Electronic Journal of Statistics
Vol. 9 (2015) 1041–1077
ISSN: 1935-7524
DOI: 10.1214/15-EJS1026

Estimation and model selection for model-based clustering with the conditional classification likelihood

Jean-Patrick Baudry[*,†]

LSTA
Université Pierre et Marie Curie – Paris 6
Boîte 158, Tour 15–25, 2e étage
4 place Jussieu
75252 Paris Cedex 05
France
e-mail: Jean-Patrick.Baudry@upmc.fr

Abstract: The Integrated Completed Likelihood (ICL) criterion was introduced by Biernacki, Celeux and Govaert (2000) in the model-based clustering framework to select a relevant number of classes and has been used by statisticians in various application areas. A theoretical study of ICL is proposed.

A contrast related to the clustering objective is introduced: the *conditional classification likelihood*. An estimator and model selection criteria are deduced. The properties of these new procedures are studied and ICL is proved to be an approximation of one of these criteria. We contrast these results with the current leading point of view about ICL, that it would not be consistent. Moreover these results give insights into the class notion underlying ICL and feed a reflection on the class notion in clustering.

General results on penalized minimum contrast criteria and upper-bounds of the bracketing entropy in parametric situations are derived, which can be useful per se.

Practical solutions for the computation of the introduced procedures are proposed, notably an adapted EM algorithm and a new initialization method for EM-like algorithms which helps to improve the estimation in Gaussian mixture models.

AMS 2000 subject classifications: Primary 62H30; Secondary 62H12..
Keywords and phrases: Bracketing entropy, ICL, model-based clustering, model selection, number of classes, penalized criteria.

Received March 2014.

1. Introduction

Model-based clustering is introduced in Sections 1.1 and 1.3. Our purpose is to better understand the ICL criterion of [9], which is presented in Section 1.4.

*Sorbonne Universités, UPMC Univ Paris 06, FRE 3684, LSTA, F-75005, Paris, France.

†This research was partly supported by Université Paris XI, Laboratoire de Mathématiques d'Orsay UMR 8628; INRIA Saclay Île-de-France, Projet SELECT and Université Paris Descartes, MAP5 UMR 8145.

The main topic of this work is the choice of the number of classes in the model-based clustering framework, and then the choice of the number of components of a Gaussian mixture. The reader is referred to [40] or [30] for comprehensive studies on Gaussian mixture models. The latter also provides (Chapter 6) an overview on the approaches for assessing the number of components, and particularly on the standard and widely used penalized likelihood criteria, such as AIC [1] or BIC [37].

The ICL criterion is an alternative to BIC. Up to now it has been widely presented as a penalized likelihood criterion, which penalty involves an "entropy" term. In contrast to this point of view, we prove that it is actually a penalized contrast criterion with a contrast which is not the standard likelihood. This justifies why this is not surprising, nor a drawback, that ICL does not asymptotically select the "true" number of components, even when the "true" model is available. Even for data arising from a mixture distribution, a relevant number of classes may differ from the true number of components of the mixture.

We prove (Section 4.2) that ICL is an approximation of a criterion deriving from a new contrast: the *conditional classification likelihood*. We introduce this contrast which is related to the clustering task. The notion of class underlying ICL is proved to be a compromise between Gaussian mixture density estimation and a strictly "cluster" point of view (Section 7).

We consider a model selection approach and technical methods as developed by [4] and [28]. All proofs are gathered in Section 8.

1.1. Gaussian mixture models

Let X be a random variable in $\mathbb{R}^d$ with distribution $f^{\wp}\cdot\lambda$, where λ is the Lebesgue measure, and $\mathbf{X} = (X_1, \ldots, X_n)$ an i.i.d. sample of the same distribution. Unless specified, all expectations $\mathbb{E}$ and probabilities $\mathbb{P}$ are taken with respect to $f^{\wp}\cdot\lambda$.

$\mathcal{M}_K$ is the Gaussian mixture model with K components:

$$\mathcal{M}_K = \left\{ f(\,.\,;\theta) = \sum_{k=1}^{K} \pi_k \phi(\,.\,;\omega_k) \,\middle|\, \theta = (\pi_1, \ldots, \pi_K, \omega_1, \ldots, \omega_K) \in \Theta_K \right\},$$

where ϕ is the Gaussian density and $\Theta_K \subset \Pi_K \times (\mathbb{R}^d \times \mathbb{S}_+^d)^K$ with $\Pi_K = \{(\pi_1, \ldots, \pi_K) \in [0,1]^K : \sum_{k=1}^K \pi_k = 1\}$ and $\mathbb{S}_+^d$ the set of symmetric positive definite $d \times d$ real matrices. Constraints can be imposed by restricting Θ_K [15]. "General" (no constraint) and "diagonal" (diagonal covariance matrices) models are considered below as examples.

Mixture models are studied here as parametric models. It is then assumed the existence of a parametrization $\varphi : \Theta_K \subset \mathbb{R}^{D_K} \to \mathcal{M}_K$. It is assumed that D_K is minimal and it is called the *dimension* of $\mathcal{M}_K$.

It will not be necessary to assume the parametrization to be identifiable, i.e. that φ is injective. Indeed our purpose is twofold: identifying a relevant number of classes to be designed; and actually designing these classes. Theorem 4.1 justifies that the first task can be achieved under a weaker "identifiability" assumption. Theorem 3.1 then guarantees that our estimator converges to the

set of the best parameters, any of which is as good as the others. There will be no "true parameter" assumption. The classes can finally be defined either through the Maximum A Posteriori (MAP) or the random labels rule (see Section 1.3). Practically, the parameters themselves are never the quantities of interest here. They only serve as a convenient notation and this is also why we expect that the assumption about the Fisher information (see Theorem 4.1) is technical and could maybe be avoided with other techniques. Please refer to [5, Chapter 4] for a more comprehensive discussion about the identifiability question.

1.2. Complete data model

Assume that the studied population consists of K sub-populations which respective proportions are $\pi_1, \ldots, \pi_K$ and which respective conditional distributions are $\phi(\,.\,;\omega_1), \ldots, \phi(\,.\,;\omega_K)$. Let the label of the sub-population to which an individual belongs be modeled by a random variable $Z \in \{0,1\}^K$: $Z_k = 1 \Leftrightarrow X$ belongs to the kth sub-population, which by an abuse of notation we will also denote by $Z = k$. Then we have $\mathbb{P}_Z(k) = \pi_k$, $f_{X|Z=k}(x;\theta) = \phi(x;\omega_k)$, and $f_{(X,Z)}(x,z;\theta) = \prod_{k=1}^K \big(\pi_k \phi(x;\omega_k)\big)^{z_k}$. Hence the *complete data* model $\{f_{(X,Z)}(\,.\,;\theta) | \theta \in \Theta_K\}$. This is a natural model for a clustering problem where X models the observation and Z the sub-population or group which has to be infered. But since the label Z is not observed, only the marginal distribution of X can be studied. Now as a matter of fact its density is $f_X(x;\theta) = \sum_{k=1}^K f_{(X,Z)}(x,k;\theta) = f(x;\theta)$ as defined in Section 1.1. This links the complete data model (with Gaussian conditional distributions) and the Gaussian mixture model $\mathcal{M}_K$.

1.3. Model-based clustering

The process is standard [see 19]:

- estimate a mixture distribution in each considered model;
- select a model and a number of components on the basis of the results of the first step;
- classify the observations through the MAP rule (recalled below) with respect to the mixture distribution estimated in the selected model.

A class is identified with each fitted Gaussian component, which is the most usual choice. Thus the number of classes is chosen at the second step. See for example [22] or [8] for alternative approaches.

Let us recall the MAP classification rule. It involves the *conditional probabilities* of the components

$$\forall \theta \in \Theta_K, \forall k, \forall x,\ \tau_k(x;\theta) = \frac{\pi_k \phi(x;\omega_k)}{\sum_{k'=1}^K \pi_{k'} \phi(x;\omega_{k'})},$$

where $\tau_k(x;\theta)$ is the probability that X belongs to the k^{th} component conditionally to $X = x$ under the distribution defined by θ ($\mathbb{P}_{Z|X=x}(k;\theta)$ in the complete

data model). Let us also denote $\tau_{ik}(\theta) = \tau_k(X_i;\theta)$. The MAP classification rule for x is

$$\widehat{z}^{\mathrm{MAP}}(\theta) = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}}\ \tau_k(x;\theta).$$

Other classification rules are possible, such as drawing the class of x with probabilities $\tau_1(x;\theta),\dots,\tau_K(x;\theta)$ (conforming then to the so-called *fuzzy* classification approach). The latter is probably a better choice when the focus is on the study of the classes (e.g. their shape, how much they overlap...) and the MAP should rather be chosen when the focus is on the classification of single observations. Remark that the classes obtained with the MAP cannot overlap, which is often not realistic. But if the main concern is the class to which an observation should be affected, then the MAP rule seems to be the most relevant choice.

Let us denote by L the *observed likelihood* associated to $\mathbf{X}$:

$$\forall\theta\in\Theta_K,\ \mathrm{L}(\theta;\mathbf{X}) = \prod_{i=1}^{n}\Big(\sum_{k=1}^{K}\pi_k\phi(X_i;\omega_k)\Big).$$

The maximum likelihood estimator in the model $\mathcal{M}_K$ is denoted by $\widehat{\theta}_K^{\mathrm{MLE}}$.

1.4. ICL

Our motivation is to better understand the ICL (Integrated Completed Likelihood) criterion. Let us introduce the *classification likelihood* associated to the complete data sample $(\mathbf{X},\mathbf{Z}) = \big((X_1,Z_1),\dots,(X_n,Z_n)\big)$ in the complete data model:

$$\forall\theta\in\Theta_K,\ \mathrm{L_c}\big(\theta;(\mathbf{X},\mathbf{Z})\big) = \prod_{i=1}^{n}\prod_{k=1}^{K}\big(\pi_k\phi(X_i;\omega_k)\big)^{Z_{ik}}. \tag{1}$$

To mimic the derivation of the BIC criterion [37] in a clustering framework, [9] approximate the integrated classification likelihood through a Laplace's approximation. Then they assume that the classification likelihood mode can be identified with $\widehat{\theta}_K^{\mathrm{MLE}}$ as n is large enough and replace the unobserved Z_{ik}'s by their MAP estimators under $\widehat{\theta}_K^{\mathrm{MLE}}$. This is questionable, notably when the components of $\widehat{\theta}_K^{\mathrm{MLE}}$ are not well separated. They derive the ICL criterion:

$$\mathrm{crit}_{\mathrm{ICL}}(K) = \log\mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}}) + \sum_{i=1}^{n}\sum_{k=1}^{K}\widehat{Z}_{ik}^{\mathrm{MAP}}(\widehat{\theta}_K^{\mathrm{MLE}})\log\tau_{ik}(\widehat{\theta}_K^{\mathrm{MLE}}) - \frac{\log n}{2}D_K.$$

[30] replace the Z_{ik}'s by their conditional expectations $\tau_{ik}(\widehat{\theta}_K^{\mathrm{MLE}})$:

$$\mathrm{crit}_{\mathrm{ICL}}(K) = \log\mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}}) + \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}(\widehat{\theta}_K^{\mathrm{MLE}})\log\tau_{ik}(\widehat{\theta}_K^{\mathrm{MLE}}) - \frac{\log n}{2}D_K. \tag{2}$$

Both versions of the ICL appear to behave analogously, and the latter is considered from now on.

The ICL differs from the standard and widely used BIC criterion of [37] through the *entropy* term (see Section 2.2):

$$\forall \theta \in \Theta_K,\ \mathrm{ENT}(\theta; \mathbf{X}) = -\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}(\theta) \log \tau_{ik}(\theta). \tag{3}$$

The BIC is known to be consistent, in the sense that it asymptotically selects the true number of components, at least when the true distribution actually lies in one of the considered models and under regularity conditions [23, 32]. This nice property may however not suit a clustering purpose. In many applications, there is no reason to assume that the distribution conditional on the (unobserved) labels Z is Gaussian. The BIC in this case tends to overestimate the number of components since several Gaussian components are necessary to approximate each non-Gaussian component of the true mixture distribution $f^{\wp}$. And the user may rather be interested in a *cluster* notion which also includes a separation notion and which is robust to non-Gaussian components. It may be of interest to discriminate into two different classes a group of observations of which the best fit is reached with a mixture of two Gaussian components having quite different parameters (we particularly think of the covariance matrices parameters). BIC tends to do so. But it may also be more relevant and it may conform to an intuitive notion of cluster, to identify two very close—or largely overlapping—Gaussian components as a single non-Gaussian shaped cluster (see for example Figure 1)...

ICL has been derived with this viewpoint. It is widely understood and explained [for instance in 9] as the BIC criterion with a supplementary penalty, which is the entropy (Section 2.2). However we do not think that the entropy should be considered as a penalty term and another point of view will be developed in this paper. A first reason supporting this point of view is that the order of the entropy is $O(n)$, such as $\log \mathrm{L}$, whereas a penalty term is expected to be of order $o(n)$, such as $\frac{\log n}{2} D_K$.

The behavior of ICL has been studied through simulations and real data studies by [9], [30, Section 6.11], [38] and in several simulation studies [see 5, Chapter 4]. Several authors chose to use it in various application areas: [20] (fMRI images); [34] (image collection automatic sorting); [21] (protein structure prediction); [16] (robots learning); [27] (uncovering groups of nodes in valued graphs and application to host-parasite interaction networks in forest ecosystems analysis); [36] (comparative genomic hybridization profile); etc. ICL appears to be more robust than BIC to non-Gaussian components.

This practical interest for ICL lets us think that it meets an interesting notion of cluster, corresponding to what some users expect. But no theoretical study is available. Our main motivation is to propose one. This leads to considering new estimation and model selection procedures for clustering. They are in the same spirit as ICL and are derived by developping the point of view underlying it to its conclusion, from the very estimation step to the model selection step, instead of introducing the MLE. We prove that ICL is an approximation of a criterion which is consistent for a particular loss function.

2. A new contrast: Conditional classification likelihood

The contrast minimization framework turns out to be a fruitful approach. It enables to fully understand that ICL is not a penalized likelihood criterion, on the contrary to the usual point of view. It should rather be linked to another contrast: the *conditional classification likelihood*.

2.1. Origin, definition

In a clustering context, the classification likelihood (1) is an interesting quantity but neither the labels $\mathbf{Z}$ are observed, nor do we even assume that they exist (in case several models with different numbers of components are considered, at most one can correspond to the true number of classes). An approach for model-based clustering consists of involving the classification likelihood instead of the observed likelihood and to estimate $\mathbf{Z}$ as well as θ, i.e. to consider them as parameters of the model (see for example [39, 14], or [2] where a *fuzzy classification likelihood*, between the classification likelihood and the observed likelihood, is considered). The classification likelihood has also been considered to select the number of classes, by estimating the labels (in [9] as mentioned in Section 1.4; see also [11]). We propose here to consider its conditional expectation given the observed sample $\mathbf{X}$.

Let us consider the following algebraic relation between L and $\mathrm{L_c}$:

$$\forall \theta \in \Theta_K,\ \log \mathrm{L_c}(\theta) = \log \mathrm{L}(\theta) + \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik} \log \tau_{ik}(\theta). \tag{4}$$

Then the Conditional Classification log Likelihood ($\log \mathrm{L_{cc}}$) is given by

$$\begin{aligned}\log \mathrm{L_{cc}}(\theta) &= \mathbb{E}_\theta\left[\log \mathrm{L_c}(\theta)|\mathbf{X}\right] \\ &= \log \mathrm{L}(\theta) + \underbrace{\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}(\theta) \log \tau_{ik}(\theta)}_{-\mathrm{ENT}(\theta;\mathbf{X})},\end{aligned}$$

which is obviously linked to the clustering objective. We consider in the following $-\log \mathrm{L_{cc}}$ as an empirical contrast to be minimized.

2.2. Entropy

The contrast $\log \mathrm{L_{cc}}$ differs from $\log \mathrm{L}$ in an *entropy* term (3).

