A sparse variable selection procedure in model-based clustering

Caroline Meynet, Cathy Maugis-Rabusseau

A sparse variable selection procedure in model-based clustering

Caroline Meynet, Cathy Maugis-Rabusseau

RESEARCH REPORT
N° ????
September 2012
Project-Team Select

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

A sparse variable selection procedure in model-based clustering

Caroline Meynet*[†], Cathy Maugis-Rabusseau[‡]

Project-Team Select

Research Report n° ???? — September 2012 — 36 pages

Abstract: Owing to the increase of high-dimensional datasets, the variable selection for clustering is an important challenge. In the context of Gaussian mixture clustering, we recast the variable selection problem into a general model selection problem. Our procedure first consists of using a ℓ_1-regularization method to build a data-driven model subcollection. Second, the maximum loglikelihood estimators (MLEs) are obtained using the EM algorithm. Next a non asymptotic penalized criterion is proposed to select the number of mixture components and the relevant clustering variables simultaneously. A general model selection theorem for MLEs with a random model collection is established. It allows one to derive the penalty shape of the criterion, which depends on the complexity of the random model collection. In practice, the criterion is calibrated using the so-called slope heuristics. The resulting procedure is illustrated on two simulated examples. Finally, an extension to a more general modeling of irrelevant clustering variables is presented.

Key-words: Variable selection, Gaussian mixture clustering, Non asymptotic penalized criterion, ℓ_1-regularization method, Maximum likelihood estimator.

* Université Paris-Sud 11, Orsay, France

[†] INRIA Saclay - Île-de-France, Orsay, France

[‡] Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse

RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

Parc Orsay Université
4 rue Jacques Monod
91893 Orsay Cedex

Une procédure parcimonieuse de sélection de variables pour la classification par mélanges gaussiens

Résumé : Au vu de l'augmentation du nombre de jeux de données de grande dimension, la sélection de variables pour la classification non supervisée est un enjeu important. Dans le cadre de la classification par mélanges gaussiens, nous reformulons le problème de sélection de variables en un problème général de sélection de modèle. Dans un premier temps, notre procédure consiste à construire une sous-collection de modèles grâce à une méthode de régularisation ℓ_1. Puis, l'estimateur du maximum de vraisemblance est déterminé via un algorithme EM pour chaque modèle. Enfin un critère pénalisé non asymptotique est proposé pour sélectionner à la fois le nombre de composants du mélange et l'ensemble des variables informatives pour la classification. D'un point de vue théorique, un théorème général de sélection de modèles dans le cadre de l'estimation par maximum de vraisemblance avec une collection aléatoire de modèles est établi. Il permet en particulier de justifier la forme de la pénalité de notre critère, forme qui dépend de la complexité de la collection de modèles. En pratique, ce critère est calibré grâce à la méthode dite de l'heuristique de pente. Cette procédure est illustrée sur deux jeux de données simulées. Finalement, une extension, associée à une modélisation plus générale des variables non informatives pour la classification, est proposée.

Mots-clés : Sélection de variables, Classification par mélanges gaussiens, Critère pénalisé non asymptotique, Méthode de régularisation ℓ_1, Estimateur du maximum de vraisemblance.

1 Introduction

The goal of clustering methods is to discover clusters among n individuals described by p variables. Many clustering methods exist and roughly fall into two categories. The first one is distance-based clustering methods, including hierarchical clusterings and K-means type algorithms. The second category is model-based clustering methods: each cluster is represented by a parametric distribution, the entire dataset is modeled by a mixture of these distributions, and a criterion is used to optimize the fit between the data and the model. An advantage of model-based clustering is to provide a rigorous statistical framework to assess the number of clusters and the role of each variable in the clustering process. In this paper, we focus on clustering with Gaussian mixtures.

Variable selection for clustering is an important challenge, motivated by the increasing study of high-dimensional datasets. Since the structure of interest may often be contained into a subset of the available variables and many attributes may be useless or even harmful to detect a reasonable clustering structure, it is important to select the relevant clustering variables. In addition, removing the irrelevant variables enables to get simpler modeling and can largely enhance interpretability. Usually, two types of variable selection approaches are envisaged. On the one hand, the "filter" approaches select the variables before the cluster analysis (see for instance Dash *et al.*, 2002; Jouve and Nicoloyannis, 2005). Their main weakness is the independence between the variable selection step and the clustering procedure. In contrast, the "wrapper" approaches combine variable selection and clustering. For distance-based clustering, one can cite the works of Fowlkes *et al.* (1988), Devaney and Ram (1997) and Brusco and Cradit (2001) for instance. Wrapper methods are also developed in the model-based clustering framework. For instance, Maugis *et al.* (2009b), which is an extension of Raftery and Dean (2006) and Maugis *et al.* (2009a), propose a general variable role modeling (relevant, redundant and irrelevant variables for clustering) and a backward stepwise algorithm is developed. In the same low dimensional context, Maugis and Michel (2011b,a) recast also the variable selection problem for clustering into a general model selection problem. For each model, the maximum likelihood estimator is considered and a data-driven penalized criterion is built for the model selection. Nevertheless, their procedure, called *MM-MLE procedure* in this paper, requires to consider the complete variable subset collection (or to preliminary order the variables). In the high dimensional context, some Bayesian methods have been developed as Tadesse *et al.* (2005) and Kim *et al.* (2006). Rather than considering a Bayesian approach, Pan and Shen (2007) propose to take advantage of the sparsity property of ℓ_1-penalization to perform automatic variable selection for high-dimensional data clustering. Their procedure, called *PS-Lasso procedure* in this paper, consists of using a Lasso method to select relevant clustering variables and estimate mixture parameters in the same exercise. Nevertheless, ℓ_1-penalization induces shrinkage of the coefficients and thus biased estimators with high estimation risk. Moreover, they use a BIC-type criterion for the model selection which can be unsuitable for high-dimensional data.

In this paper, a global model selection procedure, called *Lasso-MLE procedure*, is proposed to simultaneously choose the number of clusters and the relevant clustering variables. This procedure is aiming to take advantage of MM-MLE and PS-Lasso procedures in low- and high-dimensional cases. Following Pan and Shen (2007), an ℓ_1-penalized likelihood approach is considered to determine potential sets of relevant variables. This allows one to efficiently construct a data-driven model subcollection with reasonable complexity, even for high-dimensional situations. Contrary to Pan and Shen's approach, the evaluation of the maximum likelihood estimator (MLE) rather than the ℓ_1-penalized maximum likelihood estimator for each model is considered to avoid estimation problems due to ℓ_1-penalization shrinkage. Next, a non asymptotic penalized criterion is proposed to solve the model selection problem. Considering MLE and a random model collection require to extend the general model selection theorem of Massart (2007, Theorem 7.11) to our context. This extension allows one to justify the penalty shape of our criterion. In practice, the penalty depending on unknown constant(s) is calibrated using the so-called slope heuristics (Birgé and Massart, 2006; Baudry *et al.*, 2011). To our knowledge, the proposed Lasso-MLE estimators have never been studied in model-based clustering. Yet, this idea has emerged in other frameworks. In regression, Connault (2011) proposed such a procedure, which seems to have better performances than for the classical Lasso method. Such an estimator is also mentioned as LARS-OLS hybrid in Efron *et al.* (2004, p. 421). In the density estimation framework, Bertin *et al.* (2011) consider such an idea to estimate densities decomposed on some dictionary. Note that in this paper, only finite Gaussian mixtures with common spherical covariance matrix are considered in order to simplify the reading. This assumption allows one to reduce the relevance of variables for the clustering on the mean component vectors. The extension of the Lasso-MLE procedure to general Gaussian mixtures is discussed in Section 7.

The paper is organized as follows. Clustering with Gaussian mixtures is recalled in Section 2.1 and the model collection is described in Section 2.2. Section 3 is devoted to the description of MM-MLE, PS-Lasso and Lasso-MLE procedures. A theoretical result for the Lasso-MLE estimator is stated in Section 4.1 and the penalty calibration for the data-driven penalized criterion is explained in Section 4.2. The three procedures are compared on two simulated examples in Section 5. Section 6 is devoted to the extension of the Lasso-MLE procedure in a curve clustering context. The paper is concluded with a brief discussion in Section 7. Technical aspects are given in Appendix.

