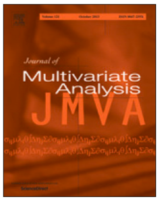
Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva



Selecting the number of clusters, clustering models, and algorithms. A unifying approach based on the quadratic discriminant score

Luca Coraggio [a,*], Pietro Coretto [b]

[a] *Dipartimento di Scienze Economiche e Statistiche, University of Naples Federico II, Italy*
[b] *Dipartimento di Scienze Economiche e Statistiche, University of Salerno, Italy*

ARTICLE INFO

Article history:
Received 3 May 2022
Received in revised form 18 March 2023
Accepted 18 March 2023
Available online 23 March 2023

AMS 2020 subject classifications:
primary 62H30
secondary 62F07

Keywords:
Cluster validation
Mixture models
Model-based clustering
Resampling methods

ABSTRACT

Cluster analysis requires fixing the number of clusters and often many hyper-parameters. In practice, one produces several partitions, and a final one is chosen based on validation or selection criteria. There exist an abundance of validation methods that, implicitly or explicitly, assume a certain clustering notion. In this paper, we focus on groups that can be well separated by quadratic or linear boundaries. The reference cluster concept is defined through the quadratic discriminant function and parameters describing clusters' size, center and scatter. We develop two cluster-quality criteria that are consistent with groups generated from a class of elliptic–symmetric distributions. Using the bootstrap resampling of the proposed criteria, we propose a selection rule that allows choosing among many clustering solutions, eventually obtained from different methods. Extensive experimental analysis shows that the proposed methodology achieves a better overall performance compared to established alternatives from the literature.

1. Introduction

The typical workflow in cluster analysis is to run one or more algorithms with various settings producing several partitions, among which a researcher needs to choose a final one. There may be multiple partitions that describe the data well according to different clusters' concepts [30]. Because of the intrinsic unsupervised nature of the clustering problem, the selection of the desired cluster solution remains a long-standing and open problem [25]. The most significant common issue to all methods and algorithms is choosing an appropriate number of groups, K. However, K is not the only relevant decision: many clustering methods and algorithms also require hyper-parameters that control the complexity level at which the data structure is represented. Similar methods with different hyper-parameters may discover different partitions of a given data set, even at the fixed "true" K. In the Supplementary Material, Section S2, we provide an example on the well-known Iris data set [3]. There is a vast catalog of methods proposed to solve the selection problem; for a recent comprehensive overview, see [21]. Traditionally, in cluster analysis, the selection of the desired partition has been treated as a validation problem rather than a model selection problem. This is probably because many clustering methods are not derived from stochastic models, although most are built around at least some implicit model assumptions. Recently, [38] attempted to categorize different types of validation approaches. Our proposal contributes to the literature on internal

[*] Correspondence to: Ufficio H11, Edificio 3, Dipartimento di Scienze Economiche e Statistiche - Università degli Studi di Napoli Federico II - Complesso Universitario di Monte Sant'Angelo, Via Cintia, 21, 80126 Napoli, Italy.
E-mail address: luca.coraggio@unina.it (L. Coraggio).

https://doi.org/10.1016/j.jmva.2023.105181

validation methods, which are methods using the same data used to fit the clusters. The advantage of internal methods is that they do not require additional information that is sometimes expensive to collect. Typically, new proposals are advertised claiming their universal ability to discover the data's "true" groups. However, in pure unsupervised contexts, true groups do not exist. Furthermore, it is often overlooked that each method pursues a specific notion of clusters, which implicitly or explicitly assumes the existence of certain structures in the data. As noted in [2], one needs to choose the validation approach that is consistent with the primary goal of the analysis. There are method-dependent validation methods, specifically designed to evaluate the output of a specific clustering method and method-independent methods that can potentially evaluate the output of any clustering methodology. However, even method-independent validation approaches privilege a certain idea of clusters.

In this paper, we take a different approach: we first define a notion of clusters that different clustering methods may retrieve and then propose a method-independent validation criterion to measure the quality of such clusters. Specifically, we look for clusters that can be well separated by quadratic boundaries or linear boundaries as a special case. These clusters are consistent with a class of elliptically-symmetric distributions (ESD), where the within-group dependence structure of the features is mainly driven by correlation. The quest for clusters of this type is rather common in applications [18].

Related literature. Model-based clustering (MBC) methods, based on ML estimation of finite mixture models of ESDs, are strong candidates for capturing the clusters mentioned above. Assuming that each of the K mixture components generates a group, the selection of the desired clustering solution is translated into a model-selection problem where, in practice, the likelihood fit is contrasted with a penalty accounting for model complexity. The most popular selection strategy is to use information criteria such as the BIC and the AIC [9,31].

Although information-type criteria are based on a solid theoretical background, there are some issues with their application to cluster selection. In [29], Keribin showed that the BIC is consistent for the number of mixture components under somewhat restrictive assumptions, but practitioners tend to believe that this result is more general, causing some faith in it. The consistency notion of [29] is for the recovery of the underlying data distribution and not for clusters. Paradoxically, these consistency results are problematic for cases where the mixture model is not meant to capture the "true" underlying distribution but rather for approximating the density regions formed by the clusters. Finite Gaussian mixtures can approximate a large class of distributions [34], implying that consistent criteria like the BIC will include additional components inflating K if, for example, a group that is only approximately normal is better fitted by more than one Gaussian component. The Integrated Complete-data Likelihood (ICL) criterion of [8] [see also 5] modifies the BIC, adapting it to solve the clustering problem. Another model-selection approach, based on likelihood-type criteria derived from mixture models, is the cross-validation method proposed in [37]. An additional drawback of information-type indexes is that they are method-dependent: they only allow to compare solutions from MBC methods because their calculation is based on likelihood quantities and models' degrees of freedom. A further issue is that, in some cases, these criteria cannot be calculated for MBC methods when the effective degrees of freedom of the underlying model cannot be derived (see the case of ML for Gaussian mixtures with the eigen-ratio regularization treated in Section 4).

Outside the MBC context, there are many method-independent internal validation indexes that mostly measure the within-cluster homogeneity contrasted to a measure of between-clusters heterogeneity. Notable examples are the popular CH index of [10] and the Average Silhouette Width criteria (ASW) of [27]. These are not genuinely model-free indexes because they purse cluster shapes that depend on the underlying dissimilarity notion. These indexes have in common with the BIC-type criteria that they implicitly attempt to contrast the cluster fit vs., the increased complexity caused by the increase in K.

Another idea from the literature that inspired some aspects of the present contribution is that of stability selection [6]. The idea taken from this literature is not the notion of stability, which is about finding similar clusterings on similar data sets [13,24], but the idea of exploring variations of the clusterings based on perturbations of the data set obtained by bootstrap resampling of the original data.

Contribution and organization of the paper. We develop a framework where each cluster is represented by a triplet of parameters representing the notions of size, location and scatter. This allows to map clusterings obtained with different methods in a form that is consistent with the notion previously discussed. The method-independent nature of our proposal is a major advantage over competitors from the MBC literature. These clusters' parameters and the quadratic score function, at the heart the Quadratic Discriminant Analysis (QDA), are used to develop two cluster quality criteria called quadratic scores. These criteria are shown to be consistent with clusters generated from a restricted class of ESDs, including the popular Gaussian model (Section 2.1). We show connections between the proposed criteria and likelihood-type quantities related to finite mixtures of ESDs and, in particular, Gaussian mixtures (see Section 2.2). In the same spirit of the pioneering work of [1] on model-selection, we propose to select a clustering solution produced by a method that achieves the largest expected score across all possible partitions of data sets sampled from the data distribution. The expected score and its confidence interval are approximated via empirical bootstrap in Section 3. Finally, in Section 4, we propose an extensive numerical analysis where the proposed method is compared against some alternatives on both real and artificial data sets. Overall, the proposed methodology shows a superior performance and proves to be able to retrieve interesting clustering solutions even in adverse circumstances. Proofs of the statements are given in Appendix A; additional examples and details are given in Supplementary Material.

2. Quadratic scoring

We fix some general notation used throughout the rest of the paper. The general clustering problem is to construct a partition $\mathcal{G}_K = \{G_k,\ k \in \{1, \dots, K\}\}$ allocating the objects $\{1, \dots, n\}$ into K groups, where K is generally unknown. Let $\mathbb{X}_n = \{\boldsymbol{x}_i,\ i \in \{1, \dots, n\}\}$ be an observed sample of p-dimensional feature vectors $\boldsymbol{x}_i \in \mathbb{R}^p$; $\mathbb{X}_n$ is the observed version of a random sample $\mathcal{X}_n = \{X_i,\ i \in \{1, \dots, n\}\}$, where $X_i \in \mathbb{R}^p$ is the p-dimensional random vector of features representing the ith unit. In clustering, a typically unsupervised task, we observe the features, but we do not observe the group memberships that we want to discover. Group memberships are introduced through the random vector of 0–1 variables $Z = (Z_1, \dots, Z_K)^\top$, where $Z_k = 1$ denotes membership to the kth group. For the ith sample point we define the group memberships as $Z_{ik} = \mathbb{I}\{i \in G_k\}$, where $\mathbb{I}\{\cdot\}$ is the usual indicator function.

2.1. The reference cluster concept

Assume that $X \sim F$, where F is the population distribution function producing K clustered regions of points. We assume that each cluster $k \in \{1, \dots, K\}$ is meaningfully described by the triplet of parameters $\boldsymbol{\theta}_k = \left\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right\}$ formalizing the notions of size, center and scatter. For the kth cluster, π_k is the expected fraction of points belonging to the kth group, $\boldsymbol{\mu}_k \in \mathbb{R}^p$ is the vector of centers and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p\times p}$ is a positive definite scatter matrix that either coincides with or is proportional to the group's covariance matrix. A cluster configuration m of K groups is represented by the parameter vector $\boldsymbol{\theta}^{(m)}$ including all unique elements of the objects $\{(\pi_k^{(m)}, \boldsymbol{\mu}_k^{(m)}, \boldsymbol{\Sigma}_k^{(m)}),\ k \in \{1, \dots, K\}\}$. Since different $\boldsymbol{\theta}^{(m)}$ may refer to cluster configurations with a different number of groups, depending on the context, we will often use the notation $K(\boldsymbol{\theta}^{(m)})$, or $K(m)$, to denote the number of groups described by $\boldsymbol{\theta}^{(m)}$. The set of possible configurations is denoted with $\mathcal{M}$. The superscript (m) is dropped if it is unnecessary to index more than one cluster configuration, $m \in \mathcal{M}$. Note that $\boldsymbol{\theta}$ is a parameter serving as a general description of the clustered region but, in general, we do not presume that F is necessarily a function of $\boldsymbol{\theta}$. Given a configuration $\boldsymbol{\theta}$, we look for clusters that form a partition of the data space into K disjoint subsets $\mathcal{Q}(\boldsymbol{\theta}) = \{Q_k(\boldsymbol{\theta}),\ k \in \{1, \dots, K\}\}$,

$$Q_k(\boldsymbol{\theta}) := \left\{\boldsymbol{x} \in \mathbb{R}^p :\ \operatorname{qs}(\boldsymbol{x}, \boldsymbol{\theta}_k) = \max_{1\leq j\leq K} \operatorname{qs}(\boldsymbol{x}, \boldsymbol{\theta}_j)\right\}, \tag{1}$$

where $\operatorname{qs}(\boldsymbol{x}, \boldsymbol{\theta}_k)$ is the quadratic score function at $\boldsymbol{x}$ according to $\boldsymbol{\theta}_k$, that is

$$\operatorname{qs}(\boldsymbol{x}, \boldsymbol{\theta}_k) := \log(\pi_k) - \frac{1}{2}\log\left(\det(\boldsymbol{\Sigma}_k)\right) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top {\boldsymbol{\Sigma}_k}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k). \tag{2}$$

From now onward, we call $\mathcal{Q}(\boldsymbol{\theta})$ the quadratic partition. A point $\boldsymbol{x}$ is defined to belong to the group for which the quadratic score is maximized. Hence, $\operatorname{qs}(\boldsymbol{x}, \boldsymbol{\theta}_k)$ can generally be interpreted as a measure of the fit of $\boldsymbol{x}$ into the kth cluster according to $\boldsymbol{\theta}_k$. Note that $\exp(\operatorname{qs}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)) \propto \pi_k\phi(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\phi(\cdot, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the multivariate normal density function with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$. The classical interpretation of (1) is that it represents the optimal classification boundaries under the Gaussian assumption. As noted in [23], in practice, the quadratic score can effectively describe partitions well beyond Gaussianity whenever quadratic and linear boundaries can adequately separate clustered regions. The following result states that the partition in (1) is consistent with a class of elliptic–symmetric models that includes the Gaussian.

Proposition 1. *Assume* $\Pr\{Z_k = 1\} = \pi_k$ *and that for all* $k \in \{1, \dots, K\}$ *the group-conditional distribution, i.e., the distribution of* $X \mid Z_k = 1$, *has the density function*

$$f(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} g\left((\boldsymbol{x} - \boldsymbol{\mu}_k)^\top {\boldsymbol{\Sigma}_k}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right), \tag{3}$$

where $g(\cdot)$ *is a strictly decreasing function on* $[0, +\infty)$, $\boldsymbol{\mu}_k \in \mathbb{R}^p$ *is the centrality parameter and* $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p\times p}$ *is a positive definite scatter matrix. Assume at least one of the following:*

(C1) $f(\cdot)$ *is the Gaussian density function (for an appropriate choice of* $g(\cdot)$*);*

(C2) $\pi_i \det(\boldsymbol{\Sigma}_i)^{-\frac{1}{2}} = \pi_j \det(\boldsymbol{\Sigma}_j)^{-\frac{1}{2}}$, $i \neq j$, $i, j = 1, \dots, K$.

