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

Jean-Patrick Baudry

Estimation and Model Selection for Model-Based Clustering with the Conditional Classification Likelihood

Jean-Patrick Baudry*
LSTA
Université Pierre et Marie Curie - Paris VI

May 31, 2012

Abstract

The Integrated Completed Likelihood (ICL) criterion has been proposed by Biernacki et al. (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 this criterion is proposed. A contrast related to the clustering objective is introduced: the *conditional classification likelihood*. This yields an estimator and a model selection criteria class. The properties of these new procedures are studied and ICL is proved to be an approximation of one of these criteria.

We oppose these results to 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 on mixture models are derived, which are interesting in their own right.

Keywords BIC; Bracketing entropy; Classification likelihood; Contrast minimization; ICL; Model-based clustering; Model selection; Mixture models; Number of classes; Penalized criteria

*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.

1 Introduction

Model-based clustering is introduced in Sections 1.1 and 1.2. Our purpose is to better understand the behavior of the ICL model selection criterion of Biernacki et al. (2000), which is presented in Section 1.3.

The main topic of this work is the choice of the number of classes in a model-based clustering framework, and then the choice of the number of components of a Gaussian mixture. The interested reader may refer to Titterington et al. (1985) or McLachlan and Peel (2000) for comprehensive studies on Gaussian mixture models. The last also provides 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 (Akaike, 1973) or BIC (Schwarz, 1978).

The ICL criterion studied here is an alternative to BIC. It was up to now widely presented as a penalized likelihood criterion, which penalty involves an "entropy" term. Here, however, we prove that it is actually a penalized contrast criterion with a criterion which is different from 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 considered. Even for data arising from a mixture distribution, a relevant number of classes may differ from the true number of components of the mixture.

The reason why we introduce this new contrast L_{cc} (Section 2.1) is not that we believe it *a priori* to be the better one for a clustering purpose, but rather that it enables to theoretically study and understand ICL. We prove (Section 4.3) that ICL is an approximation of a criterion linked to this contrast: studying further ICL then amounts to studying L_{cc}. 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 5).

Let X be a random variable in $\mathbb{R}^d$ with distribution $f^{\wp} \cdot \lambda$ and $X_1, \ldots, X_n$ an i.i.d. sample of the same distribution. Let us denote $\mathbf{X} = (X_1, \ldots, X_n)$.

All proofs are gathered in Section 6.

1.1 Gaussian Mixture Models

$\mathcal{M}_K$ 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 $\Theta_K \subset \Pi_K \times \left(\mathbb{R}^d \times \mathbb{S}_+^d\right)^K$ 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 positive definite $d \times d$ real matrices. Constraints on the model may be imposed by restricting Θ_K. We typically have in mind the decomposition suggested by Celeux and Govaert (1995). "General" (no constraint) and "diagonal" (diagonal covariance matrices) models will be considered here as examples.

Those 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 Θ_K and φ are "optimal", in the sense that D_K is minimal. D_K is the number of *free parameters* in the model $\mathcal{M}_K$ and is called the *dimension* of $\mathcal{M}_K$. For example, at most $(K-1)$ mixing proportions need to be parametrized.

It shall not be needed 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 those classes. Theorem 4.2 justifies that the first task can be achieved under a weaker "identifiability" assumption. Theorem 3.2 then guarantees that our estimator converges to the best parameters set, any of which is as good as the others. There will be no "true parameter" assumption. The classes can finally be defined through the MAP rule (see Section 1.2). Practically, the parameters themselves are never the quantities of interest here. They only stand as a convenient notation and this is also why we expect that the assumption about the Fisher information (see Theorem 4.2) is technical and could maybe be avoided with other techniques. Please refer to Baudry (2009, Chapter 4) for a more comprehensive discussion about the identifiability question.

1.2 Model-Based Clustering

Although the results are stated first for much more general situations, this paper is devoted to the question of clustering through Gaussian mixture models.

The process is standard (see Fraley and Raftery, 2002):

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

Notably, the usual choice is made here, to identify a class with each fitted Gaussian component. The number of classes to be designed is then chosen

at the second step. See for example Hennig (2010) and Baudry et al. (2010) for alternative approaches.

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

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

$\tau_k(x;\theta)$ is the probability that X arises from the k^{th} component, conditionally to $X = x$, under the distribution defined by θ. Let us also denote $\tau_{ik}(\theta) = \tau_k(X_i;\theta)$. The MAP classification rule for x is then

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

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 $\widehat{\theta}_K^{\text{MLE}}$.

1.3 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 $(\mathbf{X},\mathbf{Z})$ ($Z \in \{0,1\}^K$ is the unobserved label of X: $Z_k = 1 \Leftrightarrow X$ arises from component k):

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

To mimic the derivation of the BIC criterion (Schwarz, 1978) in a clustering framework, Biernacki et al. (2000) approximate the integrated classification likelihood through a Laplace's approximation. Then they assume that the classification likelihood mode can be identified with $\widehat{\theta}_K^{\text{MLE}}$ as n is large enough and replace the unobserved Z_{ik}'s by their MAP estimators under $\widehat{\theta}_K^{\text{MLE}}$. This is questionable, notably when the components of $\widehat{\theta}_K^{\text{MLE}}$ are not well separated. They derive the ICL criterion:

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

McLachlan and Peel (2000) replace the Z_{ik}'s by their conditional expectations $\tau_{ik}(\widehat{\theta}_K^{\mathrm{MLE}})$:

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

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 Schwarz (1978) through the *entropy* term (see Section 2.2):

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

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 (Keribin, 2000; Nishii, 1988). 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 needed to approximate each non-Gaussian component of the true mixture distribution $f^{\wp}$. And the user may rather be interested in a *cluster* notion — as opposed to this strictly *component* approach — which also includes a separation notion and which be robust to non-Gaussian components. Of course, it depends on the application, and on what a *class* should be. It may be of interest to discriminate into two different classes a group of observations 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 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 3)...

ICL has been derived with this viewpoint. It is widely understood and explained (for instance in Biernacki et al., 2000) as the BIC criterion with a supplemental penalty, which is the entropy (Section 2.2). Since the last penalizes models which maximum likelihood estimator yields an uncertain MAP classification, ICL is more robust than BIC to non-Gaussian components. However we do not think that the entropy should be considered as a penalty term and an other point of view will be developed in this paper.

The references here were found by browsing the result obtained from Google Scholar citations about Biernacki et al. (2000). Only 3 pages of 16 have been studied... The behavior of ICL has been studied through simulations and real data studies by Biernacki et al. (2000), McLachlan and Peel (2000, Section 6.11), Steele and Raftery (2010) and in several simulation studies (See Baudry, 2009, Chapter 4). Besides several authors chose to use it for the mentioned reasons in various applications area: Goutte et al. (2001) (fMRI images); Pigeau and Gelgon (2005) (image collection automatic sorting); Hamelryck et al. (2006) (protein structure prediction); De Granville et al. (2006) (robots learning); Mariadassou et al. (2010) (uncovering groups of nodes in valued graphs and application to host-parasite interaction networks in forest ecosystems analysis); Rigaill et al. (2012) (comparative genomic hybridization profile); etc.

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 go further in this direction. This leads to considering new estimation and model selection procedures for clustering, similar to ICL but for which the development of the underlying logic is driven to its conclusion, from the estimation step to the model selection step, instead of introducing the MLE. It is proved 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, as opposed to the usual point of view. It should rather be linked to an other contrast: the *conditional classification likelihood*.

2.1 Definition, Origin

In a clustering context, the classification likelihood (see (1)) is an interesting quantity but neither the labels $\mathbf{z}$ are observed, nor we assume that they even exist (think of the case several models with different number of components are fitted: then at most one can correspond to the true number of classes). Beside the first-mentioned works of Biernacki et al. (2000), Biernacki and Govaert (1997), for example, already proposed to directly involve

the classification likelihood to select the number of classes, by estimating the unobserved data. We propose here to consider its expectation conditional on the observed sample $\mathbf{X}$. In case there exists a true classification and a model with the true number of classes is considered, this conditional expectation can be interpreted as the quantity the closest to the classification likelihood, which can be considered given the available information.