2 Gaussian mixture models

2.1 Clustering with Gaussian mixtures

Consider a dataset $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ with $\boldsymbol{Y_i} \in \mathbb{R}^p$. These data come from a probability distribution with unknown density s. This target s is estimated by a finite mixture model in a clustering purpose although s is not assumed to be a Gaussian mixture density itself. Model-based clustering consists of assuming that the data come from several subpopulations and that the overall population is a mixture of these subpopulations. In this paper, each mixture component is modeled by a Gaussian density. Thus the distribution s is modeled by a finite Gaussian mixture with K components:

$$\boldsymbol{y} \in \mathbb{R}^p \mapsto \sum_{k=1}^{K} \pi_k \, \Phi(\boldsymbol{y} \mid \boldsymbol{\mu_k}, \Sigma_k).$$

The mixing proportions $(\pi_1, \ldots, \pi_K)$ belong to

$$\Pi_K = \left\{ (a_1, \ldots, a_K) \in (0,1)^K ; \sum_{k=1}^{K} a_k = 1 \right\}.$$

The density $\Phi(\cdot \mid \boldsymbol{\mu_k}, \Sigma_k)$ is the p-dimensional Gaussian density with mean $\boldsymbol{\mu_k}$ and covariance matrix Σ_k. After estimating the parameter vector, the data clustering can be obtained using the Maximum A Posteriori (MAP) principle: i is assigned to Cluster k if $\hat{\pi}_k \, \Phi\left(\boldsymbol{Y_i} \mid \boldsymbol{\hat{\mu}_k}, \hat{\Sigma}_k\right) > \hat{\pi}_\ell \, \Phi\left(\boldsymbol{Y_i} \mid \boldsymbol{\hat{\mu}_\ell}, \hat{\Sigma}_\ell\right)$ for all $\ell \neq k$.

2.2 The model collection

Currently, statistics deals with problems where data are described by many variables. In principle, the more information one has 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 determine the relevant variables for the Gaussian mixture clustering process.

In this paper, spherical Gaussian mixtures are considered where covariances fulfill $\Sigma_k = \sigma^2 I_p$ for all $k = 1, \ldots, K$ with $\sigma^2 > 0$. Then the clusters are characterized by the mean parameters and a variable j is irrelevant for the clustering if μ_{kj} is independent of k. Without loss of generality, the data are assumed to have a null expectation: $\mathbb{E}[Y_{ij}] = 0$. In practice, empirical centering of the data is performed to ensure this assumption. Since the expectation $\mathbb{E}[Y_{ij}]$ is estimated by $\sum_{k=1}^{K} \pi_k \mu_{kj}$, a variable j is called *irrelevant* for the clustering if $\mu_{kj} = 0$ for all $k = 1, \ldots, K$, otherwise it is called *relevant* for the clustering. In the sequel, $\boldsymbol{J_r}$ denotes the subset of relevant variables and $\boldsymbol{J_r^c} := \{1, \ldots, p\} \setminus \boldsymbol{J_r}$ is the irrelevant variable subset. Moreover, for all $\boldsymbol{y} \in \mathbb{R}^p$, $\boldsymbol{y}_{[\boldsymbol{J_r}]}$ denotes the restriction of $\boldsymbol{y}$ on $\boldsymbol{J_r}$.

Consequently, the following model collection indexed by the number of clusters $K \in \mathbb{N}^*$ and a relevant variable subset $\boldsymbol{J_r}$, is considered to estimate the density s:

$$\mathcal{S}_{(K,\boldsymbol{J_r})} = \left\{ \begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\boldsymbol{\theta}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_r^c}]} \mid \boldsymbol{0}, \sigma^2 I) \sum\limits_{k=1}^{K} \pi_k \, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]} \mid \boldsymbol{\mu_k}, \sigma^2 I) \\ \boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \in \Pi_K \times \left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K \times \mathbb{R}_+^* \end{array} \right\}. \tag{1}$$

The dimension of a model $\mathcal{S}_{(K,\boldsymbol{J_r})}$ corresponds to the total number of free parameters estimated in the model: $D_{(K,\boldsymbol{J_r})} = K(1 + |\boldsymbol{J_r}|)$.

3 Three competitive procedures for variable selection in clustering

In this paper, a variable selection procedure is proposed to select relevant variables for improving data clustering. This procedure is a compromise between MM-MLE and PS-Lasso procedures proposed by Maugis and Michel (2011a,b) and Pan and Shen (2007) respectively. These three variable selection procedures are presented in this section and summarized in Table 1.

3.1 MM-MLE procedure

In their procedure, Maugis and Michel (2011a,b) consider a model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}}$ where $\mathcal{S}_{(K,\boldsymbol{J_r})}$ is defined by (1) and the model collection is indexed by $\mathcal{M} = \mathbb{N}^* \times \mathcal{J}$, where $\mathcal{J}$ denotes the collection of the non empty subsets of $\{1,\dots,p\}$. For each $(K,\boldsymbol{J_r}) \in \mathcal{M}$, the maximum likelihood estimator

$$\hat{s}_{(K,\boldsymbol{J_r})} = \underset{t\in\mathcal{S}_{(K,\boldsymbol{J_r})}}{\operatorname{argmin}}\ \gamma_n(t) \text{ with } \gamma_n(t) = -\frac{1}{n}\sum_{i=1}^{n}\ln[t(\boldsymbol{Y_i})]$$

is computed using an EM algorithm (Dempster *et al.*, 1977). In practice, MIXMOD software (Biernacki *et al.*, 2006) or MCLUST (Fraley and Raftery, 2003) can be used for instance.

Next, they propose to solve the model selection problem with a data-driven penalized criterion. The theoretical construction of this criterion is the topic of Maugis and Michel (2011b). The resulting penalty is proportional to D/n, up to an unknown multiplicative constant κ. In practice, this penalty is calibrated using the slope heuristics (Birgé and Massart, 2006), recalled in Section 4.2. First, models are grouped according to their dimension D in order to obtain a model collection $\{\mathcal{S}_D\}_{D\in\mathcal{D}}$. For each dimension $D \in \mathcal{D}$, let $\hat{s}_D$ be the maximum likelihood estimator in $\mathcal{S}_D$ and $(K_D, \boldsymbol{J}_D)$ such that $\hat{s}_D = \hat{s}_{(K_D,\boldsymbol{J}_D)}$. Second, the slope $\hat{\kappa}$ is estimated to calibrate the penalty, based on the linear behavior of the function $D/n \mapsto -\gamma_n(\hat{s}_D)$ for large dimensions (see Baudry *et al.*, 2011, for more details). Third, the minimizer $\hat{D}$ of the penalized criterion

$$\hat{D} = \underset{D\in\mathcal{D}}{\operatorname{argmin}} \left\{\gamma_n(\hat{s}_D) + 2\hat{\kappa}\,\frac{D}{n}\right\} \tag{2}$$

is determined and $(\hat{K}, \hat{\boldsymbol{J}}_{\boldsymbol{r}}) := (K_{\hat{D}}, \boldsymbol{J}_{\hat{D}})$ is selected. Finally, a data clustering is derived from $\hat{s}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}})}$ by applying the MAP principle.

The MM-MLE procedure is only appropriate for low dimensional problem ($p \le n$). And, since the model collection is indexed by the complete collection of subsets $\mathcal{J}$, the procedure is time consuming. When p is very large, the variables may be ordered but this preliminary step is limiting in practice. Thus, alternative variable selection procedures have to be considered for clustering high-dimensional data.

3.2 PS-Lasso procedure

Unlike Maugis and Michel (2011a), Pan and Shen (2007) do not consider a deterministic model collection. In light of the success of variable selection via ℓ_1-penalization in regression, they rather construct a random model collection derived from a collection of potentially relevant variable subsets determined by a sparse procedure. The relevant variable selection and the parameter estimation are performed in the same process.

For all $K \in \mathbb{N}^*$, let $\mathcal{S}_K = \mathcal{S}_{(K,\{1,\dots,p\})} = \{s_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta_K\}$ according to (1) with $\Theta_K := \Pi_K \times (\mathbb{R}^p)^K \times \mathbb{R}_+^*$. To detect the relevant variables, Pan and Shen (2007) penalize the empirical contrast $\gamma_n(s_{\boldsymbol{\theta}})$ by a ℓ_1-penalty on the mean parameters proportional to $|\boldsymbol{\theta}|_1 := \sum_{j=1}^{p}\sum_{k=1}^{K}|\mu_{kj}|$. For all $\lambda \in G_K$, which is a chosen grid of regularization parameters, the Lasso estimator is defined by

$$\hat{\boldsymbol{\theta}}^L_{(K,\lambda)} = \underset{\boldsymbol{\theta}\in\Theta_K}{\operatorname{argmin}}\ \{\gamma_n(s_{\boldsymbol{\theta}}) + \lambda|\boldsymbol{\theta}|_1\}\,. \tag{3}$$

To compute $\hat{\boldsymbol{\theta}}^L_{(K,\lambda)}$, Pan and Shen (2007) construct an EM algorithm for ℓ_1-penalized model-based clustering. Then the relevant variable subset selected by the Lasso estimator $\hat{\boldsymbol{\theta}}^L_{(K,\lambda)}$ is

$$\boldsymbol{J}_{(K,\lambda)} = \{j \in \{1,\dots,p\};\ \exists k \in \{1,\dots,K\} \text{ such that } \hat{\mu}_{kj} \neq 0\}$$

and the density s is estimated by the Lasso solution $\hat{s}^L_{(K,\boldsymbol{J}_{(K,\lambda)})} := s_{\hat{\boldsymbol{\theta}}^L_{(K,\lambda)}}$.

