A non asymptotic penalized criterion for Gaussian mixture model selection

Cathy Maugis, Bertrand Michel

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A non asymptotic penalized criterion for Gaussian mixture model selection

Cathy Maugis — Bertrand Michel

N° 6549

Juin 2008

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--6549--FR+ENG

A non asymptotic penalized criterion for Gaussian mixture model selection

Cathy Maugis* , Bertrand Michel †

Thème COG — Systèmes cognitifs
Projets SELECT

Rapport de recherche n° 6549 — Juin 2008 — 33 pages

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 MLE proposed by Massart (2007) 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.

Key-words: Model-based clustering, Variable selection, Penalized likelihood criterion, Bracketing entropy.

* INRIA Futurs, Projet SELECT, Université Paris-Sud 11
† INRIA Futurs, Projet SELECT, Université Paris-Sud 11

Unité de recherche INRIA Futurs
Parc Club Orsay Université, ZAC des Vignes,
4, rue Jacques Monod, 91893 ORSAY Cedex (France)
Téléphone : +33 1 72 92 59 00 — Télécopie : +33 1 60 19 66 08

Un critère pénalisé non asymptotique pour la sélection de modèle de mélanges gaussiens

Résumé : Des mélanges gaussiens de formes spécifiques sont considérés pour résoudre un problème de sélection de variables en classification non supervisée. Un critère pénalisé non asymptotique est proposé pour sélectionner le nombre de composantes du mélange et l'ensemble des variables pertinentes pour cette classification. Le contraste Kullback-Leibler ayant un comportement non linéaire sur les mélanges gaussiens, un théorème général de sélection de modèles pour l'estimation de densités par maximum de vraisemblance du à Massart (2007) est utilisé pour déterminer la forme de la pénalité. Ce théorème nécessite le contrôle de l'entropie à crochet des familles de mélanges gaussiens étudiées. Le cas des variables ordonnées et celui des variables non ordonnées sont tous deux considérés dans cet article.

Mots-clés : Classification non supervisée, Mélanges gaussiens, Sélection de variables, Critère pénalisé, Entropie à crochet.

1 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 into account the variable roles.

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. (2004) 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 (2006) 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 in Maugis et al. (2007) where irrelevant variables can only depend on a relevant clustering variable subset and variables can have different sizes (block variables). Since all these methods are based on a variable selection procedure included into the clustering process, they do not impose specific constraints on Gaussian mixture forms. On the contrary, Bouveyron et al. (2007) consider a suitable Gaussian mixture family to take into account that data live in low-dimensional subspaces hidden in the original space. 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 this paper, a new variable method for clustering is proposed. It 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 2.1. The problem can be formulated as a choice of a model among the collection since this choice automatically leads to a data clustering and a variable selection. Thus, a data-driven criterion is needed to select the "best" model among the model collection. With a non asymptotic point of view, the "best" model is the one whose the associated maximum likelihood estimator of s gives the lowest estimation error. In this sense, the variable selection is not only beneficial to the clustering problem but also to the estimation problem.

In the density estimation framework, the principle of selecting a model by penalizing a loglikelihood type criterion has emerged during the seventies. Akaike (1973) proposed the AIC criterion (Akaike's information criterion) and Schwarz (1978) suggested the BIC (Bayesian Information Criterion). These two classical criteria assume implicitly that the true distribution belongs to the model collection (see for instance Burnham and Anderson, 2002). With a different point of view, the criterion ICL (Integrated Completed Likelihood) proposed by Biernacki et al. (2000) takes into account the clustering aim. Although the behaviours of these asymptotic criteria were tested in practice, their theoretical properties are few or not proved. For instance, the BIC consistency is only stated for cluster number under restrictive regularity assumptions and assuming that the true density belongs to the considered Gaussian mixture family (Keribin, 2000).

A non asymptotic approach for model selection via penalization has emerged during the last ten years, mainly with works of Birgé and Massart (1997) and Barron et al. (1999). An overview is available in Massart (2007). 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 parameter number 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 (2007) is a version of Theorem 2 in Barron et al. (1999). Its application requires to control the bracketing entropy of the considered Gaussian mixture models.

The paper is organized as follows: Section 2 gives the model selection principles. The Gaussian mixture models considered in this paper are described in Section 2.1 and principles of non asymptotic theory for density estimation based on Kullback-Leibler contrast are reviewed in Section 2.2. This section is completed by the statement of the general model selection theorem. The main results are stated in Section 3 and a discussion is given in Section 4. Results of Section A.1 recast the control of the bracketing entropy of mixture families into a control of the bracketing entropy of Gaussian density families. This bracketing entropy is upper bounded for Gaussian mixtures with diagonal variance matrices in Appendix A.2 and with general variance matrices in Appendix A.3. The proof of the main results is given in Appendix B.

2 Model selection principles

2.1 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, which are 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 ($\forall k = 1, \ldots, K, 0 < p_k < 1$ 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 $(p_1, \ldots, p_K, \eta_1, \ldots, \eta_K, \Lambda_1, \ldots, \Lambda_K)$.

The mixture model is an incomplete data structure model: the complete data are $((\mathbf{y}_1, \mathbf{z}_1), \ldots, (\mathbf{y}_n, \mathbf{z}_n))$ where the missing data are $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ with $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ such that $z_{ik} = 1$ iff $\mathbf{y}_i$ 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 (Dempster et al., 1977), a data clustering is deduced from the maximum a posteriori principle:

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

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 into account the variable role in the clustering process. 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 definite positive matrices. This idea is now formalized.

