subset \tilde{\Theta}_K$ such that $\sup_{\boldsymbol{\theta}\in T_K} \ell_n(\boldsymbol{\theta}) = \sup_{\boldsymbol{\theta}\in\tilde{\Theta}_K} \ell_n(\boldsymbol{\theta})$.

Step (a): Consider $\boldsymbol{\theta}$ such that $\pi_1 = 1$, $\boldsymbol{\mu}_1 = \boldsymbol{x}_1$, $\boldsymbol{\Sigma}_j = \boldsymbol{I}_p$ for all $k \in \{1, 2, \ldots, K\}$, arbitrary $\boldsymbol{\mu}_k$ and $\pi_k = 0$ for all $k \neq 1$. For such a choice of $\boldsymbol{\theta}$, $l_n(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \log f(\boldsymbol{x}_i; \boldsymbol{x}_1, \boldsymbol{I}_p) > -\infty$, thus $\sup_{\boldsymbol{\theta}\in\tilde{\Theta}_K} l_n(\boldsymbol{\theta}) > -\infty$.
Step (b): Consider a sequence $(\dot{\boldsymbol{\theta}}_m)_{m\in N}$. Lemma 1 rules out the possibility that $\dot{\lambda}_{kjm} \searrow 0$ for some index $k \in \{1, 2, \ldots, K\}$ and $j \in \{1, 2, \ldots, p\}$, because this would imply that $\sup_{\tilde{\Theta}_K} \ell_n(\dot{\boldsymbol{\theta}}_m) \longrightarrow -\infty$. Using eq. (7),

$$\ell_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \log\left(\sum_{k=1}^K \pi_k f(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \leq K \log\left(2\pi\lambda_{\min}(\boldsymbol{\theta})\right)^{-\frac{p}{2}}. \tag{14}$$

Assume that $\dot{\boldsymbol{\theta}}_m \longrightarrow \dot{\boldsymbol{\theta}}$ where $\dot{\boldsymbol{\theta}} \in \tilde{\Theta}_K$ and $\dot{\lambda}_{kjm} \longrightarrow +\infty$ for some indexes $k \in \{1, 2, \ldots, K\}$ and $j \in \{1, 2, \ldots, p\}$. Because of the ERC, $\lambda_{\min}(\dot{\boldsymbol{\theta}}_m) \longrightarrow +\infty$, and eq. (14) implies that $\sup_{\dot{\boldsymbol{\theta}}_m} \ell_n(\dot{\boldsymbol{\theta}}_m) \longrightarrow -\infty$. Therefore, the case that $\dot{\lambda}_{kjm} \longrightarrow +\infty$ is also ruled out.
Step (c): now suppose that $\|\dot{\boldsymbol{\mu}}_{km}\| \longrightarrow +\infty$ for some $k \in \{1, 2, \ldots, K\}$. W.l.o.g take $k = 1$. Choose an alternative sequence $(\ddot{\boldsymbol{\theta}}_m)_{m\in\mathbb{N}}$ that is equal to $(\dot{\boldsymbol{\theta}}_m)_{m\in\mathbb{N}}$ except now $\ddot{\boldsymbol{\mu}}_{1m} = \boldsymbol{0}$ for all $m \in \mathbb{N}$. Note that $f(\boldsymbol{x}_i; \dot{\boldsymbol{\mu}}_{1m}, \dot{\boldsymbol{\Sigma}}_1) \longrightarrow 0$ for all $i \in \{1, 2, \ldots, n\}$, which implies that $\psi(\boldsymbol{x}_i, \dot{\boldsymbol{\theta}}_m) < \psi(\boldsymbol{x}_i, \ddot{\boldsymbol{\theta}}_m)$ for large enough m for all $i \in \{1, 2, \ldots, n\}$. Hence, $\ell_n(\ddot{\boldsymbol{\theta}}_m) < \ell_n(\dot{\boldsymbol{\theta}}_m)$ for large m.

Steps (a)–(c) plus the fact that $\pi_k \in [0, 1]$ for all $k \in \{1, 2, \ldots, K\}$ imply that the vector $\boldsymbol{\theta}$ maximizing $\ell_n(\cdot)$ must lie in a compact subset $T_n \subset \tilde{\Theta}_K$, and the continuity of $\ell_n(\cdot)$) guarantees existence of $\boldsymbol{\theta}_n$. □

Remark 2 (Finite sample existence for specific models). The key Assumption 1-(b) holds depending on the density generator function $g(\cdot)$. For the Gaussian distribution, $g(t) = c_p \exp(-t/2)$, with c_p being a constant dependent on p. It is easy to see that the condition is fulfilled for all $n > K$. For the Student-t distribution with ν degrees of freedom, $g(t) = c_{p,\nu}(1 + t/\nu)^{-(\nu+p)/2}$ where $c_{p,\nu}$ is a constant that depends on both p and ν. In the latter case Assumption 1-(b) holds if $n > K(1+p/\nu)$, requiring more data points for larger p and smaller ν. Not surprisingly, if $\nu \longrightarrow +\infty$, the condition is fulfilled with $n > K$ as for the Gaussian case. Assumption 1-(b) can be easily checked for other ESDs as well.

Remark 3 (Computing). The previous statement is also relevant to ensure that applying computing algorithms searching for $\boldsymbol{\theta}_n$ on a given input data set makes sense. In the FMM context, the usual approach to compute the MLE is to run the Expectation-Maximization (EM) algorithm of Dempster et al. (1977) or some of its variants (see McLachlan and Krishnan, 1997). García-Escudero et al. (2015) and Coretto and Hennig (2017) developed EM-type algorithms implementing the ERC for the case of a Gaussian FMM; Coretto and Hennig (2017) also proved convergence results. The general structure of the EM applied to the FMM will apply to finite mixtures of ESDs; however, the generality of $g(\cdot)$ in eq. (1) does not allow to write down the M-step explicitly. Furthermore, the implementation of the ERC for the Gaussian case can be extended to those ESD that have a Gaussian representation. For instance, for mixtures of Student-t distributions, one can take the EM algorithm presented in Chapter 7 of Peel and McLachlan (2000) and add the ERC via a conditional M–step similar to the CM1-step of Algorithm 2 in Coretto and Hennig (2017). In general, an appropriate constrained M-step needs to be developed depending on the specific $g(\cdot)$, but this is outside the scope of this paper.

4. Asymptotic analysis for general P

Now consider a general P as generator of p-dimensional data to be analysed by an MLE derived from an FMM of ESDs. Treating K in the definition of ψ as fixed, define the following log-likelihood-type functionals:

$$L(\boldsymbol{\theta}, P) := \int \log \psi(\boldsymbol{x}, \boldsymbol{\theta}) dP(\boldsymbol{x}),$$

and

$$L_K := L_K(P) := \sup_{\boldsymbol{\theta} \in \tilde{\Theta}_K} L(\boldsymbol{\theta}), \qquad \boldsymbol{\theta}^\star := \boldsymbol{\theta}^\star(P) := \arg\max_{\boldsymbol{\theta} \in \tilde{\Theta}} L(\boldsymbol{\theta}, P).$$

$\boldsymbol{\theta}^\star(P)$ is not normally unique (in particular, there is the "label switching" issue, i.e., the numbering of the mixture components is arbitrary), and is taken to be any of the maximising parameter vectors. Without demanding that P is of type eq. (2), the usual interpretation of these quantities is that $L(\boldsymbol{\theta}; P)$ is proportional, up to a term that only depends on P, to the Kullback-Liebler risk (KLR) of approximating the density of P by $\psi(\cdot, \boldsymbol{\theta})$. Therefore, $\boldsymbol{\theta}^\star(P)$ provides the best approximation to P in terms of KLR obtainable from a K-components ESD mixture model. The following analysis establishes the existence of $\boldsymbol{\theta}^\star(P)$ and the convergence of $(\boldsymbol{\theta}_n)_{n \in \mathbb{N}}$ to $\boldsymbol{\theta}^\star(P)$. The notation $\mathrm{E}_P[\cdot]$ denotes the expectation with respect to the distribution P.

Assumption 2. For every set $A = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K\} \subset \mathbb{R}^p$ with at most K points: $P(A) < 1$.

Assumption 3. $\mathrm{E}_P\left[\log g(\|X\|^2)\right] > -\infty$.

Assumption 4. For fixed $\beta \in \mathbb{R}_{>0}$, $g(\beta y^{-1}) \in o(y)$ as $y \searrow 0$.

Assumption 5. $L_{K-1}(P) < L_K(P)$.

Assumption 2 is fulfilled if P does not concentrate its mass on K points. Note that this assumption is fulfilled by the empirical distribution P_n for sufficiently large n when P is absolutely continuous. Assumption 3 is needed to ensure that under P it makes sense to maximize the log-likelihood function. When $f(\cdot)$ is the Gaussian density, Assumption 3 is fulfilled if $\mathrm{E}_P \|X\|^2$ exists. Assumption 4 is the analog of Assumption 1-(b) and deals with the degeneration of the scatter matrices. It implies that, far from the centers $\boldsymbol{\mu}_k$, $f(\cdot)$ vanishes at a speed that is sufficiently fast compared to the speed at which it becomes unbounded in the center when $\lambda^*_{\min}(\boldsymbol{\Sigma}_k) \searrow 0$. Assumption 4 is fulfilled by the Gaussian model. When $f(\cdot)$ is the Student-t distribution with ν degrees of freedom, it holds if $\nu + p > 2$.

Regarding Assumption 5, note that generally $r < s \Rightarrow L_r(P) \leq L_s(P)$, because any parameter $\boldsymbol{\theta}$ with r mixture components can be reproduced with more mixture components setting some component proportions to 0. If $L_{K-1}(P) = L_K(P)$, then maxima of the likelihood exist with a component proportion of 0, and the corresponding location and scatter parameters can take any value. Particularly then the maxima of the likelihood cannot be forced into any compact set.

To prove the consistency Theorem 3, we first establish the existence of the ML functional $\boldsymbol{\theta}^\star$ (or equivalently $\boldsymbol{\theta}^\star(P) \neq \emptyset$) in Theorem 2. The existence Theorem 2 is obtained via the compactification of the parameter space based on the following Lemmas 2-4.