By varying $\lambda \in G_K$ and $K \in \mathbb{N}^*$, they get a model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}^L}$ where $\mathcal{M}^L = \{(K,\boldsymbol{J_r});\ K \in \mathbb{N}^*, \boldsymbol{J_r} \in \mathcal{J}_K\}$, $\mathcal{J}_K = \bigcup\limits_{\lambda\in G_K} \boldsymbol{J}_{(K,\lambda)}$ and $\mathcal{S}_{(K,\boldsymbol{J_r})}$ is defined by (1).

Next, the BIC criterion is used to solve the model selection problem. First, models are grouped according to their dimension D in order to obtain a model collection $\{\mathcal{S}_D\}_{D\in\mathcal{D}}$. For each dimension $D \in \mathcal{D}$, let $\hat{s}^L_D$ be the maximum likelihood estimator in $\mathcal{S}_D$ and $(K_D,\boldsymbol{J}_D)$ such that $\hat{s}^L_D = \hat{s}^L_{(K_D,\boldsymbol{J}_D)}$. Second, the minimizer $\hat{D}$ of the BIC criterion

$$\hat{D} = \underset{D\in\mathcal{D}}{\operatorname{argmin}} \left\{\gamma_n(\hat{s}^L_D) + \frac{1}{2}\frac{D}{n}\ln n\right\} \tag{4}$$

is determined and $(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}}) = (K_{\hat{D}},\boldsymbol{J}_{\hat{D}})$ is selected. Finally, a data clustering is derived from the estimated Lasso parameter vector $\hat{\boldsymbol{\theta}}^L_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}})}$ by applying the MAP principle.

3.3 Our Lasso-MLE procedure

3.3.1 Motivation

From our point of view, taking advantage of the sparsity property of ℓ_1-penalization to perform automatic variable selection in clustering for high-dimensional data is an interesting idea which is worth exploring. Nevertheless, although the model-based clustering is linked to the density estimation problem, the PS-Lasso procedure does not take the density estimation purpose into account. Let us explain the main weakness of the PS-Lasso procedure as regards estimation.

Let $\mathcal{S}$ be the set of all densities with respect to Lebesgue measure on $\mathbb{R}^p$. In a maximum likelihood approach, the loss function considered is the Kullback-Leibler information defined for all $t \in \mathcal{S}$ by

$$\mathrm{KL}(s,t) = \int_{\mathbb{R}^p} \ln\left[\frac{s(\boldsymbol{y})}{t(\boldsymbol{y})}\right] s(\boldsymbol{y})\, d\boldsymbol{y}$$

if $s\,d\boldsymbol{y}$ is absolutely continuous with respect to $t\,d\boldsymbol{y}$ and $+\infty$ otherwise. The density s is the unique minimizer of the Kullback-Leibler information on $\mathcal{S}$. Ideally, we want to estimate s by the so-called oracle $\hat{s}_{(K^\star,\boldsymbol{J}_{\boldsymbol{r}}^\star)}$ where

$$(K^\star,\boldsymbol{J}_{\boldsymbol{r}}^\star) = \underset{(K,\boldsymbol{J_r})\in\mathcal{M}}{\operatorname{argmin}}\ \mathrm{KL}(s,\hat{s}_{(K,\boldsymbol{J_r})}). \tag{5}$$

In practice, $(K^\star,\boldsymbol{J}_{\boldsymbol{r}}^\star)$ depending on the unknown density s, $\hat{s}_{(K^\star,\boldsymbol{J}_{\boldsymbol{r}}^\star)}$ is unattainable. But it is a benchmark to evaluate the quality of any estimator of s. In particular, it is the benchmark for MM-MLE procedure. Yet for high-dimensional data, the collection $\mathcal{M}$ is so rich that performing the selection of the exhaustive best subset over $\mathcal{M}$ is unfeasible. In this case, a natural idea is to consider a subset $\mathcal{M}' \subset \mathcal{M}$ so that performing best subset selection over $\mathcal{M}'$ becomes practicable by aiming at

$$\underset{(K,\boldsymbol{J_r})\in\mathcal{M}'}{\operatorname{argmin}} \ \mathrm{KL}(s, \hat{s}_{(K,\boldsymbol{J_r})}).$$

Pan and Shen (2007) construct a subset $\mathcal{M}' = \mathcal{M}^L$ of $\mathcal{M}$. But, rather than considering the family of MLE estimator $\{\hat{s}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}^L}$ by aiming at

$$\underset{(K,\boldsymbol{J_r})\in\mathcal{M}^L}{\operatorname{argmin}} \ \mathrm{KL}\left(s, \hat{s}_{(K,\boldsymbol{J_r})}\right), \tag{6}$$

they consider the family of Lasso estimators $\left\{\hat{s}^L_{(K,\boldsymbol{J_r})}\right\}_{(K,\boldsymbol{J_r})\in\mathcal{M}^L}$ and aim at

$$\underset{(K,\boldsymbol{J_r})\in\mathcal{M}^L}{\operatorname{argmin}} \ \mathrm{KL}\left(s, \hat{s}^L_{(K,\boldsymbol{J_r})}\right). \tag{7}$$

Thus the aim of the proposed procedure is to mimic the model defined by Equation (6), which is expected to be closer to the oracle (5) than the model defined by (7).

3.3.2 Description of our procedure

We propose the so-called Lasso-MLE procedure which is a compromise between the two previous procedures (see Sections 3.1 and 3.2). This procedure is decomposed into three main steps. The first step consists of constructing a subcollection of models. As Pan and Shen (2007), an ℓ_1-approach is considered to obtain $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}^L}$. For a fix number of mixture components K and a regularization parameter λ, an EM algorithm is used to compute the Lasso solution. It differs from the usual EM algorithm (for computing the ML estimator) by the update of mean parameters. In the ℓ_1 procedures, the choice of the regularization parameter grid is often a difficulty. Contrary to Pan and Shen (2007), which opt for a deterministic grid difficult to choose, we propose to construct a data-driven grid of regularization parameters by using the updating formulas of the mixture parameters in the EM algorithm computing the Lasso solutions. This construction is detailed in Appendix C. The second step consists of computing the MLE $\hat{s}_{(K,\boldsymbol{J_r})}$ using the standard EM algorithm for each model $(K, \boldsymbol{J_r}) \in \mathcal{M}^L$. The third step is devoted to model selection. As in Maugis and Michel (2011a,b), a non asymptotic penalized criterion is proposed to solve the model selection problem. Its construction requires to extend the theoretical result to determine the penalty shape in the high-dimensional context and with a random model subcollection. Next, the penalty depending on unknown multiplicative constant(s) is calibrated using the slope heuristics. The construction of this penalized criterion is the topic of Section 4.

4 Model selection

4.1 An oracle inequality for the Lasso-MLE estimator

In the third step of the Lasso-MLE procedure, a model selection criterion is required to select the number of clusters K and the relevant variables subset

Procedure	model collection	parameter estimation	model selection
PS-Lasso procedure	ℓ_1-penalization on the mean parameters to detect the irrelevant variables ⇒ data-driven collection of relevant variable subsets ⇒ model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}^L}$	Lasso	BIC
MM-MLE procedure	deterministic collection of relevant variables subsets corresponding to an exhaustive variable selection ⇒ model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}}$	MLE	non asymptotic data-driven penalized criterion
Lasso-MLE	ℓ_1-penalization on the mean parameters to detect the irrelevant variables ⇒ data-driven collection of relevant variable subsets ⇒ model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}^L}$	MLE	non asymptotic data-driven penalized criterion

Table 1: Summary of the three procedures: PS-Lasso , MM-MLE and Lasso-MLE. K (resp. $\boldsymbol{J_r}$) denotes the number of clusters (resp. a relevant variable subset).

$\boldsymbol{J_r}$ simultaneously. We follow the approach developed by Birgé and Massart (1997) and Barron *et al.* (1999) which consists of defining a non asymptotic penalized criterion leading to an oracle inequality. In the context of density estimation, Barron *et al.* (1999) and Massart (2007, Theorem 7.11) propose a general model selection theorem for maximum likelihood estimation. Since this theorem is stated for a deterministic model collection, an extension is proposed for a random model subcollection (see Theorem 2 in Appendix A). Then, by applying Theorem 2 to the considered random model collection of the finite Gaussian mixtures, the following oracle inequality for the Lasso-MLE estimator is derived.

