Journal of Statistical Theory and Practice (2022) 16:60
https://doi.org/10.1007/s42519-022-00289-1

ORIGINAL ARTICLE

Estimation of the Complexity of a Finite Mixture Distribution: From Well- to Less Known Methods

Fadoua Balabdaoui[1] · Andrei Kolar[2] · Yulia Kulagina[1] · Lilian Müller[2]

Accepted: 30 July 2022 / Published online: 25 August 2022

Abstract

Mixture models occur in numerous settings including random and fixed effects models, clustering, deconvolution, empirical Bayes problems and many others. They are often used to model data originating from a heterogeneous population, consisting of several homogeneous subpopulations, and the problem of finding a good estimator for the number of components in the mixture arises naturally. Estimation of the order of a finite mixture model is a hard statistical task, and multiple techniques have been suggested for solving it. We will concentrate on several methods that have not gained much popularity yet deserve the attention of practitioners. These can be categorized into three groups: tools built upon the determinant of the Hankel matrix of moments of the mixing distribution, minimum distance estimators, likelihood ratio tests. We will address theoretical pillars underlying each of the methods, provide some useful modifications for enhancing their performance and present the results of the comparative numerical study that has been conducted under various scenarios. According to the results, none of the methods proves to be a "magic pill". The results uncover limitations of the techniques and provide practical hints for choosing the best-suited tool under specific conditions.

Keywords Mixture complexity estimation · Moments · Minimum distance · Likelihood-ratio test · Multilayer perceptron

✉ Fadoua Balabdaoui
fadoua.balabdaoui@stat.math.ethz.ch

Andrei Kolar
andreikolartoronto@gmail.com

Yulia Kulagina
yulia.kulagina@stat.math.ethz.ch

Lilian Müller
lilian.mueller@gmail.com

1 Seminar for Statistics, D-MATH, ETH Zurich, Rämistrasse 101, Zürich 8092, Switzerland

2 D-MATH, ETH Zurich, Rämistrasse 101, Zürich 8092, Switzerland

1 Introduction

1.1 Aim and Scope

In multiple applications the collected data may be best described by a multimodal probability mass or density function meaning the empirical distribution contains several regions with high probability mass. Mixture models are a powerful mathematical tool that allows for characterizing such heterogeneous populations, which are believed to consist of multiple homogeneous subpopulations. A great multitude of statistical problems can be cast into the mixture model framework: linear inverse and deconvolution problems, random effects models, repeated measures and measurements error models, empirical and hierarchical Bayes, latent class and latent trait models, clustering, robustness and contamination models, hidden mixture structures, random coefficient regression models and many others [43]. A very important class of mixture models is the class of *finite mixtures*. These models, which assume a finite number of components, have proved to be very useful and flexible enough to model a vast range of random phenomena thus receiving much attention from both theoretical and practical viewpoints.

In some mixture models applications there is no uncertainty about the number of components in the mixture. This is the case where the components correspond to a well-known existing partition of the population. However, on many occasions this situation is far from realistic and practitioners encounter either the lack or complete absence of a priori information about the actual number of mixture components. In such cases, this number has to be inferred from the data along with the parameters of the component densities. Correct identification of mixture complexity may be of primary interest in itself or may be followed by efficient estimation of all parameters. Due to its practical importance the problem of selecting the optimal mixture complexity has been addressed in numerous statistical publications, and we will point out many seminal works as we proceed.

The objectives of the present survey are to

- provide the theoretical background of the reviewed methods for estimating the complexity of a finite mixture;
- assess the performance of these methods under various scenarios;
- suggest modifications that enhance the performance of some of the methods in particular settings;
- identify universal methods that provide stable and accurate results throughout most of the scenarios for different distribution families or single out scenarios under which certain approaches may be preferred to others.

The number of methods devoted to estimating the true number of components in a mixture is undoubtedly too large to be thoroughly described in a single survey. Thus, we restrict attention to a selected subset of approaches that have the merit of being applicable in very general settings; i.e., for wide classes of finite mixtures of distributions. One of the main goals of this survey is to uncover the extent to which each of the methods is successful in consistently estimating the true complexity for

various sample sizes. The estimation techniques reviewed below can be split into three main categories:

1. methods built upon the determinants of the Hankel matrix of moments of the mixing distribution;
2. methods based on penalized minimum distance between the unknown probability density and a consistent estimator thereof. The distances considered in this survey are the Hellinger as well as the L_2-distance;
3. likelihood ratio test (LRT) - based techniques.

Some of the key criteria we based our choice of the techniques upon were:

a) a cohesive mathematical theory behind the method, including the asymptotically consistency;
b) infrequent reference in the literature as well as relatively rare usage in practice despite of the coherent theoretical base;
c) feasibility of implementation using any programming language, e.g. Python, R, Julia, Matlab, etc.

Pseudocodes for the algorithms discussed in this work are given in Appendix D in the supplementary materials. For completeness, other interesting methods for estimating the complexity of a finite mixture are mentioned in Sect. 8.

Although not strictly a part of a survey, the performance enhancement such as the one we bring to some of the methods through specific modifications is almost inevitable. In fact, the original version of some of the approaches reviewed here cannot be of real practical value without any further adjustment. These modifications, which will be described in separate subsections, include resorting to some judicious scaling in the case of the Hankel-based-methods or using bootstrap instead of penalization for the approaches based on minimum distance estimation. In Sect. 6, we report the results of an extensive numerical study which we carried out for different mixture distributions and various number of components with the goal of comparing the performances of the techniques reviewed in this survey.

Several examples involving the estimation of mixture complexity for real data sets using the discussed methods are presented. The data sets were taken from various fields such as geology, insurance and lexicography.

1.2 Organization of the Paper

The paper is organized as follows:

- Section 2 provides some basic background on mixture models, mentions major works on mixture model estimation techniques and gives a brief overview of these approaches.
- Section 3 outlines the theoretical foundation of the original method based on the determinants of the Hankel matrix of moments of the mixing distribution as proposed in [21]. In the same section, two modifications of this approach allowing for obtaining improved results, are presented. The section also gives a concise description of a neural network extension of the Hankel matrix approach, proposing

possible working configurations of a multilayer perceptron for the mixtures of Gaussian, Poisson and geometric densities.

- Section 4 describes methods based on minimum distance estimation. The section relies to a very large extent on the works [74], [69] and [20]. We re-examine the estimation techniques that use the Hellinger and the L_2 distances when combined with two different penalties. In the same section, motivated by the idea of enhancing the original method, we propose a modification based on a bootstrap procedure instead of penalization.
- Section 5 presents the estimation approach based on the LRT combined with a bootstrap procedure as described in [38].
- Section 6 comprises the results of a comparative numerical study where all of the above mentioned techniques are tested on simulated data under various scenarios. Furthermore, the same section contains a discussion of the settings in which certain methods can be favored as they seem to outperform their counterparts.
- Section 7 encompasses several real data sets that were analysed using the studied approaches and compares the obtained results.
- Section 8 mentions a number of papers where other techniques for mixture complexity estimation, not addressed in the present survey, are considered.
- Section 9 summarizes the findings and outlines the limitations of all methods reviewed in this survey.
- Appendices A, B, C, D presented as supplementary materials to this paper include additional examples, tables with detailed simulation results, proofs of the theoretical results that are relevant for the methods described in the manuscript and pseudocodes clarifying and simplifying the implementation of the discussed techniques.

2 Finite Mixture Models: General Scope

2.1 Notation and Basic Definitions

We start with defining the terminology that will be used throughout the survey. In the sequel, a real vector of dimension r will be denoted $\boldsymbol{v}_r$ and its components by $v_1, \ldots, v_r$. A class of real vectors of dimension r will also bear the subscript r in its notation. When manipulating several vectors of dimension r we will index them as $\boldsymbol{v}_{r,1}, \boldsymbol{v}_{r,2}, \ldots$. A random sample of i.i.d. random variables are going to be denoted $(X_1, \ldots, X_n)$. Also, a sequence of random variables (for example converging weakly) will be denoted for example by $Y^{(n)}$. A class of densities which depend on some vector of parameters of dimension r will not necessarily bear the subscript r.

Suppose that some population of interest, represented abstractly by a random variable X, consists of a finite number $m \in \mathbb{N}$ of subpopulations. Each subpopulation is generated by some random process that can be modeled by an individual or component distribution, e.g. normal, exponential, Poisson, geometric, etc. We will assume that each of the component distributions admits a density with respect to some common dominating measure μ. Furthermore, the component density is assumed to be parametrized through some unknown vector $\boldsymbol{\phi}_d \in \boldsymbol{\Phi} \subseteq \mathbb{R}^d$, $d \geq 1$. To keep the

manuscript to a reasonable length, we confine our attention in this survey to the one-dimensional case; i.e. the random variable $X \in \mathbb{R}$. In the sequel, the dominating measure μ is either the Lebesgue measure in case the distribution of the components is absolutely continuous, or the counting measure in case this distribution is discrete. In the latter case, all the examples considered here treat distributions that are supported on the set of non-negative integers. Let $\mathcal{F} = \left\{ f_{\boldsymbol{\phi}_d} : \boldsymbol{\phi}_d \in \boldsymbol{\Phi} \right\}$ be the family of densities which the components belong to. If $\mathcal{X}$ is the support of X, then the distribution of X is said to have a m-component mixture distribution with density

$$f_{\boldsymbol{\theta}_{p_m}}(x) = \int_{\Phi} f_{\boldsymbol{\phi}_d}(x) dG(\boldsymbol{\phi}_d) = \sum_{j=1}^{m} \pi_j f_{\boldsymbol{\phi}_{d,j}}(x) \tag{2.1}$$

for all $x \in \mathcal{X}$, where

$$\begin{aligned} &\boldsymbol{\theta}_{p_m} = (\pi_1, \ldots, \pi_m, \boldsymbol{\phi}_{d,1}^T, \ldots, \boldsymbol{\phi}_{d,m}^T)^T \in \mathcal{S}_{m-1} \times \boldsymbol{\Phi}^m := \boldsymbol{\Theta}_{p_m}, \\ &\mathcal{S}_{m-1} = \left\{ (\pi_1, \ldots, \pi_m)^T \in [0,1]^m : \sum_{j=1}^{m} \pi_j = 1 \right\}. \end{aligned} \tag{2.2}$$

$\mathcal{S}_{m-1}$ is the $(m-1)$-dimensional simplex and $\boldsymbol{\Phi}^m$ is the Cartesian product $\{(\boldsymbol{\phi}_{d,1}, \ldots, \boldsymbol{\phi}_{d,m})^T : \boldsymbol{\phi}_{d,i} \in \boldsymbol{\Phi}, i = 1, \ldots, m\}$ with $p_m = md + m - 1$.

Above, G is a discrete distribution defined on Φ with at most m jump points at $\boldsymbol{\phi}_{d,1}, \ldots, \boldsymbol{\phi}_{d,m}$, and the integral representation in (2.1) is given here only to draw attention that finite mixtures are part of a much bigger family of mixtures where G can be any distribution function, known often under the name of "mixing distribution". In the sequel, we will refer to either the probability density or probability mass function defined in (2.1) as the mixed density and to π_j, $j = 1, \ldots, m$ as the mixing probabilities.

We define the family of m-component mixture densities as the set

$$\begin{aligned} \mathcal{F}_m &= \left\{ \pi_1 f_{\boldsymbol{\phi}_{d,1}} + \ldots + \pi_m f_{\boldsymbol{\phi}_{d,m}}, (\pi_1, \ldots, \pi_m)^T \in \mathcal{S}_{m-1}, (f_{\boldsymbol{\phi}_{d,1}}, \ldots, f_{\boldsymbol{\phi}_{d,m}}) \in \mathcal{F}^m \right\} \\ &= \left\{ f_{\boldsymbol{\theta}_{p_m}} : \boldsymbol{\theta}_{p_m} \in \boldsymbol{\Theta}_{p_m} \right\}, \end{aligned}$$

where $f_{\boldsymbol{\theta}_{pm}}$ is given by (2.1).

Suppose we observe n random variables $X_1, \ldots, X_n \in \mathcal{X}$ which are i.i.d. according to an unknown density $\sim f_0 \in \bigcup_{m \geq 1} \mathcal{F}_m$. What is the value of m that can be assigned to this density based on the observed data? It is clear that such a value needs to target the most parsimonious representation of the mixture. Estimation of the true complexity of f_0 cannot be presented without touching upon this point, which is discussed in the next section.

2.2 Identifiability and Complexity of Mixture Models

We will now touch upon the identifiability issues arising within the mixture distributions framework, which is a crucial point when the aim is to estimate the true complexity. Identifiability of general mixtures of some additively closed family of distributions was proved in the pioneer work of Teicher [65], who recognized the importance of settling the issue of identifiability before launching into estimation of the mixing distribution. Several articles have been devoted to proving identifiability of finite mixtures of some particular classes of distributions such as finite mixtures of normal or gamma distributions; see e.g. [66]. For identifiability results in other classes or review papers on the subject we can refer to [19, 35, 36, 43, 49, 68].

The identifiability of a mixture is defined as follows: a finite mixture with respect to the family $\mathcal{F}$ is said to be *identifiable* if for any $m \geq 1$ and any two elements $f_{\boldsymbol{\theta}_m}$ and $f_{\boldsymbol{\theta}'_m}$ in $\mathcal{F}_m$ satisfying the equality

$$f_{\boldsymbol{\theta}_{p_m}}(x) = f_{\boldsymbol{\theta}'_{p_m}}(x), \quad x \in \mathcal{X}$$

then there exists a permutation $\sigma : \{1, \ldots, p_m\} \mapsto \{1, \ldots, p_m\}$ such that the components of $\boldsymbol{\theta}_{p_m}$ and $\boldsymbol{\theta}'_{p_m}$ are equal up to the permutation σ; i.e., $\pi_{\sigma(i)} = \pi'_i$, $\boldsymbol{\phi}_{d,\sigma(i)} = \boldsymbol{\phi}'_{d,i}$ for $i = 1, \ldots, p_m$.

Different techniques have been developed to show identifiability. One of the most important results is the one shown in [76], which says that the characterizing condition of identifiability is linear independence of the family $\mathcal{F}$. Other characterizations or sufficient conditions could be built upon this result by resorting for example to using some additional properties of the elements of $\mathcal{F}$ or computing Fourier/ Laplace transforms (see [5, 36, 39]).

When identifiability holds, it is natural to think of the most economic representation of the finite mixture under study. Indeed, we have the inclusions

$$\mathcal{F}_m \subset \mathcal{F}_{m+1} \tag{2.3}$$

for all $m \geq 1$, and hence we can introduce the following definition: the *index of economical representation* for some finite mixture density $f \in \bigcup_{m \geq 1} \mathcal{F}_m$ is defined as

$$m(f) = \min \left\{ m \in \mathbb{N} : f \in \mathcal{F}_m \right\}.$$

This index is exactly what is called the *complexity* (or *order*). Note that this number has to be unique, an immediate consequence of identifiability. Also, from a practical point of view, $m(f)$ corresponds to the number of all the components that are actually part of the total population: all the mixing probabilities π_j, $j \in \{1, \ldots, m(f)\}$ should satisfy $\pi_j > 0$ by the very definition of $m(f)$. The term identifiability is used here with some abuse as the components of $\boldsymbol{\theta}_{p_{m(f)}}$ are unique up to some permutation (whereas the mixed density is invariant under the $m!$ permutations of the component labels). One can of course require for example that the mixing probabilities are labeled so that

$\pi_1 < \cdots < \pi_m$ in case they are all different. We will be following this convention when reporting the simulation results in Sect. 6.

The discussion above lays the ground for this survey. In the sequel, we shall assume that identifiability assumption holds. Also, the notation $m(f_0) = m_0$ will be used, where f_0 is the unknown density in $\mathcal{F}_{m_0}$ from which we observe a random sample. The true complexity or order, m_0, as well as the true parameter vector

$$\boldsymbol{\theta}_0 := \boldsymbol{\theta}_{p_{m_0}} \in \boldsymbol{\Theta}_{p_{m_0}}$$

will be assumed to be unknown. The main goal of the methods reviewed further is to consistently estimate m_0. An estimation procedure can be (but does not necessarily have to be) accompanied by the estimation of $\boldsymbol{\theta}_0$.

2.3 Popular Approaches to Mixture Model Estimation

Mixture model estimation has a long history. The early mixture model estimation techniques date back to the end of the 19-th century, when S. Newcomb [52] suggested an iterative reweighting scheme to compute the Maximum Likelihood (ML) estimator of the common mean of a mixture of a known proportions of a finite number of univariate normal populations with known variances. This scheme is regarded by many as a precursor of the well-known Expectation-Maximization (EM) algorithm.

A few years later K. Pearson [56] described an analytical and a graphical solutions to estimating the first five moments of an asymmetrical empirical distribution, which he was aiming to break up into two univariate normal curves. The graphical solutions for mixture model estimation stayed in the focus of attention until the second half of the 20-th century ([14, 34, 58]).

Between 1912 and 1922 R. Fisher [29] attempted to popularize the ML approach to fitting the mixtures. The evolution of the ML approach is considered in detail in [3]. In particular, Fisher made an analysis of the extensions of the method of moments to the likelihood equations as a way of increasing the quality of the estimates, which later caused a dispute with Pearson ([28, 57]). Around the 1950s C. R. Rao [59] used Fisher's scoring method to estimate the parameters of a mixture of two Gaussian distributions with common variance, and soon after the ML estimation for identifying the number of components as well as for parameter estimation in finite mixture models was addressed in numerous publications, such as [22, 72, 73].

