ESAIM: PS 15 (2011) 41–68
DOI: 10.1051/ps/2009004

ESAIM: Probability and Statistics
www.esaim-ps.org

A NON ASYMPTOTIC PENALIZED CRITERION FOR GAUSSIAN MIXTURE MODEL SELECTION

Cathy Maugis[1] and Bertrand Michel[2]

Abstract. Specific Gaussian mixtures are considered to solve simultaneously variable selection and clustering problems. A non asymptotic penalized criterion is proposed to choose the number of mixture components and the relevant variable subset. Because of the non linearity of the associated Kullback-Leibler contrast on Gaussian mixtures, a general model selection theorem for maximum likelihood estimation proposed by [Massart *Concentration inequalities and model selection* Springer, Berlin (2007). Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23 (2003)] is used to obtain the penalty function form. This theorem requires to control the bracketing entropy of Gaussian mixture families. The ordered and non-ordered variable selection cases are both addressed in this paper.

Mathematics Subject Classification. 62H30, 62G07.

Received September 19, 2008. Revised February 13, 2009.

Introduction

Model-based clustering methods consist of modelling clusters with parametric distributions and considering the mixture of these distributions to describe the whole dataset. They provide a rigorous framework to assess the number of mixture components and to take the variable roles into account. Currently, cluster analysis is more and more concerned with large datasets where observations are described by many variables. This large number of predictor variables could be beneficial to data clustering. Nevertheless, the useful information for clustering can be contained into only a variable subset and some of the variables can be useless or even harmful to choose a reasonable clustering structure. Several authors have suggested variable selection methods for Gaussian mixture clustering which is the most widely used mixture model for clustering multivariate continuous datasets. These methods are called "wrapper" since they are included into the clustering process. Law *et al.* [21] have introduced the feature saliency concept. Regardless of cluster membership, relevant variables are assumed to be independent of the irrelevant variables which are supposed to have the same distribution. Raftery and Dean [28] recast variable selection for clustering into a global model selection problem. Irrelevant variables are explained by all the relevant clustering variables according to a linear regression. The comparison between two nested variable subsets is performed using Bayes factor. A variation of this method is proposed by Maugis *et al.* [26] where irrelevant variables can only depend on a relevant clustering variable subset and variables can have different sizes (variable blocks). Since all these methods are based on a variable selection procedure

Keywords and phrases. Model-based clustering, variable selection, penalized likelihood criterion, bracketing entropy.

[1] Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse, 135 avenue de Rangueil, 31077 Toulouse Cedex 4, France; cathy.maugis@insa-toulouse.fr

[2] Laboratoire de Statistique Théorique et Appliquée, Université Paris 6, 175 rue du Chevaleret, 75013 Paris, France; bertrand.michel@upmc.fr

Article published by EDP Sciences

included into the clustering process, they do not impose specific constraints on Gaussian mixture forms. On the contrary, Bouveyron *et al.* [12] consider a suitable Gaussian mixture family to take that data are in low-dimensional subspaces hidden in the original space into account. However, since this dimension reduction is based on principal components, it is difficult to deduce from this approach an interpretation of the variable roles. In all these methods, an asymptotic criterion is used to solve the underlying model selection problem.

In this paper, a modelling taken the variable role for clustering process into account recasts variable selection and clustering problems into a model selection problem in a density estimation framework. Suppose that we observe a sample from an unknown probability distribution with density s. A specific collection of models is defined: a model $\mathcal{S}_{(K,\mathbf{v})}$ corresponds to a particular clustering situation with K clusters and a clustering "relevant" variable subset $\mathbf{v}$. A density t in $\mathcal{S}_{(K,\mathbf{v})}$ has the following form: its projection on the relevant variable space is a Gaussian mixture density with K components and its projection on the space of the other variables is a multidimensional Gaussian density. Definitions of models $\mathcal{S}_{(K,\mathbf{v})}$ are precised in Section 1.1. The problem can be recast into the selection of a model among the model collection since this choice automatically leads to a data clustering and a variable selection. We propose a penalized criterion to solve this model selection problem with a non asymptotic point of view. In this approach, the "best" model is the one whose the associated maximum likelihood estimator of s gives the lowest estimation error.

In the density estimation framework, the principle of selecting a model by penalizing a loglikelihood type criterion has emerged during the seventies. Akaike [1] proposed the AIC criterion (Akaike's information criterion) and Schwarz [29] suggested the BIC (Bayesian Information Criterion). These two classical criteria assume implicitly that the true distribution belongs to the model collection (see for instance [13]). With a different point of view, the criterion ICL (Integrated Completed Likelihood) proposed by Biernacki *et al.* [6] takes the clustering aim into account. Although the behaviours of these asymptotic criteria were tested in practice, there are little proved theoretical properties. For instance, the BIC consistency is only stated for the assessing cluster number under restrictive regularity assumptions and assuming that the true density belongs to the considered Gaussian mixture family [20].

A non asymptotic approach for model selection *via* penalization has emerged during the last ten years, mainly with works of Birgé and Massart [11] and Barron *et al.* [4]. An overview is available in [24]. The aim of this approach is to define penalized data-driven criteria which lead to oracle inequalities. The belonging of the true density to the model collection is not required. The penalty function depends on the number of parameters of each model and also on the complexity of the whole model collection. This approach has been carried out in several frameworks where penalty functions are explicitly assessed. In our context, a general model selection theorem for maximum likelihood estimation (MLE) is used to obtain a penalized criterion and an associated oracle inequality. This theorem proposed by Massart [24] is a version of Theorem 2 in [4]. Its application requires to control the bracketing entropy of the considered Gaussian mixture models.

The paper is organized as follows: Section 1 gives the model selection principles. The Gaussian mixture models considered in this paper are described in Section 1.1 and principles of non asymptotic theory for density estimation based on Kullback-Leibler contrast are reviewed in Section 1.2. This section is completed by the statement of the general model selection theorem. The bracketing entropy results for the Gaussian mixture families and the main results which define the non asymptotic penalized criterion for the ordered and the non-ordered cases are stated in Section 2. Appendix A is devoted to the proof of these main results. In Appendix B, the required control of the bracketing entropy of the mixture families is recast into a control of the bracketing entropy of Gaussian density families. This bracketing entropy is upper bounded for different Gaussian mixture forms. A discussion is given in Section 3.

1. Model selection principles

1.1. **Framework**

Centered observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, with $\mathbf{y}_i \in \mathbb{R}^Q$ are assumed to be a sample from a probability distribution with unknown density s. This target s is proposed to be estimated by a finite mixture model in a clustering purpose. Note that s itself is not assumed to be a Gaussian mixture density. Model-based clustering consists of assuming that the data come from a source with several subpopulations, modelled separately and the overall population is a mixture of them. The resulting model is a finite mixture model. When the data are multivariate continuous observations, the parameterized component density is usually a multidimensional Gaussian density. Thus, a Gaussian mixture density with K components is written

$$\sum_{k=1}^{K} p_k \Phi(.|\eta_k, \Lambda_k)$$

where the p_k's are the mixing proportions ($\forall k = 1, \dots, K,\ 0 < p_k < 1$ and $\sum_{k=1}^{K} p_k = 1$) and $\Phi(.|\eta_k, \Lambda_k)$ denotes the Q-dimensional Gaussian density with mean η_k and variance matrix Λ_k. The parameter vector is $(p_1, \dots, p_K, \eta_1, \dots, \eta_K, \Lambda_1, \dots, \Lambda_K)$.

The mixture model is an incomplete data structure model: The complete data are $((\mathbf{y}_1, \mathbf{z}_1), \dots, (\mathbf{y}_n, \mathbf{z}_n))$ where the missing data are $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ with $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$ such that $z_{ik} = 1$ if and only if $\mathbf{y}_i$ arises from the component k. The vector $\mathbf{z}$ defines an ideal clustering of the data $\mathbf{y}$ associated to the mixture model. After an estimation of the parameter vector thanks to the EM algorithm [17], a data clustering is deduced from the maximum a posteriori principle:

$$\hat{z}_{ik} = \begin{cases} 1 & \text{if } \hat{p}_k \Phi(\mathbf{y}_i|\hat{\eta}_k, \hat{\Lambda}_k) > \hat{p}_l \Phi(\mathbf{y}_i|\hat{\eta}_l, \hat{\Lambda}_l),\ \forall l \neq k \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

Currently, statistics deals with problems where data are explained by many variables. In principle, the more information we have about each individual, the better a clustering method is expected to perform. Nevertheless, some variables can be useless or even harmful to obtain a good data clustering. Thus, it is important to take the variable role in the clustering process into account. To this aim, Gaussian mixtures with a specific form are considered. On irrelevant variables, data are assumed to have an homogeneous behavior around the null mean (centered data) allowing not to distinguish a possible clustering. Hence the data density is modelled by a spherical Gaussian joint law with null mean vector on these variables. On the contrary, the different component mean vectors are free on relevant variables. Moreover, the variance matrices restricted on relevant variables are either taken completely free or are chosen in a specified set of positive definite matrices.

This modelling idea is now formalized. Let $\mathcal{V}$ be the collection of the nonempty subsets of $\{1, \dots, Q\}$. A Gaussian mixture family is characterized by its number of mixture components $K \in \mathbb{N}^*$ and its relevant variable index subset $\mathbf{v} \in \mathcal{V}$ whose cardinal is denoted α. In the sequel, the set of index couples $(K, \mathbf{v})$ is $\mathcal{M} = \mathbb{N}^* \times \mathcal{V}$. Consider the decomposition of a vector $x \in \mathbb{R}^Q$ into its restriction on relevant variables $x_{[\mathbf{v}]} = (x_{j_1}, \dots, x_{j_\alpha})'$ and its restriction on irrelevant variables $x_{[\mathbf{v}^c]} = (x_{l_1}, \dots, x_{l_{Q-\alpha}})'$ where $\mathbf{v} = \{j_1, \dots, j_\alpha\}$ and $\mathbf{v}^c = \{l_1, \dots, l_{Q-\alpha}\} = \{1, \dots, Q\} \backslash \mathbf{v}$. On relevant variables, a Gaussian mixture f is chosen among the following mixture family

$$\mathcal{L}_{(K,\alpha)} = \left\{ \sum_{k=1}^{K} p_k \Phi(.|\mu_k, \Sigma_k);\ \begin{array}{l} \forall k,\ \mu_k \in [-a, a]^\alpha,\ (\Sigma_1, \dots, \Sigma_K) \in \mathcal{D}^+_{(K,\alpha)} \\ 0 < p_k < 1, \sum_{k=1}^{K} p_k = 1 \end{array} \right\}$$

where $a \in \mathbb{R}^*_+$ and $\mathcal{D}^+_{(K,\alpha)}$ denotes a family of K-tuples of $\alpha \times \alpha$ symmetric positive definite matrices whose eigenvalues are bounded. The family $\mathcal{D}^+_{(K,\alpha)}$ is related to the Gaussian mixture shape and the associated set of K-tuples of Gaussian densities composing mixtures of $\mathcal{L}_{(K,\alpha)}$ is denoted $\mathcal{F}_{(K,\alpha)}$. These notations are specified hereafter. On irrelevant variables, the data density is modelled by a spherical Gaussian density g belonging to the following family

$$\mathcal{G}_{(\alpha)} = \left\{ \Phi(.|0, \omega^2 I_{Q-\alpha});\ \omega^2 \in [\lambda_m, \lambda_M] \right\} \tag{1.2}$$

where $0 < \lambda_m < \lambda_M$. Finally, the family of Gaussian mixtures associated to $(K, \mathbf{v}) \in \mathcal{M}$ is defined by

$$\mathcal{S}_{(K,\mathbf{v})} = \left\{x \in \mathbb{R}^Q \mapsto f(x_{[\mathbf{v}]})\, g(x_{[\mathbf{v}^c]});\ f \in \mathcal{L}_{(K,\alpha)},\ g \in \mathcal{G}_{(\alpha)}\right\}. \tag{1.3}$$

The dimension of the model $\mathcal{S}_{(K,\mathbf{v})}$ is denoted $D(K,\alpha)$ and corresponds to the number of free parameters common to all Gaussian mixtures in this model. It only depends on the number of components K and the number of relevant variables α. Note that a density of $\mathcal{S}_{(K,\mathbf{v})}$ can be written as a global Gaussian mixture with mean vectors $\eta_k = (\mu_k, 0, \ldots, 0)$ and block-diagonal variance matrices Λ_k with diagonal-blocks Σ_k and $\omega^2 I_{Q-\alpha}$. A data clustering can be deduced from such a Gaussian mixture using the MAP rule (see Eq. (1.1)).

In this paper, four collections of Gaussian mixtures are considered. For each collection, constraints are imposed on the variance matrices of the K Gaussian densities constituting a K-tuple of $\mathcal{F}_{(K,\alpha)}$. This implies a specific shape for mixtures of the associated family $\mathcal{L}_{(K,\alpha)}$. The Gaussian mixture notation for those four collections is taken from [7].

- For the $[L_kB_k]$ collection, the variance matrices on relevant variables are assumed to be diagonal and free, and their eigenvalues belong to the interval $[\lambda_m, \lambda_M]$. Thus the relevant variables are independent conditionally to mixture component belonging. In this context, an element of $\mathcal{F}_{(K,\alpha)}$ is composed of K Gaussian densities belonging to the following set

$$\mathcal{F}_{(\alpha)} = \left\{\Phi(.|\mu,\Sigma);\ \mu \in [-a,a]^\alpha,\ \Sigma = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_\alpha^2),\ \sigma_1^2,\ldots,\sigma_\alpha^2 \in [\lambda_m,\lambda_M]\right\} \tag{1.4}$$

and the dimension $D(K,\alpha)$ of model $\mathcal{S}_{(K,\mathbf{v})}$ is equal to $K(2\alpha+1)$.

- For the $[L_kC_k]$ collection, the variance matrices are assumed to be totally free. They belong to the set $\mathcal{D}^+_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_M)$ of $\alpha \times \alpha$ positive definite matrices with eigenvalues in the interval $[\lambda_m, \lambda_M]$. The relevant variables are thus admitted to be correlated conditionally to mixture component belonging. The set $\mathcal{F}_{(K,\alpha)}$ composing mixtures can be assimilated to the Gaussian density family

$$\mathcal{F}_{(\alpha)} = \left\{w \in \mathbb{R}^\alpha \mapsto \Phi(w|\mu,\Sigma),\ \mu \in [-a,a]^\alpha,\ \Sigma \in \mathcal{D}^+_{(\alpha)}(\lambda_{\mathrm{m}},\lambda_M)\right\} \tag{1.5}$$

and the dimension of $\mathcal{S}_{(K,\mathbf{v})}$ is equal to $D(K,\alpha) = K\left\{1+\alpha+\frac{\alpha(1+\alpha)}{2}\right\}$.

- For the $[LB_k]$ collection, the variance matrices are assumed to be diagonal and to have the same volume *i.e.* $\forall k \neq k', |\Sigma_k|^{\frac{1}{\alpha}} = |\Sigma_{k'}|^{\frac{1}{\alpha}}$. The variance matrices are decomposed into $\Sigma_k = \beta B_k$ where the common volume β belongs to $[\beta_m, \beta_M]$ and B_k is a diagonal matrix with a determinant 1 and with diagonal coefficients in the interval $[\lambda_m, \lambda_M]$. Thus the family of K-tuples of Gaussian densities composing mixtures of $\mathcal{L}_{(K,\alpha)}$ is

$$\mathcal{F}_{(K,\alpha)} = \left\{(\Phi(.|\mu_1,\beta B_1),\ldots,\Phi(.|\mu_K,\beta B_K));\ \begin{array}{l}\forall 1 \leq k \leq K,\ \mu_k \in [-a,a]^\alpha,\ B_k \in \Delta^1_{(\alpha)}(\lambda_m,\lambda_M),\\ \beta \in [\beta_m,\beta_M]\end{array}\right\} \tag{1.6}$$

where $\Delta^1_{(\alpha)}(\lambda_m,\lambda_M)$ is the set of $\alpha\times\alpha$ diagonal matrices with determinant 1 and whose eigenvalues are in the interval $[\lambda_m,\lambda_M]$ where $0<\lambda_m<\lambda_M$. Here, the model dimension is equal to $D(K,\alpha) = 2K\alpha+1$.