Let $\mathcal{V}$ be the collection of the nonempty subsets of $\{1, \ldots, Q\}$. A Gaussian mixture family is characterized by its number of mixture components $K \in \mathbb{N}^\star$ 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 $\mathcal{M} = \mathbb{N}^\star \times \mathcal{V}$. Consider the decomposition of a vector $x \in \mathbb{R}^Q$ into its restriction on relevant variables $x_{[\mathbf{v}]} = (x_{j_1}, \ldots, x_{j_\alpha})'$ and its restriction on irrelevant variables $x_{[\mathbf{v}^c]} = (x_{l_1}, \ldots, x_{l_{Q-\alpha}})'$ where $\mathbf{v} = \{j_1, \ldots, j_\alpha\}$ and $\mathbf{v}^c = \{l_1, \ldots, l_{Q-\alpha}\} = \{1, \ldots, Q\} \backslash \mathbf{v}$. On relevant variables, a Gaussian mixture f is chosen among the following mixture family

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

where $0 < a$, $0 < \lambda_{\mathrm{m}} < \lambda_{\mathrm{M}}$ and $\Delta^{+}_{(K,\alpha)}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}})$ denotes a family of K-uples of $\alpha \times \alpha$ symmetric definite positive matrices whose eigenvalues belong to the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. The family $\Delta^{+}_{(K,\alpha)}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}})$ is related to the Gaussian mixture shape specified hereafter. The associated set of Gaussian densities composing mixtures of $\mathcal{L}_{(K,\alpha)}$ is denoted $\mathcal{F}_{(\alpha)}$. On irrelevant variables spherical Gaussian density g is considered, belonging to the following family

$$\mathcal{G}_{(\alpha)} = \left\{ \Phi(.|0, \omega^2 I_{Q-\alpha});\ \omega^2 \in [\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}] \right\}. \tag{1}$$

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

$$\mathcal{S}_{(K,\mathbf{v})} = \left\{ x \in \mathbb{R}^Q \mapsto f(x_{[\mathbf{v}]})\, g(x_{[\mathbf{v}^c]});\ f \in \mathcal{L}_{(K,\alpha)},\ g \in \mathcal{G}_{(\alpha)} \right\}. \tag{2}$$

The dimension of the model $\mathcal{S}_{(K,\mathbf{v})}$ is denoted $D(K, \mathbf{v})$ and corresponds to the free parameter number common to all Gaussian mixtures in this model. It only depends on the number of components K and the number of relevant variables α.

In this paper, two collections of Gaussian mixtures are considered but the same notation $\mathcal{S}_{(K,\mathbf{v})}$ is used for the two model collections to make easier the reading of the article.

- For the diagonal collection: the variance matrices Σ_k on relevant variables are assumed to be diagonal matrices. Thus the relevant variables are independent conditionally to mixture component belonging. In this context, the set of Gaussian densities composing mixtures of $\mathcal{L}_{(K,\alpha)}$ is defined by

  $$\mathcal{F}_{(\alpha)} = \left\{ \Phi(.|\mu, \Sigma);\ \mu \in [-a, a]^{\alpha},\ \Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_\alpha^2),\ \sigma_1^2, \ldots, \sigma_\alpha^2 \in [\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}] \right\} \tag{3}$$

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

- For the general collection: the variance matrices are assumed to be totally free. The variance matrices belong to the set $\mathcal{D}^{+}_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}})$ of $\alpha \times \alpha$ positive definite matrices whose eigenvalues are to the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. The relevant variables are thus admitted to be correlated conditionally to mixture component belonging. The Gaussian density family composing mixtures is

  $$\mathcal{F}_{(\alpha)} = \left\{ w \in \mathbb{R}^{\alpha} \mapsto \Phi(w|\mu, \Sigma),\ \mu \in [-a, a]^{\alpha},\ \Sigma \in \mathcal{D}^{+}_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}) \right\} \tag{4}$$

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

Whereas a mixture with diagonal variance matrices can be seen as an element of the general collection, its number of free parameters is lower than the dimension of the corresponding general model for a couple $(K, \mathbf{v})$. It is interesting to consider the two different collections since the results obtained further are stated in function of the model dimension. Furthermore this distinction allows to have several mixture collections to cluster datasets in practice. Moreover 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 $\mathbf{v} = \{1, \ldots, \alpha\}$ and can be assimilated to its cardinal α. Thus, in order to distinguish between the two cases, Gaussian mixture families are written $\mathcal{S}_{(K,\alpha)}$ and their dimensions are denoted $D(K, \alpha)$ when variables are ordered.

These Gaussian mixture families allow to recast clustering and variable selection problems in 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 reminded in the following section.

2.2 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 obvious 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 data-driven criterion to select a model of proper dimension among a given list. A general theory on this topic, with a non asymptotic approach is proposed in the works of Birgé and Massart (see for instance Birgé and Massart, 2001a,b). 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

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

if sdx is absolutely continuous with respect to tdx and $+\infty$ otherwise. Noticing that being the unique minimizer of the Kullback-Leibler function on $\mathcal{S}$, s satisfies

$$s = \underset{t\in\mathcal{S}}{\operatorname{argmin}} \int -\ln\{t(x)\}s(x)\, dx.$$

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

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

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 now 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 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 $\text{pen} : m \in \mathcal{M} \mapsto \text{pen}(m) \in \mathbb{R}_+$ and of selecting $\hat{m}$ minimizing the associated penalized criterion

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

The resulting selected estimator is $\hat{s}_{\hat{m}}$. 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 $\inf_{m\in\mathcal{M}} \mathbb{E}[\text{KL}(s, \hat{s}_m)]$.