First, consider $\mathrm{ENT}(\theta; x_i)$ for a single observation. Figure 1 represents a dataset simulated from a four-component Gaussian mixture $f^{\wp} = f(\,.\,;\theta)$. Remark that $\mathrm{ENT}(\theta; x_i) \approx 0$ if and only if there exists k_0 such that $\tau_{ik_0} \approx 1$ and $\tau_{ik} \approx 0$ for $k \neq k_0$. There is no difficulty to classify x_i in such a case (for example x_{i_1}). But $\mathrm{ENT}(\theta; x_i)$ is all the greater that $(\tau_{i1}, \ldots, \tau_{iK})$ is closer to $(\frac{1}{K}, \ldots, \frac{1}{K})$, i.e. that the classification through the MAP or random labels rule

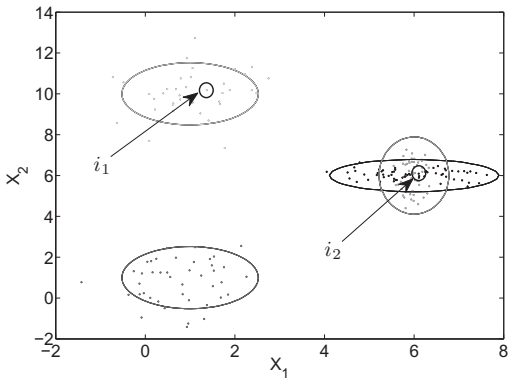


FIG 1. *A dataset example.*

is uncertain. The worst case is reached as the conditional distribution over the components $1, \ldots, K$ is uniform. For example x_{i_2} has about the same posterior probability $\frac{1}{2}$ to arise from each one of the components surrounding it. Its individual entropy is about $\log 2$.

In conclusion, the individual entropy is a measure of the *assignment confidence* of the considered observation through the MAP or random labels rule. The total entropy $\mathrm{ENT}(\theta; \mathbf{x})$ is the empirical mean assignment confidence, and then measures the MAP or random labels classification quality for the whole sample.

Involving the entropy means that one expects the classification to be confident. The class notion underlying the choice of the conditional classification likelihood as a contrast is then a compromise between the fit (and then the idea of Gaussian-shaped classes) because of the likelihood term on the one hand, and the assignment confidence because of the entropy term on the other hand (which is rather a *cluster* point of view).

Notice that $\theta \mapsto \mathrm{ENT}(\theta; \mathbf{x})$ is not differentiable, and not Lipschitz, at any point θ such that any $\tau_{ik}(\theta)$ is zero, which will be the cause of analysis difficulties which do not occur in the study of $\log \mathrm{L}$.

2.3. $\log L_{cc}$ *as a contrast*

The reader is referred to [28] for an introduction to contrast minimization. Let us consider the distribution minimizing the *loss function* associated to $\mathrm{L_{cc}}$ in a model $\mathcal{M}_m = \{f(\,.\,;\theta) : \theta \in \Theta_m\}$:

$$\theta_m \in \underbrace{\underbrace{\operatorname*{argmin}_{\theta\in\Theta_m} \left\{ d_{\mathrm{KL}}\left(f^{\wp}, f(\,.\,;\theta)\right) + \mathbb{E}\left[\mathrm{ENT}(\theta; X)\right] \right\}}_{\operatorname*{argmin}_{\theta\in\Theta_m} \mathbb{E}\left[-\log \mathrm{L_{cc}}(\theta)\right]}}_{\text{this set is denoted by } \Theta_m^0}.$$

The existence of $\mathbb{E}\left[-\log \mathrm{L}_{\mathrm{cc}}(\theta)\right]$ is a very mild assumption. The non-emptiness of Θ_m^0 may be guaranteed for example by assuming Θ_m to be compact. Let K be fixed and consider the minimization of the loss function in the model $\mathcal{M}_K$ (Section 1.1). First, remark that $\log \mathrm{L}_{\mathrm{cc}} = \log \mathrm{L}$ if $K = 1$: Θ_1^0 is the set of parameters of the distributions which minimize the Kullback-Leibler divergence to $f^{\wp}$. If $K > 1$, $\theta_K^0 \in \Theta_K^0$ may almost minimize the Kullback-Leibler divergence if the corresponding components do not overlap since then the entropy is about zero. Otherwise this is not the case (Example 2.1).

Now, from the L_{cc} point of view, the (population) best distribution is given by $\operatorname{argmin}_{\theta \in \mathcal{U}} \mathbb{E}\left[-\log \mathrm{L}_{\mathrm{cc}}(\theta)\right]$. The *universe* $\mathcal{U}$ must be chosen with care. There is no natural relevant choice, on the contrary to the density estimation framework where the set of all densities may be chosen. First the considered contrast is defined in a parametric mixture setup such as considered in this article, but not over any mixture density set (different parameterizations may lead to different values of the entropy term since the definition of each component cannot be recovered from the mixture density alone). However, this would still enable to consider mixtures much more general than mixtures of Gaussian components. To overcome some limitations of these, authors consider mixtures of different families of distributions [see for example 26]. The ideas developed in [22] or [8] may even suggest to involve mixtures which components are Gaussian mixtures. But this latter idea would not make sense here. The mixture with one component which is a mixture of K Gaussian components, and then which yields a single non-Gaussian-shaped class, always has a smaller $-\log \mathrm{L}_{\mathrm{cc}}$ value than the corresponding Gaussian mixture yielding K classes. The choice of the components is a very strong modelization choice: allowing for example the components to be any mixture of Gaussian mixtures means that a class may be almost anything and may notably contain two Gaussian-shaped clusters very far from each other! The final MAP classification depends on this choice and the same problem studied by two different researchers who made two different choices for $\mathcal{U}$ can produce two different solutions. The components should at least be chosen with respect to the corresponding cluster shape. A most natural idea is then mostly to involve only Gaussian mixtures: $\mathcal{U}$ may be chosen as $\cup_{1 \leq K \leq K_M} \mathcal{M}_K$.

Example 2.1. $f^{\wp}$ is the normal density $\mathcal{N}(0,1)$ ($d = 1$). The model $\mathcal{M}_2 = \{\frac{1}{2}\phi(\,.\,; -\mu, \sigma^2) + \frac{1}{2}\phi(\,.\,; \mu, \sigma^2); \mu \in \mathbb{R},\ \sigma^2 > 0\}$ is considered.

Let us study Θ_2^0 in this simple situation. We numerically obtain that $\Theta_2^0 = \{(-\mu_0, \sigma_0^2), (\mu_0, \sigma_0^2)\}$, so that, up to a label switch, there exists a unique minimizer of $\mathbb{E}\left[-\log \mathrm{L}_{\mathrm{cc}}(\mu, \sigma^2)\right]$ in Θ_2 in this case (see Figure 2), with $\mu_0 \approx 0.83$ and $\sigma_0^2 \approx 0.31$. This solution is obviously not the same as the one minimizing the Kullback-Leibler divergence (see Figure 3). This illustrates that the objective with the $-\log \mathrm{L}_{\mathrm{cc}}$ contrast is not to recover the true distribution, even when it is available in the considered model.

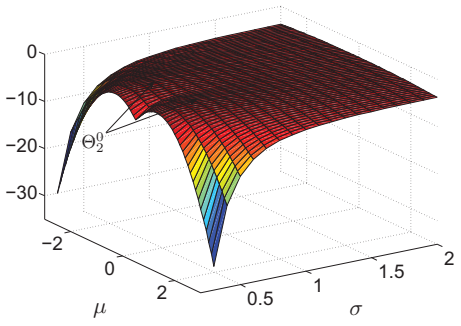


FIG 2. $\mathbb{E}\left[\log L_{cc}(\mu,\sigma^2)\right]$ *w.r.t.* μ *and* σ, *and* Θ_K^0, *for Example 2.1.*

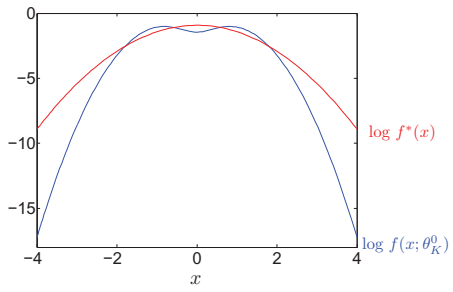


FIG 3. $\log f^{\wp}$ *(red, which is also* $\log f(\,.\,;\theta_2^{KL})$*) and* $\log f(\,.\,;\theta_2^0)$ *(blue) for Example 2.1.*

3. Estimation: MLccE

Let us fix the number of components K and the model $\mathcal{M}_K$. The subscript K is omitted in the notation of this section. A new minimum contrast estimator in $\mathcal{M}$ is considered (Definition 5). This section's main result is the consistency of this estimator (Theorem 3.1). It relies on bracketing entropy upper bounds (Lemma 3.2). Since they could be useful in other situations, general statements are provided (Section 3.2).

Let $\mathcal{M} = \{f(\,.\,;\theta);\theta \in \Theta\}$ be any parametric model with Θ compact (in Section 3.1 and when specified $\mathcal{M}$ is a Gaussian mixture model). Let $\Theta^{\mathcal{O}}$ an open subset of $\mathbb{R}^D$ such that $\Theta \subset \Theta^{\mathcal{O}}$ and $\gamma : \Theta^{\mathcal{O}} \times \mathbb{R}^d \longrightarrow \mathbb{R}$ a contrast (in Section 3.1, $\gamma = -\log \mathrm{L_{cc}}$) such that $\theta \in \Theta^{\mathcal{O}} \mapsto \gamma(\theta;x)$ be C^1 for $f^{\wp}d\lambda$-almost all x and $\theta \in \Theta \mapsto \mathbb{E}\left[\gamma(\theta;X)\right]$ be continuous. Let its empirical version be given by $\gamma_n(\theta) = \frac{1}{n}\sum_{i=1}^n \gamma(\theta;X_i)$. $\mathbb{R}^D$ is equipped with the infinite norm: $\forall\theta \in \mathbb{R}^D, \|\theta\|_\infty = \max_{1\le i\le D}|\theta_i|$. For any $r \in \mathbb{N}^* \cup \{\infty\}$ and for any measurable $g : \mathbb{R}^d \to \mathbb{R}$, $\|g\|_r = \mathbb{E}\left[|g(X)|^r\right]^{\frac{1}{r}}$ if $r < \infty$ and $\|g\|_\infty = \operatorname{ess\,sup}_{X\sim f^{\wp}}|g(X)|$. For any linear form $l : \mathbb{R}^D \to \mathbb{R}$, $\|l\|_\infty = \max_{\|\theta\|_\infty=1} l(\theta)$.

3.1. Definition, consistency

In this subsection $\mathcal{M}$ is a Gaussian mixture model with compact parameter space $\Theta \subset \Theta^{\mathcal{O}} \subset \mathbb{R}^D$ and $\gamma = -\log \mathrm{L_{cc}}$.

The Maximum conditional classification Likelihood Estimator is defined:

$$\widehat{\theta}^{\mathrm{MLccE}} \in \operatorname*{argmax}_{\theta\in\Theta} \log \mathrm{L_{cc}}(\theta). \tag{5}$$

The compactness of Θ guarantees the existence of $\widehat{\theta}^{\mathrm{MLccE}}$. This is a strong assumption, but it will be natural and necessary for the following results to hold. That the covariance matrices be bounded from below is a reasonable and necessary assumption: without it, neither L nor $\mathrm{L_{cc}}$ would be bounded (for $K \ge 2$).

Besides it is necessary to impose positive lower bounds on the proportions to guarantee the existence of $\Theta^{\mathcal{O}}$. Insights to practically choose lower bounds on the proportions and the covariance matrices are suggested in [5, Section 5.1]. The upper bound on the covariance matrices and the compactness condition on the means, although not necessary in the standard likelihood framework, do not seem to be avoidable here. This results from the behavior of the entropy term when a component vanishes.

Theorem 3.1 (Weak Consistency of MLccE)**.** *Assume that Θ is compact and contains no parameter with a zero-proportion component. Let $\Theta^0 = \operatorname{argmax}_{\theta\in\Theta} \mathbb{E}\left[\log L_{cc}(\theta)\right]$ and $\widehat{\theta}^{MLccE} \in \Theta$ an estimator s.t.:*

$$\forall \theta^0 \in \Theta^0, \forall n \in \mathbb{N}^*, \log L_{cc}(\widehat{\theta}^{MLccE}) \geq \log L_{cc}(\theta^0) + o_{\mathbb{P}}(n).$$

Then $d(\widehat{\theta}^{MLccE}, \Theta^0) \xrightarrow[n\longrightarrow\infty]{\mathbb{P}} 0$ (with $d(\theta, \Theta^0) = \inf_{\theta^0\in\Theta^0} \|\theta - \theta^0\|$).

$\widehat{\theta}^{\mathrm{MLccE}}$ is strongly consistent if it minimizes the empirical contrast almost surely. Let us insist that it converges to the set of parameters minimizing the loss function, which has no reason to contain the true distribution—except for $K = 1$—even if the latter lies in $\mathcal{M}$.

3.2. Bracketing entropy and Glivenko-Cantelli property

The results of this section hold for any model and contrast which fulfill the assumptions they involve. We will justify their application to Gaussian mixture models and the contrast $-\log \mathrm{L_{cc}}$ but they hold much more generally.

The reader is referred to [41, Chapter 19] or [18, Chapter 7] for thorough treatments of these topics. Recall a class of functions $\mathcal{G}$ is $\mathbb{P}$–*Glivenko-Cantelli* if it fulfills a uniform law of large numbers for the distribution $\mathbb{P}$ ($\sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n g(X_n) - \mathbb{E}_{\mathbb{P}}[g(X)]\right| \xrightarrow[n\to\infty]{\mathbb{P}\text{–a.s.}} 0$). A sufficient condition for a family $\mathcal{G}$ to be $\mathbb{P}$–Glivenko-Cantelli is that it is not too complex, which can be measured through its *entropy with bracketing*:

Definition 3.1 ($L_r(\mathbb{P})$-entropy with bracketing)**.** Let $r \in \mathbb{N}^*$ and $l, u \in L_r(\mathbb{P})$. The bracket $[l, u]$ is the set of all functions $g \in L_r(\mathbb{P})$ with $l \leq g \leq u$. The bracket $[l, u]$ is an ε-bracket if $\|l-u\|_r \leq \varepsilon$. The bracketing number $N_{[]}(\varepsilon, \mathcal{G}, L_r(\mathbb{P}))$ is the minimum number of ε-brackets needed to cover $\mathcal{G}$. The entropy with bracketing of $\mathcal{G}$ with respect to $\mathbb{P}$ is $\mathcal{E}_{[]}(\varepsilon, \mathcal{G}, L_r(\mathbb{P})) = \log N_{[]}(\varepsilon, \mathcal{G}, L_r(\mathbb{P}))$.

It is quite natural that the behavior of all functions lying inside a bracket can be uniformly controlled by the behavior of the extrema of the bracket. If those endpoints belong to $L_1(\mathbb{P})$, they fulfill a law of large numbers, and if the number of them needed to cover $\mathcal{G}$ is finite, then this is no surprise that $\mathcal{G}$ can be proved to fulfill a uniform law of large numbers:

Theorem 3.2. *Every class $\mathcal{G}$ of measurable functions such that $\mathcal{E}_{[]}(\varepsilon, \mathcal{G}, L_1(\mathbb{P})) < \infty$ for every $\varepsilon > 0$ is $\mathbb{P}$–Glivenko-Cantelli.*

This is a generalization of the usual Glivenko-Cantelli theorem. We shall provide $f^{\wp}\cdot\lambda$–bracketing entropy upper bounds for $\{\gamma(\,.\,;\theta):\theta\in\Theta_K\}$.

The following hypotheses will be repeatedly involved in the following:

$$\|M\|_r<\infty \text{ with } M(x)=\sup_{\theta\in\Theta}|\gamma(\theta;x)|<\infty \ f^{\wp}d\lambda\text{-a.e.} \qquad (H_{\gamma,\Theta,r})$$

$$\|M'\|_r<\infty \text{ with } M'(x)=\sup_{\theta\in\Theta}\left\|\left(\frac{\partial\gamma}{\partial\theta}\right)_{(\theta;x)}\right\|_{\infty}<\infty \ f^{\wp}d\lambda\text{-a.e.} \qquad (H'_{\gamma,\Theta,r})$$

These assumptions when $r<\infty$ are not a difficulty with $\gamma=-\log \mathrm{L}_{\mathrm{cc}}$ under the assumptions of Section 3.1. $H'_{\gamma,\Theta^{\mathcal{O}},r}$ with $\Theta^{\mathcal{O}}$ as in Section 3.1 holds too since $\Theta^{\mathcal{O}}$ itself can be chosen to be included in a compact subset of the set of possible parameters.