- For the $[LC]$ collection, the variance matrices are all equal to a free positive definite matrix Σ whose eigenvalues are assumed to be in the interval $[\lambda_m,\lambda_M]$. The set of such variance matrices is denoted $\mathcal{D}^+_{(\alpha)}(\lambda_m,\lambda_M)$. The family $\mathcal{F}_{(K,\alpha)}$ is thus defined by

$$\mathcal{F}_{(K,\alpha)} = \{(\Phi(.|\mu_1,\Sigma),\ldots,\Phi(.|\mu_K,\Sigma));\ \mu_k \in [-a,a]^\alpha,\ \Sigma \in \mathcal{D}^+_{(\alpha)}(\lambda_m,\lambda_M)\} \tag{1.7}$$

and the model dimension is $D(K,\alpha) = K(1+\alpha)+\frac{\alpha(\alpha+1)}{2}$.

Note that the family $\mathcal{F}_{(K,\alpha)}$ cannot be assimilated to a Gaussian density set $\mathcal{F}_{(\alpha)}$ for the $[LB_k]$ and $[LC]$ collections since the variance matrices have a common characteristic: the variance matrices have a common volume for the $[LB_k]$ collection and are equal for the $[LC]$ collection. This distinction with the two other collections will be important to construct the penalized criterion. It is interesting to consider these different collections since the results obtained further are stated in function of the model dimension. Whereas a mixture can belong to different collections, its number of free parameters is different according to the mixture shape. Furthermore, the consideration of different Gaussian mixture collections will allow for a larger practical use of our results. To make easier the reading of this paper, the same notation $\mathcal{S}_{(K,\mathbf{v})}$ is used for the four model collections. Finally in order to extend the application field, the cases of ordered and non-ordered variables are both addressed in this paper. If variables are assumed to be ordered, the relevant variable subset is $\mathbf{v} = \{1, \dots, \alpha\}$ and can be assimilated to its cardinal α. Thus, in order to distinguish between these two cases, Gaussian mixture families are denoted $\mathcal{S}_{(K,\alpha)}$ when variables are assumed to be ordered.

These Gaussian mixture families allow us to recast clustering and variable selection problems into a global model selection problem. A criterion is now required to select the best model according to the dataset. We propose a penalized criterion using a non asymptotic approach whose principles are given in the following section.

1.2. Non asymptotic model selection

Density estimation deals with the problem of estimating an unknown distribution corresponding to the observation of a sample $\mathbf{y}$. In many cases, it is not easy to choose a model of adequate dimension. For instance, a model with few parameters tends to be efficiently estimated whereas it could be far from the true distribution. In the opposite situation, a more complex model easily fits data but estimates have larger variances. The aim of model selection is to construct a data-driven criterion to select a model of proper dimension among a model collection. A general theory on this topic, with a non asymptotic approach is proposed in the works of Birgé and Massart (see for instance [8,9]). This model selection principle is now described in our density estimation framework.

Let $\mathcal{S}$ be the set of all densities with respect to the Lebesgue measure on $\mathbb{R}^Q$. The contrast $\gamma(t,.) = -\ln\{t(.)\}$ is considered, leading to the maximum likelihood criterion. The corresponding loss function is the Kullback-Leibler information. It is defined for two densities s and t in $\mathcal{S}$ by

$$\mathrm{KL}(s,t) = \int \ln\left\{\frac{s(x)}{t(x)}\right\} s(x)\,\mathrm{d}x$$

if $s dx$ is absolutely continuous with respect to $t dx$ and $+\infty$ otherwise. The density s being the unique minimizer of the Kullback-Leibler function on $\mathcal{S}$, it satisfies

$$s = \operatorname*{argmin}_{t\in\mathcal{S}} \int -\ln\{t(x)\}s(x)\,\mathrm{d}x.$$

Consequently, s is also a minimizer over $\mathcal{S}$ of the expectation of the empirical contrast defined by

$$\gamma_n(t) = -\frac{1}{n}\sum_{i=1}^{n} \ln\{t(\mathbf{y}_i)\}.$$

A minimizer of the empirical contrast γ_n over a model S, a subspace of $\mathcal{S}$, is denoted $\hat{s}$. Substituting the empirical criterion γ_n to its expectation and minimizing γ_n on S, it is expected to obtain a sensible estimator of s, at least if s belongs (or is close enough) to model S.

A countable collection of models $(S_m)_{m\in\mathcal{M}}$ with a corresponding collection $(\hat{s}_m)_{m\in\mathcal{M}}$ of estimators is considered. The best model is the one presenting the smallest risk

$$m(s) = \underset{m\in\mathcal{M}}{\operatorname{argmin}}\ \mathbb{E}[\mathrm{KL}(s,\hat{s}_m)].$$

However the function $\hat{s}_{m(s)}$, called oracle, is unknown since it depends on the true density s. Nevertheless, this oracle is a benchmark: a data-driven criterion is then found to select an estimator such that its risk is close to the oracle risk. The model selection *via* penalization procedure consists of considering some proper penalty function $\mathrm{pen} : m \in \mathcal{M} \mapsto \mathrm{pen}(m) \in \mathbb{R}^+$ and of selecting $\hat{m}$ minimizing the associated penalized criterion

$$\mathrm{crit}(m) = \gamma_n(\hat{s}_m) + \mathrm{pen}(m).$$

The resulting selected estimator is denoted $\hat{s}_{\hat{m}}$. The final purpose of this non asymptotic approach is to obtain a penalty function and an associated oracle inequality, allowing to compare the risk of the penalized MLE $\hat{s}_{\hat{m}}$ with the benchmark $\inf\limits_{m\in\mathcal{M}} \mathbb{E}[\mathrm{KL}(s,\hat{s}_m)]$.

Commonly, in order to find a suitable penalty function, one begins by writing the following inequality (see p. 9 in [24]): for all $m \in \mathcal{M}$ and $s_m \in S_m$,

$$\mathrm{KL}(s,\hat{s}_{\hat{m}}) \leq \mathrm{KL}(s,s_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + \bar{\gamma}_n(s_m) - \bar{\gamma}_n(\hat{s}_{\hat{m}})$$

where $\bar{\gamma}_n$ is the centered empirical process defined by $\bar{\gamma}_n(t) = \gamma_n(t) - \mathbb{E}[\gamma_n(t)]$. The penalty function has to be chosen to annihilate the fluctuation of $\bar{\gamma}_n(s_m) - \bar{\gamma}_n(\hat{s}_{\hat{m}})$. The aim is to obtain an uniform control of $\bar{\gamma}_n(s_m) - \bar{\gamma}_n(\hat{s}_{m'})$ with respect to m' in $\mathcal{M}$. This quantity is controlled by its expectation using a Talagrand's inequality (see [24,31,32] for an overview). Next, two different situations occur. In some situations, the expectation in the Talagrand's inequality can be efficiently connected to the model dimension, and an oracle inequality with explicit constants is deduced. This is the case in the works of Castellan in the context of histogram density estimation [14] and of density estimation *via* exponential model [15]. For situations when these sharp calculations are impossible to obtain, Massart [24] proposes a general theorem which gives the form of penalties and associated oracle inequalities in terms of the Kullback-Leibler and Hellinger losses. This theorem is based on the centered process control with the bracketing entropy, allowing to evaluate the "size" of models. For Gaussian mixture models, we can only follow the second alternative because of the non linear behavior of the logarithm function on Gaussian mixture densities. Moreover, being impossible to bound uniformly all the ratios of two Gaussian mixtures in our context, a hypothesis of boundness as for all $t \in S_{m'}$ $\|\bar{\gamma}_n(s_m) - \bar{\gamma}_n(t)\|_\infty$ is bounded by a constant, which is required to apply concentration inequalities, cannot be fulfilled.

Before stating the general MLE selection model theorem (Thm. 7.11 in [24]) in a restricted form which is sufficient for our study, the definition of the Hellinger distance and some notation are specified. The norm $\|\sqrt{f} - \sqrt{g}\|_2$ between two nonnegative functions f and g of $\mathbb{L}_1$ is denoted $d_H(f,g)$. We note that if f and g are two densities with respect to the Lebesgue measure on $\mathbb{R}^Q$, $d_H(f,g)$ is the Hellinger distance between f and g. In the following, $d_H(f,g)$ is improperly called Hellinger distance even if f and g are not density functions. An ε-bracketing for a subset S of $\mathcal{S}$ with respect to d_H is a set of integrable function pairs $(l_1,u_1),\ldots,(l_N,u_N)$ such that for each $f \in S$, there exists $j \in \{1,\ldots,N\}$ such that $l_j \leq f \leq u_j$ and $d_H(l_j,u_j) \leq \varepsilon$. The bracketing number $\mathcal{N}_{[.]}(\varepsilon,S,d_H)$ is the smallest number of ε-brackets necessary to cover S and the bracketing entropy is defined by $\mathcal{H}_{[.]}(\varepsilon,S,d_H) = \ln\left\{\mathcal{N}_{[.]}(\varepsilon,S,d_H)\right\}$. Since $\mathcal{S}$ is the density set, the bracket extremities can be chosen as nonnegative functions in $\mathbb{L}_1$.

Let $(S_m)_{m\in\mathcal{M}}$ be some at most countable collection of models, where for each $m \in \mathcal{M}$, the elements of S_m are assumed to be probability densities with respect to Lebesgue measure. Firstly, the following separability assumption allows to avoid measurability problems. For each model S_m, assume that there exists some countable subset S'_m of S_m such that for all $t \in S_m$, there exists a sequence $(t_k)_{k\geq 1}$ of elements of S'_m such that for $x \in \mathbb{R}^Q$, $\ln\{t_k(x)\}$ tends to $\ln\{t(x)\}$ when k tends to infinity. Secondly $\sqrt{\mathcal{H}_{[.]}(\varepsilon,S_m,d_H)}$ is assumed to be integrable at 0 for each m and we also assume that there exists a function Ψ_m on $\mathbb{R}_+$ fulfilling the following properties

[I]. Ψ_m is nondecreasing, $x \to \Psi_m(x)/x$ is nonincreasing on $]0,+\infty[$ and for $\xi \in \mathbb{R}_+$ and all $u \in S_m$, denoting $S_m(u,\xi) = \{t \in S_m ; d_H(t,u) \leq \xi\}$,

$$\int_0^{\xi} \sqrt{\mathcal{H}_{[.]}(x, S_m(u,\xi), d_H)}\,\mathrm{d}x \leq \Psi_m(\xi).$$

Theorem 1.1 (Massart [24])**.** *Let $\mathbf{y}_1,\dots,\mathbf{y}_n$ be i.i.d. random variables with unknown density s with respect to Lebesgue measure on $\mathbb{R}^Q$. Let $(S_m)_{m\in\mathcal{M}}$ be some at most countable collection of models fulfilling the previous properties and let $(\hat{s}_m)_{m\in\mathcal{M}}$ be the corresponding collection of MLEs. Let $(\rho_m)_{m\in\mathcal{M}}$ be some family of nonnegative numbers such that*

$$\sum_{m\in\mathcal{M}} \mathrm{e}^{-\rho_m} = \Upsilon < \infty.$$

*For every $m \in \mathcal{M}$, considering Ψ_m with properties **[I]**, ξ_m denotes the unique positive solution of the equation*

$$\Psi_m(\xi) = \sqrt{n}\,\xi^2.$$

Let $\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$ *and consider the penalized loglikelihood criterion*

$$\mathrm{crit}(m) = \gamma_n(\hat{s}_m) + \mathrm{pen}(m).$$

Then, there exists some absolute constants κ and C such that whenever for all $m \in \mathcal{M}$,

$$\mathrm{pen}(m) \geq \kappa \left(\xi_m^2 + \frac{\rho_m}{n}\right)$$

some random variable $\hat{m}$ minimizing crit *over $\mathcal{M}$ does exist and moreover, whatever the density s,*

$$\mathbb{E}\left[d_H^2(s,\hat{s}_{\hat{m}})\right] \leq C\left[\inf_{m\in\mathcal{M}}\{\mathrm{KL}(s,S_m) + \mathrm{pen}(m)\} + \frac{\Upsilon}{n}\right], \tag{1.8}$$

where $\mathrm{KL}(s,S_m) = \inf_{t\in S_m} \mathrm{KL}(s,t)$ *for every $m \in \mathcal{M}$.*

Inequality (1.8) is not exactly an oracle inequality since the Hellinger risk is upper bounded by the Kullback bias $\mathrm{KL}(s,S_m)$. Nevertheless, this last term is of the order of $d_H^2(s,S_m)$ if $\ln(\|s/t\|_\infty)$ is uniformly bounded on $\cup_{m\in\mathcal{M}} S_m$ according to Lemma 7.23 in [24]. In our context, this condition can be achieved if all densities are assumed to be bounded and defined on a compact support.

2. Main results

As announced previously, Theorem 1.1 is applied in our specific framework described in Section 1.1. The ensuing theoretical results are addressed, for the ordered and non-ordered variable cases in Sections 2.2 and 2.3 respectively. For each one, a non asymptotic penalized criterion is provided to select the number of mixture components K and the variable subset $\mathbf{v}$. Moreover, these results give an oracle inequality which is fulfilled by the associated penalized estimator. The main difficulty to prove these two theoretical results lies in bounding the bracketing entropy of Gaussian mixture families $\mathcal{S}_{(K,\mathbf{v})}$ in order to apply Theorem 1.1. The bracketing entropy results established in this paper for the four Gaussian mixture collections are now addressed in Section 2.1.

2.1. Bracketing entropy results

In this paper, we prove the following result allowing to control the bracketing entropy of our models $\mathcal{S}_{(K,\mathbf{v})}$ for the four multidimensional Gaussian mixture collections. Until now, bracketing entropy results have been proposed for only unidimensional Gaussian mixtures by Ghosal and van der Vaart [19] and Genovese and Wasserman [18], in order to obtain convergence rates in Hellinger distance for density estimation.

Theorem 2.1. *For the four Gaussian mixture collections, the bracketing entropy of $\mathcal{S}_{(K,\mathbf{v})}$ is upper bounded by*

$$\forall \varepsilon \in (0,1],\ \mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq \mathcal{I} + D(K,\alpha) \ln\left(\frac{1}{\varepsilon}\right)$$

where the constant $\mathcal{I}$ is an explicit function of K, α, Q and parameters λ_m, λ_M, a, and also β_m, β_M for the $[LB_k]$ collection.

Appendix B is devoted to the proof of Theorem 2.1. This proof is inspired by the work of Genovese and Wasserman [18]. The key idea, presented in Appendix B.1, consists of recasting the control of the bracketing entropy of $\mathcal{S}_{(K,\mathbf{v})}$ into the control of the bracketing entropies of the associated mixture component density families. Appendices B.2, B.4, B.3 and B.5 are then devoted to the bracketing entropy control of Gaussian density families $\mathcal{F}_{(\alpha)}$ for the $[L_kB_k]$ and $[L_kC_k]$ collections and of $\mathcal{F}_{(K,\alpha)}$ for the $[LB_k]$ and $[LC]$ collections respectively. In each of these appendices, Theorem 2.1 is stated with the explicit definition of the constant $\mathcal{I}$.

Note that in order to apply Theorem 1.1, the local bracketing entropy $\mathcal{H}_{[.]}(x, \mathcal{S}_{(K,\mathbf{v})}(u,\xi), d_H)$ has to be controlled. Nevertheless, it is difficult to characterize the subset $\mathcal{S}_{(K,\mathbf{v})}(u,\xi)$ in function of the parameters of its mixtures. Therefore a global study of the entropy bracketing is proposed in the theorem proof since $\mathcal{H}_{[.]}(x, \mathcal{S}_{(K,\mathbf{v})}(u,\xi), d_H) \leq \mathcal{H}_{[.]}(x, \mathcal{S}_{(K,\mathbf{v})}, d_H)$.

