mathematics

MDPI

Article

Asymptotic Results for the Estimation of the Quadratic Score of a Clustering

Luca Coraggio [1,*] and Pietro Coretto [2]

[1] Department of Economics and Statistics, University of Naples Federico II, 80126 Napoli, Italy
[2] Department of Economics and Statistics, University of Salerno (Italy), 84084 Fisciano, Italy; pcoretto@unisa.it
* Correspondence: luca.coraggio@unina.it

Abstract: In cluster analysis one often finds several partitions of a data set using different clustering methods and algorithms set with a variety of hyperparameters and tunings. The number of clusters K is one of the most relevant of such hyperparameters. Cluster selection is the task of choosing the desired partitions. The Bootstrap Quadratic Scoring is a recently introduced method where the cluster selection is performed by optimizing a score attached to a partition that is based on the quadratic discriminant function. Previously, we proposed the estimation of this cluster score via bootstrap resampling and investigated the proposed estimator based on numerical experiments and real data applications. However, that earlier work did not provide theoretical guarantees. In this paper, we fill that gap. We study the asymptotic behavior of the scoring method and show that the proposed estimator converges to well-defined population counterparts.

Keywords: cluster validation; model-selection; method-selection; resampling methods; asymptotic analysis

MSC: 62H30; 62F05

check for updates

Citation: Coraggio, L.; Coretto, P. Asymptotic Results for the Estimation of the Quadratic Score of a Clustering. *Mathematics* **2024**, *12*, 3417. https://doi.org/10.3390/math12213417

Academic Editor: Ionut Florescu

Received: 3 October 2024
Revised: 28 October 2024
Accepted: 29 October 2024
Published: 31 October 2024

1. Introduction

Clustering remains a fundamental challenge in the analysis of complex datasets. Researchers frequently apply multiple algorithms with varying configurations, producing several candidate partitions from which they must choose a final solution. The diversity in partitioning reflects the inherent ambiguity in defining clusters, especially given the unsupervised nature of most clustering problems (see [1]). A long-standing challenge within this context is the selection of the appropriate number of clusters, K, a decision complicated by the fact that many algorithms also require tuning of hyperparameters that affect the granularity of the data structure representation. Even at a fixed K, different hyperparameter choices can lead to different partitions ([2]).

Recent efforts, such as those of Ullmann et al. [3], have sought to categorize different validation approaches providing a complete overview of what has been done in the literature. For a comprehensive overview of the problem of cluster validation and selection see [4]. The central issue is that most clustering methods implicitly pursue particular notions of what constitutes a "good" cluster, reflecting assumptions about the structure within the data. Despite this, many new proposals claim universal applicability, suggesting that they can uncover the "true" clusters, a claim that oversimplifies the reality of unsupervised analysis. In reality, the existence of "true" groups is often an illusory concept and different methods prioritize different cluster characteristics. In a previous work [2], we introduced a novel approach that starts from a different perspective. There, we introduce the notion of clustering and we propose a method that is designed to retrieve partitions that are consistent with the target cluster concept. In particular, we design criteria, called *quadratic scores* (reviewed in Section 2), that are consistent with clusters generated from the class of elliptical-symmetric distributions that includes Gaussian models as a special case. Clusters of this type form partitions of the data space where the boundaries separating clusters can be meaningfully obtained based on the quadratic discriminant scores used in classical Quadratic Discriminant Analysis (QDA). By connecting these quadratic scores criteria to likelihood-type quantities from the model-based clustering literature (see [5]), we provided a robust framework for selecting the optimal clustering solution.

In [2], we estimate the quadratic score of clustering solutions, using sample information, via an estimation strategy based on bootstrap resampling, a novel proposal that is validated with an extensive experimental analysis. In this article, we extend the existing literature by providing a novel theoretical characterization of this cluster selection methodology. In particular, we derive asymptotic results showing the consistency of the estimation procedure. At the same time, it is acknowledged that our consistency results are based on sufficient conditions that involve assumptions on the unobservable underlying population distributions. Therefore, the results of this paper should be taken as a characterization of the statistical environment that would provide theoretical guarantees for a reasonable cluster selection.

The rest of this paper is organized as follows: Section 2 reviews the hard and smooth quadratic scores introduced in [2], as well as the bootstrap estimation strategy. Section 3 develops the novel asymptotic results, by first characterizing the asymptotic behavior of the bootstrap estimation strategy for a generic scoring function, and then, specializing the results to the hard and smooth scores. Section 4 provides a discussion of the asymptotic results and, finally, Section 5 concludes the paper.

2. Evaluating Partitions Using the Quadratic Scoring

In this section, we review the scoring method introduced in [2]. The general notation is as follows. $\mathbb{X}_n = \{\boldsymbol{x}_i, i \in \{1,\dots,n\}\}$ indicates the observed sample of size n, where each observation is a p-dimensional feature vector $\boldsymbol{x}_i \in \mathbb{R}_p$; $\mathbb{X}_n$ is the realization of a random sample, i.e., $\mathbb{X}_n = \{X_i, i \in \{1,\dots,n\}\}$, where $X_i \in \mathbb{R}^p$ is the p-dimensional random vector of features representing the i-th unit. K indicates the number of groups into which the n observations are clustered. Group memberships are introduced through the random vector of 0–1 variables $Z_i = (Z_{i1}, \dots Z_{iK})^\top$, where $Z_{ik} = 1$ if the i-th sample point belongs to the k-th group, and 0 otherwise. For clusters that are meaningfully described by a triplet, $\boldsymbol{\theta}_k$, of size, center, and scatter parameters $\boldsymbol{\theta}_k = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$, the quadratic score at point x for cluster k is defined by

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

where $\pi_k \in (0,1)$, $\sum_k \pi_k = 1$ is the size of the cluster, and $\boldsymbol{\mu}_k \in \mathbb{R}^p$, $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p \times p}$ are the cluster center and scatter, respectively. The quadratic score qs can be seen as a measure of how well point $\boldsymbol{x}$ fits into cluster k (higher values being associated with better fit) and, given clustering $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k, k \in \{1,\dots,K\}\}$, it defines a quadratic partition of the space

$$\mathcal{Q}(\boldsymbol{\theta}) = \{Q_k(\boldsymbol{\theta}), k \in \{1,\dots,K\}\}, \tag{2}$$

where the quadratic region $Q_k(\boldsymbol{\theta})$ is the region of the space where the quadratic score for cluster k is maximal:

$$Q_k(\boldsymbol{\theta}) = \left\{ \boldsymbol{x} \in \mathbb{R}^p : \mathrm{qs}(x, \boldsymbol{\theta}_k) = \max_{1 \le j \le K} \mathrm{qs}(x, \boldsymbol{\theta}_j) \right\}. \tag{3}$$

Elliptic-symmetric clusters are well described in terms of parameters $\boldsymbol{\theta}_k$, and in [2], it is shown that the quadratic score qs and the associated quadratic partition $\mathcal{Q}$ are optimal to describe a general class of elliptic-symmetric data generating processes (DGPs) in the sense that, under any DGP in this class, parametrized at $\boldsymbol{\theta}$, the quadratic partition achieves the largest probability for points generated from the k-th sub-population to fall within the k-th region $Q_k(\boldsymbol{\theta})$; any other partition of the space is sub-optimal, achieving a lower probability. The aforementioned class of elliptic-symmetric DGPs is characterized as follows.

Definition 1 (Quadratic-clustered DGPs). *DGPs compatible with elliptic-symmetric clustered regions of points that are optimally described by quadratic partitions are so characterized: for* $k \in \{1, \dots, K\}$, $\mathrm{P}(Z_k = 1) = \pi_k$ *and the group-conditional distribution* $X|Z = k$ *has density*

$$f(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} g\Big((\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) \Big), \tag{4}$$

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. Moreover, at least one of the following holds*

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

(C2) $\log(\pi_i / \pi_j) = \log\Big(\det(\boldsymbol{\Sigma}_i)^{\frac{1}{2}} / \det(\boldsymbol{\Sigma}_j)^{\frac{1}{2}}\Big)$.

Remark 1. *The group-conditional density (4) belongs to the key class of elliptical-symmetric distributions (ESDs). The ESD class includes popular models like the Gaussian, the Student-t, the Laplace, the multivariate logistic, etc. They generate groups of points lying in regions that are intersections of ellipsoids described by the pairs of centrality-scatter parameters* $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ *and, within each group, the features are connected via covariance relationships. The generating mechanism is consistent with data generated from finite mixtures of such elliptically-symmetric families. Outside the Gaussian case, DGPs are required to produce groups for which there is a balance between size and generalized variance (see (C1)).*

Most often, there is no prior knowledge on viable grouping of the data, which may also exhibit no clustering structure at all. In practice, it is common to estimate different clustering structures on the data $\mathbb{X}_n$, compare them on some goodness-of-fit measure, and select the best ones. Let $\mathcal{M}$ be a set indexing clustering solutions, with corresponding representation in terms of triplets of parameters given by $\boldsymbol{\theta}^{(m)}$, for index $m \in \mathcal{M}$. Given that the quadratic partition $\mathcal{Q}(\boldsymbol{\theta})$ optimally describes elliptic-symmetric clusters parametrized at $\boldsymbol{\theta}$, it can be shown that point-wise maximization of the quadratic score qs is achieved by those clustering solutions that better capture the main clustered regions, allowing us to define what we call the *Hard Score* criterion:

$$H_n(\boldsymbol{\theta}^{(m)}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K(\boldsymbol{\theta}^{(m)})} \mathbb{I}\Big\{ \boldsymbol{x}_i \in Q_k(\boldsymbol{\theta}^{(m)}) \Big\} \mathrm{qs}\Big(\boldsymbol{x}_i; \boldsymbol{\theta}_k^{(m)} \Big), \tag{5}$$

where $K(\boldsymbol{\theta}^{(m)})$ is the number of clusters in the m-th clustering solution, and $\mathbb{I}\{\cdot\}$ is the indicator function. Comparing $\boldsymbol{\theta}^{(m)}$ with $\boldsymbol{\theta}^{(m')}$, for $m, m' \in \mathcal{M}$, we say that solution $\boldsymbol{\theta}^{(m)}$ is preferred to $\boldsymbol{\theta}^{(m')}$ if $H_n(\boldsymbol{\theta}^{(m)}) > H_n(\boldsymbol{\theta}^{(m')})$, in line with the idea that $\boldsymbol{\theta}^{(m)}$ provides a better description of the elliptic-symmetric clustered regions in the data $\mathbb{X}_n$. $H_n(\cdot)$ attaches hard-weights (0–1 weight) to each point score $\mathrm{qs}(\cdot)$. For situations where a smooth transition between clusters is desired, ref. [2] introduced the *Smooth Quadratic* score, defined as

$$T_n(\boldsymbol{\theta}^{(m)}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K(\boldsymbol{\theta}^{(m)})} \tau_k\Big(\boldsymbol{x}_i; \boldsymbol{\theta}^{(m)} \Big) \mathrm{qs}\Big(\boldsymbol{x}_i; \boldsymbol{\theta}_k^{(m)} \Big). \tag{6}$$

The previous score rests on the same considerations that led to definition (5), but replaces the indicator function with a smooth weighting scheme, using

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

Other weighting schemes are possible, but the softmax transformation is used as it guarantees some form of optimality for Gaussian clusters. Solutions achieving a higher smooth score are preferred. With respect to (5), using (6) tends to select simpler clustering solutions in cases of strong cluster overlaps, preferring solutions that cluster together highly overlapped groups of data points. A full account of the properties of the hard and smooth scores and a comparison between them is provided in [2].