The stronger assumption $H_{\gamma,\Theta,\infty}$, which is involved below, is not always fulfilled in the general Gaussian mixtures framework. A sufficient condition for it to hold is that the support of $f^{\wp}$ be bounded. This is a reasonable modeling assumption: most phenomena are bounded. Actually, our results can be expected to hold all the same when $f^{\wp}$ has reasonably low tails (see p. 1054)... We did not prove it yet though. Another sufficient condition to guarantee this assumption is that the contrast be upper-bounded. This is actually not the case of the contrast $-\log \mathrm{L}_{\mathrm{cc}}$, but we could replace $-\log \mathrm{L}_{\mathrm{cc}}$ by $(-\log \mathrm{L}_{\mathrm{cc}}\wedge C)$ and, provided that C is large enough, this new contrast behaves like $\log \mathrm{L}_{\mathrm{cc}}$.

Lemma 3.1 guarantees that the bracketing entropy of $\{\gamma(\,.\,;\theta):\theta\in\Theta\}$ is finite for any ε, if Θ is convex and bounded. The lemma is written for any $\widetilde{\Theta}$ bounded and included in Θ (which is not assumed to be bounded itself) since it will be applied locally around θ^0 in Section 4.

For any $\widetilde{\Theta}\subset\mathbb{R}^D$, $\operatorname{diam}\widetilde{\Theta}=\sup\{\|\theta_1-\theta_2\|_\infty:\theta_1,\theta_2\in\widetilde{\Theta}\}$.

Lemma 3.1 (Bracketing Entropy, Convex Parameter Space)**.** *Let $r\in\mathbb{N}^*$ and assume that Θ is convex. Assume that $H'_{\gamma,\Theta,r}$ holds. Then*

$$\forall\widetilde{\Theta}\subset\Theta,\forall\varepsilon>0,\ N_{[]}(\varepsilon,\{\gamma(\theta):\theta\in\widetilde{\Theta}\},\|\cdot\|_r)\le\left(\frac{\|M'\|_r\ \operatorname{diam}\widetilde{\Theta}}{\varepsilon}\right)^D\vee 1.$$

Remark that Θ is not assumed to be compact. The natural parameter space of diagonal Gaussian mixture models, with equal volumes or not, for instance, is convex (Examples 8.1 and 8.2). The natural parameter space of general Gaussian mixture models is convex too, since the set of definite positive matrices is. However, Θ is not always convex.

The convexity assumption can be relaxed if we assume Θ to be compact and $H'_{\gamma,\Theta^{\mathcal{O}},r}$ to hold. The upper bound is then increased by a multiplying factor Q which only depends on Θ and roughly measures its "nonconvexity". Since our main concern is the power of ε in the upper bound, this weakens the result ever so slightly.

Lemma 3.2 (Bracketing Entropy, Compact Parameter Space)**.** *Let $r\in\mathbb{N}^*$ and assume that Θ is compact. Assume that $H'_{\gamma,\Theta^{\mathcal{O}},r}$ holds. Then*

$$\exists Q \in \mathbb{N}^*, \forall \widetilde{\Theta} \subset \Theta, \forall \varepsilon > 0,$$
$$N_{[]}(\varepsilon, \{\gamma(\theta) : \theta \in \widetilde{\Theta}\}, \|\cdot\|_r) \leq Q \left(\frac{\|M'\|_r \ \mathrm{diam}\, \widetilde{\Theta}}{\varepsilon} \right)^D \vee 1.$$

Q is a constant which depends on the geometry of Θ ($Q = 1$ if Θ is convex).

In Section 4 we need a slight modification of Lemma 3.1. Since it is applied locally there, the convexity assumption is not a problem.

Lemma 3.3 (Bracketing Entropy, Convex Parameter Space)**.** *Let $r \geq 2$, $D \in \mathbb{N}^*$ and assume that Θ is convex. Assume that $H_{\gamma,\Theta,\infty}$ and $H'_{\gamma,\Theta,2}$ hold. Then*

$$\forall \widetilde{\Theta} \subset \Theta, \forall \varepsilon > 0,$$
$$N_{[]}(\varepsilon, \{\gamma(\theta) : \theta \in \widetilde{\Theta}\}, \|\cdot\|_r) \leq \left(\frac{2^{r-2} \|M\|_\infty^{\frac{r-2}{2}} \|M'\|_2 \ \mathrm{diam}\, \widetilde{\Theta}}{\varepsilon^{\frac{r}{2}}} \right)^D \vee 1.$$

Let us remark that these results are quite general. We are interested here in their application to the conditional classification likelihood, but they hold all the same in the standard likelihood framework. [29] provide bracketing entropy results in this framework. Our results cannot be directly compared to theirs since they consider the Hellinger distance. The dependency they get on the parameter space bounds and the variable space dimension d is explicit, which is helpful to derive an oracle inequality. But they could not derive a local control of the entropy (i.e. corresponding to a small subset of Θ around some θ), hence an unpleasant logarithm term in their expression of an optimal penalty. Moreover, their results suggest the necessity of assuming the contrast to be bounded: see the discussion after Theorem 4.1 below. Our results achieve the same rate with respect to ε. They depend on more opaque quantities ($\|M\|_\infty$ and $\|M'\|_2$) but it should be possible to control them with respect to the bounds on Θ if needed. And beside their simplicity, they provide a local control of the entropy.

4. Model selection

As illustrated by Example 2.1, model selection is a crucial step. The number of classes may even be the purpose of the study.

Results are stated in Section 8.3 generally, for any contrast and family of models. Here we apply them to $-\log \mathrm{L_{cc}}$ and the Gaussian mixture models introduced in Section 1.1 and derive penalized conditional classification likelihood criteria written as

$$\mathrm{crit}(K) = -\log \mathrm{L_{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}) + \mathrm{pen}(K).$$

In Section 4.1, the consistency of a class of penalties is proved. Sufficient conditions are given in Theorem 8.1, which is applied to the framework we are interested in, in Theorem 4.1. The strongest condition of Theorem 8.1 (B4) is

discussed in Section 8.4 and can be guaranteed under regularity and (weak) identifiability assumptions. Our approach is adapted from [28] and [4] and is a first step to reach non-asymptotic results.

4.1. Consistent penalized criteria

Assume that K_0 exists such that

$$\text{and} \quad \begin{array}{l} \forall K < K_0,\ \sup\limits_{\theta\in\Theta_{K_0}} \mathbb{E}\left[\log \mathrm{L_{cc}}(\theta)\right] > \sup\limits_{\theta\in\Theta_K} \mathbb{E}\left[\log \mathrm{L_{cc}}(\theta)\right] \\ \forall K \geq K_0,\ \sup\limits_{\theta\in\Theta_{K_0}} \mathbb{E}\left[\log \mathrm{L_{cc}}(\theta)\right] \geq \sup\limits_{\theta\in\Theta_K} \mathbb{E}\left[\log \mathrm{L_{cc}}(\theta)\right] \end{array}$$

which means that the bias of the models is stationary for $K \geq K_0$: $\mathcal{M}_{K_0}$ is the "best" model. Remark that the last property should hold mostly in the mixtures framework, in particular if the models are nested. Under this assumption, a model selection procedure is expected to asymptotically select K_0, i.e. to be *consistent*. This is an *identification* aim [see 31, Chapter 1]. It would be disastrous to select a model which does not (almost) minimize the bias and then a smaller value than K_0. And it is assumed that the model $\mathcal{M}_{K_0}$ contains all the relevant information (typically, the structure of the classes).

Let us stress that the "true" number of components of $f^{\wp}$ is not of direct concern: it is not assumed that it equals K_0, and it is not even assumed to be defined ($f^{\wp}$ does not have to be a Gaussian mixture). K_0 is the best choice from the point of view introduced by using $\log \mathrm{L_{cc}}$, which is neither density estimation nor identification of the "true" number of components.

The following theorem is an application of the more general Theorem 8.1 (stated for any contrast and collection of models):

Theorem 4.1. *Assume that the support of $f^{\wp}$ is bounded.*

$(\mathcal{M}_K)_{1\leq K\leq K_M}$ are Gaussian mixture models with compact parameter spaces $\Theta_K \subset \mathbb{R}^{D_K}$ which contain no parameter with a zero-proportion component, $\Theta_K^0 = \operatorname{argmax}_{\theta\in\Theta_K} \mathbb{E}[\log L_{cc}(\theta)]$ and $K_0 = \min \operatorname{argmax}_{1\leq K\leq K_M} \mathbb{E}[\log L_{cc}(\Theta_K^0)]$.

For any K, assume that

$$\begin{array}{l} \forall \theta \in \Theta_K, \forall \theta_{K_0}^0 \in \Theta_{K_0}^0, \\ \qquad \mathbb{E}\left[\log L_{cc}(\theta)\right] = \mathbb{E}\left[\log L_{cc}(\Theta_{K_0}^0)\right] \\ \qquad \Longleftrightarrow \log L_{cc}(\theta; x) = \log L_{cc}(\theta_{K_0}^0; x) \quad f^{\wp} d\lambda - a.e. \end{array} \tag{C1}$$

Assume that $\forall \theta_K^0 \in \Theta_K^0$, $I_{\theta_K^0} = \frac{\partial^2}{\partial \theta^2}(\mathbb{E}\left[\log L_{cc}(\theta)\right])_{|\theta_K^0}$ is nonsingular.

Let $\widehat{\theta}_K^{MLccE} \in \Theta_K$ with $\log L_{cc}(\widehat{\theta}_K^{MLccE}) \geq \log L_{cc}(\theta_K^0) + o_{\mathbb{P}}(n)$.

Let pen : $\{1, \ldots, K_M\} \to \mathbb{R}^+$ *(which may depend on n, $(\Theta_K)_{1\leq K\leq K_M}$ and the data) such that*

$$\forall K \in \{1, \ldots, K_M\}, \begin{cases} \mathrm{pen}(K) > 0 \text{ and } \mathrm{pen}(K) = o_{\mathbb{P}}(n) & \text{when } n \to +\infty \\ \left(\mathrm{pen}(K) - \mathrm{pen}(K')\right) \xrightarrow[n\to+\infty]{\mathbb{P}} \infty & \text{for any } K' < K. \end{cases}$$

Define $\hat{K}$ *such that* $\hat{K} = \min \operatorname{argmin}_{1\leq K\leq K_M}\{-\log L_{cc}(\widehat{\theta}_K^{MLccE}) + \operatorname{pen}(K)\}$.
Then $\mathbb{P}[\hat{K} \neq K_0] \xrightarrow[n\longrightarrow\infty]{} 0$.

The "identifiability" assumption (C1) is reasonable: as expected the label switching phenomenon [see 30, Section 1.14] is not a problem here. But it is necessary for the identification point of view to make sense, that a single value of the *contrast function* $x \longmapsto \gamma(\theta; x)$ minimizes the loss. Remark that in the standard likelihood framework, this holds at least if any model contains the sample distribution, since it is then the unique Kullback-Leibler divergence minimizer. Several parameter values, perhaps in different models, may represent it, besides the label switching. We do not know any such result for $\log \mathrm{L_{cc}}$ and hypothesize that the assumption holds.

Hopefully the assumption about the non-singularity of I_{θ^0} could be weakened. The analogous result of [28] (Theorem 7.11) which inspires this, and is available for the standard likelihood context, does not require such an assumption since it does not rely on a study of the link between the contrast and the parameters but on a smart choice of the distance (Hellinger distance), and on properties of the log function. However, this is a usual assumption [see 35, or below].

Moreover [28] does not require the contrast he considers ($\log \mathrm{L}$) to be bounded. Remark however that the application of his Lemma 7.23 to obtain a genuine oracle inequality involves an assumption similar to the boundedness of the contrast. So that it seems reasonable that the assumptions $H_{\log \mathrm{L_{cc}},\Theta_K,\infty}$ and $H'_{\log \mathrm{L_{cc}},\Theta_K,2}$ be necessary. In the theorem, they are guaranteed by the boundedness of the support of $f^{\wp}$ but it could be replaced by other assumptions: please refer to the discussion p. 1051. Let us justify why we claim there that our results are expected to hold when $f^{\wp}$ has reasonably low tails.

This claim is supported by the fact that

$$P(\hat{K} \neq K_0) \leq P(\hat{K} \neq K_0 \text{ and } (X_1, \ldots, X_n) \in C^n) + P((X_1, \ldots, X_n) \notin C^n),$$

with C a compact. Let $\eta > 0$. Then, Theorem 4.1 holds conditionally to $(X_1, \ldots, X_n) \in C^n$ (since then, conditionally, "the support of $f^{\wp}$ is bounded") and for $n > N \in \mathbb{N}$, $P(\hat{K} \neq K_0 \text{ and } (X_1, \ldots, X_n) \in C^n) < \frac{\eta}{2}$. Now, for $n > N$ fixed, we can find C such that $P((X_1, \ldots, X_n) \notin C^n) < \frac{\eta}{2}$. The problem of course is that N depends on C, and that we need to choose C with respect to n... But we see that the conclusion of Theorem 4.1 must still hold without the support boundedness assumption, provided that C does not grow too fast as n grows... i.e. provided that "$f^{\wp}$ has reasonably low tails".

Our results can be compared to those of [32] and [23]. Both study consistency conditions for penalized criteria in the standard maximum likelihood framework. Assuming the convexity of Θ_K, strong identifiability ($\Theta_K^0 = \{\theta_K^0\}$), non-singularity of the Fisher information (analogous to our $I_{\theta_K^0}$) and other regularity conditions which are not designed for the particular case of mixture models, [32] proves procedures with penalties of the form $c_n D_K$ to be weakly consistent as soon as $\frac{c_n}{n} \to 0$ and $c_n \to \infty$. In a general mixture model framework, assuming the model family to be well-specified, the same identifiability condition as we do,

and a condition about the Fisher information which does not seem to be directly comparable to ours but sounds roughly the same (might be milder), [23] proves procedures with any penalty form to be consistent if $\frac{\text{pen}(K)}{n} \to 0$, $\text{pen}(K) \to \infty$ and $\liminf \frac{\text{pen}(K)}{\text{pen}(K')} > 1$ if $K > K'$ (which is equivalent to the conditions of [32] if $\text{pen}(K) = c_n D_K$). The assumptions are proved to hold for $\log \text{L}$ for Gaussian mixture models with lower bounded, spherical covariance matrices which are constrained to be equal for all the components and if the means belong to a compact subset.

In comparison our conditions (applied here for the $\log \text{L}_{\text{cc}}$ contrast or more generally, including the $\log \text{L}$ contrast, in Theorem 8.1) about the penalties are a little weaker than those of [23] but quite analogous, as expected. For the application to $\log \text{L}_{\text{cc}}$ we have to introduce a supplementary assumption, that the mixing proportions be kept away from zero. It does not seem easy to extend the methods used by [23] to our framework.

To get strong consistency, both [32] and [23] had to assume moreover that $\frac{\text{pen}(K)}{\log \log n} \to \infty$. By analogy it can then be conjectured that the strong version of Theorem 4.1 would probably involve penalties a little stronger.

4.2. A new light on ICL

The previous section suggests analogies between model selection penalized criteria based on L on the one hand and on L_{cc} on the other hand. Therefore, by analogy with the standard likelihood framework, it is expected that penalties proportional to D_K conform a prediction point of view (think of AIC), and that penalties proportional to $D_K \log n$ are optimal for an identification purpose (think of BIC). This possibility to derive an identification procedure from a prediction procedure by a $\log n$ factor is notified for example in [3, Section 1.2.3].

Let us then consider by analogy with BIC the penalized criterion

$$\text{crit}_{\text{L}_{\text{cc}}-ICL}(K) = \log \text{L}_{\text{cc}}(\widehat{\theta}_K^{\text{MLccE}}) - \frac{\log n}{2} D_K.$$

We almost recover ICL (replace $\widehat{\theta}_K^{\text{MLE}}$ by $\widehat{\theta}^{\text{MLccE}}$ in (2)), which may then be regarded as an approximation of this L_{cc}-ICL criterion. The corresponding penalty is $\frac{\log n}{2} D_K$, and the derivation of L_{cc}-ICL illustrates that the entropy should not be considered as part of the penalty. This notably justifies why ICL does not select the same number of components as BIC or any consistent criterion in the standard likelihood framework, even asymptotically. Actually, it should not be expected to do so.