These days the most well-studied and widely-used approach to computing ML estimates for finite mixture models as defined in (2.1), is the EM algorithm, elaborately described in [23], the seminal work that greatly exhilarated the efficient usage of mixture models. The EM algorithm is implemented by assuming that there are latent variables that link every observation to one of the components, which, together with the observed data, yield complete data.

We will summarize the main idea behind this algorithm. To that end consider two sample spaces within the mixture model framework:

1. the sample space of the incomplete observations, where only the realizations of the random variable X are observed, but no information on the mixing distribution $G(\boldsymbol{\phi}_d)$ is available;
2. the sample space of the complete observations, the estimation of which can be performed explicitly.

For the sake of simplicity consider the one-dimensional case, $\Phi \subseteq \mathbb{R}$. In this case we denote $\boldsymbol{\phi}_d$ simply by ϕ. The extension to the multidimensional case is possible but complicates the derivations.

Let $x = (x_1, \ldots, x_n)$ be the observed realizations of the random variable X, and let $z = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ denote the realizations of the corresponding unobserved (or latent) random vector $\mathbf{Z}$ indicating that the observation $x_i, i = 1, \ldots, n$ comes from the j-th component, $j = 1, \ldots, m$. In other words, $\mathbf{z}_i, i = 1, \ldots, n$ are realizations of a multinomial distribution with probabilities $\pi_1, \ldots, \pi_m$, and we have that

$$z_{ij} = \begin{cases} 1, & \text{if } x_i \in j^{th} \text{ component} \\ 0, & \text{otherwise.} \end{cases}$$

The pairs $\boldsymbol{y}_i = (x_i, \boldsymbol{z}_i)$, for $i = 1, \ldots, n$ are i.i.d. and they are usually referred to as the complete or augmented data. Let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$. For a stipulated mixture complexity $m \in \mathbb{N}$, let us denote by $l^c_{\boldsymbol{\theta}_{pm}}$ the log-likelihood of the complete data; i.e.,

$$\begin{aligned} l^c_{\boldsymbol{\theta}_{pm}}(\boldsymbol{y}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij} \log\left(\pi_j f_{\boldsymbol{\phi}_{d,j}}(x_i)\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij} \log(f_{\phi_j}(x_i)) + \sum_{j=1}^{m} \log(\pi_j) \sum_{i=1}^{n} z_{i,j}. \end{aligned}$$

On the other hand, the log-likelihood of the observed data x is given by

$$l_{\boldsymbol{\theta}_{pm}}(x) = \sum_{i=1}^{n} \log\Big(\sum_{j=1}^{m} \pi_j f_{\phi_j}(x_i)\Big).$$

It can be shown that the MLE

$$\hat{\boldsymbol{\theta}}_{pm} = \arg\max_{\boldsymbol{\theta}_{pm} \in \boldsymbol{\Theta}_{pm}} l_{\boldsymbol{\theta}_{pm}}(x). \tag{2.4}$$

can be obtained by alternating between an expectation and maximization steps involving both the complete log-likelihood $l^c_{\boldsymbol{\theta}_{pm}}$. This is precisely what the well-known EM-algorithm does. In the first step, the conditional expectation of $l^c_{\boldsymbol{\theta}_{pm}}(\boldsymbol{y})$ given the observed data x is computed under the current parameter. Then, the obtained expression is maximized over the parameter space and the maximizer becomes the new parameter. These two steps are repeated until convergence. If s is the number of the

iteration of the current E-step, then it is easy to show that this step is completed by computing the conditional expectation of the multinomial vectors z_i given the observed data x. This yields for $i = 1, \ldots, n$ and $j = 1, \ldots, m$

$$\hat{z}_{ij}^{(s} = \frac{\hat{\pi}_j^{(s-1)} f_{\hat{\phi}_j^{(s-1)}}(x_i)}{\sum_{l=1}^{m} \hat{\pi}_l^{(s-1)} f_{\hat{\phi}_l^{(s-1)}}(x_i)}$$

where $(\hat{\pi}_1^{(s-1)}, \ldots, \pi_m^{(s-1)}, \hat{\phi}_1^{(s-1)}, \ldots, \hat{\phi}_m^{(s)})$ is the MLE obtained at the $(s-1)$-th step. Note that the maximizing mixing probabilities are easily obtained and are explicitly given in the s-th M-step by the expression

$$\hat{\pi}_j^{(s)} = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{ij}^{(s)},$$

for $j = 1, \ldots, m$. To obtain $\hat{\phi}_j^{(s)}$, $j = 1, \ldots, m$, a numerical method might be required in case a a closed form is not possible. The optimization procedure then seeks to find at least the local maximum as finding the global maximum is not always possible. As noted in [50], the latter often occurs in the case of Gaussian mixtures with non-homogeneous dispersions (unequal covariance matrices). Components that have either one observation, or several identical observations or several nearly-identical observations, result in the estimated covariance matrices that are singular, which causes the likelihood function to be unbounded. Gaussian mixtures with homogeneous components result in covariance matrices that are restricted in the parameter space and thus do not have this problem. For references on the EM-algorithm, see e.g. [23] and [48].

The description given above treats one given m, a candidate for the true mixture complexity. To obtain an estimator for m_0, the true complexity, one can resort to maximizing a penalized version of the observed log-likelihood. This means that the log-likelihood will be augmented by a penalty term depending on the model complexity. Several widely used examples of this technique include Akaike Information Criterion (AIC) [2], Bayes Information Criterion (BIC) [62], Integrated Completed Likelihood (ICL) [10], Laplace-Empirical Criterion (LEC) [49], Normalized Entropy Criterion (NEC) [9] and many others [50]. These only differ in the form of the penalty function, and we will concentrate on the two criteria that have gained most popularity in practice: The Bayesian Information Criterion (BIC) and Integrated Completed Likelihood (ICL). While BIC is most widely used for performing model selection tasks, ICL is most frequently applied for solving clustering problems.

The general idea is to treat the task of choosing the number of components in the mixture as a model selection problem by considering a sequence of models $\mathcal{M}_1, \ldots, \mathcal{M}_M$ for $m = 1, \ldots, M$ with associated prior probabilities $p(\mathcal{M}_m)$, which are often taken to be equal. By the Bayes' Theorem, the posterior probability of model $\mathcal{M}_m$, given the observed data $\boldsymbol{x}$ is proportional to the probability of the data given the model multiplied by the model's prior. Under regularity assumptions, it can be show than twice the posterior probability of the mixture model with m components can be

well approximated by the

$$\mathrm{BIC}_m = 2l_{\hat{\boldsymbol{\theta}}_{p_m}}(\boldsymbol{x}) - \nu_{\mathcal{M}_m} \log n$$

where $\nu_{\mathcal{M}_m} = p_m$ is the number of independent parameters in the model and $\hat{\boldsymbol{\theta}}_{p_m}$ is the MLE of $\boldsymbol{\theta}_{p_m}$. The true complexity is then estimated by finding the integer m which maximizes BIC_m.

Given the discussion above, finding the number of components in the mixture that maximizes $m \mapsto \mathrm{BIC}_{p_m}$ is equivalent to choosing the mixture model with the greatest a posteriori probability. Some of the advantages of the BIC approach are that it is easy to implement, can be used for comparing non-nested models and was shown to be consistent for choosing the correct number of components in [40].

The ICL approach uses the log-likelihood of the complete data and replaces the unobserved labels $z_{ij}, 1 \le i \le n, 1 \le j \le m$ by their maximum a posteriori (MAP) estimator, that is

$$\hat{z}_{ij}^* = \begin{cases} 1, & \text{if } \hat{z}_{ij} = \arg\max_{1 \le k \le m} \hat{z}_{ik} \\ 0, & \text{otherwise.} \end{cases}$$

Thus, for the mixture model with m components

$$\mathrm{ICL}_m = 2l^c_{\hat{\boldsymbol{\theta}}_{p_m}}\left(\boldsymbol{x}, \boldsymbol{z}^*\right) - p_m \log n.$$

The very useful relationship between BIC_m and ICL_m can be shown:

$$\mathrm{ICL}_m = \mathrm{BIC}_m + \sum_{i=1}^{n}\sum_{j=1}^{m} \hat{z}_{ij} \log \hat{z}_{ij}.$$

It has been shown in [31] that in some cases (e.g. for the mixtures of Gaussians) evaluating the likelihood at the a maximum a posteriori (MAP) estimator instead of the MLE helps the EM algorithm to avoid singularities or degeneracies.

Regularization and variable selection techniques have also found their application in this setting. For example, [55] proposed an estimation technique for Gaussian mixtures in the context of a clustering problem, where the likelihood function is augmented by an L_1-norm penalty term $-\lambda \sum_{j=1}^{m} \sum_{k=1}^{p} |\mu_{jk}|$, where μ_{jk} is the k-th coordinate of the j-th mean vector, and derived a modification of an EM algorithm fitted for the purpose. The L_1 penalty can shrink some of the fitted means toward 0, thus leading to the most parsimonious model.

Example 1: EM solution for the mixture of Gaussian distributions. For a finite mixture of univariate Gaussian distributions with the parameter vector $\boldsymbol{\theta} =$

$\left(\pi_1, \ldots, \pi_m, (\mu_1, \sigma_1), \ldots, (\mu_m, \sigma_m)\right)$ and the mixture density given by

$$f_{\boldsymbol{\theta}}(x) = \sum_{j=1}^{m} \pi_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp^{-\frac{1}{2}\left(\frac{x-\mu_j}{\sigma_j}\right)^2},$$

the E-step at the s-th iteration will update the probabilities given the current parameter vector $\boldsymbol{\theta}^{(s-1)}$

$$\hat{z}_{ij}^{(s)} = \frac{\hat{\pi}_j^{(s-1)} \frac{1}{\sqrt{2\pi}\hat{\sigma}_j^{(s-1)}} \exp^{-\frac{1}{2}\left(\frac{x-\hat{\mu}_j^{(s-1)}}{\hat{\sigma}_j^{(s-1)}}\right)^2}}{\sum_{j'=1}^{m} \hat{\pi}_{j'}^{(s-1)} \frac{1}{\sqrt{2\pi}\hat{\sigma}_{j'}^{(s-1)}} \exp^{-\frac{1}{2}\left(\frac{x-\hat{\mu}_{j'}^{(s-1)}}{\hat{\sigma}_{j'}^{(s-1)}}\right)^2}}.$$

The M-step provides the following solutions:

$$\hat{\pi}_j^{(s)} = \frac{\sum_{i=1}^{n} \hat{z}_{ij}^{(s)}}{n}, \quad \hat{\mu}_j^{(s)} = \frac{\sum_{i=1}^{n} \hat{z}_{ij}^{(s)} x_i}{\sum_{i=1}^{n} \hat{z}_{ij}^{(s)}}, \quad \hat{\sigma}_j^{(s)} = \frac{\sum_{i=1}^{n} \hat{z}_{ij}^{(s)} (x_i - \hat{\mu}_j^{(s)})^2}{\sum_{i=1}^{n} \hat{z}_{ij}^{(s)}}.$$

Further examples (for the mixtures of geometric and Poisson distributions) can be found in Appendix A in the supplementary materials.

Concluding this section it is necessary to point out that a great amount research has been carried out in this area, and multiple software applications have been developed for working with mixture models, in particular with the Gaussian mixture models that are most frequently used in practice. Most of the software is suited for model-based classification and in particular offering the opportunity to find the ML estimates via the EM algorithm. We refer interested practitioners to R packages *Mclust* [63] and *mixtools* [7] or the MATLAB package *MIXMOD* [11].

3 Methods Based on the Hankel Matrices

The method of moments is generally considered to be less efficient when compared to maximum likelihood. Nonetheless, as justly argued in [46], there are situations where the method of moments reveals a nice mathematical structure. This is the case for the problem of estimating the true complexity of some finite mixture. As we will see below, the number of support points of a discrete mixing distribution with a finite number of jumps can be elegantly linked to whether the determinant of a special matrix of moments is equal to zero. Such a matrix is known under the name of a *Hankel matrix*.

We devote this section entirely to the estimation approaches based on the determinants of Hankel matrices of moments of the mixing distribution. The original method, with which we will start, was proposed in [21]. Additionally to the original approach

we will describe a couple of its possible extensions. [21] motivated the method with a number of appealing features:

1. it gives consistent estimators under some mild conditions;
2. it requires no a priori upper bound on the unknown order of the mixture;
3. it comes with low computational time as it does not involve estimation of the mixture parameters.

Another attracting property, not mentioned by the authors, is that the method bears a universal character and can be applied to continuous distributions as well as discrete distributions with no modifications, provided that the moment generating function of the distribution exists.

For the reader to be able to appreciate the elegant argument standing behind the method, we shall recall next the key theoretical results furnishing its basis.

3.1 The Main Theoretical Results and Basic Approach

Recall that we have confined the present study to a one-dimensional case, where $\Phi \subseteq \mathbb{R}$. For a given integer $m \geq 1$ define the set

$$\mathcal{C}_{2m} = \Big\{(c_1, \ldots, c_{2m})^T \in \mathbb{R}^{2m} : \exists \text{ some distribution function } G \text{ on } \Phi \text{ such that } c_j = \int_{\Phi} \phi^j \, dG(\phi) \text{ for } j \in \{1, \ldots, 2m\}\Big\}.$$

In other words, the component c_j is equal to the j-th moment of some distribution function G. For convenience, we will write $\boldsymbol{c}_{2m} = (c_1, \ldots, c_{2m})^T$ for any given real numbers c_j, $j = 1, \ldots, 2m$. In [21] this set is defined more generally with non-negative measure G.

For $\boldsymbol{c}_{2m} \in \mathbb{R}^{2m}$, the *Hankel matrix* associated with this vector is the $(m+1) \times (m+1)$ real symmetric matrix , denoted $H(\boldsymbol{c}_{2m})$ and given by

$$[H(\boldsymbol{c}_{2m})]_{i,j} = c_{i+j-2}, \quad 1 \leq i, j \leq m+1,$$

with $[H(\boldsymbol{c}_{2m})]_{1,1} = c_0 = 1$. More explicitly, we have that

$$H(\boldsymbol{c}) = \begin{pmatrix} 1 & c_1 & c_2 & \ldots & c_m \\ c_1 & c_2 & & & \\ c_2 & & & & \vdots \\ \vdots & & & \ddots & \\ c_m & & \ldots & & c_{2m} \end{pmatrix}.$$

Next we state the key result which links the true complexity of a finite mixture to the Hankel matrix of moments. See also PROPOSITION 1 in [21].

Theorem 3.1 *For a given $\boldsymbol{c}_{2m} \in \mathbb{R}^{2m}$, the Hankel matrix $H(\boldsymbol{c}_{2m})$ is positive semidefinite if and only if $\boldsymbol{c}_{2m} \in \mathcal{C}_{2m}$. Furthermore, $D_m := \det\big(H(\boldsymbol{c}_{2m})\big) = 0$ if and only if*

every distribution function G such that $c_j = \int_\Phi \phi^j dG(\phi)$, *G is discrete with at most m support points.*

Now we explain how the result above can be applied in the context of estimating the complexity of a finite mixture. Consider f_0, a finite mixture of densities which belong to some family $\mathcal{F}$ and let G_0 be the associated discrete distribution function with true complexity m_0. Then, Theorem 3.1 says that

$$m_0 = \inf\{m \in \mathbb{N} : D_m = 0\}, \tag{3.1}$$

where D_m, as above in Theorem 3.1, is the determinant of $H(\boldsymbol{c}_{2m})$ with

$$\boldsymbol{c}_{2m} = \left(\int_\Phi \phi dG_0(\phi), \ldots, \int_\Phi \phi^{2m} dG_0(\phi) \right).$$

In other words, the correct order of the mixture is the first integer which sets the determinant to zero. But the theorem implies also that $D_m = 0$ for all $m \geq m_0$. This characterizing feature of the true complexity is exploited to construct a sensible estimator. Indeed, assuming that it is possible based on the random sample $(X_1, \ldots, X_n)$ to obtain a strongly consistent estimator of any j-th moment of G_0, $\hat{c}_j$ say, then the Hankel estimator of m_0 proposed in [21] is given by

$$\hat{m}_n = \arg\min_{m \in \mathbb{N}} \left\{ |\widehat{D}_m| + a_m l_n \right\} \tag{3.2}$$

where

$$\widehat{D}_m = \det\left(H(\hat{\boldsymbol{c}}_{2m})\right), \quad \text{with } \hat{\boldsymbol{c}}_{2m} = \left(\hat{c}_1, \ldots, \hat{c}_{2m}\right)^T,$$

$\{a_m\}_{m \geq 1}$ is a positive and strictly increasing sequence, and $\{l_n\}_{n \geq 1}$ a positive sequence satisfying $\lim_{n \to \infty} l_n = 0$ (we have omitted writing the subscript n in the notation of the estimators of the moments and determinants).

Clearly, the term $a_m l_n$ is acting as a penalty. Adding a penalty term to $|\widehat{D}_m|$ is necessary because otherwise minimizing of $m \mapsto |\widehat{D}_m|$ alone might yield an inconsistent estimator. In fact, strong consistency of $\hat{c}_j$ implies that $|\widehat{D}_m|$ is a strongly consistent estimator of the true value $|D_m| = D_m$ (see our remark below). Since the latter is equal to 0 for all $m \geq m_0$, $|\widehat{D}_m|$ will be close to 0 for all $m \geq m_0$, which might result in choosing a value which is strictly larger than m_0. Under some additional assumptions, consistency of $\hat{m}_n$ as defined above in (3.2) can be established as shown in THEOREM 1 of [21]. We recall this result below.