Let us report 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, denoting the conditional expectation of $\log \mathrm{L_c}(\theta)$ by $\log \mathrm{L_{cc}}(\theta)$ (for Conditional Classification log Likelihood),

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

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

$\log \mathrm{L_{cc}}$ differs from $\log \mathrm{L}$ through the *entropy* (see (3)).

The behavior of the entropy is based on the properties of the function $h : t \in [0,1] \longmapsto (-t \log t)$ (with $h(0) = 0$). This nonnegative function (see Figure 1) takes zero value if and only if $t = 0$ or $t = 1$. It is continuous but not differentiable at 0, and in particular it is not Lipschitz over $[0,1]$, which will be a cause of analysis difficulties. Let us also introduce the function $h_K : (t_1, \ldots, t_K) \in \Pi_K \longmapsto \sum_{k=1}^{K} h(t_k)$. This nonnegative function (see Figure 2) then takes zero value if and only if there exists $k_0 \in \{1, \ldots, K\}$ such that $t_{k_0} = 1$ and $t_k = 0$ for $k \neq k_0$. It reaches its maximum value $\log K$ at $(t_1, \ldots, t_K) = (\frac{1}{K}, \ldots, \frac{1}{K})$ (proof in Section 6).

Now, the contribution $\mathrm{ENT}(\theta; x_i)$ of a single observation to the total entropy $\mathrm{ENT}(\theta; \mathbf{x})$ is considered. Figure 3 represents a dataset simulated from a four-component Gaussian mixture. Let θ be such that $f(\,.\,;\theta) = f^{\wp}$. First, $\mathrm{ENT}(\theta; x_i) \approx 0$ 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}). Second, $\mathrm{ENT}(\theta; x_i)$ is all the greater that $(\tau_{i1}, \ldots, \tau_{iK})$ is

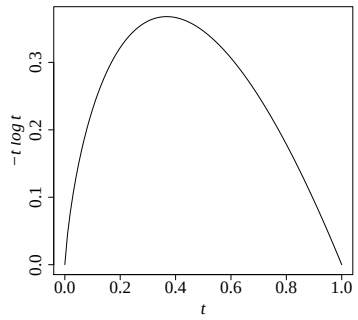

Figure 1: $h : t \mapsto -t \log t$

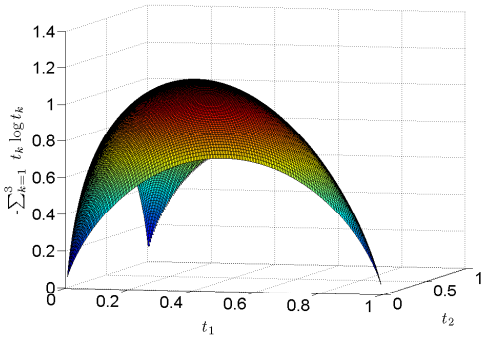

Figure 2: $h_3 : (t_1, t_2, t_3) \in \Pi_3 \mapsto \sum_{k=1}^{3} h(t_k)$

closer to $(\frac{1}{K}, \ldots, \frac{1}{K})$, i.e. that the classification through the MAP rule is uncertain. The worst case is reached as the conditional distribution over of the components $1, \ldots, K$ is uniform. The observation x_{i_2} for example 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 classification rule. The total entropy $\mathrm{ENT}(\theta; \mathbf{x})$ is the empirical mean assignment confidence, and then measures the MAP classification quality for the whole sample.

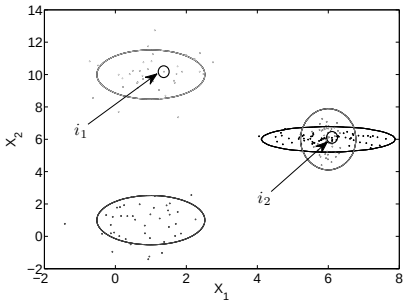

Figure 3: A dataset example

Involving this quantity in a clustering study 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).

2.3 $\log \mathbf{L_{cc}}$ as a Contrast

See for example Massart (2007) for an introduction to contrast minimization. Let us consider the best distribution from the L_{cc} point of view in a model $\mathcal{M}_m = \{f(\,.\,;\theta) : \theta \in \Theta_m\}$, namely the distribution minimizing the corresponding *loss function*

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

The existence of $\mathbb{E}_{f^{\wp}}\left[-\log \mathrm{L}_{\mathrm{cc}}(\theta)\right]$ 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 at hand in the model $\mathcal{M}_K$ (Section 1.1). First of all, 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 $f^{\wp}$. Now, if $K > 1$, $\theta_K^0 \in \Theta_K^0$ may be close to minimizing the Kullback-Leibler divergence if the corresponding components do not overlap since then, the entropy is about zero. But if those components overlap, this is not the case anymore (Example 2.1).

To completely define the loss function, and to fully understand this framework, it is necessary to consider the *best element of the universe* $\mathcal{U}$:

$$\operatorname*{argmin}_{\theta\in\mathcal{U}} \mathbb{E}_{f^{\wp}}\left[-\log \mathrm{L}_{\mathrm{cc}}(\theta)\right].$$

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 well-defined in a parametric mixture setup, and not necessarily over any mixture densities set because of the definition of the entropy term involving the definition of each component. However, this would still enable to consider mixtures much more general than mixtures of Gaussian components. The ideas developed in Baudry et al. (2010) may for example suggest to involve mixtures which components are Gaussian mixtures. But this would not make sense. The mixture with one component which is a mixture of K Gaussian components,

and which then yields a single non-Gaussian-shaped class, always has a smaller $-\log \mathrm{L_{cc}}$ value than the corresponding Gaussian mixture yielding K classes. This illustrates how carefully the components involved in the study must be chosen: involving for example any mixture of Gaussian mixtures means that one considers that a class may be almost anything and may notably contain two Gaussian-shaped clusters very far from each other! The components should in any case be chosen with respect to the corresponding cluster shape. The most natural is then to involve in the universe 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 consider Θ_2^0 in this most 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 $\mathbb{E}_{f^{\wp}}\left[-\log L_{cc}(\mu,\sigma^2)\right]$ in Θ_2 in this case (see Figure 4), 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 5). This illustrates that the objective with the $-\log L_{cc}$ contrast is not to recover the true distribution, even when it is available in the considered model.

The necessity of choosing a relevant model is striking in this example: this two-component model should obviously not be used for a clustering purpose, at least for datasets with great enough size.

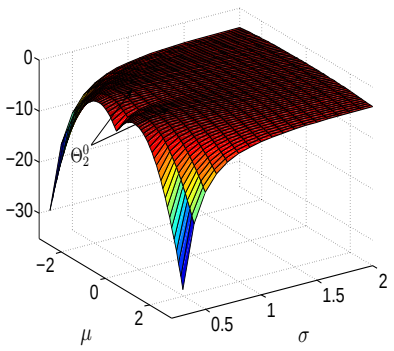

Figure 4: $\mathbb{E}_{f^{\wp}}\left[\log \mathrm{L_{cc}}(\mu,\sigma^2)\right]$ w.r.t. μ and σ, and Θ_K^0, for Example 2.1

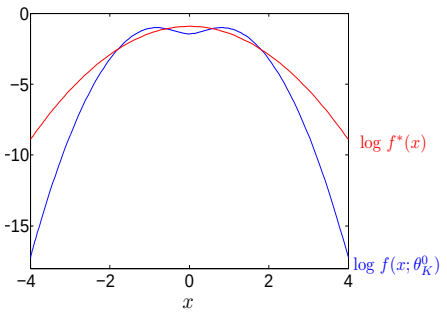

Figure 5: $\log f^{\wp}$ (red, which is also $\log f(\,.\,;\theta_2^{\mathrm{KL}})$) and $\log f(\,.\,;\theta_2^0)$ (blue) for Example 2.1

The estimator associated to the $-\log \mathrm{L_{cc}}$ contrast is now considered.