Then, for any partition of the feature space $\{A_k,\ k \in \{1, \dots, K\}\}$,

$$\Pr\left\{\bigcup_{k=1}^{K}\{Z_k = 1 \cap X \in A_k\}\right\} \leq \Pr\left\{\bigcup_{k=1}^{K}\{Z_k = 1 \cap X \in Q_k(\boldsymbol{\theta})\}\right\}, \tag{4}$$

where $Q_k(\boldsymbol{\theta}) \in \mathcal{Q}(\boldsymbol{\theta})$ *is defined in* (1).

The previous result connects and develops ideas from linear classification and its connections to elliptically-symmetric families investigated in [39].

Remark 1. The quadratic partition achieves the largest probability that its members contain points generated from the K sub-populations. The group-conditional model (3) includes popular unimodal models like the Gaussian, the Student-t, the Laplace, the multivariate logistic, etc. These models generate groups of points lying in regions that are intersections of ellipsoids described by the pairs $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and, within each group, the features are connected via their joint linear dependence. The generating mechanism assumed in Proposition 1 is consistent with data generated from finite mixtures of such elliptically-symmetric families. Outside the Gaussian case (C1), Proposition 1 is restricted to the cases where groups have a comparable square root of the generalized precision, $\det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}}$, after weighting by the cluster size π_k. A special case of (C2) is when groups are balanced (equal sizes π_k) and homoscedastic (equal dispersions $\boldsymbol{\Sigma}_k$).

2.2. Scoring cluster configurations

Given $\mathbb{X}_n$, we want to measure how well a cluster configuration $\boldsymbol{\theta}^{(m)}$ organizes these points within the quadratic partition. We want to select the "boxes" $\{Q_k(\boldsymbol{\theta}),\ k = 1, \ldots, K\}$ that best represents the clustered points. Let $B^k_\varepsilon(\boldsymbol{x}_i; \boldsymbol{\theta})$ be a ball of radius $\varepsilon > 0$, centered at $\boldsymbol{x}_i$, such that $B^k_\epsilon(\boldsymbol{x}_i; \boldsymbol{\theta}) \subset Q_k(\boldsymbol{\theta})$, i.e., $B^k_\varepsilon(\boldsymbol{x}; \boldsymbol{\theta}) := \{\boldsymbol{y} \in \mathbb{R}^p : \|\boldsymbol{y} - \boldsymbol{x}\| < \varepsilon \cap \boldsymbol{y} \in Q_k(\boldsymbol{\theta})\}$. For ε sufficiently small, the joint probability that all points in $\mathbb{X}_n$ are accommodated in the quadratic partition consistently with the underlying group memberships is

$$\prod_{i=1}^{n} \Pr\left\{Z_k = 1 \cap X_i \in B^k_\varepsilon(\boldsymbol{x}_i; \boldsymbol{\theta})\right\} = \prod_{i=1}^{n} \Pr\{Z_k = 1\} \Pr\left\{X_i \in B^k_\varepsilon(\boldsymbol{x}_i; \boldsymbol{\theta}) \mid Z_k = 1\right\}. \tag{5}$$

Under the generating process of Proposition 1, taking $\varepsilon \to 0$, the probability law (5) is transformed into its density representation

$$\mathcal{L}_n(\boldsymbol{\theta}) := \prod_{i=1}^{n} \prod_{k=1}^{K(\boldsymbol{\theta})} \left(\pi_k f(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{\mathbb{I}\{\boldsymbol{x}_i \in Q_k(\boldsymbol{\theta})\}}, \tag{6}$$

where $\mathbb{I}\{\cdot\}$ is the usual indicator function. (6) closely resembles the likelihood function for a partition model [see 19, Ch. 7]. However, this is not exactly the case: for a partition model, we would have had class membership indicators replacing $\mathbb{I}\{\boldsymbol{x}_i \in Q_k(\boldsymbol{\theta})\}$ in (6). Taking the logarithm of (6), we would like to achieve the largest

$$L_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K(\boldsymbol{\theta})} \mathbb{I}\{\boldsymbol{x}_i \in Q_k(\boldsymbol{\theta})\} \log(\pi_k f(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)). \tag{7}$$

Evaluation of (7) requires the knowledge of the specific group-conditional model $f(\cdot)$. However, we want to evaluate the quality of the partition even when the group-conditional distribution is not precisely known. Proposition 1 states that, for certain group-conditional distributions, point-wise maximization of the quadratic score in the feature space well captures the main clustered regions. We propose to rank cluster configurations based on the following hard score criterion:

$$H_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K(\boldsymbol{\theta})} \mathbb{I}\{\boldsymbol{x}_i \in Q_k(\boldsymbol{\theta})\}\, \mathrm{qs}(\boldsymbol{x}_i; \boldsymbol{\theta}_k). \tag{8}$$

We call it hard because $H_n(\cdot)$ is a weighted average of the points score with the 0–1 "hard" weights $\mathbb{I}\{\boldsymbol{x}_i \in Q_k(\boldsymbol{\theta})\}$. Interpreting $\mathrm{qs}(\boldsymbol{x}_i; \boldsymbol{\theta}_k)$ as the strength at which the object i is assigned to the kth group, (8) is the average strength achieved by a cluster configuration. Despite this qualitative interpretation of $H_n(\cdot)$, there is a connection between (7) and (8) at the population level, based on the fact that $\mathrm{qs}(\boldsymbol{x}_i; \boldsymbol{\theta}_k)$ contains the kernel of the Gaussian density. Under regularity conditions, both sample averages (7) and (8) will asymptotically approach their population counterparts

$$L(\boldsymbol{\theta}) = \sum_{k=1}^{K(\boldsymbol{\theta})} \int_{Q_k(\boldsymbol{\theta})} \log(\pi_k f(\boldsymbol{x}; \boldsymbol{\theta}_k)) dF \quad \text{and} \quad H(\boldsymbol{\theta}) = \sum_{k=1}^{K(\boldsymbol{\theta})} \int_{Q_k(\boldsymbol{\theta})} \mathrm{qs}(\boldsymbol{x}; \boldsymbol{\theta}_k) dF, \tag{9}$$

respectively. The following proposition clarifies the relationship between $H(\cdot)$ and $L(\cdot)$.

Proposition 2. *Assume that the following integrals exist and that*

(C3) $\inf_{\boldsymbol{\theta}^{(m)} \in \Theta_M} \left\{\int_{Q_k(\boldsymbol{\theta}^{(m)})} \log f(\boldsymbol{x}; \boldsymbol{\mu}^{(m)}_k, \boldsymbol{\Sigma}^{(m)}_k) dF - \int_{Q_k(\boldsymbol{\theta}^{(m)})} \log \phi(\boldsymbol{x}; \boldsymbol{\mu}^{(m)}_k, \boldsymbol{\Sigma}_k k) dF\right\} \geq 0,\ k \in \left\{1, \ldots, K(\boldsymbol{\theta}^{(m)})\right\}$.

Then

$$H(\boldsymbol{\theta}^{(m)}) = c + L(\boldsymbol{\theta}^{(m)}) - \Lambda(\boldsymbol{\theta}^{(m)}), \tag{10}$$

where c is a positive constant, and

$$\Lambda(\boldsymbol{\theta}^{(m)}) = \sum_{k=1}^{K} \int_{Q_k(\boldsymbol{\theta}^{(m)})} \log\left(\frac{f(\boldsymbol{x}; \boldsymbol{\mu}^{(m)}_k, \boldsymbol{\Sigma}^{(m)}_k)}{\phi(\boldsymbol{x}; \boldsymbol{\mu}^{(m)}_k, \boldsymbol{\Sigma}^{(m)}_k)}\right) dF \geq 0.$$

At the population level, the hard score criterion can be interpreted as the quality of the fitting of the partition, expressed by $L(\cdot)$, minus a penalty term, $\Lambda(\boldsymbol{\theta}^{(m)}) \geq 0$, that measures the departure from the Gaussian clusters' prototype model embedded into the quadratic score function. When clusters are truly Gaussian, i.e., $f(\cdot) = \phi(\cdot)$, then $\Lambda(\boldsymbol{\theta}^{(m)}) = 0$ and $H(\boldsymbol{\theta}^{(m)}) \propto L(\boldsymbol{\theta}^{(m)})$. Condition (C3) is needed to interpret the criterion: it ensures that $\Lambda(\boldsymbol{\theta}^{(m)}) \geq 0$ for any possible cluster configuration $\boldsymbol{\theta}^{(m)}$ under comparison so that it works as a penalty. (C3) obeys to the natural principle that, whenever we pick a configuration $\boldsymbol{\theta}$, the approximating Gaussian model underlying $\mathrm{qs}(\cdot)$ cannot fit the quadratic regions better than the underlying true generating model $f(\cdot)$. Indeed, (C3) is violated if there exists a configuration $\boldsymbol{\theta}^{(m)}$ for which $\int_{Q_k(\boldsymbol{\theta}^{(m)})} \log f(X; \boldsymbol{\mu}_k^{(m)}, \boldsymbol{\Sigma}_k^{(m)}) dF < \int_{Q_k(\boldsymbol{\theta}^{(m)})} \log \phi(X; \boldsymbol{\mu}_k^{(m)}, \boldsymbol{\Sigma}_k^{(m)}) dF$, where these integrals can be seen as the expected log-likelihood contribution over the kth members of the quadratic partition under $f(\cdot)$ and $\phi(\cdot)$, respectively. From Proposition 2, it immediately follows that

$$\underset{m\in\mathcal{M}}{\arg\max}\, H(\boldsymbol{\theta}^{(m)}) = \underset{m\in\mathcal{M}}{\arg\max} \left\{ L(\boldsymbol{\theta}^{(m)}) - \Lambda(\boldsymbol{\theta}^{(m)}) \right\}.$$

Since $\mathrm{qs}(\cdot)$ measures the strength at which a point is assigned to a cluster, a smooth weighting is obtained by normalizing the quadratic scores. We propose to use the softmax transformation, that is the ith point's weight into the kth cluster is

$$\tau_k(\boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{\exp(\mathrm{qs}(\boldsymbol{x}_i; \boldsymbol{\theta}_k))}{\sum_{i=1}^{n} \exp(\mathrm{qs}(\boldsymbol{x}_i; \boldsymbol{\theta}_k))}. \tag{11}$$

The corresponding smooth score criterion is defined as

$$T_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K(\boldsymbol{\theta})} \tau_k(\boldsymbol{x}_i; \boldsymbol{\theta})\, \mathrm{qs}(\boldsymbol{x}_i; \boldsymbol{\theta}_k). \tag{12}$$

Other weighting schemes are possible, but the choice of the softmax transformation is because it guarantees some form of optimality for Gaussian clusters (see the following proposition). Under regularity conditions, for sufficiently large n, (12) will approach its population counterpart

$$T(\boldsymbol{\theta}) = \sum_{k=1}^{K(\boldsymbol{\theta})} \int \tau_k(\boldsymbol{x}; \boldsymbol{\theta})\, \mathrm{qs}(\boldsymbol{x}; \boldsymbol{\theta}_k) dF. \tag{13}$$

Under the generating mechanism of Proposition 1, the unconditional distribution of X has the finite mixture density

$$\psi_f(\boldsymbol{x}; \boldsymbol{\theta}) := \sum_{k=1}^{K(\boldsymbol{\theta})} \pi_k f(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{14}$$

For a sample point $\boldsymbol{x}_i \in \mathbb{X}_n$, under (14) define the posterior weights

$$\omega_{f,k}(\boldsymbol{x}_i; \boldsymbol{\theta}) = \Pr\{Z_{ik} = 1 \mid \mathbb{X}_n\} = \frac{\pi_k f(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\psi_f(\boldsymbol{x}_i; \boldsymbol{\theta})}. \tag{15}$$

The ratios defined in (15) are central in MBC methods where the ith object is assigned to the kth component by the following rule

$$\hat{z}_k(\boldsymbol{x}_i; \boldsymbol{\theta}) = \mathbb{I}\left\{ k = \underset{1\leq j\leq K(\boldsymbol{\theta})}{\arg\max}\ \omega_{f,k}(\boldsymbol{x}_i; \boldsymbol{\theta}) \right\}; \tag{16}$$

in practice, $\boldsymbol{\theta}$ is replaced with an estimate. Typically, $\boldsymbol{\theta}$ is fitted based on an ML-type estimator, numerically approximated with the EM-algorithm [31]. The rule (16), called MAP, retrieves the unobservable membership variables $\{Z_{ik}\}$ and coincides with the optimal Bayes classifier if the group-conditional model holds. The MAP rule produces a hard assignment from the smooth (also called fuzzy) membership weights in (15). The overall uncertainty of the assignment (16) reflecting (15) is captured by

$$\mathrm{ent}(X; \boldsymbol{\theta}) = -\sum_{k=1}^{K(\boldsymbol{\theta})} \omega_{f,k}(X; \boldsymbol{\theta}) \log \omega_{f,k}(X; \boldsymbol{\theta}), \tag{17}$$

which is the entropy of the conditional distribution of $Z \mid X$. In situations where clusters are strongly separated the posteriors weights (15) will be close to either 1 or 0 for most points, and the MAP assignment will produce "clear clusters", reflecting the low entropy of $Z \mid X$. On the other hand, cluster configurations with substantial overlap will exhibit large entropy. Let ψ_ϕ be the mixture model (14) when the group-conditional model is the Gaussian density $\phi(\cdot)$, and let $\mathrm{ent}_\phi(\cdot)$ be the corresponding entropy. Moreover, let $\mathrm{d_{KL}}(f_0 \,\|\, g)$ be the Kullback–Leibler discrepancy from the approximating model g to the "true" model f_0.