Theorem 3.2 *If for all integers* $j, m \geq 1$ *we have that*

$$\hat{c}_j \to c_j \quad \textit{and} \quad \frac{\widehat{D}_m - D_m}{l_n} \to 0$$

almost surely as $n \to \infty$, *then* $\hat{m}_n \to m_0$ *a.s. as* $n \to \infty$.

Remark 3.1 Recall that $D_m = \det\big(H(\boldsymbol{c}_{2m})\big)$. Thus, D_m seen as a multivariate real function of $c_1, \ldots, c_{2m}$ (the components of $\boldsymbol{c}_{2m}$), is infinitely differentiable. Thus, if $\hat{c}_j$ is a strongly consistent estimator of c_j for any integer $j \geq 1$, then $\widehat{D}_m = \det\big(H(\hat{\boldsymbol{c}}_{2m})\big)$ is also a strongly consistent estimator of D_m. Furthermore, a multivariate weak convergence of $\hat{\boldsymbol{c}}_{2m}$ toward $\boldsymbol{c}_{2m}$ as in the case where a multivariate Central Limit Theorem applies, the estimator $\widehat{D}_m$ will converge weakly to D_m at a rate that is as fast as that of $\hat{\boldsymbol{c}}_{2m}$. Typically, the estimators $\hat{c}_j$ will result from considering some empirical estimators which we know to be asymptotically normal. Below, we will touch upon this point in some more detail.

Remark 3.2 In the light of Remark 3.1, the condition $(\widehat{D}_m - D_m)/l_n \to_{a.s.} 0, \ \forall m \in \mathbb{N}$ made in Theorem 3.2 is satisfied in case $(\hat{c}_j - c_j)/l_n \to_{a.s.} 0$ for all $j \in \mathbb{N}$. A typical situation is when $\sqrt{n}(\hat{c}_j - c_j) \to_d \mathcal{N}(0, \sigma_j^2)$ (for some $\sigma_j > 0$) and l_n is such that $\sqrt{n} l_n \to \infty$ in addition to $l_n \to_d 0$.

Without going into the full proof of Theorem 3.2, let us give some intuition for the condition $(\widehat{D}_m - D_m)/l_n \to_{a.s.} 0, \ \ \forall\, m \in \mathbb{N}$. We have that

$$|\widehat{D}_m| + a_m l_n = \begin{cases} l_n \left(\left| \frac{\widehat{D}_m - D_m}{l_n} + \frac{D_m}{l_n} \right| + a_m \right), & \text{for } m < m_0 \\ l_n \left(|\frac{\widehat{D}_m - 0}{l_n}| + a_m \right), & \text{for } m \geq m_0. \end{cases}$$

From the characterization if m_0 in (3.1), it follows that $D_m \neq 0$ for $m < m_0$ implying that

$$\left| \frac{\widehat{D}_m - D_m}{l_n} + \frac{D_m}{l_n} \right| \to \infty$$

almost surely as $n \to \infty$, whereas

$$\left| \frac{\widehat{D}_m - 0}{l_n} \right| + a_m \to a_m$$

for all $m \geq m_0$, with $a_m > a_{m_0}, \forall\, m > m_0$ since the sequence $\{a_m\}_{m \geq 1}$ is assumed to be strictly increasing. Thus, we expect that as $n \to \infty$ the minimum of the penalized criterion to be achieved at m_0.

The statement about consistency of $\hat{m}_n$ can be made more refined under additional regularity conditions. More precisely, suppose that for any integer $m \geq 1$ there exist integrable functions ψ_j and f_j for $j = 1, \ldots, 2m$ such that the j-th moment of G_0 is given by

$$c_j = f_j\big(\mathbb{E}[\boldsymbol{\psi}_{2m}(X)]\big),$$

where $\mathbb{E}[\boldsymbol{\psi}_{2m}(X)] = \left(\mathbb{E}[\psi_1(X)], \ldots, \mathbb{E}[\psi_{2m}(X)]\right)^T$. Define the estimator $\hat{m}_n$ the same way as above with

$$\hat{c}_j = f_j\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{\psi}_{2m}(X_i)\right),\ j = 1, \ldots, 2m.$$

We have the following theorem. See also Theorem 2 in [21].

Theorem 3.3 *Denote by $\boldsymbol{f}_{2m}$ the multivariate function defined as $\boldsymbol{f}_{2m}(\boldsymbol{t}_{2m}) = (f_1(\boldsymbol{t}_{2m}), \ldots, f_{2m}(\boldsymbol{t}_{2m}))$ for $\boldsymbol{t}_{2m} = (t_1, \ldots, t_{2m})^T \in \mathbb{R}^{2m}$. Suppose that for any $m \leq m_0$,*

- *$\boldsymbol{t}^{2m} \mapsto \det\left(H(\boldsymbol{f}_{2m}(\boldsymbol{t}_{2m}))\right)$ is Lipschitz with respect to some norm on $\mathbb{R}^{2m}$,*
- *for any $m \leq m_0$ the generating functions $u \mapsto \int \exp(u\psi_j(x))f_0(x)d\mu(x)$ exist in a neighborhood of 0 for all $j = 1, \ldots, 2m$.*

Furthermore, assume that $n^{1/2}l_n \to \infty$. Then, there exists a constant $d > 0$ and integer $n_0 > 0$ such that for all $n \geq n_0$

$$\mathbb{P}(\hat{m}_n \leq m_0) \leq 4m_0e^{-dnl_n^2}.$$

The main argument in the proof uses judicious upper bounds on the probabilities $\mathbb{P}(\hat{m}_n \leq m_0)$ and $\mathbb{P}(\hat{m}_n > m_0)$ based on concentration inequalities that involve the Cramer transform of the logarithm of the generating function of the centered random variables $\psi_j - \mathbb{E}[\psi_j(X)]$ for $j \in \{1, \ldots, 2m\}$ and $m \leq m_0$. Before commenting on the result itself, we would like to give some examples, which are relevant for the simulations section coming ahead.

Example 2: Mixture of Gaussian distributions. Consider a finite mixture of Gaussian distributions with density

$$f_0(x) = \pi_1\varphi(x - \theta_1) + \ldots + \pi_{m_0}\varphi(x - \theta_{m_0}),\quad x \in \mathbb{R}$$

with $\varphi(x) = 1/\sqrt{2\pi}\exp(-x^2/2)$, and $\theta_1, \ldots, \theta_{m_0} \in \mathbb{R}$. If $X \sim f_0$, then X has the same distribution as $Z + Y$ where $Z \sim \mathcal{N}(0, 1)$ and $Y \sim G_0$ with G_0 the mixing distribution with support points $\theta_1, \ldots, \theta_{m_0}$ and mixing probabilities $\pi_1, \ldots, \pi_{m_0}$ such that Y and Z are independent. Thus, for any $j \geq 1$

$$\mathbb{E}(X^j) = \sum_{k=0}^{j}\binom{j}{k}\mathbb{E}(Y^k)\mathbb{E}(Z^{j-k}) = \sum_{k=0}^{j}\binom{j}{k}c_k\mu_{j-k}$$

where $\mu_0 = 1$ and for an integer $r \geq 1$

$$\mu_r = \begin{cases} 0, & \text{if } r \text{ is odd} \\ (r-1)!!, & \text{if } r \text{ is even,} \end{cases}$$

where $x!!$ denotes the semifactorial of a number x.

Thus, the vector of moments $\boldsymbol{c}_{2m}$ satisfies the triangular linear system $\boldsymbol{c}_{2m} = BV$ where $B = A^{-1}$ and A is the lower triangular $(2m) \times (2m)$ matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \binom{2m}{2} & 0 & \binom{2m}{4} & 0 & \vdots & 1 \end{pmatrix}$$

and

$$V = \left(\mathbb{E}[X], \mathbb{E}[X^2] - 1, \dots, \mathbb{E}[X^{2m}] - (2m-1)!!\right)^T.$$

In this case, we have $c_j = \sum_{k=1}^{2m} B_{jk} \left(\mathbb{E}[X^k] - (k-1)!! \, \mathbb{I}_{k \in 2\mathbb{N}}\right)$. Thus, for location mixtures of Gaussian distributions we have shown that

$$\psi_j(x) = x^j, \quad \text{and} \quad f_j(\boldsymbol{t}_{2m}) = \sum_{k=1}^{2m} B_{jk} \left(t_k - (k-1)!! \, \mathbb{I}_{k \in 2\mathbb{N}}\right) \tag{3.3}$$

for $j \in \{1, \dots, 2m\}$.

More examples are available in Appendix A in the supplementary materials.

Now we turn to commenting on Theorem 3.3. Although the result of that theorem seems to give an actual guarantee on the consistency of $\hat{m}_n$, the exponential bound on the probability of being wrong about m_0 depends on n_0 and a constant d which are unknown. In case d is small and n_0 quite big, then consistency will not be observed for moderate and even big sample sizes: one would need an unrealistically huge number of observations to find the true complexity. Another problem is the estimation of the moments c_j, $j = 1, \dots, 2m$ for large values of m. Although the method does not require to put an upper bound on m while finding the minimum of $|\widehat{D}_m| + a_m l_n$ one has to choose some maximum admissible value for the mixture complexity. For large values of m the j-th moment c_j can become very large. When this is combined with a low quality estimator $\hat{c}_j$, $\hat{D}_m$ may be far away from 0, which is known to be the theoretical value for $m \geq m_0$. Such a phenomenon is illustrated using mixture of Gaussian distributions

$$f_0(x) = 0.3\varphi(x - 10) + 0.4\varphi(x - 13) + 0.3\varphi(x - 17). \tag{3.4}$$

In Table 1 we give the first 8 theoretical moments c_j of the mixing distribution and the mean value of their estimates $\hat{c}_j$ based on 100 replications for each of the sample sizes shown in the table. Table 2 gives the corresponding mean value of $\hat{D}_m$ as well as its penalized versions with $a_m = m$ and $l_n = \log n / \sqrt{n}$ or $l_n = \sqrt{\log n} / \sqrt{n}$ for $m \in \{1, 2, 3, 4\}$ computed on the basis of the same replications. It is clear from the values of Table 2 that $\hat{m}_n = 1$ even for this very well-separated mixture and for the large sample size $n = 10^4$.

Table 1 The true and estimated moments c_j and $\hat{c}_j$ for $j \in \{1, \ldots, 8\}$ of the mixing distribution of the 3-component mixture of Gaussian distributions given in (3.4)

Moment	1	2	3	4	5	6	7	8
True	13.3	184.3	2652.7	39480.7	604474.3	9471994.3	151201008.7	2449019520.7
n=100	13.283563	182.8470	2629.657	39006.33	595118.8	9353091	151033513	2431283379
n=1000	13.283277	184.4480	2652.951	39577.2	605040.4	9387062	150987763	2442459055
n=10000	13.305595	184.3422	2654.195	39501.67	604184.9	9493822	151502080	2446657287

Table 2 The mean value of $|\widehat{D}_m|$ and the penalized criterion $|\widehat{D}_m| + m l_n$, $m \in \{1, 2, 3, 4\}$ with the penalties $l_n = \log n/\sqrt{n}$ and $l_n = \sqrt{\log n}/\sqrt{n}$, for the 3-component mixture of Gaussian distributions given in (3.4)

$\widehat{D}_m$	1	2	3	4
$n = 100$	**7.309245**	239.1515	25703.188	81305942.21
$n = 1000$	**7.413671**	254.7251	8757.417	11693620.8
$n = 10000$	**7.414803**	254.6395	2690.280	1564924.88
$\widehat{D}_m + m \log n/\sqrt{n}$	1	2	3	4
$n = 100$	**7.769762**	240.0725	25704.57	81305944
$n = 1000$	**7.632113**	255.162	8758.072	11693622
$n = 10000$	**7.506906**	254.8237	2690.556	1564925
$\widehat{D}_m + m\sqrt{\log n}/\sqrt{n}$	1	2	3	4
$n = 100$	**7.523842**	239.5807	25703.832	81305943
$n = 1000$	**7.496784**	254.8913	8757.666	11693621
$n = 10000$	**7.445152**	254.7002	2690.371	1564925

The mean values were computed on the basis of 100 replications. In bold we indicate where the value at which the penalized criterion is minimal

Next, we examine what happens in a 2-component mixture of geometric distributions. To this aim, we consider the pmf

$$f_0(x) = 0.4(1 - 0.3)0.3^x + 0.6(1 - 0.8)0.8^x, \quad x \in \{0, 1, 2, \ldots\}. \tag{3.5}$$

The parametrization we chose implies that f_0 is a mixture of geometric distributions with success probability 0.7 and 0.2 respectively. In Table 3 one can see that the moments are accurately estimated for all sample sizes. However, the results of Table 4 indicate that the estimator $\hat{m}_n$ fails often to pick the correct mixture complexity, here $m_0 = 2$ for the penalties $a_m l_n = m \log n/\sqrt{n}$ and $a_m l_n = m\sqrt{\log n}/\sqrt{n}$.

Our decision to take the penalty $a_m l_n = m \log n/\sqrt{n}$ is based on the recommendation made in [21]. The second penalty $a_m l_n = m\sqrt{\log n}/\sqrt{n}$ was added in these simulations in order to compare the results obtained with the basic approach of [21] with the first modification we propose below and which is based on scaling the estimates of the determinants.

Table 3 The true and estimated moments c_j and $\hat{c}_j$ for $j \in \{1, \ldots, 6\}$ of the mixing distribution of the 2-component mixture of geometric distributions given in (3.5)

Moment	1	2	3	4	5	6
True	0.6	0.42	0.318	0.249	0.197	0.157578
n=100	0.59540	0.4186	0.3184	0.2448	0.1966	0.157
n=1000	0.59520	0.417	0.3174	0.255	0.1922	0.1568
n=10000	0.597200	0.4198	0.331	0.2434	0.1906	0.1552

Table 4 The mean value of $|\widehat{D}_m|$ and the penalized criterion $|\widehat{D}_m| + ml_n, m \in \{1, 2, 3\}$ with the penalties $l_n = \log n/\sqrt{n}$ and $l_n = \sqrt{\log n}/\sqrt{n}$ for the 2-component mixture of geometric distributions given in (3.5)

$\widehat{D}_m$	1	2	3
$n = 100$	0.063841	0.0024611	$\mathbf{8.2715310^{-5}}$
$n = 1000$	0.05867468	0.0006766619	$\mathbf{9.05786310^{-6}}$
$n = 10000$	0.05940753	0.0002569579	$\mathbf{1.02324210^{-6}}$
$\widehat{D}_m + m \log n/\sqrt{n}$	1	2	3
$n = 100$	**0.5243580**	0.9234951	1.3816338
$n = 1000$	**0.2771171**	0.4375615	0.6553363
$n = 10000$	**0.1515109**	0.1844638	0.2763112
$\widehat{D}_m + m\sqrt{\log n}/\sqrt{n}$	1	2	3
$n = 100$	**0.27843760**	0.43165431	0.64387252
$n = 1000$	**0.14178759**	0.16690248	0.24934778
$n = 10000$	0.08975607	**0.06095404**	0.09104665

The mean values were computed on the basis of 100 replications
Bold are those obtained for the true complexity of the mixture

The inconsistency noted in these examples, despite the nice theoretical guarantees of convergence of $\hat{m}_n$, are due to different reasons. In the Gaussian mixture, the penalty plays almost no role as $|\widehat{D}_m|$ dominates with values that are blowing up as we let m take larger values. For this reason, the estimator picks $m = 1$ in all cases. In the geometric mixture, the picture is completely reversed since the moments $c_j \in (0, 1)$ and hence get smaller for larger orders j. This causes $|\widehat{D}_m|$ to decrease with m. In this case, the penalty dominates and again $m = 1$ is often found as the minimizer of the penalized criterion.

The basic approach of [21] can suffer from serious underestimation (or overestimation) of the true mixture complexity even when the sample size is very large. Moreover, the question of how to choose the penalty term $a_m l_n$ is not really settled in [21]. In fact, a penalty which would work for a certain family of distributions could perform miserably for another. A traditional approach here would be to resort to cross-validation to decide on an optimal choice for the penalty. Although this is a very important aspect of the problem, we choose not to pursue it here.

3.2 Modification of the Basic Approach Using Scaling

The discussion above reveals that while the estimator defined by [21] is quite appealing, it is very difficult to achieve consistency in practice. The main problem resides in that the method does not exploit any knowledge about the variability of $|\widehat{D}_m|$. Without integrating the information about how these random variables behave stochastically (for n large enough), it would be almost impossible to say for example whether a small value obtained for $|\widehat{D}_m|$ can be seen as an indication that the true determinant is really equal 0. One way of circumventing the above issue is to use a rescaled version of this estimator. The starting point here is to use the already noted fact that the true determinant of the Hankel matrix of moments $\boldsymbol{c}_{2m} \mapsto D_m$ is an infinitely differentiable function on $\mathbb{R}^{2m}$. Thus, assuming that we can use the Central Limit Theorem to the vector of estimators $\hat{\boldsymbol{c}}_{2m}$, then for any fixed $m > m_0$ we get by applying the δ-method that

$$\sqrt{n}\left(\widehat{D}_1 - D_1, \ldots, \widehat{D}_{m_0-1} - D_{m_0-1}, \widehat{D}_{m_0} - 0, \ldots, \widehat{D}_m - 0\right)^T \to_d \boldsymbol{W}_m \tag{3.6}$$

where $\boldsymbol{W}_m = (W_1, \ldots, W_m)^T \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$, with Σ some nonnegative definite matrix of dimension $m \times m$.