Lemma 2. *Under Assumption 3, for all $K \geq 1$, $L_K(P) > -\infty$.*

Proof. Choose $\boldsymbol{\theta} \in \tilde{\Theta}_K$ such that $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\pi_1 = 1, \pi_2 = \ldots = \pi_K = 0$. Note that $\boldsymbol{\Sigma}_1$ fulfills the ERC. All remaining parameters of $\boldsymbol{\theta}$ are chosen arbitrarily. The statement is proven by using Assumption 3:

$$\sup_{\boldsymbol{\theta} \in \tilde{\Theta}_K} \int \log \psi(\boldsymbol{x}, \boldsymbol{\theta}) dP(\boldsymbol{x}) \geq \int \log \pi_1 f(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) dP(\boldsymbol{x}) \geq \int \log g(\|\boldsymbol{x}\|^2) dP(\boldsymbol{x}) > -\infty.$$

□

Lemma 3. *Under Assumptions 2 to 4 there exist* $\lambda^\star_{\min} > 0$, $\lambda^\star_{\max} < \infty$, $\epsilon > 0$, *so that*

(a) $L(\boldsymbol{\theta}, P) \leq L_K - \epsilon$ *for every* $\boldsymbol{\theta}$ *with* $\lambda_{\min}(\boldsymbol{\theta}) < \lambda^\star_{\min}$ *or* $\lambda_{\max}(\boldsymbol{\theta}) > \lambda^\star_{\max}$,

(b) for iid samples $X_1, X_2, \ldots$ *from* P, *for sequences* $(\dot{\boldsymbol{\theta}}_n)_{n \in \mathbb{N}}$ *with* $\lambda_{\min}(\dot{\boldsymbol{\theta}}_n) < \lambda^\star_{\min}$ *or* $\lambda_{\max}(\dot{\boldsymbol{\theta}}_n) > \lambda^\star_{\max}$ *for large enough* n: $\ell_n(\dot{\boldsymbol{\theta}}_n) \leq \ell_n(\boldsymbol{\theta}_n) - \epsilon$ P*-a.s.*

Proof. Consider a sequence $(\boldsymbol{\theta}_m)_{m \in \mathbb{N}}$ where $\boldsymbol{\theta}_m \in \tilde{\Theta}_K$ with $\lambda_{\max}(\boldsymbol{\theta}_m) \longrightarrow +\infty$ for all $m \in \mathbb{N}$. The ERC enforces $\lambda_{\min}(\boldsymbol{\theta}_m) \longrightarrow +\infty$ and therefore $\sup_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{\mu}_{km}, \boldsymbol{\Sigma}_{km}) \searrow 0$ for all $k = 1, 2, \ldots, K$, and $\sup_{\boldsymbol{x}} \psi(\boldsymbol{x}; \boldsymbol{\theta}_m) \searrow 0$. The dominated convergence theorem implies that $\mathrm{E}_P[\psi(\boldsymbol{x}; \boldsymbol{\theta}_m)] \searrow 0$, and therefore $L(\boldsymbol{\theta}_m, P) \leq \log(\mathrm{E}_P[\psi(\boldsymbol{x}; \boldsymbol{\theta}_m)]) \longrightarrow -\infty$. $L(\boldsymbol{\theta}_m, P) \searrow -\infty$ according to Lemma 2 makes it impossible that $L(\boldsymbol{\theta}_m, P)$ is close to $L_K(P)$ for m large enough. The latter proves the existence of the upper bound $\lambda^\star_{\max} < \infty$ as required in part (a).

Now consider a sequence $(\boldsymbol{\theta}_m)_{m \in \mathbb{N}}$ where $\boldsymbol{\theta}_m \in \tilde{\Theta}_K$ with $\lambda_{\min}(\boldsymbol{\theta}_m) \searrow 0$. Define

$$A_{m,\epsilon} = \left\{ \boldsymbol{x} : \min_{1 \leq k \leq K} \|\boldsymbol{x} - \boldsymbol{\mu}_{km}\| > \epsilon \right\}.$$

Assumption 2 ensures that for $\delta > 0$ there exists $\epsilon > 0$ so that $P(A_{m,\epsilon}) \geq \delta$ for all $m \in \mathbb{N}$.

$$\begin{aligned}
L(\boldsymbol{\theta}_m, P) &\leq \int_{A_{m,\epsilon}} \log \psi(\boldsymbol{x}; \boldsymbol{\theta}_m) dP(\boldsymbol{x}) + \int_{A^c_{m,\epsilon}} \log \psi(\boldsymbol{x}; \boldsymbol{\theta}_m) dP(\boldsymbol{x}) \\
&\leq \int_{A_{m,\epsilon}} \log \left(K (2\pi)^{-\frac{p}{2}} \lambda_{\min}(\boldsymbol{\theta}_m)^{-\frac{p}{2}} g(\gamma \lambda_{\min}(\boldsymbol{\theta}_m)^{-1} \epsilon^2) \right) dP(\boldsymbol{x}) + \\
&\quad + \int_{A^c_{m,\epsilon}} \log \left(K (2\pi)^{-\frac{p}{2}} \lambda_{\min}(\boldsymbol{\theta}_m)^{-\frac{p}{2}} g(0) \right) dP(\boldsymbol{x}) \\
&\leq P(A_{m,\epsilon}) \left(\log K - \frac{p}{2} \log(2\pi) - \frac{p}{2} \log \lambda_{\min}(\boldsymbol{\theta}_m) + \log(g(\gamma \lambda_{\min}(\boldsymbol{\theta}_m)^{-1} \epsilon^2) \right) + \\
&\quad + P(A^c_{m,\epsilon}) \left(\log K - \frac{p}{2} \log(2\pi) - \frac{p}{2} \log \lambda_{\min}(\boldsymbol{\theta}_m) + \log g(0) \right) \\
&\leq c_1 + c_2 \log \lambda_{\min}(\boldsymbol{\theta}_m)^{-1} + c_3 \log g \left(c_4 \lambda_{\min}(\boldsymbol{\theta}_m)^{-1} \right)
\end{aligned}$$

for positive constants c_1, c_2, c_3 and c_4 all independent of $\boldsymbol{\theta}_m$. The previous inequality can be rearranged as follows

$$L(\boldsymbol{\theta}_m, P) \leq \log g \left(c_4 \lambda_{\min}(\boldsymbol{\theta}_m)^{-1} \right) \left(\frac{c_1}{\log g \left(c_4 \lambda_{\min}(\boldsymbol{\theta}_m)^{-1} \right)} - c_2 \frac{\log \lambda_{\min}(\boldsymbol{\theta}_m)}{\log g (c_4 \lambda_{\min}(\boldsymbol{\theta}_m)^{-1})} + c_3 \right).$$

By Assumption 4, as $m \longrightarrow +\infty$, $g(c_4 \lambda_{\min}(\boldsymbol{\theta}_m)^{-1}) \searrow 0$ faster than $\lambda_{\min}(\boldsymbol{\theta}_m)$ and $L_K(P) \longrightarrow -\infty$, which contradicts $L_K(P) > -\infty$ established by Lemma 2. Therefore,

there exists a lower bound $\lambda^\star_{\min} > 0$ as stated in part (a) of the statement.

Regarding part (b), let the sequence $(\dot{\boldsymbol{\theta}}_n)_{n\in\mathbb{N}}$ be chosen as the sequence above taking $m = n \longrightarrow \infty$. Let P_n be the empirical distribution. 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). Glivenko-Cantelli enforces $P_n(A_{n,\epsilon}) - P(A_{n,\epsilon}) \to 0$ a.s. Note that $L(\dot{\boldsymbol{\theta}}_n, P_n) = \ell_n(\dot{\boldsymbol{\theta}}_n)$, and $L_K(P_n) = \ell_n(\boldsymbol{\theta}_n)$. For large enough n Lemma 2 holds with P_n replacing P. Therefore part (b) of the statement is proved by replacing P with P_n in the proof of part (a). □

Lemma 4. *Under Assumptions 2 to 5, there is a compact set $T \subset \mathbb{R}^p$, $\epsilon_1 > 0$ and $\epsilon_2 > 0$ so that*

(a) $L(\boldsymbol{\theta}, P) \leq L_K - \epsilon_1$ whenever $\boldsymbol{\mu}_k \notin T$ for some $k \in \{1, 2, \ldots, K\}$;

(b) $\sup_{\boldsymbol{\theta}\in\tilde{\Theta}_K : \boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K \in T} L(\boldsymbol{\theta}, P) = L_K(P) < +\infty$;

(c) for iid samples $X_1, X_2, \ldots$ from P, for sequences $(\dot{\boldsymbol{\theta}}_n)_{n\in\mathbb{N}}$ with $\dot{\boldsymbol{\theta}}_n \in \tilde{\Theta}_K$ and $\dot{\boldsymbol{\mu}}_{kn} \in T$ for all $k \in \{1, 2, \ldots, K\}$ for large enough n: $\sup_{\dot{\boldsymbol{\theta}}_n} \ell_n(\dot{\boldsymbol{\theta}}_n) \leq \ell_n(\boldsymbol{\theta}_n) - \epsilon_2$ P-a.s whenever $\dot{\boldsymbol{\mu}}_{kn} \notin T$ for some $k \in \{1, 2, \ldots, K\}$, and $\sup_{\dot{\boldsymbol{\theta}}_n} \ell_n(\dot{\boldsymbol{\theta}}_n) = \ell_n(\boldsymbol{\theta}_n)$ P-a.s.

Proof. Consider a sequence $(\boldsymbol{\theta}_m)_{m\in\mathbb{N}}$ where $\boldsymbol{\theta}_m \in \tilde{\Theta}_K$ for all $m \in \mathbb{N}$. Assume $\|\boldsymbol{\mu}_{km}\| \longrightarrow \infty$ for $k \in \{1, 2, \ldots, r\}$, and $\boldsymbol{\mu}_{km} \in T$ for all $k > r$. Note that $r = K$ would imply $L(\boldsymbol{\theta}_m, P) \longrightarrow -\infty$ for large enough m, therefore, we only consider the case when $r < K$. Take a sequence $(\epsilon_m)_{m\in\mathbb{N}}$ and construct the sets

$$A_m := \left\{ x : \forall k \in \{1, \ldots, r\} \; f(\boldsymbol{x}, \boldsymbol{\mu}_{km}, \boldsymbol{\Sigma}_{km}) \leq \epsilon_m \sum_{j=r+1}^{K} \pi_{jm} f(\boldsymbol{x}, \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}) \right\},$$

where $\epsilon_m \searrow 0$ slowly enough that $A_1 \supset A_2 \supset, \ldots$ and $P(A_m) \nearrow 1$. Define another sequence $(\bar{\boldsymbol{\theta}}_m)_{m\in\mathbb{N}}$ such that $\bar{\boldsymbol{\theta}}_m \in \tilde{\Theta}_{K-r}$ with $\bar{\boldsymbol{\mu}}_{km} = \boldsymbol{\mu}_{(k+r)m}$, $\bar{\boldsymbol{\Sigma}}_{km} = \boldsymbol{\Sigma}_{(k+r)m}$, and $\bar{\pi}_{km} = \pi_{(k+r)m}(\sum_{j=r+1}^{K} \pi_{jm})^{-1}$ for all $k \in \{r+1, \ldots, K\}$. By construction $\bar{\boldsymbol{\mu}}_{km} \in T$ for all $k \in \{r+1, \ldots, K\}$. Lemma 3 implies that for all $\boldsymbol{\theta} \in \tilde{\Theta}_K$ that are sufficiently close to L_L (that is $L(\boldsymbol{\theta}, P) > L_k - \epsilon$ according to Lemma 3) the mixture component densities are bounded above: $f(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \leq f_{\max} = (2\pi\lambda^\star_{\min})^{-\frac{p}{2}}$ for all k and some suitably defined $\lambda^\star_{\min} > 0$. The latter also implies that $L_K < +\infty$. $\psi(\boldsymbol{x}, \bar{\boldsymbol{\theta}}_m) \geq \psi(x, \boldsymbol{\theta}_m)$ for all $\boldsymbol{x} \in A_m$, therefore

$$\begin{aligned}
L(\boldsymbol{\theta}_m, P) &= \int_{A_m} \log(\psi(\boldsymbol{x}, \boldsymbol{\theta}_m)) dP(\boldsymbol{x}) + \int_{A_m^c} \log(\psi(\boldsymbol{x}, \boldsymbol{\theta}_m)) dP(\boldsymbol{x}) \\
&\leq \int_{A_m} \log((1+\epsilon_m)\psi(\boldsymbol{x}, \bar{\boldsymbol{\theta}}_m)) dP(\boldsymbol{x}) + P(A_m^c) \log(f_{\max}). \\
&\leq \int \mathbb{I}_{A_m}(\boldsymbol{x}) \log(\psi(\boldsymbol{x}, \bar{\boldsymbol{\theta}}_m)) dP(\boldsymbol{x})) + a_m
\end{aligned}$$

with $a_m \searrow 0$. $\log(f(\boldsymbol{x}, \bar{\boldsymbol{\mu}}_{km}, \bar{\boldsymbol{\Sigma}}_{km})$ can be bounded by $\log(f_{\max})$, and by applying the dominated convergence theorem we obtain that $\int \mathbb{I}_{A_m}(\boldsymbol{x}) \log(\psi(\boldsymbol{x}, \bar{\boldsymbol{\theta}}_m)) dP(\boldsymbol{x}) \longrightarrow L(\bar{\boldsymbol{\theta}}_m, P)$. Therefore, $L(\boldsymbol{\theta}_m, P) \leq L(\bar{\boldsymbol{\theta}}_m, P) + o(1)$ for large enough m. Because of Assumption 5, $L(\bar{\boldsymbol{\theta}}_m, P) \leq L_{K-r} < L_K$ and therefore $L(\boldsymbol{\theta}_m, P) < L_K$ for sufficiently large m, which proves part (a) of the statement.