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 is considered. Results are stated in a general parametric model setting with a general contrast γ and a model $\mathcal{M}$ with parameter space $\Theta \subset \mathbb{R}^D$, and then the conditions they involve are discussed in our framework. General conditions ensuring the consistency of such an estimator are given in Theorem 3.1. They notably involve the Glivenko-Cantelli property of the class of functions $\{\gamma(\theta) : \theta \in \Theta\}$. Sufficient conditions in terms of bracketing entropy for this property to hold are recalled and verified in the considered context in Section 3.2. Those results are also useful in the study of the model selection step (Section 4). Brought together, they provide the consistency of the estimator in Gaussian mixture models: this is Theorem 3.2.

Here and hereafter, all expectations $\mathbb{E}$ and probabilities $\mathbb{P}$ are taken with respect to $f^{\wp} \cdot \lambda$. For a general contrast γ, we write its empirical version: $\gamma_n(\theta) = \frac{1}{n}\sum_{i=1}^n \gamma(\theta; X_i)$. $\mathbb{R}^D$ is equipped with the infinite norm: $\forall \theta \in \mathbb{R}^D, \|\theta\|_\infty = \max_{1\leq i\leq D} |\theta_i|$. For any $r \in \mathbb{N}^* \cup \{\infty\}$ and for any $g : \mathbb{R}^d \to \mathbb{R}$, $\|g\|_r = \mathbb{E}_{f^{\wp}}\left[|g(X)|^r\right]^{\frac{1}{r}}$ if $r < \infty$ and $\|g\|_\infty = \operatorname{ess\,sup}_{X\sim f^{\wp}} |g(X)|$ (recall that $\operatorname{ess\,sup}_{Z\sim\mathbb{P}} Z = \inf\{z : \mathbb{P}[Z \leq z] = 1\}$ and thus: $\|g\|_\infty \leq \sup_{x\in \operatorname{supp} f^{\wp}} |g(x)|$). For any linear form $l : \mathbb{R}^D \to \mathbb{R}$, $\|l\|_\infty = \max_{\|\theta\|_\infty=1} l(\theta)$.

3.1 Definition, Consistency

The minimum contrast estimator is named MLccE (Maximum conditional classification Likelihood Estimator):

$$\widehat{\theta}^{\mathrm{MLccE}} \in \operatorname*{argmin}_{\theta\in\Theta} -\log \mathrm{L}_{\mathrm{cc}}(\theta).$$

To ensure its existence, we assume that Θ is compact. This is a heavy assumption, but it will be natural and necessary for the following results to hold. That the covariance matrices are bounded from below is a reasonable and necessary assumption in the Gaussian mixture framework: without this assumption, neither the log likelihood, nor the conditional classification likelihood would be bounded (for $K \geq 2$). Insights to choose lower bounds on the proportions and the covariance matrices are suggested in Baudry (2009, 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 (see Section 3.2).

This is a consequence of the behavior of the entropy term as a component goes to zero.

The following theorem, which is directly adapted from van der Vaart (1998, Section 5.2), gives sufficient conditions for the consistency of a minimum contrast estimator $\hat{\theta}$. We write $\forall \theta \in \Theta, \forall \widetilde{\Theta} \subset \Theta, d(\theta, \widetilde{\Theta}) = \inf_{\widetilde{\theta} \in \widetilde{\Theta}} \|\theta - \widetilde{\theta}\|_\infty$.

Theorem 3.1. *Let $\Theta \subset \mathbb{R}^D$ and $\gamma : \Theta \times \mathbb{R}^d \longrightarrow \mathbb{R}$. Assume:*

$$\exists \theta^0 \in \Theta \text{ such that } \mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right] = \min_{\theta \in \Theta} \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right] \ (\Leftrightarrow \Theta^0 \text{ is not empty}) \tag{A1}$$

$$\forall \varepsilon > 0, \quad \inf_{\{\theta : d(\theta, \Theta^0) > \varepsilon\}} \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right] > \mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right] \tag{A2}$$

$$\sup_{\theta \in \Theta} \left| \gamma_n(\theta) - \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right] \right| \xrightarrow{\mathbb{P}} 0 \tag{A3}$$

Define $\forall n, \hat{\theta} = \hat{\theta}(X_1, \ldots, X_n) \in \Theta$ such that $\gamma_n(\hat{\theta}) \leq \gamma_n(\theta^0) + o_{\mathbb{P}}(1)$.
Then $d(\hat{\theta}, \Theta^0) \xrightarrow[n \to \infty]{\mathbb{P}} 0$.

The strong consistency holds if (A3) is replaced by an almost sure convergence (this is the case under the conditions we are to define) and if the inequality in the definition of $\hat{\theta}$ holds almost surely.

Assumption (A1) is the least that can be expected. It is guaranteed if the parameter space is compact.

Assumption (A2) holds, too, under this compactness assumption: since $\theta \in \Theta_K \mapsto \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right]$ reaches its minimum value on the compact $\Theta_K \backslash \{\theta \in \Theta_K : d(\theta, \Theta_K^0) > \varepsilon\}$, it is necessarily strictly greater than $\mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right]$.

Assumption (A3) is a bit strong but it will be guaranteed under the compactness assumption through bracketing entropy arguments in Section 3.2.

Sketch of proof. The assumptions guarantee a convenient situation. With great probability as n grows, from (A3), $\gamma_n(\theta)$ is uniformly close to $\mathbb{E}_{f^\wp}\left[\gamma(\theta)\right]$: this holds for $\hat{\theta}$ and θ^0. Then, from the definition of $\hat{\theta}$, $E_{f^\wp}\left[\gamma(\hat{\theta})\right]$ cannot be much larger than $\mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right]$ which reaches the minimal value. By (A2), this implies that $\hat{\theta}$ cannot be far from Θ^0.

□

Let us apply Theorem 3.1 to Gaussian mixtures, with $\gamma = -\log \mathrm{L_{cc}}$ and $\Theta = \Theta_K$. The two following hypotheses will be involved:

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

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

Theorem 3.2 (Weak Consistency of MLccE, compact case)**.** *Let $\mathcal{M}$ be a Gaussian mixture model with compact parameter space $\Theta \subset \mathbb{R}^D$. Let $\Theta^0 = \operatorname{argmin}_{\theta\in\Theta} \mathbb{E}_{f^0}[-\log L_{cc}(\theta)]$. Let $\Theta^{\mathcal{O}} \subset \mathbb{R}^D$ open over which $\log L_{cc}$ is defined, such that $\Theta \subset \Theta^{\mathcal{O}}$ and assume that $H^{M'}_{\log L_{cc},\Theta^{\mathcal{O}},1}$ holds.*
Let $\widehat{\theta}^{MLccE} \in \Theta$ be an estimator (almost) maximizing $\log L_{cc}$:

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

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

$H^{M'}_{\log \mathrm{L_{cc}},\Theta^{\mathcal{O}},1}$ results from lemma 3.2 and shall be discussed in Section 3.2.

Under the compactness assumption, $\widehat{\theta}^{\mathrm{MLccE}}$ is then consistent. It is even strongly consistent if it minimizes the empirical contrast almost surely. Let us highlight that it then 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 last lies in $\mathcal{M}$.

3.2 Bracketing Entropy and Glivenko-Cantelli Property

Recall a class of functions over $\mathbb{R}^d$ is $\mathbb{P}$-*Glivenko-Cantelli*, with $\mathbb{P}$ a probability measure over $\mathbb{R}^d$, if it fulfills a uniform law of large numbers for the distribution $\mathbb{P}$. 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 $g \in \mathcal{G}$ with $l \leq g \leq u$. $[l, u]$ 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 $\mathcal{E}_{[]}(\varepsilon, \mathcal{G}, L_r(\mathbb{P}))$ of $\mathcal{G}$ with respect to $\mathbb{P}$ is the logarithm of $N_{[]}(\varepsilon, \mathcal{G}, L_r(\mathbb{P}))$.*

It is quite natural that the behavior of all functions lying inside an ε-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 3.3. *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.*

The reader is referred to van der Vaart (1998, Chapter 19) for accurate definitions and a proof of this result. This is a generalization of the

usual Glivenko-Cantelli theorem. We shall prove that the class of functions $\{\gamma(\,.\,;\theta) : \theta \in \Theta_K\}$ has finite ε-bracketing entropy for any $\varepsilon > 0$ and the assumption (A3) will be ensured.