When $\widehat{\theta}_K^{\text{MLccE}}$ differs from $\widehat{\theta}_K^{\text{MLE}}$, the former provides more separated clusters. The compromise between the Gaussian component and the cluster viewpoint is achieved with $\widehat{\theta}_K^{\text{MLccE}}$ from the very estimation step. However, ICL and L_{cc}-ICL usually behave analogously in the simulations (Section 6).

Finally, L_{cc}-ICL is quite close to ICL and enables to better understand the concepts underlying ICL. ICL remains attractive though, notably because it is easier to implement than L_{cc}-ICL.

4.3. Slope heuristics

Another interesting approach for model selection by penalized criteria is the slope heuristics of [12, 13]. It has been successfully applied to many situations and particularly to model-based clustering with the usual likelihood [see 7, for an overview and practical considerations].

Since we define a new contrast adapted to the clustering objective we can apply this heuristics to calibrate penalties of the form suggested in Section 4.1. No definitive theoretical justification that the penalty should be chosen proportional to the dimension of the model is available up to now, but we have hints in that direction. Simulations (Section 6) seem to confirm this choice. Criteria of the following form, with κ unknown, are then considered and expected to have an oracle-like behavior:

$$\mathrm{crit}(K) = -\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}) + \kappa D_K,$$

The slope heuristics provides a practical data-driven approach to choose κ.

The slope heuristics relies on the assumption that the bias of the models decreases as their complexity increases and is stationary for the most complex models. In our framework, this requires the family of models to be roughly nested, which does not always hold depending on the constraints on the models.

5. Practical

We introduce practical solutions for the computation of $\widehat{\theta}^{\mathrm{MLccE}}$: an algorithm adapted from the EM algorithm [17] and a new initialization procedure ("Km1") which can also be useful for the usual EM algorithm. Details are beyond the scope of this article: they are presented and discussed in [5, Chapter 5].

5.1. Definition and fundamental property of L_{cc}-EM

The L_{cc}-EM algorithm we introduce is inspired by the BEM algorithm [25].

The steps of the j^{th} algorithm iteration ($\theta^{j-1} \to \theta^j$) are:

E step For all $\theta \in \Theta_K$,

$$\begin{aligned} Q(\theta, \theta^{j-1}) &= \mathbb{E}_{\theta^{j-1}}\left[\log \mathrm{L}_{\mathrm{c}}(\theta)|\mathbf{X}\right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}(\theta^{j-1}) \log \pi_k \phi(X_i;\omega_k). \end{aligned}$$

M step Maximize $Q(\theta, \theta^{j-1}) - \mathrm{ENT}(\theta;\mathbf{X})$ with respect to $\theta \in \Theta_K$:

$$\theta^j \in \underset{\theta\in\Theta_K}{\operatorname{argmax}} \Bigg\{ \underbrace{\mathbb{E}_{\theta^{j-1}}\left[\log \mathrm{L}_{\mathrm{c}}(\theta)|\mathbf{X}\right] - \mathrm{ENT}(\theta;\mathbf{X})}_{\log \mathrm{L}(\theta) + \sum_{i=1}^{n}\sum_{k=1}^{K}(\tau_{ik}(\theta^{j-1}) + \tau_{ik}(\theta))\log \tau_{ik}(\theta)} \Bigg\}$$

On the contrary to many situations with the EM algorithm [15], we do not know any case of closed-form M step for $\mathrm{L_{cc}}$-EM. Therefore it has to be performed by means of numerical maximization.

Proposition 5.1 (Fundamental Property of the $\mathrm{L_{cc}}$-EM algorithm)**.**

$$\forall \theta, \theta' \in \Theta_K, Q(\theta', \theta) - \mathrm{ENT}(\theta'; \mathbf{X}) > Q(\theta, \theta) - \mathrm{ENT}(\theta; \mathbf{X}) \\ \Longrightarrow \log L_{cc}(\theta') > \log L_{cc}(\theta).$$

The proof is straightforward. This property suggests the interest of the algorithm to maximize $\log \mathrm{L_{cc}}$, which is assessed by simulations studies.

Note that the monotonicity of the contrast still holds if the M-step is weakened into an increase (instead of a maximization) of $Q(\theta, \theta^j) - \mathrm{ENT}(\theta; \mathbf{X})$, which is a good point about the algorithm stability despite numerical optimization.

5.2. Initialization: Known and new methods

The choice of θ^0 is crucial for the $\mathrm{L_{cc}}$-EM, as for the EM algorithm. The reader is referred to [10] for initialization methods for the standard EM. Our approach is inspired by the same idea: try different methods and keep the best result to initialize $\mathrm{L_{cc}}$-EM. We involve in our initializations:

- Solutions obtained through partial standard likelihood optimization relying on the standard EM and initialization methods from [10]:
  - Classification EM (CEM);
  - EM with random starts or small_EM starts;
  - K-means.
- Solutions obtained through a procedure directly inspired by [10] and adapted for $\log \mathrm{L_{cc}}$:
  - Small_$\mathrm{L_{cc}}$-EM. Choose the best solutions $\theta^0_{\mathrm{small}}$ among those obtained after short runs of $\mathrm{L_{cc}}$-EM from random starts.
- Solutions obtained by a new procedure called Km1 and introduced below.

Km1 "Km1" stands for "K minus 1". Figure 4 illustrates this strategy. Suppose $K \geq 2$ and $\widehat{\theta}^{\mathrm{MLccE}}_{K-1}$ is available. Then, choose one of the classes designed (through MAP) from $\widehat{\theta}^{\mathrm{MLccE}}_{K-1}$ (say, the $k_0{}^{\mathrm{th}}$) and divide it into two classes by applying to the corresponding observations the $\mathrm{L_{cc}}$-EM algorithm with a two-component Gaussian mixture model. Now, $\theta^{0k_0}_{\mathrm{Km1}}$ is built by keeping the parameters of components with label different from k_0 as in $\widehat{\theta}^{\mathrm{MLccE}}_{K-1}$ and by introducing the parameters obtained by splitting k_0 for the k_0^{th} and the K^{th} components. Run a few iterations of $\mathrm{L_{cc}}$-EM and get $\theta^{1k_0}_{\mathrm{Km1}}$.

Apply the same procedure with any k_0. Then, apply $\mathrm{L_{cc}}$-EM to the parameter which maximizes $\log \mathrm{L_{cc}}$ among $\{\theta^{11}_{\mathrm{Km1}}, \ldots, \theta^{1K-1}_{\mathrm{Km1}}\}$ and get θ^0_{Km1}.

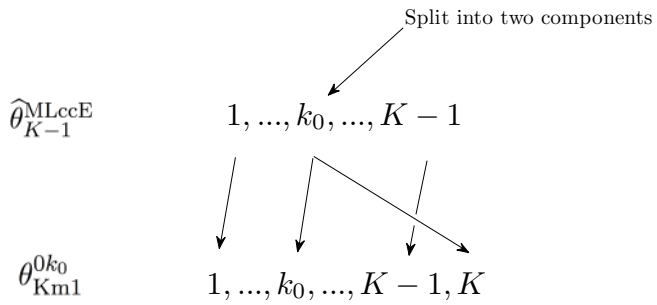


FIG 4. *The Km1 Strategy.*

This strategy can be related to the *X-means* algorithm of [33], which is an extension to the *k-means*. An adaptation of it to the *Alter* algorithm, which relies on quantization techniques, has been introduced by [24].

The Km1 method works well in practice, particularly as the number of components is larger than the sensible number of classes. When the number of components is low it is not always able to find a sensible solution since it can be far from the one with one less component. This is why we do not recommend to use Km1 as the only initialization method. Initializing L_{cc}-EM from the best solution among θ^0_{small} and θ^0_{Km1} provides sensible enough results for the application of the slope heuristics. As expected, θ^0_{small} is often more sensible than θ^0_{Km1} for small values of K and the situation is reversed for large values of K.

Let us remark that the Km1 procedure can be applied for the EM algorithm all the same and does not seem to correspond to any known procedure. It helps improving the results in particular for the slope heuristics, since models with high numbers of components have to be involved. Km1 is longer to run than small_EM though: it can be expected to be at worst K times longer (depending on both procedures parameters). Figure 5 illustrates the difference between both

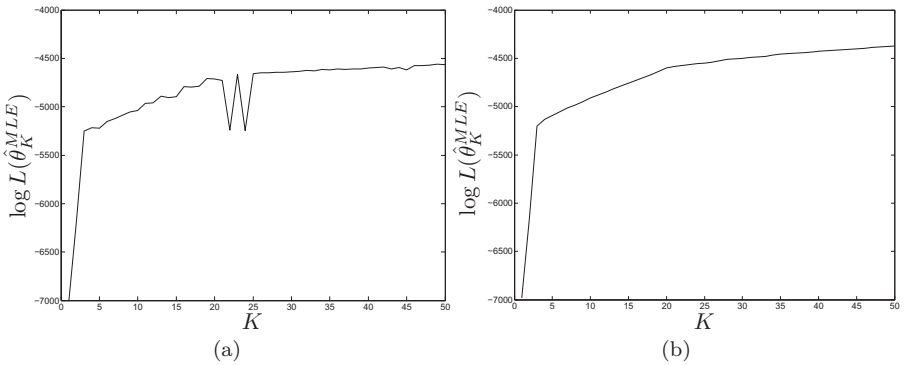


FIG 5. *Toy Dataset: Optimization of the Likelihood for Each Model. (a) Initialization Without Km1. (b) Initialization With Km1.*

procedures results (i.e. with or without Km1) on a toy dataset: we stress that the likelihood values reached with Km1 are higher than those obtained without it; that Km1 avoids some severe failures to find a good maximizer of the likelihood ($K = 21$ and $K = 24$) and that this can obviously be crucial so as to get a good estimation of the slope of the linear part of the graph.

6. Simulations

6.1. Consistency of the MLccE

Example 6.1 (Example 2.1 continued)**.** The consistency of $\widehat{\theta}^{\mathrm{MLccE}}$ stated by Theorem 3.1 can be empirically observed for Example 2.1: see Figure 6.

6.2. Model selection criteria

For each simulation setting, at least 100 datasets have been simulated (details available in [5, Section A.2]). $\widehat{\theta}_K^{\mathrm{MLE}}$ and $\widehat{\theta}_K^{\mathrm{MLccE}}$ have been computed for each $K \in \{1, \ldots, K_M\}$ and the percentage of selection of each possible number of classes is reported for each one of the following criteria:

- $\mathrm{crit}_{\mathrm{AIC}}(K) = \log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}}) - D_K$;
- $\mathrm{crit}_{\mathrm{BIC}}(K) = \log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}}) - \frac{\log n}{2} D_K$;
- $\mathrm{crit}_{\mathrm{SHL}}(K) = \log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}}) - 2 \times \widehat{\mathrm{slope}}_L \times D_K$ (Slope Heuristics applied to $\big(D_K, \log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}})\big)_{K \in \{1,\ldots,K_M\}}$);
- $\mathrm{crit}_{\mathrm{ICL}}(K) = \log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLE}}) - \frac{\log n}{2} D_K$;
- $\mathrm{crit}_{\mathrm{L_{cc}-ICL}}(K) = \log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}) - \frac{\log n}{2} D_K$;
- $\mathrm{crit}_{\mathrm{SHL_{cc}}}(K) = \log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}) - 2 \times \widehat{\mathrm{slope}}_{Lcc} \times D_K$ (Slope Heuristics applied to $\big(D_K, \log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}})\big)_{K \in \{1,\ldots,K_M\}}$).

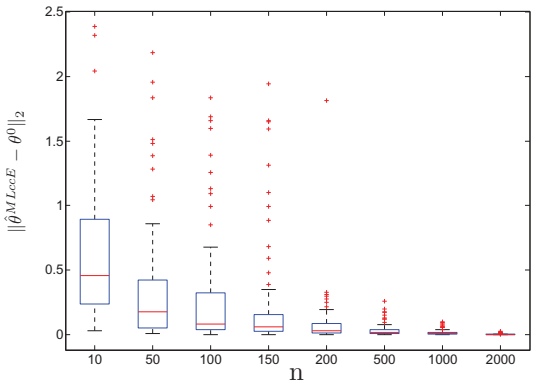


FIG 6. $\|\widehat{\theta}^{MLccE} - \theta^0\|_2$ *boxplots for* 100 *experiences with respect to* n *(Example 2.1 setting).*

According to the slope heuristics $\widehat{\text{slope}}_L$ (resp. $\widehat{\text{slope}}_{Lcc}$) is the estimated slope of the "linear part" of the graph $D_K \mapsto -\log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}})$ (resp. $D_K \mapsto -\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}})$), which occurs for large values of K.

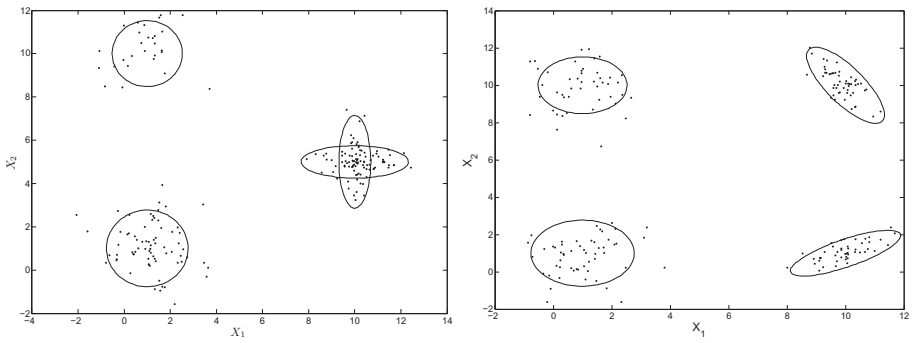
6.2.1. The "Cross" experiment

$f^{\wp}$ is a four-component Gaussian mixture in $\mathbb{R}^2$, $n = 200$ (Figure 7). Diagonal models are fitted: the true distribution is available in the model with four components.

According to Table 1, AIC clearly overestimates the number of classes; the criteria based on $\log \mathrm{L}$ (BIC, SHL) select four classes: one for each Gaussian component needed to fit the data; the criteria associated to $\log \mathrm{L}_{\mathrm{cc}}$ (L_{cc}-ICL, ICL and $\mathrm{SHL}_{\mathrm{cc}}$) select three classes: the two components of the "cross" overlap too much to be considered as two separated classes.

The expected oracle number of components for $\log \mathrm{L}$ ($\operatorname{argmin}_{1\leq K\leq 20} \mathbb{E}[d_{\mathrm{KL}}(f^{\wp}, f(.;\widehat{\theta}_K^{\mathrm{MLE}}))]$) is four and for $\log \mathrm{L}_{\mathrm{cc}}$ ($\operatorname{argmin}_{1\leq K\leq 8} \mathbb{E}_{X_1,\ldots,X_n}\mathbb{E}_X[-\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}}; X)]$) it is three.

The solutions $\widehat{\theta}_4^{\mathrm{MLE}}$ and $\widehat{\theta}_4^{\mathrm{MLccE}}$ for an example dataset (chosen for its illustrative quality) have been represented (Figure 8). Remark that $\widehat{\theta}_4^{\mathrm{MLccE}}$ does not nearly match the true distribution, on the contrary to $\widehat{\theta}_4^{\mathrm{MLE}}$. The former dislikes solutions with overlapping components.



"Cross" Experiment. "Misspecified" Experiment

FIG 7. *Simulated datasets with isodensities of* $f^{\wp}$.