Observe that Lemma 3 implies that for all $\boldsymbol{\theta} \in \tilde{\Theta}_K$ for which $L(\boldsymbol{\theta})$ is close to L_K, each eigenvalue in $\boldsymbol{\theta}$ is contained in $[\lambda^\star_{\min}, \lambda^\star_{\max}]$ with $0 < \lambda^\star_{\min} \leq \lambda^\star_{\max} < +\infty$, $\psi(\boldsymbol{x}, \boldsymbol{\theta}) \leq f_{\max}$, and therefore $L_K < +\infty$. Consider $(\boldsymbol{\theta}_m)_{m\in\mathbb{N}}$ so that $L(\boldsymbol{\theta}_m) \longrightarrow L_k$ with $\boldsymbol{\mu}_{1m}, \ldots, \boldsymbol{\mu}_{Km} \in T$, and $0 < \lambda^\star_{\min} \leq \lambda_{kjm} \leq \lambda^\star_{\max} < +\infty$ for all $k \in \{1, \ldots, K\}$, $j \in \{1, \ldots, p\}$. Because of compactness, there exists $\boldsymbol{\theta}_0$ such that $\boldsymbol{\theta}_m \to \boldsymbol{\theta}_0$ and, using Fatou's Lemma, $L_K = \lim_{m\to\infty} L(\boldsymbol{\theta}_m, P) \leq \mathrm{E}_P[\limsup_m \psi(\boldsymbol{x}, \boldsymbol{\theta}_m)] = L(\boldsymbol{\theta}_0) \leq L_K < +\infty$. The latter proves the existence of a maximum stated in part (b).

As for the proof of Lemma 3-(b), part (c) of the statement is shown by setting $n = m$, replacing the previous sequences $(\boldsymbol{\theta}_m)_{m\in\mathbb{N}}$ and $(\bar{\boldsymbol{\theta}}_m)_{m\in\mathbb{N}}$ with $(\dot{\boldsymbol{\theta}}_n)_{n\in\mathbb{N}}$ and $(\dot{\bar{\boldsymbol{\theta}}}_n)_{n\in\mathbb{N}}$, and replacing P with the empirical distribution P_n. The same steps as before can be applied taking into account the following observations. Glivenko-Cantelli enforces $P_n(A_n) - P(A_n) \longrightarrow 0$ a.s. because we can construct a sequence of closed balls $(B_n)_{n\in\mathbb{N}}$ so that $B_n \subseteq A_n$ and $P(B_n) \longrightarrow 1$ a.s. The closed balls are a Vapnik-Chervonenkis class, and $P_n(A_n) \geq P_n(B_n)$. Note that for all $\boldsymbol{\theta} \in \tilde{\Theta}_K$ with $L(\boldsymbol{\theta}, P) > L_{K-1}$: $\ell_n(\boldsymbol{\theta}) \longrightarrow L(\boldsymbol{\theta}, P)$ a.s. by the strong law of large numbers and therefore for large enough n: $\sup_{\boldsymbol{\theta}\in\tilde{\Theta}_K} \ell_n(\boldsymbol{\theta}) > L_{K-1}$ a.s. Furthermore, if $\dot{\bar{\boldsymbol{\theta}}}_n$ is chosen optimally in a compact set as in the proof of part (b), $\ell_n(\cdot)$ converges uniformly over $\tilde{\Theta}_{K-r}$ (see Theorem 2 in Jennrich (1969)) and therefore $\limsup_{n\to\infty} \ell_n(\dot{\bar{\boldsymbol{\theta}}}_n) \leq L_{K-1} < L_K$. □

Theorem 2 (Existence of the ML functional)**.** *Under Assumptions 2 to 5 there is a compact subset $T_{\tilde{\Theta}_K} \subset \tilde{\Theta}_K$ so that there exists $\boldsymbol{\theta} \in T_{\tilde{\Theta}_K}$ such that $-\infty < L(\boldsymbol{\theta}) = L_K < +\infty$, and for all $\boldsymbol{\theta} \notin T_{\tilde{\Theta}_K}$, $L(\boldsymbol{\theta})$ is bounded away from L_K.*

Proof. The statement is shown by putting together Lemmas 2–4. □

Holzmann et al. (2006) found sufficient conditions for the identifiability of certain classes of finite mixtures of ESDs. If $P = P_0$, and $\psi(x; \boldsymbol{\theta}_0)$ is its density whose mixture components also fulfill the sufficient conditions of Holzmann et al. (2006), the previous theorem guarantees that the argmax functional $\boldsymbol{\theta}^\star(P)$ exists and is unique up to label switching. This is a consequence of Lemma 1 of Wald (1949). But here we are interested in the more realistic case when ψ is not necessarily a density of P. In this case neither $L(\boldsymbol{\theta})$ nor $\ell_n(\boldsymbol{\theta})$ can be expected to have a unique maximum, not even up to label switching. Based on a technique inspired by Redner (1981), we show that the sequence of constrained ML estimators is asymptotically close to one of the parameters $\boldsymbol{\theta}^\star$ giving the best approximation of P in terms of the KLR. This technique provides consistency on the quotient space topology identifying all population log-likelihood maxima. Define the sets

$$S(\dot{\boldsymbol{\theta}}) := \left\{ \boldsymbol{\theta} \in \Theta_K(P) : \int \log \psi(\boldsymbol{x}; \boldsymbol{\theta}) dP(\boldsymbol{x}) = \int \log \psi(\boldsymbol{x}; \dot{\boldsymbol{\theta}}) dP(\boldsymbol{x}) \right\},$$

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

Note that $S(\dot{\boldsymbol{\theta}})$ and $\mathcal{T}(\dot{\boldsymbol{\theta}}, \varepsilon)$ do not depend on the specific likelihood maximiser if $\dot{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star$.

Theorem 3 (Consistency)**.** *Under Assumptions 2 to 5, for every $\varepsilon > 0$ and every sequence of maximizers $\boldsymbol{\theta}_n$ of $\ell_n(\cdot)$:* $\lim_{n\to\infty} \Pr[\boldsymbol{\theta}_n \in \mathcal{T}(\boldsymbol{\theta}^\star, \varepsilon)] = 1$.

Proof. Because of Lemmas 3-(b) and 4-(c) there is a compact set $T_{\tilde{\Theta}_K} \subset \tilde{\Theta}_K$ so that all $\boldsymbol{\theta}_n \in T_{\tilde{\Theta}_K}$ for large enough n a.s. Using Lemma 3, $|\log \psi(\boldsymbol{x}, \boldsymbol{\theta})| \leq C$ for some finite constant C for all $\boldsymbol{\theta} \in T_{\tilde{\Theta}_K}$. Sufficient conditions for Theorem 2 in Jennrich (1969) are

satisfied, and therefore $\sup_{\boldsymbol{\theta}\in T_{\tilde{\Theta}_K}} |\ell_n(\boldsymbol{\theta}) - L(\boldsymbol{\theta})| \longrightarrow 0$ P-a.s. Applying the same argument as in the proof of Theorem 5.7 in van der Vaart and Wellner (1996), we obtain $L(\boldsymbol{\theta}_n) \longrightarrow L(\boldsymbol{\theta}^\star)$ P-a.s. By continuity of $L(\cdot)$ and Theorem 2 we have that for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $L(\boldsymbol{\theta}^\star) - L(\boldsymbol{\theta}) > \delta$ for all $\boldsymbol{\theta} \in T_{\tilde{\Theta}_K} \setminus \mathcal{T}(\boldsymbol{\theta}^\star, \varepsilon)$. Denote $(\Omega, \mathcal{A}, P)$ the probability space where the sample random variables are defined, and consider the following events

$$A_n := \left\{\omega \in \Omega : \ \boldsymbol{\theta}_n \in T_{\tilde{\Theta}_K} \setminus \mathcal{T}(\boldsymbol{\theta}^\star, \varepsilon)\right\},$$

and

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

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

5. The MLE functional in the mixture case

After having showed the nonparametric consistency of the MLE for its own canonical functional, here we investigate the canonical functional for a P that is a mixture of K well enough separated nonparametric components $Q_1, \ldots, Q_K$.

Such a P can be interpreted as generating K well separated clusters, even though the Q_k, $k = 1, \ldots, K$, can be chosen so flexibly that one or more of them themselves could be multimodal or even a mixture of well separated components generating homogeneous data. As the MLE functional of the mixture of ESDs enforces K components to be fitted to P, it seems reasonable, interpreting these as corresponding to K clusters, to expect that they align with $Q_1, \ldots, Q_K$ also in the latter case if the separation between $Q_1, \ldots, Q_K$ is stronger than the separation between any "subcomponents" of any Q_k.

For P of this type, the clusters generated by the K ESD mixture components of the MLE functional (ESDC) will indeed approximately correspond to the "central regions" of $Q_1, \ldots, Q_K$, and the parameters of the ESDC will approximate the parameters of the MLE canonical functionals of separate ESDs evaluated at each of $Q_1, \ldots, Q_K$ alone. For example, if the ESD family is chosen as Gaussian with flexible means and covariance matrices, the corresponding parameters of the ESDC will approximate the mean and covariance matrix functionals of $Q_1, \ldots, Q_K$.

Here are some definitions and assumptions. Let $Q_1, \ldots, Q_K$ be distributions on $\mathbb{R}^p$ (generally the same notation refers to distributions and their cumulative distribution functions) parameterized in such a way that 0 is their "center" in some sense; it could be the mode, the mean, the multivariate median or quantile; important is only that Q_k is defined relative to 0. Let $\xi_1, \ldots, \xi_K > 0$ mixture proportions with $\sum_{k=1}^K \xi_k = 1$. For $m \in \mathbb{N}$, $k \in \{1, \ldots, K\}$ let $\boldsymbol{\rho}_{mk} \in \mathbb{R}^p$ sequences so that

$$\lim_{m\to\infty} \min_{k_1 \neq k_2 \in \{1,\ldots,K\}} \|\boldsymbol{\rho}_{mk_1} - \boldsymbol{\rho}_{mk_2}\| = \infty.$$

Define a sequence of distributions P_m on $\mathbb{R}^p$ by $P_m(\boldsymbol{x}) := \sum_{k=1}^K \xi_k Q_k(\boldsymbol{x} - \boldsymbol{\rho}_{mk})$. The mixture P_m is constructed in such a way that its mixture components for increasing m become better and better separated, although they are nonparametric and may have non-vanishung densities, so there may be overlap betwen them even for arbitrarily large m.