2.2. Ordered variable case

In this section, variables are assumed to be ordered and the model collection is denoted $(\mathcal{S}_{(K,\alpha)})_{(K,\alpha)\in\mathcal{M}}$. In the four cases of Gaussian mixtures, the following theorem gives the form of penalty functions and the associated oracle inequalities. This theorem is proved in Appendix A.1.

Theorem 2.2. *For the four Gaussian mixture collections, there exists two absolute constants κ and C such that, if*

$$\text{pen}(K,\alpha) \geq \kappa \frac{D(K,\alpha)}{n} \left\{ 1 + 2\,\mathcal{A}^2 + \ln\left(\frac{1}{1 \wedge \frac{D(K,\alpha)}{n} \mathcal{A}^2} \right) \right\}$$

where the constant $\mathcal{A}$ is a function of Q, λ_m, λ_M, a, and also β_m, β_M for the $[LB_k]$ collection, such that $\mathcal{A}^2 = O(\sqrt{\ln Q})$ as Q tends to infinity, then the model $(\hat{K}, \hat{\alpha})$ minimizing

$$\text{crit}(K,\alpha) = \gamma_n(\hat{s}_{(K,\alpha)}) + \text{pen}(K,\alpha)$$

over $\mathcal{M}$ exists and

$$\mathbb{E}\left[d_H^2(s, \hat{s}_{(\hat{K},\hat{\alpha})})\right] \leq C \left[\inf_{(K,\alpha)\in\mathcal{M}} \{\text{KL}(s, \mathcal{S}_{(K,\alpha)}) + \text{pen}(K,\alpha)\} + \frac{1}{n} \right] \cdot$$

Several remarks can be given about this result. First, the deduced penalty function has an expected form since it is proportional to the model dimension $D(K,\alpha)$. This shows the interest of considering separately the four collections since the model dimensions are different. For instance, for the $[L_kB_k]$ mixture family, the risk bound is less accurate when this family is considered as a subset of the $[L_kC_k]$ collection. Second, the constant $\mathcal{A}$ is made explicit in the theorem proof (see Appendix A.1) and its expression is different for each mixture collection (see Eqs. (A.1), (A.2) and (A.3)). It depends on parameters λ_m, λ_M, a, Q and also β_m, β_M for the $[LB_k]$ collection and $\mathcal{A}^2 = O(\sqrt{\ln Q})$ as Q tends to infinity. This number of variables Q has to have a reasonable order in the constant $\mathcal{A}$ so that the upper bound in the oracle inequality remains meaningful. Contrary to classical criteria for which Q is fixed and n tends to infinity, our result allows to study cases for which Q increases with n. For specific clustering problems where the number of variables Q is of the order of n or even larger than n, the oracle inequality is still significant. Third, since the multiplicative constants are not explicit, a practical method is necessary to calibrate the penalty function. This is addressed in our companion paper [27].

2.3. Non-ordered variable case

Theorem 2.2 can be generalized to the non-ordered variable case. In this context, a model $\mathcal{S}_{(K,\mathbf{v})}$ is characterized by its number of mixture components K and its subset $\mathbf{v} \in \mathcal{V}$ of relevant variable indexes. This model is related to the model $\mathcal{S}_{(K,\alpha)}$ of the ordered case by

$$\mathcal{S}_{(K,\mathbf{v})} = \{x \in \mathbb{R}^Q \mapsto f \circ \tau(x),\ f \in \mathcal{S}_{(K,\alpha)}\}$$

where τ is a permutation such that $(\tau(x)_1, \ldots, \tau(x)_\alpha)' = x_{[\mathbf{v}]}$ and has the same dimension $D(K, \alpha)$. Consequently, the model $\mathcal{S}_{(K,\mathbf{v})}$ has the same complexity as $\mathcal{S}_{(K,\alpha)}$ and thus has the same bracketing entropy. However, the model set $\{\mathcal{S}_{(K,\mathbf{v})}\}_{(K,\mathbf{v})\in\mathcal{M}}$ contains more models per dimension than in the ordered case. This richness of the model family involves to define following new weights:

$$\rho_{(K,\mathbf{v})} = \frac{D(K,\alpha)}{2} \ln \left[\frac{8eQ}{\{D(K,\alpha) - 1\} \wedge (2Q - 1)} \right].$$

Consequently, in the following theorem which is the analog of Theorem 2.2 for the non-ordered case, the associated penalty functions have an additional logarithm term depending on the dimension.

Theorem 2.3. *For the four Gaussian mixture collections, there exists two absolute constants κ and C such that, if*

$$\mathrm{pen}(K,\mathbf{v}) \geq \kappa \frac{D(K,\mathbf{v})}{n} \left(2\,\mathcal{A}^2 + \ln \left\{ \frac{1}{1 \wedge \frac{D(K,\mathbf{v})}{n} \mathcal{A}^2} \right\} + \frac{1}{2} \ln \left[\frac{8eQ}{\{D(K,\mathbf{v}) - 1\} \wedge (2Q-1)} \right] \right)$$

where $\mathcal{A}$ is the same constant as in the ordered case, then the model $(\hat{K}, \hat{\mathbf{v}})$ minimizing $\mathrm{crit}(K,\mathbf{v}) = \gamma_n(\hat{s}_{(K,\mathbf{v})}) + \mathrm{pen}(K,\mathbf{v})$ *on $\mathcal{M}$ exists and*

$$\mathbb{E}\left[d_H^2(s, \hat{s}_{(\hat{K},\hat{\mathbf{v}})}) \right] \leq C \left[\inf_{(K,\mathbf{v})\in\mathcal{M}} \{\mathrm{KL}(s, \mathcal{S}_{(K,\mathbf{v})}) + \mathrm{pen}(K,\mathbf{v})\} + \frac{2}{n} \right].$$

The theorem proof given in Appendix A.2 only consists of justifying the form of new weights and finding an upper bound of the weight sum since $\mathcal{S}_{(K,\mathbf{v})}$ has the same bracketing entropy as $\mathcal{S}_{(K,\alpha)}$. This non-ordered case is more attractive for practical use but the results of Theorem 2.3 are difficult to apply when the number of variables becomes too large since an exhaustive research of the best model is then untractable.

3. Discussion

In this paper, specific Gaussian mixtures are considered to take the role of variables in the clustering process into account. Main results are stated for four Gaussian mixture forms for ordered and non-ordered variables. A non asymptotic penalized criterion is proposed to select the number of clusters and the clustering relevant variable subset. Oracle inequalities satisfied by the associated estimator $\hat{s}_{(\hat{K},\hat{\mathbf{v}})}$ are also obtained. The main interest of these results is to give the shape of an adequate penalty in this particular framework. Proofs of these results require to control the bracketing entropy of multidimensional Gaussian density families and to determine weights taking the richness of the model collection into account. Similar results for non-Gaussian mixtures can be obtained as soon as the bracketing entropy of the new component density family can be controlled.

Usually, the Gaussian mixture clustering problem is recast into a selection problem of the number of mixture components and besides of the mixture shape. A complete collection of twenty eight parsimonious models is available, used for instance in Mixmod software [7]. These models are obtained by imposing conditions on the proportions and the elements of variance matrix eigenvalue decomposition (see [3,16]). Commonly, an asymptotic criterion as BIC [29] or ICL [6] is used to solve this model selection problem. In this paper, our main results allow us to propose a non asymptotic criterion to select the number of clusters, the subset $\mathbf{v}$ being fixed to the complete variable set. Moreover, we focus on four mixture forms but similar results can be stated for several of the mixture shapes. It is thus possible to obtain a non asymptotic criterion which besides allows us to select the mixture shape. A comparison of our criterion with BIC, ICL and AIC is proposed in our framework in the companion paper [27] and also for the selection of the mixture component number in [5].

For practical purposes, theoretical results stated in this paper are not immediately usable since they depend on unknown constants and mixture parameters are not bounded. Nevertheless, they are required to justify the shape of penalties and allow that the number of variables Q can be large. Birgé and Massart [10] propose their so-called "slope heuristics" (see also Sect. 8.5 in [24]) to calibrate these constants. This heuristics consists of assuming that twice the minimal penalty is almost the optimal penalty. Theoretically, this rule of thumb is proved by Birgé and Massart [10] in the framework of Gaussian regression in a homoscedastic fixed design and generalized by Arlot and Massart [2] in the heteroscedastic random design case for histograms. The slope heuristics is also the subject of several practical studies. For example, it has been successfully applied for multiple change point detection in [22], for clustering [5], for estimation of oil reserves [23] and genomics [33]. In our context of Gaussian mixture clustering, this heuristics is carried out to calibrate our non asymptotic penalized criterion. Although our modelling is based on strong assumptions on variable relationships, our calibrated penalized criterion allows us to obtain suitable results on simulated and real datasets, presented in our companion paper [27].

A. Proofs of Theorems 2.2 and 2.3

A.1. Proof of Theorem 2.2

We consider the Gaussian mixture models $\mathcal{S}_{(K,\alpha)}$ for which the variables are assumed to be ordered. The aim of this section is to apply the general MLE selection model theorem (Thm. 1.1) in order to prove Theorem 2.2. According to the bracketing entropy upper bound result given in Theorem 2.1, the following proposition gives a convenient function $\Psi_{(K,\alpha)}$, fulfilling properties **[I]** for applying Theorem 1.1.

Proposition A.1. *For the four mixture collections, for all $\xi > 0$*

$$\int_0^{\xi} \sqrt{\mathcal{H}_{[.]}(x, \mathcal{S}_{(K,\alpha)}, d_H)}\, \mathrm{d}x \leq \xi \sqrt{D(K,\alpha)} \left\{ \mathcal{A} + \sqrt{\ln\left(\frac{1}{1 \wedge \xi}\right)} \right\}$$

where the constant $\mathcal{A}$ is given by:

- *for the $[L_k B_k]$ collection:*

$$\mathcal{A} := \sqrt{\pi} + \sqrt{\ln\left(18\pi \mathrm{e}^2\right)} + \sqrt{\ln\left(a\sqrt{\frac{8}{c_1\lambda_m}}\right)} + \sqrt{\ln\left(8\frac{\lambda_M}{\lambda_m}\right)} + \sqrt{\ln(9\sqrt{2}\,Q)}; \tag{A.1}$$

- *for the $[L_k C_k]$ and $[LC]$ collections:*

$$\mathcal{A} = \sqrt{\ln(18\pi \mathrm{e}^2)} + \sqrt{\ln(Q^2)} + \sqrt{\pi} + \sqrt{\ln\left(\frac{24\sqrt{2}\,\lambda_M}{\lambda_m}\right)} + \sqrt{\ln\left(\frac{54\sqrt{3}\lambda_M}{\lambda_m}\right)} + \sqrt{\ln\left(\frac{54\,a}{\sqrt{\lambda_m}}\right)}; \tag{A.2}$$

- *for the* $[LB_k]$ *collection:*

$$\mathcal{A}=\sqrt{\ln(18\pi e^2)}+\sqrt{\pi}+\sqrt{\ln(Q)}+\sqrt{\ln\left(\frac{216\sqrt{2}\lambda_M}{\lambda_m}\right)}+\sqrt{\ln\left(\frac{6a}{\sqrt{\beta_M(1-2^{-\frac{1}{4}})}}\right)}+\sqrt{\ln\left(\frac{24\beta_M}{\beta_m}\right)}. \qquad (A.3)$$

Proof. According to Theorem 2.1, $\forall x\in(0,1]$, $\mathcal{H}_{[.]}(x,\mathcal{S}_{(K,\alpha)},d_H)\leq\mathcal{I}+D(K,\alpha)\ln\left(\frac{1}{x}\right)$ hence for all positive real number ξ,

$$\int_0^{\xi}\sqrt{\mathcal{H}_{[.]}(x,\mathcal{S}_{(K,\alpha)},d_H)}\,\mathrm{d}x\leq\xi\sqrt{\mathcal{I}}+\int_0^{\xi\wedge 1}\sqrt{D(K,\alpha)\ln\left(\frac{1}{x}\right)}\,\mathrm{d}x. \qquad (A.4)$$

In order to control the last term of the right-hand side of Inequality (A.4), the following technical result is considered:

Lemma A.2. *For all* $\varepsilon\in(0,1]$, $\int_0^{\varepsilon}\sqrt{\ln\left(\frac{1}{x}\right)}\,\mathrm{d}x\leq\varepsilon\left\{\sqrt{\ln\left(\frac{1}{\varepsilon}\right)}+\sqrt{\pi}\right\}$.

Proof. This inequality is deduced from an integration by part and the following concentration inequality (see [24], p. 19): if Z is a centered standard Gaussian variable then $P(Z\geq c)\leq\mathrm{e}^{-\frac{c^2}{2}}$ for all $c>0$. □

Next, using Lemma A.2, we get

$$\begin{aligned}\int_0^{\xi}\sqrt{\mathcal{H}_{[.]}(x,\mathcal{S}_{(K,\alpha)},d_H)}\,\mathrm{d}x &\leq \xi\sqrt{\mathcal{I}}+\xi\sqrt{D(K,\alpha)}\left\{\sqrt{\ln\left(\frac{1}{1\wedge\xi}\right)}+\sqrt{\pi}\right\}\\ &\leq \xi\sqrt{D(K,\alpha)}\left\{\sqrt{\frac{\mathcal{I}}{D(K,\alpha)}}+\sqrt{\pi}+\sqrt{\ln\left(\frac{1}{1\wedge\xi}\right)}\right\}.\end{aligned}$$

For the $[L_kB_k]$ collection, according to the explicit expression of $\mathcal{I}$ given in Proposition B.4,

$$\sqrt{\frac{\mathcal{I}}{D(K,\alpha)}}\leq\sqrt{\frac{C(K)}{D(K,\alpha)}}+\sqrt{\frac{K\alpha}{D(K,\alpha)}\ln\left(a\sqrt{\frac{8}{c_1\lambda_m}}\right)}+\sqrt{\frac{(K\alpha+1)}{D(K,\alpha)}\ln\left(8\frac{\lambda_M}{\lambda_m}\right)}+\sqrt{\frac{(2K\alpha+1)}{D(K,\alpha)}\ln(9\sqrt{2}\,Q)}.$$

Moreover, since $\frac{C(K)}{D(K,\alpha)}\leq\ln(18\pi e^2)$ and $\frac{K\alpha}{D(K,\alpha)}$, $\frac{K\alpha+1}{D(K,\alpha)}$ and $\frac{2K\alpha+1}{D(K,\alpha)}$ are all smaller than 1, $\sqrt{\frac{\mathcal{I}}{D(K,\alpha)}}+\sqrt{\pi}$ is upper bounded by the following constant

$$\mathcal{A}:=\sqrt{\pi}+\sqrt{\ln\left(18\pi e^2\right)}+\sqrt{\ln\left(a\sqrt{\frac{8}{c_1\lambda_{\mathrm{m}}}}\right)}+\sqrt{\ln\left(8\frac{\lambda_M}{\lambda_m}\right)}+\sqrt{\ln(9\sqrt{2}\,Q)}.$$

In the same way, we obtain the explicit expression of the constant $\mathcal{A}$ for the three other Gaussian collections according to Propositions B.5, B.6 and B.11. □

Consequently, for the four collections, the following function

$$\Psi_{(K,\alpha)}:\xi\in\mathbb{R}_+^{\star}\mapsto\xi\sqrt{D(K,\alpha)}\left\{\mathcal{A}+\sqrt{\ln\left(\frac{1}{1\wedge\xi}\right)}\right\}$$

which satisfies properties **[I]** of Theorem 1.1 can be considered. To continue the proof, we need to find $\xi_\star$ such that $\Psi_{(K,\alpha)}(\xi_\star) = \sqrt{n}\,\xi_\star^2$ to deduce the penalty function. This is equivalent to solving

$$\sqrt{\frac{D(K,\alpha)}{n}}\left\{\mathcal{A} + \sqrt{\ln\left(\frac{1}{1\wedge\xi_\star}\right)}\right\} = \xi_\star.$$