Theorem 1. *Let $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}}$ be the model collection defined by (1). Let $\mathcal{M}^L$ be a random subcollection of index sets (selected by the Lasso) included in the whole collection $\mathcal{M}$. Let A_μ, a_σ and A_σ be absolute positive constants and consider the collection of bounded models $\{\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\mathcal{M}^L}$ defined by*

$$\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})} = \left\{ s_{\boldsymbol{\theta}} \in \mathcal{S}_{(K,\boldsymbol{J_r})};\quad \boldsymbol{\theta} \in \Pi_K \times \left([-A_\mu, A_\mu]^{|\boldsymbol{J_r}|} \right)^K \times\ [a_\sigma, A_\sigma] \right\}. \tag{8}$$

Consider the maximum likelihood estimator

$$\hat{s}_{(K,\boldsymbol{J_r})} = \underset{s_{\boldsymbol{\theta}}\in\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}}{argmin}\ \gamma_n(s_{\boldsymbol{\theta}}).$$

Denote by $D_{(K,\boldsymbol{J_r})} = K(1 + |\boldsymbol{J_r}|)$ the dimension of the model $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}$. Define

$$B(A_\mu, A_\sigma, a_\sigma, p) := 1 + \sqrt{\ln\left[\frac{A_\sigma}{a_\sigma}\left(1 + \frac{A_\mu}{a_\sigma}\right)\right]} + \sqrt{\ln p}.$$

Let $s_{(K,\boldsymbol{J_r})} \in \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}$ *such that* $KL(s, s_{(K,\boldsymbol{J_r})}) \leq 2 \inf_{s_{\boldsymbol{\theta}} \in \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}} KL(s, s_{\boldsymbol{\theta}})$ *and let* $\tau > 0$ *such that*

$$s_{(K,\boldsymbol{J_r})} \geq e^{-\tau} s. \tag{9}$$

Let pen : $\mathcal{M} \mapsto \mathbb{R}_+$. *Suppose that there exists an absolute constant* $\kappa > 0$ *such that, for all* $(K, \boldsymbol{J_r}) \in \mathcal{M}$,

$$\begin{aligned}\text{pen}\,(K, \boldsymbol{J_r}) \geq \kappa \frac{D_{(K,\boldsymbol{J_r})}}{n} \Bigg[B^2(A_\mu, A_\sigma, a_\sigma, p) + \ln\left(\frac{1}{1 \wedge B^2(A_\mu, A_\sigma, a_\sigma, p)\frac{D_{(K,\boldsymbol{J_r})}}{n}}\right) \\ +(1 \vee \tau) \ln\left(\frac{p}{D_{(K,\boldsymbol{J_r})} \wedge p}\right)\Bigg] .\end{aligned} \tag{10}$$

Then, the estimator $\hat{s}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}})}$ *with*

$$(\hat{K}, \hat{\boldsymbol{J}}_{\boldsymbol{r}}) = \underset{(K,\boldsymbol{J_r}) \in \mathcal{M}^L}{\textit{argmin}} \left\{\gamma_n(\hat{s}_{(K,\boldsymbol{J_r})}) + \text{pen}(K, \boldsymbol{J_r})\right\}$$

satisfies

$$\begin{aligned} & \mathbb{E}\left[d_H^2\left(s, \hat{s}_{(\hat{K},\hat{\boldsymbol{J}}_{\boldsymbol{r}})}\right)\right] \\ \leq \quad & C\left(\mathbb{E}\left[\inf_{(K,\boldsymbol{J_r}) \in \mathcal{M}^L}\left\{\inf_{s_{\boldsymbol{\theta}} \in \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}} KL(s, s_{\boldsymbol{\theta}}) + \text{pen}(K, \boldsymbol{J_r})\right\}\right] + \frac{1 \vee \tau}{n}\right) \end{aligned} \tag{11}$$

for some absolute positive constant C.

The proof of Theorem 1 is given in Appendix B.

Note that Condition (9) is required to control the second moment of log-likelihood ratios in order to apply Bernstein's inequality to bound the empirical process of $\ln(s/s_{(K,\boldsymbol{J_r})})$ (see Lemma 1 in Appendix A.). The larger the parameter τ, the larger the minimal penalty (10) and the less accurate Inequality (11). Since $s_{(K,\boldsymbol{J_r})}$ is positive, there always exists some $\tau > 0$ fulfilling Condition (17) in Theorem 2. It seems difficult to have an idea of the minimal convenient value of τ since it depends on the unknown true density s. Nonetheless, we may think that Condition (9) is satisfied for reasonable values of τ because $s_{(K,\boldsymbol{J_r})}$ is expected to be close to s. Note that the constant 2 in the Kullback-Leibler constraint for $s_{(K,\boldsymbol{J_r})}$ can be replaced by $1 + \varepsilon$ with $\varepsilon > 0$.

As in Maugis and Michel (2011b), the mixture parameters are bounded (see the model collection (8)) in order to construct brackets over $\mathcal{S}_{(K,\boldsymbol{J_r})}$ and thus to upper bound the entropy number. The bracket construction is adapted from the one of Maugis and Michel (2011b) for our specific Gaussian mixtures (see Appendix B). We also obtain an Inequality (11) which is not exactly an oracle inequality since the Hellinger risk is upper bounded by the Kullback-Leibler bias. But contrary to MM-MLE procedure, our Lasso-MLE procedure runs on a small random subcollection of models and it remains feasible even for large p. Thus our estimator $\hat{s}_{(K,\boldsymbol{J_r})}$ is attainable in practice. Moreover, contrary to classical asymptotic criteria for which p is fixed and n tends to infinity, our result is non asymptotic and allows to study cases for which p increases with n. Since the ratio $\ln(p)/n$ appears in the right hand-side of Inequality (11) through the term pen$(K, \boldsymbol{J_r})$, Theorem 1 ensures that our estimator achieves good performance compared with the oracle as long as $p < \mathrm{e}^n$, which allows to consider many situations with $p \gg n$.

As expected, the penalty (10) is proportional to the model dimension and thus penalizes models with high complexities. It also involves two additional logarithm terms. On the one hand, the first logarithm term is due to a lack of accuracy in the proof of Theorem 1. Specifically, only a global entropy bracketing control is obtained while a local version is only required for applying Theorem 2. It is sufficient since the local entropy is upper bounded by the global entropy but it is not optimal and yields extra logarithm terms. On the other hand, the second logarithm term quantifies the complexity of the model collection by taking into account the possible large number of models with identical dimension. In regression, Birgé and Massart (2006) prove that, for complete variable selection, a penalty proportional to the dimension selects too complex models with high probability and that a logarithm term is necessary to select smaller models. This has been practically checked in many situations: for multiple change points detection in a regression framework (Lebarbier, 2005) or for histogram selection in a density estimation framework Castellan (1999) for instance. Nevertheless this logarithm term becomes unnecessary if the number of models with the same dimension is small enough. For instance, for finite Gaussian mixture models in a low-dimensional setting, Maugis and Michel (2011a) observe that a penalty proportional to the dimension,with no logarithm term, is sufficient to select a model close to the oracle. But in our high-dimensional context, the number of models having the same dimension is expected to grow. Nonetheless, thanks to the random preselection of relevant variables subsets, a complete variable selection is not performed here. However, it is difficult to know how rich is the random model collection. Thus, according to (10), we may retain that the penalty is $\mathrm{pen}(K, \boldsymbol{J_r}) = \kappa_1 \, \mathrm{pen}_{\text{shape}}(K, \boldsymbol{J_r})$ with

$$\mathrm{pen}_{\text{shape}}(K, \boldsymbol{J_r}) = \frac{D_{(K,\boldsymbol{J_r})}}{n} \left[1 + \kappa_2 \ln\left(\frac{p}{D_{(K,\boldsymbol{J_r})}}\right)\right] \tag{12}$$

where κ_1 and κ_2 are two unknown constants. But if our random model subcollection is much poorer than the whole model collection and contains few models with the same dimension, the penalty (12) may be too pessimistic. In this case, a penalty proportional to the dimension

$$\mathrm{pen}_{\text{shape}}(K, \boldsymbol{J_r}) = \frac{D_{(K,\boldsymbol{J_r})}}{n} \tag{13}$$

might be sufficient to select a model with proper dimension.