From now on, since Θ is typically assumed to be compact, it is assumed that $\Theta \subset \Theta^{\mathcal{O}} \subset \mathbb{R}^D$ with $\Theta^{\mathcal{O}}$ open over which γ is defined and C^1 for $f^{\wp}d\lambda$-almost all x. This is no problem for Gaussian mixture models with $\log \mathrm{L}_{\mathrm{cc}}$ (or the standard likelihood by the way), for example with the general or diagonal model. But this requires (with the $\log \mathrm{L}_{\mathrm{cc}}$ contrast) the proportions to be positive. Actually, this could be avoided here, but we will need this assumption for the definition of M' (Hypothesis $H^{M'}_{\gamma,\Theta,r}$). As already mentioned, components going to zero must be avoided. For the same technical reason, we have to assume the mean parameters to be bounded.

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 assumption about the differential of the contrast is not a difficulty in our framework, provided that non-zero lower bounds over Θ on the proportions and the covariance matrices are imposed. The lemma is written for any $\tilde{\Theta}$ bounded and included in Θ (which is not assumed to be bounded itself) since it will be applied locally around θ^0 in the Section 4.

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

Lemma 3.1 (Bracketing Entropy, Convex Case)**.** *Let $r \in \mathbb{N}^*$, $D \in \mathbb{N}^*$ and $\Theta \subset \mathbb{R}^D$ assumed to be convex. Let $\Theta^{\mathcal{O}} \subset \mathbb{R}^D$ open such that $\Theta \subset \Theta^{\mathcal{O}}$ and $\gamma : \Theta^{\mathcal{O}} \times \mathbb{R}^d \longrightarrow \mathbb{R}$. $\theta \in \Theta^{\mathcal{O}} \longmapsto \gamma(\theta; x)$ is assumed to be C^1 for $f^{\wp}d\lambda$-almost all x. Assume that $H^{M'}_{\gamma,\Theta,r}$ holds. Then*

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

Remark that Θ does not have to be compact. Its proof is a calculation which relies on the mean value theorem, hence the convexity assumption. The natural parameter space of diagonal Gaussian mixture models, with equal volumes (if $d > 1$) or not, for instance, is convex (see Examples 6.1 and 6.2, p. 26). General mixture models have a convex natural parameter space, too, since the set of definite positive matrices is convex. However, there is no reason that the parameter space Θ should be convex in general.

Lemma 3.1 can then be generalized at the price of assuming Θ to be compact, and included in an open set $\Theta^{\mathcal{O}}$ such that $H^{M'}_{\gamma,\Theta^{\mathcal{O}},r}$ holds. This is no difficulty for the mixture models we consider, under the same lower bounds constraints as before (since $\Theta^{\mathcal{O}}$ itself can be chosen to be included

in a compact subset of the set of possible parameters). The entropy is then increased by a multiplying factor Q, which only depends on Θ and roughly measures its "nonconvexity". Since the exponential behavior of the entropy with respect to ε is of concern, this does not make the result really weaker.

Lemma 3.2 (Bracketing Entropy, Compact Case)**.** *Let $r \in \mathbb{N}^*$, $D \in \mathbb{N}^*$ and $\Theta \subset \mathbb{R}^D$ assumed to be compact. Let $\Theta^{\mathcal{O}} \subset \mathbb{R}^D$ open such that $\Theta \subset \Theta^{\mathcal{O}}$ and $\gamma : \Theta^{\mathcal{O}} \times \mathbb{R}^d \longrightarrow \mathbb{R}$. $\theta \in \Theta^{\mathcal{O}} \longmapsto \gamma(\theta; x)$ is assumed to be C^1 for $f^{\wp} d\lambda$-almost all x. Assume that $H^{M'}_{\gamma,\Theta^{\mathcal{O}},r}$ holds.*

Then

$$\exists Q \in \mathbb{N}^*, \forall \widetilde{\Theta} \subset \Theta, \forall \varepsilon > 0,$$

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

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

This lemma is proved by applying Lemma 3.1 since Θ is still locally convex. Since it is compact, it can be covered with a finite number Q of open balls, which are convex. Lemma 3.1 then applies to the convex hull of the intersection of Θ with each one of them. The supremum of M' is taken over $\Theta^{\mathcal{O}}$ — instead of Θ — to make sure that the assumptions of Lemma 3.1 are fulfilled over those entire balls, which may not be included in Θ.

The result we need for Section 4 is Lemma 3.3, obtained from Lemma 3.1 by a slight modification. Since it is applied locally there, the convexity assumption is no problem. A supplementary and strong assumption $H^{M}_{\gamma,\Theta,\infty}$ is made. This is not fulfilled in the general Gaussian mixtures framework. A sufficient condition is that the support of $f^{\wp}$ is bounded. This is false of course for most usual distributions we may have in mind, but this is a reasonable modeling assumption: most modeled phenomena are bounded. Another sufficient condition to guarantee this assumption is that the contrast is bounded from above. This is actually not the case of the contrast $-\log \mathrm{L}_{\mathrm{cc}}$, but this can be imposed: 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 $\log \mathrm{L}_{\mathrm{cc}}$. This is a supplemental difficulty in practice to choose a relevant C value, though.

Lemma 3.3 (Bracketing Entropy, Convex Case)**.** *Let $r \geq 2$, $D \in \mathbb{N}^*$ and $\Theta \subset \mathbb{R}^D$ assumed to be convex. Let $\Theta^{\mathcal{O}} \subset \mathbb{R}^D$ open such that $\Theta \subset \Theta^{\mathcal{O}}$ and $\gamma : \mathbb{R}^D \times \Theta^{\mathcal{O}} \longrightarrow \mathbb{R}$. $\theta \in \Theta^{\mathcal{O}} \longmapsto \gamma(\theta; x)$ is assumed to be C^1 for $f^{\wp} d\lambda$-almost*

all x. Assume that $H^{M}_{\gamma,\Theta,\infty}$ and $H^{M'}_{\gamma,\Theta,2}$ hold. Then

$$\forall \widetilde{\Theta} \subset \Theta, \forall \varepsilon > 0,$$

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

Let us remark that those 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. Maugis and Michel (2011) already 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. This is helpful to derive an oracle inequality. But they could not derive a local control of the entropy, hence an unpleasant logarithm term in the expression of the optimal penalty they get. Their results also suggest the necessity of assuming the contrast to be bounded: see the discussion after the Theorem 4.2. The results we propose achieve the same rate with respect to ε. They depend on more opaque quantities ($\|M\|_\infty$ and $\|M'\|_2$). This notably implies, from this first step already, the assumption that the contrast is bounded — over the true distribution support. However, it could be expected to control those quantities with respect to the parameter space bounds. Moreover, beside their simplicity, they straightforwardly enable to derive 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 target of the study. Anyhow, a relevant number of classes must obviously be chosen so as to design a good classification.

Model selection procedures introduced here are penalized conditional classification likelihood criteria:

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

Most results are stated for a general contrast γ and any family of models $\{\mathcal{M}_K\}_{1 \leq K \leq K_M}$ and then applied to $-\log \mathrm{L_{cc}}$ and the Gaussian mixtures family of models $\{\mathcal{M}_K\}_{1 \leq K \leq K_M}$ introduced in Section 1.1.

In Section 4.1, the consistency of such a model selection procedure ("identification" point of view) is proved for a class of penalties. Sufficient

conditions are given in the general Theorem 4.1, which is applied to the framework we are interested in in Theorem 4.2. The heaviest condition of Theorem 4.1 (B4) may be guaranteed under regularity and (weak) identifiability assumptions, and is discussed in Section 4.2. Our approach is inspired from works of Massart (2007) and is the first step to reach non-asymptotic results.