Noticing that the quantity $\tilde{\xi} = \sqrt{\frac{D(K,\alpha)}{n}}\,\mathcal{A}$ satisfies $\tilde{\xi} \leq \xi_\star$, we get $\xi_\star \leq \sqrt{\frac{D(K,\alpha)}{n}}\left\{\mathcal{A} + \sqrt{\ln\left(\frac{1}{1\wedge\tilde{\xi}}\right)}\right\}$ and thus

$$\xi_\star^2 \leq \frac{D(K,\alpha)}{n}\left\{2\,\mathcal{A}^2 + \ln\left(\frac{1}{1\wedge\frac{D(K,\alpha)}{n}\,\mathcal{A}^2}\right)\right\}.$$

Finally, according to the lower bound of penalty functions in Theorem 1.1, it remains to define the weights $\rho_{(K,\alpha)}$. The considered weights $\rho_{(K,\alpha)} = D(K,\alpha)$ depend on the model dimension and their sum Υ is equal to 1 since

$$\text{card}\,\{(K,\alpha)\in\mathbb{N}^\star\times\{1,\ldots,Q\};\ D(K,\alpha)=D\}\leq D$$

and $\sum\limits_{(K,\alpha)} \mathrm{e}^{-\rho_{(K,\alpha)}} \leq \sum\limits_{D\geq 1} D\,e^{-D} \leq 1$. Therefore according to Theorem 1.1, if the penalty function satisfies the inequality

$$\text{pen}(K,\alpha) \geq \kappa\frac{D(K,\alpha)}{n}\left\{1 + 2\,\mathcal{A}^2 + \ln\left(\frac{1}{1\wedge\frac{D(K,\alpha)}{n}\,\mathcal{A}^2}\right)\right\},$$

a minimizer $(\hat{K},\hat{\alpha})$ of $\text{crit}(K,\alpha) = \gamma_n(\hat{s}_{(K,\alpha)}) + \text{pen}(K,\alpha)$ on $\mathcal{M}$ exists and

$$\mathbb{E}\left[d_H^2(s,\hat{s}_{(\hat{K},\hat{\alpha})})\right] \leq C\left[\inf_{(K,\alpha)\in\mathcal{M}}\{\text{KL}(s,\mathcal{S}_{(K,\alpha)}) + \text{pen}(K,\alpha)\} + \frac{1}{n}\right].$$

A.2. Proof of Theorem 2.3

Apart from the weight definition step, the proof of Theorem 2.3 is the same as in the ordered case. The following lemma is used to define weights for this richer family. Recall that $D(K,\alpha)$ denotes the dimension of $\mathcal{S}_{(K,\mathbf{v})}$ and is given for each collection in Section 1.1.

Lemma A.3. *For the four collections, $\mathcal{C}(D) := \text{card}\,\{(K,\mathbf{v})\in\mathbb{N}^\star\times\mathcal{V};\ D(K,\alpha)=D\}$ is upper bounded by*

$$\begin{cases} 2^Q & \textit{if } Q \leq \frac{D-1}{2} \\ \left(\frac{2\mathrm{e}Q}{D-1}\right)^{\frac{D-1}{2}} & \textit{otherwise.}\end{cases}$$

Proof. For the $[L_kB_k]$ collection,

$$\begin{aligned}\mathcal{C}(D) &= \text{card}\,[(K,\mathbf{v})\in\mathbb{N}^\star\times\mathcal{V};\ K\{2\,\text{card}(\mathbf{v})+1\}=D]\\ &= \sum_{K=1}^{\infty}\sum_{\alpha=1}^{Q}\binom{Q}{\alpha}\mathbb{I}_{K(2\alpha+1)=D}\\ &\leq \sum_{\alpha=1}^{\infty}\binom{Q}{\alpha}\ \mathbb{I}_{\alpha\leq Q\wedge\lfloor\frac{D-1}{2}\rfloor}.\end{aligned}$$

If $Q \leq \lfloor \frac{D-1}{2} \rfloor$, $\sum\limits_{\alpha=1}^{\infty} \binom{Q}{\alpha} \mathbb{1}_{\alpha \leq Q \wedge \lfloor \frac{D-1}{2} \rfloor} = 2^Q$. Otherwise, according to Proposition 2.5 in [24],

$$\sum_{\alpha=1}^{\infty} \binom{Q}{\alpha} \mathbb{1}_{\alpha \leq Q \wedge \lfloor \frac{D-1}{2} \rfloor} \leq f\left(\left\lfloor \frac{D-1}{2} \right\rfloor \right)$$

where $f(x) = \left(\frac{\mathrm{e}Q}{x}\right)^x$ is an increasing function on $[1, Q]$. Noticing that Q is an integer, it leads to

$$\sum_{\alpha=1}^{Q \wedge \lfloor \frac{D-1}{2} \rfloor} \binom{Q}{\alpha} \leq \begin{cases} 2^Q & \text{if } Q \leq \frac{D-1}{2} \\ \left(\frac{2\mathrm{e}Q}{D-1}\right)^{\frac{D-1}{2}} & \text{otherwise.} \end{cases}$$

For the $[L_k C_k]$ collection, since $D(K, \mathbf{v}) = K \left[1 + \operatorname{card}(\mathbf{v}) + \frac{\operatorname{card}(\mathbf{v})\{\operatorname{card}(\mathbf{v})+1\}}{2}\right]$, $\mathcal{C}(D)$ is upper bounded by

$$\sum_{\alpha=1}^{Q} \binom{Q}{\alpha} \mathbb{1}_{1+\frac{3}{2}\alpha+\frac{\alpha^2}{2} \leq D} \leq \sum_{\alpha=1}^{Q} \binom{Q}{\alpha} \mathbb{1}_{\alpha \leq \frac{D-1}{2}}$$

hence the result is the same as for the $[L_k B_k]$ collection. An analogous proof gives the result for the two other collections $[L B_k]$ and $[LC]$. □

Proposition A.4. *Consider the following weight family* $\left\{\rho_{(K,\mathbf{v})}\right\}_{(K,\mathbf{v}) \in \mathbb{N}^\star \times \mathcal{V}}$ *defined by*

$$\rho_{(K,\mathbf{v})} = \frac{D(K,\alpha)}{2} \ln \left[\frac{8\mathrm{e}Q}{\{D(K,\alpha)-1\} \wedge (2Q-1)} \right].$$

Then we have $\sum\limits_{(K,\mathbf{v}) \in \mathbb{N}^\star \times \mathcal{V}} \mathrm{e}^{-\rho_{(K,\mathbf{v})}} \leq 2.$

Proof. According to Lemma A.3,

$$\begin{aligned}
\sum_{(K,\mathbf{v}) \in \mathbb{N}^\star \times \mathcal{V}} \mathrm{e}^{-\rho_{(K,\mathbf{v})}} &= \sum_{D=3}^{\infty} \exp\left[-\frac{D}{2} \ln \left\{ \frac{8\mathrm{e}Q}{(D-1) \wedge (2Q-1)} \right\} \right] \operatorname{card}\{(K,\mathbf{v});\ D(K,\mathbf{v}) = D\} \\
&\leq \sum_{D=3}^{\infty} \exp\left[-\frac{D}{2} \ln \left\{ \frac{8\mathrm{e}Q}{(D-1) \wedge (2Q-1)} \right\} \right] \left\{ 2^Q \mathbb{1}_{Q \leq \frac{D-1}{2}} + \left(\frac{2\mathrm{e}Q}{D-1} \right)^{\frac{D-1}{2}} \mathbb{1}_{\frac{D-1}{2} < Q} \right\} \\
&\leq \sum_{D=3}^{2Q} \exp\left\{ -\frac{D}{2} \ln \left(\frac{8\mathrm{e}Q}{D-1} \right) + \frac{D-1}{2} \ln \left(\frac{2\mathrm{e}Q}{D-1} \right) \right\} \\
&\quad + \sum_{D=2Q+1}^{\infty} \exp\left\{ -\frac{D}{2} \ln \left(\frac{8\mathrm{e}Q}{2Q-1} \right) + Q \ln(2) \right\}.
\end{aligned}$$

For the term in the exponential function of the first sum,

$$\begin{aligned}
-\frac{D}{2} \ln \left(\frac{8\mathrm{e}Q}{D-1} \right) + \frac{D-1}{2} \ln \left(\frac{2\mathrm{e}Q}{D-1} \right) &= -\frac{D}{2} \ln(4) - \frac{1}{2} \ln \left(\frac{2\mathrm{e}Q}{D-1} \right) \\
&\leq -(D-1)\ln(2)
\end{aligned}$$

since $D \leq 2Q$. For the term in the exponential function of second sum, since $D \geq 2Q+1$,

$$\begin{aligned} -\frac{D}{2}\ln\left(\frac{8\mathrm{e}Q}{2Q-1}\right) + Q\ln(2) &= -\frac{3D}{2}\ln(2) + Q\ln(2) - \frac{D}{2}\ln\left(\frac{\mathrm{e}Q}{2Q-1}\right) \\ &\leq \left(Q - \frac{D-1}{2}\right)\ln(2) - (D-1)\ln(2) \\ &\leq -(D-1)\ln(2). \end{aligned}$$

Then

$$\begin{aligned} \sum_{(K,\mathbf{v})\in\mathbb{N}^\star\times\mathcal{V}} \mathrm{e}^{-\rho(K,\mathbf{v})} &\leq \sum_{D=3}^{\infty}\left(\frac{1}{2}\right)^{D-1} \\ &\leq 2. \end{aligned}$$

□

B. Tools: bound on bracketing entropies of mixture density families

This appendix is devoted to the proof of Theorem 2.1. Since the model $\mathcal{S}_{(K,\mathbf{v})}$ has the same complexity as $\mathcal{S}_{(K,\alpha)}$, the proof is only given for the ordered variable case. Section B.1 gives the general principle for the control of the bracketing entropy of mixture model families. The four following sections detail the proof of Theorem 2.1 for each of the four Gaussian mixture collections.

B.1. General principle for the control of the bracketing entropy of mixture model families

Ghosal and van der Vaart [19] and Genovese and Wasserman [18] have proposed an upper bound of the bracketing entropy of unidimensional Gaussian mixtures in order to obtain convergence rates in Hellinger distance for density estimation using the Gaussian mixtures. We first tried to follow the strategy proposed in [19] to control the bracketing entropy of our multidimensional Gaussian mixture models. But the control obtained this way has a too large dependency in Q: this implies that the constant $\mathcal{A}$ in the penalty function depends on a power of Q, allowing not that Q is of the order of n or even larger than n in particular. We propose instead a method inspired by the work of Genovese and Wasserman [18]. The key idea is given by their theorem stated hereafter: the control of the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$ can be recast into the control of the bracketing entropies of the associated mixture component density families. For all k in $\{1,\ldots,K\}$, let $\mathcal{C}_k = \{f_{\theta_k}, \theta_k \in \Theta_k\}$ be a family of densities with respect to Lebesgue measure on $\mathbb{R}^Q$. The following family of mixture distributions based on $\mathcal{C}_k$ is considered

$$\mathcal{W}_K := \left\{\sum_{k=1}^{K} p_k f_{\theta_k},\ \theta_k \in \Theta_k\ \forall k=1,\ldots,K,\ \mathbf{p}=(p_1,\ldots,p_K)\in\mathcal{P}_{K-1}\right\}$$

where $\mathcal{P}_{K-1}$ is the $K-1$ dimensional simplex defined by

$$\mathcal{P}_{K-1} := \left\{\mathbf{p}=(p_1,\ldots,p_K),\ \forall k=1,\ldots,K,\, p_k\geq 0,\, \sum_{k=1}^{K} p_k = 1\right\}.$$

Theorem B.1. *With the previous notation, for all K and all $\varepsilon \in (0,1]$,*

$$\mathcal{N}_{[.]}(\varepsilon,\mathcal{W}_K,d_H) \leq \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3},\mathcal{P}_{K-1},d_H\right)\prod_{k=1}^{K}\mathcal{N}_{[.]}\left(\frac{\varepsilon}{3},\mathcal{C}_k,d_H\right)$$

where

$$\mathcal{N}_{[.]}\left(\frac{\varepsilon}{3},\mathcal{P}_{K-1},d_H\right) \leq K(2\pi\,\mathrm{e})^{\frac{K}{2}}\left(\frac{3}{\varepsilon}\right)^{K-1}.$$

In our context, we want to take the specific form of the studied multidimensional mixtures into account. Recall that two situations occur, depending on whether $\mathcal{F}_{(K,\alpha)}$ can be written as the cartesian product of K times a set $\mathcal{F}_{(\alpha)}$ or not. For these two situations, a new result is deduced from Theorem B.1.

For the $[L_kB_k]$ and $[L_kC_k]$ collections, equation (1.3) gives that an element $f \in \mathcal{S}_{(K,\mathbf{v})}$ can be written, for all $x \in \mathbb{R}^Q$,

$$f(x) = \Phi(x_{[\mathbf{v}^c]}|0, \omega^2 I_{Q-\alpha}) \sum_{k=1}^{K} p_k \Phi(x_{[\mathbf{v}]}|\mu_k, \Sigma_k)$$

where $\Phi(.|0, \omega^2 I_{Q-\alpha})$ belongs to $\mathcal{G}_{(\alpha)}$ and where Gaussian densities $\Phi(.|\mu_k, \Sigma_k)$ belong to $\mathcal{F}_{(\alpha)}$ (see Sect. 1.1). According to Theorem B.1, the bracketing entropy of the mixture family $\mathcal{L}_{(K,\alpha)}$ is related to the one of $\mathcal{F}_{(\alpha)}$ ($\mathcal{C}_k = \mathcal{F}_{(\alpha)},\ \forall k \in \{1, \ldots, K\}$). The following proposition is deduced from Theorem B.1. It allows us to bound the bracketing entropy of the mixture family $\mathcal{S}_{(K,\mathbf{v})}$ by a product of the bracketing entropies of the simplex $\mathcal{P}_{K-1}$, of $\mathcal{G}_{(\alpha)}$ and of $\mathcal{F}_{(\alpha)}$.

Proposition B.2. *For the $[L_kB_k]$ and $[L_kC_k]$ mixture collections, for all $\varepsilon \in (0, 1]$, the bracketing number of the density family $\mathcal{S}_{(K,\mathbf{v})}$ is bounded by*

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq K(2\pi\,\mathrm{e})^{\frac{K}{2}} \left(\frac{9}{\varepsilon}\right)^{K-1} \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) \mathcal{N}_{[.]}\left(\frac{\varepsilon}{9}, \mathcal{F}_{(\alpha)}, d_H\right)^K.$$

It is then deduced that the bracketing entropy of $\mathcal{S}_{(K,\mathbf{v})}$ is bounded by

$$\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq C(K) + (K-1)\ln\left(\frac{1}{\varepsilon}\right) + \mathcal{H}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) + K\,\mathcal{H}_{[.]}\left(\frac{\varepsilon}{9}, \mathcal{F}_{(\alpha)}, d_H\right) \tag{B.1}$$

with $C(K) = \ln(K) + \frac{K}{2}\ln(2\pi\mathrm{e}) + (K-1)\ln(9)$.

Proof. According to Theorem B.1, for all $\delta \leq 1$,

$$\mathcal{N}_{[.]}(\delta, \mathcal{L}_{(K,\alpha)}, d_H) \leq K(2\pi\mathrm{e})^{\frac{K}{2}} \left(\frac{3}{\delta}\right)^{K-1} \prod_{k=1}^{K} \mathcal{N}_{[.]}\left(\frac{\delta}{3}, \mathcal{F}_{(\alpha)}, d_H\right).$$

If we prove that for all $\varepsilon \leq 1$,

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{L}_{(K,\alpha)}, d_H\right) \tag{B.2}$$

then we obtain

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq K(2\pi\,\mathrm{e})^{\frac{K}{2}} \left(\frac{9}{\varepsilon}\right)^{K-1} \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) \mathcal{N}_{[.]}\left(\frac{\varepsilon}{9}, \mathcal{F}_{(\alpha)}, d_H\right)^K.$$

Thus, it remains to check Inequality (B.2). It is done by the following adaptation of a result proof given in [18].