4.2 Practical penalty calibration

According to (12) and (13), the penalty is known up to multiplicative constant(s) and in practice, we have to choose between these two penalty shapes. To fill in the gap between theory on penalization and practical calibration penalty, Birgé and Massart (2006) proposed the so-called *slope heuristics*. In the context of Gaussian homoscedastic least squares regression with fixed design, they show that there exists a minimal penalty, namely such that the dimension and the risk of the models selected with smaller penalties become large. Moreover, they propose the rule of thumb of the slope heuristics: a penalty equal to twice the minimal penalty allows us to select a model close to the oracle model. This rule of thumb is theoretical proved in few specific frameworks (Birgé and Massart, 2006; Arlot and Massart, 2009; Arlot and Bach, 2010; Lerasle, 2011, 2012) and largely valid in practice (see for instance Lebarbier, 2005; Maugis and Michel, 2011a; Caillerie and Michel, 2011; Verzelen, 2010). For determining the minimal penalty, the data-driven slope estimation (DDSE) procedure proposed by Baudry *et al.* (2011) is used. This method is based on the existence of a linear behavior between the penalty shape and the contrast value for the most complex models (the empirical bias gets stable for the most complex models and the minimal penalty, corresponding to an empirical "estimation error" term, is given by the behavior of $-\gamma_n(\hat{s}_{(K,\boldsymbol{J_r})})$). In our context, the DDSE procedure first provides a graphical way to preliminary choose the penalty shape between (12) and (13), and second allows us to calibrate the penalty by estimating the linear slope. The reader is referred to Baudry *et al.* (2011) for more details about the DDSE procedure. Note that if the penalty shape (12) is required, the DDSE procedure is adapted with a double regression for calibrating the two constants κ_1 and κ_2 (instead of one constant usually).

5 Applications

In this section, the Lasso-MLE procedure is studied on two simulated examples and is compared with PS-Lasso and MM-MLE procedures. The Adjusted Rand Index (ARI) is used to measure the similarity between data clusterings and the variables declared relevant for the clustering by the three procedures are compared. We also compare the data-driven model selection criterion with two widely-used criteria: the Akaike Information Criterion (AIC) of Akaike (1973) and defined by $\gamma_n(\hat{s}_D)+\frac{D}{n}$; the Bayesian Information Criterion (BIC) of Schwarz (1978) defined by $\gamma_n(\hat{s}_D)+\frac{\ln n}{2}\,\frac{D}{n}$. Moreover, since the density s is known for simulated datasets, the model selected by one criterion can be compared with the associated oracle model to judge the quality of this criterion. The oracle model is defined by

$$D_{\text{oracle}} = \underset{D\in\mathcal{D}}{\operatorname{argmin}}\, \mathrm{KL}\,(s,\hat{s}_D) = \underset{D\in\mathcal{D}}{\operatorname{argmin}} \left\{ -\int_{\boldsymbol{x}\in\mathbb{R}^p} \ln\left[\hat{s}_D(\boldsymbol{x})\right] s(\boldsymbol{x})\; d\boldsymbol{x} \right\}.$$

In practice, the oracle model is obtained by approximating the integral by a Monte Carlo procedure.

5.1 First simulated example

This first simulated dataset is in the spirit of an example in Maugis and Michel (2011a). The dataset consists of $n = 200$ observations described by p variables. The data are simulated according to a mixture of four spherical Gaussian distributions $\Phi(\cdot \mid \boldsymbol{\mu_k}, I)$ where

$$\begin{aligned}
&\boldsymbol{\mu_1} = -\boldsymbol{\mu_3} = (3, 2, 1, 0.7, 0.3, 0.2, 0.1, 0.07, 0.05, 0.025, \mathbf{0_{p-10}}),\\
&\boldsymbol{\mu_2} = \mathbf{0_p},\\
&\boldsymbol{\mu_4} = (3, -2, 1, -0.7, 0.3, -0.2, 0.1, -0.07, -0.05, -0.025, \mathbf{0_{p-10}}).
\end{aligned}$$

The vector $\mathbf{0}_{\boldsymbol{l}}$ denotes the null vector of length l. The mixing proportions are $(\pi_1, \pi_2, \pi_3, \pi_4) = (0.3, 0.2, 0.2, 0.3)$. The relevant clustering variables are the first ten variables ($\boldsymbol{J}_{\boldsymbol{r}}^{\star} = \{1, \dots, 10\}$). Note that the four subpopulations are progressively gathered together into a unique Gaussian distribution, thus the discriminant power of the relevant variables decreases with respect to the variable index (see Figure 1). In this simulation study 20 datasets are simulated for each value of $p \in \{30, 200, 1000\}$, and mixtures with $K \in \{2, \dots, 6\}$ components are considered.

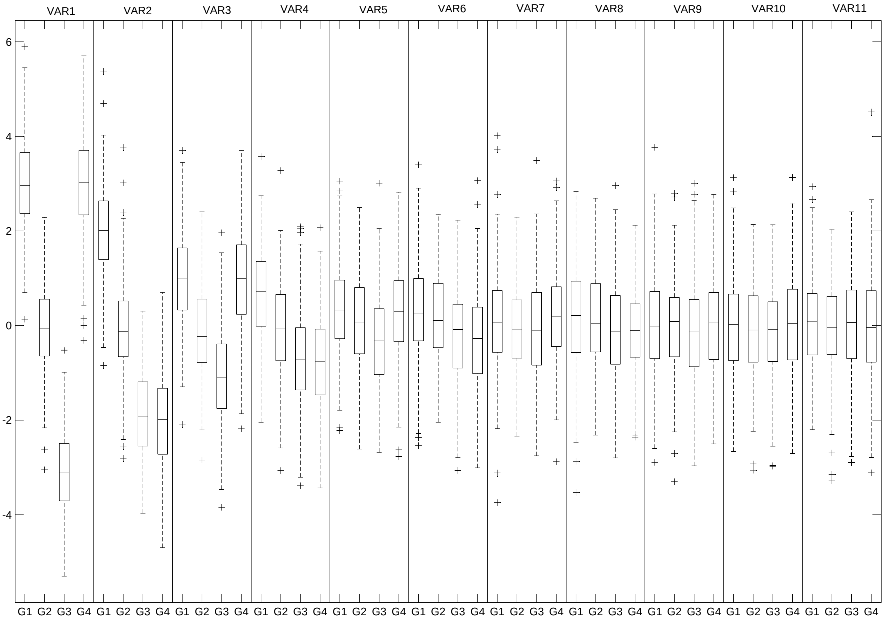

Figure 1: Boxplots of the first eleven variables (VAR1,...,VAR11) on the four mixture components (G1,G2,G3,G4).

The Lasso-MLE procedure is compared with PS-Lasso and MM-MLE procedures in the low-dimensional case ($p = 30$) and only with PS-Lasso procedure in the high-dimensional cases ($p = 200, 1000$). Note that for the MM-MLE procedure, the variables are preliminary ordered (the collection of relevant variable subsets is $\{(1, 2, \dots, d), 1 \leq d \leq p\}$) as in Maugis and Michel (2011a) because the model collection is too rich to perform complete variable selection. Figure 2 checks that the function $D/n \in [0, 1] \mapsto -\gamma_n(\hat{s}_D)$ has a linear behavior for most complex models, which justifies the use of penalty shape (13). Note that the values of the estimated slopes by DDSE for Lasso-MLE and MM-MLE procedures are very similar here. The results are summarized in Table 2.

Low-dimensional case ($p = 30 \ll n$)

With the Lasso-MLE procedure, the data-driven criterion (DDSE), BIC and the oracle globally select the true number of cluster $K^{\star} = 4$. With the variable selection, the DDSE criterion allows us to select a relevant variable subset closer to the oracle than BIC. Moreover, note that the model collections considered

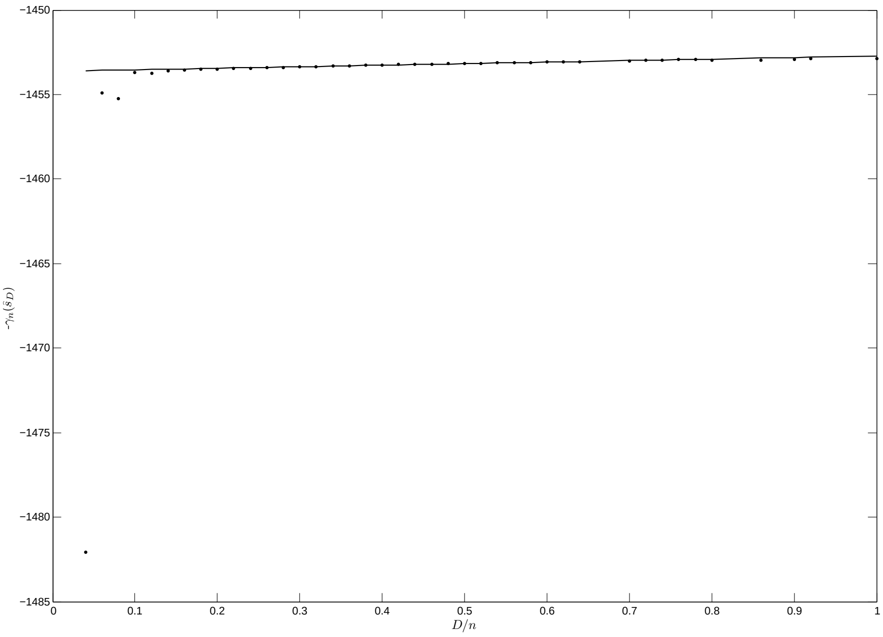

Figure 2: For one simulation with $p = 1000$, graphical validation of the penalty shape ($\text{pen}(D) = 0.91D/n$).