4.1 Consistent Penalized Criteria

Assume that K_0 exists such that

$$\text{and} \quad \begin{array}{l} \forall K < K_0, \inf\limits_{\theta\in\Theta_{K_0}} \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right] < \inf\limits_{\theta\in\Theta_K} \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right] \\ \forall K \geq K_0, \inf\limits_{\theta\in\Theta_{K_0}} \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right] \leq \inf\limits_{\theta\in\Theta_K} \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right] \end{array}$$

which means that the bias of the models is stationary from the model $\mathcal{M}_{K_0}$: it is the "best" model. Remark that the last property should hold mostly in the mixtures framework, and notably if the models were not constrained, and then were nested. Under this assumption, a model selection procedure is expected to asymptotically recover K_0, i.e. to be *consistent*. This is an *identification* aim (see McQuarrie and Tsai, 1998, Chapter 1). It would be disastrous to select a model which does not (almost) minimize the bias. And it is besides assumed that the model $\mathcal{M}_{K_0}$ contains all the interesting information (typically, the structure of the classes).

Let us stress that the "true" number of components of $f^\wp$ is not directly of concern: it is in particular not assumed that it equals K_0, and 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 particular point of view introduced by using the $\log \mathrm{L_{cc}}$ contrast, which is not density estimation, neither is it identification of the "true" number of components.

Theorem 4.1. *$\{\Theta_K\}_{1\leq K\leq K_M}$ 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 $\Theta_K^0 = \operatorname{argmin}\limits_{\theta\in\Theta_K} \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right]$. Assume*

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

$$\begin{array}{c} \forall K,\ \hat{\theta}_K \in \Theta_K \text{ defined such that } \gamma_n(\hat{\theta}_K) \leq \gamma_n(\theta_K^0) + o_{\mathbb{P}}(1) \\ \text{fulfills } \gamma_n(\hat{\theta}_K) \xrightarrow{\mathbb{P}} \mathbb{E}_{f^\wp}\left[\gamma(\theta_K^0)\right] \end{array} \tag{B2}$$

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

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

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

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

Sketch of proof. First prove that $\hat{K}$ cannot asymptotically "underestimate" K_0. Suppose $\mathbb{E}_{f^{\wp}}\left[\gamma(\theta_K^0)\right] > \mathbb{E}_{f^{\wp}}\left[\gamma(\theta_{K_0}^0)\right]$. From (B2), $\big(\gamma_n(\hat{\theta}_K) - \gamma_n(\hat{\theta}_{K_0})\big)$ is asymptotically of order $\mathbb{E}_{f^{\wp}}\left[\gamma(\theta_K^0)\right] - \mathbb{E}_{f^{\wp}}\left[\gamma(\theta_{K_0}^0)\right] > 0$. Since the penalty is $o_{\mathbb{P}}(1)$ from (B3), $\text{crit}(K_0) < \text{crit}(K)$ asymptotically and $\hat{K} > K$.

That $\hat{K}$ does not asymptotically "overestimate" K_0, involves the heaviest assumption (B4). It is more involved since then $\big(\mathbb{E}_{f^{\wp}}\left[\gamma(\theta_K^0)\right] - \mathbb{E}_{f^{\wp}}\left[\gamma(\theta_{K_0}^0)\right]\big)$ is zero. The fluctuations of $\big(\gamma_n(\hat{\theta}_K) - \gamma_n(\hat{\theta}_{K_0})\big)$ around zero then have to be evaluated and canceled by a penalty large enough. According to (B4), a penalty larger than $\frac{1}{n}$ should suffice. (B3) guarantees this condition. □

Assumption (B3) defines the range of possible penalties; Assumption (B2) is guaranteed under assumption (A3) of Theorem 3.1:

Lemma 4.1. *For a fixed* K*, assume* (A3)*. Then* (B2) *holds.*

Indeed, asymptoticaly, minimizing $\theta \mapsto \gamma_n(\theta)$ cannot differ much from minimizing $\theta \mapsto \mathbb{E}_{f^{\wp}}\left[\gamma(\theta)\right]$ if they are uniformly close to each other (A3).

Assumption (B4) is the heaviest assumption. Section 4.2 is devoted to deriving sufficient conditions so that it holds. This will justify the

Theorem 4.2. $(\mathcal{M}_K)_{1\le K\le K_M}$ *Gaussian mixture models with compact parameter space* Θ_K *and* $\Theta_K^0 = \operatorname{argmin}_{\theta\in\Theta_K} \mathbb{E}_{f^{\wp}}\left[-\log L_{cc}(\theta)\right]$ *for any* K*. Let* $K_0 = \min \operatorname*{argmin}_{1\le K\le K_M} \mathbb{E}_{f^{\wp}}\left[-\log L_{cc}(\Theta_K^0)\right]$*. Assume* $\forall K, \forall\theta \in \Theta_K, \forall\theta_{K_0}^0 \in \Theta_{K_0}^0$,

$$\mathbb{E}_{f^{\wp}}\left[-\log L_{cc}(\theta)\right] = \mathbb{E}_{f^{\wp}}\left[-\log L_{cc}(\Theta_{K_0}^0)\right]$$
$$\iff -\log L_{cc}(\theta; x) = -\log L_{cc}(\theta_{K_0}^0; x) \quad f^{\wp}d\lambda - a.e. \tag{C1}$$

For any K, let $\Theta_K^{\mathcal{O}} \subset \mathbb{R}^{D_K}$ open over which $\log L_{cc}$ is defined, such that $\Theta_K \subset \Theta_K^{\mathcal{O}}$; assume that $H^M_{\log L_{cc},\Theta_K^{\mathcal{O}},\infty}$ and $H^{M'}_{\log L_{cc},\Theta_K^{\mathcal{O}},2}$ hold and that $\forall \theta_K^0 \in \Theta_K^0$, $I_{\theta_K^0} = \frac{\partial^2}{\partial \theta^2}\left(\mathbb{E}_{f^{\wp}}\left[-\log L_{cc}(\theta)\right]\right)_{|\theta_K^0}$ is nonsingular; let $\widehat{\theta}_K^{MLccE} \in \Theta_K$ with

$$-\log L_{cc}(\widehat{\theta}_K^{MLccE}) \leq -\log L_{cc}(\theta_K^0) + o_{\mathbb{P}}(n).$$

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

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

Select $\hat{K}$ *such that* $\hat{K} = \min \underset{1 \leq K \leq K_M}{\operatorname{argmin}} \left\{-\log L_{cc}(\widehat{\theta}_K^{MLccE}) + \mathrm{pen}(K)\right\}$.

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

If Θ_K is convex, M and M' can be defined as suprema over Θ_K instead of $\Theta_K^{\mathcal{O}}$ and there is no need to introduce the sets $\Theta_K^{\mathcal{O}}$. The new "identifiability" assumption (C1) introduced is reasonable: as expected the label switching phenomenon is no 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 the unique Kullback-Leibler divergence minimizer. Obviously, several parameter values, perhaps in different models, may represent it, besides the label switching. We do not know any such result with the $-\log \mathrm{L_{cc}}$ contrast and hypothesize that the assumption holds.

The assumption about the nonsingularity of I_{θ^0} is unpleasant, since it is hard to be guaranteed. Hopefully, it could be weakened. The result of Massart (2007) (Theorem 7.11) which inspires this, and is available in a standard likelihood context, does not require such an assumption since it does not rely on the study of this link between the contrast and the parameters but on a clever choice of the involved distances (Hellinger distances), and on particular properties of the log function. However, this is a usual assumption (see Redner and Walker, 1984, or below).

Massart (2007) moreover does not require the contrast (i.e. the likelihood) to be bounded, as we have to. 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 about M and M' (the last is much milder than the former) be necessary. They are typically ensured if either the contrast is bounded or if the support of $f^{\wp}$ is bounded.

The conditions about the penalty form are analogous to that of Nishii (1988) or Keribin (2000), which are both derived in the standard maximum likelihood framework. As those of Keribin (2000), they can be regarded as generalizing those of Nishii (1988) when the considered models are Gaussian mixture models. Indeed, Nishii (1988) considers penalties of the form $c_n D_K$ and proves the model selection procedure to be weakly consistent if $\frac{c_n}{n} \to 0$ and $c_n \to \infty$. Note that Nishii (1988) assumes the parameter space to be convex. He moreover notably assumes that $\Theta_K^0 = \{\theta_K^0\}$ and that the counterpart of $I_{\theta_K^0}$ is nonsingular, together with other regularity assumptions. Those results are not particularly designed for mixture models. Instead, as we do, Keribin (2000) considers general penalty forms and proves the procedure 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 $K > K'$. These conditions are equivalent to Nishii's if $\text{pen}(K) = c_n D_K$. In a general mixture model framework, she assumes the model family to be well-specified, the same notion of identifiability as we do, and a condition which does not seem to be directly comparable to ours about $I_{\theta_K^0}$ but which tastes roughly the same. It might be milder. Those assumptions are proved to hold with the standard likelihood contrast for Gaussian mixture models with lower bounded, spherical covariance matrices which are the same for all components, and if the means belong to a compact. Our conditions about the penalty are a little weaker than Keribin's, but they still are quite analogous. Moreover, as compared to those results, we notably have to keep the proportions away from zero. This is necessary because the entropy term must be handled. It does not seem easy to extend the methods used by Keribin (2000) to our framework.