Commonly, in order to find a suitable penalty function, one begins by writing the following inequality (see Massart, 2007, p.9): For all $m \in \mathcal{M}$ and $s_m \in S_m$,

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

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 (Talagrand, 1995, 1996; Massart, 2007, 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 context of histogram density estimation (Castellan, 1999) and of density estimation via exponential model (Castellan, 2003). For situations when these sharp calculations are impossible to obtain, Massart (2007, Section 7.4) 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. Another requirement to apply concentration inequalities is to fulfill an hypothesis of boundness as $\|\bar{\gamma}_n(s_m) - \bar{\gamma}_n(t)\|_\infty$ is bounded by a constant for all $t \in S_{m'}$. But in our context, it is impossible to bound uniformly all the ratios of two Gaussian mixtures. After remembering the definition of the Hellinger distance and specifying some notation, this general MLE selection model theorem (Massart, 2007, Theorem 7.11) is stated in a restricted form, which is sufficient for our study.

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 be 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 $S_m(u, \xi) = \{t \in S_m; d_H(t, u) \leq \xi\}$,

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

Theorem 1 (Massart (2007))**.** *Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ 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*

$$\sum_{m\in\mathcal{M}} e^{-\rho_m} = \Upsilon < \infty. \tag{5}$$

*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,*

$$\mathbb{E}\left[d_H^2(s, \hat{s}_{\hat{m}})\right] \leq C\left[\inf_{m\in\mathcal{M}}\{\mathrm{KL}(s, S_m) + \mathrm{pen}(m)\} + \frac{\Upsilon}{n}\right], \tag{6}$$

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

Inequality (6) 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 Massart (2007). In our context, this condition can be achieved even if it means assuming that all densities are defined on compact support.

3 Main results

As announced previously, Theorem 1 is applied to our specific framework described in Section 2.1. The ensuing theoretical results are now addressed, for the ordered and non-ordered variable cases separately. These provide a non asymptotic penalized criterion to select the number of clusters K and the variable subset $\mathbf{v}$ used in the Gaussian mixtures. Moreover, these results give an oracle inequality which is fulfilled by the associated penalized estimator.

3.1 Ordered variable case

In this section, variables are assumed to be ordered and we recall that the model collection is denoted $(\mathcal{S}_{(K,\alpha)})_{(K,\alpha)\in\mathcal{M}}$ in this case. In the two types of Gaussian mixtures, the following theorem gives the form of penalty functions and the associated oracle inequalities for this model collection.

Theorem 2. *For the diagonal and general mixture collections, there exists two absolute constants κ and C such that, if*

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

where A is a function of Q, λ_m, λ_M and a such that $A = O(\sqrt{\ln Q})$ as Q goes to infinity , then the model $(\hat{K}, \hat{\alpha})$ minimizing

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

over $\mathcal{M}$ exists and

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

This theorem is proved in Section B.1. It requires to control the bracketing entropy of Gaussian mixture families. In Section A.1, this problem is recast into the control for Gaussian density families. Section A.2 and Section A.3 are then devoted to the bracketing entropy control of Gaussian density families in the diagonal and the general cases respectively. Note that in order to apply Theorem 1, the local bracketing entropy $\mathcal{H}_{[.]}(x, \mathcal{S}_{(K,\alpha)}(u,\xi), d_H)$ has to be controlled. Nevertheless, it is difficult to characterize the subset $\mathcal{S}_{(K,\alpha)}(u,\xi)$ in

function of the parameters of its mixtures. Therefore a global study of the entropy bracketing is proposed in the theorem proof since $\mathcal{H}_{[.]}(x, \mathcal{S}_{(K,\alpha)}(u,\xi), d_H) \leq \mathcal{H}_{[.]}(x, \mathcal{S}_{(K,\alpha)}, d_H)$.

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 two collections since the model dimensions are different. For the diagonal mixture family, the risk bound is more accurate when this family is the diagonal collection than if it is considered as a subset of the general collection. Second, the constant A is made explicit in the theorem proof (see Section B.1) and its expression is different between the diagonal case and the general case (see Equations (22) and (23)). In the both cases, it depends on parameters $\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}, a$ and Q with $A = O(\sqrt{\ln Q})$ as Q goes to infinity. This number of variables Q has to have a reasonable order in the constant 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. Thirds, since the multiplicative constants are not explicit, a practical method is necessary. This will be addressed in the discussion section.

3.2 Non-ordered variable case

Theorem 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 $(\tau(x)_1, \ldots, \tau(x)_\alpha)' = x_{[\mathbf{v}]}$. Moreover, the dimension $D(K,\mathbf{v})$ of $\mathcal{S}_{(K,\mathbf{v})}$ is equal to $D(K,\alpha)$. 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 in penalty function:

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

Consequently, in the following theorem which is the analog of Theorem 2 for the non-ordered case, the associated penalty functions have an additional logarithm term depending on the dimension.

Theorem 3. *For both diagonal and general mixture collections, there exists two absolute constants κ and C such that, if*

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

where A is the same constant as 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 Section B.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 this result is difficult to apply when the number of variables becomes too large since an exhaustive research of the best model is then untractable.

4 Discussion

In this paper, specific Gaussian mixtures are considered to take into account the role of variables in the clustering process. Main results are stated for diagonal and general Gaussian mixture forms for ordered and non-ordered variables. A data-driven 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 in which loglikelihoods are difficult to control. Proofs of these results require to control the bracketing entropy of multidimensional Gaussian density families and to determinate weights taking into account the richness of the model collection. 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 as a selection problem of the number of mixture components and besides, of the mixture shape among a mixture shape collection. A complete collection of twenty eight parsimonious models is available, used for instance in MIXMOD software (Biernacki et al., 2006). These models are obtained by imposing conditions on the proportions and the elements of variance matrix eigenvalue decomposition (see Banfield and Raftery, 1993; Celeux and Govaert, 1995). Commonly, an asymptotic criterion as BIC (Schwarz, 1978) or ICL (Biernacki et al., 2000) is used to solve this model selection problem. In this paper, our main results allows to propose a non asymptotic criterion to select the number of clusters (the subset $\mathbf{v}$ is fixed to the complete variable set). Moreover, we focus on two 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 to select the mixture form. A comparison of our criterion with BIC, ICL and AIC is proposed in our framework in Maugis and Michel (2008) and also for the selection of the mixture component number in Baudry (2007).