Let $\delta \in [0, 1]$ and $h \in \mathcal{S}_{(K,\mathbf{v})}$, decomposed into $h(x) = f(x_{[\mathbf{v}]})g(x_{[\mathbf{v}^c]})$ where $f \in \mathcal{L}_{(K,\alpha)}$ and $g \in \mathcal{G}_{(\alpha)}$. Let $[l, u]$ and $[\tilde{l}, \tilde{u}]$ be two δ-brackets of $\mathcal{L}_{(K,\alpha)}$ and $\mathcal{G}_{(\alpha)}$ containing f and g respectively. Then, the two functions defined by

$$L(x) = l(x_{[\mathbf{v}]})\tilde{l}(x_{[\mathbf{v}^c]}) \text{ and } U(x) = u(x_{[\mathbf{v}]})\tilde{u}(x_{[\mathbf{v}^c]}) \tag{B.3}$$

constitute a bracket of $\mathcal{S}_{(K,\mathbf{v})}$ containing h. The size of this bracket is now calculated. First of all, Lemma 3 from [18] gives that

$$\begin{cases} \int u(x_{[\mathbf{v}]})\mathrm{d}x_{[\mathbf{v}]} \leq 1 + 3\delta \\ \int \tilde{u}(x_{[\mathbf{v}^c]})\mathrm{d}x_{[\mathbf{v}^c]} \leq 1 + 3\delta. \end{cases} \tag{B.4}$$

Then the squared Hellinger distance between L and U is equal to

$$\begin{aligned}
d_H^2(L,U) &= \int \left\{ \sqrt{u(x_{[\mathbf{v}]})\tilde{u}(x_{[\mathbf{v}^c]})} - \sqrt{l(x_{[\mathbf{v}]})\tilde{l}(x_{[\mathbf{v}^c]})} \right\}^2 \mathrm{d}x \\
&= \int \left[\sqrt{\tilde{u}(x_{[\mathbf{v}^c]})} \left\{ \sqrt{u(x_{[\mathbf{v}]})} - \sqrt{l(x_{[\mathbf{v}]})} \right\} + \left\{ \sqrt{\tilde{u}(x_{[\mathbf{v}^c]})} - \sqrt{\tilde{l}(x_{[\mathbf{v}^c]})} \right\} \sqrt{l(x_{[\mathbf{v}]})} \right]^2 \mathrm{d}x \\
&= \int \tilde{u}(x_{[\mathbf{v}^c]})\, \mathrm{d}x_{[\mathbf{v}^c]} \int \left\{ \sqrt{u(x_{[\mathbf{v}]})} - \sqrt{l(x_{[\mathbf{v}]})} \right\}^2 \mathrm{d}x_{[\mathbf{v}]} \\
&\quad + \int \left\{ \sqrt{\tilde{u}(x_{[\mathbf{v}^c]})} - \sqrt{\tilde{l}(x_{[\mathbf{v}^c]})} \right\}^2 \mathrm{d}x_{[\mathbf{v}^c]} \int l(x_{[\mathbf{v}]})\, \mathrm{d}x_{[\mathbf{v}]} \\
&\quad + 2 \int \sqrt{\tilde{u}(x_{[\mathbf{v}^c]})} \left\{ \sqrt{\tilde{u}(x_{[\mathbf{v}^c]})} - \sqrt{\tilde{l}(x_{[\mathbf{v}^c]})} \right\} \mathrm{d}x_{[\mathbf{v}^c]} \times \int \left\{ \sqrt{u(x_{[\mathbf{v}]})} - \sqrt{l(x_{[\mathbf{v}]})} \right\} \sqrt{l(x_{[\mathbf{v}]})}\, \mathrm{d}x_{[\mathbf{v}]}.
\end{aligned}$$

According to Cauchy-Schwarz inequality and (B.4),

$$\begin{aligned}
\int \left\{ \sqrt{u(x_{[\mathbf{v}]})} - \sqrt{l(x_{[\mathbf{v}]})} \right\} \sqrt{l(x_{[\mathbf{v}]})}\, \mathrm{d}x_{[\mathbf{v}]} &\leq 1 \times d_H(l,u) \\
&\leq \delta
\end{aligned}$$

and

$$\begin{aligned}
\int \sqrt{\tilde{u}(x_{[\mathbf{v}^c]})} \left\{ \sqrt{\tilde{u}(x_{[\mathbf{v}^c]})} - \sqrt{\tilde{l}(x_{[\mathbf{v}^c]})} \right\} \mathrm{d}x_{[\mathbf{v}^c]} &\leq \sqrt{1+3\delta} \times d_H(\tilde{l},\tilde{u}) \\
&\leq 2\,\delta.
\end{aligned}$$

Thus,

$$\begin{aligned}
d_H^2(L,U) &\leq d_H^2(l,u) \int \tilde{u}(x_{[\mathbf{v}^c]})\, \mathrm{d}x_{[\mathbf{v}^c]} + d_H^2(\tilde{l},\tilde{u}) + 4\delta^2 \\
&\leq (1+3\delta)\,\delta^2 + \delta^2 + 4\delta^2 \\
&\leq 9\delta^2.
\end{aligned}$$

Finally, with $\delta = \varepsilon/3$ and according to the bracket definition (B.3), the number of brackets for $\mathcal{S}_{(K,\mathbf{v})}$ is upper bounded by $\mathcal{N}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) \times \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{L}_{(K,\alpha)}, d_H\right)$. □

As explained in Section 1.1, the variance matrices have a common element in the $[LB_k]$ and $[LC]$ collections: They have the same volume for the $[LB_k]$ collection and are equal for the $[LC]$ collection. The family $\mathcal{F}_{(K,\alpha)}$ cannot be assimilated to one Gaussian density family and thus Theorem B.1 cannot be applied in this case. Nevertheless the following proposition is a variant, allowing us to take the specific form of studied mixtures into account. Its proof is established along the line of the proof of Theorem 2 in [18]. This proposition recasts the problem to upper bound the bracketing entropy of $\mathcal{S}_{(K,\mathbf{v})}$ into the study of the bracketing entropy of the simplex, $\mathcal{G}_{(\alpha)}$ and $\mathcal{F}_{(K,\alpha)}$.

Proposition B.3. *For the $[LB_k]$ and $[LC]$ mixture collections, for all $\varepsilon \in (0,1]$, the bracketing number of $\mathcal{S}_{(K,\mathbf{v})}$ is upper bounded by*

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq K(2\pi\,\mathrm{e})^{\frac{K}{2}} \left(\frac{9}{\varepsilon}\right)^{K-1} \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) \mathcal{N}_{[.]}\left(\frac{\varepsilon}{9}, \mathcal{F}_{(K,\alpha)}, d_H\right).$$

Hence with $C(K) = \ln(K) + \frac{K}{2}\ln(2\pi \mathrm{e}) + (K-1)\ln(9)$,

$$\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq C(K) + (K-1)\ln\left(\frac{1}{\varepsilon}\right) + \mathcal{H}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) + \mathcal{H}_{[.]}\left(\frac{\varepsilon}{9}, \mathcal{F}_{(K,\alpha)}, d_H\right). \tag{B.5}$$

Proof. According to (B.2) in the proof of Proposition B.2,

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right)\mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{L}_{(K,\alpha)}, d_H\right).$$

Then we can prove along the line of the proof of Theorem 2 in [18] that

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{L}_{(K,\alpha)}, d_H) \leq \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{P}_{K-1}, d_H\right)\mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{F}_{(K,\alpha)}, d_H\right)$$

where $\mathcal{P}_{K-1}$ is the $K-1$ dimensional simplex. Sketch of the proof: consider an $\varepsilon/3$-bracketing $\{[a_1 b_1], \ldots, [a_N, b_N]\}$ with $a_j, b_j \in [0,1]^K$, for the simplex $\mathcal{P}_{K-1}$ and an $\varepsilon/3$-bracketing for $\mathcal{F}_{(K,\alpha)}$. This last family is a set of K-tuples $([l_1, u_1], \ldots, [l_K, u_K])$ such that $d_H(l_k, u_k) \leq \varepsilon/3$ for all k. Then the family of brackets $[L, U]$ defined by $L(x) = \sum_{k=1}^K a_{jk} l_k(x)$ and $U(x) = \sum_{k=1}^K b_{jk} u_k(x)$ is an ε-bracketing of $\mathcal{L}_{(K,\alpha)}$. □

Finally, the control of the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$ is recast into the one of $\mathcal{G}_{(\alpha)}$ and $\mathcal{F}_{(\alpha)}$ or $\mathcal{F}_{(K,\alpha)}$ according to the considered mixture collection. This control is stated in Propositions B.4, B.5, B.6 and B.11 for the $[L_k B_k]$, $[LB_k]$, $[L_k C_k]$ and $[LC]$ collections respectively. We are now in position to prove Theorem 2.1 for the four collections.

B.2. Control of the bracketing entropy for the $[L_k B_k]$ collection

In this section, we consider the case of the $[L_k B_k]$ Gaussian mixture family (see the description in Sect. 1.1). The following proposition gives an upper bound of the bracketing entropy of the two families $\mathcal{F}_{(\alpha)}$ and $\mathcal{G}_{(\alpha)}$ defined by (1.4) and (1.2) respectively. It allows us to deduce an upper bound of the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$ according to Inequality (B.1).

Proposition B.4. *Set* $c_1 = 5\left(1 - 2^{-\frac{1}{4}}\right)/8$. *For all* $\varepsilon \in (0,1]$,

$$\mathcal{H}_{[.]}(\varepsilon, \mathcal{F}_{(\alpha)}, d_H) \leq \alpha \ln\left(2\,a\sqrt{\frac{2}{c_1\,\lambda_m}}\right) + \alpha\ln\left(8\frac{\lambda_M}{\lambda_m}\right) + 2\alpha\ln(\sqrt{2}\,Q) + 2\alpha\ln\left(\frac{1}{\varepsilon}\right) \tag{B.6}$$

and

$$\mathcal{H}_{[.]}(\varepsilon, \mathcal{G}_{(\alpha)}, d_H) \leq \ln\left(8\frac{\lambda_M}{\lambda_m}\right) + \ln\left(\sqrt{2}\,Q\right) + \ln\left(\frac{1}{\varepsilon}\right)\cdot \tag{B.7}$$

Thus,

$$\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\alpha)}, d_H) \leq \mathcal{I} + D(K,\alpha)\ln\left(\frac{1}{\varepsilon}\right) \tag{B.8}$$

where

$$\mathcal{I} = C(K) + (2K\alpha + 1)\ln(9\sqrt{2}\,Q) + (K\alpha + 1)\ln\left(8\frac{\lambda_M}{\lambda_m}\right) + K\alpha\ln\left(a\sqrt{\frac{8}{c_1\lambda_m}}\right)$$

with $C(K) = \ln(K) + \frac{K}{2}\ln(2\pi e) + (K-1)\ln(9)$.

Proof. According to Assertion (B.1), the upper bound of the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$, given by Inequality (B.8), is deduced from upper bounds of the bracketing entropy of $\mathcal{F}_{(\alpha)}$ and $\mathcal{G}_{(\alpha)}$, respectively expressed in Inequalities (B.6) and (B.7). These two inequalities are now proved.

The proof of Inequality (B.6) is adapted from the work of Genovese and Wasserman [18] who prove similar results for unidimensional Gaussian mixture families. The main idea is to define a lattice over the parameter space $\mathcal{B} = \{(\mu, \sigma_1^2, \dots, \sigma_\alpha^2) \in [-a, a]^\alpha \times [\lambda_m, \lambda_M]^\alpha\}$ and next to deduce a bracket covering of $\mathcal{F}_{(\alpha)}$ according to the Hellinger distance.

First, consider $\varepsilon \in (0, 1]$ and $\delta = \varepsilon/(\sqrt{2}Q)$. For all $j \in \{2, \dots, r\}$, set

$$b_j^2 = (1+\delta)^{1-\frac{j}{2}}\, \lambda_M$$

with $r = \left\lceil 2\, \frac{\ln\left\{\frac{\lambda_M (1+\delta)}{\lambda_{\mathrm{m}}}\right\}}{\ln(1+\delta)} \right\rceil$ in order to have $b_r^2 \leq \lambda_m < \lambda_M = b_2^2$ ($\lceil h \rceil$ denotes the smallest integer greater than or equal to h). Then, for all $J = (j(1), \dots, j(\alpha)) \in \{2, \dots, r\}^\alpha$, a diagonal matrix B_J is defined by

$$B_J = \mathrm{diag}(b_{j(1)}^2, \dots, b_{j(\alpha)}^2).$$

We also consider vectors

$$\nu_J = (\nu_1^{(J)}, \dots, \nu_\alpha^{(J)}) \in [-a, a]^\alpha$$

such that

$$\forall q \in \{1, \dots, \alpha\},\ \nu_q^{(J)} = \sqrt{c_1\, \lambda_M}\, \delta\, (1+\delta)^{\frac{1-j(q)}{4}}\, s_q,$$

where $s_q \in \mathbb{Z} \cap [-A, A]$ with $A = \left\lfloor \frac{a\, \delta^{-1}\, (1+\delta)^{-\frac{1-j(q)}{4}}}{\sqrt{c_1\, \lambda_M}} \right\rfloor$. Thus, the set $\mathcal{R}(\varepsilon, \alpha)$ of all such couples (ν_J, B_J) forms a lattice on $\mathcal{B}$.

This set $\mathcal{R}(\varepsilon, \alpha)$ allows to construct brackets that cover $\mathcal{F}_{(\alpha)}$. For a function $f(.) = \Phi(.|\mu, \Sigma)$ of $\mathcal{F}_{(\alpha)}$, the two following functions are considered:

$$\begin{cases} l(x) = (1+\delta)^{-\alpha} \Phi(x|\nu_J, (1+\delta)^{-\frac{1}{4}}\, B_{J+1}) \\ u(x) = (1+\delta)^{\alpha} \Phi(x|\nu_J, (1+\delta)\, B_J). \end{cases}$$

The index set $J = (j(1), \dots, j(\alpha))$ is taken to satisfy $b_{j(q)+1}^2 \leq \sigma_q^2 \leq b_{j(q)}^2$ for all q in $\{1, \dots, \alpha\}$ and ν_J can be chosen such that

$$(\mu - \nu_J)' B_{J+1}^{-1} (\mu - \nu_J) \leq c_1\, \alpha\, \delta^2 \tag{B.9}$$

where $J+1 := (j(1)+1, \dots, j(\alpha)+1)$. Then we check that the bracket $[l, u]$ contains f. Inequality (B.9) implies that

$$(\mu - \nu_J)' B_J^{-1} (\mu - \nu_J) \leq \frac{\alpha}{4} \delta^2. \tag{B.10}$$

The use of Corollary C.2, which allows to bound the ratio of two Gaussian densities with diagonal variance matrices, together with (B.10) leads to

$$\begin{aligned} \frac{f(x)}{u(x)} &= \frac{\Phi(x|\mu, B)}{(1+\delta)^\alpha\, \Phi(x|\nu_J, (1+\delta)\, B_J)} \\ &\leq (1+\delta)^{-\frac{\alpha}{4}} \exp\left[\frac{1}{2\,\delta} (\mu - \nu_J)'\, B_J^{-1}\, (\mu - \nu_J)\right] \\ &\leq 1. \end{aligned}$$

The function $h : \delta \mapsto 1 - (1+\delta)^{-\frac{1}{4}}$ being concave, it yields $1 - (1+\delta)^{-\frac{1}{4}} \geq \delta(1 - 2^{-\frac{1}{4}})$. With Corollary C.2 and (B.9), this shows that $l \leq f$ since

$$\begin{aligned} \frac{l(x)}{f(x)} &= \frac{(1+\delta)^{-\alpha} \Phi(x|\nu_J, (1+\delta)^{-\frac{1}{4}}\, B_{J+1})}{\Phi(x|\mu, B)} \\ &\leq (1+\delta)^{-\frac{5\alpha}{8}} \exp\left[\frac{(\mu - \nu_J)'\, B_{J+1}^{-1} (\mu - \nu_J)}{2[1 - (1+\delta)^{-\frac{1}{4}}]}\right] \\ &\leq 1. \end{aligned}$$

Therefore, $[l,u]$ contains the function f. To prove that $[l,u]$ is an ε-bracket, it remains to check that $d_H(l,u) \leq \varepsilon$. According to Corollary C.4,

$$\begin{aligned}
d_H^2(l,u) &= d_H^2\Big((1+\delta)^{-\alpha}\Phi(.|\nu_J,(1+\delta)^{-\frac{1}{4}}\,B_{J+1}),(1+\delta)^{\alpha}\Phi(.|\nu_J,(1+\delta)\,B_J)\Big)\\
&= (1+\delta)^{-\alpha}+(1+\delta)^{\alpha}-2\left\{\frac{2}{(1+\delta)^{-\frac{7}{8}}+(1+\delta)^{\frac{7}{8}}}\right\}^{\frac{\alpha}{2}}\\
&= \underbrace{2\cosh(\alpha\ln[1+\delta])-2}_{(i)}+\underbrace{2-2\left[\cosh\left\{\frac{7}{8}\ln(1+\delta)\right\}\right]^{-\frac{\alpha}{2}}}_{(ii)}.
\end{aligned}$$

The upper bounds of terms (i) and (ii) separately lead to

$$\begin{aligned}
d_H^2(l,u) &\leq \left\{\sinh(1)+\frac{49}{128}\right\}\alpha^2\delta^2\\
&\leq 2\,\alpha^2\delta^2\\
&\leq \varepsilon^2.
\end{aligned}$$

Consequently, the parameter family $\mathcal{R}(\varepsilon,\alpha)$ induces an ε-bracketing family over $\mathcal{F}_{(\alpha)}$.