Although the covariance matrix Σ is unknown it can be estimated using resampling techniques. Here, we focus only on estimating the diagonal elements of Σ, $\sigma_1^2, \ldots, \sigma_m^2$. By sampling B times with replacement from the original sample $(X_1, \ldots, X_n)$ we obtain a new sample $(X_1^*, \ldots, X_n^*)$ which can be used to compute the bootstrap determinants $\{\widehat{D}_1^*, \ldots, \widehat{D}_m^*\}$. Repeating this procedure B times allows us to estimate $\sigma_j/\sqrt{n}$ by computing the standard deviation $\hat{\sigma}_j^*$ of the bootstrap sample $(\widehat{D}_j^{*b}, b = 1, \ldots, B, j = 1, \ldots, m)$. As the setting here is very standard, it follows that as $n, B \to \infty$ $\hat{\sigma}_j^* \approx \sigma_j$, $j = 1, \ldots, m$ in probability.

As mentioned in the previous section, the true order of the mixture can be assumed to be smaller than some given value $m = m_{max}$; i.e., the search of the minimizer of the penalized criterion will be performed in the set $\{1, \ldots, m_{max}\}$. Assuming that $m_{max} > m_0$, define the rescaled vector

$$\left(\frac{\widehat{D}_1}{\hat{\sigma}_1^*}, \ldots, \frac{\widehat{D}_{m_0-1}}{\hat{\sigma}_{m_0-1}^*}, \frac{\widehat{D}_{m_0}}{\hat{\sigma}_{m_0}^*}, \ldots, \frac{\widehat{D}_{m_{max}}}{\hat{\sigma}_{m_{max}}^*}\right)^T := \left(Y_1^{(n)}, \ldots, Y_{m_0-1}^{(n)}, Y_{m_0}^{(n)}, \ldots, Y_{m_{max}}^{(n)}\right)^T.$$

Thus, we redefine the estimator $\hat{m}_n$ as

$$\hat{m}_n = \arg\min_{m \in \{1, \ldots, m_{max}\}} \left\{|Y_m^{(n)}| + a_m \sqrt{n} l_n\right\}. \tag{3.7}$$

We will not give a formal proof of consistency of $\hat{m}_n$. The latter, however, can be intuitively seen to hold since it follows from the weak convergence in (3.6) that

$$|Y_m^{(n)}| \to \infty, \quad \text{for } m = 1, \ldots, m_0 - 1,$$

Table 5 Proportion of the time the scaled Hankel estimator is equal to $m_0 = 3$ in the example of the finite mixture of Gaussian densities given in (3.4)

n	100	1000	10000
$l_n = \log n/\sqrt{n}$	0	0.98	1
$l_n = \sqrt{\log n}/\sqrt{n}$	0.63	0.99	1

The proportions are computed on the basis of $B = 500$ and 100 independent replications

Table 6 Proportion of the time the scaled Hankel estimator is equal to $m_0 = 2$ in the example of the finite mixture of geometric probability mass functions given in (3.5)

n	100	1000	10000
$l_n = \log n/\sqrt{n}$	0	0	1
$l_n = \sqrt{\log n}/\sqrt{n}$	0.16	0.75	1

The proportion is computed on the basis of $B = 500$ and 100 independent replications

and

$$(|Y_{m_0}^{(n)}|, \ldots, |Y_{m_{max}}^{(n)}|)^T \rightarrow_d (|Y_1|, \ldots, |Y_{m_{max}-m_0+1}|)^T$$

where $(Y_1, \ldots, Y_{m_{max}-m_0+1})^T$ is a multivariate Gaussian vector with a covariance matrix having all its diagonal terms equal to 1. One the one hand, this implies that for any integer $m \in \{1, \ldots, m_0 - 1\}$ the probability that m is the location of the minimum should decrease as $n \rightarrow \infty$. On the other hand, for $m \geq m_0$ the penalty $a_m \sqrt{n} l_n$ becomes the dominating term. Since a_m increases with m, the minimum is achieved at m_0 with increasing probability.

In the following, the examples considered above will be revisited using this modified approach to see to what extent the estimation accuracy is ameliorated. More specifically, we use the scaling approach described above to compute the proportion of times the alternative estimator $\hat{m}_n$ defined (3.7) is equal to the true complexity in 100 independent replications. In both examples, we have taken $m_{max} = 8$. The number of bootstrap replications was taken to be $B = 500$. From Table 5 and 6 , we see how the results drastically improve with the scaling method for the sample sizes $n = 1000, 10000$ with 100% or close for the recovery of the true complexity. The improvement seems to be more pronounced with the choice of penalty $m\sqrt{\log n}/\sqrt{n}$. Thus, one conclusion that can be drawn here is that the method would greatly benefit from comparing the performance of different penalties. As mentioned above, such a comparison can be done using some cross-validation approach.

3.3 Modification of the Basic Approach Using Bootstrap

A specific feature of the Hankel matrix of moments methods discussed previously is the possibility to estimate the order of the mixture without estimating the parameters. However, it seems that there might be a high price to pay for avoiding this part: some of the essential features of the mixture may not be captured by the determinant alone,

Table 7 Proportion of the time the modified Hankel estimator is equal to $m_0 = 3$ in the example of the finite mixture of Gaussian densities given in (3.4)

n	100	1000	10000
Proportion of times $\hat{m}_n = 3$	0	0	0

The proportions are computed on the basis of $B = 500$ and 100 independent replications

which can lead to the wrong answer with a "bad" penalty, even for very large sample sizes. Furthermore, in many applications it may be desirable to obtain the estimator of the order of the mixture as well as the estimators of all the parameters. We describe here another modification that is suited for this purpose. It is in essence a sequential testing procedure in which some statistic computed from the data is compared with a critical value obtained e.g. by re-sampling from the assumed theoretical model.

The said statistic can be taken to be either the determinant of the Hankel matrix $\widehat{D}_m$ as in the basic approach proposed by [21] or its rescaled version as described in the previous section. The idea is to replace minimizing the objective function in (3.2) or (3.7) by taking a reject/accept decision regarding whether the current value of m is equal to the true complexity. We describe this procedure only when the statistic is taken to be equal to $\widehat{D}_m$ since the modifications are rather obvious for the scaled version thereof:

- for $m \in \{1, \ldots, m_{max}\}$, compute $\widehat{D}_m$ and the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}_m$ of $\boldsymbol{\theta}_{m_0}$ based on the given sample $X_1, \ldots, X_n$;
- from $f_{\hat{\boldsymbol{\theta}}_m}$, the corresponding estimate of the mixture density, generate a large number, B, of samples of size n to obtain a sequence of statistics $\{\widehat{D}_m^{*b}\}_{1 \le b \le B}$. Let $\hat{q}_{m,B,\alpha/2}$ and $\hat{q}_{m,B,1-\alpha/2}$ be the the empirical $(\alpha/2)$- and $(1-\alpha/2)-$quantiles based on this bootstrap sample;
- if $m = m_{max}$ or $\hat{q}_{m,B,\alpha/2} \le \widehat{D}_m \le \hat{q}_{m,B,1-\alpha/2}$, then take $\hat{m}_n = m$, otherwise repeat the previous steps with $m + 1$.

The procedure described above is not new in the context of estimating a mixture complexity. In fact, a similar approach will be encountered below with the only difference that it is based either on some minimum distance estimation or likelihood ratio statistic (see Sects. 4 and 5). In a nutshell, one sequentially tests

$$H_0^m : m_0 = m \text{ versus } H_1^m : m_0 > m \tag{3.8}$$

and declares as an estimate for m_0 the first value of m for which H_0^m is not rejected.

In Table 7 and 8 we report the proportion of the time the sequential procedure described above gives the correct mixture complexity for the same finite Gaussian and geometric mixtures given in (3.4) and (3.5) respectively.

From the simulations results obtained in Table 7 and 8 we can see that this other modification of the original Hankel matrix method is less successful for the Gaussian mixture but still works well for the geometric one. This might be explained again by the large values of the higher-degree moments of the Gaussian distribution which

Table 8 Proportion of the time the modified Hankel estimator is equal to $m_0 = 2$ in the example of the finite mixture of geometric probability mass functions given in (3.5)

n	100	1000	10000
Proportion of times $\hat{m}_n = 2$	0.25	0.96	0.94

The proportion is computed on the basis of $B = 500$ and 100 independent replications

impact heavily the quality of estimating the determinants. This is not at all an issue with the geometric distribution whose moments are much easier to estimate.

3.4 Extension of the Hankel Matrix Approach using Neural Networks

The conclusions achieved on the basis of the simulation study, summarized in Sect. 6, stipulate that there is a need of search for a more reliable and universal mixture order estimation technique that would yield more precise estimates irrespectively of the underlying scenario. Obviously, the approaches we have already examined yield estimators which depend on the features involved in a non-linear fashion. To this extent we decided to turn our attention the popular statistical tool designed specifically for modelling nonlinearities between the sets of input and output variables - Neural Networks (NNs), in the hope that they might identify patterns and relationships that the other approaches cannot capture.

For the past decade the amount of research carried out in the field of NNs has experienced exponential growth, and a great multitude of NN types and classes have been designed to successfully solve a wide range of problems. The simplest of these tasks like image labelling or pattern recognition are usually solved by feed-forward network architectures such as the multi-layer perceptron (MLP), convolutional neural network (CNN) or radial basis function network (RBFN). More sophisticated tasks such as speech recognition or text translation require more complex interactions between the layers of the network, which are implemented in such architectures as long-short-term memory (LSTM).

At this stage of our research we are not aiming at estimating the whole mixture model (finding the optimal complexity as well as all its parameters) but only pursue the goal of identifying the number of subgroups in the population based on the observed data. Thus, when cast into the NN framework, the problem of estimating the number of components in a mixture can be viewed as a supervised multiclass classification problem and the relevant questions that need to be addressed are

- discovering the most informative features to be fed into the NN and
- devising an adequate design for the training set;
- proposing the optimal NN architecture;
- choosing the learning algorithm;
- determine whether a universal architecture for multiple families of distributions can be found;
- understanding whether using NNs is beneficial compared to the other methods.

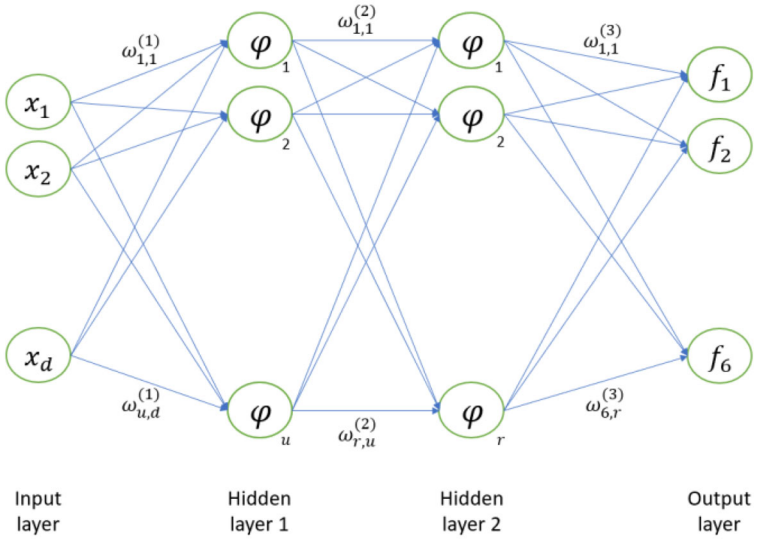

Fig. 1 Sample MLP architecture

It seems natural to start the search for a well-suited network by browsing first thorough the simplest class of the NNs: the multilayer perceptron (MLP). Despite of their relative simplicity, networks with just two layers can approximate any continuous functional mapping [12]. One of the simplest possible architectures of the considered model with only two hidden layers is depicted in Fig. 1.

The input features $x_1, \ldots, x_d$, $d \geq 1$, the first and the second hidden layers consist of u and r neurons respectively, the weights vector $\boldsymbol{\omega}$ is learned by optimizing the loss function $J(\boldsymbol{\omega} = (\omega_{1,1}^{(1)}, \ldots, \omega_{6,r}^{(3)})^T)$, which is taken to be the cross entropy function, which is most often used in multiclass classification tasks such as ours:

$$J(\boldsymbol{\omega}) = \frac{1}{m} \sum_{k=1}^{m} \left[z_k \log\left(\hat{p}_k(\boldsymbol{\omega})\right) + (1 - z_k) \log\left(1 - \hat{p}_k(\boldsymbol{\omega})\right) \right],$$

where

$$z_k = \begin{cases} 1, & \text{if } m_0 = k, \text{ with } k = 1, 2, 3, 4, 5, 6, \\ 0, & \text{otherwise} \end{cases}$$

are known labels for each of the generated vector of features in the training sample.

To this end, the choice over the optimal configurations is restricted to deciding on the number of hidden layers in the network, the number of neurons in each layer and the corresponding activation functions, the loss function and the learning algorithm.

For estimating the order of a mixture using a NN, one needs an appropriate assumption on the possible maximal number of components. We take the maximum number of components to be 6 for our task.

Recall from Sect. 3 that Hankel's criterion is backed by elegant, orderly statistical theory, however the method's performance turns out to be rather poor in practice. We use Hankel matrix determinants as inputs to the MLP to try to improve the estimation results by exploiting the information concentrated in the determinants without resorting to the use of any penalty function. Our experience reveals that using sequences of the first 6 Hankel matrix determinants of the mixing distribution as inputs leads to improved results when compared to other tested options. For this reason we regard this approach as an extension to the original Hankel technique.

Using the sequence of the Hankel determinants as inputs produced resulted in high performance for the Geometric mixtures, but showed poorer performance for the Poisson mixture due to the fact that the determinants for the Poisson mixtures tend to blow up while those for the Geometric mixtures stay bounded within a [0, 1] interval. For that reason the inputs for the Poisson mixture had to be modified. One of the modifications led to good performance was the relative changes in the absolute values of the Hankel determinants:

$$x_k = \begin{cases} |\hat{D}_1|, k = 1 \\ \frac{|\hat{D}_k| - |\hat{D}_{k-1}|}{|\hat{D}_{k-1}|}, k \in 2, \ldots, 6. \end{cases} \tag{3.9}$$

To ensure the variety of the training examples, the characteristics and structure of the data that is used for prediction should be scrutinized and taken into account. The training set should be designed to be as representative as possible of the data of interest. The following procedure can serve as a useful example. For the considered mixtures distributions the parameters for each mixture component is chosen without replacement from a grid with a pre-specified step to insure that these are distinct. A grid on [0.05, 1] with step 0.05 for the geometric distribution should do well in most of the applications. A grid on [1, 20] with step 1 for the Poisson can be taken if the expected rate of occurrences is believed not to exceed 20 by much and the parameters of subpopulations are separated by at least 1 unit. The step value can be reduced if more precision is desired. The mixing proportions can be taken on a gird with an appropriate step size in a similar way, in this case replacement is allowed and the generated results should be normalized. The mixtures with the parameters obtained in this way are then used for generating samples. It seems to be useful to enrich the training data set by simulating several times from each of the resulting mixtures to account for possible variation in the sample populations during the training.

The 6 output neurons of the output layer is further processed by softmax activation function $\sigma : \mathbb{R}^6 \longrightarrow [0, 1]^6$, which ensures that the estimated class probabilities live between 0 and 1 and sum up to 1:

$$\hat{p}_k = \sigma(\hat{q})_k = \frac{e^{\hat{q}_k}}{\sum_{k'=1}^{6} e^{\hat{q}_k'}},$$

where $\hat{q}_k$ us the resulting value of the k-th output neuron.

Whenever optimizing the loss function, the value of the learning rate becomes of importance: when too small, the weights of the NN are hardly updated, and much time

is needed for the network to find a solution; when too large, the weights are updated in large increments, and overshooting the optimum becomes highly probable. In the case of mixture order estimation the value of the learning rate is influenced by the values of the parameters of the mixtures in the training set: an efficient learning rate for the geometric mixtures, where the parameters lie within (0, 1] and thus higher precision is required to separate the components, will be smaller than for Poisson or Gaussian mixtures where the parameters can range from 0 to 20.

The output of the MLP is a vector of estimates of the class probabilities, that is, the probabilities of an observation (represented by either a sample from a mixture distribution or a vector of alternative relevant features) belonging to each of possible 6 classes: $\hat{p}_k, k \in \{1, 2, 3, 4, 5, 6\}$. The predicted number of components in the mixture is taken to be the class with the highest estimated probability:

$$\hat{m} = \arg\max_k \hat{p}_k.$$

The search for a successful model (done using KerasTuner library [54]) requires examination of a large parameter space even for a simple network such as a MLP. Combinations of several hyperparameters of the NN were kept track of in order to identify the optimal NN architecture:

- activation function: relu, tanh, sigmoid
- number of layers: $1, 2, \ldots, 9, 10$
- neurons in each hidden layer: $10, 25, 40, \ldots, 280, 295, 310$
- dropout layer after the last hidden layer: dropout rate between 0 and 0.1
- learning rate for the optimization algorithm: $10^{-2}, 10^{-3}, 10^{-4}$

A set of 10000 samples was used for training the NN, the motivation being that taking the mo- ments and determinants as the features requires quality estimation thereof. Therefore, while a sample of this size is rarely available in real-world datasets, the emphasis was placed on finding a neural network that would perform well if the estimates are good. In practice, a neural net- work trained with 10000 samples still predicts well when the test sample size is much smaller.