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 ensure that the number of variables Q can be large. Birgé and Massart (2006) propose their so-called "slope

heuristics" (see also Massart, 2007, Section 8.5) 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 in Birgé and Massart (2006) in the framework of Gaussian regression in a homoscedastic fixed design and generalized by Arlot and Massart (2008) 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 by Lebarbier (2005), for clustering (Baudry, 2007), for Gaussian Markov random fields (Verzelen, 2008), for estimation of oil reserves (Lepez, 2002) and genomics (Villers, 2007). In the context of Gaussian mixtures, this heuristics is carried out on simulated and real datasets in Maugis and Michel (2008). It is shown that the developed procedure allows to obtain efficient clusterings and variable selection, for instance in the curve clustering context.

Acknowledgements

The authors are grateful to Sylvain Arlot (Université Paris-Sud 11) and Pascal Massart (Université Paris-Sud 11) for helpful discussions and valuable comments.

Appendices

A Tools: bound on bracketing entropies

A.1 The bracketing entropy of mixture density family

In order to use Theorem 1, it is necessary to control the bracketing entropy of Gaussian mixture families to define a suitable function $\psi_{(K,\alpha)}$ fulfilling properties **[I]**. Ghosal and van der Vaart (2001) and Genovese and Wasserman (2000) 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 Ghosal and van der Vaart (2001) to control the bracketing entropy of our multidimensional Gaussian mixture models. But the control obtained this way has a too large dependency in Q. We propose instead a method inspired by the work of Genovese and Wasserman (2000). They state the next theorem allowing to bound the bracketing number of a mixture set according to the product of the bracketing numbers of the associated mixture component families. For all k in $\{1,\dots,K\}$, 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

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

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

$$\mathcal{T}_{K-1} := \Big\{ \mathbf{p} = (p_1,\dots,p_K),\ \forall k = 1,\dots,K,\ p_k \geq 0,\ \sum_{k=1}^{K} p_k = 1 \Big\}.$$

Theorem 4. *With the previous notation, for all K and all $\varepsilon \in (0,1]$,*

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{W}_K, d_H) \leq K(2\pi\, e)^{\frac{K}{2}} \left(\frac{3}{\varepsilon}\right)^{K-1} \prod_{k=1}^{K} \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{C}_k, d_H\right) .$$

Here we want to take into account the specific form of the multidimensional mixtures studied in this paper. A new result is deduced from Theorem 4 for the general and diagonal cases. For all $f \in \mathcal{S}_{(K,\mathbf{v})}$ and $x \in \mathbb{R}^Q$, Equation (2) gives that

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

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 Section 2.1). In order to bound the bracketing entropy of the mixture family $\mathcal{S}_{(K,\mathbf{v})}$ by a sum with the bracketing entropies of the simplex, of $\mathcal{G}_{(\alpha)}$ and of $\mathcal{F}_{(\alpha)}$, the following proposition is stated.

Proposition 1. *For the diagonal and general cases, for all $\varepsilon \in (0,1]$, the bracketing number of the density family $\mathcal{S}_{(K,\mathbf{v})}$ is bounded by*

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

Proposition 1 allows to deduce that

$$\mathcal{H}_{[.]}(\varepsilon,\mathcal{S}_{(K,\,\mathbf{v})},d_H) \leq C(K)+(K-1)\ln\left(\frac{1}{\varepsilon}\right)+\mathcal{H}_{[.]}\left(\frac{\varepsilon}{3},\mathcal{G}_{(\alpha)},d_H\right)+K\,\mathcal{H}_{[.]}\left(\frac{\varepsilon}{9},\mathcal{F}_{(\alpha)},d_H\right) \quad (7)$$

with $C(K) = \ln(K) + \frac{K}{2}\ln(2\pi e) + (K-1)\ln(9)$. Therefore, Inequality (7) recasts the bracketing entropy control problem of $\mathcal{S}_{(K,\mathbf{v})}$ in the bracketing entropy control of the two families $\mathcal{F}_{(\alpha)}$ and $\mathcal{G}_{(\alpha)}$. Bracketing entropy upper bounds of these two sets are assessed for two mixture forms in Section A.2 and Section A.3.

Proof. According to Theorem 4, for all $\delta \leq 1$,

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

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

$$\mathcal{N}_{[.]}(\varepsilon,\mathcal{S}_{(K,\mathbf{v})},d_H) \leq \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3},\mathcal{G}_{(\alpha)},d_H\right) \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3},\mathcal{L}_{(K,\alpha)},d_H\right) \quad (8)$$

then we obtain the result

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

Thus, it remains to check Inequality (8). It is done by the following adaptation of a result proof given by Genovese and Wasserman (2000).