An upper bound of the bracketing number of $\mathcal{F}_{(\alpha)}$ is then deduced from an upper bound of the cardinal of $\mathcal{R}(\varepsilon,\alpha)$

$$\begin{aligned}
\mathcal{N}_{[.]}(\varepsilon,\mathcal{F}_{(\alpha)},d_H) &\leq \operatorname{Card}(\mathcal{R}(\varepsilon,\alpha))\\
&\leq \sum_{J\in\{2,\ldots,r\}^{\alpha}}\prod_{q=1}^{\alpha}\left\{\frac{2\,a}{\sqrt{c_1\,\lambda_M}\,\delta\,(1+\delta)^{\frac{1-j(q)}{4}}}\right\}\\
&\leq \left\{\frac{2\,a(1+\delta)^{\frac{r-1}{4}}}{\sqrt{c_1\,\lambda_M}\delta}\right\}^{\alpha}(r-1)^{\alpha}.
\end{aligned}$$

According to the definition of r, $(1+\delta)^{\frac{r-1}{4}} \leq \sqrt{\lambda_M(1+\delta)/\lambda_m}$. Hence,

$$\begin{aligned}
\mathcal{N}_{[.]}(\varepsilon,\mathcal{F}_{(\alpha)},d_H) &\leq \left(\frac{2\,a}{\delta}\sqrt{\frac{1+\delta}{c_1\,\lambda_m}}\right)^{\alpha}\left[2\,\frac{\ln\left\{\frac{\lambda_M}{\lambda_m}(1+\delta)\right\}}{\ln(1+\delta)}\right]^{\alpha}\\
&\leq \left(\frac{2\sqrt{2}\,a}{\sqrt{c_1\lambda_m}}\right)^{\alpha}\left(\frac{8\,\lambda_M}{\lambda_m}\right)^{\alpha}\ \delta^{-(2\alpha)}\\
&\leq \left(\frac{2\sqrt{2}\,a}{\sqrt{c_1\lambda_m}}\right)^{\alpha}\left(\frac{8\,\lambda_M}{\lambda_m}\right)^{\alpha}\ \left(\frac{\sqrt{2}Q}{\varepsilon}\right)^{2\alpha}
\end{aligned}$$

that implies Inequality (B.6).

Using a similar proof, the upper bound of the bracketing entropy of $\mathcal{G}_{(\alpha)}$ given by Inequality (B.7) is obtained. To check this result, the variance family

$$\{b_j^2 = (1+\delta)^{1-\frac{j}{2}}\,\lambda_M,\ \forall\, 2 \leq j \leq r\}$$

and brackets $[\tilde{l}, \tilde{u}]$ defined on $\mathbb{R}^{Q-\alpha}$ by

$$\left\{ \begin{array}{l} \tilde{l}(x) = (1+\delta)^{-(Q-\alpha)}\, \Phi(x|0, (1+\delta)^{-\frac{1}{4}}\, b_{j+1}^2 I_{Q-\alpha}) \\ \tilde{u}(x) = (1+\delta)^{Q-\alpha}\, \Phi(x|0, (1+\delta)\, b_j^2 I_{Q-\alpha}) \end{array} \right.$$

are considered. □

B.3. Control of the bracketing entropy for the $[LB_k]$ collection

In this section, an upper bound for the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$ is given according to Inequality (B.5). Recall that it is sufficient to control the bracketing entropy of the family $\mathcal{F}_{(K,\alpha)}$, defined by (1.6), and used the one of the family $\mathcal{G}_{(\alpha)}$ given by (B.7).

Proposition B.5. *For all $\varepsilon \in (0,1]$, the bracketing number of the set $\mathcal{F}_{(K,\alpha)}$ is upper bounded by*

$$\mathcal{N}_{[.]}\left(\mathcal{F}_{(K,\alpha)}\right) \leq \left(\frac{24\lambda_M}{\lambda_m}\right)^{K(\alpha-1)} \left(\frac{6a}{\sqrt{\beta_M c_1}}\right)^{K\alpha} \left(\frac{24\beta_M}{\beta_m}\right)^{\frac{K\alpha}{2}+1} \left(\frac{\alpha}{\varepsilon}\right)^{K(2\alpha-1)+1}$$

where $c_1 = 1 - 2^{-\frac{1}{4}}$. Hence

$$\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\alpha)}, d_H) \leq \mathcal{I} + D(K,\alpha) \ln\left(\frac{1}{\varepsilon}\right)$$

where

$$\mathcal{I} = C(K) + \{K(\alpha-1)+1\} \ln\left(\frac{216\sqrt{2}\lambda_M}{\lambda_m}\right) + K\alpha \ln\left(\frac{54a}{\sqrt{\beta_M c_1}}\right) + \left(\frac{K\alpha}{2}+1\right) \ln\left(\frac{24\beta_M}{\beta_m}\right) + \{K(2\alpha-1)+2\} \ln(Q)$$

and $D(K,\alpha) = 2K\alpha + 1$.

Proof. The proof for the unidimensional case ($\alpha = 1$) is already available in [18]. Let $\varepsilon \in (0,1]$ and assume $K \geq 2$ and $\alpha \geq 2$ fixed. Let $\delta = \varepsilon/(3\alpha)$. For $j \in \{2, \ldots, r\}$, we define

$$b_j^2 = (1+\delta)^{1-\frac{j}{2}}\, \lambda_M$$

where $r = \left\lceil 2\, \frac{\ln\left\{\frac{\lambda_M}{\lambda_m}(1+\delta)\right\}}{\ln(1+\delta)} \right\rceil$ in order to have $b_r^2 \leq \lambda_m \leq b_2^2 = \lambda_M$. For $z \in \{0, \ldots, r'\}$, $\beta_z = (1+\delta)^{-z}\beta_M$ is considered where $r' = \left\lceil \frac{\ln\left\{\frac{\beta_M}{\beta_m}\right\}}{\ln(1+\delta)} \right\rceil$ in order to have $\beta_{r'} \leq \beta_m \leq \beta_0 = \beta_M$.

For a vector $J = (j(1), \ldots, j(\alpha-1)) \in \{2, \ldots, r\}^{\alpha-1}$, the diagonal matrices B_J^l and B_J^u are defined by

$$B_J^l = \operatorname{diag}\left(b_{j(1)+1}^2, \ldots, b_{j(\alpha-1)+1}^2, \lambda_M^{1-\alpha}(1+\delta)^{\frac{S_J}{2}-(\alpha-1)}\right)$$

and

$$B_J^u = \operatorname{diag}\left(b_{j(1)}^2, \ldots, b_{j(\alpha-1)}^2, \lambda_M^{1-\alpha}(1+\delta)^{\frac{S_J}{2}-\frac{\alpha-1}{2}}\right)$$

with $S_J = \sum_{q=1}^{\alpha-1} j(q)$. The qth diagonal coefficients of these matrices are denoted $B_{J,q}^l$ and $B_{J,q}^u$ respectively.

First, a function $\Phi(.|\mu, \beta B)$ such that $\beta \in [\beta_m, \beta_M]$, $\mu \in [-a,a]^\alpha$ and $B \in \Delta_{(\alpha)}^1(\lambda_m, \lambda_M)$ is considered. Let $z \in \{0, \ldots, r'\}$ be the unique integer of $\{0, \ldots, r'\}$ such that $\beta_{z+1} < \beta \leq \beta_z$ and let J be the unique vector of $\{2, \ldots, r\}^{\alpha-1}$ such that $\forall q \in \{1, \ldots, \alpha-1\}$, $B_{J,q}^l \leq B_{qq} \leq B_{J,q}^u$. Hence for all $q \in \{1, \ldots, \alpha\}$,

$$\beta_{z+1}\, B_{J,q}^l \leq \beta\, \Sigma_{qq} \leq \beta_z\, B_{J,q}^u.$$

For a couple (z, J), we also consider a regular lattice of mean vector $\nu^{(z,J)} = \left(\nu_1^{(z,J)}, \dots, \nu_\alpha^{(z,J)}\right) \in [-a, a]^\alpha$ such that for all $q \in \{1, \dots, \alpha - 1\}$,

$$\nu_q^{(z,J)} = (1+\delta)^{-\frac{j(q)+1}{4} - \frac{z}{2}} \sqrt{\lambda_M \, \beta_M \, c_1} \, \delta s_q,$$

with $s_q \in \{-N_q, \dots, N_q\}$ where $N_q = \left\lfloor \frac{a(1+\delta)^{\frac{j(q)+1}{4} + \frac{z}{2}}}{\sqrt{\lambda_M \, \beta_M \, c_1} \, \delta} \right\rfloor$,

$$\nu_\alpha^{(z,J)} = (1+\delta)^{\frac{S_J}{4} - \frac{\alpha+z}{2}} \sqrt{\lambda_M^{1-\alpha} \beta_M \, c_1} \, \delta s_\alpha,$$

with $s_\alpha \in \{-N_\alpha, \dots, N_\alpha\}$ where $N_\alpha = \left\lfloor \frac{a(1+\delta)^{\frac{\alpha+z}{2} - \frac{S_J}{4}}}{\sqrt{\lambda_M^{1-\alpha} \beta_M \, c_1} \delta} \right\rfloor$ and $c_1 := 1 - 2^{-\frac{1}{4}}$. For a given couple (z, J), this insures that for all vectors $\mu \in [-a, a]^\alpha$, there exists a vector $\nu^{(z,J)}$ of this lattice such that

$$\frac{(1+\delta)^z}{\beta_M \lambda_M} \left\{ \sum_{q=1}^{\alpha-1} \left(\nu^{(z,J)} - \mu\right)_q^2 (1+\delta)^{\frac{j(q)+1}{2}} + \left(\nu^{(z,J)} - \mu\right)_\alpha^2 (1+\delta)^{-\frac{S_J}{2} + \alpha} \lambda_M^\alpha \right\} \leq c_1 \, \alpha \, \delta^2. \tag{B.11}$$

For a couple (z, J) and a vector $\nu^{(z,J)}$ defined as before, the two following associated functions are considered

$$\begin{cases} l(x) = (1+\delta)^{-2\alpha} \, \Phi\left(x|\nu^{(z,J)}, (1+\delta)^{-\frac{1}{4}} \, \beta_{z+1} \, B_J^l\right) \\ u(x) = (1+\delta)^{2\alpha} \, \Phi\left(x|\nu^{(z,J)}, (1+\delta) \, \beta_z \, B_J^u\right). \end{cases} \tag{B.12}$$

According to Proposition C.1, one can easily check first that $l(x) \leq \Phi(x|\mu, \beta B) \leq u(x)$ using Condition (B.11) and second that $d_H(l, u) \leq \varepsilon$ according to Proposition C.3. Details are available in [25]. Finally, we can construct an ε-bracket family (with respect to d_H) to cover $\mathcal{F}_{(K,\alpha)}$. Let $(\Phi(.|\mu_1, \beta B_1), \dots, \Phi(.|\mu_K, \beta B_K))$ be an element of $\mathcal{F}_{(K,\alpha)}$. Let $z \in \{0, \dots, r'\}$ and $J_1, \dots, J_K$ in $\{2, \dots, r\}^{\alpha-1}$ such that for all $k \in \{1, \dots, K\}$ and all $q \in \{1, \dots, \alpha\}$,

$$\beta_{z+1} \, B_{J_k,q}^l \leq \beta \, B_{k,qq} \leq \beta_z \, B_{J_k,q}^u.$$

For all k, there exists a vector $\nu^{(z,J_k)}$ such that condition (B.11) is satisfied for the mean vector μ_k. For z, J_k and $\nu^{(zJ_k)}$, the two associated functions defined by (B.12) are denoted u_k and l_k. Then we define $L := (l_1, \dots, l_K)$ and $U := (u_1, \dots, u_K)$. The set of all such brackets $[L, U]$ covers the family $\mathcal{F}_{(K,\alpha)}$ and is denoted $\mathcal{R}(\varepsilon, K, \alpha)$. An upper bound of the bracketing number of $\mathcal{F}_{(K,\alpha)}$ is thus determined by the computation of the cardinal of $\mathcal{R}(\varepsilon, K, \alpha)$. Then $\mathcal{N}_{[.]}(\varepsilon, \mathcal{F}_{(K,\alpha)}, d_H)$ is upper bounded by (see [25])

$$\text{card } \mathcal{R}(K, \varepsilon, \alpha) \leq \left(\frac{24\lambda_M}{\lambda_m}\right)^{K(\alpha-1)} \left(\frac{6a}{\sqrt{\beta_M c_1}}\right)^{K\alpha} \left(\frac{24\beta_M}{\beta_m}\right)^{\frac{K\alpha}{2}+1} \left(\frac{\alpha}{\varepsilon}\right)^{K(2\alpha-1)+1}. \qquad \square$$

B.4. **Control of the bracketing entropy for the $[L_k C_k]$ collection**

We now consider the case of the $[L_k C_k]$ Gaussian mixture collection. The following proposition gives an upper bound of the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$, based on the Gaussian density family $\mathcal{F}_{(\alpha)}$ defined by (1.5).

Proposition B.6. *For all $\varepsilon \in (0, 1]$,*

$$\begin{aligned} \mathcal{H}_{[.]}\left(\varepsilon, \mathcal{F}_{(\alpha)}, d_H\right) \quad \leq \quad & \left\{\frac{\alpha(\alpha+1)}{2} + \alpha\right\} \ln(Q^2) + \left\{\frac{\alpha(\alpha+1)}{2} + \alpha\right\} \ln\left(\frac{1}{\varepsilon}\right) \\ & + \frac{\alpha(\alpha+1)}{2} \ln\left(\frac{6\sqrt{3}\lambda_M}{\lambda_m}\right) + \alpha \ln\left(\frac{6\,a}{\sqrt{\lambda_m}}\right). \end{aligned} \tag{B.13}$$

Thus,

$$\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\alpha)}, d_H) \leq \mathcal{I} + D(K,\alpha) \ln\left(\frac{1}{\varepsilon}\right) \tag{B.14}$$

where

$$\mathcal{I} = C(K) + \ln\left(\frac{24\sqrt{2}\,\lambda_M}{\lambda_m}\right) + K\frac{\alpha(\alpha+1)}{2}\ln\left(\frac{54\sqrt{3}\lambda_M}{\lambda_m}\right) + K\alpha\ln\left(\frac{54\,a}{\sqrt{\lambda_m}}\right) + \left\{K\,\frac{\alpha(\alpha+1)}{2} + K\alpha + 1\right\}\ln(Q^2)$$

with $C(K) = \ln(K) + \frac{K}{2}\ln(2\pi e) + (K-1)\ln(9)$.