The strong version of Theorem 4.2, which would state the almost sure consistency of $\hat{K}$ to K_0, would then probably involve penalties a little heavier, as Nishii (1988) and Keribin (2000) proved in their respective frameworks: both had to assume that $\frac{\text{pen}(K)}{\log \log n} \to \infty$.

Theorem 4.2 is a direct consequence of Theorem 4.1, Lemma 4.1, Theorem 3.2, which can be applied under those assumptions, and of Corollary 4.2 below and the discussion about its assumptions along the lines of Section 4.2.

4.2 Sufficient Conditions to Ensure Assumption (B4)

Let us introduce the notation $S_n\gamma(\theta) = n\big(\gamma_n(\theta) - \mathbb{E}_{f^{\wp}}[\gamma(\theta)]\big)$. The main result of this section is Lemma 4.2. Some intermediate results which en-

able to link Lemma 4.2 to Theorem 4.1 via Assumption (B4) are stated as corollaries and proved subsequently. Lemma 4.2 povides a control of $\sup_{\theta\in\Theta} \frac{S_n\big(\gamma(\theta^0)-\gamma(\theta)\big)}{\|\theta^0-\theta\|_\infty^2+\beta^2}$ (with respect to β) and then of $\frac{S_n\big(\gamma(\theta^0)-\gamma(\hat\theta)\big)}{\|\theta^0-\hat\theta\|_\infty^2+\beta^2}$. 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 4.1 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 4.2 yields (Corollary 4.2) $n\big(\gamma_n(\theta^0_K)-\gamma_n(\hat\theta_K)\big) = O_{\mathbb{P}}(1)$ for any model $K \in \operatorname{argmin}_{1\le K\le K_M} \mathbb{E}_{f^\wp}\left[\gamma(\theta^0_K)\right]$ 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 4.2. *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 $\mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right] = \inf_{\theta\in\Theta} \mathbb{E}_{f^\wp}\left[\gamma(\theta)\right]$.*
Assume that $H^M_{\gamma,\Theta,\infty}$ and $H^{M'}_{\gamma,\Theta,2}$ hold.
Then $\exists \alpha > 0 / \forall n, \forall \beta > 0, \forall \eta > 0$, with probability larger than $(1-e^{-\eta})$,

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

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

Sketch of proof. The proof relies on results of Massart (2007) and on the evaluation of the bracketing entropy of the class of functions at hand. Lemma 3.3 provides a local control of the entropy and hence, through Theorem 6.8 in Massart (2007), 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 in Massart (2007) 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 Massart (2007), which can be thought of as an application of Markov's inequality. □

Corollary 4.1. *Same assumptions as Lemma 4.2, but the convexity of Θ. Besides assume that $I_{\theta^0} = \frac{\partial^2}{\partial\theta^2}\left(\mathbb{E}_{f^\wp}\left[\gamma(\theta)\right]\right)_{|\theta^0}$ 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 4.2: 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: the order of $\|\theta - \theta^0\|_\infty^2$ with respect to n when the model is fixed is of concern.

The assumption that I_{θ^0} is nonsingular plays an analogous role as Assumption (A2) in Theorem 3.1: this ensures that $\mathbb{E}_{f^\wp}[\gamma(\theta)]$ cannot be close to $\mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right]$ if θ is not close to θ^0. But this stronger assumption is necessary to strengthen the conclusion: the rate of the relation between $\mathbb{E}_{f^\wp}[\gamma(\theta)] - \mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right]$ and $\|\theta - \theta^0\|$ can then be controlled...

Should this assumption fail, $\exists \tilde{\theta} \in \Theta / \tilde{\theta}' I_{\theta^0} \tilde{\theta} = 0 \Rightarrow \mathbb{E}_{f^\wp}\left[\gamma(\theta^0 + \lambda\tilde{\theta})\right] = \mathbb{E}_{f^\wp}\left[\theta^0\right] + o(\lambda^2)$ and then there is no hope to have $\alpha > 0$ such that $\mathbb{E}_{f^\wp}[\gamma(\theta)] - \mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right] > \alpha\|\theta - \theta^0\|^2$: this approach cannot be applied without this — admittedly unpleasant — assumption. Perhaps an other approach (with distances not involving the parameters but directly the contrast values) might enable to avoid it, as Massart (2007) did in the likelihood framework.

Corollary 4.2. *Let $(\Theta_K)_{1 \le K \le K_M}$ be models with, for any K, $\Theta_K \subset \mathbb{R}^{D_K}$. Assume that $D_1 \le \cdots \le D_{K_M}$. 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 $H^M_{\gamma,\Theta^{\mathcal{O}},\infty}$ and $H^{M'}_{\gamma,\Theta^{\mathcal{O}},2}$ hold. Let, for any K, $\Theta_K^0 = \operatorname{argmin}_{\theta \in \Theta_K} \mathbb{E}_{f^\wp}[\gamma(\theta)]$ and $\theta_K^0 \in \Theta_K^0$.*
Let $K_0 = \min \operatorname{argmin}_{1 \le K \le K_M} \mathbb{E}_{f^\wp}\left[\gamma(\Theta_K^0)\right]$ and assume $\forall K, \forall \theta \in \Theta_K$,

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

Let $\mathcal{K} = \left\{K \in \{1, \dots, K_M\} : \mathbb{E}_{f^\wp}\left[\gamma(\theta_K^0)\right] = \mathbb{E}_{f^\wp}\left[\gamma(\theta_{K_0}^0)\right]\right\}$.
For any $K \in \mathcal{K}$, let $\hat{\theta}_K \in \Theta_K$ such that

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

Assume that $I_{\theta_K^0} = \frac{\partial^2}{\partial \theta^2}\left(\mathbb{E}_{f^\wp}[\gamma(\theta)]\right)_{|\theta_K^0}$ 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 4.1 is ensured.

4.3 A New Light on ICL

The previous section suggests links between model selection penalized criteria with the standard likelihood on the one hand and with the conditional classification likelihood we defined on the other hand. Indeed penalties with the same form as those given by Nishii (1988) or Keribin (2000) with the standard likelihood are proved to be "consistent" in our framework. Therefore, by analogy with the standard likelihood framework, it is expected that penalties proportional to D_K conform an efficiency 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 an efficient procedure by a $\log n$ factor is notified for example by Arlot (2007).

Let us then consider by analogy with BIC the penalized criterion

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

The point is that we almost recover ICL (replace $\widehat{\theta}_K^{\mathrm{MLE}}$ by $\widehat{\theta}^{\mathrm{MLccE}}$ in (2)), which may then be regarded as an approximation of this $\mathrm{L_{cc}}$-ICL criterion. The corresponding penalty is $\frac{\log n}{2} D_K$, and the derivation of $\mathrm{L_{cc}}$-ICL illustrates that the entropy should not be considered as a 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 $\widehat{\theta}_K^{\mathrm{MLccE}}$ differs from $\widehat{\theta}_K^{\mathrm{MLE}}$, the former provides more separated clusters. The compromise between the Gaussian component and the cluster viewpoint is achieved with $\widehat{\theta}_K^{\mathrm{MLccE}}$ from the very estimation step. The user is provided a solution which aims at this compromise for each number of classes K. However, the number of classes selected through $\mathrm{L_{cc}}$-ICL differs seldom from the one selected by ICL in simulations (See Baudry, 2009, Chapter 4).