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

$$L(x) = l(x_{[\mathbf{v}]})\tilde{l}(x_{[\mathbf{v}^c]}) \text{ and } U(x) = u(x_{[\mathbf{v}]})\tilde{u}(x_{[\mathbf{v}^c]}) \quad (9)$$

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 Genovese and Wasserman (2000) gives that

$$\begin{cases} \int u(x_{[\mathbf{v}]})\mathrm{d}x_{[\mathbf{v}]} \leq 1+3\delta \\ \int \tilde{u}(x_{[\mathbf{v}^c]})\mathrm{d}x_{[\mathbf{v}^c]} \leq 1+3\delta. \end{cases} \quad (10)$$

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

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

According Cauchy-Schwarz inequality and (10),

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

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 (9), the number of brackets for $\mathcal{S}_{(K,\mathbf{v})}$ is upper bounded by $\mathcal{N}_{[.]}(\varepsilon, \mathcal{S}_{(K,\mathbf{v})}, d_H) \leq \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{G}_{(\alpha)}, d_H\right) \times \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{L}_{(K,\alpha)}, d_H\right)$. □

A.2 The bracketing entropy of Gaussian mixture family with diagonal variance matrices

In this section, we consider the case where variance matrices of the Gaussian densities in the mixtures are diagonal. The following proposition gives an upper bound of the bracketing entropy of the two families $\mathcal{F}_{(\alpha)}$ and $\mathcal{G}_{(\alpha)}$ defined by (3) and (1) respectively. It allows us to deduce an upper bound of the bracketing entropy of $\mathcal{S}_{(K,\mathbf{v})}$ according to Inequality (7).

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

$$\mathcal{H}_{[.]}(\varepsilon, \mathcal{F}_{(\alpha)}, d_H) \leq \alpha \ln\left(2\,a\sqrt{\frac{2}{c_1\,\lambda_m}}\right) + \alpha \ln\left(8\frac{\lambda_M}{\lambda_m}\right) + 2\alpha \ln(\sqrt{2}\,Q) + 2\alpha \ln\left(\frac{1}{\varepsilon}\right) \tag{11}$$

and

$$\mathcal{H}_{[.]}(\varepsilon, \mathcal{G}_{(\alpha)}, d_H) \leq \ln\left(8\frac{\lambda_M}{\lambda_m}\right) + \ln\left(\sqrt{2}\,Q\right) + \ln\left(\frac{1}{\varepsilon}\right). \tag{12}$$

Thus,

$$\begin{aligned}\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\alpha)}, d_H) \quad \leq \quad & C(K) + \alpha(2K\alpha+1)\ln(9\sqrt{2}\,Q) + (K\alpha+1)\ln\left(8\frac{\lambda_M}{\lambda_m}\right) \\ & + K\alpha \ln\left(a\sqrt{\frac{8}{c_1\lambda_m}}\right) + D(K,\alpha)\ln\left(\frac{1}{\varepsilon}\right)\end{aligned} \tag{13}$$

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

Proof. According to Assertion (7), the upper bound on the bracketing entropy of $\mathcal{S}_{(K,\alpha)}$, given by Inequality (13), is deduced from upper bounds on the bracketing entropy of $\mathcal{F}_{(\alpha)}$ and $\mathcal{G}_{(\alpha)}$, respectively expressed in Inequalities (11) and (12). These two inequalities are now proved successively.

The proof of Inequality (11) is adapted from Genovese and Wasserman (2000) who prove similar results for unidimensional Gaussian mixture families. The main idea is to define a lattice over the parameter space $\mathcal{B} = \{(\mu, \sigma_1^2, \ldots, \sigma_\alpha^2) \in [-a,a]^\alpha \times [\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]^\alpha\}$ 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 $j \in \{2, \ldots, r\}$, set

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

with $r = \left\lceil 2\,\frac{\ln\left\{\frac{\lambda_{\mathrm{M}}(1+\delta)}{\lambda_{\mathrm{m}}}\right\}}{\ln(1+\delta)} \right\rceil$ 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 $J = (j(1), \ldots, j(\alpha)) \in \{2, \ldots, r\}^\alpha$, a diagonal matrix B_J is defined by

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

We also consider vectors

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

such that

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

where $s_q \in \mathbb{Z} \cap [-A, A]$ with $A = \left\lfloor \frac{a\,\delta^{-1}\,(1+\delta)^{-\frac{1-j(q)}{4}}}{\sqrt{c_1\,\lambda_{\mathrm{M}}}} \right\rfloor$. 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:

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

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

$$(\mu-\nu_J)' B^{-1}_{J+1}(\mu-\nu_J) \leq c_1\,\alpha\,\delta^2 \tag{14}$$

where $J+1 := (j(1)+1,\ldots,j(\alpha)+1)$. Then we check that the bracket $[l,u]$ contains f. Inequality (14) implies that

$$(\mu-\nu_J)' B^{-1}_{J}(\mu-\nu_J) \leq \frac{\alpha}{4}\delta^2. \tag{15}$$

The use of Corollary 2, which allows to bound the ratio of two Gaussian densities with diagonal variance matrices, together with (15) leads to

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

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 2 and (14), this shows that $l \leq f$ since

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

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 3,

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

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

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

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

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

that implies Inequality (11).

Using a similar proof, the upper bound of the bracketing entropy of $\mathcal{G}_{(\alpha)}$ given by Inequality (12) is obtained. To check this result, the variance family

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

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.

□

A.3 The bracketing entropy of Gaussian mixture family with general variance matrices

We now consider the case of general Gaussian mixture collection. The following proposition gives an upper bound of the bracketing entropy of the Gaussian density family $\mathcal{F}_{(\alpha)}$ defined by (4).

Proposition 3. *For all* $\varepsilon \in (0,1]$,

$$\begin{aligned}\mathcal{H}_{[.]}\left(\varepsilon, \mathcal{F}_{(\alpha)}, d_H\right) &\leq \frac{\alpha(\alpha+1)}{2}\ln\left(\frac{6\sqrt{3}\,\lambda_{_M}}{\lambda_{_m}}\right) + \alpha\ln\left(\frac{6\,a}{\sqrt{\lambda_{_m}}}\right) \\ &+ \left\{\frac{\alpha(\alpha+1)}{2}+\alpha\right\}\ln(Q^2) + \left\{\frac{\alpha(\alpha+1)}{2}+\alpha\right\}\ln\left(\frac{1}{\varepsilon}\right).\end{aligned}$$

This result together with Inequality (7) and the upper bound of $\mathcal{G}_{(\alpha)}$ (see Inequality (12)) gives the following corollary.