by both procedures are different. With the MM-MLE procedure, the ordered variable subsets is considered, containing the true relevant variable subset $\boldsymbol{J_r^\star}$. This explains in particular why the selected models never contains false relevant variables. On the contrary, the Lasso-MLE procedure considers data-driven relevant variable subsets for each simulation. In particular, $\boldsymbol{J_r^\star}$ may not belong to this data-driven collection. Moreover, the relevant variable subsets collection obtained by varying the regularization parameter in the Lasso procedure may remove the least relevant variables ($j \in \{7, \dots, 10\}$, see Figure 1) before some noisy variables. With both procedures, the oracle model does not coincide with the true model. Furthermore, AIC often selects too many components and more relevant variables than the other criteria.

With the PS-Lasso procedure, the oracle model sometimes overestimates the number of mixture components. It contains most (yet not all) true relevant variables but also many false relevant variables. This tendency to select too many variables has already been widely noted in the regression framework (Zhao and Yu, 2007; Zou, 2006; Yuan and Lin, 2007; Bach, 2008; Connault, 2011). For this procedure, the errors of variable selection are reduced by using BIC.

High-dimensional case ($p = 200, 1000$)

For the two high-dimensional scenarios where $p = 200$ and $p = 1000$, the Lasso-MLE procedure and PS-Lasso procedure are compared. First, results in Table 2 show that the variable selection problem for the PS-Lasso procedure (oracle and BIC) gets worse when p grows: the selected variable subset contains more and more false relevant variables at the expense of true relevant variables.

p	procedure	estimator	TR	FR	$\hat{K}$					ARI
					2	3	4	5	6	
30	PS-Lasso	oracle	9.1 (0.8)	14.2 (1.3)	0	0	14	6	0	0.90 (0.05)
		BIC	6.2 (0.8)	2.3 (1.7)	0	0	14	6	0	0.87 (0.06)
	MM-MLE	oracle	6.2 (1.6)	0.0 (0.0)	0	0	20	0	0	0.88 (0.05)
		AIC	7.3 (1.3)	0.0 (0.0)	0	0	8	6	6	0.88 (0.04)
		BIC	4.9 (0.8)	0.0 (0.0)	0	0	20	0	0	0.89 (0.05)
		DDSE	6.0 (1.3)	0.0 (0.0)	0	0	18	2	0	0.89 (0.05)
	Lasso-MLE	oracle	6.1 (1.4)	0.4 (0.7)	0	0	20	0	0	0.89 (0.06)
		AIC	8.2 (1.4)	2.4 (1.3)	0	0	8	8	4	0.89 (0.06)
		BIC	5.1 (1.2)	0.0 (0.0)	0	0	20	0	0	0.89 (0.06)
		DDSE	6.3 (1.4)	1.3 (1.5)	0	0	18	2	0	0.90 (0.05)
200	PS-Lasso	oracle	8.3 (1.4)	61.8 (8.7)	0	0	14	4	2	0.84 (0.04)
		BIC	5.8 (1.4)	4.1 (3.6)	0	0	14	4	2	0.79 (0.08)
	Lasso-MLE	oracle	5.9 (1.2)	0.4 (0.6)	0	0	20	0	0	0.85 (0.05)
		AIC	7.3 (1.4)	10.7 (5.7)	0	0	10	8	2	0.82 (0.04)
		BIC	5.2 (1.1)	1.1 (0.7)	0	0	20	0	0	0.84 (0.05)
		DDSE	6.2 (1.5)	1.6 (1.3)	0	0	20	0	0	0.84 (0.05)
1000	PS-Lasso	oracle	6.2 (1.4)	99.7 (18.8)	0	0	8	8	4	0.83 (0.04)
		BIC	5.3 (0.8)	12.1 (2.7)	0	0	10	6	4	0.77 (0.07)
	Lasso-MLE	oracle	5.4 (0.9)	1.2 (0.5)	0	0	19	1	0	0.84 (0.04)
		AIC	6.3 (1.2)	13.4 (8.0)	0	0	10	6	4	0.81 (0.09)
		BIC	5.6 (1.3)	5.2 (3.6)	0	0	19	1	0	0.83 (0.05)
		DDSE	5.6 (1.4)	2.3 (1.6)	0	0	19	1	0	0.84 (0.06)

Table 2: Averaged number of true relevant (TR) and false relevant (FR) variables (± standard deviation); number of times a clustering with $\hat{K} = 2$, 3, 4, 5 or 6 components is selected; Averaged ARI (± standard deviation) over the 20 simulations.

Moreover, the selection of the component number is deteriorated. With the Lasso-MLE procedure, the models selected by AIC and, to a lesser extend by BIC, contain more (true relevant and false relevant) variables when p grows. On the opposite, the oracle model and the model selected by the data-driven penalized criterion remain stable. Moreover, the average multiplicative factor $2\hat{\kappa}$ in the data-driven penalty $\text{pen}(D) = 2\hat{\kappa}\frac{D}{n}$ is equal to $1.47(\pm 0.10)$, $2.27(\pm 0.19)$ and $3.68(\pm 0.57)$ when $p = 30$, 200 and 1000 respectively. This factor globally increases with respect to p and thus the associated penalty becomes stronger as p increases, whereas the fixed BIC penalty $(0.5 \ln(n) = 2.65)$ tends to underpenalize as p grows.

Globally the data clustering slightly deteriorates as p grows. The biggest deterioration is for PS-Lasso procedure. The ARI for each model in PS-Lasso and Lasso-MLE procedures is represented in Figure 3. For PS-Lasso procedure, the models achieving the best clustering have moderate or high dimension while BIC selects a less complex model. On the opposite, for Lasso-MLE procedure, small models achieve good clustering. In particular, the DDSE, which selects such a small model, leads to a satisfactory data clustering. Figure 3 also confirms that variable selection is useful to get a better data clustering: introducing the relevant variables into the models improves the clustering but adding after the irrelevant variables deteriorates the clustering.

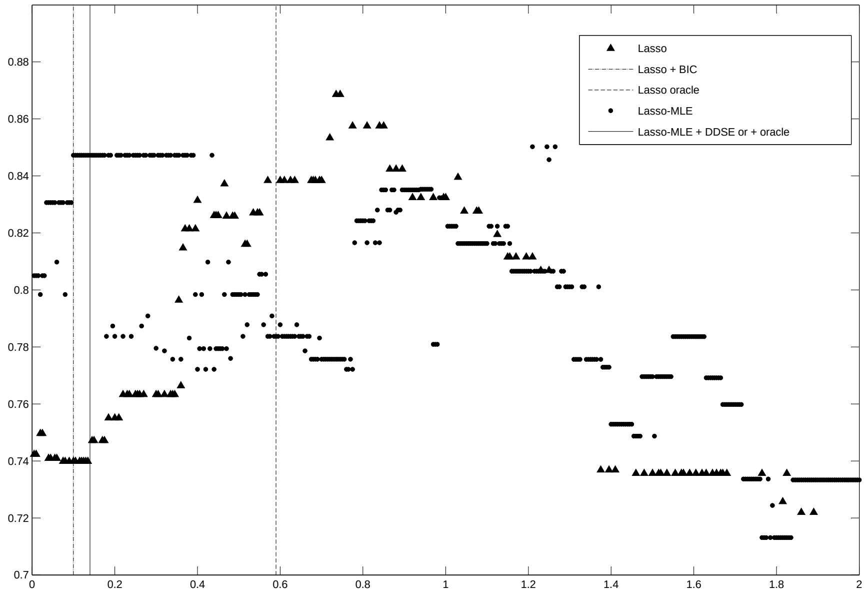

Figure 3: For one simulation with $p = 200$, ARI values for each model with $K = 4$ clusters obtained with the PS-Lasso and Lasso-MLE procedures are represented according to D/n. D/n is equal to 0.1 (resp. 0.59) for the model selected by PS-Lasso procedure with BIC (resp. by the Lasso oracle) and equal 0.14 for the Lasso-MLE procedure with DDSE and the oracle.