Finally, $\mathrm{L_{cc}}$-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 $\mathrm{L_{cc}}$-ICL.

5 Discussion

Two families of criteria, in the clustering framework, are distinguished: it is shown that ICL's purpose is of different nature than that of BIC or AIC. The identification theory for the criteria based on the conditional classification likelihood is — not surprisingly — very similar to the one for the

standard likelihood. A major interest of the newly introduced estimator and criteria is to better understand the ICL criterion and the underlying notion of class. This is nor a simple notion of cluster — as for example for the k-means procedure — neither a pure notion of "component" — as underlying the MLE/BIC approach — but a compromise between both. ICL leads to discovering classes matching a subtle combination of the notions of well separated, compact, clusters, and (Gaussian) mixture components. It then enjoys the flexibility and modeling possibilities of the model-based clustering approach, but does not break the expected notion of cluster. Better understanding of the ICL criterion now means better understanding the newly involved contrast $\mathrm{L_{cc}}$.

The choice of the involved mixture components must be handled with care in this framework since it leads the cluster shape underlying the study. Several forms of Gaussian mixtures may be involved: for example, spherical and general models may be compared, or models with free proportions may be compared with models with equal proportions.

Besides it should be further studied how the complexity of the models should be measured when several model kinds are compared. The dimension of the model as a parametric space works for the reported theoretical results. But we are not completely 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 natural.

A further theoretical step would be to drive non-asymptotic results and oracle inequalities. This may give more precise insights about the best penalty shape to use, and then justify the use of the slope heuristics of Birgé and Massart (2007) (see also Baudry et al., 2011 or Baudry, 2009 for simulations and discussions on this topic).

A practical challenge is to provide efficient optimization algorithms. Some work has been done in this direction already: see Baudry et al. (2008) and Baudry (2009, Section 5.1). But they need be improved to be more reliable, and above all to run much faster, which would obviously be a condition for a spread practical use of the new contrast.

A possibility to make this contrast more flexible would be to assign different weights to the log likelihood and the entropy: $\log \mathrm{L_{cc}}_\alpha = \alpha \log \mathrm{L} + (1-\alpha)\,\mathrm{ENT}$, with $\alpha \in [0;1]$. This would enable to tune how important the assignment confidence is with respect to the Gaussian fit... The difficulty would then be to choose α. A first insight which comes in mind is to calibrate α from simulations of situations in which the user knows what solution he expects.

6 Proofs

Proof (Max of h_K, p. 7). If h_K reaches a max. value at $(t_1^0, \dots, t_K^0)$ under the constraint $\sum_{k=1}^K t_k^0 = 1$ then, with $S : (t_1, \dots, t_K) \mapsto \sum_{k=1}^K t_k$,

$$\exists \lambda \in \mathbb{R} / dh_K(t_1^0, \dots, t_K^0) = \lambda dS(t_1^0, \dots, t_K^0).$$

This is equivalent to $\forall k, \log t_k^0 + 1 = \lambda$. Then, $\forall k, \forall k', t_k^0 = t_{k'}^0$ and since $\sum_{k=1}^K t_k^0 = 1$, this yields $t_k^0 = \frac{1}{K}$. □

Proof of Theorem 3.1. Let $\varepsilon > 0$ and $\eta = \inf_{d(\theta, \Theta^0) > \varepsilon} \mathbb{E}_{f^\wp}[\gamma(\theta)] - \mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right] > 0$ (from assumption (A2)). For n large enough and with large probability, from assumption (A3) and the definition of $\hat{\theta}$,

$$\sup_{\theta \in \Theta} \left| \gamma_n(\theta) - \mathbb{E}_{f^\wp}[\gamma(\theta)] \right| < \frac{\eta}{3} \quad \text{and} \quad \gamma_n(\hat{\theta}) \leq \gamma_n(\theta^0) + \frac{\eta}{3}.$$

Then

$$\begin{aligned} \mathbb{E}_{f^\wp}\left[\gamma(\hat{\theta})\right] - \mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right] &\leq \mathbb{E}_{f^\wp}\left[\gamma(\hat{\theta})\right] - \gamma_n(\hat{\theta}) + \gamma_n(\hat{\theta}) - \gamma_n(\theta^0) \\ &\quad + \gamma_n(\theta^0) - \mathbb{E}_{f^\wp}\left[\gamma(\theta^0)\right] \\ &< \eta. \end{aligned}$$

And $d(\hat{\theta}, \Theta^0) < \varepsilon$ with great probability, as n is large enough. □

Proof of Lemma 3.1. 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. For the sake of simplicity, it is assumed without loss of generality, that $\widetilde{\Theta}_\varepsilon \subset \widetilde{\Theta}$. 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 $\left(\frac{\operatorname{diam}\widetilde{\Theta}}{\varepsilon}\right)^D \vee 1$ and covers $\left\{\gamma(\tilde{\theta}) : \tilde{\theta} \in \widetilde{\Theta}\right\}$. □

Example 6.1 (Diagonal Gaussian Mixture Model Parameter Space is Convex)**.** *Following Celeux and Govaert (1995), 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*

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

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 6.2 (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. Let $O_1, \ldots, O_Q$ be a finite covering of Θ consisting of open balls such that $\cup_{q=1}^{Q} O_q \subset \Theta^{\mathcal{O}}$. Such a covering always exists since Θ is assumed to be compact. 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 $\sup_{\theta \in \mathrm{conv}(O_q \cap \Theta)} \left\| \left(\frac{\partial \gamma}{\partial \theta}\right)_{(\theta;x)} \right\|_\infty \leq M'(x)$ since $\mathrm{conv}(O_q \cap \Theta) \subset O_q \subset \Theta^{\mathcal{O}}$. Therefore, Lemma 3.1 applies to $O_q \cap \widetilde{\Theta} \subset \mathrm{conv}(O_q \cap \Theta)$:

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

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,

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

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

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

(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 $\left(\frac{\operatorname{diam}\widetilde{\Theta}}{\varepsilon}\right)^D\vee 1$ and covers $\left\{\gamma(\tilde{\theta}):\tilde{\theta}\in\widetilde{\Theta}\right\}$, which yields Lemma 3.3. □

Proof of Theorem 4.1. Let $\mathcal{K}=\operatorname{argmin}_{1\le K\le K_M}\mathbb{E}_{f^\wp}\left[\gamma(\theta_K^0)\right]$. By assumption, $K_0=\min\mathcal{K}$.

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

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

Then

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

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

Let now $K\in\mathcal{K}$ (hence $K>K_0$). This part of the result is more involved than the first one but at this stage, it is not more difficult to derive: all the difficulty is hidden in the strong assumption (B4)... Indeed, it implies that $\exists V>0$, such that for n large enough and with large probability,

$$n\big(\gamma_n(\hat{\theta}_{K_0})-\gamma_n(\hat{\theta}_K)\big)\le V.$$

Increase n enough so that $n\big(\operatorname{pen}(K)-\operatorname{pen}(K_0)\big)>V$ with large probability (which is possible from assumption (B4)). Then, for n large enough and with large probability,

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

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

Finally, since $\mathbb{P}[\hat{K}\neq K_0]=\sum_{K\notin\mathcal{K}}\mathbb{P}[\hat{K}=K]+\sum_{K\in\mathcal{K},\ K\neq K_0}\mathbb{P}[\hat{K}=K]$, the result follows. □

Proof of Lemma 4.1. For any $\varepsilon > 0$, with large probability and for n large enough:

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

Proof of Lemma 4.2. 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 Massart (2007) to hold). But it can be checked that both those 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, $\mathbb{E}_{f^{\wp}}\left[\sup_{\theta\in\Theta} g(\theta; \mathbf{X})\right] = \mathbb{E}_{f^{\wp}}\left[\sup_{\theta\in\Theta^{\text{count}}} g(\theta; \mathbf{X})\right]$. Remark however that the models which are actually considered are discrete, because of the computation limitations.