Corollary 1. *For all* $\varepsilon \in (0,1]$,

$$\begin{aligned}\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,\alpha)}, d_H) &\leq C(K) + \ln\left(\frac{8\sqrt{2}\ \lambda_{_M}}{\lambda_{_m}}\right) + K\frac{\alpha(\alpha+1)}{2}\ln\left(\frac{54\sqrt{3}\,\lambda_{_M}}{\lambda_{_m}}\right) + K\alpha\ln\left(\frac{54\,a}{\sqrt{\lambda_{_m}}}\right) \\ &+ V(K,\alpha)\ln(Q^2) + D(K,\alpha)\ln\left(\frac{1}{\varepsilon}\right)\end{aligned}$$

where $C(K) = \ln(K) + \frac{K}{2}\ln(2\pi e) + (K-1)\ln(9)$ *and* $V(K,\alpha) = K\,\frac{\alpha(\alpha+1)}{2} + K\alpha + 1$.

To prove Proposition 3, the method used in the diagonal case 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 in the diagonal case. 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 $\mathcal{D}^+_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}})$ with respect to the uniform norm, and then of using it to construct a bracket covering of $\mathcal{F}_{(\alpha)}$. The following notation is used for matrix norms: $\|B\|_\infty = \max_{1\leq i,j\leq\alpha}|B_{ij}|$ and $|||B||| = \sup_{\|x\|_2=1}|x'Bx| = \sup_{\lambda\in \mathrm{vp}(B)}|\lambda|$ 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 $\mathcal{D}^+_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}})$ for the uniform norm $\|.\|_\infty$, composed of symmetric matrices and defined by

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

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

$$\|A - \Sigma\|_\infty \leq \beta. \tag{16}$$

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

Lemma 1. *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 Serre (2002, Theorem 3.3.2 p49),

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

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 $\Sigma \in \mathcal{D}^{+}_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}})$. For $\beta > 0$, there exists a matrix $A \in \mathcal{R}(\beta)$ such that $\|A - \Sigma\|_\infty \leq \beta$ according to (16). Then the two following functions are considered

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

and

$$l(x) = (1 + 2\delta)^{-\alpha} \, \Phi\left(x|\,\nu, (1+\delta)^{-1}A\right) \tag{18}$$

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 6. The resulting bounds on Gaussian density ratios are given in Lemma 3.

Lemma 2. *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 $\mathbb{R}^\alpha$,*

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

and

$$x'\{\Sigma - (1+\delta)^{-1}A\}x \geq \beta\alpha\|x\|_2^2. \tag{20}$$

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

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

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

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

□

Lemma 3. *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 6, since $(1+\delta)A - \Sigma$ is a positive definite matrix from Lemma 2,

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

Inequality (19) 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 1,

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

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 6, (20) and Lemma 1, 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 finishes the construction of an ε-bracket covering of $\mathcal{F}_{(\alpha)}$.

Proposition 4. *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 $\Sigma\in\mathcal{D}^{+}_{(\alpha)}(\lambda_{\mathrm{m}},\lambda_{\mathrm{M}})$. There exists A in $\mathcal{R}(\beta)$ such that $\|\Sigma-A\|_{\infty}\leq\beta$ and ν in $\mathcal{X}(\varepsilon,a,\lambda_{\mathrm{m}},\alpha)$ satisfying, for all q in $\{1,\ldots,\alpha\}$, $|\mu_q-\nu_q|\leq\sqrt{\lambda_{\mathrm{m}}}\varepsilon/(3\alpha)$. Consider the two associated functions l and u defined in (17) and (18) respectively. Since $\|\mu-\nu\|_2^2\leq\lambda_{\mathrm{m}}\,\varepsilon^2/(9\alpha)$, using Lemma 3,

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

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.7,

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

□

Proof of Proposition 3

Proof. Since the set of ε-brackets over $\mathcal{F}_{(\alpha)}$, described in the Proposition 4 is totally defined by the parameter spaces $\mathcal{R}(\beta)$ and $\mathcal{X}(\varepsilon,a,\lambda_{\mathrm{m}},\alpha)$, an upper bound of the bracketing number of $\mathcal{F}_{(\alpha)}$ is deduced from an upper bound of the two set cardinals.

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

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

$$\begin{aligned}\mathcal{H}_{[.]}\left(\varepsilon,\mathcal{F}_{(\alpha)},d_H\right) &\leq \frac{\alpha(\alpha+1)}{2}\ln\left(\frac{6\sqrt{3}\,\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)+\alpha\ln\left(\frac{6\,a}{\sqrt{\lambda_{\mathrm{m}}}}\right)\\ &+ \left\{\frac{\alpha(\alpha+1)}{2}+\alpha\right\}\ln(Q^2)+\left\{\frac{\alpha(\alpha+1)}{2}+\alpha\right\}\ln\left(\frac{1}{\varepsilon}\right).\end{aligned}$$

□

B Proofs of the main results

B.1 Proof of Theorem 2

Theorem 1 is applied in order to prove Theorem 2. It requires to find a convenient function $\psi_{(K,\alpha)}$ which is deduced from the bracketing entropy upper bounds established in Section A. We state the following technical result which is used hereafter: For all $\varepsilon \in (0,1]$,

$$\int_0^{\varepsilon} \sqrt{\ln\left(\frac{1}{x}\right)}\,\mathrm{d}x \leq \varepsilon\left\{\sqrt{\ln\left(\frac{1}{\varepsilon}\right)} + \sqrt{\pi}\right\}. \tag{21}$$

This inequality is deduced from an integration by part and the following concentration inequality (Massart, 2007, p.19): If Z is a centered standard Gaussian variable then $P(Z \geq c) \leq e^{-\frac{c^2}{2}}$ for all $c > 0$.