Remark 2. *When clusters are generated by the ESD group-conditional model $f(\cdot)$, by selecting a partition described by a $\boldsymbol{\theta}^{(m)}$ maximizing (5) or (6), it is shown that one finds an optimal quadratic partition. In practice, for real data applications, one cannot expect the previous assumption to hold precisely. In fact, our work [2], based on a massive numerical experiment, showed that the criteria proposed work when: (i) clusters are well-described by size-centrality-scatter parameters $\boldsymbol{\theta}$, and (ii) whenever the clusters can be reasonably separated by linear and quadratic boundaries.*

Bootstrap Resampling of the Scores

In the previous section, we assumed the existence of a set of clustering solutions, indexed by set $\mathcal{M}$, from which an optimal solution needs to be selected, glossing over the origin of the set itself. In fact, such a set needs to be estimated in practice. The usual approach is to identify a pool of clustering approaches, and then define settings and hyperparameters for each of them. A clustering approach, its hyperparameters, and its algorithmic controls define a clustering method fitting the data, thereby producing a clustering solution. With a slight abuse of notation, we use the set $\mathcal{M}$ to index the clustering methods to be fit on the data. For each clustering solution, we denote its representation in terms of triplet parameters as $\hat{\boldsymbol{\theta}}_n^{(m)} = \left\{ \hat{\boldsymbol{\theta}}_{n,k}^{(m)}, k \in \{1, \ldots, K(m)\} \right\}$, where $K(m)$ is the number of clusters implied by method $m \in \mathcal{M}$, and the subscript n highlights the dependence on sample data and its size.

Once the set of clustering solutions is estimated, we can then use criterion (5) or (6) to pick the desired solution. Unfortunately, this strategy is likely to return an over-optimistic assessment of the quality of fit. Estimates of (5) or (6) will tend to favor more complex clustering solutions as these have higher degrees of freedom to adapt to the data better. However, overly complex solutions might capture not only patterns from the unknown data-generating process but also artifacts of the sampling process. In this case, the selection problem is affected by the bias–variance trade-off that arises with clustering methods of different complexity, and can not be decided using the same set of data both to fit the clustering solution and to score it. This fact is well known in the supervised learning context and is well documented in [2].

More formally, let us generalize our notation by calling $s(x, \boldsymbol{\theta})$ the point-wise score, that is the score attached to the point x based on the solution described by $\boldsymbol{\theta}$. Let us rewrite the score attached to the partition as

$$\text{hard score:} \quad s(x,\boldsymbol{\theta}) = \sum_{k=1}^{K(\boldsymbol{\theta})} \mathbb{I}\{x \in Q_k(\boldsymbol{\theta})\}\mathrm{qs}(x,\boldsymbol{\theta}_k), \qquad S_n(\mathbb{X}_n,\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} s(x_i,\boldsymbol{\theta});$$

$$\text{smooth score:} \quad s(x,\boldsymbol{\theta}) = \sum_{k=1}^{K(\boldsymbol{\theta})} \tau_k(x,\boldsymbol{\theta})\mathrm{qs}(x,\boldsymbol{\theta}_k), \qquad S_n(\mathbb{X}_n,\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} s(x_i,\boldsymbol{\theta}).$$

To ease the notation, $S_n(\cdot)$ identifies both the hard and the smooth score depending on the weighting scheme in the point-wise score $s(\cdot)$. When it is important to distinguish the two types of score, $S_n(\cdot)$ will be denoted specifically as $H_n(\cdot)$ (hard type) or $T_n(\cdot)$ (smooth type). Further, fixing method $m \in \mathcal{M}$, let us write as $\hat{\boldsymbol{\theta}}(\mathbb{X}_n) = \hat{\boldsymbol{\theta}}_n$ the clustering solution obtained fitting method m on the sample data $\mathbb{X}_n$. The naive strategy described above that results in over-estimating the performance of solution $\hat{\boldsymbol{\theta}}_n$ consists of computing the score

$$S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}(\mathbb{X}_n)) = \frac{1}{n}\sum_{i=1}^{n} s(x_i, \hat{\boldsymbol{\theta}}(\mathbb{X}_n)), \tag{7}$$

where it is evident that the same data source is used both to estimate the clustering solution and to evaluate the score.

In the spirit of Akaike's seminal work [6], in order to avoid the over-fitting bias described above, one would like to marginalize out the randomness of the clustering solutions. Indeed, assuming that $X \sim F$, and $\hat{\boldsymbol{\theta}}_n \sim G$, for some distribution G that represents the randomness of the clustering output, the target quantity of interest, at the population level, is $W = \mathbb{E}_G\, \mathbb{E}_F[s(X, \hat{\boldsymbol{\theta}}_n)]$. W is the expectation of the average score s over all possible realization of $\hat{\boldsymbol{\theta}}_n \sim G$. However, in large samples, if $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ for $n \to \infty$, W will be evaluated at the limit clustering $\boldsymbol{\theta}_0$.

Estimating W with in-sample quantities would require the observation of multiple independent samples from F and G. As this is not possible in practice, we propose to approximate it using resampling strategies. Here, we focus on the empirical bootstrap ([7]). Let $\mathbb{X}_n$ be the observed data and $\mathbb{F}_n$ be the corresponding empirical cumulative distribution function (ECDF) of the sample. Denoting with $\mathbb{X}_n^{(b)}$ a bootstrap sample from $\mathbb{F}_n$, we propose to use bootstrap resamples to approximate the outer expectation over G, and the distribution $\mathbb{F}_n$ to approximate for the inner expectation over F. That is, we propose to estimate W with the quantity

$$\frac{1}{B}\sum_{b=1}^{B} S\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}(\mathbb{X}_n^{(b)})\right) = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n}\sum_{i=1}^{n} s\left(x_i, \hat{\boldsymbol{\theta}}_n^{(b)}\right), \tag{8}$$

where $\hat{\boldsymbol{\theta}}_n^{(b)}$ represents the clustering solution fitted on the b-th sample (to be precise, the clustering solution is obtained by fitting method $m \in \mathcal{M}$, but we suppress the dependence from m in the notation), and its randomness may arise both due sampling variability and by the algorithmic procedure that estimates the clustering solution. Under appropriate regularity conditions, (8) can be used to estimate W in large samples:

$$\lim_{n,B\to\infty}\frac{1}{B}\sum_{b=1}^{B} S\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^{(b)}\right) = W; \tag{9}$$

in the case of a degenerate G, we are also interested in confidence intervals $\mathbb{E}_F[s(X, \boldsymbol{\theta}_0)]$. The exact procedure introduced in [2] to estimate (9) is described in Algorithm 1.

Algorithm 1 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)  $\mathbb{X}_n^{(b)} \leftarrow \left\{x_i^{(b)}; i \in \{1, \dots, n\}\right\} \overset{iid}{\sim} \mathbb{F}_n$
    (Step 1.2)  $\hat{\boldsymbol{\theta}}^{(b)} \leftarrow$ fit method $m$ on $\mathbb{X}_n^{(b)}$
    (Step 1.3)  $S_n^{(b)} \leftarrow S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}^{(b)}) = n^{-1}\sum_{i=1}^{n} s\left(x_i; \hat{\boldsymbol{\theta}}^{(b)}\right)$
end for
(Step 2)  $\widetilde{W}_n \leftarrow B^{-1}\sum_{b=1}^{B} S_n^{(b)}$
(Step 3)  Let $R_n^{(b)} = \sqrt{n}\left(S_n^{(b)} - \widetilde{W}_n\right)$
(Step 3.1)  Compute