5.2 Second simulated example

This second simulated example is proposed in Pan and Shen (2007). The dataset consists of $n = 200$ observations described by $p = 1000$ variables. The data are simulated according to a mixture of two Gaussian distributions $\pi_1 \Phi(\cdot|\mathbf{0_p}, I) + (1-\pi_1)\Phi(\cdot|\boldsymbol{\mu_2}, I)$ where $\boldsymbol{\mu_2} = (1.5, \ldots, 1.5, \mathbf{0_{950}})$ and $\pi_1 = 0.85$. The relevant variables are the first fifty variables ($\boldsymbol{J_r^\star} = \{1, \ldots, 50\}$). We perform 20 simulations of the dataset. For each simulation, models with $K \in \{1, 2, 3\}$ clusters are considered. In this high-dimensional context, the Lasso-MLE procedure is compared with the PS-Lasso procedure. The results are summarized in Table 3.

Table 3 shows that the Lasso oracle model, and to a lesser extend the model selected by BIC, contain many false relevant variables and may overestimate the number of mixture components. This confirms that the PS-Lasso procedure is not suited to recover the true model and the true relevant variables. Moreover, BIC data clustering is disappointing. In contrast, the Lasso-MLE oracle model always coincides with the true model and leads to a very good data clustering. The DDSE achieves better performance than BIC and AIC.

procedure	estimator	TR	FR	K			ARI
				1	2	3	
PS-LASSO	oracle	50.3 (0.2)	214.6 (79.0)	0	16	4	0.90 (0.03)
	BIC	49.7 (0.8)	14.3 (3.4)	0	18	2	0.86 (0.02)
Lasso-MLE	oracle	50.0 (0.0)	0.2 (0.2)	0	20	0	0.95 (0.02)
	AIC	50.0 (0.0)	17.1 (4.2)	0	14	6	0.90 (0.04)
	BIC	49.8 (0.4)	4.4 (2.2)	0	20	0	0.92 (0.02)
	DDSE	50.0 (0.0)	2.4 (1.7)	0	20	0	0.94 (0.02)

Table 3: Averaged number of true relevant (TR) and false relevant (FR) variables (± standard deviation); number of times a clustering with $\hat{K} = 1$, 2 and 3 components is selected; Averaged ARI (± standard deviation) over the 20 simulations.

6 Extension for curve clustering

For some clustering problems, one may be interested in providing a sparse parameter estimation for each cluster besides clustering the data. This is typically the case for curve clustering problems involving a smooth representative curve per cluster (see for instance Misiti *et al.*, 2007a; Auder and Fischer, 2011). This sparse curve reconstruction allows to improve the curve clustering interpretability and may be used for prediction. For instance, this problem is encountered in the study of electricity consumption where representative curves are desirable to forecasting.

For such problems, curves are preliminary decomposed into some appropriate basis such as a wavelet basis. Then, the data are the coefficients associated to the curve decomposition into the basis. If the Lasso-MLE procedure is used on the empirically centered coefficient matrix to obtain a curve partition, representative curves per cluster are not directly available. One solution to estimate a representative profile for a given cluster would be to take the mean of the observed curves assigned to this cluster. Nonetheless, this process produces noised profiles and it is not able to provide smooth profiles. To overcome this problem, an alternative to the Lasso-MLE procedure which avoids empirical centering of the data and which is suited to clustering involving smooth representative curves is proposed. This alternative is based on a more general variable role modeling, implying a richer model collection than (1). The theoretical model selection criterion is adapted and we highlight that a logarithm factor in the penalty is now observed, due to the model collection richness.

6.1 Description of the alternative procedure

In this section, data are not assumed to fulfill the assumption $\mathbb{E}[Y_{ij}] = 0$. A variable j is now irrelevant for the clustering if $\mu_{1j} = \ldots = \mu_{Kj} := \nu_j$, ν_j being not necessary equal to zero. With the previous modeling, given a number K of clusters and relevant variables subset $\boldsymbol{J_r}$, $K|\boldsymbol{J_r}| + |\boldsymbol{J_r^c}| \geq p$ mean parameters have to be estimated, even when $p \gg n$.

Therefore, selecting the relevant clustering variables $\boldsymbol{J_r}$ is not sufficient to get sparse models and an additional dimensional reduction step is needed before the estimation step. Here, the irrelevant variables $\boldsymbol{J_r^c}$ are assumed to be decomposed into the *zero irrelevant* variables $\boldsymbol{J_0}$ ($j \in \boldsymbol{J_0}$ if $\nu_j = 0$) and *non zero irrelevant* variables $\boldsymbol{J_0^c}$ ($j \in \boldsymbol{J_0^c}$ if $\nu_j \neq 0$). Moreover, we assume that the irrelevant variables are predominantly zero irrelevant ($|\boldsymbol{J_0}| \gg |\boldsymbol{J_0^c}|$). In particular, this assumption is fulfilled for curve clustering using sparse representations of signals in some appropriate basis such as a wavelet basis. This new variable role modeling leads to the new model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_0})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_0}) \in \mathcal{M}^{LL}}$ where $\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_0})}$ is defined by

$$\left\{\begin{array}{l} \boldsymbol{y} \in \mathbb{R}^p \mapsto s_{\boldsymbol{\theta}}(\boldsymbol{y}); \\ s_{\boldsymbol{\theta}}(\boldsymbol{y}) = \Phi(\boldsymbol{y}_{[\boldsymbol{J_0}]}|\boldsymbol{0}, \sigma^2 I)\ \Phi(\boldsymbol{y}_{[\boldsymbol{J_0^c}]}|\boldsymbol{\nu}, \sigma^2 I)\ \sum\limits_{k=1}^{K} \pi_k\, \Phi(\boldsymbol{y}_{[\boldsymbol{J_r}]}|\boldsymbol{\mu_k}, \sigma^2 I) \\ \boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\nu}, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma) \in \Pi_K \times \mathbb{R}^{|\boldsymbol{J_0^c}|} \times \left(\mathbb{R}^{|\boldsymbol{J_r}|}\right)^K \times\ \mathbb{R}_+^* \end{array}\right\} \qquad (14)$$

The dimension of the model $\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_0})}$ is $D_{(K,\boldsymbol{J_r},\boldsymbol{J_0})} = K(1 + |\boldsymbol{J_r}|) + |\boldsymbol{J_0^c}|$. In practice, the random model collection $\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_0})}$ is determined using the following alternative procedure. First, a model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r}) \in \mathcal{M}^L}$ is constructed using the Lasso-MLE procedure on the centered data. Second, for each $\boldsymbol{J_r^c}$, the empirical contrast restricted to the $\boldsymbol{Y_{i[J_r^c]}}$'s is penalized by an ℓ_1-penalty proportional to $\|\boldsymbol{\nu}\|_1$ with various regularization parameter values, in order to detect the zero irrelevant variables. Next, Theorem 1 can be easily adapted for the model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_0})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_0}) \in \mathcal{M}^{LL}}$ in order to construct a data-driven non asymptotic penalized criterion to solve the model selection problem. As in Section 4, the same sufficient penalty shape (10) is established with model dimensions $D_{(K,\boldsymbol{J_r},\boldsymbol{J_0})}$. But this alternative procedure depending on two embedding Lasso algorithms, the model collection is richer than the model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r}) \in \mathcal{M}^L}$ obtained with the Lasso-MLE procedure. As a consequence, the logarithm term, which takes into account the richness of the model collection and the number of models having the same dimension, is now observed in the penalty shape (see Figure 5). As explained in Section 4.2, the penalty is calibrated using the DDSE method.

6.2 Functional data clustering example

The functional dataset consists of $n = 200$ noisy curves simulated as follows: the wavelet coefficients $\{Y_{ij}, i = 1, \ldots, n, j = 1, \ldots, p\}$ with $p = 1086$ are simulated according to the Gaussian mixture $0.85\,\Phi(\cdot \mid \boldsymbol{\mu_1}, \boldsymbol{I}) + 0.15\,\Phi(\cdot \mid \boldsymbol{\mu_2}, \boldsymbol{I})$ with mean vectors $\boldsymbol{\mu_1} = (\mathbf{0_{25}}, \mathbf{1.5_{25}}, \mathbf{0_{p-50}})$ and $\boldsymbol{\mu_2} = (\mathbf{1.5_{50}}, \mathbf{0_{p-50}})$ ($\boldsymbol{a_l}$ denotes the vector of length l whose coordinates equal a). Each wavelet coefficients vector $\boldsymbol{Y_i}$ corresponds to the decomposition of a curve $\boldsymbol{g_i}$ in the symmlet-4 basis at level 10 (Misiti *et al.*, 2007b). In particular, $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ are the decomposition of $\boldsymbol{f_1}$ and $\boldsymbol{f_2}$ which are the discretization of two functions on a fine time grid $\{t_1, \ldots, t_T\}$ containing $T = 1024$ points, represented at Figure 4 (top left). Thus, the first 25 variables are relevant for the clustering while the variables 26 to 50 are non zero irrelevant.