Unfortunately, we were not able to find a single MLP specification that would work equally well for all considered families of distributions - Gaussian, geometric and Poisson. Table 9 presents three different MLP configurations for Gaussian, geometric and Poisson mixtures that achieved satisfactory performance in our simulations. The predicted class probabilities as well as prediction accuracy on a number of test cases for networks with the denoted specifications can be found in Sect. 6.

4 Methods Based on Minimum Distance Estimators

4.1 General Setting

The estimation techniques discussed in this section are mainly based on the works [69, 74, 75]. Additional relevant references will be mentioned below. In a nutshell, these techniques use the minimal distance between a consistent estimator of f_0, $\hat{f}_n$

Table 9 NN configurations for Gaussian, geometric and Poisson mixtures

NN characteristics	Gaussian	Geometric	Poisson
Inputs	Relative changes in Hankel determinants	Moments	Relative changes in Hankel determinants
Activation function	relu	relu	sigmoid
Hidden Layers	4	3	2
Neurons in Layer 1	205	200	200
Neurons in Layer 2	115	150	200
Neurons in Layer 3	70	100	–
Neurons in Layer 4	70	100	–
Dropout rate	0.03	0.03	–
Learning rate	0.0001	0.0001	0.001

say, and an estimator of the latter in the class of finite mixtures with m components $m \geq 1$. Using the notation of Sect. 2, this means that the estimator of the true m_0 will be based on the projection of $\hat{f}_n$ on the class $\mathcal{F}_m, m \geq 1$ in a sense that will be determined. As noted above, these classes are nested; i.e., $\mathcal{F}_m \subset \mathcal{F}_{m+1}$. Thus, the basic approach stops at the first m where the projection on $\mathcal{F}_{m+1}$ does not bring a substantial improvement over the projection on $\mathcal{F}_m$. One main feature of the methods investigated in this section is that one performs estimation of the parameters of the mixture as well as the mixture complexity.

In the above mentioned papers, the projection of $\mathcal{F}_m$ makes use of the Hellinger or the L_2 distances. The estimation method suggested in [74] is built upon a model selection procedure by sequentially fitting the nested mixture models. Thus, the method is reminiscent of the sequential testing approach already encountered above for the second modified version of the determinant of the Hankel matrix of moments. The procedure allows at each iteration to search over a higher class by adding one component to the mixture and find the best model within each class until adding another component brings no more benefit in the sense that it decreases the objective loss function by an amount smaller than a specified tolerance level. In the sequel, we consider only the situation where the dominating measure, μ, is either the counting or Lebesgue measure. In the first case, the nonparametric estimator of f_0 is the empirical probability mass function given by

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i = x), \ x \in \mathcal{X}. \tag{4.1}$$

In the second one where X is absolutely continuous, we consider a kernel density estimator with fixed or random bandwidth c_n:

$$\hat{f}_n(x) = \frac{1}{nc_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{c_n}\right), \tag{4.2}$$

where K is some standard kernel function.

4.2 The Minimum Distance Estimator: The Basic Approach

In the following, let $\mathcal{F}$ denote the class of densities that are mixed. Also, let $\mathcal{D}$ be either the Hellinger or L_2 distance, that is for two densities f and g with respect to μ we have either

$$\begin{aligned}\mathcal{D}^2(f,g) &= \frac{1}{2}\int_{\mathcal{X}}\left(\sqrt{f(x)}-\sqrt{g(x)}\right)^2 d\mu(x)\\ &= 1-\int_{\mathcal{X}}\sqrt{f(x)}\sqrt{g(x)}d\mu(x) = \mathcal{H}^2(f,g)\end{aligned}$$

or

$$\mathcal{D}^2(f,g) = \int_{\mathcal{X}}(f(x)-g(x))^2\,d\mu(x) = L_2^2(f,g).$$

Recall the following notation from Sect. 2: For a given

$$\boldsymbol{t}_{p_m} = (\pi_1,\ldots,\pi_m,\boldsymbol{\phi}_{d,1},\ldots,\boldsymbol{\phi}_{d,m})^T \in \Theta_{p_m}$$

where $p_m = m(d+1)-1$, we denote by $f_{\boldsymbol{t}_{p_m}}$ the m-component mixture density given by $f_{\boldsymbol{t}_{p_m}}(x) = \pi_1 f_{\boldsymbol{\phi}_{d,1}}(x)+\ldots+\pi_m f_{\boldsymbol{\phi}_{d,m}}(x)$. For a given density f, we now consider the functional

$$\boldsymbol{\theta}_{p_m}^{\mathcal{D}}(f) = \left\{\boldsymbol{\theta}_{p_m}\in\boldsymbol{\Theta}_{p_m} : \mathcal{D}(f_{\boldsymbol{\theta}_{p_m}},f) = \min_{\boldsymbol{t}_{p_m}\in\boldsymbol{\Theta}_{p_m}}\mathcal{D}(f_{\boldsymbol{t}_{p_m}},f)\right\}.$$

provided that the minimum exists. Here, $\boldsymbol{\theta}_{p_m}^{\mathcal{D}}(f)$ denotes the set of all minimizers as uniqueness of the solution is not guaranteed. Note that our notation is different from the one used in [8], [74], [75] and [69]. In case $f=\hat{f}_n$, we have the following definition: for a given m and the non-parametric estimator $\hat{f}_n$ defined above in (4.1) or (4.2), the *minimum distance estimator with respect to* $\mathcal{D}$ of $\boldsymbol{\theta}_{p_m}$ is defined as

$$\hat{\boldsymbol{\theta}}_{p_m}^{\mathcal{D}} = \boldsymbol{\theta}_{p_m}^{\mathcal{D}}(\hat{f}_n) \tag{4.3}$$

provided that a minimizer exists.

Proving existence of a minimizer $\boldsymbol{\theta}_{p_m}^{\mathcal{D}}(f)$ for some given density f requires careful argumentation under some regularity conditions. When $\mathcal{D}$ is the Hellinger distance, Theorem 1 in [8] gives a proof of this existence under the condition that $\boldsymbol{t}_{p_m}\mapsto f_{\boldsymbol{t}_{p_m}}(x)$ is continuous for almost every $x\in\mathcal{X}$, that the mixture is identifiable and the parameter space $\Theta_{p_m} = \mathcal{S}_{m-1}\times\Phi^m$ is compact. Also, the same theorem proves uniqueness of $\boldsymbol{\theta}_{p_m}^{\mathcal{D}}(f)$ in case f is itself a finite mixture. In order words, if $f=f_{\boldsymbol{\theta}_{p_m}}$, then $\boldsymbol{\theta}_{p_m}^{\mathcal{D}}(f) = \boldsymbol{\theta}_{p_m}$. This can be easily seen as an immediate consequence of identifiability. When $\mathcal{D}$ is the L_2 distance and μ is the counting measure on the set of non-negative integers,

then [69] show a similar theorem while relaxing the condition of compactness on the parameter space. The main building block in the proof is to show that the mapping

$$\boldsymbol{t}_{p_m} \mapsto \sum_{x=0}^{\infty} \big(f(x) - f_{\boldsymbol{t}_{p_m}}(x)\big)^2 = \|f - f_{\boldsymbol{t}_{p_m}}\|_2^2$$

is continuous. In the discrete setting considered in [69], a proof of the continuity property can be based on a slightly different argument. Indeed, if $\boldsymbol{t}_{p_m}^{(k)}$ is a sequence converging to $\boldsymbol{t}_{p_m}$ as $k \to \infty$, then by Minkowski's inequality (also used in page 4252 of [69])

$$\begin{aligned}\Big|\|f - f_{\boldsymbol{t}_{p_m}}\|_2 - \|f - f_{\boldsymbol{t}_{pm}^{(k)}}\|_2\Big| &\le \|f_{\boldsymbol{t}_{p_m}} - f_{\boldsymbol{t}_{pm}^{(k)}}\|_2 \\ &\le \sum_{x=0}^{\infty} |f_{\boldsymbol{t}_{p_m}}(x) - f_{\boldsymbol{t}_{pm}^{(k)}}(x)|, \\ &\qquad \text{using that a pmf is always bounded by 1.}\end{aligned}$$

The latter sum converges to 0 by continuity of $\boldsymbol{t}_{p_m} \mapsto f_{\boldsymbol{t}_{p_m}}(x)$ and application of the Sheffé's Theorem. For existence of a minimizer when $\mathcal{D}$ is the L_2-distance, [69] makes the assumption that for the pmf f to be projected, for any m there exist some compact C (which depends on m but we omit writing this dependence explicitly), and $\boldsymbol{\theta}_{p_m}^*$ such that

$$\inf_{g \in \Theta_{p_m} \setminus C} \mathcal{D}(f, g) > \mathcal{D}(f, f_{\boldsymbol{\theta}_{p_m}^*}).$$

Such an assumption is not needed in case Θ_{p_m} is itself compact. Also, the compact C is rather abstract and one only needs to exhibit its existence in some way. It is clear that even when $\boldsymbol{\theta}_{p_m}^{\mathcal{D}}(f)$ is not a singleton, we have $f_{\boldsymbol{\theta}_{p_m}} = f_{\boldsymbol{\theta}'_{p_m}}$ a.e. for two minimizers $\boldsymbol{\theta}_{p_m}, \boldsymbol{\theta}'_{p_m} \in \boldsymbol{\theta}_{p_m}^{\mathcal{D}}(f)$. When $f = \hat{f}_n$, we will denote the corresponding density $f_{\boldsymbol{\theta}_{p_m}}$ (or $f_{\boldsymbol{\theta}'_{p_m}}$) by $\widehat{f}_m^{\mathcal{D}}$. Note that we have omitted the subscript n, and replaced p_m by m for the sake of a lighter notation. By definition of $\boldsymbol{\theta}_{p_m}^{\mathcal{D}}(f)$ we have

$$\widehat{f}_m^{\mathcal{D}} = \arg\min_{g \in \mathcal{F}_m} \mathcal{D}(\hat{f}_n, g). \tag{4.4}$$

For $f = f_0 \in \mathcal{F}_{m_0}$ we write

$$f_m^{\mathcal{D}} = \arg\min_{g \in \mathcal{F}_m} \mathcal{D}(f_0, g). \tag{4.5}$$

Note that $f_m^{\mathcal{D}} = f_0$ for all $m \ge m_0$. The roles of $\widehat{f}_m^{\mathcal{D}}$ and $f_m^{\mathcal{D}}$ will become clear below. Although our notation for those projections is different from the one used in the aforementioned papers, our choice is driven by our desire to maintain some notational coherence throughout this survey.

Now, we describe how the basic approach works with the minimum distance estimators. The estimation procedure as outlined in [74] is much inspired by the work of

[37]. The latter paper is however mainly focused on estimating the true complexity of a finite mixture of Gaussian distributions using the Kullback-Leibler divergence instead of the Hellinger (or L_2) distance. There are two starting points of the basic approach. The first one is to note that the true mixture complexity m_0 satisfies

$$\begin{aligned} m_0 &= \min\{m : \mathcal{D}^2(f_0, f_m^{\mathcal{D}}) = 0\} \qquad (4.6)\\ &= \min\{m : \mathcal{D}^2(f_0, f_m^{\mathcal{D}}) = \mathcal{D}^2(f_0, f_{m+1}^{\mathcal{D}})\} \\ &= \min\{m : \mathcal{D}^2(f_0, f_m^{\mathcal{D}}) \leq \mathcal{D}^2(f_0, f_{m+1}^{\mathcal{D}})\}. \qquad (4.7) \end{aligned}$$

While the identity in (4.6) is a direct consequence of identifiability, the one in (4.7) is less obvious. Note that this identity is proved if we show that for $m \leq m_0 - 1$, $\mathcal{D}(f_0, f_m^{\mathcal{D}}) > \mathcal{D}(f_0, f_{m+1}^{\mathcal{D}})$. A proof of this fact when $\mathcal{D}$ is the Hellinger distance can be found in p. 1485 of [74], where an auxiliary lemma (Lemma A.4) was used. For the case where $\mathcal{D}$ is the L_2 distance, and for the sake of completeness, we give a proof in Appendix C in the supplementary materials.

Based on the discussion above, it seems natural to search for the first m which minimizes some empirical version of the distance $\mathcal{D}^2(f_0, f_m^{\mathcal{D}})$, namely $\mathcal{D}(\hat{f}_n, \hat{f}_m^{\mathcal{D}})$. The second one is to recall once more the inclusion $\mathcal{F}_m \subset \mathcal{F}_{m+1}$. This implies that

$$\mathcal{D}(\hat{f}_n, \hat{f}_{m+1}^{\mathcal{D}}) \leq \mathcal{D}(\hat{f}_n, \hat{f}_m^{\mathcal{D}}).$$

The inequality above means that without penalization it is in principle impossible to find a finite order which can be taken as an estimator of m_0. This overfitting is accounted for by adding a penalty term which is proportional to the number of parameters in the mixture model. This yields the following criterion

$$\mathcal{D}^2(\hat{f}_n, \hat{f}_m^{\mathcal{D}}) + b_n v_m, \qquad (4.8)$$

for some chosen sequences $\{v_m\}_m$ and $\{b_n\}_n$ such that the former is increasing and the latter satisfies $\lim_{n\to\infty} b_n = 0$. Note that in the works [74], [75] and [69], $b_n v_m/n$ is taken instead. Now, mimicking the property in (4.7) gives rise to the following definition of the minimum distance estimator

$$\begin{aligned} \hat{m}_n &= \min\left\{m : \mathcal{D}^2(\hat{f}_n, \hat{f}_m^{\mathcal{D}}) + b_n v_m \leq \mathcal{D}^2(\hat{f}_n, \hat{f}_{m+1}^{\mathcal{D}}) + b_n v_{m+1}\right\} \\ &= \min\left\{m : \mathcal{D}^2(\hat{f}_n, \hat{f}_m^{\mathcal{D}}) \leq \mathcal{D}^2(\hat{f}_n, \hat{f}_{m+1}^{\mathcal{D}}) + \alpha_{n,m}\right\}, \\ &\qquad \text{with } \alpha_{n,m} = b_n(v_{m+1} - v_m). \end{aligned}$$

If this minimum does not exist, then $\hat{m}_n = \infty$. The term $\alpha_{n,m}$ can be seen as a threshold so that an integer m is declared to be the estimator when the projection of $\hat{f}_n$ on the class $\mathcal{F}_{m+1}$ yields an insignificant change in comparison with its projection on the previous class $\mathcal{F}_m$.

Strong consistency of the minimum distance estimator defined above result was stated in Theorem 1 in [75] for the Hellinger distance and in the Consistency Theorem Section in [69]. In the former, the nonparametric estimator $\hat{f}_n$ is taken

to be a kernel estimator, as defined above in (4.2), with bandwidth c_n such that $\lim_{n\to\infty}\left(c_n + (nc_n)^{-1}\right) = 0$. In the latter, $\hat{f}_n$ is the empirical probability mass function as given in (4.1). Now, the only condition assumed on $\alpha_{n,m}$ in these convergence theorems is that $\alpha_{n,m} \to 0$ as $n \to \infty$. This is of course guaranteed by the fact that $\lim_{n\to\infty} b_n = 0$. However, we believe that this statement is not accurate as is, since the penalty needs to depend on the rate of convergence of $\hat{f}_n$ to the true density f_0. This rate of convergence is known to depend on the smoothness of f_0 and the kernel K. In Appendix C in the supplementary materials, we explain why the condition $\lim_{n\to\infty} \alpha_{n,m} = 0$ is not enough.

4.3 Modification of the Basic Approach Via Bootstrap

In this section we propose a modification of the basic approach based on minimal distance between the projection of a non-parametric estimator on the class of m-component mixtures and this estimator augmented with some given threshold. We resort in this modification to a parametric bootstrap procedure in order to avoid the bad choice of a threshold. For a given integer $m \geq 1$, consider the hypothesis testing as in (3.8)

$$H_0^m : m_0 = m \quad \text{vs.} \quad H_1^m : m_0 > m,$$

and recall the estimator $\widehat{f}_m^{\mathcal{D}}$ as defined in (4.3). Let us now define

$$\Delta_m = \mathcal{D}(\widehat{f}_m^{\mathcal{D}}, \hat{f}_n) - \mathcal{D}(\widehat{f}_{m+1}^{\mathcal{D}}, \hat{f}_n).$$

To obtain the distribution of Δ_m under the null hypothesis H_0^m, we draw B independent samples of size n from the fitted density $\widehat{f}_m^{\mathcal{D}}$. For $b = 1, \ldots, B$, we compute the non-parametric estimators $\hat{f}_n^{(b)}$, $\widehat{f}_m^{\mathcal{D},(b)}$ and $\widehat{f}_{m+1}^{\mathcal{D},(b)}$ and the corresponding difference

$$\Delta_m^{(b)} = \mathcal{D}(\widehat{f}_m^{\mathcal{D},(b)}, \hat{f}_n^{(b)}) - \mathcal{D}(\widehat{f}_{m+1}^{\mathcal{D},(b)}, \hat{f}_n^{(b)}).$$

If $\hat{q}_{B,\alpha/2}$ and $\hat{q}_{B,1-\alpha/2}$ denote the empirical $\alpha/2$ and $(1-\alpha/2)$-quantiles based on the bootstrap sample $(\Delta_m^{(1)}, \ldots, \Delta_m^{(B)})$, then H_0^m is rejected if

$$\Delta_m \notin [\hat{q}_{B,\alpha/2}, \hat{q}_{B,1-\alpha/2}]$$

and the current candidate m is replaced by $m+1$. We take as our estimator for m_0 the first m for which the null hypothesis is not rejected. Here, we report simulations results for the two examples for finite mixtures considered above; see (3.4) and (3.5). In these simulations, $B = 500$ and the sample sizes considered are $n = 100, 1000, 10000$. The performance of the method proposed in this section is assessed through the proportion of times the estimator is equal to the true complexity (3 and 2 respectively). We used Hellinger distance for the Gaussian mixture and L_2 distance for the geometric mixture.