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

$BQH = \arg\max_{m\in\mathcal{M}}\left\{\widetilde{L}_n\right\}$ 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
```

Using Algorithm 1, multiple clustering methods $m \in \mathcal{M}$ can be scored, and the solution achieving the highest bootstrap lower confidence interval is selected. The *BQH* and *BQS* validation criteria are defined according to whether (5) or (6) is used in place of scoring criterion S. Both prove to be effective in solving the selection problem, with the smooth score having an edge over the hard score in cases of strong overlap of (some of) the clusters.

3. Theoretical Analysis

In this section, we analyze the bootstrap estimator presented in Algorithm 1. We use the same notation as in [8] to represent probability measures. In this notation, the original probability measure is denoted as P, and the original probability space is $(\Omega, \mathcal{F}, P)$. This is the probability measure governing the behavior of the data-generating process. The probability measure induced by the bootstrap algorithm is denoted as P^*. To be precise, $P^*_{n,\omega}$ depends upon a realization $\omega \in \Omega$ and on the sample size n; however, we are going to omit this in our notation. Dependency on n will be made explicit in the notation of quantities estimated on bootstrap resamples (e.g., $\hat{\theta}^*_n$). In this notation, a bootstrap statistic T^*_n converges in "probability–P^*, almost sure–P" to T if, for any $\epsilon > 0$, there exists $F \in \mathcal{F}$ such that $P(F) = 1$ and for all $\omega \in F$: $\lim_{n\to\infty} P^*(\|T^*_n - T\| > \epsilon) = 0$. Similarly, T^*_n converges in "probability–P^*, probability–P" to T if, for any $\epsilon > 0$ and $\delta > 0$: $\lim_{n\to\infty} P(P^*(\|T^*_n - T\| > \epsilon) > \delta) = 0$. Using subsequences arguments (e.g., [9]), "convergence in probability–P^*, probability–P" means that for any subsequence $\{n'\}$, there exists a further subsequence $\{n''\}$ where the convergence holds in "probability–P^*, almost sure–P". Convergence of T^*_n to T in distribution–P^*, probability–P means that T^*_n converges weakly to the law of T on a set with probability P converging to 1.

3.1. Analysis of the Resampling Algorithm

In this section, we present the main theoretical results on the bootstrap resampling procedure illustrated in Algorithm 1. We note that, in Algorithm 1, step 3, the quantities $R^{(b)}_n$ are centered with respect to a bootstrap average

$$\widetilde{W}_n = \frac{1}{B}\sum_{b=1}^{B} S_n\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}^{(b)}_n\right) = \frac{1}{B}\sum_{b=1}^{B}\sum_{\boldsymbol{x}_i \in \mathbb{X}_n} s\left(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^{(b)}_n\right), \tag{10}$$

$$R^{(b)}_n = \sqrt{n}\left(S^{(b)}_n - \widetilde{W}_n\right), \tag{11}$$

where we emphasized the dependence on sample size n of the estimated clustering solution and dropped the dependence from a specific method $m \in \mathcal{M}$, which is assumed to be fixed. The scoring function s is a generic one, and in the application above coincides either with the hard or smooth score. The centering used in (11) is different to what is typically done in standard bootstrap theory, where the centering value is the in-sample counterpart of the bootstrapped statistic ([10]).

In order to establish asymptotic properties for (10) and (11), we need two preliminary results. The first establishes the convergence of each term of the sum in (10). The second shows the convergence in the distribution of (11).

Proposition 1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let X be a random variable with distribution F on this space, with $X(\omega) \in \mathbb{R}^p$ for some finite integer p. Let $X_1(\omega), X_2(\omega), \ldots$ be an infinite sequence of independent and identically distributed random variables; $\mathbb{X}_n = \{X_1, \ldots X_n\}$ being the first n terms. Let F_n be the ECDF of $\mathbb{X}_n$, and $\mathbb{X}^*_l$ be a bootstrap sample from $\mathbb{X}_n$, of size l. Let $\Theta \subseteq \mathbb{R}^d$, for some finite integer d; $S_n : \prod_{i=1}^{n} \mathbb{R}^p \times \Theta \to \mathbb{R}$ random functions; $S : \Theta \to \mathbb{R}$ an almost sure continuous random function over Θ. Let $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}(\mathbb{X}_n)$ and $\hat{\boldsymbol{\theta}}^*_l = \hat{\boldsymbol{\theta}}(\mathbb{X}^*_l)$ be fitted clustering solutions, with $\boldsymbol{\theta}_0$, $\hat{\boldsymbol{\theta}}_n$, $\hat{\boldsymbol{\theta}}^*_l \in \mathbb{R}^d$. Assume that:*

(A1) for all $l, n \in \mathbb{N}$, $P(P^*(\hat{\boldsymbol{\theta}}^*_l \in \Theta)) = 1$, $P(\hat{\boldsymbol{\theta}}_n \in \Theta) = 1$ and $P(\boldsymbol{\theta}_0 \in \Theta) = 1$.

(A2) *(Convergence of the estimator in conditional probability) for any $\epsilon > 0$, $\delta > 0$ as $n, l \to \infty$:*

$$\lim_{n\to\infty} P\Big\{\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| > \epsilon\Big\} = 0,$$
$$\lim_{n,l\to\infty} P\Big\{P^*\{\|\hat{\boldsymbol{\theta}}_l^* - \hat{\boldsymbol{\theta}}_n\| > \epsilon\} > \delta\Big\} = 0.$$

(A3) *(Uniform convergence of S_n) for any $\epsilon > 0$:*

$$\lim_{n\to\infty} P\Big\{\sup_{\boldsymbol{\theta}\in\Theta} |S_n(\mathbb{X}_n, \boldsymbol{\theta}) - S(\boldsymbol{\theta})| > \epsilon\Big\} = 0.$$

Then, for any $\epsilon > 0$, $\delta > 0$:

$$\lim_{n,l\to\infty} P\Big\{P^*\{|S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_l^*) - S(\boldsymbol{\theta}_0)| > \epsilon\} > \delta\Big\} = 0, \tag{12}$$

which is a convergence in probability–P^*, probability–P.

Remark 3. *In general, the rate at which l goes to ∞ is determined as a function of n and depends by the particular result applied to show validity of assumption (A2). A typical choice is $l = n$.*

Remark 4. *Assumption (A2) can be replaced by any result stating the convergence of the bootstrapped quantity $\hat{\boldsymbol{\theta}}_l^*$, as for example, for any $\epsilon > 0$, $\delta > 0$,*

$$\lim_{n,l\to\infty} P\{P^*\{\|\hat{\boldsymbol{\theta}}_l^* - \boldsymbol{\theta}\| > \epsilon\} > \delta\} = 0.$$

Now, we move to the second result, showing the convergence in distribution of the quantity (11). This convergence justifies the confidence intervals estimated in Algorithm 1, Step 3.1. The proof requires additional assumptions and uses the delta method applied to the root (11).

Proposition 2. *Let assumptions (A1), (A2), and (A3) be satisfied, and assume for convenience that $l = n$. Additionally, assume that:*

(A4) *(bootstrap estimator's convergence in distribution): for any $\epsilon > 0$:*

$$\lim_{n\to\infty} P\Big\{\sup_t \Big| P^*\{a_n(\hat{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n) \le t\} - P\{a_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \le t\}\Big| > \epsilon\Big\} = 0, \tag{13}$$

for some rate a_n, $a_n \to \infty$ as $n \to \infty$; assuming $a_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges in distribution to a continuous distribution function, $\mathcal{T}$, and call T a random element such that $T \sim \mathcal{T}$.

(A5) *(uniform convergence of the first derivative of S_n over Θ): $S_n(\mathbb{X}_n, \boldsymbol{\theta})$ is twice differentiable in $\boldsymbol{\theta}$ with uniformly converging first derivatives over Θ, in probability–P. That is, assume that for any $\epsilon > 0$:*

$$\lim_{n\to\infty} P\Big\{\sup_{\boldsymbol{\theta}\in\Theta} \|\nabla_{\boldsymbol{\theta}} S_n(\mathbb{X}, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} S(\boldsymbol{\theta})\| > \epsilon\Big\} = 0.$$

Then, as $n \to \infty$, in distribution–P^*, probability–P:

$$a_n\left(S_n\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^{(b)*}\right) - \frac{1}{B}\sum_{b=1}^{B} S_n\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^{(b)*}\right)\right) \xrightarrow{d} \nabla_{\boldsymbol{\theta}} S(\boldsymbol{\theta}_0)^{\mathsf{T}} T + \nabla_{\boldsymbol{\theta}} S(\boldsymbol{\theta}_0)^{\mathsf{T}}\, \mathbb{E}\, T, \tag{14}$$

or, if $\mathbb{E}\,T = 0$,

$$a_n\left(S_n\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^{(b)*}\right) - \frac{1}{B}\sum_{b=1}^{B} S_n\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^{(b)*}\right)\right) \xrightarrow{d} \nabla_{\boldsymbol{\theta}} S(\boldsymbol{\theta}_0)^{\mathsf{T}} T. \tag{15}$$

3.2. Properties of the Quadratic Scoring Function

Proposition 1 and Proposition 2 provide conditions that characterize the asymptotic behavior of the resampling strategy proposed in Algorithm 1, for a generic function S. We now focus on providing results for the hard and smooth scoring functions (5) and (6), which we will use later to show that some of the assumptions (A1) to (A5) hold for this particular choice of S.

The following results for the hard (5) and smooth (6) scores are based on an equovalent formulation in terms of Gaussian densities. The next result establishes the equivalence with Gaussian densities up to a constant term. Proofs of statements are deferred to Appendix A.

Proposition 3. *Consider $m,\ m' \in \mathcal{M}$, and observed data $\mathbb{X}_n$; define the scores h_n and t_n as*

$$h_n\left(\boldsymbol{\theta}^{(m)}\right) := \frac{1}{n}\sum_{x_i \in \mathbb{X}_n}\sum_{k=1}^{K(m)} \mathbb{I}\left\{k = \arg\max_k\left\{\pi_k^{(m)}\phi_k(x_i;m)\right\}\right\}\log\left(\pi_k^{(m)}\phi_k(x_i;m)\right) \tag{16}$$