Result (B.13) together with Inequality (B.1) and the upper bound of $\mathcal{G}_{(\alpha)}$ (see Inequality (B.7)) gives the upper bound (B.14). To prove Inequality (B.13), the method used for the $[L_k B_k]$ collection cannot be extended to this general situation. Considering the eigenvalue decomposition of the variance matrices, a countable covering on the spectrum could be build as previously. An explicit countable covering over the orthogonal matrix set is also necessary to obtain an upper bound of the bracketing entropy of $\mathcal{F}_{(\alpha)}$. Nevertheless, this last point is tricky thus an alternative method is proposed. It consists of defining an adequate covering over the space $\mathcal{D}^+_{(\alpha)}(\lambda_m, \lambda_M)$ with respect to the uniform norm, and then using it to construct a bracket covering of $\mathcal{F}_{(\alpha)}$. The following notation is used for matrix norms: $\|B\|_\infty = \max\limits_{1\leq i,j\leq\alpha} |B_{ij}|$ and $|||B||| = \sup\limits_{\|x\|_2=1} |x'Bx| = \sup\limits_{\lambda\in \mathrm{vp}(B)} |\lambda|$ where $\mathrm{vp}(B)$ denotes the spectrum of B.

The variance matrix lattice. Let $\beta > 0$ and let $\mathcal{R}(\beta)$ be a β-covering on $\mathcal{D}^+_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_M)$ for the uniform norm $\|.\|_\infty$, composed of symmetric matrices and defined by

$$\mathcal{R}(\beta) = \left\{A = (A_{ij})_{1\leq i,j\leq\alpha};\ A_{ij} = a_{ij}\beta;\ a_{ij} = a_{ji} \in \mathbb{Z} \cap \left[-\left\lfloor\frac{\lambda_M}{\beta}\right\rfloor, \left\lfloor\frac{\lambda_M}{\beta}\right\rfloor\right]\right\}.$$

Thus, for all Σ in $\mathcal{D}^+_{(\alpha)}(\lambda_m, \lambda_M)$, there exists A in $\mathcal{R}(\beta)$ such that

$$\|A - \Sigma\|_\infty \leq \beta. \tag{B.15}$$

The following lemma allows to compare the eigenvalues of Σ with respect to those of its associated matrix A.

Lemma B.7. *Let* $\Sigma \in \mathcal{D}^+_{(\alpha)}(\lambda_m, \lambda_M)$ *and* $A \in \mathcal{R}(\beta)$ *such that* $\|\Sigma - A\|_\infty \leq \beta$. *Let* $\lambda_1, \ldots, \lambda_\alpha$ *and* $\tau_1, \ldots, \tau_\alpha$ *be respectively the eigenvalues of* Σ *and* A*, ranked in increasing order and counted with their multiplicity.Then, for all* $q \in \{1, \ldots, \alpha\}$,

$$\tau_q - \beta\alpha \leq \lambda_q \leq \tau_q + \beta\alpha.$$

Proof. Since $\|\Sigma - A\|_\infty \leq \beta$, we have $|||\Sigma - A||| \leq \beta\alpha$. Moreover, according to theorem of Rayleigh, given for instance in Theorem 3.3.2, p. 49 in [30],

$$\lambda_q = \min_{\dim(F)=q}\ \max_{x\in F\setminus\{0\}} \frac{x'\Sigma x}{\|x\|_2^2} \text{ and } \tau_q = \min_{\dim(F)=q}\ \max_{x\in F\setminus\{0\}} \frac{x'Ax}{\|x\|_2^2}$$

where F is a linear subspace of $\mathbb{R}^\alpha$. Then, for all $q \in \{1, \ldots, \alpha\}$, $\tau_q - \beta\alpha \leq \lambda_q \leq \tau_q + \beta\alpha$. □

Covering $\mathcal{F}_{(\alpha)}$ with a family of ε-brackets. Based on the set $\mathcal{R}(\beta)$, ε-brackets for the Gaussian density family $\mathcal{F}_{(\alpha)}$ are now constructed. Consider $f = \Phi(.|\,\mu, \Sigma)$ be a function of $\mathcal{F}_{(\alpha)}$ with $\mu \in [-a, a]^\alpha$ and $\Sigma \in \mathcal{D}^+_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_M)$. For $\beta > 0$, there exists a matrix $A \in \mathcal{R}(\beta)$ such that $\|A - \Sigma\|_\infty \leq \beta$ according to (B.15). Then the two following functions are considered

$$u(x) = (1 + 2\delta)^\alpha\, \Phi\left(x|\,\nu, (1+\delta)A\right) \tag{B.16}$$

and

$$l(x) = (1+2\delta)^{-\alpha}\,\Phi\left(x|\,\nu,(1+\delta)^{-1}A\right) \tag{B.17}$$

where the vector ν and the positive number δ are adjusted later in order that $[l,u]$ is an ε-bracket of $\mathcal{F}_{(\alpha)}$ containing the function f.

Next lemma allows to fulfill hypothesis necessary to use Proposition C.1. The resulting bounds on Gaussian density ratios are given in Lemma B.9.

Lemma B.8. *Assume that* $0<\beta<\lambda_m/(3\alpha)$ *and set* $\delta = 3\beta\alpha/\lambda_m$. *Then,* $(1+\delta)A-\Sigma$ *and* $\Sigma-(1+\delta)^{-1}A$ *are both positive definite matrices. Moreover, for all* x *in* $\mathbb{R}^\alpha$,

$$x'\{(1+\delta)A-\Sigma\}x \geq \beta\alpha\|x\|_2^2 \tag{B.18}$$

and

$$x'\{\Sigma-(1+\delta)^{-1}A\}x \geq \beta\alpha\|x\|_2^2. \tag{B.19}$$

Proof. For all $x\neq 0$, since $|||A-\Sigma|||\leq\alpha\beta$,

$$\begin{aligned}
x'\{(1+\delta)A-\Sigma\}x &= (1+\delta)x'(A-\Sigma)x+\delta x'\Sigma x\\
&\geq -(1+\delta)\,|||A-\Sigma|||\,\|x\|_2^2+\delta\lambda_m\|x\|_2^2\\
&\geq \{\delta\lambda_m-(1+\delta)\alpha\beta\}\|x\|_2^2\\
&\geq \left(\frac{2}{3}\delta\lambda_m-\alpha\beta\right)\|x\|_2^2
\end{aligned}$$

because $\alpha\beta\leq\lambda_m/3$. Then $x'\{(1+\delta)A-\Sigma\}x\geq\alpha\beta\|x\|_2^2>0$ according to the definition of δ. Similarly,

$$\begin{aligned}
x'\{\Sigma-(1+\delta)^{-1}A\}x &= (1+\delta)^{-1}x'(\Sigma-A)x+\{1-(1+\delta)^{-1}\}x'\Sigma x\\
&\geq\left(\frac{\delta\lambda_m-\alpha\beta}{1+\delta}\right)\|x\|_2^2\\
&\geq\frac{2\alpha\beta}{1+\delta}\|x\|_2^2\\
&\geq\alpha\beta\|x\|_2^2>0.
\end{aligned}$$

□

Lemma B.9. *Assume that* $\beta<\lambda_m/(3\alpha)$ *and set* $\delta=3\beta\alpha/\lambda_m$. *Then,*

$$\frac{f(x)}{u(x)}\leq(1+2\delta)^{-\frac{\alpha}{2}}\exp\left(\frac{\|\mu-\nu\|_2^2}{2\beta\alpha}\right)$$

and

$$\frac{l(x)}{f(x)}\leq(1+2\delta)^{-\frac{\alpha}{2}}\,\exp\left(\frac{\|\mu-\nu\|_2^2}{2\beta\alpha}\right).$$

Proof. According to Proposition C.1, since $(1+\delta)A-\Sigma$ is a positive definite matrix from Lemma B.8,

$$\frac{f(x)}{u(x)}\leq(1+2\delta)^{-1}\sqrt{\frac{|(1+\delta)A|}{|\Sigma|}}\;\exp\left[\frac{1}{2}(\mu-\nu)'\{(1+\delta)A-\Sigma\}^{-1}(\mu-\nu)\right].$$

Inequality (B.18) implies that $|||\{(1+\delta)A-\Sigma\}^{-1}|||=\{\inf\lambda\}^{-1}\leq(\beta\alpha)^{-1}$ where the infimum is taken over all eigenvalues of $(1+\delta)A-\Sigma$. Then, since

$$(\mu-\nu)'\{(1+\delta)A-\Sigma\}^{-1}(\mu-\nu)\leq|||\{(1+\delta)A-\Sigma\}^{-1}|||\,\|\mu-\nu\|_2^2,$$

this leads to

$$(\mu-\nu)'\{(1+\delta)A-\Sigma\}^{-1}(\mu-\nu) \leq \frac{\|\mu-\nu\|_2^2}{\alpha\beta}.$$

Moreover, according to Lemma B.7,

$$\begin{aligned}\frac{|(1+\delta)A|}{|\Sigma|} &= (1+\delta)^{\alpha}\prod_{q=1}^{\alpha}\frac{\tau_q}{\lambda_q}\\ &\leq (1+\delta)^{\alpha}\prod_{q=1}^{\alpha}\left(1+\frac{\beta\alpha}{\lambda_q}\right)\\ &\leq (1+\delta)^{\alpha}\left(1+\frac{\beta\alpha}{\lambda_m}\right)^{\alpha}\\ &\leq (1+2\delta)^{\alpha}.\end{aligned}$$

Then

$$\frac{f(x)}{u(x)} \leq (1+2\delta)^{-\frac{\alpha}{2}}\exp\left(\frac{\|\mu-\nu\|_2^2}{2\beta\alpha}\right).$$

Similarly, using Proposition C.1, Inequality (B.19) and Lemma B.7, we obtain

$$\frac{l(x)}{f(x)} \leq (1+2\delta)^{-\frac{\alpha}{2}}\ \exp\left(\frac{\|\mu-\nu\|_2^2}{2\beta\alpha}\right).$$ □

Next proposition terminates the construction of an ε-bracket covering of $\mathcal{F}_{(\alpha)}$.

Proposition B.10. *For all $\varepsilon\in(0,1]$, we define $\delta=\varepsilon/(\sqrt{3}\,\alpha)$ and $\beta=\lambda_m\varepsilon/(3\sqrt{3}\alpha^2)$. The following set*

$$\left\{[l,u];\ \begin{array}{l}u(x)=(1+2\delta)^{\alpha}\,\Phi\left(x|\,\nu,(1+\delta)A\right)\\ l(x)=(1+2\delta)^{-\alpha}\,\Phi\left(x|\,\nu,(1+\delta)^{-1}A\right)\end{array}\ ;\ A\in\mathcal{R}(\beta),\ \nu\in\mathcal{X}(\varepsilon,a,\lambda_m,\alpha)\right\}$$

where

$$\mathcal{X}(\varepsilon,a,\lambda_m,\alpha)=\left\{\nu=(\nu_1,\ldots,\nu_\alpha);\ \nu_q=\frac{\sqrt{\lambda_m}\,\varepsilon}{3\alpha}s_q;\ s_q\in\mathbb{Z}\cap\left[-\left\lfloor\frac{3\,a\,\alpha}{\sqrt{\lambda_m}\,\varepsilon}\right\rfloor,\left\lfloor\frac{3\,a\,\alpha}{\sqrt{\lambda_m}\,\varepsilon}\right\rfloor\right]\right\},$$

is an ε-bracket set over $\mathcal{F}_{(\alpha)}$.

Proof. Let $f(x)=\Phi(x|\mu,\Sigma)$ be a function of $\mathcal{F}_{(\alpha)}$ where $\mu\in[-a,a]^{\alpha}$ and $\Sigma\in\mathcal{D}^{+}_{(\alpha)}(\lambda_m,\lambda_M)$. There exists A in $\mathcal{R}(\beta)$ such that $\|\Sigma-A\|_{\infty}\leq\beta$ and ν in $\mathcal{X}(\varepsilon,a,\lambda_m,\alpha)$ satisfying, for all q in $\{1,\ldots,\alpha\}$, $|\mu_q-\nu_q|\leq\sqrt{\lambda_m}\varepsilon/(3\alpha)$. Consider the two associated functions l and u defined in (B.16) and (B.17) respectively. Since $\|\mu-\nu\|_2^2\leq\lambda_m\varepsilon^2/(9\alpha)$, using Lemma B.9,

$$\frac{f(x)}{u(x)} \leq (1+2\delta)^{-\frac{\alpha}{2}}\exp\left(\frac{\sqrt{3}\,\varepsilon}{6}\right).$$

Thus, noting that for all x in $[0,2]$, $\ln(1+x)\geq x/2$, it leads to

$$\begin{aligned}\ln\left\{\frac{f(x)}{u(x)}\right\} &\leq -\frac{\alpha}{2}\ln\left(1+\frac{2\,\varepsilon}{\sqrt{3}\alpha}\right)+\frac{\sqrt{3}\,\varepsilon}{6}\\ &\leq -\frac{\alpha}{2}\frac{\varepsilon}{\sqrt{3}\alpha}+\frac{\varepsilon}{2\,\sqrt{3}}\leq 0.\end{aligned}$$

Similarly, $\ln\{l(x)/f(x)\} \leq 0$ and thus for all $x \in \mathbb{R}^\alpha$, $l(x) \leq f(x) \leq u(x)$. It remains to bound the size of bracket $[l, u]$ with respect to Hellinger distance. According to Proposition C.3,

$$\begin{aligned} d_H^2(l,u) &= (1+2\delta)^\alpha + (1+2\delta)^{-\alpha} \\ &\quad - \left\{2 - d_H^2(\Phi(.|\nu,(1+\delta)A), \Phi(.|\nu,(1+\delta)^{-1}A))\right\} \\ &= 2\left(\cosh\{\alpha\ln(1+2\delta)\} - 1 + 1 - [\cosh\{\ln(1+\delta)\}]^{-\frac{\alpha}{2}}\right) \\ &\leq 2\left(\sinh(1)\alpha^2\delta^2 + \frac{1}{4}\alpha^2\delta^2\right) \\ &\leq 3\alpha^2\delta^2 = \varepsilon^2. \end{aligned}$$

□

Proof of Proposition B.6. Since the set of ε-brackets over $\mathcal{F}_{(\alpha)}$, described in the Proposition B.10 is totally defined by the parameter spaces $\mathcal{R}(\beta)$ and $\mathcal{X}(\varepsilon, a, \lambda_m, \alpha)$, an upper bound of the bracketing entropy of $\mathcal{F}_{(\alpha)}$ is deduced from an upper bound of the two set cardinals.

$$\begin{aligned} \mathcal{N}_{[.]}\left(\varepsilon, \mathcal{F}_{(\alpha)}, d_H\right) &\leq \operatorname{card}\{\mathcal{R}(\beta)\} \times \operatorname{card}\{\mathcal{X}(\varepsilon, a, \lambda_m, \alpha)\} \\ &\leq \left(\frac{2\lambda_M}{\beta}\right)^{\frac{\alpha(\alpha+1)}{2}} \left(\frac{6\,a\alpha}{\sqrt{\lambda_m}\,\varepsilon}\right)^\alpha \\ &\leq \left(\frac{6\sqrt{3}\lambda_M\alpha^2}{\varepsilon\lambda_m}\right)^{\frac{\alpha(\alpha+1)}{2}} \left(\frac{6\,a\alpha}{\sqrt{\lambda_m}\,\varepsilon}\right)^\alpha. \end{aligned}$$

Thus, since $\ln(\alpha)$ and $\ln(\alpha^2)$ are bounded by $\ln(Q^2)$,

$$\mathcal{H}_{[.]}\left(\varepsilon, \mathcal{F}_{(\alpha)}, d_H\right) \leq \frac{\alpha(\alpha+1)}{2}\ln\left(\frac{6\sqrt{3}\lambda_M}{\lambda_m}\right) + \alpha\ln\left(\frac{6\,a}{\sqrt{\lambda_m}}\right) + \left\{\frac{\alpha(\alpha+1)}{2} + \alpha\right\}\ln(Q^2) + \left\{\frac{\alpha(\alpha+1)}{2} + \alpha\right\}\ln\left(\frac{1}{\varepsilon}\right).$$

□

B.5. Control of the bracketing entropy for the $[LC]$ collection

This section is devoted to upper bound the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$ for the $[LC]$ Gaussian mixture collection. According to Inequality (B.5), it remains to control the bracketing entropy of the family $\mathcal{F}_{(K,\alpha)}$ defined by (1.7).