Consider, for $\epsilon > 0$, the "central set" $\{\boldsymbol{x} : \ \|\boldsymbol{x}\| < \epsilon\}$ about 0. ϵ can be chosen large enough

that for arbitrarily small $\eta > 0$:

$$\forall k \in \{1, \ldots, K\} : \ Q_k\{\|\boldsymbol{x}\| < \epsilon\} \geq 1 - \eta, \tag{15}$$

which in particular implies that

$$\exists \delta > 0 : \ \forall k \in \{1, \ldots, K\} : \ \xi_k Q_k\{\|\boldsymbol{x}\| < \epsilon\} \geq \delta. \tag{16}$$

The following theorem states that in this setup, when evaluating $\boldsymbol{\theta}^\star(P_m)$, eventually the different clusters include the full (arbitrarily large) central sets of the different mixture components, and in this sense the clustering corresponds to the mixture structure. We require:

Assumption 6. For any sequence $(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)_{n\in\mathbb{N}}$, Q_k, $k = 1, \ldots, K$ for which $\frac{\lambda^*_{\max}(\boldsymbol{\Sigma}_n)}{\lambda^*_{\min}(\boldsymbol{\Sigma}_n)} \leq \gamma$:

$$\lim_{n\to\infty} \lambda^*_{\min}(\boldsymbol{\Sigma}_n) = 0 \Rightarrow \lim_{n\to\infty} \int \log f(\boldsymbol{x}; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) dQ_k(\boldsymbol{x}) = -\infty.$$

Assumption 7. $\exists c_0 > -\infty$ so that for all $k \in \{1, \ldots, K\}$: $\int \log g(\|\boldsymbol{x}\|) dQ_k(\boldsymbol{x}) \geq c_0$.

Remark 4. Assumption 6 is e.g. fulfilled for Q_k if Q_k is not concentrated on a single point and $g(x) \in o(x^{-\delta})$, where $\delta > \frac{p}{2\epsilon}$, $\epsilon > 0$ so that there are two disjoint sets A and B : $Q_k(A) \geq \epsilon$, $Q_k(B) \geq \epsilon, \inf_{x\in A, y\in B} \frac{\|x-y\|}{2} =: \eta > 0$. This is because in that case, using eq. (6),

$$\begin{gathered}\int \log f(\boldsymbol{x}; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) dQ_k(\boldsymbol{x}) \leq \\ \epsilon \left[\log g\left(\lambda^*_{\max}(\boldsymbol{\Sigma}_n)^{-1}\eta^2\right) - \tfrac{p}{2}\log \lambda^*_{\min}(\boldsymbol{\Sigma}_n)\right] + (1-\epsilon)\left[\log g(0) - \tfrac{p}{2}\log \lambda^*_{\min}(\boldsymbol{\Sigma}_n)\right] \\ \in o\left((\epsilon\delta - \tfrac{p}{2}) \log \lambda^*_{\max}(\boldsymbol{\Sigma}_n)\right).\end{gathered}$$

With similar reasoning, Q_k continuous and $g(x) \in o(x^{-\delta})$ with $\delta > \frac{p}{2}$ will fulfill Assumption 6.

Assumption 7 may be violated if Q_k has far heavier tails than f; e.g., if f is Gaussian, it amounts to Q_k having an existing covariance matrix.

Theorem 4. *With the above definitions, under Assumption 6, for large enough m, the clusters of $\boldsymbol{\theta}^\star(P_m)$ can be numbered in such a way that for $k \in \{1, \ldots, K\}$:*

$$B_\epsilon(\boldsymbol{\rho}_{mk}) := \{\boldsymbol{x} : \ \|\boldsymbol{x} - \boldsymbol{\rho}_{mk}\| < \epsilon\} \subseteq C_{mk} = \{\boldsymbol{x} : \ \mathrm{cl}(\boldsymbol{x}, \boldsymbol{\theta}^\star(P_m)) = k\}.$$

Proof. Show the following statements:

S1 For $\tilde{\boldsymbol{\theta}}_m$ defined by $\tilde{\pi}_{mk} = \xi_k$, $\tilde{\boldsymbol{\mu}}_{mk} = \boldsymbol{\rho}_{mk}$, $\tilde{\boldsymbol{\Sigma}}_{mk} = \boldsymbol{I}_p$, $k = 1, \ldots, K$:

$$\exists m^- > -\infty \ \forall m : \ L(\tilde{\boldsymbol{\theta}}_m, P_m) \geq m^-. \tag{17}$$

S2 For a sequence $(\boldsymbol{\theta}_m)_{m\in\mathbb{N}}$ let

$$D_m(\boldsymbol{\theta}_m) := \max_{1\leq k\leq K} \min_{1\leq j\leq K} \|\boldsymbol{\rho}_{mk} - \boldsymbol{\mu}_{mj}\|.$$

If $\limsup_{m\to\infty} D_m(\boldsymbol{\theta}_m) = \infty$, then $\liminf_{m\to\infty} L(\boldsymbol{\theta}_m, P_m) = -\infty$.

S3 The following hold for $\boldsymbol{\theta}_m := \boldsymbol{\theta}^\star(P_m)$: There are constants $0 < c_1 < c_2 < \infty$ independent of m so that $c_1 < \lambda_{\min}(\boldsymbol{\theta}_m) < c_2$, and there is a constant $c_3 > 0$ so that for large enough m, $k = 1, \ldots, K : \pi_{mk} \geq c_3$.

S4 If $\exists m_D < \infty$ so that $\forall m : D_m(\boldsymbol{\theta}_m) \leq m_D$, and S3 holds for $(\boldsymbol{\theta}_m)_{m\in\mathbb{N}}$, then for large enough m the components of $\boldsymbol{\theta}_m$ can be numbered so that for $k = 1, \ldots, K$:

$$B_\epsilon(\boldsymbol{\rho}_{mk}) \subseteq C_{mk}. \tag{18}$$

S1 together with S2 imply that $\limsup_{m\to\infty} D_m(\boldsymbol{\theta}^\star(P_m)) < \infty$, so that, together with S3, $(\boldsymbol{\theta}^\star(P_m))_{m\in\mathbb{N}}$ fulfills the condition for S4, from which the theorem follows.

Proof of S1:

$$\begin{aligned}
L(\tilde{\boldsymbol{\theta}}_m, P_m) &= \int \log\left(\sum_{j=1}^K \tilde{\pi}_{mj} f(\boldsymbol{x}; \tilde{\boldsymbol{\mu}}_{mj}, \tilde{\boldsymbol{\Sigma}}_{mj})\right) d\sum_{k=1}^K \xi_k Q_k(\boldsymbol{x} - \boldsymbol{\rho}_{mk}) \\
&= \sum_{k=1}^K \xi_k \int \log\left(\sum_{j=1}^K \xi_j f(\boldsymbol{x}; \boldsymbol{\rho}_{mj}, \boldsymbol{I}_p)\right) dQ_k(\boldsymbol{x} - \boldsymbol{\rho}_{mk}) \\
&\geq \sum_{k=1}^K \xi_k \int \log\left(\xi_k f(\boldsymbol{x}; \boldsymbol{\rho}_{mk}, \boldsymbol{I}_p)\right) dQ_k(\boldsymbol{x} - \boldsymbol{\rho}_{mk}) \\
&= \sum_{k=1}^K \xi_k \int \log\left(\xi_k g(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})\right)\Big) dQ_k(\boldsymbol{x}) = m^-,
\end{aligned}$$

independently of m. $m^- > -\infty$ follows from Assumption 7.

Proof of S2:
It suffices to consider $\lim_{m\to\infty} D_m(\boldsymbol{\theta}_m) = \infty$ because in case this holds for the lim sup, there is a subsequence diverging to ∞. W.l.o.g., number mixture components so that $D_m(\boldsymbol{\theta}_m) = \|\boldsymbol{\rho}_{mk^*} - \boldsymbol{\mu}_{mk^*}\|$ for a fixed $k^* \in \{1, \ldots, K\}$.

The following is required for showing that $\lim_{m\to\infty} L(\boldsymbol{\theta}_m, P_m) = -\infty$: Because of the properties of f and g, for $k = 1, \ldots, K$ and any sequence $(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)_{n\in\mathbb{N}}$ for which $\frac{\lambda^*_{\max}(\boldsymbol{\Sigma}_n)}{\lambda^*_{\min}(\boldsymbol{\Sigma}_n)} \leq \gamma$:

$$\lim_{n\to\infty} \lambda^*_{\min}(\boldsymbol{\Sigma}_n) = \infty \Rightarrow \int \log f(\boldsymbol{x}; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) dQ_k(\boldsymbol{x}) = -\infty. \tag{19}$$

Furthermore,

$$\lim_{n\to\infty} |\boldsymbol{\mu}_n| = \infty \Rightarrow \int \log f(\boldsymbol{x}; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) dQ_k(\boldsymbol{x}) = -\infty, \tag{20}$$

regardless of whether $\lambda^*_{\min}(\boldsymbol{\Sigma}_n)$ is bounded (in which case $\log f(\boldsymbol{x}; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \longrightarrow -\infty$ for $\boldsymbol{x} \in S$ with $Q_k(S)$ arbitrarily close to 1), diverges to ∞ (in which case eq. (19) obtains), or converges to zero (Assumption 6).

Observe

$$L(\boldsymbol{\theta}_m, P_m) = \xi_{k^*} \int \log\left(\sum_{j=1}^K \pi_{mj} f(\boldsymbol{x}; \boldsymbol{\mu}_{mj}, \boldsymbol{\Sigma}_{mj})\right) dQ_{k^*}(\boldsymbol{x} - \boldsymbol{\rho}_{mk^*}) + L^*_m, \tag{21}$$

where

$$L_m^* := \sum_{k \neq k^*} \xi_k \int \log \left(\sum_{j=1}^{K} \pi_{mj} f(\boldsymbol{x}; \boldsymbol{\mu}_{mj}, \boldsymbol{\Sigma}_{mj}) \right) dQ_k(\boldsymbol{x} - \boldsymbol{\rho}_{mk}).$$

Because of eq. (20), the first term of eq. (21) diverges to $-\infty$, and so does $L(\boldsymbol{\theta}_m, P_m)$ if L_m^* is bounded from above. Assumption 6 implies that $\lambda_{\min}(\boldsymbol{\theta}_m) \geq c > 0$, therefore $f(\boldsymbol{x}; \boldsymbol{\mu}_{mk}, \boldsymbol{\Sigma}_{mk}) \leq c^{-p/2} g(0) < \infty$, and $L_m^* \leq c^{-p/2} g(0)$. This proves S2.

Proof of S3:
$c_1 < \lambda_{\min}(\boldsymbol{\theta}^\star(P_m)) < c_2$ holds by S1, Assumption 6, and eq. (19).