Proposition 3. *Let f_0 be the density function corresponding to the "true" underlying population distribution function F. Then*

$$\underset{m\in\mathcal{M}}{\arg\max}\ T(\boldsymbol{\theta}^{(m)}) = \underset{m\in\mathcal{M}}{\arg\min}\ \left\{\mathrm{d}_{\mathrm{KL}}(f_0 \parallel \psi_\phi(\cdot;\boldsymbol{\theta}^{(m)})) + \mathrm{E}_F\left[\mathrm{ent}_\phi(X;\boldsymbol{\theta}^{(m)})\right]\right\}, \tag{18}$$

where all expectations are assumed to exist and $\mathrm{E}_F[\cdot]$ *denotes the expectation under* F.

Proposition 3 clarifies that $T(\cdot)$ looks for a compromise between the best approximation of f_0, in the sense of ψ_ϕ, and the lowest entropy of the resulting assignment under the Gaussian prototype model. The entropy term discourages the criteria from focusing on too complex clustering structures. The term $\mathrm{d}_{\mathrm{KL}}(f_0 \parallel \psi_\phi(\cdot;\boldsymbol{\theta}^{(m)}))$ can be made arbitrarily small if $\boldsymbol{\theta}^{(m)}$ is an overly rich description of the density regions produced by F. Indeed, finite Gaussian mixtures can approximate any continuous distribution in a nonparametric sense [34]. However, an overly complex $\boldsymbol{\theta}^{(m)}$ (e.g. $K(\boldsymbol{\theta}^{(m)})$ is large) that describes the density regions too locally would imply a strong overlap and therefore a large $\mathrm{ent}_\phi(\cdot)$.

Propositions 2 and 3 clarify the type of model reference-concept driving the proposed score selection. [5] formulated a parameter estimation criterion based on the right-hand side of (18) to perform MBC. In contrast, here $H(\cdot)$ and $T(\cdot)$ are not meant to be parameter estimation criteria, as the "true" underlying generating model F may well be not a function of the $\boldsymbol{\theta}^{(m)}$ for $m \in \mathcal{M}$. This will become clearer in the examples of Section 2.3, where we show an example where the maximum score cannot identify the true underlying distribution even in the Gaussian case.

Under the Gaussian assumption, there is a further connection between the sample scores $H_n(\cdot)$ and $T_n(\cdot)$ and what is called observed complete data log-likelihood into the MBC literature. For details, we refer the reader to Supplementary Material, Section S3.

2.3. Clusters' boundaries

To see how $H(\cdot)$ and $T(\cdot)$ define the clusters' boundary in Gaussian and non-Gaussian settings, consider the following examples. We define two data generating processes (dgp):

dgpG F is a mixture of two spherical Gaussians in dimension $p = 2$ with equal sizes $\pi_1 = \pi_2 = 0.5$ and equal identity covariance matrix. The first Gaussian component is centered at $\boldsymbol{\mu}_1 = (0, 0)^\mathsf{T}$, while the second component has mean $\boldsymbol{\mu}_2 = (d, 0)^\mathsf{T}$, for some fixed $d > 0$.

dgpU F is a mixture of two uniform distributions with equal volume in dimension $p = 2$ and $\pi_1 = \pi_2 = 0.5$. The first uniform distribution has support on the square $[-1, 1]^2$ with center at $\boldsymbol{\mu}_1 = (0, 0)^\mathsf{T}$. The second uniform distribution takes value on the square $[d-1, d+1] \times [-1, 1]$ with center at $\boldsymbol{\mu}_2 = (d, 0)^\mathsf{T}$, for some fixed $d > 0$.

In both cases, d is the Euclidean distance between the clusters' centers. For $d \in [0, 10]$ we have different data generating processes. For each value of d, we have a different generating distribution function, F_d, that is a mixture of: two Gaussian components in dgpG; two uniform components in dgpU. The dgpU is introduced as a substantial departure from the elliptic assumption of Proposition 1.

We recall that the cluster configuration parameter $\boldsymbol{\theta}^{(m)}$ represents the mth configuration collecting the triplets $(\pi_k^{(m)}, \boldsymbol{\mu}_k^{(m)}, \boldsymbol{\Sigma}_k^{(m)})$ representing the kth cluster size, center and scatter. At each d, we want to compare the population version of the score for two alternative cluster configurations $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}\}$, where $K(\boldsymbol{\theta}^{(1)}) = 1$ and $K(\boldsymbol{\theta}^{(2)}) = 2$. The number of possible choices of such configurations is infinite. Hence, we compare two possible specifications, $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$, that try to reflect the group-conditional distributions corresponding to F_d. The problem here is that the two types of F_d considered in the example are not always a function of cluster configuration parameters. In the dgpG case with $K = 2$, the generating distribution F_d is exactly specified in terms of proportion, mean and covariance parameters of the two Gaussian components. However, for all the remaining cases, this is not true. For example, in the dgpG case with $K = 1$, we need to define $\boldsymbol{\theta}^{(1)} = (\pi^{(m)}, \boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)})$ that does not coincide with the parameters of the corresponding F_d. In each case, we defined competing cluster configuration parameters that are natural descriptions of the group-conditional distributions. We have three different cases.

- dgpG and dgpU with $K = 1$: We set $\boldsymbol{\theta}^{(1)} = \left(\pi_1^{(1)}, \boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)}\right)$ as follows

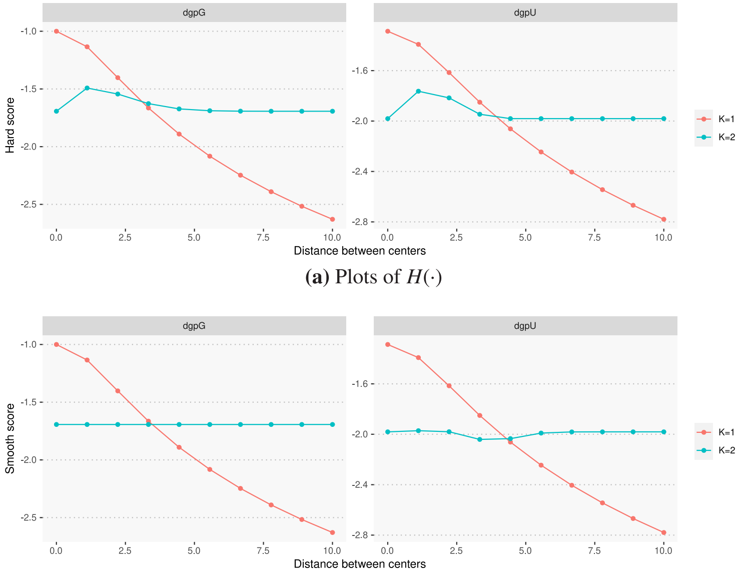
$$\pi_1^{(1)} = 1, \qquad \boldsymbol{\mu}^{(1)} = \int \boldsymbol{x} dF_d, \qquad \boldsymbol{\Sigma}^{(1)} = \int \left(\boldsymbol{x} - \boldsymbol{\mu}^{(1)}\right)\left(\boldsymbol{x} - \boldsymbol{\mu}^{(1)}\right)^\mathsf{T} dF_d. \tag{19}$$

- dgpG with $K = 2$: This is the easiest case, because as previously noted, the parameters of F_d coincide with the parameters of the two groups. In this case, $\boldsymbol{\theta}^{(2)}$ is defined as follows

$$\pi_1^{(2)} = \pi_2^{(2)} = 0.5, \quad \boldsymbol{\mu}_1^{(2)} = (0, 0)^\mathsf{T}, \ \boldsymbol{\mu}_2^{(2)} = (d, 0)^\mathsf{T}, \quad \boldsymbol{\Sigma}_1^{(2)} = \boldsymbol{\Sigma}_2^{(2)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{20}$$

- dgpU with $K = 2$: The main problem for this case is that a uniform distribution is not a function of a scatter parameter. Both uniform components in dgpU have the same volume and, apart from their center, they would produce the same scatter of points. First, we computed

$$\boldsymbol{V}_U = \int \boldsymbol{x}\boldsymbol{x}^\mathsf{T} dU,$$



(a) Plots of $H(\cdot)$

(b) Plots of $T(\cdot)$

Fig. 1. Population score vs within-cluster distance, for a 2-components Gaussian mixture (dgpG) and a 2-components uniform mixture (dgpU). The distance between true components' centers is indicated on the x-axis; y-axis shows $H(\cdot)$ and $T(\cdot)$ scores (top and bottom panels, respectively), when either $K = 1$ or $K = 2$ groups are used to cluster the data.

where U is the distribution function of a random variable X uniformly distributed on the square $[-1, 1]^2$. $\boldsymbol{V}_U$ would be the covariance of such X. The parameter $\boldsymbol{\theta}^{(2)}$ is set as follows

$$\pi_1^{(2)} = \pi_2^{(2)} = 0.5, \quad \boldsymbol{\mu}_1^{(2)} = (0, 0)^\mathsf{T}, \ \boldsymbol{\mu}_2^{(2)} = (d, 0)^\mathsf{T}, \quad \boldsymbol{\Sigma}_1^{(2)} = \boldsymbol{\Sigma}_2^{(2)} = \boldsymbol{V}_U. \tag{21}$$

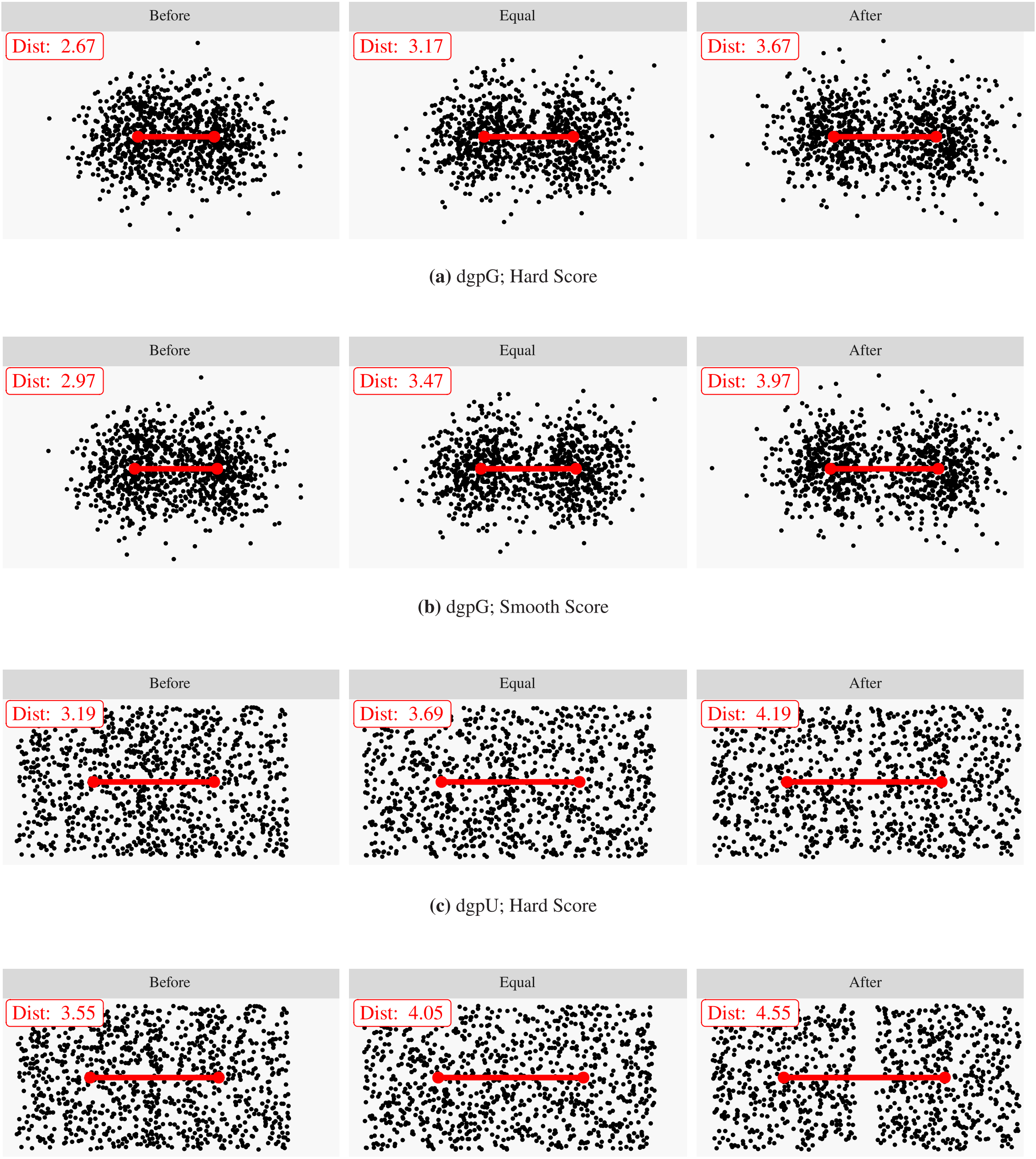
Since some of the previous integrals, including those defining $H(\cdot)$ and $T(\cdot)$, cannot be calculated analytically we computed their approximation (for each value of d) using Monte Carlo integration; all the integrals involved in the example are computed on completely independent experiments with 10^6 random draws. Each integral has been computed 100 times, and the results were averaged to obtain a Monte Carlo standard error consistently below 10^{-5}.