$$t_n\left(\boldsymbol{\theta}^{(m)}\right) := \frac{1}{n}\sum_{x_i \in \mathbb{X}_n}\sum_{k=1}^{K(m)} \frac{\pi_k^{(m)}\phi_k(x_i;m)}{\sum_{k=1}^{K(m)}\pi_k^{(m)}\phi_k(x_i;m)}\log\left(\pi_k^{(m)}\phi_k(x;m)\right), \tag{17}$$

where $\phi_k(\cdot, m)$ is a p-dimensional Gaussian density parametrized at $\boldsymbol{\theta}^{(m)}$:

$$\phi_k(x,m) = (2\pi)^{-p/2}\det\left(\boldsymbol{\Sigma}_k^{(m)}\right)^{-1/2}\exp\left\{-\frac{1}{2}\left(x-\boldsymbol{\mu}_k^{(m)}\right)^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}\left(x-\boldsymbol{\mu}_k^{(m)}\right)\right\}.$$

Then, referring to (5) *and* (6), *for any $m,\ m' \in \mathcal{M}$,*

$$H_n\left(\boldsymbol{\theta}^{(m)}\right) > H_n\left(\boldsymbol{\theta}^{(m')}\right) \iff h_n\left(\boldsymbol{\theta}^{(m)}\right) > h_n\left(\boldsymbol{\theta}^{(m')}\right)$$
$$T_n\left(\boldsymbol{\theta}^{(m)}\right) > T_n\left(\boldsymbol{\theta}^{(m')}\right) \iff t_n\left(\boldsymbol{\theta}^{(m)}\right) > t_n\left(\boldsymbol{\theta}^{(m')}\right).$$

With a slight abuse of notation, in the following, we will indicate with h and t (without subscript) the generic term of the sums in in h_n and t_n, respectively, (compare with (16) and (17)).

$$h(\boldsymbol{x},\boldsymbol{\theta}) = \sum_{k=1}^{K}\mathbb{I}\left\{k = \arg\max_k\{\pi_k\phi_k(\boldsymbol{x})\}\right\}\log(\pi_k\phi_k(\boldsymbol{x})), \tag{18}$$

$$t(\boldsymbol{x},\boldsymbol{\theta}) = \sum_{k=1}^{K}\frac{\pi_k\phi_k(\boldsymbol{x})}{\sum_{k=1}^{K}\pi_k\phi_k(\boldsymbol{x})}\log(\pi_k\phi_k(\boldsymbol{x})). \tag{19}$$

The following propositions characterize some properties of h_n and t_n.

Proposition 4 (continuity)**.** *$h(x,\theta)$ and $t(x,\theta)$ defined in* (18) *and* (19) *are continuous both in $\boldsymbol{x}$ and $\boldsymbol{\theta}$.*

The proof is trivial, h and t being composition of continuous functions in both argument. It follows immediately that both h_n and t_n are continuous.

Proposition 5 (Bounded from above)**.** *If $\pi_k \in (0,1)$ and $\det\boldsymbol{\Sigma}_k > 0$ for all $k \in \{1,\dots,K\}$, then $h(\boldsymbol{x},\boldsymbol{\theta})$ and $t(\boldsymbol{x},\boldsymbol{\theta})$ are bounded from above: $\exists M \in \mathbb{R} : h(\boldsymbol{x},\boldsymbol{\theta}) < M$ for any $\boldsymbol{x} \in \mathbb{R}^p$ (analogously for t).*

Now, we show the existence of the second moment of h and t, with respect to random variable X, for all possible values of $\boldsymbol{\theta} \in \Theta$. These are needed to establish uniform convergence required by assumption (A3), on which Propositions 1 and 2 heavily rely. This is an essential regularity condition, and amounts to shape the degree of smoothness required for the scoring function used in the resampling scheme.

Proposition 6 (Second moment of t)**.** *Assume the following hold, for every $\boldsymbol{\theta} \in \Theta$*

(B1) *X is a p-valued random variable with $X \sim F$, where F is a continuous distribution function such that the fourth moment exists: $\mathbb{E}(X_i X_j X_l X_m) < \infty$, for any $i, j, l, m \in \{1, \dots, p\}$;*

(B2) *$\pi_k > 0$ for every k and $\sum_{k=1}^{K} \pi_k = 1$;*

(B3) *$\|\boldsymbol{\mu}_k\|_2 \le M$ for some large M for every k;*

(B4) *$\boldsymbol{\Sigma}_k$ is non singular for every k.*

Then, the first two moments of t with respect to F exist and are finite:

$$\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \Big(t(X, \boldsymbol{\theta})^2 \Big) < \infty.$$

Proposition 7 (Second moment of h)**.** *Assume (B1), (B2), (B3), (B4) hold. Then, the first two moments of h under F exist and are finite:*

$$\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \Big(h(X, \boldsymbol{\theta})^2 \Big) < \infty.$$

Proposition 8 (Bounded in probability)**.** *Let (B1) to (B4) hold. Then both $h_m(\boldsymbol{\theta})$ and $t_n(\boldsymbol{\theta})$ are bounded in probability: for any $\epsilon > 0$, $\exists M \in \mathbb{R} : P(|h_n(\boldsymbol{\theta})| > M_\epsilon) \le \epsilon$ (analogously for t_n).*

The next proposition is similar to the two above and it is required to show the validity of assumption (A5), but demands for stronger assumptions on the data distribution. This is needed to ensure regularity conditions to apply delta method in Proposition 2. This is shown for smooth scoring s only.

Proposition 9 (Existence of $\nabla_{\boldsymbol{\theta}} t$ second moment)**.** *Let assumptions (B2) to (B4) hold. Let assumption (B1) be strengthened by the following:*

(B1)* *X is p-dimensional random variable, $X \sim F$, and F is such that*

$$\mathbb{E}\left(X_{i_1} X_{i_2} X_{i_3} X_{i_4} X_{i_5} X_{i_6} X_{i_7} X_{i_8}\right) < \infty; \quad \forall i_1, i_2, i_3, i_4, i_5, i_6, i_7, i_8 \in \{1, \dots, p\}$$

Then:

$$\mathbb{E}\left(\sup_{\boldsymbol{\theta} \in \Theta} \|\nabla_\theta s(X, \boldsymbol{\theta})\|^2 \right) < \infty. \tag{20}$$

3.3. Consistency of the Hard and Smooth Scores

In this section, we bring together results provided in Sections 3.1 and 3.2 to show conditions under which (10) consistently estimates its population counterpart and the validity of confidence intervals based on (11).

Assuming $\mathbb{X}_n$ is a sequence of i.i.d. random variables from F and that (A1), (A2) and (B1) to (B4) hold. These assumptions ensure that $s(X, \boldsymbol{\theta})$ (where s is either h or t) is a continuous function in both arguments, Θ is a compact set, and $\mathbb{E}(\sup_{\boldsymbol{\theta} \in \Theta} s(X, \boldsymbol{\theta})^2) < \infty$ (by Propositions 6 and 7). Then, by a straightforward application of ([11] Theorem 2.7.5), we have that, for any $\epsilon > 0$:

$$P\Big\{ \lim_{n\to\infty} \sup_{\boldsymbol{\theta} \in \Theta} \Big| \frac{1}{n} \sum_{i=1}^{n} s(X_i, \boldsymbol{\theta}) - \mathbb{E}\, s(X, \boldsymbol{\theta}) \Big| > \epsilon \Big\} = P\Big\{ \lim_{n\to\infty} \sup_{\boldsymbol{\theta} \in \Theta} \Big| S_n(\mathbb{X}_n, \boldsymbol{\theta}) - S(\boldsymbol{\theta}) \Big| > \epsilon \Big\} = 0, \tag{21}$$

which is an almost sure uniform convergence of S_n over Θ to S, and is a stronger condition than that required by assumption (A3). Hence, if $\mathbb{X}_n$ is a sequence of i.i.d. random variables from F and (A1), (A2), (B1), (B2), (B3) and (B4) hold, then (A3) is implied by (21), allowing to apply Proposition 1, which yields the convergence in probability–P^*, probability–P for S_n: for any $\epsilon > 0$, $\delta > 0$, and any $b = 1, \ldots, B$:

$$\lim_{n\to\infty} P\left\{ P^*\{|S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^{*(b)}) - S(\boldsymbol{\theta})| > \epsilon\} > \delta \right\} = 0.$$

Thus, applying the law of large numbers to the bootstrap probability–P^*, we have the following:

$$\lim_{B\to\infty} \lim_{n\to\infty} \frac{1}{B} \sum_{b=1}^{B} S_n^{*(b)} \to \mathbb{E}\, S(\boldsymbol{\theta}) \equiv \mathbb{E}_{\boldsymbol{\theta}}\, \mathbb{E}_X\, s(X, \boldsymbol{\theta}), \tag{22}$$

where the convergence happens in probability–P^*, in probability–P. This result states that we can consistently estimate (9) with the bootstrap strategy in Algorithm 1. Next, we show the adequacy of the confidence interval.

Now, we turn to the validity of confidence interval based on (11). The following discussion consider s to be specified as the smooth score (17). Assume that $\mathbb{X}_n$ is a sequence of i.i.d. random variables from F and that (A1), (A2), (A4), (B1*), (B2), (B3), and (B4) hold. Using the argument made to derive (21), these assumptions imply (A3), as (B1*) implies (B1). Moreover, using the same argument, they also imply (A5) by using Proposition 9 in place of Proposition 6. Thus, we can apply Proposition 2 together with ([12] Lemma 23.3), yielding the consistency of the bootstrapped quantiles for the distribution $\nabla_{\boldsymbol{\theta}} S(\boldsymbol{\theta}) T$ (possibly shifted by a bias term). We denote the quantiles with q_α. More precisely, L_n^* and U_n^*, as defined in Algorithm 1, converges to L and U, where:

$$\begin{aligned} L &= \inf_t \left\{ t : P(\nabla_{\boldsymbol{\theta}} S(\boldsymbol{\theta}) T \le t) \ge \frac{\alpha}{2} \right\} + \nabla_{\boldsymbol{\theta}} S(\boldsymbol{\theta})\, \mathbb{E}\, T = q_{\frac{\alpha}{2}} + \text{constant}; \\ U &= \inf_t \left\{ t : P(\nabla_{\theta} S(\theta) T \le t) \ge 1 - \frac{\alpha}{2} \right\} + \nabla_{\theta} S(\theta)\, \mathbb{E}\, T = q_{1-\frac{\alpha}{2}} + \text{constant}; \end{aligned}$$

the convergence occurs as $B \to \infty$, $n \to \infty$, in probability–P^*, in probability–P.

Now, for simplicity, we argue along subsequences n, where the convergence in Proposition 2 can be taken to hold in distribution–P^*, almost sure–P Thus, with probability–$P = 1$, as $B, n \to \infty$:

$$\begin{aligned} &P^*\left\{ S_n(\mathbb{X}_n, \boldsymbol{\theta}) \ge S_n^* - \frac{L_n^*}{a_n} \right\} = P^*\left\{ a_n \left(\frac{1}{B} \sum_{b=1}^{B} S_n^{*(b)} - S_n(\mathbb{X}_n, \boldsymbol{\theta}) \right) \le L_n^* \right\} = \\ &P^*\left\{ a_n \left(\frac{1}{B} \sum_{b=1}^{B} S_n^{*(b)} - S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n) \right) + a_n \left(S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n) - S_n(\mathbb{X}_n, \boldsymbol{\theta}) \right) \le L_n^* \right\} \xrightarrow{p^*} \\ &P\left\{ \left(\nabla_{\boldsymbol{\theta}} S(\boldsymbol{\theta})\, \mathbb{E}\, T + \nabla_{\boldsymbol{\theta}} S(\boldsymbol{\theta}) T \right) \le \nabla_{\boldsymbol{\theta}} S(\boldsymbol{\theta})\, \mathbb{E}\, T + q_{\frac{\alpha}{2}} \right\} = P\left\{ \nabla_{\boldsymbol{\theta}} S(\boldsymbol{\theta}) T \le q_{\frac{\alpha}{2}} \right\} = \frac{\alpha}{2}, \end{aligned}$$