Because of S2, $\exists m_D < \infty$ so that $\forall m: \; D_m(\boldsymbol{\theta}^\star(P_m)) \leq m_D$. Number the components of $\boldsymbol{\theta}_m$ so that for $k = 1, \ldots, K$:

$$\|\boldsymbol{\rho}_{mk} - \boldsymbol{\mu}_{mk}\| = \min_{1 \leq j \leq K} \|\boldsymbol{\rho}_{mk} - \boldsymbol{\mu}_{mj}\|. \tag{22}$$

This is possible because of $D_m(\boldsymbol{\theta}_m) \leq m_D$.

Assume w.l.o.g. that $\pi_{m1} \longrightarrow 0$. Then, eq. (21) holds with $k^* = 1$. L_m^* is once more bounded from above, and

$$\lim_{m \to \infty} \int \log \left(\sum_{j=1}^{K} \pi_{mj} f(\boldsymbol{x}; \boldsymbol{\mu}_{mj}, \boldsymbol{\Sigma}_{mj}) \right) dQ_1(\boldsymbol{x} - \boldsymbol{\rho}_{m1}) = -\infty, \tag{23}$$

because $\pi_{m1} \longrightarrow 0$, $f(\boldsymbol{x}; \boldsymbol{\mu}_{m1}, \boldsymbol{\Sigma}_{m1}) \leq c_1^{-p/2} g(0) < \infty$ as in the proof of S2, and for $j \neq 1: \; f(\boldsymbol{x}; \boldsymbol{\mu}_{mj}, \boldsymbol{\Sigma}_{mj}) \longrightarrow 0$ for $\boldsymbol{x} \in S$ with $\int_S dQ_1(\boldsymbol{x} - \boldsymbol{\rho}_{m1})$ arbitrarily close to 1. But for $\boldsymbol{\theta}_m = \boldsymbol{\theta}^\star(P_m)$, eq. (23) is in contradiction to S1.

Proof of S4:
For large enough m number the components of $\boldsymbol{\theta}_m$ according to eq. (22). For $\tilde{\boldsymbol{x}} \in B_\epsilon(\boldsymbol{0})$ so that $\boldsymbol{x} = \tilde{\boldsymbol{x}} + \boldsymbol{\rho}_{mk} \in B_\epsilon(\boldsymbol{\rho}_{mk})$ consider

$$\tau_k(\boldsymbol{x}; \boldsymbol{\theta}_m) = \frac{\pi_{mk} f(\boldsymbol{x}; \boldsymbol{\mu}_{mk}, \boldsymbol{\Sigma}_{mk})}{\sum_{j=1}^{K} \pi_{mj} f(\boldsymbol{x}; \boldsymbol{\mu}_{mj}, \boldsymbol{\Sigma}_{mj})}.$$

Because of S3, $\pi_{mk} \geq c_3$, and $f(\boldsymbol{x}; \boldsymbol{\mu}_{mk}, \boldsymbol{\Sigma}_{mk}) \geq c_4 > 0$, whereas for $j \neq k$: $f(\boldsymbol{x}; \boldsymbol{\mu}_{mj}, \boldsymbol{\Sigma}_{mj}) \longrightarrow 0$ uniformly for $\tilde{\boldsymbol{x}} \in B_\epsilon(\boldsymbol{0})$, so that $\tau_k(\boldsymbol{x}; \boldsymbol{\theta}_m) \longrightarrow 1$, and for large enough $m: \; \forall \boldsymbol{x}: \; \mathrm{cl}(\boldsymbol{x}, \boldsymbol{\theta}_m) = k$. □

Remark 5. It is not essential that $B_\epsilon(\boldsymbol{\rho}_{mk})$ is defined based on the Euclidean distance; other L_p-distances will work as well as Mahalanobis distances with any fixed covariance matrix.

Remark 6. For m large, the different components Q_k, $k = 1, \ldots, K$, are shifted by $\boldsymbol{\rho}_{mk}$ and their central sets $B_\epsilon(\boldsymbol{\rho}_{mk})$ are therefore very far away from each other. It may therefore not seem surprising that these ultimately belong to different clusters. But the statement is not trivial. The Q_k can have densities that are nowhere zero and even have heavy tails (Assumption 7 will then require f to have heavy tails, too), so that there will always be overlap between the Q_k. There are clustering methods with fixed K that for $n \longrightarrow \infty$ will not match the K different clusters to $B_\epsilon(\boldsymbol{\rho}_{m1}), \ldots, B_\epsilon(\boldsymbol{\rho}_{mK})$:

- If the Q_k (or even at least one of them) are chosen so that for $n \longrightarrow \infty$, the

largest distance $D_{max,n}$ of an observation to its nearest neighbour goes to ∞ in probability (which holds for distributions with heavy enough tails, see Theorem 5.3 in Jammalamadaka and Janson (2015)), then for any given but arbitrarily large m, n can be chosen large enough so that $D_{max,n}$ is arbitrarily much larger than $\max_{1\leq j,k\leq K}\|\boldsymbol{\rho}_{mj}-\boldsymbol{\rho}_{mk}\|$ with arbitrarily large probability. Then, the Single Linkage dendrogram based on the Euclidean distance will merge different central sets earlier than connecting the observation that is farthest from its nearest neighbour with anything else.

- A trimmed clustering method trimming a proportion of α observations can trim a complete central set if $\alpha \geq \xi_k$ for some k. Similarly, RIMLE (Coretto and Hennig (2017)) may classify a complete central set as noise.

Arguably there are situations in which the behaviour of trimmed clustering and RIMLE as explained above may be seen as desirable, namely if one of the Q_k looks like modelling more than one cluster; e.g., it can be bimodal with two clearly separated modes. It may then be seen as appropriate if this attracts more than one of the K clusters, whereas another Q_k with either low probability or low (smoothed) density may be classified as generating outliers.

The following theorem states that for the setup of Theorem 4, the estimators of $\boldsymbol{\mu}_{mk}$ and $\boldsymbol{\Sigma}_{mk}$ converge to the estimators for the individual mixture components Q_{mk}, so that the increasing separation of mixture components for $m \longrightarrow \infty$ implies that ultimately the characteristics of Q_{mk} defined in terms of the ESD densities f can be estimated without influence of the other mixture components. For example, if f defines a Gaussian location-scale family, $\boldsymbol{\mu}^\star_{mk}$ will converge toward the mean of Q_{mk}, and $\boldsymbol{\Sigma}^\star_{mk}$ will converge to the covariance matrix of Q_{mk}, see Corollary 1, which shows that the required Assumption 8 below holds at least in this case.

For $k = 1, \ldots, K$, let

$$\tilde{\boldsymbol{\kappa}}_k := (\tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k) = \arg\max_{\boldsymbol{\kappa}} \tilde{L}(\boldsymbol{\kappa}, Q_k),\ \tilde{L}(\boldsymbol{\kappa}, Q) := \int \log f(\boldsymbol{x}; \boldsymbol{\kappa}) dQ_k(\boldsymbol{x}).$$

The corresponding functionals for $Q_{mk} = Q_k(\bullet - \boldsymbol{\rho}_{mk})$ are

$$\tilde{\boldsymbol{\mu}}_{mk} := \tilde{\boldsymbol{\mu}}_k + \boldsymbol{\rho}_{mk},\ \tilde{\boldsymbol{\Sigma}}_{mk} := \tilde{\boldsymbol{\Sigma}}_k. \tag{24}$$

Assumption 8. For given Q_k,

$$\forall \varepsilon > 0\ \exists \beta > 0 :\ \|\boldsymbol{\kappa} - \tilde{\boldsymbol{\kappa}}_k\| > \varepsilon \Rightarrow L(\tilde{\boldsymbol{\kappa}}_k, Q_k) - L(\boldsymbol{\kappa}, Q_k) > \beta.$$

Assumption 9. For $\tilde{\boldsymbol{\theta}} = (\xi_1, \ldots, \xi_k, \tilde{\boldsymbol{\kappa}}_1, \ldots, \tilde{\boldsymbol{\kappa}}_K)$: $\frac{\lambda_{\min}(\tilde{\boldsymbol{\theta}})}{\lambda_{\max}(\tilde{\boldsymbol{\theta}})} \leq \gamma$.

Theorem 5. *With the above definitions, under Assumptions 6 and 7, for large enough m, the clusters of $\boldsymbol{\theta}^\star(P_m)$ can be numbered in such a way that for $k \in \{1, \ldots, K\}$, with $\boldsymbol{\theta}^\star_m := \boldsymbol{\theta}^\star(P_m)$.*

$$\lim_{m\to\infty} \|\pi^\star_{mk} - \xi_k\| = 0, \tag{25}$$

and if $Q_1, \ldots, Q_K$ fulfill Assumption 9, then for those Q_k that fulfill Assumption 8:

$$\lim_{m\to\infty} \|\boldsymbol{\kappa}^\star_{mk} - \tilde{\boldsymbol{\kappa}}_{mk}\| = 0. \tag{26}$$

Proof. Assume throughout that mixture components are numbered according to eq. (22). Show the following statements:

S1 For $\boldsymbol{\theta}_m^\star = \boldsymbol{\theta}^\star(P_m)$, using the notation of eq. (3), and $\boldsymbol{\theta}_{mk}^\star$ denoting the parameters in $\boldsymbol{\theta}_m^\star$ belonging to mixture component k, $k = 1, \ldots, K$,

$$\boldsymbol{\theta}_{mk}^\star = \arg\max_{\boldsymbol{\theta}} \int \tau_k(\boldsymbol{x}; \boldsymbol{\theta}_{mk}^\star)(\log \pi_k + \log f(\boldsymbol{x}; \boldsymbol{\kappa}_k))dP_m(\boldsymbol{x}).$$

S2 eq. (25) follows from S1.

S3 From S1,

$$\boldsymbol{\kappa}_{mk}^\star = \arg\max_{\boldsymbol{\kappa}} q_{mk}(\boldsymbol{\kappa}),$$

where $q_{mk}(\boldsymbol{\kappa})$ is a function so that $q_{mk}(\boldsymbol{\kappa}) - \int \log f(\boldsymbol{x}; \boldsymbol{\kappa})dQ_k(\boldsymbol{x} - \boldsymbol{\rho}_{mk})$ converges uniformly to 0.

S4 eq. (26) follows from S3 and Assumption 8.

Proof of S1:
The proof is based on the two-step form of a mixture model involving for each observed X an unobserved random variable ζ indicating the mixture component that has generated X, see Section 2.1. S1 is the population version of the fixed point equation on which the EM-algorithm is based, see Section 3 and in particular Corollary 2 of Dempster et al. (1977).