Proposition B.11. *For all $\varepsilon \in (0,1]$, the bracketing number of $\mathcal{F}_{(K,\alpha)}$ is controlled by*

$$\mathcal{N}_{[.]}\left(\varepsilon, \mathcal{F}_{(K,\alpha)}, d_H\right) \leq \left(\frac{6\sqrt{3}\lambda_M\alpha^2}{\lambda_m}\right)^{\frac{\alpha(\alpha+1)}{2}} \left(\frac{6a\alpha}{\sqrt{\lambda_m}}\right)^{K\alpha} \left(\frac{1}{\varepsilon}\right)^{K\alpha + \frac{\alpha(\alpha+1)}{2}}$$

where $K\alpha + \frac{\alpha(\alpha+1)}{2}$ is the dimension of $\mathcal{F}_{(K,\alpha)}$. Hence the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$ is upper bounded by

$$\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\alpha)}, d_H) \leq \mathcal{I} + D(K,\alpha)\ln\left(\frac{1}{\varepsilon}\right)$$

where

$$\mathcal{I} = C(K) + \ln\left(\frac{24\sqrt{2}\,\lambda_M}{\lambda_m}\right) + K\frac{\alpha(\alpha+1)}{2}\ln\left(\frac{54\sqrt{3}\lambda_M}{\lambda_m}\right) + K\alpha\ln\left(\frac{54\,a}{\sqrt{\lambda_m}}\right) + \left\{\frac{\alpha(\alpha+1)}{2} + K\alpha + 1\right\}\ln(Q^2)$$

with $C(K) = \ln(K) + \frac{K}{2}\ln(2\pi \mathrm{e}) + (K-1)\ln(9)$.

Proof. This result is obtained by considering the following bracket family. Its construction is inspired by the bracket family used in the study of the bracketing entropy of $\mathcal{F}_{(\alpha)}$ for the $[L_kC_k]$ collection (see Appendix B.4).

For all $\varepsilon \in (0,1]$, let $\delta = \frac{\varepsilon}{\sqrt{3}\,\alpha}$ and $\beta = \frac{\lambda_m \varepsilon}{3\sqrt{3}\alpha^2}$. The following set

$$\left\{ ([l_1,u_1],\ldots,[l_K,u_K]);\ \begin{array}{l} u_k(x) = (1+2\delta)^{\alpha}\Phi\left(x|\nu_k,(1+\delta)A\right) \\ l_k(x) = (1+2\delta)^{-\alpha}\Phi\left(x|\nu_k,(1+\delta)^{-1}A\right) \\ A \in \mathcal{R}(\beta), \nu_k \in \mathcal{X}(\varepsilon,a,\lambda_m,\alpha) \end{array} \right\}$$

where

$$\mathcal{R}(\beta) = \left\{ A = (A_{ij})_{1\le i,j\le \alpha};\ A_{ij} = a_{ij}\beta;\ a_{ij} = a_{ji} \in \mathbb{Z} \cap \left[-\left\lfloor \frac{\lambda_M}{\beta} \right\rfloor, \left\lfloor \frac{\lambda_M}{\beta} \right\rfloor \right] \right\}$$

and

$$\mathcal{X}(\varepsilon,a,\lambda_m,\alpha) = \left\{ \nu = (\nu_1,\ldots,\nu_\alpha);\ \nu_q = \frac{\sqrt{\lambda_m}\,\varepsilon}{3\alpha} s_q;\ s_q \in \mathbb{Z} \cap \left[-\left\lfloor \frac{3\,a\,\alpha}{\sqrt{\lambda_m}\,\varepsilon} \right\rfloor, \left\lfloor \frac{3\,a\,\alpha}{\sqrt{\lambda_m}\,\varepsilon} \right\rfloor \right] \right\}$$

is an ε-bracket set over $\mathcal{F}_{(K,\alpha)}$. Finally the bracketing number of $\mathcal{F}_{(K,\alpha)}$ is upper bounded by

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{F}_{(K,\alpha)}, d_H) \le \left(\frac{6\sqrt{3}\lambda_M \alpha^2}{\lambda_m} \right)^{\frac{\alpha(\alpha+1)}{2}} \times \left(\frac{6a\alpha}{\sqrt{\lambda_m}} \right)^{K\alpha} \times \left(\frac{1}{\varepsilon} \right)^{K\alpha + \frac{\alpha(\alpha+1)}{2}}.$$

This allows to end the proof. □

C. Results for multivariate Gaussian densities

C.1. Ratio of two Gaussian densities

Proposition C.1. *Let $\Phi(.|\mu_1,\Sigma_1)$ and $\Phi(.|\mu_2,\Sigma_2)$ be two Gaussian densities. If $\Sigma_2 - \Sigma_1$ is a positive definite matrix then for all $x \in \mathbb{R}^Q$,*

$$\frac{\Phi(x|\mu_1,\Sigma_1)}{\Phi(x|\mu_2,\Sigma_2)} \le \sqrt{\frac{|\Sigma_2|}{|\Sigma_1|}} \exp\left\{ \frac{1}{2}(\mu_1-\mu_2)'(\Sigma_2-\Sigma_1)^{-1}(\mu_1-\mu_2) \right\}.$$

Proof. The ratio between the two Gaussian densities is equal to

$$\frac{\Phi(x|\mu_1,\Sigma_1)}{\Phi(x|\mu_2,\Sigma_2)} = \sqrt{\frac{|\Sigma_2|}{|\Sigma_1|}} \exp\left[-\frac{1}{2} \left\{ (x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) - (x-\mu_2)'\Sigma_2^{-1}(x-\mu_2) \right\} \right].$$

Proposition C.1 is proved if

$$(x-\mu_2)'\Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) \le (\mu_1-\mu_2)'(\Sigma_2-\Sigma_1)^{-1}(\mu_1-\mu_2). \qquad \text{(C.1)}$$

The matrix $\Sigma_1^{-1} - \Sigma_2^{-1} = \Sigma_1^{-1}(\Sigma_2-\Sigma_1)\Sigma_2^{-1}$ is a positive definite matrix as the product of three positive definite matrices. Defining $\mu^\star = (\Sigma_1^{-1} - \Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2)$,

$$\begin{aligned} (x-\mu_2)'\Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) &= (\mu_1-\mu_2)'(\Sigma_2-\Sigma_1)^{-1}(\mu_1-\mu_2) \\ &\quad -(x-\mu^\star)'(\Sigma_1^{-1}-\Sigma_2^{-1})(x-\mu^\star). \end{aligned}$$

Since the matrix $(\Sigma_1^{-1} - \Sigma_2^{-1})$ is a positive definite matrix, $(x-\mu^\star)'(\Sigma_1^{-1}-\Sigma_2^{-1})(x-\mu^\star) \ge 0$ and it leads to Inequality (C.1). □

Corollary C.2. *Let $\Phi(.|\mu_1, B_1)$ and $\Phi(.|\mu_2, B_2)$ be two Gaussian densities. Their variance matrice are assumed to have the diagonal form $B_i = \operatorname{diag}(b_{i1}^2, \ldots, b_{iQ}^2)$ for all $i = 1, 2$ such that $b_{2q}^2 > b_{1q}^2 > 0$ for all $q \in \{1 \ldots, Q\}$. Then, for all $x \in \mathbb{R}^Q$, the ratio of the two densities is bounded by*

$$\left(\prod_{q=1}^{Q} \frac{b_{2q}}{b_{1q}}\right) \exp\left\{\frac{1}{2}(\mu_1 - \mu_2)' \operatorname{diag}\left(\frac{1}{b_{21}^2 - b_{11}^2}, \ldots, \frac{1}{b_{2Q}^2 - b_{1Q}^2}\right)(\mu_1 - \mu_2)\right\}.$$

C.2. Hellinger distance between two Gaussian densities

The following proposition gives the expression of the Hellinger distance between two Gaussian densities.

Proposition C.3. *Let $\Phi(.|\mu_1, \Sigma_1)$ and $\Phi(.|\mu_2, \Sigma_2)$ be two Gaussian densities. The squared Hellinger distance between these two densities has the following expression:*

$$2\left[1 - 2^{\frac{Q}{2}} |\Sigma_1 \Sigma_2|^{-\frac{1}{4}} |\Sigma_1^{-1} + \Sigma_2^{-1}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{4}(\mu_1 - \mu_2)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)\right\}\right].$$

Proof. According to the definition of the Hellinger distance,

$$d_H^2(\Phi(.|\mu_1, \Sigma_1), \Phi(.|\mu_2, \Sigma_2)) = 2 - 2\int \sqrt{\Phi(x|\mu_1, \Sigma_1)\, \Phi(x|\mu_2, \Sigma_2)}\mathrm{d}x.$$

Furthermore,

$$\Phi(x|\mu_1, \Sigma_1)\, \Phi(x|\mu_2, \Sigma_2) = (2\pi)^{-Q}\ |\Sigma_1 \Sigma_2|^{-\frac{1}{2}}\ \exp\left[-\frac{1}{2}\left\{(x - \mu_1)'\Sigma_1^{-1}(x - \mu_1) + (x - \mu_2)'\Sigma_2^{-1}(x - \mu_2)\right\}\right].$$

Defining $\mu^\star = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2)$, we deduce that

$$(x - \mu_1)'\Sigma_1^{-1}(x - \mu_1) + (x - \mu_2)'\Sigma_2^{-1}(x - \mu_2) = (x - \mu^\star)'(\Sigma_1^{-1} + \Sigma_2^{-1})(x - \mu^\star) + (\mu_1 - \mu_2)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2).$$

Finally, the squared distance $d_H^2(\Phi(.|\mu_1, \Sigma_1), \Phi(.|\mu_2, \Sigma_2))$ is equal to

$$2 - 2(2\pi)^{-\frac{Q}{2}} |\Sigma_1 \Sigma_2|^{-\frac{1}{4}} \exp\left\{-\frac{1}{4}(\mu_1 - \mu_2)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)\right\} \int \exp\left\{-\frac{1}{4}(x - \mu^\star)'(\Sigma_1^{-1} + \Sigma_2^{-1})(x - \mu^\star)\right\} \mathrm{d}x =$$

$$2 - 2(2\pi)^{-\frac{Q}{2}} |\Sigma_1 \Sigma_2|^{-\frac{1}{4}} \exp\left\{-\frac{1}{4}(\mu_1 - \mu_2)'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)\right\} \times (4\pi)^{\frac{Q}{2}} |\Sigma_1^{-1} + \Sigma_2^{-1}|^{-\frac{1}{2}}$$

that entails the concluding result. □

Corollary C.4. *Using the notation of Corollary C.2, the Hellinger distance of two Gaussian densities with diagonal variance matrices is given by the following expression*

$$2 - 2\left(\prod_{q=1}^{Q} \frac{2\, b_{1q}\, b_{2q}}{b_{1q}^2 + b_{2q}^2}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{4}(\mu_1 - \mu_2)' \operatorname{diag}\left\{\left(\frac{1}{b_{1q}^2 + b_{2q}^2}\right)_{1 \le q \le Q}\right\}(\mu_1 - \mu_2)\right].$$

Acknowledgements. The authors are grateful to Sylvain Arlot (CNRS Ulm), Gilles Celeux (INRIA), Marie-Laure Martin-Magniette (INRA) and Pascal Massart (Université Paris-Sud 11) for helpful discussions and valuable comments.

References

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

[2] S. Arlot and P. Massart, Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.* (2008) (to appear).

[3] J.D. Banfield and A.E. Raftery, Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** (1993) 803–821.

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

[5] J.-P. Baudry, *Clustering through model selection criteria*. Poster session at One Day Statistical Workshop in Lisieux. http://www.math.u-psud.fr/~baudry, June (2007).

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

[7] C. Biernacki, G. Celeux, G. Govaert and F. Langrognet, Model-based cluster and discriminant analysis with the MIXMOD software. *Comput. Stat. Data Anal.* **51** (2006) 587–600.

[8] L. Birgé and P. Massart, Gaussian model selection. *J. Eur. Math. Soc.* **3** (2001) 203–268.

[9] L. Birgé and P. Massart, *A generalized C_p criterion for Gaussian model selection*. Prépublication n° 647, Universités de Paris 6 et Paris 7 (2001).

[10] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Prob. Th. Rel. Fields* **138** (2007) 33–73.

[11] L. Birgé and P. Massart, From model selection to adaptive estimation, in *Festschrift for Lucien Le Cam*. Springer, New York (1997) 55–87.

[12] C. Bouveyron, S. Girard and C. Schmid, High-Dimensional Data Clustering. *Comput. Stat. Data Anal.* **52** (2007) 502–519.

[13] K.P. Burnham and D.R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York, 2nd edition (2002).

[14] G. Castellan, *Modified Akaike's criterion for histogram density estimation*. Technical report, Université Paris-Sud 11 (1999).

[15] G. Castellan, Density estimation via exponential model selection. *IEEE Trans. Inf. Theory* **49** (2003) 2052–2060.

[16] G. Celeux and G. Govaert, Gaussian parsimonious clustering models. *Pattern Recogn.* **28** (1995) 781–793.

[17] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc, Ser. B.* **39** (1977) 1–38.

[18] C.R. Genovese and L. Wasserman, Rates of convergence for the Gaussian mixture sieve. *Ann. Stat.* **28** (2000) 1105–1127.

[19] S. Ghosal and A.W. van der Vaart, Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Stat.* **29** (2001) 1233–1263.

[20] C. Keribin, Consistent estimation of the order of mixture models. *Sankhyā. The Indian Journal of Statistics. Series A* **62** (2000) 49–66.

[21] M.H. Law, M.A.T. Figueiredo and A.K. Jain, Simultaneous feature selection and clustering using mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **26** (2004) 1154–1166.

[22] E. Lebarbier, Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Proc.* **85** (2005) 717–736.

[23] V. Lepez, *Potentiel de réserves d'un bassin pétrolier: modélisation et estimation*. Ph.D. thesis, Université Paris-Sud 11 (2002).

[24] P. Massart, *Concentration inequalities and model selection*. Springer, Berlin (2007). Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23 (2003).

[25] C. Maugis, *Sélection de variables pour la classification non supervisée par mélanges gaussiens. Applications à l'étude de données transcriptomes*. Ph.D. thesis, University Paris-Sud 11 (2008).

[26] C. Maugis, G. Celeux and M.-L. Martin-Magniette, Variable Selection for Clustering with Gaussian Mixture Models. *Biometrics* (2008) (to appear).

[27] C. Maugis and B. Michel, *Slope heuristics for variable selection and clustering via Gaussian mixtures*. Technical Report 6550, INRIA (2008).

[28] A.E. Raftery and N. Dean, Variable Selection for Model-Based Clustering. *J. Am. Stat. Assoc.* **101** (2006) 168–178.

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

[30] D. Serre, *Matrices*. Springer-Verlag, New York (2002).

[31] M. Talagrand, Concentration of measure and isoperimetric inequalities in product spaces. *Publ. Math., Inst. Hautes Étud. Sci.* **81** (1995) 73–205.

[32] M. Talagrand, New concentration inequalities in product spaces. *Invent. Math.* **126** (1996) 505–563.

[33] F. Villers, *Tests et sélection de modèles pour l'analyse de données protéomiques et transcriptomiques*. Ph.D. thesis, University Paris-Sud 11 (2007).