Let us introduce the centered empirical process

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

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

$$\forall \theta \in \Theta(\sigma), \mathbb{E}_{f^{\wp}}\left[|\gamma(\theta^0) - \gamma(\theta)|^r\right] \leq \|M'\|_2^2 \|\theta^0 - \theta\|_\infty^2 (2\|M\|_\infty)^{r-2}$$
$$\leq \frac{r!}{2}(\|M'\|_2\sigma)^2 \left(\frac{2\|M\|_\infty}{3}\right)^{r-2}. \quad (5)$$

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 $\mathrm{e}^{H(\delta,\Theta(\sigma))}$ the minimal cardinal of such a C_δ,

$$\mathrm{e}^{H(\delta,\Theta(\sigma))} \leq \left(\frac{\overbrace{\operatorname{diam}\Theta(\sigma)}^{\leq 2\sigma} \|M'\|_2}{\delta} \right)^D \vee 1. \tag{6}$$

Then, according to Theorem 6.8 in Massart (2007),
$\exists \alpha, \forall \varepsilon \in]0,1], \forall A$ measurable such that $\mathbb{P}[A] > 0$,

$$\begin{aligned} \mathbb{E}^A \left[\sup_{\theta\in\Theta(\sigma)} S_n\big(\gamma(\theta^0) - \gamma(\theta)\big) \right] \leq \frac{\alpha}{\varepsilon}\sqrt{n} \int_0^{\varepsilon\|M'\|_2\sigma} \sqrt{H\big(u, \Theta(\sigma)\big)}\, du \\ + 2\Big(\frac{4}{3}\|M\|_\infty + \|M'\|_2\sigma\Big) H\big(\|M'\|_2\sigma, \Theta(\sigma)\big) \\ + (1+6\varepsilon)\|M'\|_2\sigma \sqrt{2n \log\frac{1}{\mathbb{P}[A]}} + \frac{8}{3}\|M\|_\infty \log\frac{1}{\mathbb{P}[A]}. \end{aligned} \tag{7}$$

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 (6), this yields

$$\begin{aligned} \forall t \in \mathbb{R}^+, \int_0^t \sqrt{H(u,\Theta(\sigma))}du \leq \sqrt{D}\int_0^t \sqrt{\log\frac{2\|M'\|_2\sigma}{u} \vee 0}\, du \\ \leq \sqrt{D}\big(t \wedge 2\|M'\|_2\sigma\big)\sqrt{\log\frac{e}{\frac{t}{2\|M'\|_2\sigma}\wedge 1}}, \end{aligned} \tag{8}$$

after a simple substitution.
Next, let us apply Lemma 4.23 in Massart (2007): From (6), (7) and (8),

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

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

As required for Lemma 4.23 in Massart (2007) to hold, $\frac{\varphi(t)}{t}$ is nonincreasing. It follows

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

We then choose $\varepsilon = 1$ and apply Lemma 2.4 in Massart (2007): for any $\eta > 0$ and any $\beta > 0$, with probability larger than $1 - \mathrm{e}^{-\eta}$,

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

Proof of Corollary 4.1. 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 large probability, for $n \ge n_0$, $\hat\theta_n \in B(\theta^0,\varepsilon)$. Now, $B(\theta^0,\varepsilon)$ is convex and the assumptions of the corollary guarantee that Lemma 4.2 applies. Let us apply it to $\hat\theta_n$: $\forall n \ge n_0, \forall \beta > 0$, with great probability as η is large,

$$\frac{S_n\big(\gamma(\theta^0)-\gamma(\hat\theta_n)\big)}{\|\theta^0-\hat\theta_n\|_\infty^2 + \beta^2} \le \frac{\alpha}{\beta^2}\Bigg(\sqrt{nD}\|M'\|_2\beta + \big(\|M\|_\infty + \|M'\|_2\beta\big)D$$
$$+ \|M'\|_2\beta\sqrt{n\eta} + \|M\|_\infty\eta\Bigg). \quad (9)$$

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

$$\mathbb{E}_{f^\wp}[\theta] - \mathbb{E}_{f^\wp}\left[\theta^0\right] = (\theta-\theta^0)' I_{\theta^0}(\theta-\theta^0) + r(\|\theta-\theta^0\|_\infty)\|\theta-\theta^0\|_\infty^2$$
$$\ge \big(2\alpha' + r(\|\theta-\theta^0\|_\infty)\big)\|\theta-\theta^0\|_\infty^2$$

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...),

$$\forall\theta \in B(\theta^0,\varepsilon), \mathbb{E}_{f^\wp}[\theta] - \mathbb{E}_{f^\wp}\left[\theta^0\right] \ge \alpha'\|\theta-\theta^0\|_\infty^2. \quad (10)$$

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

(9) together with (10) 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 large 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 4.2. This is a direct application of Corollary 4.1. Let $K\in\mathcal{K}$: $\mathbb{E}_{f^\wp}\big[\gamma(\theta^0_K)\big]=\mathbb{E}_{f^\wp}\big[\gamma(\theta^0_{K_0})\big]$. Θ_K can be assumed to be convex: if it is not, $\hat\theta_K$ lies in $B(\theta^0_{K_0},\varepsilon)\subset\Theta^{\mathcal{O}}$ with large probability for large n and Θ_K may be replaced by $B(\theta^0_{K_0},\varepsilon)$. According to Lemma 4.2, with probability larger than $(1-\mathrm{e}^{-\eta})$ for n large, with $\beta=\frac{\beta_0}{\sqrt{n}}$ for any $\beta_0>0$:

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

But, according to Corollary 4.1, $n\|\theta^0_K-\hat\theta_K\|_\infty^2 = O_{\mathbb{P}}(1)$. Moreover, by definition,

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

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

References

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

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

Baudry, J., Raftery, A., Celeux, G., Lo, K., and Gottardo, R. (2010). Combining mixture components for clustering. *J. Comput. Graph. Statist.*, 19(2):332–353.

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/`.

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*, pages 337–348, Heidelberg. Physica-Verlag.

Baudry, J.-P., Maugis, C., and Michel, B. (2011). Slope heuristics: overview and implementation. *Statist. Comput.*, 22(2):455–470.

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.

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

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781 – 793.

De Granville, C., Southerland, J., and Fagg, A. (2006). Learning grasp affordances through human demonstration. In *Proceedings of the International Conference on Development and Learning, electronically published.*

Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.*, 97:611–631.

Goutte, C., Hansen, L., Liptrot, M., and Rostrup, E. (2001). Feature-space clustering for fMRI meta-analysis. *Human Brain Mapping*, 13(3):165–183.

Hamelryck, T., Kent, J. T., and Krogh, A. (2006). Sampling realistic protein conformations using local structural bias. *PLoS Comput. Biol.*, 2:e131.

Hennig, C. (2010). Methods for merging Gaussian mixture components. *Adv. Data Anal. Classif.*, 4(1):3–34.

Keribin (2000). Consistent estimation of the order of mixture models. *Sankhya A*, 62(1):49–66.

Mariadassou, M., Robin, S., and Vacher, C. (2010). Uncovering latent structure in valued graphs: a variational approach. *Ann. Appl. Stat.*, 4(2):715–742.

Massart, P. (2007). *Concentration Inequalities and Model Selection.* Lecture Notes in Math. Springer.

Maugis, C. and Michel, B. (2011). A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM Probab. Stat.*, 15:41–68.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* New York : Wiley.

McQuarrie, A. and Tsai, C. (1998). *Regression and time series model selection.* World Scientific.

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

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*, pages 141–150. ACM New York, NY, USA.

Redner, R. and Walker, H. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, 26(2):195 – 239.

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

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

Steele, R. J. and Raftery, A. (2010). Performance of bayesian model selection criteria for gaussian mixture models. In *Frontiers of Statistical Decision Making and Bayesian Analysis*, pages 113–130. Springer.

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite mixture Distributions*. New York : Wiley.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

Acknowledgements I am deeply grateful to G. Celeux (INRIA Saclay Île-de-France and Université Paris-Sud), J.-M. Marin (Université Montpellier II) and P. Massart (Université Paris-Sud) for their essential help.

Jean-Patrick Baudry
Université Pierre et Marie Curie - Paris VI
Boîte 158, Tour 15-25, 2^{e} étage
4 place Jussieu, 75252 Paris Cedex 05
France.
Jean-Patrick.Baudry@upmc.fr
`http://www.lsta.upmc.fr/Baudry`