Regarding the mixture eq. (2), on which the ML-estimation is based, let $\tilde{\psi}(\bullet; \boldsymbol{\theta})$ be the joint density of $Y = (X, \zeta)$, and $\bar{\psi}(\bullet|\boldsymbol{x}; \boldsymbol{\theta})$ be the conditional density of Y given $X = \boldsymbol{x}$ so that for all $\boldsymbol{y}$, $\boldsymbol{x}$:

$$\bar{\psi}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}) = \frac{\tilde{\psi}(\boldsymbol{y}; \boldsymbol{\theta})}{\psi(\boldsymbol{x}; \boldsymbol{\theta})}. \tag{27}$$

Define

$$\begin{aligned} G(\boldsymbol{\theta}'|\boldsymbol{\theta}) &:= \int\int \log(\tilde{\psi}(\boldsymbol{y}; \boldsymbol{\theta}'))\bar{\psi}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta})d\boldsymbol{y}dP(\boldsymbol{x}) \\ H(\boldsymbol{\theta}'|\boldsymbol{\theta}) &:= \int\int \log(\bar{\psi}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}'))\bar{\psi}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta})d\boldsymbol{y}dP(\boldsymbol{x}). \end{aligned}$$

Then, following Dempster et al. (1977), eq. (27) still holds averaged over the $y|\boldsymbol{x}$ assumed distributed according to $\bar{\psi}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta})$, and then averaging over $P_m(\boldsymbol{x})$ yields

$$L(\boldsymbol{\theta}', P) = G(\boldsymbol{\theta}'|\boldsymbol{\theta}) - H(\boldsymbol{\theta}'|\boldsymbol{\theta}).$$

For given $\boldsymbol{x}$, formula (1e6.6) in Rao (1965) implies

$$\begin{aligned} \int (\log \bar{\psi}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}) - \log \bar{\psi}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}'))\bar{\psi}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta})dy &\geq 0 \\ \Rightarrow H(\boldsymbol{\theta}|\boldsymbol{\theta}) &\geq H(\boldsymbol{\theta}'|\boldsymbol{\theta}). \end{aligned}$$

Let $M(\boldsymbol{\theta}_m^\star) := \arg\max_{\boldsymbol{\theta}} G(\boldsymbol{\theta}|\boldsymbol{\theta}_m^\star)$. Then

$$L(M(\boldsymbol{\theta}_m^\star)) = Q(M(\boldsymbol{\theta}_m^\star)|\boldsymbol{\theta}_m^\star) - H(M(\boldsymbol{\theta}_m^\star)|\boldsymbol{\theta}_m^\star) \geq G(\boldsymbol{\theta}_m^\star|\boldsymbol{\theta}_m^\star) - H(\boldsymbol{\theta}_m^\star|\boldsymbol{\theta}_m^\star) = L(\boldsymbol{\theta}_m^\star),$$

implying $\boldsymbol{\theta}^\star_m = \arg\max_{\boldsymbol{\theta}} G(\boldsymbol{\theta}|\boldsymbol{\theta}^\star_m)$ (otherwise the "$\geq$" above would be "$>$" and $M(\boldsymbol{\theta}^\star_m)$ would improve L). Note that

$$G(\boldsymbol{\theta}|\boldsymbol{\theta}^\star_m) = \int \sum_{k=1}^{K} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})(\log\pi_k + \log f(\boldsymbol{x};\boldsymbol{\kappa}_k))dP_m(\boldsymbol{x}),$$

which can be maximised for each k separately, leading to S1.

Proof of S2:
$\int \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})(\log\pi_k + \log f(\boldsymbol{x};\boldsymbol{\kappa}_k))dP_m(\boldsymbol{x})$ can be maximised separately over $\pi_1,\ldots,\pi_K$ and $\boldsymbol{\kappa}_1,\ldots,\boldsymbol{\kappa}_K$. Maximising $\int \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})\log\pi_k dP_m(\boldsymbol{x})$ yields

$$\pi^\star_{mk} = \int \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})dP_m(\boldsymbol{x}) = \int_{B_\varepsilon(\boldsymbol{\rho}_{mk})} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})dP_m(\boldsymbol{x}) + \int_{B_\varepsilon(\boldsymbol{\rho}_{mk})^c} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})dP_m(\boldsymbol{x}), \tag{28}$$

$B_\varepsilon(\boldsymbol{\rho}_{mk})$ chosen as in the proof of S4 of Theorem 4. From there, $\tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk}) \longrightarrow 1$ for $\tilde{\boldsymbol{x}} \in B_\epsilon(\mathbf{0})$, i.e., $\boldsymbol{x} \in B_\epsilon(\boldsymbol{\rho}_{mk})$, therefore

$$\lim_{m\to\infty} \left| \int_{B_\varepsilon(\boldsymbol{\rho}_{mk})} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})dP_m(\boldsymbol{x}) - \int_{B_\varepsilon(\boldsymbol{\rho}_{mk})} d\sum_{k=1}^{K} \xi_k Q_{mk}(\boldsymbol{x}) \right| = 0, \tag{29}$$

and furthermore, for $j \neq k$, $m \longrightarrow \infty$, because of eq. (15),

$$Q_{mj}(B_\varepsilon(\boldsymbol{\rho}_{mk})) \longrightarrow 0 \Rightarrow \int_{B_\varepsilon(\boldsymbol{\rho}_{mk})} d\xi_j Q_{mj}(\boldsymbol{x}) \longrightarrow 0. \tag{30}$$

ε can be chosen so large that $Q_{mk}(B_\varepsilon(\boldsymbol{\rho}_{mk}))$ is arbitrarily close to 1, and due to eq. (30), $\int_{B_\varepsilon(\boldsymbol{\rho}_{mk})} d\sum_{k=1}^{K} \xi_k Q_{mk}(\boldsymbol{x})$ is arbitrarily close to ξ_k. Furthermore, assuming m large enough that $B_\varepsilon(\boldsymbol{\rho}_{mj})$, $j = 1,\ldots,K$, do not intersect, with $\tilde{B} = \left(\bigcup_{j=1}^{K} B_\varepsilon(\boldsymbol{\rho}_{mj})\right)^c$,

$$\int_{B_\varepsilon(\boldsymbol{\rho}_{mk})^c} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})dP_m(\boldsymbol{x}) = \sum_{j\neq k} \int_{B_\varepsilon(\boldsymbol{\rho}_{mj})} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})dP_m(\boldsymbol{x}) + \int_{\tilde{B}} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})dP_m(\boldsymbol{x}).$$

For the same reason as eq. (29) and eq. (30) (with roles of j and k inverted),

$$\sum_{j\neq k} \int_{B_\varepsilon(\boldsymbol{\rho}_{mj})} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})dP_m(\boldsymbol{x}) \longrightarrow 0.$$

Because of eq. (15), $P_m(\tilde{B})$ is arbitrarily small for ε large enough, as is $\int_{\tilde{B}} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})dP_m(\boldsymbol{x})$. Putting everything together, eq. (28) implies eq. (25).

Proof of S3:
From S1, $\boldsymbol{\kappa}^\star_{mk} = \arg\max_{\boldsymbol{\kappa}} \tilde{q}_{mk}(\boldsymbol{\kappa})$ with

$$\tilde{q}_{mk}(\boldsymbol{\kappa}) = \int \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})\log f(\boldsymbol{x};\boldsymbol{\kappa}_k)dP_m(\boldsymbol{x}) = \int \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})\log f(\boldsymbol{x};\boldsymbol{\kappa}_k)d\sum_{k=1}^{K} \xi_k Q_{mk}(\boldsymbol{x}).$$

As in eq. (28) and the proof of S2,

$$\tilde{q}_{mk}(\boldsymbol{\kappa}) = \int_{B_\varepsilon(\boldsymbol{\rho}_{mk})} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})\log f(\boldsymbol{x};\boldsymbol{\kappa}_k)dP_m(\boldsymbol{x}) + \int_{B_\varepsilon(\boldsymbol{\rho}_{mk})^c} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk})\log f(\boldsymbol{x};\boldsymbol{\kappa}_k)dP_m(\boldsymbol{x}),$$

and

$$\int_{B_\varepsilon(\boldsymbol{\rho}_{mk})^c} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk}) \log f(\boldsymbol{x};\boldsymbol{\kappa}_k) dP_m(\boldsymbol{x}) = \sum_{j\neq k} \int_{B_\varepsilon(\boldsymbol{\rho}_{mj})} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk}) \log f(\boldsymbol{x};\boldsymbol{\kappa}_k) dP_m(\boldsymbol{x}) + \int_{\tilde{B}} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk}) \log f(\boldsymbol{x};\boldsymbol{\kappa}_k) dP_m(\boldsymbol{x}).$$

Because of Assumption 6, regarding $\arg\max_{\boldsymbol{\kappa}} q_{mk}(\boldsymbol{\kappa})$, it suffices to consider $\boldsymbol{\kappa}$ with $\lambda^*_{\min}(\boldsymbol{\Sigma}) > c_1$ for suitable $c_1 > 0$, $c_2 = \log(c_1^{-p/2} g(0))$, so that $\log f(\boldsymbol{x};\boldsymbol{\kappa}_k) < c_2$. On $B_\varepsilon(\boldsymbol{\rho}_{mk})$, according to the proof of Theorem 4, $\tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk}) \longrightarrow 1$ and, for $j \neq k$, $\tau_j(\boldsymbol{x};\boldsymbol{\theta}^\star_{mj}) \longrightarrow 0$, and furthermore, for $m \longrightarrow \infty$,

$$\int_{B_\varepsilon(\boldsymbol{\rho}_{mk})} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk}) \log f(\boldsymbol{x};\boldsymbol{\kappa}_k) dQ_{mj}(\boldsymbol{x}) < c_2 \int_{B_\varepsilon(\boldsymbol{\rho}_{mk})} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk}) dQ_{mj}(\boldsymbol{x}) \longrightarrow 0.$$

Therefore,

$$\lim_{m\to\infty} \left| \int_{B_\varepsilon(\boldsymbol{\rho}_{mk})} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk}) \log f(\boldsymbol{x};\boldsymbol{\kappa}_k) dP_m(\boldsymbol{x}) - \int_{B_\varepsilon(\boldsymbol{\rho}_{mk})} \log f(\boldsymbol{x};\boldsymbol{\kappa}_k) d\xi_k Q_{mk}(\boldsymbol{x}) \right| = 0. \tag{31}$$

For $j \neq k$, on $B_\varepsilon(\boldsymbol{\rho}_{mj})$: $\tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk}) \longrightarrow 0$, therefore

$$\int_{B_\varepsilon(\boldsymbol{\rho}_{mj})} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk}) \log f(\boldsymbol{x};\boldsymbol{\kappa}_k) dP_m(\boldsymbol{x}) < c_2 \int_{B_\varepsilon(\boldsymbol{\rho}_{mj})} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk}) dP_m(\boldsymbol{x}) \longrightarrow 0, \tag{32}$$

and

$$\int_{\tilde{B}} \tau_k(\boldsymbol{x};\boldsymbol{\theta}^\star_{mk}) \log f(\boldsymbol{x};\boldsymbol{\kappa}_k) dP_m(\boldsymbol{x}) < c_2 P_m(\tilde{B}). \tag{33}$$

Consider $q_{mk}(\boldsymbol{\kappa}) = \frac{\tilde{q}_{mk}(\boldsymbol{\kappa})}{\xi_k}$ in order to drop ξ_k from eq. (31). Once more ε can be chosen large enough that $P_m(\tilde{B})$ becomes arbitrarily small, and $Q_{mk}(B_\varepsilon(\boldsymbol{\rho}_{mk}))$ becomes arbitrarily large, so that eq. (31), eq. (32), and eq. (33) together imply that

$$q_{mk}(\boldsymbol{\kappa}) - \int \log f(\boldsymbol{x};\boldsymbol{\kappa}) dQ_{mk}(\boldsymbol{x}) \longrightarrow 0,$$

and the convergence is uniform over $\boldsymbol{\kappa}$ as neither $\tau_j(\boldsymbol{x};\boldsymbol{\theta}^\star_{mj})$ for any j nor ϵ nor c_2 depend on $\boldsymbol{\kappa}$.