where we used $\xrightarrow{p^*}$ to indicate that the convergence is in probability–P^*, and is motivated by an application of Slutsky's theorem to the bootstrap probability P^*. The same reasoning applies to U_n^*, therefore $S_n(\mathbb{X}_n, \boldsymbol{\theta})$:

$$P^*\left\{ S_n^* - \frac{U_n^*}{a_n} \le S_n(\mathbb{X}_n, \boldsymbol{\theta}) \le S_n^* - \frac{L_n^*}{a_n} \right\} \to 1 - \alpha,$$

where the convergence is in probability–P^*, probability–P (as $B, n \to \infty$).

4. Discussion

In Section 3.1, we analyzed the asymptotic behavior of the bootstrap strategy proposed in Algorithm 1 for a generic scoring function, and characterized assumptions (A1) to (A5) that are required for asymptotic convergence. Then, in Section 3.2, we derived some useful properties for the hard and smooth scores, and defined under which conditions of the data-generating process these hold. Conditions (B1) to (B4), essentially require the existence of moments of the distribution of X, and some restrictions for the set Θ containing the solutions $\boldsymbol{\theta}(m)$, that roughly amount to requiring that the clustering method under study does not output singular clusters, and that none of the $K(m)$ cluster is returned empty. Finally, in Section 3.3, we linked results in Sections 3.1 and 3.2, showing that (i) for the hard and smooth quadratic scoring functions some of the assumptions required by Proposition 1 and Proposition 2 hold under some regularity conditions on F, and reasonable assumptions on the space Θ; and (ii) the bootstrap strategy proposed in Algorithm 1 can consistently estimate the asymptotic targets, by applying Propositions 1 and 2.

Unfortunately, no further insight is available for assumptions (A2) and (A4), although they play a key role in the proof of the two propositions. These assumptions essentially require that the clustering method behave smoothly, ensuring convergence of both in-sample estimates, $\hat{\boldsymbol{\theta}}_n(m)$, and their bootstrap counterparts, $\hat{\boldsymbol{\theta}}^*_n(m)$, to well-defined values, with the latter accurately approximating the former. Andrews [13] shows that, when $\hat{\boldsymbol{\theta}}_n$ is an argmax functional (such as an MLE) and F has sufficient smoothness beyond the second order, these conditions are satisfied. However, verifying such conditions is generally intractable, except in basic cases, which are often of limited interest for clustering analysis.

Even if possible, finding sufficient conditions for Propositions 1 and 2 may still be insufficient in practical clustering scenarios, as in some cases the functional that maps observations into a cluster is hard to frame into a well-understood mathematical object. For example, this is the case for k-means clustering: while $\hat{\boldsymbol{\theta}}_n(m)$ can theoretically be represented as a maximum likelihood estimator [14], in practice, solutions are obtained using heuristic algorithms (like Lloyd's), which adds a further level of complication.

This suggests that experimental analysis might be a practical way to understand the bootstrap behavior of $\hat{\boldsymbol{\theta}}_n(m)$ in applied clustering settings. For instance, O'Hagan et al. [15] show empirically that with roughly balanced clusters, the bootstrap estimated parameters precisely estimate the true parameters for Gaussian mixture models and that bootstrap confidence intervals achieve good coverage, close to the nominal one.

5. Conclusions

In this work, we reviewed the bootstrap quadratic hard (BQH) and smooth (BQS) scores introduced in [2]. These are validation indexes that adapt ideas from Quadratic Discriminant Analysis (QDA), combined with resampling schemes, to the problem of model selection in cluster analysis, and target clustering solutions that optimally represent a well-defined concept of elliptic-symmetric clusters.

In an extensive experimental analysis, using both simulated and real data sets in small sample sizes settings, and considering a wide range of clustering methods, the BQH and BQS criteria proved to achieve state-of-the-art performance, often providing better results than recognized competitors from the literature. However, a formal investigation of their asymptotic behavior was not at that time pursued.

Our contribution in this work is to provide a theoretical characterization of the asymptotic behavior of the bootstrap estimation strategy upon which BQH and BQS are based. We do so by first defining the set of assumptions under which the bootstrap estimation strategy can be expected to produce scores that are asymptotically consistent and then showing that some of these assumptions are verified for the hard and smooth scores. In doing this, we also derive some formal properties for the two quadratic scoring functions and further define conditions for the data-generating process that ensure these hold.

Overall, this allows us not only to characterize the limit quantity estimated by the BQH and BQS indexes but also to define the statistical framework in which these scores can be expected to produce asymptotically consistent results. Unfortunately, the clustering problem is a very complex one, and it is often very difficult to formally assess properties of the clustering method under study against all of the assumptions required for BQH and

BQS consistency. However, in most cases, it is possible to investigate the validity of these hypotheses through experimental studies, and the theoretical characterization we provide can help the researcher orient their or her investigation.

Author Contributions: Conceptualization,L.C. and P.C.; methodology, L.C.; formal analysis, L.C.; investigation, L.C.; writing—original draft preparation, L.C.; writing—review and editing, L.C. and P.C.; project administration, L.C. and P.C. All authors have read and agreed to the published version of the manuscript.

Funding: Luca Coraggio acknowledges financial support from the European Union—NextGenerationEU, in the framework of the GRINS—Growing Resilient, INclusive and Sustainable project (GRINS PE00000018—CUP E63C22002140007). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them. Pietro Coretto acknowledges financial support from the PRIN 2022 grant Methodological and computational issues in large-scale time series models for economics and finance (20223725WE)—CUP: D53D23006100006-M4.C2.1.1.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Proofs

Proof of Proposition 1. First, note that convergence of $\hat{\boldsymbol{\theta}}_l^*$ to $\boldsymbol{\theta}_0$ in conditional probability P^* follows from assumption (A2). Indeed:

$$P\left\{P^*\{\|\hat{\boldsymbol{\theta}}_l^* - \boldsymbol{\theta}_0\| > 2\epsilon\} > \delta\right\} \le P\left\{P^*\{\|\hat{\boldsymbol{\theta}}_l^* - \hat{\boldsymbol{\theta}}_n\| + \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| > 2\epsilon\} > \delta\right\},$$

since the event $\|\hat{\boldsymbol{\theta}}_l^* - \hat{\boldsymbol{\theta}}_n\| + \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| > \epsilon$ contains the event $\|\hat{\boldsymbol{\theta}}_l^* - \boldsymbol{\theta}_0\| > \epsilon$ for any value of ϵ. Now, due to the convergence of $\hat{\boldsymbol{\theta}}_n$ to $\boldsymbol{\theta}_0$, for any subsequence $\{n'\}$, we can find a further subsequence $\{n''\}$ where the convergence happens almost surely. Since the latter is true for any sequence, consider directly the almost sure argument. As n grows, the term $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|$ can be made arbitrarily small, for example, no bigger than ϵ. Hence, for some integer $\bar{n}$, for $n > \bar{n}$, we have:

$$\begin{aligned} P\left\{P^*\{\|\hat{\boldsymbol{\theta}}_l^* - \boldsymbol{\theta}_0\| > 2\epsilon\} > \delta\right\} &\le P\left\{P^*\{\|\hat{\boldsymbol{\theta}}_l^* - \hat{\boldsymbol{\theta}}_n\| + \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| > 2\epsilon\} > \delta\right\} \\ &= P\left\{P^*\{\|\hat{\boldsymbol{\theta}}_l^* - \hat{\boldsymbol{\theta}}_n\| > 2\epsilon - \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|\} > \delta\right\} \le P\left\{P^*\{\|\hat{\boldsymbol{\theta}}_l^* - \hat{\boldsymbol{\theta}}_n\| > \epsilon\} > \delta\right\}. \end{aligned}$$

We note that the term $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|$ is constant when considering the probability P^*. Moreover, the last inequality is justified by the fact that the set $\|\hat{\boldsymbol{\theta}}_l^* - \hat{\boldsymbol{\theta}}_n\| > \epsilon$ contains the set $\|\hat{\boldsymbol{\theta}}_l^* - \hat{\boldsymbol{\theta}}_n\| > \epsilon'$ if $\epsilon < \epsilon'$. The last term of the inequality above goes to 0 by assumption if $n, l \to \infty$. So, we have that, for any $\epsilon > 0$:

$$\lim_{n,l \to \infty} P\left\{P^*\{\|\hat{\boldsymbol{\theta}}_l^* - \boldsymbol{\theta}_0\| > \epsilon\} > \delta\right\} = 0, \tag{A1}$$

which is a convergence in probability–P, probability–P^*.