First, we consider the diagonal mixture form with ordered variables. For all positive real number ξ, using (21),

$$\begin{aligned}
\int_0^{\xi}\sqrt{\mathcal{H}_{[.]}(x,\mathcal{S}_{(K,\alpha)},d_H)}\,\mathrm{d}x &\leq \xi\left\{\sqrt{C(K)} + \sqrt{K\alpha\ln\left(a\sqrt{\frac{8}{c_1\lambda_{\mathrm{m}}}}\right)} + \sqrt{(K\alpha+1)\ln\left(8\frac{\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)}\right\} \\
&\quad + \xi\sqrt{(2K\alpha+1)\ln(9\sqrt{2}\,Q)} + \int_0^{\xi\wedge 1}\sqrt{D(K,\alpha)\ln\left(\frac{1}{x}\right)}\,\mathrm{d}x \\
&\leq \xi\left\{\sqrt{C(K)} + \sqrt{K\alpha\ln\left(a\sqrt{\frac{8}{c_1\lambda_{\mathrm{m}}}}\right)} + \sqrt{(K\alpha+1)\ln\left(8\frac{\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)}\right\} \\
&\quad + \xi\left\{\sqrt{(2K\alpha+1)\ln(9\sqrt{2}\,Q)}\right\} \\
&\quad + \xi\sqrt{D(K,\alpha)}\left\{\sqrt{\ln\left(\frac{1}{1\wedge\xi}\right)} + \sqrt{\pi}\right\}.
\end{aligned}$$

According to Inequality (21) and Proposition 2, we get

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

with

$$\begin{aligned}
(\square) &= \sqrt{\frac{C(K)}{D(K,\alpha)}} + \sqrt{\frac{K\alpha}{D(K,\alpha)}\ln\left(a\sqrt{\frac{8}{c_1\lambda_{\mathrm{m}}}}\right)} + \sqrt{\frac{(K\alpha+1)}{D(K,\alpha)}\ln\left(8\frac{\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)} \\
&\quad + \sqrt{\frac{(2K\alpha+1)}{D(K,\alpha)}\ln(9\sqrt{2}\,Q)} + \sqrt{\pi}.
\end{aligned}$$

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, $(\square)$ is bounded by a constant $\mathcal{A}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}, a, Q)$ denoted only $\mathcal{A}$ hereafter and defined by

$$\mathcal{A}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}, a, Q) := \sqrt{\pi} + \sqrt{\ln\left(18\pi e^2\right)} + \sqrt{\ln\left(a\sqrt{\frac{8}{c_1\lambda_{\mathrm{m}}}}\right)} + \sqrt{\ln\left(8\frac{\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)} + \sqrt{\ln(9\sqrt{2}\,Q)}. \quad (22)$$

In the same way, we obtain in the general case that for all $\xi > 0$

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

where the constant is equal to

$$\mathcal{A} = \sqrt{\ln(18\pi e^2)} + \sqrt{\ln\left(\frac{8\sqrt{2}\ \lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)} + \sqrt{\ln\left(\frac{54\sqrt{3}\,\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}\right)} + \sqrt{\ln\left(\frac{54\,a}{\sqrt{\lambda_{\mathrm{m}}}}\right)} + \sqrt{\ln(Q^2)} + \sqrt{\pi}. \quad (23)$$

Consequently in the two cases 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 condition **[I]** of Theorem 1 can be considered. Next 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 solve

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

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 so

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

Finally, according to the low bound of penalty functions in Theorem 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

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

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

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

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

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

B.2 Proof of Theorem 3

Apart from the weight definition step, the proof of Theorem 3 is the same as in the ordered case. The following Lemma is used to define weights for this family which has a greater richness. Recall that $D(K,\mathbf{v})$ denotes the dimension of $\mathcal{S}_{(S,\mathbf{v})}$ which is equal to $K\{2\,\mathrm{card}(\mathbf{v})+1\}$ and $K\left\{1+\mathrm{card}(\mathbf{v})+\frac{\mathrm{card}(\mathbf{v})(\mathrm{card}(\mathbf{v})+1)}{2}\right\}$ in diagonal and general cases respectively.

Lemma 4. *The quantity* $\mathrm{card}\,\{(K,\mathbf{v}) \in \mathbb{N}^\star \times \mathcal{V};\ D(K,\mathbf{v}) = D\}$ *is upper bounded by*

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

Proof. 1. In the diagonal case,

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

If $Q \leq \lfloor \frac{D-1}{2} \rfloor$, $\sum_{\alpha=1}^{\infty} \binom{Q}{\alpha}\, \mathbb{1}_{\alpha \leq Q \wedge \lfloor \frac{D-1}{2} \rfloor} = 2^Q$. Otherwise, according to Proposition 2.5 in Massart (2007),

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

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

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

2. In the general case, $\text{card}\left\{(K,\mathbf{v}) \in \mathbb{N}^\star \times \mathcal{V};\ K\left[1+\text{card}(\mathbf{v})+\frac{\text{card}(\mathbf{v})\{\text{card}(\mathbf{v})+1\}}{2}\right]=D\right\}$ 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 the diagonal case.