Fig. 1(a) reports $H(\cdot)$ vs., d. For both dgpG and dgpU, the hard score prefers a single cluster for low values of d. The two clusters are split at $d = 3.173$ for dgpG and $d = 3.694$ for dgpU. Fig. 2 shows examples of data produced by the two sampling designs around the point d where $H(\cdot)$ splits a single cluster into two clusters. The general behavior of $H(\cdot)$ is similar for both sampling designs. Under $K = 2$, for both dgpG and dgpU, there is evidence of a non-monotonic behavior of the criterion due to the hard weighting nature of $H(\cdot)$. Taking dgpG with $K = 2$, d only changes the position of the second group, and this is precisely reflected in the definition of $\boldsymbol{\theta}^{(2)}$. We have the same quadratic regions accommodating data points in the same manner. The only difference introduced by d is their location. Therefore, one may expect a monotonic behavior of $H(\cdot)$. However, when d decreases, overlapping the tail regions of the two distributions, both $Q_1(\boldsymbol{\theta}^{(2)})$ and $Q_2(\boldsymbol{\theta}^{(2)})$ start to lose tail points in favor of more central points where $\mathrm{qs}(\cdot)$ is larger. Given the symmetric nature of the setup, for all larger values of d both centers remain at an equal distance from the clusters' boundary. Indeed, notice that the boundaries between $Q_1(\boldsymbol{\theta}^{(2)})$ and $Q_2(\boldsymbol{\theta}^{(2)})$ do not change at changing d, in this particular example. This causes the tendency to split overlapped regions of points that one would not qualify as separate clusters. This may be problematic in cases of strong overlap (as shown later, in Section 4). The behavior of $T(\cdot)$ is reported in Fig. 1(b). For $K = 2$, $T(\cdot)$ is flat for dgpG and almost flat for dgpU. $T(\cdot)$ splits the two groups at slightly larger separation now: $d = 3.47$ for dgpG and $d = 4.05$ for dgpU. $T(\cdot)$ does not attempt to split close clusters and is more appropriate to handle overlapped groups. Scatter plots of data sets around the transition are shown in Fig. 2. Finally, we observe that for all d, the true underlying model corresponds to $K = 2$, but both scores will prefer $K = 1$ for low values of d. The latter implies that the maximum score cannot identify the true underlying distribution even in the Gaussian case.

3. Score selection via resampling

The following discussion applies to both hard and smooth score criteria, therefore we unify the notation. Rewrite both (8) and (12) as the average

$$S_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} s(\boldsymbol{x}_i; \boldsymbol{\theta}), \qquad s(\boldsymbol{x}; \boldsymbol{\theta}) := \sum_{k=1}^{K(\boldsymbol{\theta})} w_k(\boldsymbol{x}; \boldsymbol{\theta})\, \mathrm{qs}(\boldsymbol{x}; \boldsymbol{\theta}), \tag{22}$$



(**a**) dgpG; Hard Score

(**b**) dgpG; Smooth Score

(**c**) dgpU; Hard Score

(**d**) dgpU; Smooth Score

Fig. 2. Scatter plots of data sampled under the dgpG and dgpU sampling designs. In each of the four raw panels (a)–(d), we show the same sample design for three different values of d: the center plot refers to a value of d such that the criterion ($H(\cdot)$ or $T(\cdot)$) values both cases (i.e., $K = 1$ and $K = 2$) equally; the left plot refers to a value of d, where the criterion prefers $K = 1$; the right plot refers to a value of d, where the criterion prefers $K = 2$.

where $s(\boldsymbol{x};\boldsymbol{\theta})$ is the cluster-weighted point-score. With $w_k(\boldsymbol{x};\boldsymbol{\theta}) = \mathbb{I}\{\boldsymbol{x} \in Q_k(\boldsymbol{\theta})\}$ we obtain the hard scoring, while $w_k(\boldsymbol{x};\boldsymbol{\theta}) = \tau_k(\boldsymbol{x};\boldsymbol{\theta})$ returns the smooth score. In Section 2, we assumed a fixed list of candidate configurations, $\mathcal{M}$. In practice, we work with a list of solutions obtained from applying different algorithms (and their various settings) to the only available data set $\mathbb{X}_n$. Let $\hat{\boldsymbol{\theta}}_n = \text{Clust}(\mathbb{X}_n)$ be a cluster configuration obtained by running a specific algorithm on $\mathbb{X}_n$; $\hat{\boldsymbol{\theta}}_n$ reflects the sampling variability, the fitting method's variance and often an error equal to the difference between the method's true solution and its algorithmic approximation. The clustering problem is affected by a mechanism similar to that of the bias–variance trade-off in predictive tasks. Computing both $\hat{\boldsymbol{\theta}}_n$ and $S_n(\cdot)$ using the same observed sample is not ideal because it will lead to an over-optimistic fitting: increasing the solution's complexity (e.g. increasing K) improves the fit on the observed data, but does not necessarily guarantee a more coherent representation of the underlying clustering structure. One way to overcome the previous issue is to make the fitting step independent of the validation step via resampling. We explore two methodologies: cross-validation and bootstrap.

Algorithm 1 *k*-folds cross-validation of quadratic scores (CVQH, CVQS)

```
input: observed sample $\mathbb{X}_n$, clustering method $m \in \mathcal{M}$.
output: $\widetilde{CV}^{(m)}$.

(to ease notation, dependence on $m$ is dropped and reintroduced in step 3.1)

(step 1)   randomly partition $\mathbb{X}_n$ into $k$ folds $\left\{\mathbb{X}^{(t)},\ t \in \{1, \ldots, k\}\right\}$, each with (approximately) $n/k$ data points.

for $t = 1, \ldots, k$ do
    (step 2.1)   $\hat{\boldsymbol{\theta}}^{(t)} \leftarrow \text{Clust}_m(\widehat{\mathbb{X}})$,    where $\widehat{\mathbb{X}} \leftarrow \bigcup_{j \neq t} \mathbb{X}^{(j)}$
    (step 2.2)   $S^{(t)} \leftarrow \frac{1}{\#\mathbb{X}^{(t)}} \sum_{\boldsymbol{y} \in \mathbb{X}^{(t)}} s(\boldsymbol{y}; \hat{\boldsymbol{\theta}}^{(t)})$
end for

(step 3)   Compute:    $\bar{S} \leftarrow \frac{1}{k} \sum_{t=1}^{k} S^{(t)}$;   $\hat{\sigma}_S \leftarrow \frac{1}{k-1} \sum_{t=1}^{k} \left(S^{(t)} - \bar{S}\right)^2$

(step 3.1)   Compute:    $\widetilde{CV}^{(m)} \leftarrow \bar{S} - \delta \, \frac{\hat{\sigma}_S}{\sqrt{k}}$
```

$CVQH = \arg\max_{m \in \mathcal{M}} \left\{\widetilde{CV}^{(m)}\right\}$, when $s(\cdot)$ corresponds to the hard quadratic score

$CVQS = \arg\max_{m \in \mathcal{M}} \left\{\widetilde{CV}^{(m)}\right\}$, when $s(\cdot)$ corresponds to the smooth quadratic score

3.1. *Cross-validation*

Cross-validation (CV) is probably the most popular resampling method to perform model selection by separating the fitting and the testing step. CV has been proposed to estimate K in the MBC framework by [37]. The authors in [20] proposed the CV to select K with the k-means algorithm. The random CV method of [37] produces an estimate of expected Kullback–Leibler information loss under a reference mixture model over an independent test set. Therefore, it is appropriate for tuning the mixture order for density approximation rather than clustering. We consider estimating the expected score (22) via the k-folds CV Algorithm 1. A clustering solution is selected by maximizing $\widetilde{CV}$; this defines the criteria CVQH and CVQS according to s being the hard and smooth scores, respectively. Rather than maximizing the average score criterion $\bar{S}$ computed in step 3, we look at the lower limit of an approximate confidence interval whose size depends on δ. Assuming the approximate normality of $\bar{S}$, $\delta = 1.96$ would determine an approximate 95% confidence interval. Although this may result in crude approximation due to the well-known difficulty to estimate the risk variance via CV [7], it allows to take into account the uncertainty about the estimated mean score, and it is rather popular in applications [see 22]. In the numerical experiments, the selection based on the average criterion $\bar{S}$ led to inferior results compared with the approximate confidence interval rule of step 3. Based on the experimental evidence we suggest $k = 10$ folds and $\delta = 1.96$. The user may tune the constant δ, but in our experiments, it produced relatively better results compared to the more common 1-standard-error rule, that is $\delta = 1$. Additional details about the CV-based selection methods (including the original proposal by [37]) are given in the Supplementary Material, Section S1.

Overall, CV-based methods did not perform well in the following experiments except for some specific cases. The latter is because the application of the CV framework to the clustering task is problematic. CV is designed to estimate the prediction error of a model conditional on the training set, although [4] recently proved that CV does not achieve this goal in general. However, clustering is not a prediction problem. We want to assess how a certain $\boldsymbol{\theta}^{(m)}$ describes the clustered structure produced by the underlying F. Therefore, we need the fitted $\hat{\boldsymbol{\theta}}_n^{(m)}$ and the sample on which the score is computed to convey the same information about the underlying F. The CV aims to estimate a conditional prediction error, requiring that the train and the test set do not overlap, which often causes the two subsamples' structures to differ substantially in finite samples. The latter is the primary motivation for introducing the following bootstrap method.

Algorithm 2 bootstrap scoring (BQH, BQS)

```
input: observed sample $\mathbb{X}_n$ (with ecdf $\mathbb{F}_n$), $\alpha \in (0, 1)$; clustering method $m \in \mathcal{M}$.
output: $\widetilde{W}_n$, $\widetilde{L}_n$, $\widetilde{U}_n$.

(to ease notation, dependence on $m$ is dropped and reintroduced in step 3.1)

for $b \in \{1, \dots, B\}$ do
    (step 1.1)  $\mathbb{X}_n^{*(b)} \leftarrow \left\{x_i^{*(b)};\ i \in \{1, \dots, n\}\right\} \overset{\text{iid}}{\sim} \mathbb{F}_n$
    (step 1.2)  $\hat{\boldsymbol{\theta}}_n^{*(b)} \leftarrow \text{Clust}_m(\mathbb{X}_n^{*(b)})$
    (step 1.3)  $S_n^{*(b)} \leftarrow S_n(\hat{\boldsymbol{\theta}}_n^{*(b)}) = n^{-1} \sum_{i=1}^n s(x_i; \hat{\boldsymbol{\theta}}_n^{*(b)})$
end for
(step 2)  $\widetilde{W}_n \leftarrow \frac{1}{B} \sum_{b=1}^B S_n^{*(b)}$
(step 3)  Let $R_n^{*(b)} = \sqrt{n}\left(S_n^{*(b)} - \widetilde{W}_n\right)$
(step 3.1)  Compute

    $$\widetilde{L}_n^{(m)} \leftarrow \inf_t \left\{t : \frac{1}{B} \sum_{b=1}^B \mathbb{I}\left\{R_n^{*(b)} \le t\right\} \ge \frac{\alpha}{2}\right\}; \quad \widetilde{U}_n^{(m)} \leftarrow \inf_t \left\{t : \frac{1}{B} \sum_{b=1}^B \mathbb{I}\left\{R_n^{*(b)} \le t\right\} \ge 1 - \frac{\alpha}{2}\right\}$$

$BQH = \arg\max_{m \in \mathcal{M}} \left\{\widetilde{L}_n\right\}$ when $s(\cdot)$ corresponds to the hard quadratic score
$BQS = \arg\max_{m \in \mathcal{M}} \left\{\widetilde{L}_n\right\}$ when $s(\cdot)$ corresponds to the smooth quadratic score
```

3.2. Bootstrap

Assume that $\hat{\boldsymbol{\theta}}_n \sim G$, where G reflects the randomness of the clustering output. Assuming that $\hat{\boldsymbol{\theta}}_n$ is independent of X, we want to construct a selection criterion that, at the population level, targets the quantity $W = \mathrm{E}_G[\mathrm{E}_F[s(X; \hat{\boldsymbol{\theta}}_n)]]$. W is the expectation over all possible realization of $\hat{\boldsymbol{\theta}}_n$ of the expected cluster-weighted point score (22). This approach is inspired by the seminal work of [1] on model selection. In practical situations, G is not available, but the variations induced by $\hat{\boldsymbol{\theta}}_n$ can be reproduced by repeating the clustering step on resampled versions of the data. Let $\mathbb{F}_n$ be the ecdf of the sample; we propose to approximate W using multiple independent samples obtained from $\mathbb{F}_n$. The proposed estimation procedure is described in Algorithm 2, and it is based on the classical Efron's non-parametric bootstrap idea. In steps (1.1)–(1.2) of Algorithm 2, independent bootstrap samples from the original data are used to reproduce the variations of $\hat{\boldsymbol{\theta}}_n$. In step (1.3), the original sample is used to compute the empirical approximation of the inner expectation of W at the specific $\hat{\boldsymbol{\theta}}_n^{*(b)}$. Step (2) of Algorithm 2 computes an estimate $\widetilde{W}_n$ of W obtained as the expectation of the Monte Carlo approximation of the bootstrap distribution of $S_n^{*(b)}$. Step (3) corresponds to the percentile method calculation of an approximate $(1-\alpha)$–confidence interval for W. Calculation of a confidence interval for W can be used to consider the uncertainty about W, reflecting both the sample variations and the variance of $\hat{\boldsymbol{\theta}}_n$. Let $W^{(m)}$ be the value of the expected score, W, produced by the mth method/algorithm under comparison. Rather than selecting the cluster configurations achieving the largest estimated $\widetilde{W}^{(m)}$, we propose to select the clustering corresponding to $\hat{\boldsymbol{\theta}}_n^{(m^*)}$ where, for a fixed level of $\alpha \in (0, 1)$,

$$m^* = \underset{m \in \mathcal{M}}{\arg\max}\ \widetilde{L}_n^{(m)}; \tag{23}$$

this defines the BQH and BQS criteria when s is the hard and smooth score, respectively. In principle, one should fix B large enough. The main drawback of Algorithm 2 is that it requires refitting the clusters B times for each clustering configuration $m \in \mathcal{M}$. With the k-folds CV, one also needs to refit the clusters k times, but the number of folds k is usually much smaller than the number of required bootstrap data sets B. In the large experiment shown in Section 4, we set $B = 1000$. In the Supplementary Material, Section S5.2, we also provide evidence that even choosing $B = 100$ did not change results substantially.