For the remaining part of the proof, consider the following chain of inequalities:

$$P\Big\{P^*\{|S_n(\mathbb{X}_n,\hat{\boldsymbol{\theta}}_l^*)-S(\boldsymbol{\theta}_0)|>\epsilon\}>\delta\Big\}\leq$$
$$P\Big\{P^*\{|S_n(\mathbb{X}_n,\hat{\boldsymbol{\theta}}_l^*)-S(\hat{\boldsymbol{\theta}_l}^*)|+|S(\hat{\boldsymbol{\theta}}_l^*)-S(\boldsymbol{\theta}_0)|>\epsilon\}>\delta\Big\}\leq$$
$$P\Big\{P^*\{\sup_{\theta\in\Theta}|S_n(\mathbb{X}_n,\boldsymbol{\theta})-S(\boldsymbol{\theta})|+|S(\hat{\boldsymbol{\theta}}_l^*)-S(\boldsymbol{\theta}_0)|>\epsilon\}>\delta\Big\}\quad w.p.1,$$

where the last inequality holds with probability 1 due to assumption (A1), which ensures $\boldsymbol{\theta}_0$, $\hat{\boldsymbol{\theta}}_n$, and $\hat{\boldsymbol{\theta}}_l^*$ are in Θ with probability 1. Now, consider any subsequence $\{n'\}$. For this sequence, there is a further subsequence, for example, $\{n_1''\}$, such that assumption (A3) holds almost sure–P (e.g., [9] Theorem 20.5). This implies that there is an integer $\bar{n}_1'$ such that for any $\epsilon'>0$, $\sup_{\boldsymbol{\theta}\in\Theta}|S_n(\mathbb{X}_n,\boldsymbol{\theta})-S(\boldsymbol{\theta})|\leq\epsilon'$ if $n>\bar{n}_1'$ but for a null set N_2. For the sequence $\{n_1''\}$, it is possible to find a further subsequence, $\{n_2''\}$ (Subsequence $\{n_2''\}$ might possibly coincide with $\{n_1''\}$, and its existence is ensured by the same argument that holds for sequences $\{n'\}$ and $\{n_1''\}$), such that the convergence in (A1) holds in probability–P^*, almost surely–P. That is, there is a null set N_3 such that for $\omega\in\Omega\backslash N_3$, and any $\epsilon>0$: $\lim_{n,l\to\infty}P^*(\|\hat{\boldsymbol{\theta}}_l^*-\boldsymbol{\theta}_0\|>\epsilon)\to 0$. Now, consider the set $N_1=\{\bigcup_l\bigcup_n N_{1,l,n}^*\}\bigcup\{\bigcup_n N_{1,n}'\}\bigcup N_1''$, where $N_{1,l,n}^*$, $N_{1,n}'$ and N_1'' are the null sets where assumption (A1) fails to hold for $\hat{\boldsymbol{\theta}}_l^*$, $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_0$, respectively. Being the countable union of null sets, $N_1\in\mathcal{F}$ and is null. Let N_4 be the null set where $S(\cdot)$ fails to be continuous, and let $N=N_1\bigcup N_2\bigcup N_3\bigcup N_4$. N is the countable union of null sets in $\mathcal{F}$, thus, it is a null set, $N\in\mathcal{F}$ and $P(\Omega\backslash N)=1$. For $\omega\in\Omega\backslash N$, and $n,l>\bar{n}'$, we have:

$$P^*\{\sup_{\boldsymbol{\theta}\in\Theta}|S_n(\mathbb{X}_n,\boldsymbol{\theta})-S(\boldsymbol{\theta})|+|S(\hat{\boldsymbol{\theta}}_l^*)-S(\boldsymbol{\theta}_0)|>\epsilon\}=$$
$$P^*\{|S(\hat{\boldsymbol{\theta}}_l^*)-S(\boldsymbol{\theta}_0)|>\epsilon-\sup_{\boldsymbol{\theta}\in\Theta}|S_n(\mathbb{X}_n,\boldsymbol{\theta})-S(\boldsymbol{\theta})|\}\leq$$
$$P^*\{|S(\hat{\boldsymbol{\theta}}_l^*)-S(\boldsymbol{\theta}_0)|>\epsilon-\epsilon'\}.$$

By the continuity of S and the convergence of $\hat{\boldsymbol{\theta}}_l^*\to\boldsymbol{\theta}_0$ in probability–P^*, almost sure–P as $n,l\to\infty$, the last term goes to 0. Thus, the term $S_n(X_n,\hat{\boldsymbol{\theta}}_l^*)\to S(\boldsymbol{\theta}_0)$ in probability–P^*, almost surely–P for the subsequence $\{n'\}$. Since this argument holds for any subsequence $\{n'\}$, this implies that the convergence $S_n(X_n,\hat{\boldsymbol{\theta}}_l^*)\to S(\boldsymbol{\theta}_0)$ is in probability–P^*, probability–P or equivalently:

$$\lim_{n,l\to\infty}P\Big\{P^*\{|S_n(\mathbb{X}_n,\hat{\boldsymbol{\theta}}_l^*)-S(\boldsymbol{\theta}_0)|>\epsilon\}>\delta\Big\}=0$$

□

Proof of Proposition 2. The requirement in assumption (A4) is equivalent to the following ([12] Ch. 23):

$$\lim_{n\to\infty}P\{a_n(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)\leq t\}=\mathcal{T}(t)$$
$$\lim_{n\to\infty}P\Big\{\Big|P^*\{a_n(\hat{\boldsymbol{\theta}}_n^*-\hat{\boldsymbol{\theta}}_n)\leq t\}-\mathcal{T}(t)\Big|>\epsilon\Big\}=0,$$

for all t and any $\epsilon>0$, for some distribution $\mathcal{T}$. That is $a_n(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)$ converges in distribution to $T\sim\mathcal{T}$ and this can be approximated by the bootstrap distribution of $a_n(\hat{\boldsymbol{\theta}}_n^*-\hat{\boldsymbol{\theta}}_n)$ that converges in distribution–P^*, probability–P to $T\sim\mathcal{T}$.

By arguing along subsequences, we consider the almost sure argument. That is, for any subsequence $\{n'\}$, there is a further subsequence $\{n''\}$ where the convergence in (A4), (A5) are almost sure–P. Then, consider the following:

$$\begin{aligned} a_n\left(S_n\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^{*(b)}\right) - \frac{1}{B}\sum_{b=1}^{B} S_n\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^{*(b)}\right)\right) = \\ a_n\left(S_n\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^{*(b)}\right) - \frac{1}{B}\sum_{b=1}^{B} S_n\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^{*(b)}\right) + S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n) - S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n)\right) = \\ a_n\left(S_n\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^{*(b)}\right) - S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n)\right) - \frac{1}{B}\sum_{b=1}^{B} a_n\left(S_n\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^{*(b)}\right) - S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n)\right). \end{aligned} \quad \text{(A2)}$$

Now, consider the expansion of $S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^*)$ around $\hat{\boldsymbol{\theta}}_n$:

$$S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^*) = S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n) + \nabla_{\boldsymbol{\theta}} S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n)^{\mathsf{T}}(\hat{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n) + (\hat{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n)^{\mathsf{T}} \Delta_{\boldsymbol{\theta}} S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n) + \ldots,$$

where $\Delta_{\boldsymbol{\theta}}$ indicates the matrix of second derivatives of S_n. Rearranging the terms and multiplying by a_n yields:

$$a_n\left(S_n\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^*\right) - S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n)\right) = \nabla_{\boldsymbol{\theta}} S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n)^{\mathsf{T}} a_n(\hat{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n) + o_{P^*}(a_n),$$

where we indicated by o_{P^*} a term that goes to 0 when $n \to \infty$, at a rate a_n in probability–P^*, probability–P, which follows from assumption (A2). Then, as $n \to \infty$, using (A5) and an application of Slustky's theorem:

$$a_n\left(S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^*) - S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n)\right) = \nabla_{\boldsymbol{\theta}} S_n(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n)^{\mathsf{T}} a_n(\hat{\boldsymbol{\theta}}_n^* - \hat{\boldsymbol{\theta}}_n) + o_{P^*}(a_n) \xrightarrow{d} \nabla_{\boldsymbol{\theta}} S(\boldsymbol{\theta}_0)^{\mathsf{T}} T, \quad \text{(A3)}$$

Letting $B \to \infty$, using the result in (A3), together with the strong law of large numbers for i.i.d. random variables, the last term in Equation (A2) goes to $\nabla_{\boldsymbol{\theta}} S(\boldsymbol{\theta})^{\mathsf{T}} \, \mathbb{E}\, T$. Finally, by an application of Slustky's theorem to the bootstrap probability P^* and again (A3), as $B, n \to \infty$:

$$a_n\left(S_n\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^{*(b)}\right) - \frac{1}{B}\sum_{b=1}^{B} S_n\left(\mathbb{X}_n, \hat{\boldsymbol{\theta}}_n^{*(b)}\right)\right) \xrightarrow{d} \nabla_{\boldsymbol{\theta}} S(\boldsymbol{\theta}_0)^{\mathsf{T}} T + \nabla_{\boldsymbol{\theta}} S(\boldsymbol{\theta}_0)^{\mathsf{T}} \, \mathbb{E}\, T$$

in distribution–P^*, almost sure–P.

As a last point, we note that the above argument, which is based on almost sure convergence in P, is valid for any subsequence $\{n'\}$. Thus, for the stated assumptions, the actual convergence is in distribution–P^*, probability–P. □

Proof of Proposition 3. For any p-dimensional point $\boldsymbol{x}$, and triplet $\boldsymbol{\theta}_k$, the quadratic score is defined as

$$\mathrm{qs}(\boldsymbol{x}, \boldsymbol{\theta}_k) = \log(\pi_k) - \frac{1}{2}\log(\det(\boldsymbol{\Sigma}_k)) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^{\mathsf{T}} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k).$$

Let us denote with $\phi_k(x)$ the Gaussian density with parameters $\boldsymbol{\theta}_k$, evaluated at x, that is

$$\phi_k(\boldsymbol{x}) = \phi(\boldsymbol{x}, \boldsymbol{\theta}_k) = (2\pi)^{-p/2} \, \det(\Sigma)^{-1/2} \, \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^{\mathsf{T}} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_K)\right).$$

Thus, it follows

$$\mathrm{qs}(\boldsymbol{x}, \boldsymbol{\theta}_k) = \log(\pi_k \phi_k(\boldsymbol{x})) + c, \quad \text{(A4)}$$

where $c = -(p/2)\log(2\pi)$ is a constant term.

Now, consider the hard score criterion

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

The indicator function weighting for point $\boldsymbol{x}$ is equivalent to

$$\mathbb{I}\{\boldsymbol{x} \in Q_k(\boldsymbol{\theta})\} = 1 \iff k = \underset{j\in\{1,\dots,K\}}{\arg\max}\, \mathrm{qs}(\boldsymbol{x}, \boldsymbol{\theta}_j) \iff k = \underset{j\in\{1,\dots,K\}}{\arg\max}\left\{\pi_j^{(m)}\phi_j(\boldsymbol{x})\right\},$$