TABLE 1
"Cross" Experiment. Blank cells: L_{cc}*-based criteria computed for* $K \in \{1,\ldots,8\}$

Selected Number of Components	2	3	4	5	6	7	8	9	10–20
AIC	0	0	1	1	2	2	3	3	88
BIC	0	4	91	5	0	0	0	0	0
SHL	0	2	84	10	3	0	0	0	1
ICL	0	96	3	1	0	0	0	0	0
L_{cc}-ICL	0	99	1	0	0	0	0		
$\mathrm{SHL}_{\mathrm{cc}}$	2	79	8	8	3	0	0		

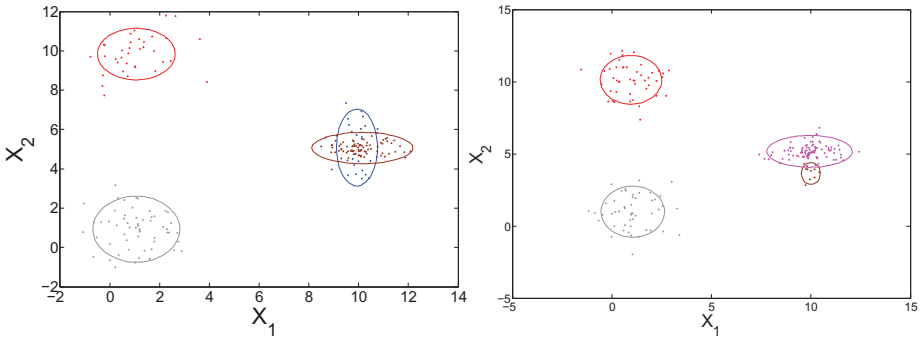


FIG 8. *"Cross" Experiment. MAP classifications.*

TABLE 2
"Cross" Experiment. Risk of each criterion in terms of Kullback-Leibler divergence to the true distribution, estimated by Monte Carlo simulations

	Risk $\times 10^3$	$\frac{\text{Risk of the criterion}}{\text{Risk of the oracle}}$
Oracle	59	1
AIC	506	8.03
BIC	65	1.10
(ICL)	156	2.62
SHL	69	1.17

TABLE 3
"Cross" Experiment. "Risk" of each criterion for the L_{cc} contrast, estimated by Monte Carlo simulations

	"Risk" $\times 10^3$
Oracle	3618
ICL	3622
L_{cc}-ICL	3623
SHL_{cc}	3632

Tables 2 and 3 compare the risk of each criterion with the corresponding trajectory oracle ($K_{\text{oracle}} = \text{argmin}_{1\leq K\leq 20} d_{\text{KL}}(f^{\wp}, f(\,.\,;\widehat{\theta}_K^{\text{MLE}}))$ for $\log \text{L}$ and $K_{\text{oracle}} = \text{argmin}_{1\leq K\leq 8} \mathbb{E}_X[-\log \text{L}_{\text{cc}}(\widehat{\theta}_K^{\text{MLccE}}; X)]$ for $\log \text{L}_{\text{cc}}$). Remark that ICL should rather be compared to the L_{cc} oracle.

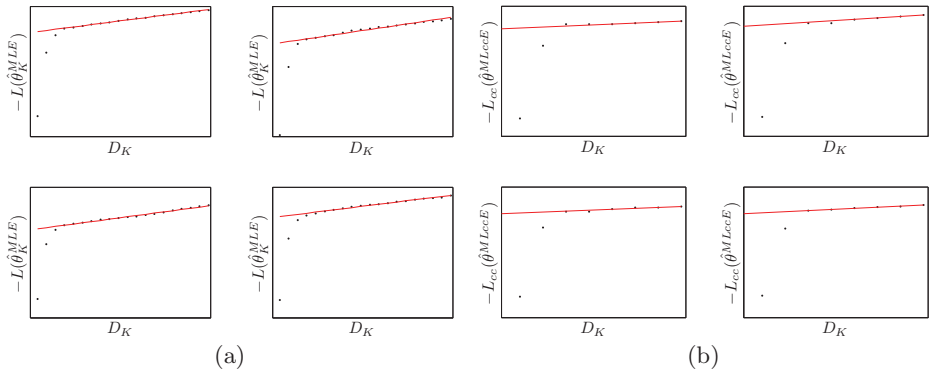


FIG 9. *"Cross" Experiment. Examples of* $D_K \mapsto \log L(\widehat{\theta}_K^{MLE})$ *(a) and of* $D_K \mapsto \log L_{cc}(\widehat{\theta}_K^{MLccE})$ *(b) graphs.*

A few graphs $D_K \mapsto \log \mathrm{L}(\widehat{\theta}_K^{\mathrm{MLE}})$ and $D_K \mapsto \log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}_K^{\mathrm{MLccE}})$ are plotted (Figure 9) to check that linear parts occur for large values of K. This is a necessary condition for the slope heuristics to be confidently applied.

6.2.2. Misspecified models

$f^{\wp}$ is a four-component Gaussian mixture in $\mathbb{R}^2$, $n = 200$ (Figure 7). The two left-hand side components are diagonal, but the two others are not. Diagonal models are fitted: this experiment illustrates a misspecified models situation.

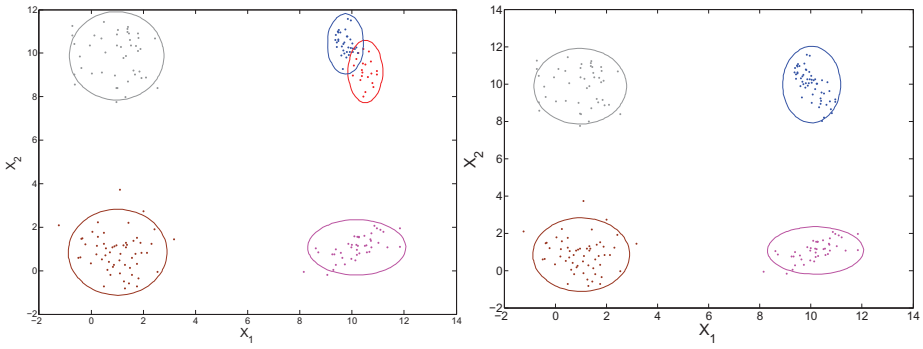
From Table 4, BIC tends to select quite a high number of components. Indeed, the number of diagonal components needed to approximate $f^{\wp}$ is larger than four, because of the two non-diagonal components. See Figure 10 for an example. SHL yields the selections of $\hat{K}$ the closest to the oracle's (Table 4). However, it does not yield better risk results than BIC (Table 5): both get good results. ICL and L_{cc}-ICL select the expected four classes half of the time. The number of observations does not always enable them to decide that some components of the five- or six-component fitted solution overlap. $\mathrm{SHL}_{\mathrm{cc}}$ reaches the "best" results (from the clustering point of view), in the sense that it recovers the expected four classes the most often.

The expected oracle number of components is six or seven for $\log \mathrm{L}$ and four for $\log \mathrm{L}_{\mathrm{cc}}$.

TABLE 4

"Misspecified" Experiment. Blank cells: L_{cc}-based criteria computed for $K \in \{1, \ldots, 8\}$

Selected number of components	4	5	6	7	8	9–16	17	18	19	20
Oracle (for $\log \mathrm{L}$)	4	10	30	43	12	1	0	0	0	0
AIC	0	0	0	0	0	20	14	12	26	28
BIC	3	43	38	13	3	0	0	0	0	0
SHL	2	19	26	32	11	10	0	0	0	0
ICL	49	35	9	5	2	0	0	0	0	0
L_{cc}-ICL	54	29	13	4	0					
$\mathrm{SHL}_{\mathrm{cc}}$	81	17	2	0	0					



MLE ($\hat{K}^{\mathrm{BIC}} = 5$) MLccE ($\hat{K}^{\mathrm{ICL}} = 4$)

FIG 10. *"Misspecified" Experiment. MAP classifications.*

TABLE 5
"Misspecified" Experiment. Risk of each criterion in terms of Kullback-Leibler divergence to the true distribution, estimated by Monte Carlo simulations

	Risk $\times 10^3$	$\frac{\text{Risk of the criterion}}{\text{Risk of the oracle}}$
Oracle	206	1
AIC	712	3.45
BIC	240	1.16
(ICL)	272	1.32
SHL	249	1.21

TABLE 6
"Misspecified" Experiment. "Risk" of each criterion for the L_{cc} contrast, estimated by Monte Carlo simulations

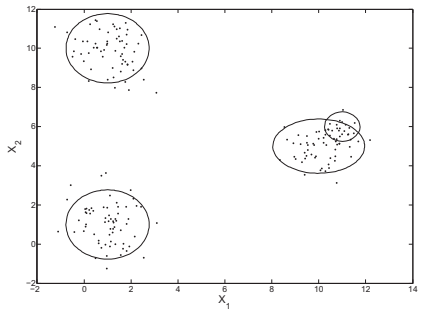
	"Risk" $\times 10^3$
Oracle	3910
ICL	3926
L_{cc}-ICL	3928
SHL_{cc}	3915

6.2.3. Distorted component

$f^{\wp}$ is a four-component Gaussian mixture in $\mathbb{R}^2$, $n = 200$ (Figure 11). The fourth is smaller than the others (size: $\pi = 0.1$ against 0.3 and volume: $\det \Sigma = 0.01$ against 1 or 0.5). Diagonal mixture models are fitted: the true distribution is available for $K = 4$.

From the table in Figure 11, BIC and SHL mostly recover the four Gaussian components. This is what they are expected to do. ICL, L_{cc}-ICL and SHL_{cc} mostly select three classes, as expected too.

The expected oracle number of components is four for $\log L$ and three for $\log L_{cc}$.



K	3	4	5	6	7	8
AIC	0	24	30	23	3	20
BIC	42	57	0	0	0	1
SHL	22	67	10	1	0	0
ICL	93	7	0	0	0	0
L_{cc}-ICL	98	2	0	0	0	0
SHL_{cc}	78	17	4	1	0	0

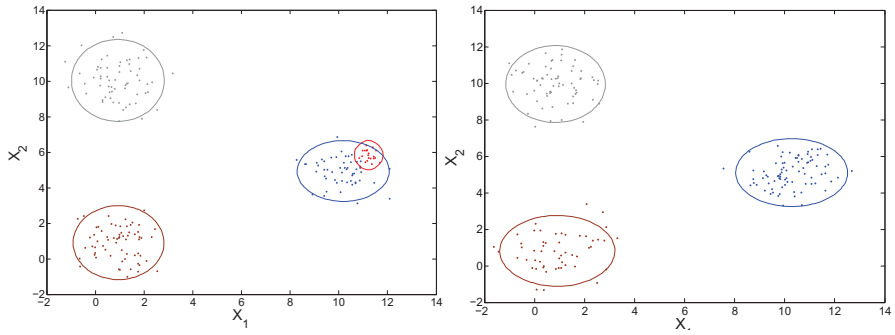
FIG 11. *"Distorted" Experiment. Simulated dataset with isodensity of $f^{\wp}$ (left) and Selected numbers of components for each criterion (right).*

TABLE 7
"Distorted" Experiment. Risk of each criterion in terms of Kullback-Leibler divergence to the true distribution, estimated by Monte Carlo simulations

	Risk $\times 10^3$	$\frac{\text{Risk of the criterion}}{\text{Risk of the oracle}}$
Oracle	58.3	1
AIC	108.5	1.9
BIC	73.7	1.3
(ICL)	99.9	1.7
SHL	68.0	1.2

TABLE 8
"Distorted" Experiment. "Risk" of each criterion for the L_{cc} contrast, estimated by Monte Carlo simulations

	"Risk" $\times 10^3$
Oracle	3857
ICL	3859
L_{cc}-ICL	3857
SHL_{cc}	3863



MLE ($\hat{K}^{\mathrm{BIC}} = 4$) MLccE ($\hat{K}^{\mathrm{ICL}} = 3$)

FIG 12. *"Distorted" Experiment. MAP classifications.*

7. Discussion

The Simulations section illustrates what the theoretical works suggest: BIC and ICL belong to two different families of criteria. Before the choice of the penalty, the most decisive choice is the contrast. The purpose of L-based criteria (BIC, SHL...) is of a different nature than that of $\mathrm{L_{cc}}$-criteria ($\mathrm{L_{cc}}$-ICL, SHLcc...). The former rely on a density estimation approach and thus the assumption that the conditional distributions of the classes are Gaussian is strong: each group of observations that requires a Gaussian component to be fitted deserves to be considered as a class. The latter aim at discovering classes matching a subtle compromise between the notions of well separated clusters and Gaussian-shaped classes: observations which cannot be confidently discriminated from each other, should belong to the same class. They enjoy the flexibility and modeling possibilities of the model-based clustering approach and still do not break an expected notion of cluster.

Once a contrast is chosen, then a penalty has to be chosen. We want to stress that the choice among criteria based on a same contrast, though important, is a second order choice. As expected, BIC and ICL (regarded as an approximation of $\mathrm{L_{cc}}$-ICL) in their respective families of criteria, perform pretty well, while being quite easy to run, notably as compared to the slope heuristics methods. But the latter are interesting for their data-driven and non-asymptotic properties, at least in some experiments.

As a practical conclusion a user should rather consider ICL if he is interested in finding well-separated clusters, which conditional distribution can possibly not clearly be Gaussian-shaped, whereas a user who is confident that the conditional distribution of the components are Gaussian and for whom this is not a problem to get overlapping clusters should rather consider BIC. [8] propose a way to get the most of both criteria.

The first aim of this study was to better understand the ICL criterion. A step further in this direction now means better understanding the contrast $\mathrm{L_{cc}}$.

A further theoretical step would be to derive non-asymptotic results and oracle inequalities. This should give more precise insights about the optimal penalty shape to use, and then justify the use of the slope heuristics of [13] (see also [7] or [5] for partial results, simulations, and discussions on this).

It should also be further studied how the complexity of the models should be measured, particularly when several kinds of models are compared. The dimension of the model as a parametric space works for the reported theoretical results. But we are not convinced that it is the finest measure of the complexity of Gaussian mixture models. As a matter of fact this simple parametric point of view amounts to considering that all parameters play an analogous role. This is not really intuitive.

Finally, let us stress that the implementation of the considered procedures is quite challenging. Some work has been done in this direction already: see [6] and Section 5 above. But the resulting algorithms are longer to run than standard EM procedures up to now: further work in that direction is necessary and ICL in the standard MLE version can be a good practical compromise for now.

8. Proofs

8.1. Proof of Theorem 3.1

Proof of Theorem 3.1. Let $\varepsilon > 0$. Since $\theta \mapsto \mathbb{E}\left[-\log \mathrm{L}_{\mathrm{cc}}(\theta)\right]$ is continuous and $\{\theta \in \Theta : d(\theta, \Theta^0) \geq \varepsilon\}$ is compact, $\eta = \frac{1}{n}\mathbb{E}\left[\log \mathrm{L}_{\mathrm{cc}}(\theta^0)\right] - \frac{1}{n}\inf_{\theta\in\Theta: d(\theta,\Theta^0)\geq\varepsilon} \mathbb{E}\left[\log \mathrm{L}_{\mathrm{cc}}(\theta)\right] > 0$. Since from Lemma 3.2 (under the assumption $H'_{\log \mathrm{L}_{\mathrm{cc}},\Theta^{\mathcal{O}},1}$ which holds since the components conditional probabilities are kept away from zero by the assumption that Θ is compact and contains no zero-proportion component, as discussed p. 1051) and Theorem 3.2, $\{-\log \mathrm{L}_{\mathrm{cc}}(\theta;\cdot) : \theta \in \Theta\}$ is $f^{\wp}\cdot\lambda$–Glivenko-Cantelli on the one hand; by the definition of $\widehat{\theta}^{\mathrm{MLccE}}$ on the other hand, with high probability for n large enough,

$$\frac{1}{n}\sup_{\theta\in\Theta}\left|-\log \mathrm{L}_{\mathrm{cc}}(\theta) - \mathbb{E}\left[-\log \mathrm{L}_{\mathrm{cc}}(\theta)\right]\right| < \frac{\eta}{3}$$
$$\frac{1}{n}\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}^{\mathrm{MLccE}}) \geq \frac{1}{n}\log \mathrm{L}_{\mathrm{cc}}(\theta^0) - \frac{\eta}{3}$$

for any $\theta^0 \in \Theta^0$ and then

$$\begin{aligned}\mathbb{E}\left[\log \mathrm{L}_{\mathrm{cc}}(\theta^0)\right] &- \mathbb{E}\left[\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}^{\mathrm{MLccE}})\right]\\ &= \left(\mathbb{E}\left[\log \mathrm{L}_{\mathrm{cc}}(\theta^0)\right] - \log \mathrm{L}_{\mathrm{cc}}(\theta^0)\right) + \left(\log \mathrm{L}_{\mathrm{cc}}(\theta^0) - \log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}^{\mathrm{MLccE}})\right)\\ &\quad + \left(\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}^{\mathrm{MLccE}}) - \mathbb{E}\left[\log \mathrm{L}_{\mathrm{cc}}(\widehat{\theta}^{\mathrm{MLccE}})\right]\right)\\ &< n\eta,\end{aligned}$$

hence

$$\forall \varepsilon > 0, \forall u \in [0,1], \exists N \in \mathbb{N} \text{ s.t. } \forall n \geq N, \mathbb{P}\left(d(\widehat{\theta}^{\mathrm{MLccE}}, \Theta^0) < \varepsilon\right) \geq u,$$

i.e. $d(\widehat{\theta}^{\mathrm{MLccE}}, \Theta^0) \xrightarrow{\mathbb{P}} 0$. □