4. Experimental analysis

In this section, we present an extensive experimental analysis of the selection problem. The complexity of the following setting aims at offering a neutral comparison, where each competing method is expected to perform well in certain scenarios. This is of utmost importance to achieve scientific progress in unsupervised learning, where global theoretical guarantees are rare, and most of the performances are shown via experimental studies [32]. Experiments are conducted on

Table 1
Data sets analyzed in the experimental section. For each, the table shows: the number of sample points (n), the data dimensionality (d), the "true" number of clusters (K), and a short description. For real data sets (top sub-table) the reported K reflects the original classification(s) of data points; for simulated designs (bottom sub-table), K coincides with the number of mixture components of the data generating process.

	Data	n	d	K	Short Description
Real	Iris	150	4	3	Measurements on Iris flowers; two classes show substantial overlap.
	Banknote	200	6	2	Measurements on original and counterfeit bills; the latter class is usually split in more groups due to the high variability of the measurements.
	Olive	572	8	3/9	Olive oils' fatty acids. Features two different classifications; some classes scatters are concentrated on lower dimensional hyperplanes and show substantial overlap.
	Wine	178	13	3	Chemical analysis of wines grown from three different cultivars. High dimensions; balanced classes.
Simulated	Pentagon5	300	2	5	Mixture of highly unbalanced Gaussian distributions; strong pairwise overlap of 4 of the 5 components.
	T52D	300	2	5	Mixture of 5 equal-proportions, well-separated Student-t components.
	T510D	300	10	5	Adds 8 unclustered dimensions to T52D, increasing dimensionality without adding new clustering information.
	Flower2	300	2	5	Mixture of 2 Student-t, 2 uniform and 1 spherical Gaussian; features regions of strong cluster overlaps.
	Uniform	300	2	1	Uniform distribution; many criteria are not able to identify the unclustered case.

both real and simulated data. The latter are analyzed using Monte Carlo replicates, as explained later. Table 1 summarizes the different settings, giving a short description of the challenges that each setting poses for the selection problem. A detailed discussion of the data is given in Supplementary Material, Section S4. In what follows, we describe the general aspects that are applied to all data sets in the experiments.

4.1. Clustering methods and algorithms

For each data set, the set of candidate solutions for the selection problem is obtained by fitting a clustering method, $m \in \mathcal{M}$, to the data. Each member $m \in \mathcal{M}$ is a solution obtained by an algorithm implementing a clustering method with a set of its specific hyper-parameters. Hyper-parameters are the number of clusters $K \in \{1, \dots, 10\}$; often, restrictions and regularizers for clusters' covariance matrices (whenever possible); algorithmic initialization (for a subset of methods). For each data set we consider $|\mathcal{M}| = 440$ candidate solutions including: K-means and K-medoids partitions; ML for Gaussian mixtures with covariance matrix restrictions (as implemented in mclust software [36]) or eigen-ratio regularization (as implemented in otrimle software [11,12]); ML for Student-t and Skew Student-t mixtures (as implemented in EMMIXskew software [40]). Both Gaussian and Student-t based MBC methods are natural candidates to discover the cluster concept presented in Section 2. On the other hand, we also consider Skewed Student-t models to assess the ability of the selection procedure to tame the overfitting issue usually arising with additional complexity. In fact, the Skewed Student-t family contains both Gaussian and Student-t models as special cases. In what follows, we occasionally refer to subsets of solutions in $\mathcal{M}$, named after the implementing software: K-MEANS, K-MEDOIDS, MCLUST, RIMLE and EMMIX. More details on the clustering methods are given in Supplementary Material, Section S1.1.

4.2. Selection methods

We compare a large number of selection criteria over $\mathcal{M}$. The list of existing criteria is vast. Thus, we restrict the comparison to classical internal validation criteria routinely used by practitioners or those criteria rooted into the MBC literature that are more appropriate for pursuing the cluster notion of interest (for a detailed description see Supplementary Material, Section S1.2).

Method-independent criteria. We consider the Caliński–Harabasz (CH; [10]) and Average Silhouettes Width (ASW; [28]) indexes based on Euclidean distances. While not designed to pursue the cluster's notion investigated in this paper, they are rather popular, and practitioners use them in various settings. The bootstrap stability method of [13], labeled as FW, is introduced in the comparison as another bootstrap-based alternative. It pursues a stability notion rather than the validation philosophy developed here.

Method-dependent criteria. The strongest candidates to discover the cluster concept of interest are information criteria: AIC [1], BIC [35], and ICL [8]). They cannot be computed for members of $\mathcal{M}$ not derived from a probability model, or when the underlying model does not easily map into degrees-of-freedom (e.g. K-MEANS, K-MEDOIDS and RIMLE); this is summarized in Table 2. We also consider the methodology of [37], labeled as CVLK, which minimizes a cross-validated risk based on the data likelihood. CVLK also requires the definition of a models' likelihood function, therefore it can be applied to MBC methods only: MCLUST, RIMLE and EMMIX.

Table 2
Available clustering quality criteria for each type of configuration in $\mathcal{M}$. A tick mark (✓) in a cell indicates that the corresponding quality criterion (column) can be calculated for solutions obtained with a certain class of clustering configurations (row).

Configuration	AIC	BIC	ICL	ASW	CH	FW	CVLK	QH	QS	CVQH	CVQS	BQH	BQS
K-MEANS				✓	✓	✓		✓	✓	✓	✓	✓	✓
K-MEDOIDS				✓	✓	✓		✓	✓	✓	✓	✓	✓
MCLUST	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RIMLE				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EMMIX	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Proposed selection criteria. QH and QS select the clustering solution maximizing (8) and (12), respectively. They exploit in-sample information only, using the observed data both to estimate and score the solution; their results will motivate the need for resampling strategies as discussed above. BQH and BQS methods are the bootstrapped version of the quadratic score method. They correspond to the maximization of (23) using hard and smooth scores, respectively. For both BQH and BQS we set $B = 1000$ for real data; this may be a demanding computing load for large data sets but, in practice, setting a much lower $B = 100$ left the results almost unaltered (see Supplementary Material, Section S5). For the Monte Carlo experiments with simulated data, due to the higher computational load, we set $B = 100$.

4.3. Performance measures

We measure the quality of the selected solutions in terms of: (i) agreement with respect to true clusters' memberships; (ii) the selected number of clusters compared to the ground truth. Point (i) captures similarity between true and fitted groups, and it is measured using the Adjusted Rand Index (ARI) of [26] and the Variation of Information Criterion (VIC) of [33]. ARI $\in [0, 1]$, where ARI = 1 means perfect agreement. Originally, VIC $\in [0, \infty)$; however, we compute and report the negative of the VIC, so that a larger value means better agreement as for the ARI. ARI and VIC are not only different in scales, but they capture the similarity differently. The data sets present different challenges in retrieving the true classes. We design situations where, even for some artificial data, the "true clustering" is not obvious and none of the 440 methods in $\mathcal{M}$ is able to reach near-to-perfect performances (e.g. ARI ≈ 1 and/or VIC ≈ 0). Nevertheless, here we do not compare clustering methods. In contrast, we study the problem of selecting the best available partition. For this reason, besides comparing with the ground truth, we benchmark the 13 selection methods against the "two best feasible partitions", labeled as BEST ARI and BEST VIC. These are obtained running the 440 methods' configurations on a data set and choosing the partitions achieving the best ARI and VIC, respectively. Note that for some data sets, there are multiple members of $\mathcal{M}$ that give the same best feasible partition.

5. Discussion of the results

Table 3 summarizes the results on both real and simulated data, which are discussed in 5.1 and 5.2, respectively. Additional results and comments are given in Supplementary Material, Section S5.

5.1. Results on real data sets

The data sets analyzed in this study are (Table 1, top sub-table): the Iris data set of [3,14]; the Banknote data set of [15]; the Olive data set of [16], for which there are two possible true partitions (a coarser one with 3 classes corresponding to Italian geographical macro-regions, and a finer classification with 9 narrower geographical regions); the Wine data set of [17]. Additional description and visualization is given in the Supplementary Material, Section S4.1. Results presented in this section use $B = 1000$ bootstrap resamples. These are almost unaltered setting a much lower $B = 100$ (see Supplementary Material, Section S5.2).

Fig. 3 provides a graphical representation of the results for the proposed smooth score on the four data sets. Similar displays for the other data sets are shown in the Supplementary Material, Section S5.2, using $B = 100$. For all the clustering methods, there is remarkable evidence that in-sample estimates of the score (QH and QS) become overly optimistic as the complexity of the clustering solutions increases. Indeed, considering the Iris data, for $K > 3$ (true number of groups) and increased model complexity, both QH and QS leave the scores' confidence intervals. Moreover, as soon as $K(m)$ exceeds the true $K = 3$, the more complex members of $\mathcal{M}$ also produce wider confidence bands, confirming the well-known pattern in the model selection that unnecessary additional model complexity introduces additional uncertainty. An analogous pattern is found for the other data sets, although for Olive and Wine the vertical scale dominates the plots. These results are robust to a lower $B = 100$.

Table 3, top sub-table, summarizes the selected solutions for all the clustering selection criteria on the four data sets (details of the selected solutions are shown in Supplementary Material, Section S5.2). First, note that the best feasible partitions available from $\mathcal{M}$ (BEST) do not always retrieve the underlying clusters perfectly, although they catch the true K but for the Olive data with 9 classes. In this case, the best feasible solution corresponds to a configuration fitted by the mclust software with $K = 8$ groups.

Table 3
Summary of the experimental results. Top sub-table: selected solutions for real data; cell value: selected solution's number of groups (K), ARI (A) and negative VIC (V). Bottom sub-table: aggregated results for selected solutions on 100 MC replicates for each simulated design; cell value: most frequently selected number of groups, with frequency in parentheses (K); ARI (A) and negative VIC (V), with standard errors in parentheses. Column BEST: in both tables it refers to the BEST ARI (A) and BEST VIC (V) solutions; when BEST ARI and BEST VIC select different number of groups or with different frequency (bottom sub-table), both solutions are shown separated by a hyphen, reporting BEST ARI first. The best results are highlighted in bold.