where the third implication follows from (A4), c being a constant, and log being a monotonically increasing function. Hence, we can write

$$\begin{aligned} H_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K(\boldsymbol{\theta})} \mathbb{I}\{\boldsymbol{x}_i \in Q_k(\boldsymbol{\theta})\}\mathrm{qs}(\boldsymbol{x}_i, \boldsymbol{\theta}_k) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K(\boldsymbol{\theta})} \mathbb{I}\left\{k = \underset{k}{\arg\max}(\pi_k\phi_k(\boldsymbol{x}))\right\}\log(\pi_k\phi_k(\boldsymbol{x})) + \tilde{c} = h_n(\boldsymbol{\theta}) + c_h, \end{aligned}$$

where c_h is a constant term. Thus, for any two $m,\ m' \in \mathcal{M}$, it follows

$$H_n\left(\boldsymbol{\theta}^{(m)}\right) > H_n\left(\boldsymbol{\theta}^{(m')}\right) \iff h_n\left(\boldsymbol{\theta}^{(m)}\right) > h_n\left(\boldsymbol{\theta}^{(m')}\right).$$

Analogously, for the smooth score

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

consider the equivalence

$$\tau_k(\boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{\exp \mathrm{qs}(\boldsymbol{x}, \boldsymbol{\theta}_k)}{\sum_k \exp \mathrm{qs}(\boldsymbol{x}, \boldsymbol{\theta}_k)} = \frac{\pi_k\phi_k(\boldsymbol{x})e^c}{\sum_k \pi_k\phi_k(\boldsymbol{x})e^c} = \frac{\pi_k\phi_k(\boldsymbol{x})}{\sum_k \pi_k\phi_k(\boldsymbol{x})}.$$

So that the following equivalence holds

$$\begin{aligned} T_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K(\boldsymbol{\theta})} \tau_k(\boldsymbol{x}_i\boldsymbol{\theta})\mathrm{qs}(\boldsymbol{x}_i, \boldsymbol{\theta}_k) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K(\boldsymbol{\theta})} \frac{\pi_k\phi_k(\boldsymbol{x})}{\sum_k \pi_k\phi_k(\boldsymbol{x})}\log(\pi_k\phi_k(\boldsymbol{x})) + c_t, \end{aligned}$$

where c_t is a constant term. Then, for any two $m,\ m' \in \mathcal{M}$

$$T_n\left(\boldsymbol{\theta}^{(m)}\right) > T_n\left(\boldsymbol{\theta}^{(m')}\right) \iff t_n\left(\boldsymbol{\theta}^{(m)}\right) > t_n\left(\boldsymbol{\theta}^{(m')}\right).$$

□

Proof of Proposition 4. $t(\boldsymbol{x}, \boldsymbol{\theta})$ is obviously continuous in both arguments as it is obtained as the product of continuous functions (see (19)).

Consider $s_h(x, \theta)$. Discontinuity points might occur when the indicator function switches from one component to another. Without loss of generality, we treat the case where $K = 2$. Consider the following:

$$h(\boldsymbol{x}, \boldsymbol{\theta}) = \begin{cases} \log(\pi_1\phi_1(x)) \text{ if } 1 = \arg\max_{1,2} \pi_i\phi_i(x) \\ \log(\pi_2\phi_2(x)) \text{ if } 2 = \arg\max_{1,2} \pi_i\phi_i(x) \end{cases} = \max\{\log(\pi_1\phi_1(x)), \log(\pi_2\phi_2(x))\}.$$

Noting that the maximum of continuous functions is itself continuous, the statement follows. □

Proof of Proposition 5. We give the proof for t. The proof for h follows by the same argument, replacing the smooth weight with the indicator weight. For a given number of K, assume without loss of generality that $\pi_k \in (0,1)$, for any $k \in \{1,\ldots,K\}$ (if one of the π_k is equal to 0 we are in a case with $K-1$ components; if one of the π_k is equal to 1 we are in the case of $K=1$). Moreover, $\phi_k(\boldsymbol{x})$ for finite $\boldsymbol{\mu}_k$ and non-singular Σ_k is bounded by 0 from below and by $\phi_k(\boldsymbol{\mu}_k)$ from above. As a consequence:

$$\frac{\pi_k\phi_k(\boldsymbol{x})}{\sum_{k=1}^K \pi_k\phi_k(\boldsymbol{x})} \in (0,1); \quad \sum_{k=1}^K \frac{\pi_k\phi_k(\boldsymbol{x})}{\sum_{k=1}^K \pi_k\phi_k(\boldsymbol{x})} = 1.$$

Consider $\log(\pi_k\phi_k(\boldsymbol{x}))$, this quantity belongs to the interval $(-\infty, \log(\pi_k\phi_k(\boldsymbol{\mu}_k)))$. As a consequence, it is easy to see that:

$$\sum_{k=1}^K \frac{\pi_k\phi_k(\boldsymbol{x})}{\sum_{k=1}^K \pi_k\phi_k(\boldsymbol{x})} \log(\pi_k\phi_k(\boldsymbol{x})) \leq \sum_{k=1}^K \log(\pi_k\phi_k(\boldsymbol{x})) \leq \sum_{k=1}^K \log(\pi_k\phi_k(\boldsymbol{\mu}_k)) \leq \infty.$$

However, it is not bounded from below, as it may happen that as $\|\boldsymbol{x}\| \to \infty$, $\frac{\pi_k\phi_k(\boldsymbol{x})}{\sum_{k=1}^K \pi_k\phi_k(\boldsymbol{x})} > 0$ and $\log(\pi_k\phi_k(\boldsymbol{x})) \to -\infty$. □

Proof of Proposition 6. Consider a partition of $\mathbb{R}^p$, $\{A_k\}_{k=1\ldots K}$, where:

$$A_i := \{\boldsymbol{x} \in \mathbb{R}^p : \log(\pi_i\phi_i(\boldsymbol{x})) \geq \log(\pi_k\phi_k(\boldsymbol{x}))\ \forall k \neq i\}; \tag{A5}$$

Note that due to the continuity of the functions involved, such a partition can always be found for any value of $\boldsymbol{\theta} \in \Theta$. Then:

$$\begin{aligned}
\mathbb{E}\left(t(X,\theta)^2\right) &= \int_{\mathbb{R}^p} \left(\sum_{k=1}^K \frac{\pi_k\phi_k(\boldsymbol{x})}{\sum_{k=1}^K \pi_k\phi_k(\boldsymbol{x})} \log(\pi_k\phi_k(\boldsymbol{x}))\right)^2 dF(\boldsymbol{x}) \leq \\
&\int_{\mathbb{R}^p} \left(\sum_{k=1}^K \log(\pi_k\phi_k(\boldsymbol{x}))\right)^2 dF(\boldsymbol{x}) = \sum_{i=1}^K \int_{A_i} \left(\sum_{k=1}^K \log(\pi_k\phi_k(\boldsymbol{x}))\right)^2 dF(\boldsymbol{x}) \leq \\
&\sum_{i=1}^K \int_{A_i} (K\log(\pi_i\phi_i(\boldsymbol{x})))^2 dF(\boldsymbol{x}) \leq \sum_{i=1}^K \int_{\mathbb{R}^p} (K\log(\pi_i\phi_i(\boldsymbol{x})))^2 dF(\boldsymbol{x}) = \\
&\sum_{i=1}^K \int_{\mathbb{R}^p} K^2 \log(\pi_i)^2 dF(\boldsymbol{x}) + \sum_{i=1}^K \int_{\mathbb{R}^p} K^2 \log(\phi_i(\boldsymbol{x}))^2 dF(\boldsymbol{x}) + \\
&\sum_{i=1}^K \int_{\mathbb{R}^p} K^2 2\log(\pi_i)\log(\phi_i(\boldsymbol{x}))dF(\boldsymbol{x}).
\end{aligned} \tag{A6}$$

Note that the inequality passing back from A_i to $\mathbb{R}^p$ is due to the positiveness of the integrand (which is a squared function). Now, we analyze in turn the finiteness of the last three terms.

Consider the first term. It is constant and clearly finite if and only if we have a finite number of components K and $\pi_i > 0$ for each i, which is assumed in (B2).

The second term can be rewritten more explicitly as:

$$\sum_{i=1}^{K} \int_{\mathbb{R}^p} K^2 \log(\phi_i(\boldsymbol{x}))^2 dF(\boldsymbol{x}) = K^2 \sum_{i=1}^{K} \int_{\mathbb{R}^p} \left(C_i + \frac{-(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(-\boldsymbol{x}-\boldsymbol{\mu}_i)}{2} \right)^2 dF(\boldsymbol{x})$$
$$= K^2 \sum_{i=1}^{K} C_i^2 + \frac{K^2}{4} \sum_{i=1}^{K} \mathbb{E}\left(X'\boldsymbol{\Sigma}_i^{-1}X + \boldsymbol{\mu}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - 2\boldsymbol{\mu}_i'\boldsymbol{\Sigma}_i^{-1}X\right)^2 -$$
$$K^2 \sum_{i=1}^{K} C_i \, \mathbb{E}\left(X'\boldsymbol{\Sigma}_i^{-1}X + \mu_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - 2\boldsymbol{\mu}_i'\boldsymbol{\Sigma}_i^{-1}X\right). \quad \text{(A7)}$$

where: $C_i = \log(2\pi^{-d/2}|\boldsymbol{\Sigma}_i|^{-1/2})$. Consider now the second term of the expansion above, (i) all the terms involving parameters in $\boldsymbol{\theta}$ are finite due to assumptions (B3) and (B4); and (ii) the only term that might trouble the finiteness of this term is the one containing the random variable X. The term $\mathbb{E}(X'\boldsymbol{\Sigma}_i^{-1}X)^2$ involves computing the expected value of linear combinations of the components of X up to the fourth power. Let p be the dimension of X. Define a set of integer vectors $I = \{(i_1, i_2, \dots, i_p) : 0 \le i_j \le 4, \sum_{j=1}^{p} i_j = 4\}$. Denote with I^* the set of unique elements of I, and let i^* denote elements in I^*. Simple algebra shows that we can arrange the previous term as follows:

$$(X'\boldsymbol{\Sigma}_i^{-1}X)^2 = \sum_{i^* \in I^*} \gamma_{i^*} X_1^{i_1} X_2^{i_2} \dots X_d^{i_p}, \quad \text{(A8)}$$

where γ_{i^*} is a finite coefficient depending on i^*. Note that in (A8) at most four distinct component of X can be present in each summand. Thus, a sufficient condition for (A8) to be bounded is:

$$\mathbb{E}\, X_i X_j X_l X_m < \infty; \quad \forall i, j, l, m = 1, \dots, p. \quad \text{(A9)}$$

which is assumed in (B1). Summarizing, for the second term in (A7), considering each of the terms in the brackets: (i) the first term is finite due to (B1); (ii) the second term, does not depend on X, and (B3) and (B4) ensures it is finite as well; and (iii) the third term requires boundedness of $\mathbb{E}(X_i X_j)$ to be finite, which is already implied by (B1). The finiteness of the third term in (A7) follows from that of the second term, due to the less restrictive conditions on moments required by the former. Analogously, the finiteness of (A7) implies that of the third term in (A6).

Finally, consider taking the superior over Θ. Since the integrand function is continuous in $\boldsymbol{\theta}$, and Θ is assumed compact, the superior is equal to the maximum of the integrand function in $\boldsymbol{\theta}$. Call $\boldsymbol{\theta}_0$ the maximizer, then

$$\mathbb{E}\left(\sup_{\boldsymbol{\theta} \in \Theta} s(X, \boldsymbol{\theta})^2\right) = \mathbb{E}\left(s(X, \boldsymbol{\theta}_0)^2\right) < \infty, \quad \text{(A10)}$$

since the argument above applies to any $\boldsymbol{\theta} \in \Theta$. □

Proof of Proposition 7. Consider a partition of $\mathbb{R}^p$, $\{A_k\}_{k=1\dots K}$, where:

$$A_k := \{\boldsymbol{x} \in \mathbb{R}^p : \log(\pi_k \phi_k(\boldsymbol{x})) \ge \log(\pi_i \phi_i(\boldsymbol{x})) \; \forall i \ne k\}. \quad \text{(A11)}$$