8.2. Proofs of the bracketing entropy upper-bounds

Proof of Lemma 3.1. This is a calculation which relies on the mean value theorem, hence the convexity assumption. Let $\varepsilon > 0$, and $\widetilde{\Theta} \subset \Theta$, with $\widetilde{\Theta}$ bounded. Let $\widetilde{\Theta}_\varepsilon$ be a grid in Θ which "ε-covers" $\widetilde{\Theta}$ in any dimension with step ε. $\widetilde{\Theta}_\varepsilon$ is for example $\widetilde{\Theta}_\varepsilon^1 \times \cdots \times \widetilde{\Theta}_\varepsilon^D$ with

$$\forall i \in \{1, \dots, D\}, \widetilde{\Theta}_\varepsilon^i = \left\{\widetilde{\theta}_{\min}^i, \widetilde{\theta}_{\min}^i + \varepsilon, \dots, \widetilde{\theta}_{\max}^i\right\},$$

where

$$\forall i \in \{1, \dots, D\}, \left\{\theta^i : \theta \in \widetilde{\Theta}\right\} \subset \left[\widetilde{\theta}_{\min}^i - \frac{\varepsilon}{2}, \widetilde{\theta}_{\max}^i + \frac{\varepsilon}{2}\right].$$

This is always possible since Θ is convex. With the $\|\cdot\|_\infty$ norm, the step of the grid $\widetilde{\Theta}_\varepsilon$ is the same as the step over each dimension, ε:

$$\forall \tilde{\theta} \in \widetilde{\Theta}, \exists \tilde{\theta}_\varepsilon \in \widetilde{\Theta}_\varepsilon / \|\tilde{\theta} - \tilde{\theta}_\varepsilon\|_\infty \leq \frac{\varepsilon}{2}.$$

And the cardinal of $\widetilde{\Theta}_\varepsilon$ is at most

$$\prod_{i=1}^{D} \frac{(\sup_{\theta \in \widetilde{\Theta}} \theta^i - \inf_{\theta \in \widetilde{\Theta}} \theta^i)}{\varepsilon} \vee 1 \leq \left(\frac{\operatorname{diam} \widetilde{\Theta}}{\varepsilon}\right)^D \vee 1.$$

Now, let θ_1 and θ_2 in Θ and $x \in \mathbb{R}^d$.

$$\begin{aligned} \left|\gamma(\theta_1; x) - \gamma(\theta_2; x)\right| &\leq \sup_{\theta \in [\theta_1; \theta_2]} \left\| \left(\frac{\partial \gamma}{\partial \theta}\right)_{(\theta; x)} \right\|_\infty \|\theta_1 - \theta_2\|_\infty \\ &\leq \underbrace{\sup_{\theta \in \Theta} \left\| \left(\frac{\partial \gamma}{\partial \theta}\right)_{(\theta; x)} \right\|_\infty}_{M'(x)} \|\theta_1 - \theta_2\|_\infty, \end{aligned}$$

since Θ is convex. Let $\tilde{\theta} \in \widetilde{\Theta}$ and choose $\tilde{\theta}_\varepsilon \in \widetilde{\Theta}_\varepsilon$ such that $\|\tilde{\theta} - \tilde{\theta}_\varepsilon\|_\infty \leq \frac{\varepsilon}{2}$. Then

$$\forall x \in \mathbb{R}^d, \left|\gamma(\tilde{\theta}_\varepsilon; x) - \gamma(\tilde{\theta}; x)\right| \leq M'(x)\frac{\varepsilon}{2}$$

and

$$\gamma(\tilde{\theta}_\varepsilon; x) - \frac{\varepsilon}{2} M'(x) \leq \gamma(\tilde{\theta}; x) \leq \gamma(\tilde{\theta}_\varepsilon; x) + \frac{\varepsilon}{2} M'(x).$$

The set of $\varepsilon \|M'\|_r$-brackets (for the $\|\cdot\|_r$-norm)

$$\left\{[\gamma(\tilde{\theta}_\varepsilon) - \frac{\varepsilon}{2} M'; \gamma(\tilde{\theta}_\varepsilon) + \frac{\varepsilon}{2} M'] : \tilde{\theta}_\varepsilon \in \widetilde{\Theta}_\varepsilon\right\}$$

then has cardinal at most $(\frac{\operatorname{diam} \widetilde{\Theta}}{\varepsilon})^D \vee 1$ and covers $\{\gamma(\tilde{\theta}) : \tilde{\theta} \in \widetilde{\Theta}\}$. □

Example 8.1 (Diagonal Gaussian Mixture Model Parameter Space is Convex)**.** Following [15], we write $[p\lambda_k B_k]$ for the model of Gaussian mixtures with diagonal covariance matrices and equal mixing proportions. To keep simple notation, let us consider the case $d = 2$ and $K = 2$ ($d = 1$ or $K = 1$ are obviously particular cases!). A natural parametrization of this model (which dimension is 8) is

$$\theta \in \mathbb{R}^4 \times \mathbb{R}^{+*4} \overset{\varphi}{\longmapsto} \frac{1}{2}\phi\left(\,.\,; \begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}, \begin{pmatrix}\theta_5 & 0\\ 0 & \theta_6\end{pmatrix}\right) + \frac{1}{2}\phi\left(\,.\,; \begin{pmatrix}\theta_3\\ \theta_4\end{pmatrix}, \begin{pmatrix}\theta_7 & 0\\ 0 & \theta_8\end{pmatrix}\right)$$

Then $[p\lambda_k B_k] = \varphi(\mathbb{R}^4 \times \mathbb{R}^{+*4})$ and the parameter space $\mathbb{R}^4 \times \mathbb{R}^{+*4}$ is convex.

Example 8.2 (The Same Model with Equal Volumes is Convex, too...)**.** $[p\lambda B_k]$ is the same model as in the previous example, but the covariance matrices determinants have to be equal. With $d = 2$ and $K = 2$, a natural parametrization of this model of dimension 7 is

$$\theta \in \mathbb{R}^4 \times \mathbb{R}^{+*3} \overset{\varphi}{\longmapsto} \frac{1}{2}\phi\left(\,.\,; \begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}, \sqrt{\theta_7}\begin{pmatrix}\theta_5 & 0\\ 0 & \frac{1}{\theta_5}\end{pmatrix}\right) + \frac{1}{2}\phi\left(\,.\,; \begin{pmatrix}\theta_3\\ \theta_4\end{pmatrix}, \sqrt{\theta_7}\begin{pmatrix}\theta_6 & 0\\ 0 & \frac{1}{\theta_6}\end{pmatrix}\right)$$

Then $[p\lambda B_k] = \varphi(\mathbb{R}^4 \times \mathbb{R}^{+*3})$ and the parameter space $\mathbb{R}^4 \times \mathbb{R}^{+*3}$ is convex.

Proof of Lemma 3.2. Θ is not supposed to be convex as in Lemma 3.1 but since it is compact it can be covered with a finite number Q of open balls. Let $O_1, \ldots, O_Q$ be such a covering of Θ consisting of open balls such that $\Theta \subset \cup_{q=1}^{Q} O_q \subset \Theta^{\mathcal{O}}$. Remark that

$$\Theta = \cup_{q=1}^{Q}(O_q \cap \Theta) \subset \cup_{q=1}^{Q}\mathrm{conv}(O_q \cap \Theta).$$

Now, for any q, $\mathrm{conv}(O_q \cap \Theta)$ is convex and $\sup_{\theta\in\mathrm{conv}(O_q\cap\Theta)} \|(\frac{\partial\gamma}{\partial\theta})_{(\theta;x)}\|_\infty \leq M'(x)$ since $\mathrm{conv}(O_q \cap \Theta) \subset O_q \subset \Theta^{\mathcal{O}}$ (O_q may not be included in Θ, hence the introduction of $\Theta^{\mathcal{O}}$ in the assumptions of this lemma). Therefore, for any $\widetilde{\Theta} \subset \Theta$, Lemma 3.1 applies to $O_q \cap \widetilde{\Theta} \subset \mathrm{conv}(O_q \cap \Theta)$:

$$\forall \varepsilon > 0,\ N_{[]}(\varepsilon, \{\gamma(\theta) : \theta \in \widetilde{\Theta} \cap O_q\}, \|\cdot\|_r) \leq \left(\frac{\|M'\|_r\ \mathrm{diam}\,\widetilde{\Theta}}{\varepsilon}\right)^D \vee 1.$$

Since $N_{[]}(\varepsilon, \{\gamma(\theta) : \theta \in \widetilde{\Theta}\}, \|\cdot\|_r) \leq N_{[]}(\varepsilon, \cup_{q=1}^{Q}\{\gamma(\theta) : \theta \in \widetilde{\Theta} \cap O_q\}, \|\cdot\|_r)$, the result follows. □

Proof of Lemma 3.3. Consider the grid $\widetilde{\Theta}_\varepsilon$ of the proof of Lemma 3.1. Let θ_1 and θ_2 in Θ and $x \in \mathbb{R}^d$. Since Θ is convex,

$$\begin{aligned}\left|\gamma(\theta_1;x) - \gamma(\theta_2;x)\right|^r &\leq \sup_{\theta\in[\theta_1;\theta_2]}\left\|\left(\frac{\partial\gamma}{\partial\theta}\right)_{(\theta;x)}\right\|_\infty^2 \|\theta_1 - \theta_2\|_\infty^2 \left(2\sup_{\theta\in\{\theta_1,\theta_2\}}|\gamma(\theta;x)|\right)^{r-2}\\ &\leq M'(x)^2\|\theta_1 - \theta_2\|_\infty^2(2\|M\|_\infty)^{r-2} \quad f^{\wp}d\lambda\text{-a.e.}\end{aligned}$$

Let $\tilde{\theta} \in \widetilde{\Theta}$ and choose $\tilde{\theta}_\varepsilon \in \widetilde{\Theta}_\varepsilon$ such that $\|\tilde{\theta} - \tilde{\theta}_\varepsilon\|_\infty \leq \frac{\varepsilon}{2}$. Then

$$\left|\gamma(\tilde{\theta}_\varepsilon; x) - \gamma(\tilde{\theta}; x)\right| \leq M'(x)^{\frac{2}{r}} \left(\frac{\varepsilon}{2}\right)^{\frac{2}{r}} (2\|M\|_\infty)^{\frac{r-2}{r}} \quad f^{\wp}\text{-a.e.}$$

and the set of brackets

$$\left\{\left[\gamma(\tilde{\theta}_\varepsilon; x) - \varepsilon^{\frac{2}{r}} M'(x)^{\frac{2}{r}} \|M\|_\infty^{\frac{r-2}{r}} 2^{1-\frac{4}{r}}; \gamma(\tilde{\theta}_\varepsilon; x) + \varepsilon^{\frac{2}{r}} M'(x)^{\frac{2}{r}} \|M\|_\infty^{\frac{r-2}{r}} 2^{1-\frac{4}{r}}\right] : \tilde{\theta} \in \widetilde{\Theta}_\varepsilon\right\}$$

(of $\|\cdot\|_r$-norm length $(2^{2-\frac{4}{r}})\|M\|_\infty^{\frac{r-2}{r}} \|M'\|_2^{\frac{2}{r}} \varepsilon^{\frac{2}{r}}$) has cardinal at most $(\frac{\operatorname{diam} \widetilde{\Theta}}{\varepsilon})^D \vee 1$ and covers $\{\gamma(\tilde{\theta}) : \tilde{\theta} \in \widetilde{\Theta}\}$, which yields Lemma 3.3. □

8.3. *Proof of Theorem 4.1*

Theorem 4.1 is an application of Theorem 8.1, written for a general contrast and family of models:

Theorem 8.1. *$\{\Theta_K\}_{1 \leq K \leq K_M}$ a collection of models with $\Theta_K \subset \mathbb{R}^{D_K}$ ($D_1 \leq \cdots \leq D_{K_M}$) and let $\theta_K^0 \in \Theta_K^0$, with $\Theta_K^0 = \operatorname{argmin}_{\theta \in \Theta_K} \mathbb{E}[\gamma(\theta)]$. Assume*

$$K_0 = \min \operatorname*{argmin}_{1 \leq K \leq K_M} \mathbb{E}\left[\gamma(\Theta_K^0)\right] \tag{B1}$$

$$\forall K,\ \hat{\theta}_K \in \Theta_K \text{ is such that } \gamma_n(\hat{\theta}_K) \leq \gamma_n(\theta_K^0) + o_{\mathbb{P}}(1) \text{ and fulfills } \gamma_n(\hat{\theta}_K) \xrightarrow{\mathbb{P}} \mathbb{E}\left[\gamma(\theta_K^0)\right] \tag{B2}$$

$$\forall K,\ \begin{cases} \operatorname{pen}(K) > 0 \text{ and } \operatorname{pen}(K) = o_{\mathbb{P}}(1) & \text{when } n \to +\infty \\ n\big(\operatorname{pen}(K) - \operatorname{pen}(K')\big) \xrightarrow[n \to +\infty]{\mathbb{P}} \infty & \text{when } K > K' \end{cases} \tag{B3}$$

$$n\big(\gamma_n(\hat{\theta}_{K_0}) - \gamma_n(\hat{\theta}_K)\big) = O_{\mathbb{P}}(1) \text{ for any } K \in \operatorname*{argmin}_{1 \leq K \leq K_M} \mathbb{E}\left[\gamma(\Theta_K^0)\right]. \tag{B4}$$

Define $\hat{K}$ such that

$$\hat{K} = \min \operatorname*{argmin}_{1 \leq K \leq K_M} \Big\{ \underbrace{\gamma_n(\hat{\theta}_K) + \operatorname{pen}(K)}_{\operatorname{crit}(K)} \Big\}.$$

Then $\mathbb{P}[\hat{K} \neq K_0] \xrightarrow[n \to \infty]{} 0$.

Assumption (B3) defines the range of possible penalties. Regarding (B2),

Lemma 8.1. *For a fixed K, assume $\sup_{\theta \in \Theta_K} \left|\gamma_n(\theta) - \mathbb{E}[\gamma(\theta)]\right| \xrightarrow{\mathbb{P}} 0$. Then (B2) holds.*

Proof of Lemma 8.1. For any $\varepsilon > 0$, with high probability for n large enough:

$$\underbrace{\gamma_n(\hat{\theta}) - \mathbb{E}_{f^{\wp}}\left[\gamma(\hat{\theta})\right]}_{\geq -\varepsilon} + \underbrace{\mathbb{E}_{f^{\wp}}\left[\gamma(\hat{\theta})\right] - \mathbb{E}\left[\gamma(\theta^0)\right]}_{\geq 0} = \gamma_n(\hat{\theta}) - \mathbb{E}\left[\gamma(\theta^0)\right]$$
$$= \underbrace{\gamma_n(\hat{\theta}) - \gamma_n(\theta^0)}_{\leq \varepsilon} + \underbrace{\gamma_n(\theta^0) - \mathbb{E}\left[\gamma(\theta^0)\right]}_{\leq \varepsilon}.$$

□

Section 8.4 is devoted to deriving sufficient conditions so that (B4) holds.

Theorem 4.1 is a direct consequence of Theorem 8.1, Lemma 8.1, Theorem 3.1, and Corollary 8.2 (below).

Proof of Theorem 8.1. Let $\mathcal{K} = \operatorname{argmin}_{1 \leq K \leq K_M} \mathbb{E}\left[\gamma(\theta_K^0)\right]$. $K_0 = \min \mathcal{K}$.