		BEST	AIC	BIC	ICL	QH	QS	CH	ASW	FW	CVLK	CVQH	CVQS	BQH	BQS
Iris	K	3	6	2	2	7	7	3	2	2	4	4	4	**3**	**3**
	A	0.94	0.57	0.57	0.57	0.42	0.42	0.73	0.57	0.57	0.81	0.81	0.81	**0.9**	**0.9**
	V	−0.26	−1.52	−0.67	−0.67	−1.56	−1.56	−0.76	−0.67	−0.67	−0.57	−0.57	−0.58	−**0.32**	−**0.32**
Banknote	K	2	6	3	3	10	10	**2**	**2**	2	3	3	3	3	3
	A	1	0.6	0.84	0.84	0.26	0.26	**1**	**1**	0.98	0.85	0.78	0.78	0.86	0.86
	V	0	−1.16	−0.43	−0.43	−2.14	−2.14	**0**	**0**	−0.08	−0.42	−0.62	−0.62	−0.37	−0.37
Olive $(K = 3)$	K	3	10	6	6	10	10	3	2	**2**	10	7	7	8	8
	A	1	0.33	0.52	0.52	0.29	0.29	0.32	0.39	**0.82**	0.3	0.28	0.28	0.49	0.49
	V	−0.03	−1.74	−1.42	−1.42	−1.84	−1.84	−1.88	−1.28	−**0.42**	−1.81	−2.04	−2.04	−1.28	−1.28
Olive $(K = 9)$	K	8	10	6	6	10	10	3	2	2	10	7	7	**8**	**8**
	A	0.88	0.47	0.76	0.76	0.54	0.54	0.42	0.29	0.36	0.58	0.44	0.44	**0.86**	**0.86**
	V	−0.65	−1.77	−1.32	−1.32	−1.26	−1.26	−2.28	−2.28	−1.84	−1.18	−1.96	−1.96	−0.74	−0.74
Wine	K	3	3	3	3	8	8	10	2	3	6	5	5	**3**	**3**
	A	0.98	0.44	0.84	0.84	0.46	0.46	0.15	0.37	0.87	0.59	0.64	0.64	**0.9**	**0.9**
	V	−0.08	−1.42	−0.58	−0.58	−1.53	−1.53	−3.15	−1.41	−0.48	−1.37	−1.11	−1.11	−0.38	−0.38
Pentagon5	K	5 (83%)-4 (52%)	5 (59%)	**5 (45%)**	3 (88%)	10 (31%)	4 (39%)	**3 (100%)**	**3 (100%)**	**3 (99%)**	6 (32%)	4 (42%)	3 (61%)	3 (90%)	**3 (98%)**
	A	0.92 (0.02)	0.82 (0.11)	**0.88 (0.04)**	0.85 (0.03)	0.76 (0.12)	0.84 (0.07)	**0.84 (0.02)**	**0.84 (0.03)**	**0.84 (0.06)**	0.73 (0.13)	0.82 (0.09)	0.84 (0.05)	0.85 (0.03)	**0.84 (0.03)**
	V	−0.36 (0.07)	−0.66 (0.31)	−**0.43 (0.09)**	−0.42 (0.06)	−0.89 (0.36)	−0.57 (0.25)	−**0.43 (0.05)**	−**0.43 (0.06)**	−**0.43 (0.1)**	−0.85 (0.33)	−0.58 (0.25)	−0.48 (0.13)	−0.43 (0.06)	−**0.43 (0.06)**
T52D	K	5 (99%–97%)	6 (21%)	5 (90%)	5 (95%)	10 (41%)	10 (21%)	7 (40%)	5 (86%)	2 (81%)	6 (41%)	4 (43%)	4 (54%)	**5 (98%)**	**5 (98%)**
	A	0.99 (0.01)	0.84 (0.13)	0.97 (0.04)	0.98 (0.01)	0.85 (0.1)	0.91 (0.08)	0.7 (0.1)	0.92 (0.15)	0.59 (0.18)	0.84 (0.13)	0.9 (0.11)	0.93 (0.08)	**0.99 (0.01)**	**0.99 (0.01)**
	V	−0.06 (0.05)	−0.5 (0.32)	−0.12 (0.1)	−0.11 (0.08)	−0.55 (0.26)	−0.35 (0.25)	−0.7 (0.26)	−0.26 (0.32)	−0.93 (0.36)	−0.44 (0.27)	−0.32 (0.23)	−0.27 (0.19)	−**0.08 (0.06)**	−**0.08 (0.06)**
T510D	K	5 (99%)	9 (24%)	6 (50%)	**5 (83%)**	10 (98%)	10 (98%)	2 (100%)	2 (100%)	2 (94%)	6 (45%)	5 (36%)	5 (38%)	5 (69%)	**5 (85%)**
	A	0.99 (0.01)	0.7 (0.13)	0.86 (0.1)	**0.94 (0.08)**	0.55 (0.08)	0.55 (0.07)	0.51 (0.03)	0.51 (0.03)	0.53 (0.11)	0.74 (0.13)	0.79 (0.13)	0.8 (0.13)	0.91 (0.12)	**0.94 (0.09)**
	V	−0.09 (0.06)	−1.06 (0.43)	−0.35 (0.18)	−**0.23 (0.15)**	−1.23 (0.32)	−1.2 (0.23)	−1.1 (0.05)	−1.1 (0.05)	−1.05 (0.21)	−0.75 (0.43)	−0.62 (0.33)	−0.61 (0.34)	−0.28 (0.26)	−**0.2 (0.18)**
Flower2	K	5 (73%–77%)	8 (24%)	2 (58%)	2 (65%)	10 (85%)	10 (59%)	10 (87%)	**5 (85%)**	5 (86%)	7 (43%)	6 (27%)	5 (23%)	**5 (74%)**	**5 (72%)**
	A	0.68 (0.06)	0.48 (0.1)	0.32 (0.1)	0.35 (0.12)	0.47 (0.07)	0.46 (0.08)	0.44 (0.04)	**0.45 (0.17)**	0.45 (0.1)	0.49 (0.11)	0.47 (0.13)	0.43 (0.14)	**0.53 (0.09)**	**0.46 (0.11)**
	V	−1.21 (0.17)	−1.91 (0.34)	−1.88 (0.22)	−1.8 (0.26)	−1.96 (0.21)	−1.88 (0.22)	−1.98 (0.17)	−**1.58 (0.29)**	−1.6 (0.21)	−1.8 (0.27)	−1.73 (0.27)	−1.74 (0.27)	−**1.51 (0.25)**	−**1.58 (0.24)**
Uniform	K	1 (100%)	10 (51%)	4 (65%)	1 (77%)	10 (90%)	10 (71%)	10 (46%)	4 (74%)	4 (64%)	8 (30%)	7 (22%)	1 (85%)	10 (82%)	**1 (96%)**
	A	1 (0)	0 (0)	0 (0)	0.77 (0.42)	0 (0)	0.16 (0.37)	0 (0)	0 (0)	0 (0)	0 (0)	0.06 (0.24)	0.85 (0.35)	0 (0)	**0.96 (0.19)**
	V	0 (0)	−3.01 (0.22)	−1.94 (0.4)	−0.22 (0.45)	−3.1 (0.11)	−2.61 (1.14)	−3.06 (0.41)	−2.1 (0.37)	−2.38 (0.59)	−2.84 (0.28)	−2.3 (0.81)	−0.25 (0.73)	−3.14 (0.31)	−**0.1 (0.54)**

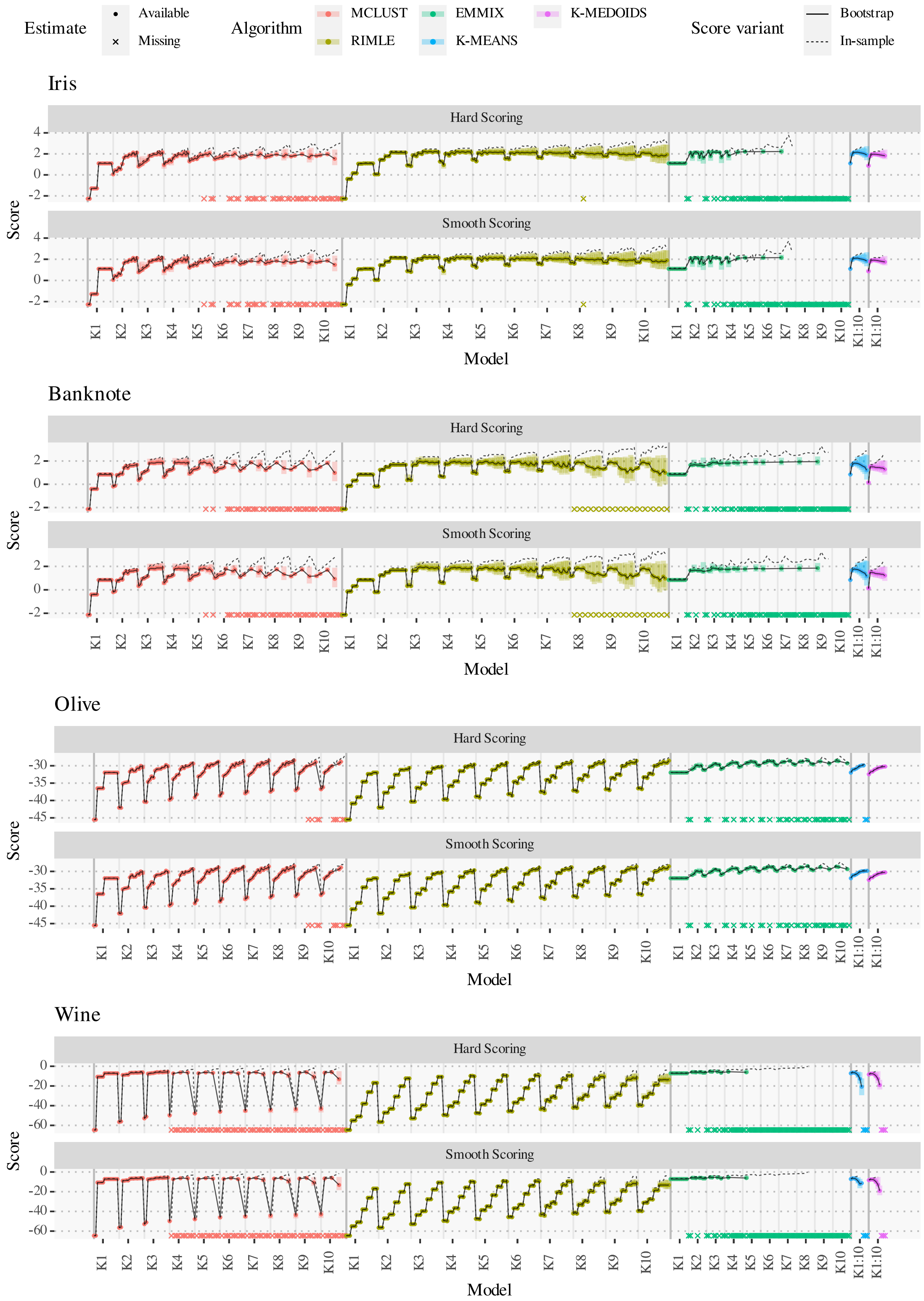


Fig. 3. Results for the Quadratic Smooth score criteria QS and BQS. Horizontal axes: the 440 $m \in \mathcal{M}$ are sorted by: clustering method (colors); increasing K (axis ticks); increasing complexity (fewer restrictions on scatter matrices). Vertical axes: QH and QS (dashed lines); bootstrap estimated $\widetilde{W}_n$ (solid lines) with estimated confidence intervals at 95% (shaded areas). Lower band corresponds to BQH and BQS. Missing solutions are reported with an (×)-symbol (bottom of the plot).

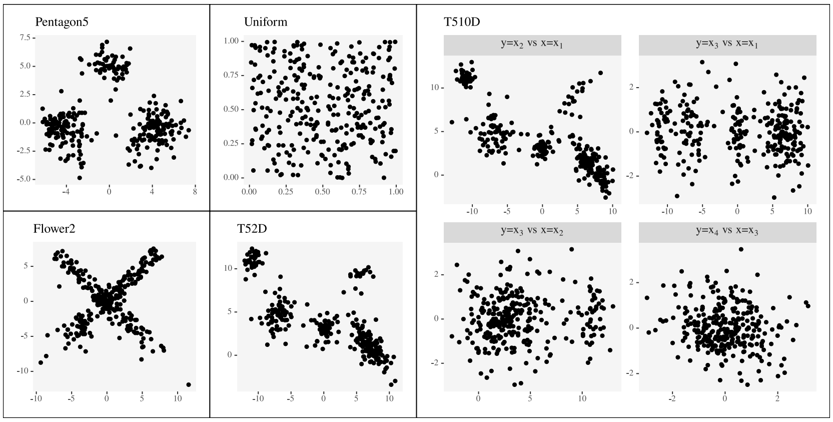


Fig. 4. Scatters produced by the 5 DGPs with $n = 300$ in each case. For T510D (right) we plot the first two marginals (x_1 and x_2), a combination of them with an uninformative marginal (x_3) and two uninformative marginals (x_3 and x_4).

Iris: The selected solutions include partitions with a K ranging from 2 to 6. The true $K = 3$ is detected by BQH, BQS and CH. However, only BQH and BQS selected partition is very close to the best available.

Banknote: The top performers are CH and ASW that discover the true partition exactly, with FW reporting a close performance. In this case, some methods, including BQH, BQS and ICL, provided a second-best performance fitting $K = 3$ groups. This is due to the heterogeneity of the "counterfeit" class, which a single ESD component cannot adequately capture.

Olive: Assuming 3-classes, only CH discovers 3 groups, but these are unrelated to the ground truth; FW reports the best, reasonably good ARI and VIC, with 2 groups mixing some of the underlying 3 classes. Assuming $K = 9$ classes, none of the selection methods discovers 9 groups: BQH and BQS retrieve two partitions that are close to the best available in $\mathcal{M}$, while all other methods select solutions that are far away from the ground truth.

Wine: Within the set of considered methods, it is almost possible to retrieve the true classes exactly. Nonetheless, all the methods show disappointing performances but for BIC, ICL, BQH and BQS. These four criteria select solutions with correct number of classes, but BQH and BQS outperform the other two, achieving better ARI and VIC, close to the optimal ones.

The overall conclusion are: (i) BQH and BQS offer a similar performance, finding the best feasible partition or a partition close to it; (ii) the in-sample versions of the quadratic score criteria, QH and QS, dramatically over-estimate K in all situations; (iii) all cross-validation alternatives showed a poor performance; (iv) information-based criteria showed a mixed evidence. AIC tends to select too complex solutions, while both BIC and the ICL select less complex solutions as expected. BIC and ICL show a similar performance, selecting a reasonable partition in the case of the Banknote and Wine data.

5.2. Monte Carlo experiments

In this section, we present experiments with data simulated from 5 different data generating processes (DGP), shown in Fig. 4. The DGPs are labeled as (Table 1, bottom sub-table): Pentagon5, T52D, T510D, Flower2 and Uniform. All DGPs produce data in dimension $p = 2$ except for T10D, where $p = 10$. The Uniform design generates points drawn from a single 2-dimensional uniform distribution to test the behavior with unclustered data. For all other DGPs, points are drawn from finite mixtures with 5 components. Pentagon5 generates points form Gaussian components, some of which are strongly overlapped and unbalanced. T52D generates points from reasonably separated Student-t components. T510D generates the same clusters as T52D on the first two coordinates while the remaining 8 dimensions are "noisy features" with a joint spherical distribution that does not carry any clustering information. Finally, Flower2 generates points from both uniform and ESD components. A detailed description of the DGPs and additional data visualizations are available in the Supplementary Material, Section S4.2. The "true" cluster membership of a point is identified with the corresponding mixture component generating it. However, some DGPs produce situations that are not always in line with this ground truth definition. For example, one may want to look for 3 clusters in Pentagon5, while 5 groups may not be necessarily the only appropriate description of Flower2's structure. Some DGPs contain substantial departures from elliptic shapes, unbalanced groups and strong between-scatter discrepancies. This is for testing the robustness of the proposed method in situations where the assumptions in Proposition 1 are not exactly fulfilled. Moreover, we fix $n = 300$ for all simulated data sets. The latter choice challenges some resampling criteria due to the strong stress it imposes on bootstrap resampling. Indeed, empirical bootstrap may fail to replicate the distribution of small clusters when n is small.

For each of the 5 simulated designs, we simulate 100 independent data sets from the DGP, and run the model selection experiment on each, in a Monte Carlo (MC) fashion. In this section, we report results for the MC experiments, aggregated for each sample design. Due to the computational complexity of this exercise, we limit the bootstrap replicate to $B = 100$ for all the experiments. Results for all the designs are summarized in Fig. 5, showing boxplots of the Monte Carlo distribution of the ARI and the VIC, and Table 3, bottom sub-table. The ARI and the VIC compare the selected partition to the ground truth previously defined.