Then,

$$\mathbb{E}(h(X, \boldsymbol{\theta})^2) = \sum_{k=1}^{K} \int_{A_k} \log(\pi_k \phi_k(\boldsymbol{x}))^2 dF(\boldsymbol{x}) \le$$
$$\sum_{k=1}^{K} \int_{A_k} \log(\phi_k(\boldsymbol{x}))^2 dF(\boldsymbol{x}) = \sum_{k=1}^{K} \int_{A_k} \left(C_k - \frac{(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}{2} \right)^2 dF(\boldsymbol{x})$$

where $C_k = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_k|)$. The rest of the proof is identical to that of Proposition 6 (compare with (A6) and (A7)). □

Proof of Proposition 8. We give the proof for h_n. The proof for t_n follows by the same argument, and both are an immediate consequence of Proposition 7 and Proposition 6. By Markov's inequality

$$P(|h_n(\boldsymbol{\theta})| > M) \leq \frac{\mathbb{E}(h_n(\boldsymbol{\theta}))}{M} = \frac{\sum_{i=1}^{n} \mathbb{E}(h(X_i, \boldsymbol{\theta}))}{nM}.$$

For independent and identically distributed X_i,

$$\frac{\sum_{i=1}^{n} \mathbb{E}(h(X_i, \boldsymbol{\theta}))}{nM} = \frac{\mathbb{E}(h(X_i, \boldsymbol{\theta}))}{M}.$$

Proposition 7 ensures $\mathbb{E}(h(X_i, \boldsymbol{\theta}))$ exists and is finte. Choosing $M = \mathbb{E}(h(X_i, \boldsymbol{\theta}))/\epsilon$ completes the proof. □

Proof of Proposition 9. For simplicity, we consider the case of diagonal covariance matrices, $\boldsymbol{\Sigma}_k$, $k \in \{1, \ldots, K\}$, where $\sigma_{k,i}$ indicates the i-th diagonal term in the k-th covariance matrix. This is without loss of generality, since the result can also be shown in the more general case of positive definite variance matrices. Indeed, the expansion of the derivative in this latter case includes at most quadratic terms in x_i and the argument used does not change. Consider the typical components of $\nabla_{\boldsymbol{\theta}} t(\boldsymbol{x}, \boldsymbol{\theta})$:

$$t(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \frac{\pi_k \phi_k(\boldsymbol{x})}{f(\boldsymbol{x}, \boldsymbol{\theta})} \log(\pi_k \phi_k(\boldsymbol{x})); \quad f(\boldsymbol{x}, \boldsymbol{\theta}) \sum_{k=1}^{K} \pi_k \phi_k(\boldsymbol{x});$$

$$\frac{\partial}{\partial \pi_k} t(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\phi_k(\boldsymbol{x})}{f(\boldsymbol{x}, \boldsymbol{\theta})} \Big(\log(\pi_k \phi_k(\boldsymbol{x})) + 1 - t(\boldsymbol{x}, \boldsymbol{\theta}) \Big); \tag{A12}$$

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} t(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\pi_k \phi_k(\boldsymbol{x})}{f(\boldsymbol{x}, \boldsymbol{\theta})} \Big(\log(\pi_k \phi_k(\boldsymbol{x})) + 1 - t(\boldsymbol{x}, \boldsymbol{\theta}) \Big) \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) \tag{A13}$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} t(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\pi_k \phi_k(\boldsymbol{x})}{f(\boldsymbol{x}, \boldsymbol{\theta})} \Big(\log(\pi_k \phi_k(\boldsymbol{x})) + 1 - t(\boldsymbol{x}, \boldsymbol{\theta}) \Big) \left(-\frac{1}{2} \boldsymbol{\Sigma}_k^{-1} \Big(I_p - (\boldsymbol{x} - \mu_k)(\boldsymbol{x} - \mu_k)^\top \boldsymbol{\Sigma}_k^{-1} \Big) \right); \tag{A14}$$

Note that the above equations are of the form:

$$\frac{\pi_k \phi_k(\boldsymbol{x})}{f(\boldsymbol{x}, \boldsymbol{\theta})} \Big(\log(\pi_k \phi_k(\boldsymbol{x})) + 1 - t(\boldsymbol{x}, \boldsymbol{\theta}) \Big) g(\boldsymbol{x}, \boldsymbol{\theta}), \tag{A15}$$

where $g(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{\pi_k}$ for Equation (A12); $g(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)$ for (A13); and for (A14)

$$g(\boldsymbol{x}, \boldsymbol{\theta}) = \left(-\frac{1}{2} \boldsymbol{\Sigma}_k^{-1} \Big(I_p - (\boldsymbol{x} - \boldsymbol{\mu}_k)(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} \Big) \right).$$

Now, to bound each term in $\mathbb{E}\,\|\nabla_{\boldsymbol{\theta}} t(X, \boldsymbol{\theta})\|^2$, we need that the term involving higher moments in X exist finite. Clearly, from (A14), the terms involving higher moments of X are

$$\mathbb{E}\left(\frac{\pi_k \phi_k(X)}{f(X, \boldsymbol{\theta})} \Big(- t(X, \boldsymbol{\theta}) \Big) \Big(\boldsymbol{\Sigma}_k^{-1} X X^\top \boldsymbol{\Sigma}_k^{-1} \Big)_{ij} \right)^2, \quad i, j \in \{1, \ldots, p\}; \tag{A16}$$

However, using the same line of proof in proposition 6, we can expand this further in

$$\mathbb{E}\left(\frac{\pi_k\phi_k(X)}{f(X,\boldsymbol{\theta})}\Big(-t(X,\boldsymbol{\theta})\Big)\Big(\boldsymbol{\Sigma}_k^{-1}XX^\top\boldsymbol{\Sigma}_k^{-1}\Big)_{ij}\right)^2 \leq \mathbb{E}\left(\Big(t(X,\boldsymbol{\theta})\Big)^2\Big(\boldsymbol{\Sigma}_k^{-1}XX^\top\boldsymbol{\Sigma}_k^{-1}\Big)^2_{ij}\right) \leq$$
$$\sum_{j=1}^{K}\int_{\mathbb{R}^p} K^2\log(\pi_j)^2\Big(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma}_k^{-1}\Big)_{ij} dF(\boldsymbol{x})+$$
$$\sum_{j=1}^{K}\int_{\mathbb{R}^p} K^2\log(\phi_j(\boldsymbol{x}))^2\Big(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma}_k^{-1}\Big)_{ij} dF(\boldsymbol{x})+$$
$$\sum_{j=1}^{K}\int_{\mathbb{R}^p} K^2 2\log(\pi_j)\log(\phi_j(\boldsymbol{x}))\Big(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma}_k^{-1}\Big)_{ij} dF(\boldsymbol{x}),$$

where the last inequality is motivated by using the sets A_i as in Proposition 6. Using the same argument of Proposition 6, the term that involves higher moments in term of X is the third one, and in particular, after expansion (compare with (A7)):

$$\mathbb{E}\left(\Big(X^\top\boldsymbol{\Sigma}_k^{-1}X\Big)^2\Big(\boldsymbol{\Sigma}_k^{-1}XX^\top\boldsymbol{\Sigma}_k^{-1}\Big)^2_{ij}\right). \tag{A17}$$

Up to multiplicative constants, the term $(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma}_k^{-1})^2$ has elements of type (A9), thus a sufficient condition to bound the generic element (A17) is

$$\mathbb{E}\left((X_{i_1}X_{i_2}X_{i_3}X_{i_4}X_{i_5}X_{i_6}X_{i_7}X_{i_8})\right) < \infty; \quad \forall i_1, i_2, i_3, i_4, i_5, i_6, i_7, i_8 \in \{1,\ldots,p\}$$

p being the dimension of X. However, (B1*) ensure this condition is satisfied. Now, this condition implies that all the terms appearing in $\|\nabla_\theta s(X,\theta)\|^2$ have a finite expectation. Similarly, because these terms are a continuous function in $\boldsymbol{\theta}$, Θ is assumed compact, and by assumptions (B2) to (B4), the argument above is valid for any $\boldsymbol{\theta} \in \Theta$ and by linearity of the integral, it follows:

$$\mathbb{E}(\sup_{\theta\in\Theta}\|\nabla_\theta s(x,\Theta)\|^2) \leq \mathbb{E}(\|\sup_{\theta\in\Theta}\nabla_\theta s(x,\Theta)\|^2) < \infty.$$

□

References

1. von Luxburg, U.; Williamson, R.C.; Guyon, I. Clustering: Science or Art? In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*; Proceedings of Machine Learning Research; Guyon, I., Dror, G., Lemaire, V., Taylor, G., Silver, D., Eds.; Bellevue: Washington, DC, USA, 2012; Volume 27, pp. 65–79.
2. Coraggio, L.; Coretto, P. Selecting the number of clusters, clustering models, and algorithms. A unifying approach based on the quadratic discriminant score. *J. Multivar. Anal.* **2023**, *196*, 105181. https://doi.org/10.1016/j.jmva.2023.105181.
3. Ullmann, T.; Hennig, C.; Boulesteix, A.L. Validation of cluster analysis results on validation data: A systematic framework. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2022**, *12*, e1444.
4. Hennig, C.; Meila, M.; Murtagh, F.; Rocci, R. (Eds.) *Handbook of Cluster Analysis*; Chapman & Hall/CRC Handbooks of Modern Statistical Methods; CRC Press: Boca Raton, FL, USA, 2016; pp. xx+730.
5. Bouveyron, C.; Celeux, G.; Murphy, T.B.; Raftery, A.E. *Model-Based Clustering and Classification for Data Science*; Cambridge Series in Statistical and Probabilistic Mathematics; Cambridge University Press: Cambridge, UK, 2019; pp. xviii+427. https://doi.org/10.1017/9781108644181.
6. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*; Akad. Kiadó: Budapest, Hungary, 1973; pp. 267–281.
7. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26.
8. Gonçalves, S.; White, H. Maximum likelihood and the bootstrap for nonlinear dynamic models. *J. Econom.* **2004**, *119*, 199–219. https://doi.org/10.1016/S0304-4076(03)00204-5.
9. Billingsley, P. *Probability and Measure*, 3rd ed.; Wiley Series in Probability and Mathematical Statistics; Wiley: New York, NY, USA, 1995.

10. Bickel, P.J.; Freedman, D.A. Some Asymptotic Theory for the Bootstrap. *Ann. Stat.* **1981**, *9*, 1196–1217. https://doi.org/10.1214/aos/1176345637.
11. Bierens, H.J. *Topics in Advanced Econometrics*; Cambridge Unversity Press: New York, NY, USA, 1994.
12. van der Vaart, A.W. *Asymptotic Statistics*; Cambridge Series in Statistical and Probabilistic Mathematics; Cambridge University Press: New York, NY, USA, 1998.
13. Andrews, D.W.K. Higher-Order Improvements of a Computationally Attractive k-Step Bootstrap for Extremum Estimators. *Econometrica* **2023**, *70*, 119–162.
14. Pollard, D. Strong consistency of k-means clustering. *Ann. Stat.* **1981**, *9*, 135–140.
15. O'Hagan, A.; Murphy, T.B.; Scrucca, L.; Gormley, I.C. Investigation of parameter uncertainty in clustering using a Gaussian mixture model via jackknife, bootstrap and weighted likelihood bootstrap. *Comput. Stat.* **2019**, *34*, 1779–1813. https://doi.org/10.1007/s00180-019-00897-9.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.