It is first proved that $\hat{K}$ does not asymptotically "underestimate" K_0. Let $K \notin \mathcal{K}$. Let $\varepsilon = \frac{1}{2}\left(\mathbb{E}\left[\gamma(\theta_K^0)\right] - \mathbb{E}\left[\gamma(\theta_{K_0}^0)\right]\right) > 0$. From (B2) and (B3) ($\operatorname{pen}(K) = o_{\mathbb{P}}(1)$), with high probability for n large enough:

$$\left|\gamma_n(\hat{\theta}_K) - \mathbb{E}\left[\gamma(\theta_K^0)\right]\right| \leq \frac{\varepsilon}{3}; \quad \left|\gamma_n(\hat{\theta}_{K_0}) - \mathbb{E}\left[\gamma(\theta_{K_0}^0)\right]\right| \leq \frac{\varepsilon}{3}; \quad \operatorname{pen}(K_0) \leq \frac{\varepsilon}{3}.$$

Then

$$\operatorname{crit}(K) = \gamma_n(\hat{\theta}_K) + \operatorname{pen}(K) \geq \mathbb{E}\left[\gamma(\theta_K^0)\right] - \frac{\varepsilon}{3} + 0$$
$$= \mathbb{E}\left[\gamma(\theta_{K_0}^0)\right] + \frac{5\varepsilon}{3} \geq \underbrace{\gamma_n(\hat{\theta}_{K_0}) + \operatorname{pen}(K_0)}_{\operatorname{crit}(K_0)} + \varepsilon.$$

Then, with high probability for n large enough, $\hat{K} \neq K$.

Let now $K \in \mathcal{K}$ (hence $K > K_0$). Assumption (B4) implies that $\exists V > 0$ such that with high probability for n large enough,

$$n\left(\gamma_n(\hat{\theta}_{K_0}) - \gamma_n(\hat{\theta}_K)\right) \leq V.$$

Increase n enough so that $n\left(\operatorname{pen}(K) - \operatorname{pen}(K_0)\right) > V$ with high probability (B3). Then, with high probability for n large enough,

$$\operatorname{crit}(K) = \gamma_n(\hat{\theta}_K) + \operatorname{pen}(K) \geq \gamma_n(\hat{\theta}_{K_0}) - \frac{V}{n} + \operatorname{pen}(K) > \operatorname{crit}(K_0).$$

And then, with high probability for n large enough, $\hat{K} \neq K$.

Conclude: $\mathbb{P}[\hat{K} \neq K_0] = \sum_{K \notin \mathcal{K}} \mathbb{P}[\hat{K} = K] + \sum_{K \in \mathcal{K},\ K \neq K_0} \mathbb{P}[\hat{K} = K]$. □

8.4. Sufficient conditions to ensure Assumption (B4)

Let us introduce the notation $S_n\gamma(\theta) = n\left(\gamma_n(\theta) - \mathbb{E}\left[\gamma(\theta)\right]\right)$. The main result of this section is Lemma 8.2. Some intermediate results which enable to link Lemma

8.2 to Theorem 8.1 via Assumption (B4) are stated as corollaries and proved subsequently. Lemma 8.2 povides a control of $\sup_{\theta\in\Theta}\frac{S_n(\gamma(\theta^0)-\gamma(\theta))}{\|\theta^0-\theta\|_\infty^2+\beta^2}$ (with respect to β) and then of $\frac{S_n(\gamma(\theta^0)-\gamma(\hat\theta))}{\|\theta^0-\hat\theta\|_\infty^2+\beta^2}$. With a good choice of β, and if $S_n\big(\gamma(\theta^0)-\gamma(\hat\theta)\big)$ can be linked to $\|\theta^0-\hat\theta\|_\infty^2$, it is proved in Corollary 8.1 that it may then be assessed that $n\|\hat\theta-\theta^0\|_\infty^2=O_{\mathbb{P}}(1)$. Plugging this last property back into the result of Lemma 8.2 yields (Corollary 8.2) $n\big(\gamma_n(\theta^0_K)-\gamma_n(\hat\theta_K)\big)=O_{\mathbb{P}}(1)$ for any model $K\in\operatorname{argmin}_{1\le K\le K_M}\mathbb{E}\left[\gamma(\theta^0_K)\right]$ and then, under mild identifiability condition, $n\big(\gamma_n(\theta^0_{K_0})-\gamma_n(\hat\theta_K)\big)=O_{\mathbb{P}}(1)$, which is Assumption (B4).

Lemma 8.2. *Let $D\in\mathbb{N}^*$ and $\Theta\subset\mathbb{R}^D$ convex. Let $\Theta^{\mathcal{O}}\subset\mathbb{R}^D$ open such that $\Theta\subset\Theta^{\mathcal{O}}$ and $\gamma:\Theta^{\mathcal{O}}\times\mathbb{R}^d\to\mathbb{R}$. $\theta\in\Theta^{\mathcal{O}}\mapsto\gamma(\theta;x)$ is assumed to be C^1 over $\Theta^{\mathcal{O}}$ for $f^{\wp}d\lambda$-almost all x. Let $\theta^0\in\Theta$ such that $\mathbb{E}\left[\gamma(\theta^0)\right]=\inf_{\theta\in\Theta}\mathbb{E}\left[\gamma(\theta)\right]$. Assume that $H_{\gamma,\Theta,\infty}$ and $H'_{\gamma,\Theta,2}$ hold.*

Then $\exists\alpha>0/\forall n,\forall\beta>0,\forall\eta>0$, with probability larger than $(1-\exp-\eta)$,

$$\sup_{\theta\in\Theta}\frac{S_n(\gamma(\theta^0)-\gamma(\theta))}{\|\theta^0-\theta\|_\infty^2+\beta^2}\le\frac{\alpha}{\beta^2}\Big(\|M'\|_2\beta\sqrt{nD}+\big(\|M\|_\infty+\|M'\|_2\beta\big)D+\|M'\|_2\sqrt{n\eta}\beta+\|M\|_\infty\eta\Big)$$

Note that α is an absolute constant which notably does not depend on θ^0.

Sketch of the proof of Lemma 8.2. The proof relies on results of [28]. Lemma 3.3 provides a local control of the bracketing entropy of the class of functions we consider and hence, through Theorem 6.8 in [28], a control of the supremum of $S_n(\gamma(\theta^0)-\gamma(\theta))$ as $\|\theta-\theta^0\|_\infty^2<\sigma$, with respect to σ. The "peeling" Lemma 4.23 [28] then enables to take advantage of this local control to derive a fine global control of $\sup_{\theta\in\Theta}\frac{S_n(\gamma(\theta^0)-\gamma(\theta))}{\|\theta-\theta^0\|^2+\beta^2}$, for any β. This control in expectation, which can be derived conditionally to any event A, yields a control in probability thanks to Lemma 2.4 in [28], which can be thought of as an application of Markov's inequality. □

Corollary 8.1. *Same assumptions as Lemma 8.2, but the convexity of Θ. Besides assume that $I_{\theta^0}=\frac{\partial^2}{\partial\theta^2}\left(\mathbb{E}\left[\gamma(\theta)\right]\right)_{|\theta^0}$ is nonsingular. Let $(\hat\theta_n)_{n\ge1}$ such that $\hat\theta_n\in\Theta$, $\gamma_n(\hat\theta_n)\le\gamma_n(\theta^0)+O_{\mathbb{P}}(\frac{1}{n})$ and $\hat\theta_n\xrightarrow[n\to\infty]{\mathbb{P}}\theta^0$. Then*

$$n\|\hat\theta_n-\theta^0\|_\infty^2=O_{\mathbb{P}}(1).$$

The constant involved in $O_{\mathbb{P}}(1)$ depends on D, $\|M\|_\infty$, $\|M'\|_2$ and I_{θ^0}.

This is a direct consequence of Lemma 8.2: it suffices to choose β well. The dependency of $O_{\mathbb{P}}(1)$ in D, $\|M\|_\infty$, $\|M'\|_2$ and I_{θ^0} is not a problem since we aim at deriving an asymptotic result: we are interested in the order of $\|\theta-\theta^0\|_\infty^2$ with respect to n when the model is fixed.

The assumption that I_{θ^0} is nonsingular ensures that $\mathbb{E}\left[\gamma(\theta)\right]$ cannot be close to $\mathbb{E}\left[\gamma(\theta^0)\right]$ if θ is not close to θ^0. It guarantees that the rate of the relation

between $\mathbb{E}\left[\gamma(\theta)\right] - \mathbb{E}\left[\gamma(\theta^0)\right]$ and $\|\theta - \theta^0\|$ can then be controlled... Should this assumption fail, $\exists\tilde{\theta} \in \Theta / \tilde{\theta}' I_{\theta^0}\tilde{\theta} = 0 \Rightarrow \mathbb{E}[\gamma(\theta^0 + \lambda\tilde{\theta})] = \mathbb{E}\left[\theta^0\right] + o(\lambda^2)$ and then there is no hope to have $\alpha > 0$ such that $\mathbb{E}\left[\gamma(\theta)\right] - \mathbb{E}\left[\gamma(\theta^0)\right] > \alpha\|\theta - \theta^0\|^2$: this approach cannot be applied without this—admittedly unpleasant—assumption. Perhaps another approach (with distances not involving the parameters but directly the contrast values) might enable to avoid it, as [28] did in the likelihood framework.

Corollary 8.2. *Let $(\Theta_K)_{1\leq K\leq K_M}$ be models with, for any K, $\Theta_K \subset \mathbb{R}^{D_K}$. Assume that $D_1 \leq \cdots \leq D_{K_M}$. For any K, assume there exists an open set $\Theta_K^{\mathcal{O}} \subset \mathbb{R}^{D_K}$ such that $\Theta_K \subset \Theta_K^{\mathcal{O}}$ and such that with $\Theta^{\mathcal{O}} = \Theta_1^{\mathcal{O}} \cup \cdots \cup \Theta_{K_M}^{\mathcal{O}}$, $\gamma : \Theta^{\mathcal{O}} \times \mathbb{R}^d \longrightarrow \mathbb{R}$ is defined and C^1 for $f^{\wp}d\lambda$-almost all x. Assume that $H_{\gamma,\Theta^{\mathcal{O}},\infty}$ and $H'_{\gamma,\Theta^{\mathcal{O}},2}$ hold. Let, for any K, $\Theta_K^0 = \operatorname{argmin}_{\theta\in\Theta_K}\mathbb{E}\left[\gamma(\theta)\right]$ and $\theta_K^0 \in \Theta_K^0$.*

Let $K_0 = \min\operatorname{argmin}_{1\leq K\leq K_M}\mathbb{E}\left[\gamma(\Theta_K^0)\right]$ and assume $\forall K, \forall\theta \in \Theta_K$,

$$\mathbb{E}\left[\gamma(\theta)\right] = \mathbb{E}\left[\gamma(\theta_{K_0}^0)\right] \Longleftrightarrow \gamma(\theta) = \gamma(\theta_{K_0}^0) \quad f^{\wp}d\lambda - a.e.$$

Let $\mathcal{K} = \{K \in \{1,\ldots,K_M\} : \mathbb{E}\left[\gamma(\theta_K^0)\right] = \mathbb{E}\left[\gamma(\theta_{K_0}^0)\right]\}$.

For any $K \in \mathcal{K}$, let $\hat{\theta}_K \in \Theta_K$ such that

$$\gamma_n(\hat{\theta}_K) \leq \gamma_n(\theta_K^0) + O_{\mathbb{P}}\Big(\frac{1}{n}\Big) \qquad \text{and} \qquad \hat{\theta}_K \xrightarrow[n\to\infty]{\mathbb{P}} \theta_K^0.$$

Assume that $I_{\theta_K^0} = \frac{\partial^2}{\partial\theta^2}\left(\mathbb{E}[\gamma(\theta)]\right)_{|\theta_K^0}$ is nonsingular for any $K \in \mathcal{K}$.

Then $\forall K \in \mathcal{K}$, $n\big(\gamma_n(\hat{\theta}_{K_0}) - \gamma_n(\hat{\theta}_K)\big) = O_{\mathbb{P}}(1)$.

This last corollary states conditions under which assumption (B4) of Theorem 8.1 is ensured.

Proof of Lemma 8.2. Actually, the proof as it is written below holds for an at most countable model (because this assumption is necessary for Lemma 4.23 and Theorem 6.8 in [28] to hold). But it can be checked that both these results may be applied to a dense subset of $\{\gamma(\theta) : \theta \in \Theta\}$ containing θ^0 and their respective conclusions generalized to the entire set: choose Θ^{count} a countable dense subset of Θ. Then, for any $\theta \in \Theta$, let $\theta_n \in \Theta^{\text{count}} \xrightarrow[n\to\infty]{} \theta$. Then, $\gamma(\theta_n; X) \xrightarrow[n\to\infty]{a.s.} \gamma(\theta; X)$. Now, whatever $g : \mathbb{R}^D \times \left(\mathbb{R}^d\right)^n \to \mathbb{R}$ such that $\theta \in \mathbb{R}^D \mapsto g(\theta, \mathbf{X})$ continue a.s., $\sup_{\theta\in\Theta} g(\theta; \mathbf{X}) = \sup_{\theta\in\Theta^{\text{count}}} g(\theta; \mathbf{X})$ a.s. Hence, $\mathbb{E}\left[\sup_{\theta\in\Theta} g(\theta; \mathbf{X})\right] = \mathbb{E}\left[\sup_{\theta\in\Theta^{\text{count}}} g(\theta; \mathbf{X})\right]$. Remark that the models which are actually considered in practice are discrete anyway, because of the computation limitations.

Let us introduce the centered empirical process

$$S_n\gamma(\theta) = n\gamma_n(\theta) - n\mathbb{E}\left[\gamma(\theta; X)\right].a(\theta; X_i) - \mathbb{E}_{f^{\wp}}\left[\gamma(\theta; X)\right]\big).$$

Here and hereafter, α stands for a generic absolute constant, which may differ from a line to another. Let $\theta^0 \in \Theta$ such that $\mathbb{E}\left[\gamma(\theta^0)\right] = \inf_{\theta\in\Theta}\mathbb{E}\left[\gamma(\theta)\right]$. Let us

define

$$\forall \sigma > 0, \Theta(\sigma) = \{\theta \in \Theta : \|\theta - \theta^0\|_\infty \leq \sigma\}.$$

On the one hand, for all $r \in \mathbb{N}^* \backslash \{1\}$,

$$\forall \theta \in \Theta(\sigma), \left|\gamma(\theta^0; x) - \gamma(\theta; x)\right|^r \leq M'(x)^2 \|\theta^0 - \theta\|_\infty^2 (2M(x))^{r-2}$$

since $\Theta(\sigma) \subset \Theta$ is convex. And thus,

$$\begin{aligned} \forall \theta \in \Theta(\sigma), \mathbb{E}\left[|\gamma(\theta^0) - \gamma(\theta)|^r\right] &\leq \|M'\|_2^2 \|\theta^0 - \theta\|_\infty^2 (2\|M\|_\infty)^{r-2} \\ &\leq \frac{r!}{2} (\|M'\|_2 \sigma)^2 \left(\frac{2\|M\|_\infty}{2}\right)^{r-2}. \end{aligned} \tag{6}$$

On the other hand, from Lemma 3.3, for any $r \in \mathbb{N}^* \backslash \{1\}$, for any $\delta > 0$, there exists C_δ a set of brackets which cover $\{(\gamma(\theta^0) - \gamma(\theta)) : \theta \in \Theta(\sigma)\}$ (deduced from a set of brackets which cover $\{\gamma(\theta) : \theta \in \Theta(\sigma)\}$...) such that:

$$\forall r \in \mathbb{N}^* \backslash \{1\}, \forall [g_l, g_u] \in C_\delta, \|g_u - g_l\|_r \leq \left(\frac{r!}{2}\right)^{\frac{1}{r}} \delta^{\frac{2}{r}} \left(\frac{4\|M\|_\infty}{3}\right)^{\frac{r-2}{r}}$$

and such that, writing $\exp H(\delta, \Theta(\sigma))$ the minimal cardinal of such a C_δ,

$$\exp H(\delta, \Theta(\sigma)) \leq \left(\frac{\overbrace{\operatorname{diam} \Theta(\sigma)}^{\leq 2\sigma} \|M'\|_2}{\delta}\right)^D \vee 1. \tag{7}$$

Then, according to Theorem 6.8 in [28], $\exists \alpha, \forall \varepsilon \in]0, 1], \forall A$ measurable such that $\mathbb{P}[A] > 0$,

$$\begin{aligned} \mathbb{E}^A \left[\sup_{\theta \in \Theta(\sigma)} S_n\big(\gamma(\theta^0) - \gamma(\theta)\big)\right] &\leq \frac{\alpha}{\varepsilon} \sqrt{n} \int_0^{\varepsilon \|M'\|_2 \sigma} \sqrt{H\big(u, \Theta(\sigma)\big)}\, du \\ &\quad + 2\Big(\frac{4}{3}\|M\|_\infty + \|M'\|_2 \sigma\Big) H\big(\|M'\|_2 \sigma, \Theta(\sigma)\big) \\ &\quad + (1 + 6\varepsilon)\|M'\|_2 \sigma \sqrt{2n \log \frac{1}{\mathbb{P}[A]}} + \frac{8}{3}\|M\|_\infty \log \frac{1}{\mathbb{P}[A]}. \end{aligned} \tag{8}$$

Now, we have

$$\begin{aligned} \forall t \in \mathbb{R}^+, \int_0^t \sqrt{\log \frac{1}{u} \vee 0}\, du &= \int_0^{t \wedge 1} \sqrt{\log \frac{1}{u}}\, du \\ &\leq \sqrt{t \wedge 1} \sqrt{\int_0^{t \wedge 1} \log \frac{1}{u}\, du} = (t \wedge 1) \sqrt{\log \frac{e}{t \wedge 1}} \end{aligned}$$

by the Cauchy-Schwarz inequality. Together with (7), this yields

$$\begin{aligned}\forall t \in \mathbb{R}^+, \int_0^t \sqrt{H(u,\Theta(\sigma))}du &\leq \sqrt{D}\int_0^t \sqrt{\log\frac{2\|M'\|_2\sigma}{u} \vee 0}\, du \\ &\leq \sqrt{D}\big(t \wedge 2\|M'\|_2\sigma\big)\sqrt{\log\frac{e}{\frac{t}{2\|M'\|_2\sigma}\wedge 1}}\end{aligned} \tag{9}$$

after a simple substitution.