Pentagon5: All methods do well. The AIC and the BIC selected 5 groups in roughly 50% of the experiments. This confirms the tendency of such criteria to recover the underlying true DGP rather than the clustering structure. In fact, for this DGP, 3 groups are what one would suggest by visual inspection of the scatter plot in Fig. 4. The other well-performing criteria typically prefer the 3-clusters solution. BQH, BQS, ICL, ASW, CH and FW also excelled for the stability of the results.

T52D: The top performers are BIC, ICL, ASW, BQH and BQS. All of these criteria fit 5 clusters on average, selecting partitions of $\mathcal{M}$ that are close to the best available in the set. This is not surprising given the strong between-cluster separation. BQH and BQS do marginally better, showing the most stable selection. It is worth noting that ASW does well, even if it is not specifically designed to handle DGPs of this type. Whenever clusters are well separated, the intuition is that a distance-based index like ASW can retrieve the true clusters if it uses an appropriate metric.

T510D: The addition of uninformative noisy features in T510D changes the results dramatically: only ICL and BQS maintain excellent performances, with BQS doing slightly better overall in terms of ARI and VIC. In our experiments (see Supplementary Material, Section S5.3) ICL never selects solutions having a number of groups extremely different from that of the ground truth partition, in contrast with BQS, which selects $K > 7$ groups in rare cases. However, it is worth noting that information-type criteria select over a smaller subset of $\mathcal{M}$, not including K-MEANS, K-MEDOIDS and RIMLE solutions, which may produce less variability in the selection.

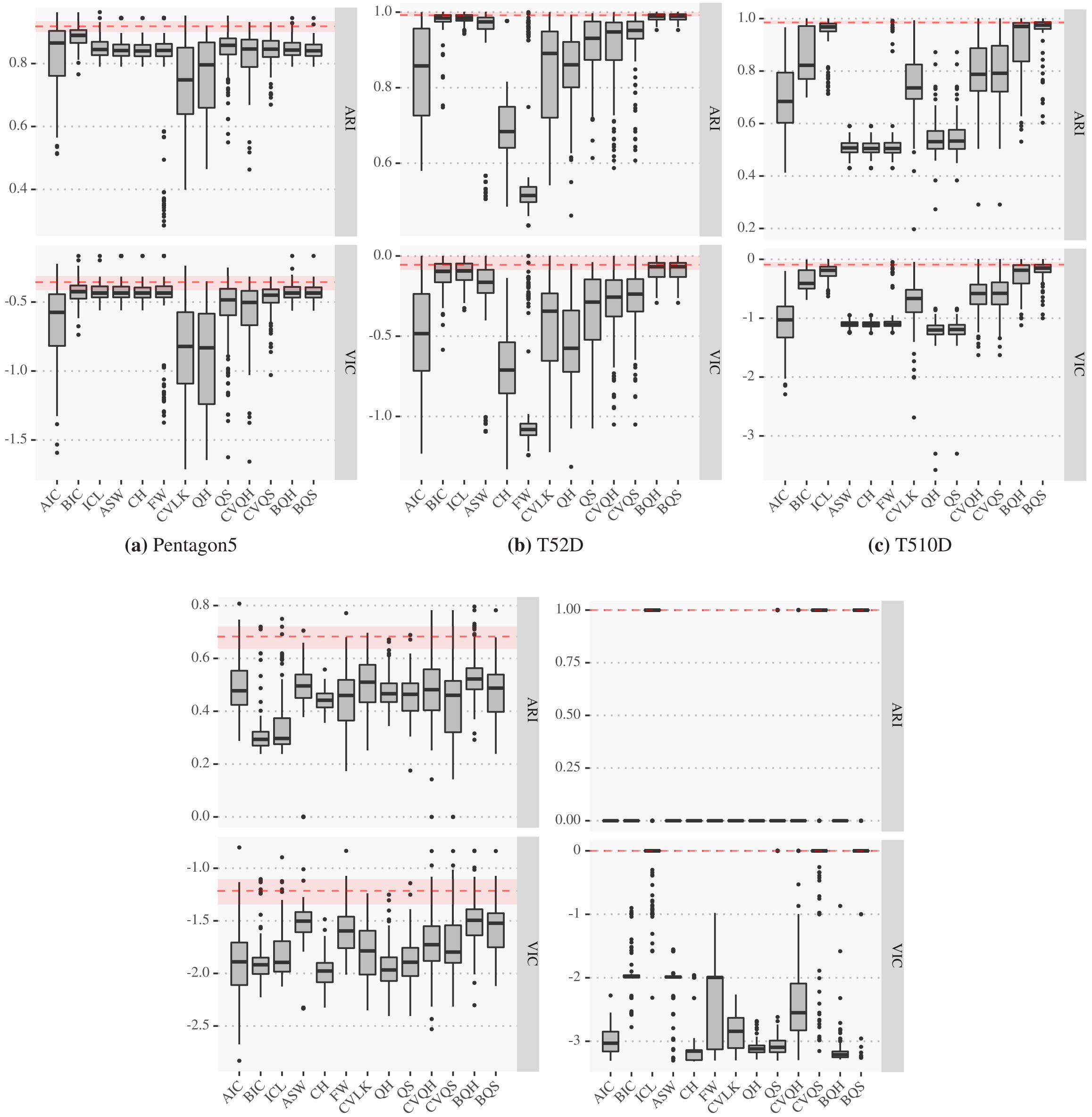
Flower2: This is probably the most challenging case. The best feasible solutions in $\mathcal{M}$ achieve modest levels of average ARI and VIC. A 5 cluster solution achieves the best ARI and VIC roughly 77% of the time, and the methods identifying 5 clusters more often are ASW, FW, BQH, and BQS. The latter two more closely match the frequency with which BEST ARI and BEST VIC selects 5 groups. BQH does only marginally better than its competitors in terms of ARI and VIC, but we can see that the performance of BQS, ASW, and FW are equally good.

Uniform: this sampling design is more of a clear-cut: ICL, BQS, and CVQS are all able to correctly identify no clustering structure. In this case, the clear winner is BQS, selecting a single cluster in 96% of the replicates compared to the 85% of CVQS and 77% of ICL. All the other methods wrongly identify clustering structures in the data (note that FW cannot be directly used to handle the unclustered case). Here, we can see that the AIC looks for the best distribution fit rather than accommodating clustered regions. Indeed, AIC prefers a large number of mixture components to fit the highly unstructured uniform scatter. The BIC mitigates this tendency, but it is not enough. It is also remarkable to see the difference between the top performer, BQS, and its close cousin BQH failing miserably. The explanation of such a bad performance is the tendency of the hard scoring approach to split close groups of points (as shown in Section 2.3). In this case, with a small $n = 300$, the uniform DGP (see Fig. 4) creates many small groups of data points with minimal within-distance, which encourages the hard score to identify many groups.

Overall, experiments on simulated data confirm the analysis on real data. BQS has shown a best or second-best performance in all cases, also yielding better results overall than BQH, proving to be more robust to diverse settings than the latter. ICL is undoubtedly the strongest competitor, although its performance is far from optimal on some occasions. Method-independent criteria like the ASW and the CH, routinely used by practitioners, sometimes completely miss the underlying structure. However, they selected meaningful solutions occasionally, depending on the underlying clustering structure. As already noted for the real data sets, the in-sample estimates QH and QS show a strong selection bias and variance for all data sets. All the methods based on cross-validation exhibit disappointing performances.

6. Conclusions and final remarks

We introduce a unifying framework for treating the problem of cluster selection and validation in the context of clusters generated from elliptic–symmetric families. Within this framework, we propose a novel method for selecting an appropriate clustering for a given data set over a set of candidate partitions (potentially obtained with any clustering method). An extensive comparative experimental study shows that the proposed methodology improves upon popular existing alternatives. In particular, the smooth score criterion with resampling (BQS) consistently provides the best or second-best results in all the considered settings and is thus the authors' advocated criterion. Due to the resampling–refit strategy, the method can be computationally demanding in some circumstances, but this drawback is offset by improved performances and a visualization method that can be used to inspect for unnecessary complexity of the solutions.



(**a**) Pentagon5

(**b**) T52D

(**c**) T510D

(**d**) Flower2

(**e**) Uniform

Fig. 5. Boxplots of Monte Carlo distribution of ARI and VIC performance measures (y-axis; higher value is better), for the considered selection criteria (x-axis). Each sub-plot (a–e) refers to a specific experimental design, and it is further split into two sub-figures representing ARI and VIC performance, respectively. As a benchmark, we report information on ARI and VIC's Monte Carlo distribution for the BEST ARI and BEST VIC solutions, respectively: the red dashed line denotes average values, and the shaded area represents the interval ranging from the first to the third quartile. Overall, the proposed BQH and BQS criteria consistently show close-to-best performance, with less variability across Monte Carlo replicates, as compared to the other criteria. In particular, only the smooth score (CVQS and BQS) and the ICL criteria are able to automatically distinguish the unclustered case (panel e).

CRediT authorship contribution statement

Luca Coraggio: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Pietro Coretto:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration.

Appendix A. Proofs of statements

Proof of Proposition 1. The problem is the analogue of showing the optimality of the Bayes classifier. However, this is conceptually different due to the unsupervised nature of the clustering problem, where a natural notion of loss does not exist. Consider any partition $\{A_k,\ k = 1, \dots, K\}$, then

$$\Pr\left\{\bigcup_{k=1}^{K}\{Z_k = 1 \cap X \in A_k\}\right\} = \sum_{k=1}^{K}\Pr\{Z_k = 1\}\Pr\{X \in A_k \mid Z_k = 1\} = \sum_{k=1}^{K}\int_{A_k}\pi_k f(\boldsymbol{x};\, \boldsymbol{\mu}_k,\, \boldsymbol{\Sigma}_k)d\boldsymbol{x}. \tag{A.1}$$

In order to maximize (A.1) it suffices to choose the partition $\left\{A_k^*,\ k \in \{1, \dots, K\}\right\}$

$$A_k^* = \left\{\boldsymbol{x} \in \mathbb{R}^p :\ \pi_k f(\boldsymbol{x};\, \boldsymbol{\mu}_k,\, \boldsymbol{\Sigma}_k) = \max_{1\le j\le K}\ \pi_j f(\boldsymbol{x};\, \boldsymbol{\mu}_j,\, \boldsymbol{\Sigma}_j)\right\}.$$

Under (C1), $\pi_k f(\boldsymbol{x};\, \boldsymbol{\mu}_k,\, \boldsymbol{\Sigma}_k) = \pi_k\phi(\boldsymbol{x};\, \boldsymbol{\mu}_k,\, \boldsymbol{\Sigma}_k)$, and it is immediate to see that A_k^* coincides with Q_k, proving (4). Denote $\delta_k = (\boldsymbol{x} - \boldsymbol{\mu}_k)^{\mathsf{T}}{\boldsymbol{\Sigma}_k}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)$. Since both $g(t)$ and $\exp(-t/2)$ are monotonically decreasing for $t \in [0, +\infty)$, under (C2), for any $\boldsymbol{x} \in \mathbb{R}^p$,

$$\pi_k f(\boldsymbol{x};\, \boldsymbol{\mu}_k,\, \boldsymbol{\Sigma}_k) \ge \max_{1\le j\le K}\pi_j f(\boldsymbol{x};\, \boldsymbol{\mu}_j,\, \boldsymbol{\Sigma}_j) \iff g(\delta_k) \ge \max_{1\le j\le K}\{g(\delta_j)\} \iff \exp(-\delta_k/2) \ge \max_{1\le j\le K}\{\exp(-\delta_j/2)\}$$
$$\iff \operatorname{qs}(\boldsymbol{x},\, \boldsymbol{\theta}_k^{(m)}) \ge \max_{1\le j\le K}\left\{\operatorname{qs}(\boldsymbol{x},\, \boldsymbol{\theta}_j^{(m)})\right\}.$$

This means that $A_k^* = Q_k(\boldsymbol{\theta}) \in \mathcal{Q}(\boldsymbol{\theta})$, $k \in \{1, \dots, K\}$. □

Proof of Proposition 2. First, note that

$$\operatorname{qs}(\boldsymbol{x},\, \boldsymbol{\theta}_k^{(m)}) = c + \log(\pi_k^{(m)}\phi(\boldsymbol{x};\, \boldsymbol{\mu}_k^{(m)},\, \boldsymbol{\Sigma}_k^{(m)})),$$

where $c = p\log(\sqrt{2\pi})/2$, with π here being the mathematical constant. Since $\sum_{k=1}^{K(\boldsymbol{\theta}^{(m)})}\int_{Q_k(\boldsymbol{\theta}^{(m)})} c\ dF = c$, then

$$H(\boldsymbol{\theta}^{(m)}) = c + \sum_{k=1}^{K(\boldsymbol{\theta}^{(m)})}\int_{Q_k(\boldsymbol{\theta}^{(m)})}\log(\pi_k^{(m)})dF + \sum_{k=1}^{K(\boldsymbol{\theta}^{(m)})}\int_{Q_k(\boldsymbol{\theta}^{(m)})}\log(\phi(\boldsymbol{x};\, \boldsymbol{\mu}_k^{(m)},\, \boldsymbol{\Sigma}_k^{(m)}))dF. \tag{A.2}$$

Using the expression for $L(\boldsymbol{\theta})$ from (9), we can write

$$\sum_{k=1}^{K(\boldsymbol{\theta}^{(m)})}\int_{Q_k(\boldsymbol{\theta}^{(m)})}\log(\pi_k^{(m)})dF = L(\boldsymbol{\theta}^{(m)}) - \sum_{k=1}^{K(\boldsymbol{\theta}^{(m)})}\int_{Q_k(\boldsymbol{\theta}^{(m)})}\log(f(\boldsymbol{x};\, \boldsymbol{\mu}_k^{(m)},\, \boldsymbol{\Sigma}_k^{(m)}))dF.$$