Table 10 Proportion of the time the modified minimum-distance estimator is equal to $m_0 = 3$ in the example of the finite mixture of Gaussian densities given in (3.4)

n	100	1000	10000
Proportion of times $\hat{m}_n = 3$	0.27	0.94	1.00

The proportions are computed on the basis of $B = 500$ and 100 independent replications

Table 11 Proportion of the time the modified minimum-distance estimator is equal to $m_0 = 2$ in the example of the finite mixture of geometric densities given in (3.5)

n	100	1000	10000
Proportion of times $\hat{m}_n = 2$	0.47	0.96	0.99

The proportions are computed on the basis of $B = 500$ and 100 independent replications

From Tables 10 and 11 one can see that coupling the bootstrap with distance minimization gives very promising results. The downsize of this modification remains essentially the computational burden which comes with the resampling step.

5 Sequential Likelihood Ratio Tests with Bootstrap

Likelihood-based methods play a central role in statistical inference. Among these methods the likelihood ratio test (LRT) is one of the most widely used in practice. The LRT has a simple interpretation and enjoys the property of being invariant under re-parametrization; see for example [41]. Furthermore, whenever certain regularity conditions are satisfied Wilks' theorem [71] says that, under the null hypothesis, it converges weakly to a chi-square distribution as the sample size grows to ∞. The degrees of freedom of the limiting chi-square distribution are determined by the difference of the dimension of the whole parameter space and that of the null space. However, and as already mentioned above, the regularity conditions required for the asymptotic theory to hold are not satisfied in the mixture problem. The issue is that it is always possible to write a mixture model with m components as a model with $m + 1$ components (or more) by setting some of the mixing probabilities to 0. Hence, the true parameter under the null hypothesis lies on the boundary of the alternative space. To better explain the problem, let us consider the example of testing whether a random sample was generated from a (single) Gaussian distribution $\mathcal{N}(\theta, 1)$ versus a 2-component Gaussian mixture, where each of the components has variance equal to 1. If f_0 denotes the true density, then the goal is to test

$$H_0 : f_0(x) = \varphi(x - \theta) \text{ vs. } H_1 : f_0(x) = \pi_1\varphi(x - \theta_1) + (1 - \pi_1)\varphi(x - \theta_2)$$

for some $\theta \in \mathbb{R}$ and $\theta_1 \neq \theta_2 \in \mathbb{R}$ and $\pi_1 \in (0, 1)$. If we define here the likelihood ratio as the ratio of the maximum likelihoods over the null and the whole space, then the maximization needs to be done on $\mathbb{R}$ (the null space) and

$$\Big\{(\theta_1, \theta_2, \pi_1) : \theta_1, \theta_2 \in \mathbb{R}, \quad \pi_1 \in [0, 1]\Big\} = \mathbb{R}^2 \times [0, 1]$$

(equal to the whole parameter space; i.e., the union of the null and alternative spaces). Thus, a density under H_0 lies on the boundary in the sense that f_0 results from either setting the mixing probability π_1 to 1 or 0 and taking $\theta_1 = \theta$ (or $\theta_2 = \theta$), or letting $\theta_1 = \theta_2$ while π_1 is arbitrary. In the classical setting, one of the main arguments which leads to the chi-square distribution as the weak limit is the use of Taylor expansion up to the second order of the log-likelihood at the global MLE around the true parameter. We can refer here to the proof of Theorem 22 in [26] for well-explained and rigorous arguments. Things then work because the true parameter is an interior point and has a unique representation as an element in the whole parameter space. In the mixture model setting, this argument does not work because, as it can be seen from the example above, the true parameter has different representations under the null hypothesis where it is on the boundary. This non-standard situation has triggered a strong interest for either computational or theoretical investigation of the limit distribution of the LR statistic under the null hypothesis. In this context, we can refer to [1], [64], [13], and [47]. For a nice review of the papers on this subject, we can refer to Section 6.5 in [49] among others. The main message that one can take from these works is that when the limit distribution of the LRT can be simply described, it is a mixture of chi-square distributions. In more complicated cases, this limit distribution is given more abstractly by $\sup_{s\in\mathcal{S}} \max(0, Y_s)^2$ where Y is some well-defined centered Gaussian process and $\mathcal{S}$ is some suitable set. See for example [47] where it is shown that $\mathcal{S}$ is the ensemble of cluster points of some generalized score.

In this section we review the sequential testing procedure based on LRT and use of resampling for approximating the distribution of the test statistic under the null hypothesis. This method was proposed in [38] for estimating the true complexity of a finite mixture of Poisson. Although the focus there was put on that family, the approach can certainly be extended to other distributions. Consider again the hypothesis testing problem

$$H_0^m : m_0 = m \quad \text{vs.} \quad H_1^m : m_0 > m.$$

Using the same notation as above, we define the maximum likelihood under H_0^m and H_1^m as

$$L_{\boldsymbol{X}}(\hat{\boldsymbol{\theta}}_{p_m}) = \sup_{\boldsymbol{\theta}_{p_m} \in \Theta_{p_m}} L_{\boldsymbol{X}}(\boldsymbol{\theta}_{p_m}), \quad \text{and} \quad L_{\boldsymbol{X}}(\hat{\boldsymbol{\theta}}_{p_{m+1}}) = \sup_{\boldsymbol{\theta}_{p_{m+1}} \in \Theta_{p_{m+1}}} L_{\boldsymbol{X}}(\boldsymbol{\theta}_{p_{m+1}})$$

where $L_{\boldsymbol{X}}$ denotes the likelihood function based on the sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from the unknown mixture, and $p_m = m(d + 1) - 1$. As in [38], consider the log-likelihood ratio statistic

$$\lambda = -2\left(\log L_{\boldsymbol{X}}(\hat{\boldsymbol{\theta}}_{p_m}) - \log L_{\boldsymbol{X}}(\hat{\boldsymbol{\theta}}_{p_{m+1}})\right) \tag{5.1}$$

A mixture with smaller number of components will be rejected in favor of the larger model whenever the log-likelihood ratio statistic is large. Otherwise, the mixture will

Table 12 Proportion of the time the LRT with bootstrap estimator is equal to $m_0 = 3$ in the example of the finite mixture of Gaussian densities given in (3.4)

n	100	1000	10000
Proportion of times $\hat{m}_n = 3$	0.26	0.97	1.00

The proportions are computed on the basis of $B = 500$ and 100 independent replications

Table 13 Proportion of the time the LRT with bootstrap estimator is equal to $m_0 = 2$ in the example of the finite mixture of Gaussian densities given in (3.5)

n	100	1000	10000
Proportion of times $\hat{m}_n = 2$	0.92	0.95	0.98

The proportions are computed on the basis of $B = 500$ and 100 independent replications

be declared to have m component in the absence of a strong evidence against it. Exactly as in Sects. 3 and 4 , the decision against H_0^m is taken sequentially starting with $m = 1$ until H_0^m cannot be rejected, in which case m will be declared as the estimator of m_0. In view of the issues related with deriving the asymptotic distribution of the log-likelihood ratio statistic, one can resort to a parametric bootstrap. As this is already done above, the description of the procedure is omitted.

We give the results of this procedure for the two finite mixtures with $m_0 = 3$ and 2 already considered above.

6 Simulation Results

In this section we describe the simulation results obtained for sample sizes $n \in \{50, 100, 500, 1000, 5000, 10000\}$ using the procedures described in the previous sections for finite mixtures of Gaussian, geometric and Poisson distributions with $m_0 \in \{2, 3, 4\}$. Throughout this study, unless explicitly stated otherwise, the standard deviations for Gaussian mixture components are assumed to be 1. For the minimum distance-based methods, we follow the recommendations made in the relevant papers [75] and [69] and use the following two penalty functions $\alpha_{n,m}$, with m denoting the stipulated mixture order and n the sample size:

- the penalty based on the Akaike Information Criterion (AIC) $\alpha_{n,m} = \frac{0.6}{n} \log\left(\frac{m+1}{m}\right)$ for the L_2 distance and $\alpha_{n,m} = \frac{2}{n}$ for the Hellinger distance
- the penalty based on the Schwarz Bayesian Criterion (SBC) $\alpha_{n,m} = 0.6\frac{\log n}{n} \log\left(\frac{m+1}{m}\right)$ for the L_2 distance and $\alpha_{n,m} = \frac{\log n}{n}$ for the Hellinger distance.

For all simulations, we used 500 replications to compute the frequencies of an exact recovery of the true complexity. These are displayed in Tables 28, 29, 30, 31, 32, 33 and 34 to be found in Appendix B in the supplementary materials.

Clearly, the performance of the NN's cannot be directly compared to the performance of the other methods we have considered. One of the reasons being that the

NN framework is very much different from the other approaches in terms of training and testing procedures applied. The neural network requires to encounter as many different mixtures as possible to be able to learn the latent structure, while the other techniques (except for the Hankel approach) are seeking the best fit for a given sample from the target mixture. We still report the NN results along with the performances of the other methods with a few reservations so that the reader could put together the full picture. Several points need to be borne in mind when reading and interpreting the NN performance:

- the values reported are the predicted class probabilities that reflect how sure the network is that an observation (one of the tested mixtures) belongs to each of the represented classes;
- bin ≥ 5 can be slightly misleading in this case as it reflects the summed probability for classes with 5 components and 6 components and in a few instances the probability corresponding to one of the classes is larger then two of the probabilities for 5 and 6 when considered separately but less, when they are so combined; still the results as given provide an insight into how well the NN can tell the classes apart;
- sample size of 10000 was used for computing the input features for the NN thus the results are reported in the respective table cell;
- 10000 samples were used to train the NN;
- accuracy measure that shows in how many instances the NN was able to correctly predict the true number of components cases will be reported separately.

The accuracy measure was computed by generating 100 different samples from one of the selected mixtures (these did not occur in the training set), and the percentage of times the network gave out the correct prediction was computed.

It follows from the simulations that the bootstrap modification of the Hankel matrix determinants method shows reasonable results for the geometric and Poisson mixtures but fails to produce accurate estimates for all 2-component and 3-component Gaussian mixtures, persistently underestimating the number of components when compared to the truth. This underestimation is likely to be the result of the "exploding"moments and determinants issue inherent in the Hankel matrices of the mixing distribution for the finite normal mixtures. The issue was covered in Sect. 3, and the simulation study results illustrate it once again.

For a well separated 2-component Gaussian mixture all methods (except for the Hankel matrix determinants approach with bootstrap) perform well even for sample size as small as $n = 50$. The BIC and ICL methods demonstrate high performance in this setting (estimating the number of components correctly in more than 95% of the cases for small sample sizes of $n = 50$, 100 and 100% for larger ones) and so do the minimum Hellinger distance methods. For the estimation of mixture complexity via the minimum Hellinger distance procedures (the L_2 distance is not applicable in the case of continuous distributions) we used a KDE with bandwidth of 0.5. The Hankel matrix determinants modification with bootstrap and scaling show slightly inferior results for small and average sample sizes when compared to BIC, ICL and the minimal Hellinger distance methods but performs well for sample sizes of $n \geq 5000$. The LRT approach for the available sample sizes of $n \in \{50, 100, 500, 1000\}$ shows more modest results

in this setting although the relative frequency of the correctly estimated cases is still high. For a well-separated Gaussian mixture with a low mixing proportion for one of the components the LRT method works well irrespective of the sample size estimating the number of components correctly in more than 90% of the instances for $n \geq 100$ and outperforming many other methods (except for BIC and ICL) for small sample sizes. The performance of all techniques drops significantly for not well-separated mixtures (e. g. a 2-component Gaussian mixture with overlapping regions). The NN extension of the Hankel method demonstrates good results in all 3 cases for the 2-component Gaussian mixtures even when the modes are located close to each other, which is manifested in the estimated class probabilities as well as in high accuracy rates of 100%, 96% and 100% respectively.

The BIC and ICL methods that proved to be supreme when compared to other methods in many scenarios for the Gaussian mixtures are not performing as well for the mixtures of geometric distributions. For the 2-component mixtures the simulation results indicate that the minimum Hellinger distance with bootstrap and LRT techniques tend to outperform all other tested methods. The penalized Hellinger and L_2-distance methods perform well whenever the mixture is well-separated (the weights do not have to be well-balanced however) and $n \geq 500$. For small and medium sample sizes the minimum distance-based methods with penalties show rather poor results. The techniques using the Hankel matrix determinants also enjoy high performance for well-separated mixtures when the mixing proportions are similar or not of observations is large, identifying the number of components correctly in 95% of the cases. The NNs also demonstrate less confidence when applied to the geometric mixtures as indicated by a higher level of spread in the estimated class probabilities although the accuracy rate of 59% can be considered a rather decent success measure for the challenging first scenario. The accuracy amounts to 100% for a less demanding case with the well-separated mixture.

Generally, for $n \geq 5000$ and more observations all methods with an exception of the minimal L_2 distance approach using the SBC-based penalty allow for the correct estimation in at least 90% of the cases.

For well-separated 2-component Poisson mixtures all methods perform well even for very small sample sizes. The BIC, ICL and the two penalized MHD methods often demonstrate accuracy which is close to absolute. The minimum distance-based approaches with the AIC-based term seem to work better than the minimum distance-based approaches with the SBC-based penalty whenever the mixtures are either not well separated or when one of the mixtures is scarcely represented. Whenever the mixture is less challenging, the approaches involving the SBC-based penalty tends to lead to higher accuracy than those involving the AIC penalty. Also, choosing the Hellinger distance for the minimal distance techniques seems to achieve slightly better performance than using the L_2 distance (whenever the components are not too close). In the case of well-separated 2-component Poisson mixtures the scaled bootstrap version of the Hankel matrix approach seems to outperform the bootstrap modification. For a well-separated mixture with a low mixing proportion of one of the components the LRT and the scaled version of the Hankel matrix determinants techniques seem to be an appropriate choice. Approaches using the Minimum L_2 distance perform rather poorly for all sample sizes in this setting while all Hellinger distance-based approaches

provide good estimates when the number of observations is large. The NNs seem to be able to learn well the 2-component Poisson mixtures. The accuracy rates for 3 out of 4 scenarios are 100%. The only exception being the mixture with very closely located components, where the NN fails.

For a well-separated 3-component Gaussian mixture with similar mixing proportions the BIC and ICL methods show outstanding performance even for small sample sizes. All methods (except for the Hankel matrix determinants with bootstrap) perform relatively poorly for small sample sizes and very well for sample size of $n \geq 5000$ where their accuracy achieves 95%. The Hankel matrix method with bootstrap and scaling shows again slightly worse results for small and average sample sizes but performs well for sample sizes of $n \geq 5000$. The minimum Hellinger distance approach with AIC-based term shows better performance than other estimators from the same group for average sample sizes. The LRT method delivers poor results for small sample sizes but achieves mor tha 95% accuracy already for $n = 500$. For a well-separated mixture with a low mixing proportion for one of the components the picture is similar, with a slightly lower performance. The BIC method outperforms the other methods, although giving correct predictions in more than 80% of the cases only for $n \geq 5000$. The LRT method seems to outperform all other methods (except for the BIC) for small sample sizes, providing the correct estimate in almost 70% of the instances for $n = 500$. In general the results for almost all methods are rather poor for small sample sizes, improving as the number of observations grows. For average sample sizes the minimum Hellinger distance with AIC-based term shows a significant improvement, estimating the mixture complexity correctly in more than 70% of the cases, and all Hellinger distance-based methods perform well for sample sizes $n \geq 5000$. The NN demonstrates again good results in all 3 cases for the 3-component Gaussian mixtures achieving accuracy rates of 90%, 96% and 100% respectively for the given cases. For a not well-separated mixture all methods perform quite poorly, being able to identify only 2 components out of 3 in most of the instances. The LRT method for the available sample sizes shows estimation results which are only marginally better than those achieved by the other methods. The BIC approach also performs poorly in this case, often either underestimating or overestimating the number of components in the mixture, but still shows better results when compared to other methods.

Geometric mixtures with 3 components seem to be a challenging task for all methods considered in the survey. The minimum Hellinger distance approach with bootstrap and the LRT technique are still able to show better results for small and average sample sizes when compared to other methods whenever the mixtures are well-separated. It is likely that these methods could show better performance for large sample sizes. For the first 3-components mixture, which is not well-separated, none of the methods estimate the complexity correctly, systematically underestimating it. In this scenario the methods penalized with AIC-based penalty once again show better performance than their counterparts using the SBC-based thresholds. The group of distance-based approaches as well as the cluster of Hankel matrix techniques also perform poorly in this setting even when the sample size is large and the mixture is well-separated. The 3-component mixtures seem to be challenging also for the NNs, causing the accuracy rate for the two scenarios to drop to 65% and 44% respectively.