Next, let us apply Lemma 4.23 in [28]: From (7), (8) and (9),

$$\forall \sigma > 0, \mathbb{E}\left[\sup_{\theta\in\Theta(\sigma)} S_n\big(\gamma(\theta^0)-\gamma(\theta)\big)\right] \leq \varphi(\sigma)$$

$$\begin{aligned}\text{with } \varphi(t) = \frac{\alpha}{\not{\sigma}}\sqrt{n}\sqrt{D}\not{\sigma}\|M'\|_2 t\sqrt{\log\frac{2e}{\varepsilon}} + 2\Big(\frac{4}{3}\|M\|_\infty + \|M'\|_2 t\Big)D\log 2 \\ + (1+6\varepsilon)\|M'\|_2 t\sqrt{2n\log\frac{1}{\mathbb{P}[A]}} + \frac{8}{3}\|M\|_\infty \log\frac{1}{\mathbb{P}[A]}.\end{aligned}$$

As required for Lemma 4.23 in [28] to hold, $\frac{\varphi(t)}{t}$ is nonincreasing. It follows

$$\forall \beta > 0, \mathbb{E}^A\left[\sup_{\theta\in\Theta}\frac{S_n\big(\gamma(\theta^0)-\gamma(\theta)\big)}{\|\theta^0-\theta\|_\infty + \beta^2}\right] \leq 4\beta^{-2}\varphi(\beta).$$

We then choose $\varepsilon = 1$ and apply Lemma 2.4 in [28]: for any $\eta > 0$ and any $\beta > 0$, with probability larger than $1 - \exp -\eta$,

$$\begin{aligned}\sup_{\theta\in\Theta}\frac{S_n\big(\gamma(\theta^0)-\gamma(\theta)\big)}{\|\theta^0-\theta\|_\infty^2+\beta^2} \leq \frac{\alpha}{\beta^2}\Bigg(\sqrt{nD}\|M'\|_2\beta\sqrt{\log 2e} \\ + \big(\|M\|_\infty + \|M'\|_2\beta\big)D\log 2 + \|M'\|_2\beta\sqrt{n\eta} + \|M\|_\infty\eta\Bigg).\end{aligned}$$

□

Proof of Corollary 8.1. Let $\varepsilon > 0$ such that $B(\theta^0,\varepsilon) \subset \Theta^{\mathcal{O}}$. Then, since $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta^0$, there exists $n_0 \in \mathbb{N}^*$ such that, with high probability, for $n \geq n_0$, $\hat{\theta}_n \in B(\theta^0,\varepsilon)$. Now, $B(\theta^0,\varepsilon)$ is convex and Lemma 8.2 applies to $\hat{\theta}_n$: $\forall n \geq n_0, \forall \beta > 0$, with great probability as η is large,

$$\begin{aligned}\frac{S_n\big(\gamma(\theta^0)-\gamma(\hat{\theta}_n)\big)}{\|\theta^0-\hat{\theta}_n\|_\infty^2+\beta^2} \leq \frac{\alpha}{\beta^2}\Bigg(\sqrt{nD}\|M'\|_2\beta + \big(\|M\|_\infty + \|M'\|_2\beta\big)D \\ + \|M'\|_2\beta\sqrt{n\eta} + \|M\|_\infty\eta\Bigg).\end{aligned} \tag{10}$$

But since I_{θ^0} is supposed to be nonsingular, $\forall\theta \in B(\theta^0,\varepsilon)$,

$$\begin{aligned}\mathbb{E}[\theta] - \mathbb{E}\left[\theta^0\right] &= (\theta-\theta^0)' I_{\theta^0}(\theta-\theta^0) + r(\|\theta-\theta^0\|_\infty)\|\theta-\theta^0\|_\infty^2 \\ &\geq \big(2\alpha' + r(\|\theta-\theta^0\|_\infty)\big)\|\theta-\theta^0\|_\infty^2\end{aligned}$$

where $\alpha' > 0$ depends on I_{θ^0} and $r : \mathbb{R}^+ \longrightarrow \mathbb{R}$ fulfills $r(x) \xrightarrow[x\to 0]{} 0$. Then, for $\|\theta - \theta^0\|_\infty$ small enough (ε may be decreased...),

$$\forall \theta \in B(\theta^0, \varepsilon), \mathbb{E}\left[\theta\right] - \mathbb{E}\left[\theta^0\right] \geq \alpha' \|\theta - \theta^0\|_\infty^2. \tag{11}$$

$$\begin{aligned} S_n\big(\gamma(\theta^0) - \gamma(\hat{\theta}_n)\big) &= n\big(\gamma_n(\theta^0) - \gamma_n(\hat{\theta}_n)\big) + n\mathbb{E}_{f^\wp}\big[\gamma(\hat{\theta}_n) - \gamma(\theta^0)\big] \qquad \text{Since} \\ &\geq O_{\mathbb{P}}(1) + n\mathbb{E}_{f^\wp}\big[\gamma(\hat{\theta}_n) - \gamma(\theta^0)\big], \end{aligned}$$

(10) together with (11) leads (with great probability) to

$$n\|\hat{\theta}_n - \theta^0\|_\infty^2 \leq \frac{\|M'\|_2(\sqrt{nD} + \sqrt{\eta n} + D)\beta + \|M\|_\infty(D+\eta) + O_{\mathbb{P}}(1)}{\frac{\alpha'}{\alpha} - \frac{1}{n\beta^2}\Big(\|M'\|_2(\sqrt{nD} + \sqrt{\eta n} + D)\beta + \|M\|_\infty(D+\eta)\Big)},$$

as soon as the denominator of the right-hand side is positive. It then suffices to choose β such that this condition is fulfilled and such that the right-hand side is upper-bounded by a quantity which does not depend on n to get the result. Let us try $\beta = \frac{\beta_0}{\sqrt{n}}$ with β_0 independent of n:

$$n\|\hat{\theta}_n - \theta^0\|_\infty^2 \leq \frac{\|M'\|_2(\sqrt{D} + \sqrt{\eta} + D)\beta_0 + \|M\|_\infty(D+\eta) + O_{\mathbb{P}}(1)}{\frac{\alpha'}{\alpha} - \frac{1}{\beta_0^2}\Big(\|M'\|_2(\sqrt{D} + \sqrt{\eta} + D)\beta_0 + \|M\|_\infty(D+\eta)\Big)}.$$

This only holds if the denominator is positive. Choose β_0 large enough so as to guarantee this, which is always possible. The result follows: with high probability and for n larger than n_0, we have $n\|\hat{\theta}_n - \theta^0\|_\infty^2 = CO_{\mathbb{P}}(1)$ with C depending on D, $\|M\|_\infty$, $\|M'\|_2$, I_{θ^0} and η. □

Proof of Corollary 8.2. This is a direct application of Corollary 8.1. Let $K \in \mathcal{K}$: $\mathbb{E}\left[\gamma(\theta_K^0)\right] = \mathbb{E}\left[\gamma(\theta_{K_0}^0)\right]$. Θ_K can be assumed to be convex: if it is not, $\hat{\theta}_K$ lies in $B(\theta_{K_0}^0, \varepsilon) \subset \Theta^{\mathcal{O}}$ with high probability for large n and Θ_K may be replaced by $B(\theta_{K_0}^0, \varepsilon)$. According to Lemma 8.2, with probability larger than $(1 - \exp -\eta)$ for n large, with $\beta = \frac{\beta_0}{\sqrt{n}}$ for any $\beta_0 > 0$:

$$\begin{aligned} S_n\big(\gamma(\theta_K^0) - \gamma(\hat{\theta}_K)\big) \leq \alpha\frac{n\|\theta_K^0 - \hat{\theta}_K\|_\infty^2 + \beta_0^2}{\beta_0^2}\Bigg(\|M'\|_2\Big(\sqrt{D_K} + \sqrt{\eta} + \frac{\overbrace{D_K}^{\leq D_K}}{\sqrt{n}}\Big)\beta_0 \\ + \|M\|_\infty(D_K + \eta)\Bigg). \end{aligned}$$

But, according to Corollary 8.1, $n\|\theta_K^0 - \hat{\theta}_K\|_\infty^2 = O_{\mathbb{P}}(1)$. Moreover, by definition,

$$S_n\big(\gamma(\theta_K^0) - \gamma(\hat{\theta}_K)\big) = n\big(\gamma_n(\theta_K^0) - \gamma_n(\hat{\theta}_K)\big) + n\underbrace{\Big(\mathbb{E}_{f^\wp}\left[\gamma(\hat{\theta}_K)\right] - \mathbb{E}\left[\gamma(\theta_K^0)\right]\Big)}_{\geq 0}$$

Thus, $n\big(\gamma_n(\theta_K^0) - \gamma_n(\hat{\theta}_K)\big) = O_{\mathbb{P}}(1)$. This holds for any $K \in \mathcal{K}$ and then in particular for K_0 and K. Besides $\gamma_n(\theta_K^0) = \gamma_n(\theta_{K_0}^0)$ since, by assumption, $\gamma(\theta_K^0) = \gamma(\theta_{K_0}^0)$ $f^\wp d\lambda$-a.e. Hence $n\big(\gamma_n(\hat{\theta}_{K_0}) - \gamma_n(\hat{\theta}_K)\big) = O_{\mathbb{P}}(1)$. □

Acknowledgements

The author is deeply grateful to G. Celeux, J.-M. Marin and P. Massart for their essential help and sincerely thanks the editors and reviewers for their valuable comments, which were of great help in revising and improving the manuscript. The author thanks Kat for helping to improve the writing.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings, 2nd Internat. Symp. on Information Theory* 267–281. MR0483125

[2] Ambroise, C. and Govaert, G. (2000). Clustering by maximizing a fuzzy classification maximum likelihood criterion. In *COMPSTAT* 187–192. Springer.

[3] Arlot, S. (2007). Resampling and model selection. PhD thesis, Univ. Paris-Sud.

[4] Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields* **113** 301–413. MR1679028

[5] Baudry, J. P. (2009). Model selection for clustering. Choosing the number of classes. PhD thesis, Univ. Paris-Sud. http://tel.archives-ouvertes.fr/tel-00461550/fr/.

[6] Baudry, J. P., Celeux, G. and Marin, J. M. (2008). Selecting models focussing on the modeler's purpose. In *COMPSTAT 2008: Proceedings in Computational Statistics* 337–348. Physica-Verlag, Heidelberg. MR2509588

[7] Baudry, J.-P., Maugis, C. and Michel, B. (2011). Slope heuristics: overview and implementation. *Statist. Comput.* **22** 455–470. MR2865029

[8] Baudry, J. P., Raftery, A. E., Celeux, G., Lo, K. and Gottardo, R. (2010). Combining mixture components for clustering. *J. Comput. Graph. Statist.* **19** 332–353. MR2758307

[9] Biernacki, C., Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. PAMI* **22** 719–725.

[10] Biernacki, C., Celeux, G. and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis* **41** 567–575. MR1968069

[11] Biernacki, C. and Govaert, G. (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics* **29** 451–457.

[12] Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society* **3** 203–268. MR1848946

[13] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138** 33–73. MR2288064

[14] Celeux, G. and Govaert, G. (1993). Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of Statistical Computation and simulation* **47** 127–146.
[15] Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition* **28** 781–793.
[16] De Granville, C., Southerland, J. and Fagg, A. H. (2006). Learning grasp affordances through human demonstration. In *Proceedings of the International Conference on Development and Learning*, electronically published.
[17] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society. Series B* **39** 1–38. MR0501537
[18] Dudley, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge Univ Press. MR1720712
[19] Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635
[20] Goutte, C., Hansen, L. K., Liptrot, M. G. and Rostrup, E. (2001). Feature-space clustering for fMRI meta-analysis. *Human Brain Mapping* **13** 165–183.
[21] Hamelryck, T., Kent, J. T. and Krogh, A. (2006). Sampling realistic protein conformations using local structural bias. *PLoS Comput. Biol.* **2** e131.
[22] Hennig, C. (2010). Methods for merging Gaussian mixture components. *Adv. Data Anal. Classif.* **4** 3–34. MR2639661
[23] Keribin (2000). Consistent estimation of the order of mixture models. *Sankhya A* **62** 49–66. MR1769735
[24] Laloë, T. and Servien, R. (2013). The X-alter algorithm: a parameter-free method of unsupervised clustering. *Journal of Modern Applied Statistical Methods* **12** 14.
[25] Lange, K. (1999). *Numerical Analysis for Statisticians*. Springer-Verlag, New-York. MR1681963
[26] Lee, S. X. and McLachlan, G. J. (2013). Model-based clustering and classification with non-normal mixture distributions. *Statistical Methods & Applications* **22** 427–454. MR3127088
[27] Mariadassou, M., Robin, S. and Vacher, C. (2010). Uncovering latent structure in valued graphs: a variational approach. *Ann. Appl. Stat.* **4** 715–742. MR2758646
[28] Massart, P. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in Math.* Springer. MR2319879
[29] Maugis, C. and Michel, B. (2011). A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM Probab. Stat.* **15** 41–68. MR2870505
[30] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York. MR1789474

[31] McQuarrie, A. D. R. and Tsai, C. L. (1998). *Regression and Time Series Model Selection*. World Scientific. MR1641582

[32] Nishii, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *J. Multivariate Anal.* **27** 392–403. MR0970962

[33] Pelleg, D., Moore, A. W. et al. (2000). X-means: extending K-means with efficient estimation of the number of clusters. In *ICML* 727–734.

[34] Pigeau, A. and Gelgon, M. (2005). Building and tracking hierarchical geographical & temporal partitions for image collection management on mobile devices. In *Proceedings 13th Annual ACM Internat. Conf. on Multimedia* 141–150. ACM, New York, NY, USA.

[35] Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239. MR0738930

[36] Rigaill, G., Lebarbier, E. and Robin, S. (2012). Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statist. Comput.* 1–13. MR2913792

[37] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

[38] Steele, R. J. and Raftery, A. E. (2010). Performance of Bayesian model selection criteria for Gaussian mixture models. In *Frontiers of Statistical Decision Making and Bayesian Analysis* 113–130. Springer.

[39] Symons, M. J. (1981). Clustering criteria and multivariate normal mixtures. *Biometrics* **37** 35–43. MR0673031

[40] Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York. MR0838090

[41] van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press. MR1652247