Replace the right-hand side of the previous equation into (A.2) to obtain (10). Under (C3), for any choice of $\boldsymbol{\theta}^{(m)}$ and k,

$$\int_{Q_k(\boldsymbol{\theta}^{(m)})}\log\left(\frac{f(\boldsymbol{x};\, \boldsymbol{\mu}_k^{(m)},\, \boldsymbol{\Sigma}_k^{(m)})}{\phi(\boldsymbol{x};\, \boldsymbol{\mu}_k^{(m)},\, \boldsymbol{\Sigma}_k^{(m)})}\right)dF \ge 0,$$

which proves that $\Lambda(\boldsymbol{\theta}^{(m)}) \ge 0$. □

Proof of Proposition 3. The posterior weights (15) under the Gaussian group-conditional model coincide with the smooth score weights, in fact

$$\omega_{\phi,k}(\boldsymbol{x};\, \boldsymbol{\theta}^{(m)}) = \frac{\pi_k^{(m)}\phi(\boldsymbol{x};\, \boldsymbol{\mu}_k^{(m)},\, \boldsymbol{\Sigma}_k^{(m)})}{\sum_{k=1}^{K(\boldsymbol{\theta}^{(m)})}\pi_k^{(m)}\phi(\boldsymbol{x};\, \boldsymbol{\mu}_k^{(m)},\, \boldsymbol{\Sigma}_k^{(m)})} = \tau_k(\boldsymbol{x};\, \boldsymbol{\theta}^{(m)})$$

for all k. Use the same arguments as in the proof of Proposition 2 and write

$$T(\boldsymbol{\theta}^{(m)}) = c + \sum_{k=1}^{K(\boldsymbol{\theta}^{(m)})}\int\omega_{\phi,k}(\boldsymbol{x};\, \boldsymbol{\theta}^{(m)})\log(\pi_k^{(m)}\phi(\boldsymbol{x};\, \boldsymbol{\mu}_k^{(m)},\, \boldsymbol{\Sigma}_k^{(m)}))dF, \tag{A.3}$$

for an appropriate constant c that does not depend on $\boldsymbol{\theta}^{(m)}$. Since $\sum_{k=1}^{K(\boldsymbol{\theta}^{(m)})}\omega_{\phi,k}(\boldsymbol{x};\, \boldsymbol{\theta}^{(m)}) = 1$, the right-hand-side of (A.3), neglecting the constant term, can be expressed as

$$\sum_{k=1}^{K(\boldsymbol{\theta}^{(m)})}\int\omega_{\phi,k}(\boldsymbol{x};\, \boldsymbol{\theta}^{(m)})\log(\pi_k^{(m)}\phi(\boldsymbol{x};\, \boldsymbol{\mu}_k^{(m)},\, \boldsymbol{\Sigma}_k^{(m)}))dF = A(\boldsymbol{\theta}^{(m)}) - B(\boldsymbol{\theta}^{(m)}), \tag{A.4}$$

where

$$A(\boldsymbol{\theta}^{(m)}) = \int \log(\psi_\phi(\boldsymbol{x}; \boldsymbol{\theta}^{(m)}))dF = \int \log\left(\sum_{k=1}^{K(\boldsymbol{\theta}^{(m)})} \pi_k^{(m)} \phi(\boldsymbol{x}; \boldsymbol{\mu}_k^{(m)}, \boldsymbol{\Sigma}_k^{(m)})\right) dF, \tag{A.5}$$

and

$$B(\boldsymbol{\theta}^{(m)}) = -\sum_{k=1}^{K(\boldsymbol{\theta}^{(m)})} \int \omega_{\phi,k}(\boldsymbol{x}; \boldsymbol{\theta}^{(m)}) \log \omega_{\phi,k}(\boldsymbol{x}; \boldsymbol{\theta}^{(m)})dF. \tag{A.6}$$

The term $A(\boldsymbol{\theta}^{(m)})$ is the expected log-likelihood under the Gaussian mixture model. Since f_0 by assumption is the density of F, then

$$A(\boldsymbol{\theta}^{(m)}) = -\,\mathrm{d}_{\mathrm{KL}}(f_0 \parallel \psi(\cdot; \boldsymbol{\theta}^{(m)})) + \int \log(f_0(\boldsymbol{x}))dF,$$

where the last integral depends only on unknown population objects, and therefore does not depend on $\boldsymbol{\theta}^{(m)}$. (A.6) is the expectation under F of

$$\mathrm{ent}_\phi(Z \mid X; \boldsymbol{\theta}^{(m)}) = -\sum_{k=1}^{K(\boldsymbol{\theta}^{(m)})} \omega_{\phi,k}(X; \boldsymbol{\theta}^{(m)}) \log\left(\omega_{\phi,k}(X; \boldsymbol{\theta}^{(m)})\right).$$

We can now conclude that

$$\begin{aligned} \underset{1\le m\le M}{\arg\max}\; T(\boldsymbol{\theta}^{(m)}) &= \underset{1\le m\le M}{\arg\max}\; A(\boldsymbol{\theta}^{(m)}) - B(\boldsymbol{\theta}^{(m)}), \\ &= \underset{1\le m\le M}{\arg\min}\; \mathrm{d}_{\mathrm{KL}}(f_0 \parallel \psi(\cdot; \boldsymbol{\theta}^{(m)})) + \mathrm{E}_F\left[\mathrm{ent}_\phi(\boldsymbol{\theta}^{(m)})\right]. \end{aligned}$$

The latter proves the desired result (18). □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2023.105181.

References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: Second International Symposium on Information Theory (Tsahkadsor, 1971), Akadémiai Kiadó, Budapest, 1973, pp. 267–281.

[2] S.E. Akhanli, C. Hennig, Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes, Stat. Comput. 30 (5) (2020) 1523–1544, http://dx.doi.org/10.1007/s11222-020-09958-2.

[3] E. Anderson, The species problem in Iris, Ann. Missouri Bot. Gard. 23 (3) (1936) 471–483, https://www.jstor.org/stable/pdf/2394164.pdf?seq=1#page%5Fscan%5Ftab%5Fcontents.

[4] S. Bates, T. Hastie, R. Tibshirani, Cross-validation: What does it estimate and how well does it do it?, 2021, Available from: ArXiv:2104.00673. (Accessed 31 May 2021).

[5] J.-P. Baudry, Estimation and model selection for model-based clustering with the conditional classification likelihood, Electron. J. Stat. 9 (1) (2015) 1041–1077, http://dx.doi.org/10.1214/15-EJS1026.

[6] S. Ben-David, U. von Luxburg, D. Pál, A sober look at clustering stability, in: G. Lugosi, H.U. Simon (Eds.), Learning Theory, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 5–19.

[7] Y. Bengio, Y. Grandvalet, No unbiased estimator of the variance of K-fold cross-validation, J. Mach. Learn. Res. (JMLR) 5 (2003) 1089–1105.

[8] C. Biernacki, G. Celeux, G. Govaert, Assessing a mixture model for clustering with the integrated completed likelihood, IEEE Trans. Pattern Anal. Mach. Intell. 22 (7) (2000) 719–725.

[9] C. Bouveyron, G. Celeux, T.B. Murphy, A.E. Raftery, Model-based clustering and classification for data science, in: Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2019, p. xvii+427, http://dx.doi.org/10.1017/9781108644181, 3967046.

[10] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Comm. Statist. Theory Methods 3 (1) (1974) 1–27.

[11] P. Coretto, C. Hennig, Consistency, breakdown robustness, and algorithms for robust improper maximum likelihood clustering, J. Mach. Learn. Res. 18 (142) (2017) 1–39, http://jmlr.org/papers/v18/16-382.html.

[12] P. Coretto, C. Hennig, OTRIMLE: Robust model-based clustering, 2021, R package version 2.0.

[13] Y. Fang, J. Wang, Selection of the number of clusters via the bootstrap method, Comput. Statist. Data Anal. 56 (3) (2012) 468–477.

[14] R.A. Fisher, The use of multiple measurments in taxonomic problems, Ann. Eugen. (1936) http://rcs.chemometrics.ru/Tutorials/classification/Fisher.pdf.

[15] B. Flury, H. Riedwyl, Multivariate Statistics. A Practical Approach, Chapman & Hall, London, 1988.

[16] M. Forina, C. Armanino, S. Lanteri, E. Tiscornia, Classification of olive oils from their fatty acid composition, Food Res. Data Anal. (1983) 189–214.

[17] M. Forina, R. Leardi, A. C, S. Lanteri, PARVUS: AN Extendable Package of Programs for Data Exploration, Elsevier, Amsterdam, 1988.

[18] C. Fraley, A.E. Raftery, How many clusters? Which clustering method? Answers via model-based cluster analysis, Comput. J. 41 (8) (1998) 578–588.

[19] S. Frühwirth-Schnatter, G. Celeux, C.P. Robert (Eds.), Handbook of mixture analysis, in: Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press, Boca Raton, FL, 2019, p. xxiii+497.

[20] W. Fu, P.O. Perry, Estimating the number of clusters using cross-validation, J. Comput. Graph. Statist. 29 (1) (2020) 162–173, http://dx.doi.org/10.1080/10618600.2019.1647846.

[21] M. Halkidi, M. Vazirgiannis, C. Hennig, Method-independent indices for cluster validation and estimating the number of clusters, in: C. Hennig, M. Meila, F. Murtagh, R. Rocci (Eds.), Handbook of Cluster Analysis, CRC Press, Boca Raton, FL, 2015, pp. 595–618.

[22] T. Hastie, R.J. Tibshirani, J. Friedman, The Elements of Statistical Learning, second ed., Springer New York, 2009, http://dx.doi.org/10.1007/978-0-387-84858-7.

[23] T.J. Hastie, M. Zhu, Discussion of dimension reduction and visualization in discriminant analysis (with discussion), by Cook and yin, Aust. N. Z. J. Stat. 43 (2) (2001) 147–199, http://dx.doi.org/10.1111/1467-842x.00164.

[24] C. Hennig, Cluster-wise assessment of cluster stability, Comput. Statist. Data Anal. 52 (1) (2007) 258–271, http://dx.doi.org/10.1016/j.csda.2006.11.025, https://www.sciencedirect.com/science/article/pii/S0167947306004622.

[25] C. Hennig, Clustering strategy and method selection, in: C. Hennig, M. Meila, F. Murtagh, R. Rocci (Eds.), Handbook of Cluster Analysis, CRC Press, Boca Raton, FL, 2015, pp. 703–730.

[26] L. Hubert, P. Arabie, Comparing partitions, J. Classification 2 (1) (1985) 193–218.

[27] L. Kaufman, P.J.R. Rousseeuw, Finding groups in data, in: Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, Inc., New York, 1990.

[28] L. Kaufman, P.J.R. Rousseeuw, Partitioning around medoids (program PAM), in: Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, Inc., New York, 1990, pp. 68–125, http://dx.doi.org/10.1002/9780470316801.ch2.

[29] C. Keribin, Consistent estimate of the order of mixture models, Compte. Rendus de L Acad. Des Sci. Ser. I Math. 326 (2) (1998) 243–248.

[30] U. von Luxburg, R.C. Williamson, I. Guyon, Clustering: Science or art? in: I. Guyon, G. Dror, V. Lemaire, G. Taylor, D. Silver (Eds.), Proceedings of ICML Workshop on Unsupervised and Transfer Learning, in: Proceedings of Machine Learning Research, vol. 27, PMLR, Bellevue, Washington, USA, 2012, pp. 65–79, https://proceedings.mlr.press/v27/luxburg12a.html.

[31] G. McLachlan, D. Peel, Finite mixture models, in: Wiley Series in Probability and Statistics: Applied Probability and Statistics, John Wiley & Sons, Inc., New York, 2000, p. xxii+419, http://dx.doi.org/10.1002/0471721182.

[32] I.V. Mechelen, A.-L. Boulesteix, R. Dangl, N. Dean, I. Guyon, C. Hennig, F. Leisch, D. Steinley, Benchmarking in cluster analysis: A white paper, 2018, Available from: arXiv:1809.10496.

[33] M. Meilă, Comparing clusterings—an information based distance, J. Multivariate Anal. 98 (5) (2007) 873–895, http://dx.doi.org/10.1016/j.jmva.2006.11.013.

[34] T.T. Nguyen, H.D. Nguyen, F. Chamroukhi, G.J. McLachlan, Approximation by finite mixtures of continuous density functions that vanish at infinity, in: L. Liu (Ed.), Cogent Math. Stat. 7 (1) (2020) 1750861, http://dx.doi.org/10.1080/25742558.2020.1750861.

[35] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (2) (1978) 461–464.

[36] L. Scrucca, M. Fop, T.B. Murphy, A.E. Raftery, mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models, The R J. 8 (1) (2016) 205–233, https://journal.r-project.org/archive/2016-1/scrucca-fop-murphy-etal.pdf.

[37] P. Smyth, Model selection for probabilistic clustering using cross-validated likelihood, Stat. Comput. 10 (1) (2000) 63–72, http://dx.doi.org/10.1023/a:1008940618127.

[38] T. Ullmann, C. Hennig, A.-L. Boulesteix, Validation of cluster analysis results on validation data: A systematic framework, WIREs Data Min. Knowl. Discov. 12 (3) (2022) e1444, http://dx.doi.org/10.1002/widm.1444.

[39] S. Velilla, A. Hernández, On the consistency properties of linear and quadratic discriminant analyses, J. Multivariate Anal. 96 (2) (2005) 219–236.

[40] K. Wang, A. Ng, G.J. McLachlan, EMMIXskew: The EM algorithm and skew mixture distributions, 2018, https://CRAN.R-project.org/package=EMMIXskew.