Proof of S4:
Given arbitrarily small ε and the corresponding β from Assumption 8, according to S3, for m large enough, uniformly

$$|q_{mk}(\boldsymbol{\kappa}) - \tilde{L}(\boldsymbol{\kappa}, Q_k)| < \frac{\beta}{2}.$$

This means that $\boldsymbol{\kappa}$ with $\|\boldsymbol{\kappa} - \tilde{\boldsymbol{\kappa}}_{mk}\| > \epsilon$ cannot maximise $q_{mk}(\boldsymbol{\kappa})$, and therefore $\boldsymbol{\kappa}^\star_{mk} = \arg\max_{\boldsymbol{\kappa}} q_{mk}(\boldsymbol{\kappa})$ will become arbitrarily close to $\tilde{\boldsymbol{\kappa}}_{mk}$, because Assumption 9 enforces $\frac{\lambda_{\min}(\tilde{\boldsymbol{\theta}})}{\lambda_{\max}(\tilde{\boldsymbol{\theta}})} < \gamma$, which is also required to hold for $\boldsymbol{\theta}^\star(P^m)$. This proves eq. (26). □

Corollary 1. *In the situation of Theorem 5, requiring Assumptions 6 and 7, if g is chosen so that $f(\bullet;\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the density of a p-variate Gaussian distribution with mean $\boldsymbol{\mu}$*

and covariance matrix $\boldsymbol{\Sigma}$*, then*

$$\lim_{m\to\infty}\left\|\boldsymbol{\mu}_{mk}^{\star}-\int \boldsymbol{x}dQ_k(\boldsymbol{x})-\boldsymbol{\rho}_{mk}\right\|=0. \tag{34}$$

If additionally Assumption 9 holds, then

$$\lim_{m\to\infty}\left\|\boldsymbol{\Sigma}_{mk}^{\star}-\int(\boldsymbol{x}-\tilde{\boldsymbol{\mu}}_k)(\boldsymbol{x}-\tilde{\boldsymbol{\mu}}_k)^{\intercal}dQ_k(\boldsymbol{x})\right\|=0.$$

In this case, Assumption 6 will hold if Q_k is not concentrated on a single point, see Remark 4. Assumption 7 amounts to $\int\|\boldsymbol{x}^{\intercal}\boldsymbol{x}\|dQ_k(\boldsymbol{x})<\infty$.

Proof. Due to Assumption 6, consider $\boldsymbol{\kappa}$ with $\lambda_{\min}^{*}(\boldsymbol{\Sigma}_k)>c_1$ for $k=1,\ldots,K,\ c_1>0$. For the multivariate Gaussian (see, e.g., Anderson and Olkin (1985))

$$\begin{aligned}
\log f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= c-\frac{1}{2}\log|\boldsymbol{\Sigma}|-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}),\\
\frac{\partial}{\partial\boldsymbol{\mu}}\log f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{x}),\\
\frac{\partial}{\partial\boldsymbol{\Sigma}^{-1}}\log f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{2}\boldsymbol{\Sigma}-\frac{1}{2}(\boldsymbol{x}-\mu)(\boldsymbol{x}-\mu)^{\intercal}.
\end{aligned}$$

For $\boldsymbol{\mu}\in B_{\varepsilon}(\boldsymbol{\mu}_0)$ for any $\boldsymbol{\mu}_0$,

$$\left\|\frac{\partial}{\partial\boldsymbol{\mu}}\log f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\right\|\leq c_1^{-p}(\|\boldsymbol{\mu}_0-\boldsymbol{x}\|+\varepsilon).$$

For $\boldsymbol{\Sigma}^{-1}\in B_{\varepsilon}(\boldsymbol{\Sigma}_0{}^{-1})$ for any $\boldsymbol{\Sigma}_0$,

$$\left\|\frac{\partial}{\partial\boldsymbol{\Sigma}^{-1}}\log f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\right\|\leq\frac{1}{2}\left(\boldsymbol{\Sigma}_0+\varepsilon-(\boldsymbol{x}-\mu)(\boldsymbol{x}-\mu)^{\intercal}\right).$$

Both these bounding functions are integrable w.r.t. P_m because of Assumption 7. This means that $\log f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ can be differentiated under the integral sign. Standard algebra yields

$$\tilde{\boldsymbol{\mu}}_{mk}=\int xdQ_{mk}(x),\ \tilde{\boldsymbol{\Sigma}}_{mk}=\int(\boldsymbol{x}-\tilde{\boldsymbol{\mu}}_{mk})(\boldsymbol{x}-\tilde{\boldsymbol{\mu}}_{mk})^{\intercal}dQ_{mk}(\boldsymbol{x}).$$

Assumption 8 follows because $\log f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ is strictly concave in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{-1}$ (Anderson and Olkin (1985)), and the Corollary follows from Theorem 5. Assumption 9 is not required for eq. (34), because $\int xdQ_{mk}(x)$ maximises the Gaussian likelihood regardless of $\boldsymbol{\Sigma}$, cp. S3 and S4 in the proof of Theorem 5. □

6. Numerical experiments

In order to illustrate the results in Section 5, we present a small simulation study. We computed MLEs for mixtures of two different families of ESDs, namely Gaussian (MLE-N) and multivariate t_5 (MLE-T5), and applied them to mixtures with components that deviate from those assumed. We generated data from three different mixture models, each with $K=3$, namely a mixture of three bivariate t_3-distributions, a mixture of three skew normal distributions (Lee and McLachlan (2013)), and a mixture of one skew

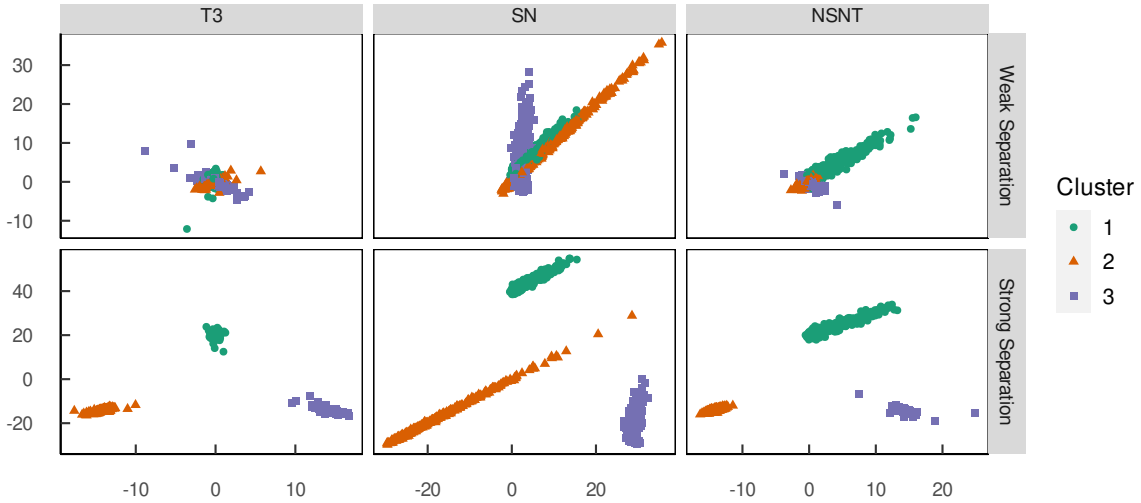

Figure 1: Data sets generated from mixture models T3, SN, and NSNT. "*Weak Separation*" is obtained setting $s = 0.5$ for T3 and NSNT, and $s = 2$ for SN. "*Strong Separation*" is obtained setting $s = 20$ for T3 and NSNT, and $s = 40$ for SN. The sample size is set to $n = 1000$ in all the six cases.

normal, one Gaussian and one t_3-distribution. The previous models are labeled T3, SN, and NSNT respectively. For each model we varied the amount of separation between mixture components. Parameters were chosen as follows. For all models $\pi_1 = \pi_2 = \pi_3 = 1/3$. Given a separation parameter $s > 0$, let

$$\boldsymbol{\mu}_1 = (0, s)^\intercal, \quad \boldsymbol{\mu}_2 = (-s/\sqrt{2}, -s/\sqrt{2})^\intercal, \quad \boldsymbol{\mu}_3 = (s/\sqrt{2}, -s/\sqrt{2})^\intercal,$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.55 & 0.45 \\ 0.45 & 0.55 \end{pmatrix}, \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 0.55 & -0.45 \\ -0.45 & 0.55 \end{pmatrix},$$

$$\boldsymbol{\delta}_1 = (5, 5)^\intercal, \ \boldsymbol{\delta}_2 = (15, 15)^\intercal, \ \boldsymbol{\delta}_3 = (1, 10)^\intercal.$$

In the T3 model, the three mixture components had means $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_3$ and covariance matrices $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, $\boldsymbol{\Sigma}_3$. s was chosen to range from 0.5 to 20. Skew normal distributions in the SN model were generated according to $\boldsymbol{\mu}_k + \boldsymbol{\delta}_k|Z_0| + Z_1$, where $Z_0 \sim \mathcal{N}_1(0, 1)$, $Z_1 \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}_k)$, $k = 1, 2, 3$, where $\mathcal{N}_p$ denotes the p-dimensional Gaussian distribution. If the k-th component of the mixture has a skew normal distribution, its expectation is $\boldsymbol{\mu}_k + \boldsymbol{\delta}_k\sqrt{2/\pi}$, and its covariance matrix is $\boldsymbol{\Sigma}_k + \boldsymbol{\delta}_k\boldsymbol{\delta}_k^\intercal(1 - 2/\pi)$. For this model, the separation parameter s was chosen from the range $s = 2$ to $s = 40$. For the mixed mixture model NSNT, the skew normal component was generated using the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\delta}_1$, the Gaussian component was generated using the mean $\boldsymbol{\mu}_2$ and the covariance matrix $\boldsymbol{\Sigma}_2$, and the t_3-component was generated using the mean $\boldsymbol{\mu}_3$ and the covariance matrix $\boldsymbol{\Sigma}_3$. For this model, s was chosen in the interval [1,20]. For each of the three models, we considered a grid of five (logarithmically equispaced) values of s for a total of 15 data generating processes. We also considered two sample sizes, $n = 1000, 10000$, for a total of 30 sample designs. Figure 1 shows examples of datasets with $n = 1000$ generated by the three models, with two different choices of the separation parameter s.