In the case of well-separated 3-component Poisson mixtures the BIC method achieves absolute performance for $n \geq 5000$, while the LRT technique outperforms all other tested methods for small sample sizes. One could suppose that it would be very effective for larger sample sizes. The ICL persistently underestimates the number of components in the 3-component mixture even when the components are far apart. Whenever the Poisson mixture is well-separated the modification of the Hankel matrix approach with bootstrap tends to be more accurate than the Hankel matrix approach with bootstrap and scaling. All methods relying on the minimum Hellinger distance estimate the complexity correctly in more than 90% of the instances for $n \geq 1000$ and the LRT techniques perform well for average sample sizes. The bootstrap modification of the Hankel matrix determinants approach and the minimal L_2 distance-based methods also perform well whenever the mixture is well-separated and the mixing proportions are approximately the same for all components, with no component dominating over the others provided the number of observations is large enough. For average sample sizes and "unbalanced" mixtures the results of these methods are rather poor when compared to the previously listed approaches. The NNs achieves accuracy rates of 83% and 76% for the two well-separated mixture, but is not able to correctly predict the order of the mixture with very closely located components.

Poisson mixtures with 4 components seem to be quite a challenge for all methods, and a large number of observations is needed to obtain decent performance.

7 Applications to Real Data

In this Section we demonstrate how the approaches discussed in the previous sections perform on real data. Wherever applicable the size of the bootstrap sample size is taken to be 1000, for the LRT-based approach, the observed LRT statistic is always compared with the 95%-quantile, for other methods using bootstrap (the methods relying on the Hankel matrix determinants and the minimum distance-based procedures), 2.5%- and 97.5%-quantiles are used.

We will begin the analysis by considering the Old Faithful Geyser Data, first published in [4]. The data comprises waiting times between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. The waiting times between the eruptions of the Old Faithful geyser are assumed to be well described by a finite Gaussian mixture model.

Unfortunately, the techniques based on the Hankel matrix determinants do not appear to be an appropriate choice in this setting. To make use of the translation nonparametric Hankel method approach (as outlined in the Section 3.1 of [21]) one should make sure that the assumption of equal standard deviations throughout the mixture components holds. This assumption does not seem to hold for the Old Faithful data. The NN extension cannot be used for these data either as for the NN training we assumed a known variance of 1 for all components in the mixture, which is not the case here.

The BIC and ICL method applied to the Old Faithful data result in different estimates of the optimal number of components. The maximal BIC values corresponds to 3

Table 14 BIC and ICL values for the Old Faithful Data

Method	1	2	3	4	5
BIC	5426	4787	**4229**	4242	4174
ICL	5426	3819	3724	**3455**	3500

Bold are those obtained for the true complexity of the mixture

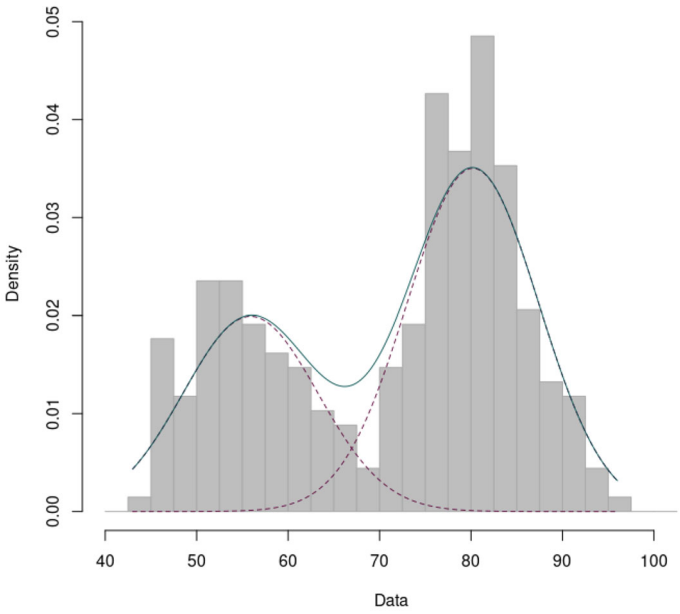

Fig. 2 Estimated mixture model for the Old Faithful dataset (MHD_{bt})

groups while the ICL chooses 4. The BIC and ICL values for this data can be found in Table 14.

Implementing the minimum Hellinger distance method with an automatically selected bandwidth for the KDE and AIC-based penalty one obtains that the optimal mixture should have 2 components. The resulting estimated 2-component mixture (in green) as well as the 2 components (in dark red) are plotted in Fig. 2 along with the empirical distribution of the waiting times.

Changing the penalty term to the SBC-based from the AIC-based yields a slightly different parameter vector estimate, however the choice of the number of components remains the same. The differences of the squared Hellinger distance values and the AIC/SBC-based thresholds can be found in Table 15. The minimum Hellinger distance method with bootstrap and the LRT approach also yield a 2-component mixture.

The minimum distance methods and the LRT attain slightly different parameter estimates (the estimated parameters for BIC and ICL approaches are similar to those of the LRT as in all these cases the MLE is used). The results have been gathered into a single table (Table 16) for the ease of comparison.

Table 15 Hellinger distance measures and thresholds for the Old Faithful dataset

Method	m	$\Delta_m = \mathcal{D}^2(\hat{f}_n, \hat{f}_m^{\mathcal{D}}) - \mathcal{D}^2(\hat{f}_n, \hat{f}_{m+1}^{\mathcal{D}})$	Threshold
MHD_{AIC}	1	0.065318	0.011029
	2	−0.002220	0.011029
MHD_{SBC}	1	0.065417	0.030914
	2	−0.022900	0.030914

Table 16 Estimated parameter vectors for the distance-based and likelihood-based approaches

Method	Estimated parameters
MHD_{AIC}	$\big(0.65, 0.35, (79.72, 7.01), (54.88, 6.21)\big)$
MHD_{SBC}	$\big(0.68, 0.32, (80.03, 6.73), (54.62, 7.21)\big)$
MHD_{bt}	$\big(0.65, 0.35, (79.21, 6.91), (54.32, 6.45)\big)$
LRT/BIC/ICL	$\big(0.64, 0.36, (80.09, 5.87), (54.61, 5.87)\big)$

Table 17 BIC and ICL values for the Children Data

Method	1	2	3	4	5
BIC	7289	**6727**	6743	6760	6777
ICL	**7289**	8256	9160	11022	10413

Bold are those obtained for the true complexity of the mixture

We now consider the data taken from the 1952 Annual Report of a pension fund that contains the information on the number of children of 4075 widows entitled to the fund support. The dataset first appeared in [67]. The data do not appear to be simply a random sample from a Poisson distribution as the number of zeros (widows with no children) appears to be too large. This issue was treated in [67] by fitting a mixture of two processes, one of which is a Dirac distribution at 0 while the other follows a Poisson distribution. We attempted to fit a mixture of Poisson distributions to the data using several of the discussed approaches to verify the above mentioned population heterogeneity assumption.

The BIC and the ICL methods estimate 2 and 1 components respectively for this data set, the values are given in Table 17.

The estimated parameters are (0.66, 0.34; 0.0311.115) for the BIC approach and (1, 0.4) for the ICL.

Non-parametric non-scaled and scaled Hankel matrix approaches with respective penalty terms $\frac{m\log(n)}{\sqrt{n}}$ and $m\log(n)$ yield the estimated number of components in the Poisson mixture equal to 2, as can be seen from Table 18.

The parametric non-scaled Hankel matrix determinants approach with bootstrap as well as the corresponding scaled version yield a 2-component mixture of Poisson distributions with the estimated by the maximum likelihood parameters (0.66, 0.34, 0.03, 1.11). This agrees with the population's heterogeneity hypothesis

Table 18 Absolute values of non-scaled Hankel matrix determinants for the Children Data

m	1	2	3	4	5
$\lvert\hat{D}_m\rvert$	0.3932	**0.2659**	0.3907	0.5209	0.6544
$\lvert\hat{D}_m\rvert/\sigma_m^*$	21.2339	**16.6435**	26.7833	35.3900	41.6757

Bold are those obtained for the true complexity of the mixture

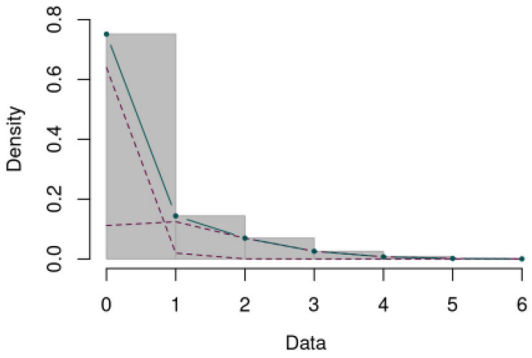

Fig. 3 Estimated Mixture Model for the Children Dataset (HM_{bt})

Table 19 Δ_m for Hellinger and L_2 distances and thresholds for the Children dataset

Method	m	$\Delta_m = \mathcal{D}^2(\hat{f}_n, \hat{f}_m^{\mathcal{D}}) - \mathcal{D}^2(\hat{f}_n, \hat{f}_{m+1}^{\mathcal{D}})$	Threshold
MHD_{AIC}	1	0.0030988	0.000491
	2	−2.417e−10	0.000491
MHD_{SBC}	1	0.030988	0.002040
	2	−2.417e−10	0.002040
L_2E_{AIC}	1	0.005977	0.000102
	2	9.848e−08	5.970e−05
L_2E_{SBC}	1	0.005977	0.000848
	2	9.848e−08	0.000496

found in [67]. The estimated Poisson mixture along with the empirical distribution are shown in Fig. 3.

The estimated parameters for the minimum Hellinger and L_2 distance methods with AIC-based and SBC-based penalties and bootstrap as well as the LRT approach also yield the same number of components and very similar parameter estimates. Table 19 displays the differences of the squared Hellinger and L_2 distances along with the corresponding thresholds.

The NN predicts 2 components with very high probability, the estimated class probabilities output by the NN being as reported in Table 20.

Thus all methods (except for ICL) applied to these data agree on the optimal number of components in the mixture, also confirming the hypothesis, identifying the optimal number of components as 2.

Table 20 Predicted class probabilities for the Children Data

Number of components	1	2	3	4	5	6
Class probabilities	0.00	**0.95**	0.05	0.00	0.00	0.00

Table 21 BIC and ICL values for the Shakespeare Data

Method	1	2	3	4	5
BIC	179890	**167095**	167116	167136	167157
ICL	179890	**174784**	192055	189458	202613

Bold are those obtained for the true complexity of the mixture

Table 22 Absolute values of non-scaled Hankel matrix determinants for the Shakespeare Data

m	1	2	3	4	5
$\lvert\hat{D}_m\rvert$	0.1668	**0.1182**	0.1767	0.2356	0.2945
$\lvert\hat{D}_m\rvert/\sigma_m^*$	62.8608	**33.0276**	41.0747	50.11900	55.1286

Bold are those obtained for the true complexity of the mixture

The dataset used for fitting a mixture of geometric distributions is the Shakespeare dataset, analyzed in the seminal work [24], which comprises counts of the number of times certain words that William Shakespeare used in his writings. The data is designed as follows: the number of times Shakespeare used a word only once is 14376, the number of times the same word occurred exactly 10 times in his writings is 363 an so on. The goal set in [24] was to use the observed frequencies words, to estimate the unobserved number of words that Shakespeare knew but did not use in his writings. This problem is known under the name of "species richness"and can be solved using a variety of approaches. One of the approaches was considered in [6], where the theoretical rationale for using a mixture of geometric distributions in such a setting is laid out.

The BIC and ICL methods both select 2 as the optimal complexity for the Shakespeare dataset as can be deduced from Table 21 containing the BIC and ICL values for number of components $1, \ldots, 5$.

Both non-parametric non-scaled and scaled Hankel matrix techniques estimated the number of components as 2, as follows from Table 22.

The parametric Hankel matrix determinants approach with bootstrap applied to the Shakespeare data yields the estimated number of components in the mixture of geometric distributions equal to 2. The same number of components is obtained when the scaled version of the parametric Hankel matrix determinants approach with bootstrap is used.

The estimated mixture for the minimum Hellinger and L_2 distance methods with AIC-based and SBC-based penalties and bootstrap suggest 3 components unlike the Hankel matrix determinants procedures, yielding slightly different estimates $(0.4148, 0.3758, 0.2094, 0.8406, 0.2902, 0.0490)$ (for Hellinger) and $(0.3839, 0.3870, 0.2291, 0.8622, 0.3203, 0.0523)$ (for L_2). The squared differences

Table 23 Δ_m for Hellinger and L_2 distances and thresholds for the Shakespeare dataset

Method	m	$\Delta_m = \mathcal{D}^2(\hat{f}_n, \hat{f}_m^{\mathcal{D}}) - \mathcal{D}^2(\hat{f}_n, \hat{f}_{m+1}^{\mathcal{D}})$	Threshold
MHD_{AIC}	1	0.128804	6.495e−05
	2	0.008953	6.495e−05
	3	−3.276e−10	6.495e−05
MHD_{SBC}	1	0.128804	0.000336
	2	0.008953	0.000336
	3	7.497e−11	0.000336
L_2E_{AIC}	1	0.016010	1.351e−05
	2	0.000386	7.901e−06
	3	6.306e−08	7.901−06
L_2E_{SBC}	1	0.016010	0.000140
	2	0.000386	8.165e−05
	3	1.214e−06	5.793e−05

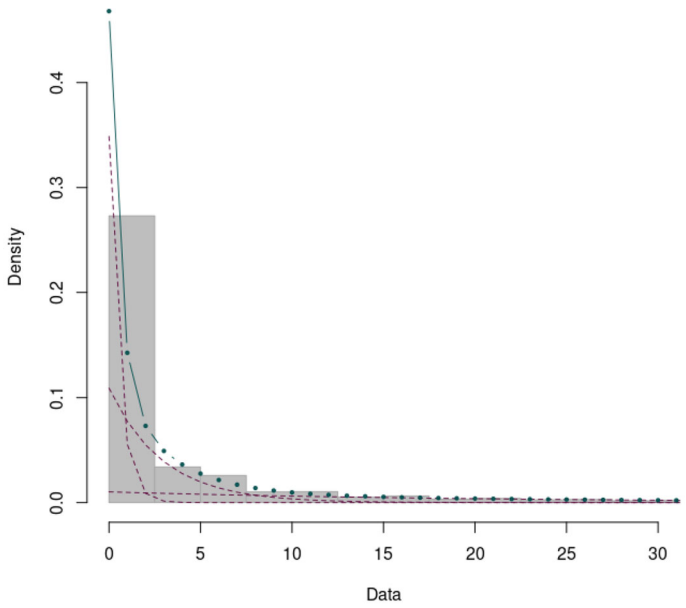

Fig. 4 Estimated Mixture Model for the Shakespeare Dataset (L_2E_{AIC})

for Hellinger and L_2 distances and the BIC/AIC-based thresholds can be found in Table 23.

The estimated mixture of geometric distributions using the Hellinger distance approach with bootstrap, all of its components and the empirical distribution are plotted in Fig. 4.

The LRT approach results in a 3-component mixture with the estimated parameter vector (0.4270, 0.3744, 0.1986, 0.8311, 0.2741, 0.0446).

Table 24 Predicted class probabilities for the Shakespeare Data

Number of components	1	2	3	4	5	6
Class probabilities	0.00	0.00	0.02	0.11	0.32	**0.55**

Bold are those obtained for the true complexity of the mixture

Table 25 Results of all reviewed methods used on the real-world datasets

Method	Old Faithful	Children	Shakespeare
BIC	3	2	2
ICL	4	1	2
HM	–	2	2
HM_{sc}	–	2	2
HM_{bt}	–	2	2
HM_{btsc}	–	2	2
L_2E_{AIC}	–	2	3
L_2E_{SBC}	–	2	3
MHD_{AIC}	2	2	3
MHD_{SBC}	2	2	3
MHD_{bt}	2	2	3
LRT	2	2	3
NN	–	2	6

The estimated parameters for the minimum Hellinger and L_2 distance methods with AIC-based and SBC-based penalties and bootstrap as well as the LRT approach also yield the same number of components and very similar parameter estimates.

The NN predicts the maximal possible number of components for the Shakespeare data, which is 6, with the class probabilities designated in Table 24:

Thus for the Shakespeare data various methods do not agree on the optimal number of components, which ranges from 2 components to 6, also providing different estimates for the model parameters. The final model choice in this case is left to the researcher who might have an insight on which number of component makes the most sense (e.g. according to the parts of speech different words belong to). Table 25 summarizes the estimates for all reviewed methods and all datasets.

8 Other Work on Mixture Complexity Estimation

The current survey certainly does not cover all the existing methods for estimating the complexity of a finite mixture.