□

Proposition 5. *Consider the following weight family* $\left\{\rho_{(K,\mathbf{v})}\right\}_{(K,\mathbf{v})\in\mathbb{N}^\star\times\mathcal{V}}$ *defined by*

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

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

Proof. According to Lemma 4,

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

In the first sum,

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

since $D\leq 2Q$. In the second sum, since $D\geq 2Q+1$,

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

Then

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

□

C Results for multivariate Gaussian densities

C.1 Ratio of two Gaussian densities

Proposition 6. *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 $x\in\mathbb{R}^Q$,*

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

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

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

The matrix $\Sigma_1^{-1}-\Sigma_2^{-1}=\Sigma_1^{-1}(\Sigma_2-\Sigma_1)\Sigma_2^{-1}$ is nonsingular since Σ_1, Σ_2 and $\Sigma_2-\Sigma_1$ are positive definite matrices. Defining $\mu^\star=(\Sigma_1^{-1}-\Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\mu_1-\Sigma_2^{-1}\mu_2)$,

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

Finally,

$$\begin{aligned}\frac{\Phi(x|\mu_1,\Sigma_1)}{\Phi(x|\mu_2,\Sigma_2)} &= \sqrt{\frac{|\Sigma_2|}{|\Sigma_1|}}\,\exp\left[-\frac{1}{2}\left\{(x-\mu^\star)'(\Sigma_1^{-1}-\Sigma_2^{-1})(x-\mu^\star)+(\mu_1-\mu_2)'(\Sigma_2-\Sigma_1)^{-1}(\mu_1-\mu_2)\right\}\right]\\ &\leq \sqrt{\frac{|\Sigma_2|}{|\Sigma_1|}}\,\exp\left\{\frac{1}{2}(\mu_1-\mu_2)'(\Sigma_2-\Sigma_1)^{-1}(\mu_1-\mu_2)\right\}.\end{aligned}$$

□

Corollary 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 following 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*

$$\frac{\Phi(x\,|\,\mu_1,B_1)}{\Phi(x\,|\,\mu_2,B_2)} \leq \left(\prod_{q=1}^{Q}\frac{b_{2q}}{b_{1q}}\right)\exp\left\{\frac{1}{2}(\mu_1-\mu_2)'\operatorname{diag}\left(\frac{1}{b_{21}^2-b_{11}^2},\ldots,\frac{1}{b_{2Q}^2-b_{1Q}^2}\right)(\mu_1-\mu_2)\right\}.$$

C.2 Hellinger distance between two Gaussian densities

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

Proposition 7. *Let $\Phi(.|\mu_1,\Sigma_1)$ and $\Phi(.|\mu_2,\Sigma_2)$ be two Gaussian densities. The Hellinger distance between these two densities $d_H^2(\Phi(.|\mu_1,\Sigma_1),\Phi(.|\mu_2,\Sigma_2))$ 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 distance is equal to

$$\begin{aligned}
d_H^2(\Phi(.|\mu_1,\Sigma_1),\Phi(.|\mu_2,\Sigma_2)) &= 2-2(2\pi)^{-\frac{Q}{2}}|\Sigma_1\Sigma_2|^{-\frac{1}{4}}\,\exp\left\{-\frac{1}{4}(\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)\right\}\\
&\quad\times\int\exp\left\{-\frac{1}{4}(x-\mu^\star)'(\Sigma_1^{-1}+\Sigma_2^{-1})(x-\mu^\star)\right\}\mathrm{d}x\\
&= 2-2(2\pi)^{-\frac{Q}{2}}|\Sigma_1\Sigma_2|^{-\frac{1}{4}}\,\exp\left\{-\frac{1}{4}(\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)\right\}\\
&\quad\times(4\pi)^{\frac{Q}{2}}|\Sigma_1^{-1}+\Sigma_2^{-1}|^{-\frac{1}{2}}
\end{aligned}$$

that entails the concluding result. □

Corollary 3. *Using the notation of the Corollary 2, the Hellinger distance of two Gaussian densities with diagonal variance matrices is given by the following expression*

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

References

Akaike, H. (1973). In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest.

Arlot, S. and Massart, P. (2008). Slope heuristics for heteroscedastic regression on a random design. Submitted to the Annals of Statistics.

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

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

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

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

Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics and Data Analysis*, 51(2):587–600.

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

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

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

Birgé, L. and Massart, P. (2006). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1-2):33–73.

Bouveyron, C., Girard, S., and Schmid, C. (2007). High-dimensional data clustering. *Computational Statistics and Data Analysis*, 52(1):502–519.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference*. Springer-Verlag, New York, second edition. A practical information-theoretic approach.

Castellan, G. (1999). Modified Akaike's criterion for histogram density estimation. Technical report, University Paris-Sud 11.

Castellan, G. (2003). Density estimation via exponential model selection. *IEEE Trans. Inform. Theory*, 49(8):2052–2060.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological*, 39(1):1–38.

Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *The Annals of Statistics*, 28(4):1105–1127.

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5):1233–1263.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā. The Indian Journal of Statistics. Series A*, 62(1):49–66.

Law, M. H., Jain, A. K., and Figueiredo, M. A. T. (2004). Simultaneous Feature Selection and Clustering Using Mixture Models. *IEEE*, 26:1154–1166.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal Processing*, 85:717–736.

Lepez, V. (2002). *Potentiel de réserves d'un bassin pétrolier : modélisation et estimation*. PhD thesis, University Paris-Sud 11.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2007). Variable selection for clustering with Gaussian mixture models. Technical Report 6211, INRIA.

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

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Serre, D. (2002). *Matrices*, volume 216 of *Graduate Texts in Mathematics*. Springer-Verlag, New York. Theory and applications.

Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Institut des Hautes Études Scientifiques. Publications Mathématiques*, (81):73–205.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563.

Verzelen, N. (2008). Adaptive estimation of regular Gaussian Markov random fields. In preparation.

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

Unité de recherche INRIA Futurs
Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399