We ran 1000 Monte Carlo replicates with i.i.d. sampling. For each sample, we computed MLE-N and MLE-T5 with $\gamma = 100$. After computing the MLEs, the observations were assigned to clusters using the MAP rule. These were compared with the true mixture components using the Adjusted Rand Index (ARI; Hubert and Arabie (1985)). Results

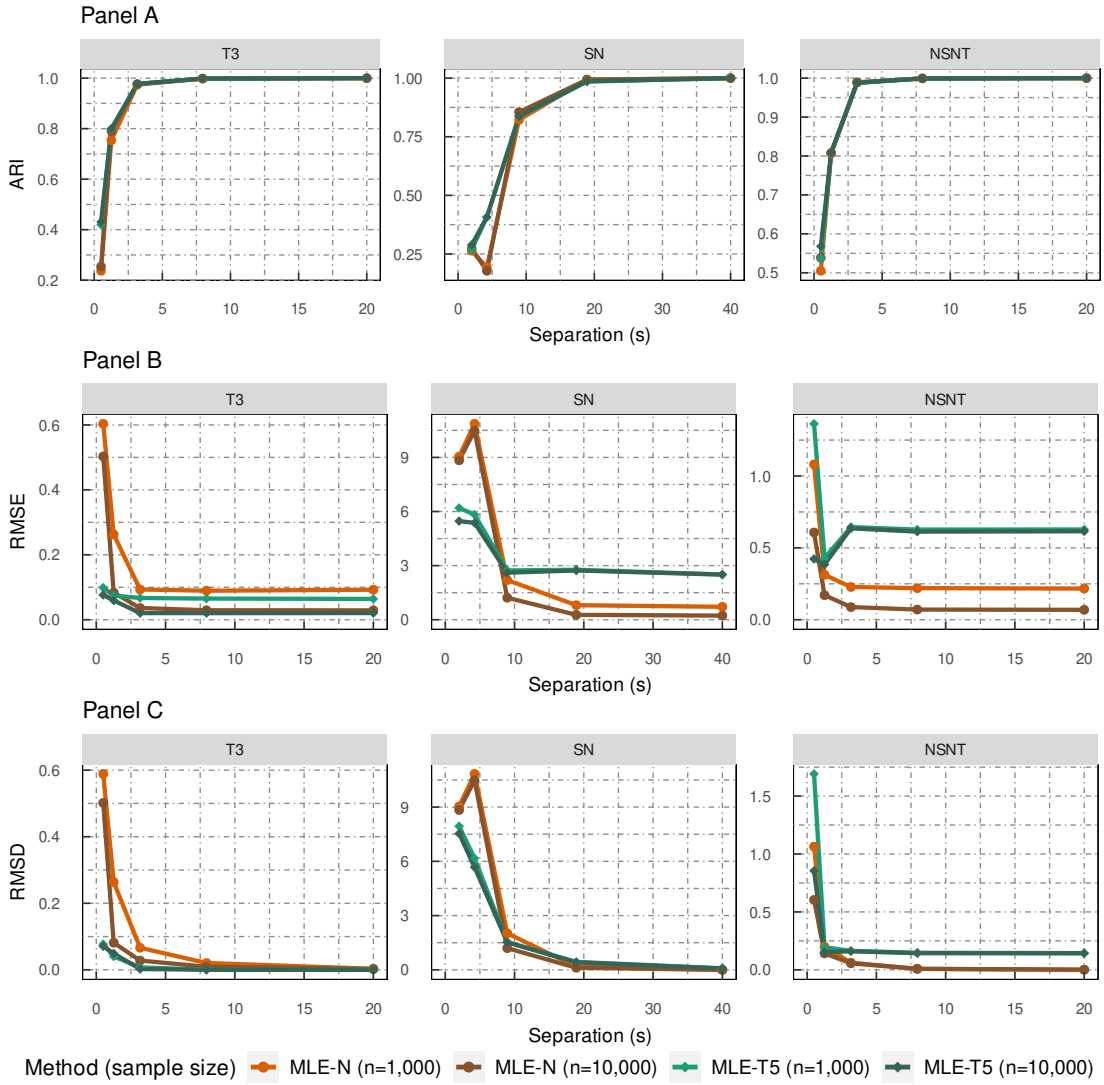

Figure 2: Monte Carlo averages for ARI (Panel A), RMSE for mean parameters (Panel B) and RMSD for ML mean functionals (Panel C).

are shown in Figure 2. Panel A shows the ARI averages over the Monte Carlo runs. Panel B shows the root mean square error (RMSE) of the mean parameters fitted by MLN-N and MLE-T5 compared to the mean of the generating mixture components. Panel C shows the root mean square deviation (RMSD) between the mean parameters estimated by the fit of MLE-N and MLE-T5, respectively, and the corresponding estimators computed on only the observations from the true mixture components, i.e., the sample mean (as to be compared to the MLE-N component mean estimators), and the ML estimator of location of a t_5-distribution (as to be compared to the MLE-T5 component location estimators). The latter would indicate the best performance one could expect, using true component information, estimating $\tilde{\boldsymbol{\kappa}}_{mk}$ from eq. (26) (Theorem 5) by ML.

The results show that with low separation s estimation and clustering do not work well. For the larger values of s, the ARI becomes almost perfect, which means that the clustering from the misspecified mixture model matches the true component memberships. This illustrates Theorem 4, although with $n = 1000$ already such a good classification is achieved that any improvement with $n = 10000$ is not visible.

The influence of growing n becomes clear looking at Panel B, where the improvement of the RMSE between $n = 1000$ and $n = 10000$ is obvious. For MLE-N the RMSE for $n = 10000$ is close to zero for all three setups. We are interested in particular in examining the RMSE for estimating the mean with the aim of illustrating Corollary 1. For t_3-distributions, the mean is the center of symmetry, and therefore the ML-estimator that treats the data as t_5 estimates the t_3-mean as well, and does this better (with lower variance) than the arithmetic mean. Correspondingly, MLE-T5 achieves a lower RMSE than MLE-N for the model T3, but both will converge to zero for $n \to \infty$. The models SN and NSNT involve asymmetric mixture components for which the ML location functional based on the t_5-distribution is not the same as the mean. For this reason, the RMSE for MLE-T5 comparing it to the mean will not converge to zero for $n \to \infty$, and consequently it seems to hit a nonzero floor in Panel B.

Panel C shows that for strong enough separation the location estimators from MLE-N and MLE-T5 become indistinguishable from the corresponding estimators computed based on the true component information for models T3 and SN, as should be expected from Theorem 5 and Panel A. For model NSNT, however, the RMSD of MLE-T5 does not seem to converge to zero. Surprised by this, we found numerically that the ML functional based on the t_5-distribution computed for the components separately violates $\frac{\lambda_{\min}(\tilde{\boldsymbol{\theta}})}{\lambda_{\max}(\tilde{\boldsymbol{\theta}})} \leq \gamma = 100$, i.e., Assumption 9, and therefore it cannot be recovered even asymptotically by MLE-T5. In fact, also the Gaussian ML functional violates $\frac{\lambda_{\min}(\tilde{\boldsymbol{\theta}})}{\lambda_{\max}(\tilde{\boldsymbol{\theta}})} \leq 100$ here, but this does not affect the consistency of the means from MLE-N, as Assumption 9 is not required for eq. (34).

7. Conclusion

Statistical model assumptions are not normally fulfilled in real life situations, and it is of interest how statistical methodology behaves in cases in which the model assumptions are not fulfilled. Consistency for the canonical functional means that for large data sets the estimator will stabilize even if its distributional assumptions are not fulfilled, although still assuming i.i.d. data. Results of this kind can be found in many places. Additionally here we look at specific underlying distributions P that are of such a kind that using the mixture MLE could be of interest for clustering, namely a mixture with K well separated components that can have a different distributional shape from what is assumed. The components of the canonical functional will then correspond in a well defined sense to the components of P, although potentially very strong separation is required. Without requiring strong enough separation, such a result cannot be had, and actually it would not be desirable, as the components Q_k of P can themselves be heterogeneous. If different Q_k are not well separated, from the point of view of clustering it may be more sensible to split up a heterogeneous, potentially multimodal, Q_k into subgroups rather than separate it from a $Q_l \neq Q_k$ from which it is not separated. That said, investigating the canonical functional in more general situations is certainly of interest. Furthermore, performance guarantees for fixed n would be welcome, although these will require more restrictive assumptions.

The presented results concern the global optimum of the likelihood function whereas existing algorithms are not guaranteed to find it. Similar results for local optima as found by the EM-algorithm would also be a desirable extension.

References

Anderson, T. and I. Olkin (1985). Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear Algebra and its Applications 70*, 147–171.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 719–725.

Chen, J. (2017). Consistency of the MLE under mixture models. *Statistical Science 32*(1), 47–63.

Ciuperca, G., A. Ridofi, and J. Idier (2003). Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics 30*, 45–59.

Coretto, P. and C. Hennig (2017). Consistency, breakdown robustness, and algorithms for robust improper maximum likelihood clustering. *Journal of Machine Learning Research 18*(142), 1–39.

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

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

Frühwirth-Schnatter, S., G. Celeux, and C. P. Robert (Eds.) (2019). *Handbook of Mixture Analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Boca Raton, FL: CRC Press.

Gallegos, M. T. and G. Ritter (2009). Trimmed ML estimation of contaminated mixtures. *Sankhyā. The Indian Journal of Statistics 71*(2, Ser.A), 164–220.

García-Escudero, L. A., A. Gordaliza, F. Greselin, S. Ingrassia, and A. Mayo-Iscar (2018). Eigenvalues and constraints in mixture modeling: geometric and computational issues. *Advances in Data Analysis and Classification 12*(2), 203–233.

García-Escudero, L. A., A. Gordaliza, C. Matrán, and A. Mayo-Iscar (2015). Avoiding spurious local maximizers in mixture modeling. *Statistics and Computing 25*, 1–15.

Genton, M. G. (Ed.) (2004). *Skew-elliptical distributions and their applications*. Boca Raton, FL: Chapman & Hall CRC.

Hartigan, J. A. (1981). Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association 76*(374), 388–394.

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

Hennig, C. (2010). Methods for merging gaussian mixture components. *Advances in Data Analysis and Classification 4*, 3–34.

Hennig, C., M. Meila, F. Murtagh, and R. Rocci (Eds.) (2016). *Handbook of cluster analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Boca Raton, FL: CRC Press.

Holzmann, H., A. Munk, and T. Gneiting (2006). Identifiability of finite mixtures of elliptical distributions. *Scandinavian Journal of Statistics. Theory and Applications 33*(4), 753–763.

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

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

Jammalamadaka, S. R. and S. Janson (2015). Asymptotic distribution of the maximum interpoint distance in a sample of random vectors with a spherically symmetric distribution. *The Annals of Applied Probability 25*(6), 3571–3591.

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

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā (Statistics). The Indian Journal of Statistics. Series A 32*, 419–438.

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

Lee, S. X. and G. J. McLachlan (2013). Model-based clustering and classification with non-normal mixture distributions (with discussion). *Statistical Methods and Applications 22*, 427–454.

McLachlan, G. and D. Peel (2000). *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. New York: John Wiley & Sons, Inc.

McLachlan, G. J. and T. Krishnan (1997). *The EM Algorithm and Extensions*. New York: Wiley.

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

Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics 9*(1), 135–140.

Rao, C. R. (1965). *Linear Statistical Inference and its Applications*. New York: Wiley-Interscience.

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

Redner, R. A. and H. F. Walker (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review 26*, 195–239.

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

Sriperumbudur, B. and I. Steinwart (2012). Consistency and rates for clustering with dbscan. In N. D. Lawrence and M. Girolami (Eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, Volume 22 of *Proceedings of Machine Learning Research*, La Palma, Canary Islands, pp. 1090–1098. PMLR.

van der Vaart, A. and J. A. Wellner (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. New York: Springer-Verlag.

von Luxburg, U., M. Belkin, and O. Bousquet (2008). Consistency of spectral clustering. *The Annals of Statistics 36*(2), 555–586.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics 20*, 595–601.