Before mentioning some other interesting references, we would like to draw the reader's attention to the seminal work of Bruce Lindsay which, among other things, brought a very novel way of viewing the nonparametric maximum likelihood estimator (NPMLE) of a general mixture distribution; see [44] and [45]. The novelty resides in considering this estimator from a geometric perspective. One of the most important

results which derive from this is that, under some simple conditions, the NPMLE of a general mixture is a finite mixture with complexity not exceeding the number of distinct observations. Not surprisingly, all the papers on the methods reviewed and implemented in this survey refer to one of Lindsay's work on mixture models. This continues to hold true for the other approaches we would like now to mention and which we believe to be worthwhile to be brought to the reader's attention. There has been a lot of papers where penalization of some goodness-of-fit criterion was proposed with rigorous proofs of consistency or some other guarantee of the resulting estimator of the true number of components. In [42] the penalized NPMLE was considered with a penalization function $\alpha_{n,m}$ satisfying $\alpha_{n,m+1} > \alpha_{n,m}$ and $\limsup_{n\to\infty} \alpha_{n,m}/n = 0$. Under some regularity conditions on the mixture model, it is very rigorously shown that the estimator is at least equal to the true number of components as the sample $n \to \infty$ with probability 1. The method requires only computation of the NPMLE for a number of values of m, which can be done using for example a support reduction algorithm as described in [70] or [33]. As it is not known whether this penalized NPMLE is actually consistent, [18] constructed a penalized minimum-distance estimator which could be thought as a precursor of the minimum distance estimators reviewed in Sect. 4. Two main differences are to be noted though: Firstly, [18] consider distances between distribution functions instead densities. Secondly, the penalization function takes the form of $-c_n \sum_{j=1}^{m} \log \pi_j$ where π_j, $j = 1, \ldots, m$ are the mixing probabilities and $(c_n)_n$ a sequence converging to 0 as $n \to \infty$. Consistency of the penalized minimum distance estimator of the true complexity is shown when one chooses c_n such that the distance between the empirical distribution function and the true mixture distribution is $O(c_n)$ almost surely. In the special case where one wants to decide between $m_0 = 1$ (homogeneity) and $m_0 = 2$, [17] propose a method based on modifying the likelihood ratio test. The modification operates first on the log-likelihood function by adding a negative penalty of the form $C \log(4\pi(1-\pi))$ with $\pi > 0$ the mixing probability and $C > 0$ some chosen constant. The penalty clearly discourages the MLE, under the null hypothesis, from fitting a mixing probability that is close to 0. The modification is motivated by the desire of overcoming the issues associated with the nesting $\mathcal{F}_1 \subset \mathcal{F}_2$ (boundary problem and non-uniqueness of representing the null hypothesis) already described in some details in Sect. 5. If $\pi f_{\phi_1} + (1-\pi) f_{\phi_2}$ is the mixture density, and if we denote by l_n the modified log-likelihood, the the modified LRT is given by $2\left(l_n(\hat{\pi}, \hat{\phi}_1, \hat{\pi}_2) - l_n(1/2, \hat{\phi}, \hat{\phi})\right)$, where $(\hat{\pi}, \hat{\phi}_1, \hat{\phi}_2)$ and $\hat{\phi}$ are the MLE under the alternative and null hypotheses. One very interesting theoretical result is that the LRT is shown to converge weakly to mixture $(1/2) : (1/2)$ of a Dirac at 0 and $\chi^2_{(1)}$. This limit distribution can be then used to construct asymptotic critical region for rejecting homogeneity. As for the constant C, it is recommended to take $\log M$ if it is believed that $\phi_1, \phi_2 \in [-M, M]$ although the results do not seem to be too much sensitive to taking other values for C.

Bayesian approaches were also used to make inference about the number of components of a mixture. In this framework, this number is viewed as a random variable drawn from some prior distribution and the corresponding posterior distribution is then derived and subsequently used for inference purposes. For mixtures of Gaussian distributions whose means and variances are regarded as random variables drawn

from a Dirichlet process, we refer to work of [25]. The posterior distribution can be approximated using Monte Carlo. In [60] the authors make use of reversible jump Markov chain Monte Carlo methods to conduct a a more general Bayesian analysis while restricting attention to Gaussian mixtures. In their analysis, the authors put themselves in the setting where no strong prior information on the components of such a mixture is available. In their work, the reader finds a thorough sensitivity analysis including the dependence of the posterior distribution of the number of components on the chosen prior for the means and variances. In the very interesting work of [16] the authors study the frequentist properties of Bayesian estimators of the true order in nested models including mixture models. Bounds on underestimation and overestimation of the Bayesian estimators are obtained. In particular, it is shown that, under some regularity conditions, the probability of underestimation decays exponentially. For further articles using Bayesian theory for clustering, we can refer to [51], [61], [30], [53] and the references therein.

We finish this section by drawing the reader's attention to existence of whole body of literature on mixture estimation and clustering, at the intersection of Statistics and Computing. This includes research papers on extensions or modifications of the famous EM-algorithm and numerical implementation of various information criteria. We refer to [15] where an entropy criteria was considered, which is derived from a simple relationship between the likelihood and classification likelihood of a mixture. In [27] an algorithm based on the Minimum Message Length (MML) criterion was implemented with the aim of selecting the best overall mixture model given the observed data (using a variant of the EM-algorithm). The very recent paper of [32] presents a novel form of cross-validation approach which is adaptive to the data. The paper contains an excellent literature review and the ideas discussed there are highly relevant, especially in connection with the question of how to choose the penalty function in the penalized methods reviewed above.

9 Some Conclusions

As acknowledged in the literature on the mixture model estimation and is supported by the results presented in Sect. 6, estimation of the number of components in a mixture distribution is a challenging task. None of the methods examined in the previous sections of the present survey can be regarded as a reliable universal tool that can be applied in any setting without second thoughts.

The widespread use among practitioners of BIC and ICL techniques is justified. These methods are easy to implement, do not require much computational time and outperform in many settings the other methods we have reviewed here. Whenever ICL tends to underestimate the number of components, which happens when they are not well separated, BIC does not seem to ehibit the same behavior producing more reliable estimates of the mixture order.

The LRT approach appears to be beneficial in terms of accuracy for small and average sample sizes, in particular in the settings when the mixture is not well-separated or when some components in the mixture noticeably dominate the other components or whenever the actual number of components in the mixture is large (e.g. 3 or 4

components). The disadvantage of this technique is the large amount of time needed for the computation.

The bootstrap methods on the average perform better than their penalty-based counterparts. In most of the settings MHD approach with bootstrap is more accurate than both MHD with the AIC-based penalty term and MHD with SBC-based term. The undeniable advantage of the procedures using the bootstrap is that the obtained estimates do not depend on the form of the penalty term, thus the errors resulting from the poor choice thereof can be avoided. The disadvantage however is the computational intensity of this procedure.

In the settings where the mixtures have well separated components LRT and MHD with bootstrap approaches provide an improvement over the other methods whenever the number of the components is more than 2. If a mixture comprises only 2 components and many observations are available, any of the methods can be applied.

The distance-based methods and the LRT approach can be used in the setting where the parameter values need to be estimated. If there is no such requirement, methods based on the Hankel matrix of moments of the mixing distributions can be used. For small sample sizes the scaled version of the Hankel matrix-based method achieves better performance than other methods. But generally it does not provide better accuracy than either the other methods, nor one can benefit significantly in terms of computational time.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42519-022-00289-1.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich.

Declarations

Conflict of interest statement On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Aitkin M, Anderson D, Hinde J (1981) Statistical modelling of data on teaching styles. J R Stat Soc Ser A (General) 144(4):419–448. https://doi.org/10.2307/2981826
2. Akaike H (1998) Information theory and an extension of the maximum likelihood principle. In: Selected papers of Hirotugu Akaike, pp 199–213. Springer
3. Aldrich J (1997) Ra fisher and the making of maximum likelihood 1912–1922. Stat Sci 12(3):162–176

4. Adelchi A, Bowman Adrian W (1990) A look at some data on the old faithful geyser. J R Stat Soc Ser C (Appl Stat) 39(3):357–365. https://doi.org/10.2307/2347385
5. Balabdaoui F, Butucea C (2014) On location mixtures with pólya frequency components. Stat Probab Lett 95:144–149. https://doi.org/10.1016/j.spl.2014.08.013
6. Balabdaoui F, de Fournas-Labrosse G (2020) Least squares estimation of a completely monotone pmf: from analysis to statistics. J Stat Plan Inference 204:55–71. https://doi.org/10.1016/j.jspi.2019.04.006
7. Benaglia T, Chauveau D, Hunter DR, Young DS (2010) mixtools: an r package for analyzing mixture models. J Stat Softw 32:1–29
8. Beran R (1977) Minimum hellinger distance estimates for parametric models. Ann Stat 5:445–463
9. Biernacki C, Celeux G, Govaert G (1999) An improvement of the nec criterion for assessing the number of clusters in a mixture model. Pattern Recogn Lett 20(3):267–272
10. Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans Pattern Anal Mach Intell 22(7):719–725. https://doi.org/10.1109/34.865189
11. Christophe B, Gilles C, Gérard G, Florent L (2006) Model-based cluster and discriminant analysis with the mixmod software. Comput Stat Data Anal 51(2):587–600. https://doi.org/10.1016/j.csda.2005.12.015
12. Bishop CM et al (1995) Neural networks for pattern recognition. Oxford University Press, Oxford
13. Böhning D, Dietz E, Schaub R, Schlattmann P, Lindsay BG (1994) The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. Ann Inst Stat Math 46(2):373–388
14. Richard Morrison Cassie (1954) Some uses of probability paper in the analysis of size frequency distributions. Mar Freshw Res 5(3):513–522. https://doi.org/10.1071/MF9540513
15. Celeux G, Soromenho G (1996) An entropy criterion for assessing the number of clusters in a mixture model. J Classif 13(2):195–212
16. Chambaz A, Rousseau J (2008) Bounds for Bayesian order identification with application to mixtures. Ann Stat 36(2):938–962. https://doi.org/10.1214/009053607000000857
17. Chen H, Chen J, Kalbfleisch JD (2001) A modified likelihood ratio test for homogeneity in finite mixture models. J R Stat Soc Ser B (Stat Methodol) 63(1):19–29. https://doi.org/10.1111/1467-9868.00273
18. Chen J, Kalbfleisch JD (1996) Penalized minimum-distance estimates in finite mixture models. Can J Stat 24(2):167–175. https://doi.org/10.2307/3315623
19. Crawford SL (1994) An application of the laplace method to finite mixture distributions. J Am Stat Assoc 89(425):259–267
20. Cutler A, Cordero-Brana OI (1996) Minimum hellinger distance estimation for finite mixture models. J Am Stat Assoc 91(436):1716–1723
21. Dacunha-Castelle D, Gassiat E (1997) The estimation of the order of a mixture model. Bernoulli. https://doi.org/10.2307/3318593
22. Day NE (1969) Estimating the components of a mixture of normal distributions. Biometrika 56(3):463–474
23. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc Ser B (Methodol) 39(1):1–22
24. Efron B, Thisted R (1976) Estimating the number of unseen species: How many words did shakespeare know? Biometrika 63(3):435–447
25. Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. J Am Stat Assoc 90(430):577–588
26. Ferguson TS (2017) A course in large sample theory. Routledge, London
27. Figueiredo MAT, Jain AK (2002) Unsupervised learning of finite mixture models. IEEE Trans Pattern Anal Mach Intell 24(3):381–396. https://doi.org/10.1109/34.990138
28. Fisher RA (1937) Professor Karl Pearson and the method of moments. Ann Eugen 7(4):303–318
29. Fisher RA (1997) On an absolute criterion for fitting frequency curves. Stat Sci 12(1):39–41
30. Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97(458):611–631. https://doi.org/10.1198/016214502760047131
31. Fraley C, Raftery AE (2007) Bayesian regularization for normal mixture estimation and model-based clustering. J Classif 24(2):155–181
32. Fu W, Perry PO (2020) Estimating the number of clusters using cross-validation. J Comput Gr Stat 29(1):162–173. https://doi.org/10.1080/10618600.2019.1647846

33. Groeneboom P, Jongbloed G, Wellner JA (2008) The support reduction algorithm for computing non-parametric function estimates in mixture models. Scand J Stat 35(3):385–399. https://doi.org/10.1111/j.1467-9469.2007.00588.x
34. Harding JP (1949) The use of probability paper for the graphical analysis of polymodal frequency distributions. J Mar Biol Assoc UK 28(1):141–153
35. Holzmann H, Munk A, Stratmann B (2004) Identifiability of finite mixtures-with applications to circular distributions. Sankhya Indian J Stat 5:440–449
36. Holzmann H, Munk A, Gneiting T (2006) Identifiability of finite mixtures of elliptical distributions. Scand J Stat 33(4):753–763. https://doi.org/10.1111/j.1467-9469.2006.00505.x
37. James LF, Marchette DJ, Priebe CE (2001) Consistent estimation of mixture complexity. Ann Stat 29(5):1281–1296. https://doi.org/10.1214/aos/1013203454
38. Karlis D, Xekalaki E (1999) On testing for the number of components in a mixed Poisson model. Ann Inst Stat Math 51(1):149–162
39. Kent JT (1983) Identifiability of finite mixtures for directional data. Ann Stat 2:984–988
40. Keribin C (2000) Consistent estimation of the order of mixture models. Sankhya Indian J Stat Ser A 2:49–66
41. Lehmann EL (2012) Some principles of the theory of testing hypotheses. In: Selected works of EL Lehmann, pp 139–164. Springer
42. Leroux BG (1992) Consistent estimation of a mixing distribution. Ann Stat 2:1350–1360
43. LINDSAY BG (1995) Mixture models: theory, geometry, and applications. In: NSFCBMS regional conference series in probability and statistics, vol 5. Institute of Mathematical Statistics
44. Lindsay BG (1983) The geometry of mixture likelihoods: a general theory. Ann Stat 2:86–94
45. Lindsay BG (1983) The geometry of mixture likelihoods, part ii: the exponential family. Ann Stat 11(3):783–792
46. Lindsay BG (1989) Moment matrices: applications in mixtures. Ann Stat 17(2):722–740. https://doi.org/10.1214/aos/1176347138
47. Liu X, Shao Y (2003) Asymptotics for likelihood ratio tests under loss of identifiability. Ann Stat 31(3):807–832. https://doi.org/10.1214/aos/1056562463
48. McLachlan GJ, Krishnan T (2007) The EM algorithm and extensions, vol 382. Wiley, London
49. McLachlan GJ, Peel D (2004) Finite mixture models. Wiley, London
50. Melnykov V, Maitra R (2010) Finite mixture models and model-based clustering. Stat Surv 4:80–116. https://doi.org/10.1214/09-SS053
51. Mengersen KL (1996) Testing for mixtures: a bayesian entropic approach. Bayesian Stat 3:255–276
52. Newcomb S (1886) A generalized theory of the combination of observations so as to obtain the best result. Am J Math 2:343–366
53. Nobile A (2004) On the posterior distribution of the number of components in a finite mixture. Ann Stat 32(5):2044–2073. https://doi.org/10.1214/009053604000000788
54. O'Malley T, Bursztein E, Long J, Chollet F, Jin H, Invernizzi L et al. (2019) Kerastuner. https://github.com/keras-team/keras-tuner
55. Pan W, Shen X (2007) Penalized model-based clustering with application to variable selection. J Mach Learn Res 8(5):5528
56. Pearson K (1894) Contributions to the mathematical theory of evolution. Philos Trans R Soc Lond A 185:71–110
57. Pearson K (1936) Method of moments and method of maximum likelihood. Biometrika 28(1/2):34–59
58. Preston EJ (1953) A graphical method for the analysis of statistical distributions into two normal components. Biometrika 40(3/4):460–464
59. Rao CR (1948) The utilization of multiple measurements in problems of biological classification. J R Stat Soc Ser B (Methodol) 10(2):159–203
60. Richardson S, Green PJ (1997) On bayesian analysis of mixtures with an unknown number of components (with discussion). J R Stat Soc Ser B (Stat Methodol) 59(4):731–792. https://doi.org/10.1111/1467-9868.00095
61. Roeder K, Wasserman L (1997) Practical bayesian density estimation using mixtures of normals. J Am Stat Assoc 92(439):894–902. https://doi.org/10.1080/01621459.1997.10474044
62. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 2:461–464
63. Scrucca L, Fop M, Murphy TB, Raftery AE (2016) mclust 5: clustering, classification and density estimation using gaussian finite mixture models. R J 8(1):289

64. Self SG, Liang K-Y (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 82(398):605–610
65. Teicher H (1961) Identifiability of mixtures. Ann Math Stat 32(1):244–248
66. Teicher H (1963) Identifiability of finite mixtures. Ann Math Stat 5:1265–1269
67. Thisted RA (1996) Elements of statistical computing. Numer Comput 2:89
68. Titterington DM, Afm S, Smith AFM, Makov UE et al (1985) Statistical analysis of finite mixture distributions, vol 198. Wiley, London
69. Umashanger T, Sriram TN (2009) L2e estimation of mixture complexity for count data. Comput Stat Data Anal 53(12):4243–4254. https://doi.org/10.1016/j.csda.2009.05.013
70. Wang Y (2007) On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. J R Stat Soc Ser B (Stat Methodol) 69(2):185–198. https://doi.org/10.1111/j.1467-9868.2007.00583.x
71. Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Stat 9(1):60–62
72. Wolfe JH (1965) A computer program for the maximum likelihood analysis of types. Technical report, Naval Personnel Research Activity San Diego USA
73. Wolfe JH (1967) Normix: computational methods for estimating the parameters of multivariate normal mixtures of distributions. Technical report, Naval Personnel Research Activity San Diego Calif
74. Woo M-J, Sriram TN (2006) Robust estimation of mixture complexity. J Am Stat Assoc 101(476):1475–1486. https://doi.org/10.1198/016214506000000555
75. Woo M-J, Sriram TN (2007) Robust estimation of mixture complexity for count data. Comput Stat Data Anal 51(9):4379–4392. https://doi.org/10.1016/j.csda.2006.06.006
76. Yakowitz SJ, Spragins JD (1968) On the identifiability of finite mixtures. Ann Math Stat 39(1):209–214