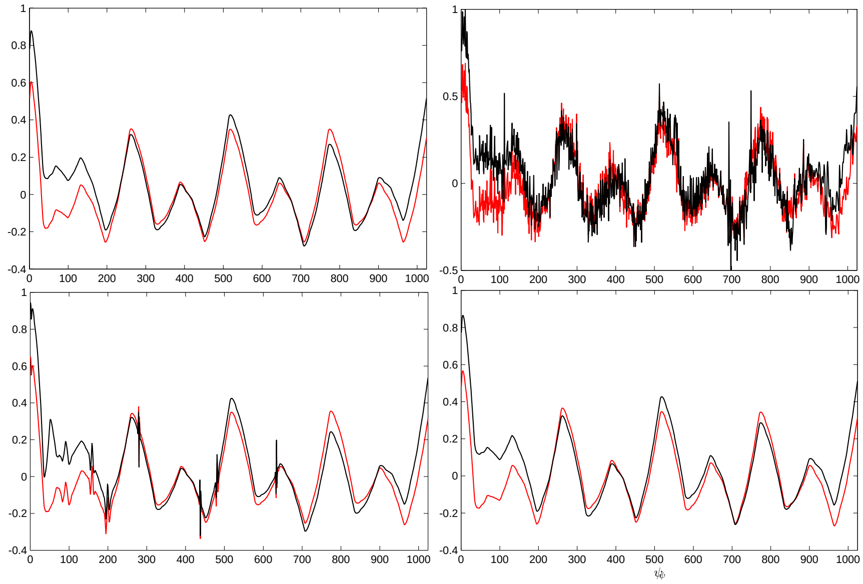

Figure 4: Top left : true curves for the two clusters. Top right : curve estimations obtained by the Lasso-MLE procedure with a penalty without logarithm term. Bottom : curve estimations obtained by the alternative procedure with a penalty without (left) and with (right) a logarithm term.

In order to obtain a clustering of this 200 curves, our alternative method is used on the wavelet coefficients. Next, an inverse wavelet transform of the estimated mean vectors is performed to obtain a denoised representative curve for each cluster. Models with $K \in \{1, 2, 3\}$ mixture components are considered. The alternative procedure is compared with the Lasso-MLE procedure. Both procedures select a model with two classes. The two estimated representative curves obtained by the Lasso-MLE procedure with model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r})}\}_{(K,\boldsymbol{J_r})\in\mathcal{M}^L}$ described in Section 3.3.2 and by the alternative procedure with model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_0})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_0})\in\mathcal{M}^{LL}}$ described in Section 6.1 are displayed in Figure 4. As expected, the curve estimations obtained by the

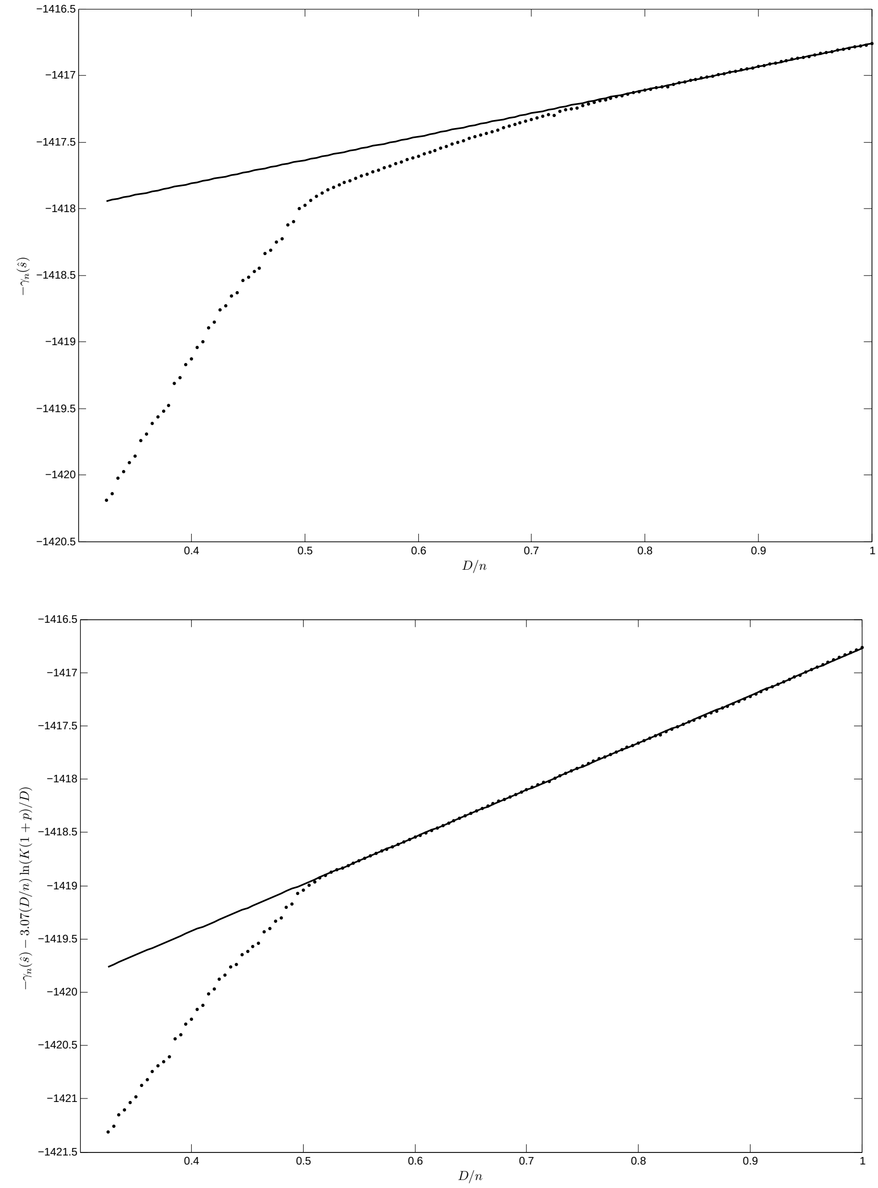

Figure 5: For one simulation, slope graphs obtained by considering the model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_0})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_0})\in\mathcal{M}^{LL}}$ with a penalty without (top) and with (bottom) a logarithm term. The resulting calibrated penalties are $\text{pen}(D) = 2\times 1.75D/n$ and $\text{pen}(D) = 2(4.42D/n + 3.07(D/n)\ln(K(1+p)/D))$ respectively.

Lasso-MLE procedure is very noisy because no thresholding of the smallest mean coefficients is performed. Note that the PS-Lasso procedure suffers from the same drawback and moreover, it gives shrunk curve profile estimations since the non zero mean wavelet coefficients are shrunk due to ℓ_1-penalization. On the contrary, the alternative procedure leads to smoother estimated representative curves. This is all the more true for the estimations obtained by the penalty shape with the logarithm term rather than without the logarithm term. This highlights that the model collection $\{\mathcal{S}_{(K,\boldsymbol{J_r},\boldsymbol{J_0})}\}_{(K,\boldsymbol{J_r},\boldsymbol{J_0})\in\mathcal{M}^{LL}}$ is rich enough to require a penalty with a logarithm factor and that a penalty proportional to the dimension underpenalizes, resulting in a less smooth curve estimation with some extra peaks. The slope graphs obtained for one simulation of this dataset are represented in Figure 5.

7 Discussion

In this paper, the Lasso-MLE procedure is proposed to cluster data, detecting the relevant clustering variables. This procedure is especially suited to study high-dimensional datasets. It is based on a compromise between a ℓ_1-regularization procedure and a MLE procedure. The variable selection and the data clustering problems are recast into a general model selection problem. A ℓ_1-approach is used to perform automatic variable selection and thus to deduce a reasonable random model collection. A data-driven penalized criterion is then built to solve the model selection problem and a data clustering is deduced to the associated ML estimator using the MAP principle.

Only spherical Gaussian mixtures are considered in this paper to focus on the mean vectors of mixtures. Nevertheless the Lasso-MLE procedure may be extended for Gaussian mixtures with more general variance matrices. First a more general Lasso procedure can be adapted to the ℓ_1-procedure of Zhou *et al.* (2009). Second, Theorem 1 may be extended for a random collection of general Gaussian mixture models by using the control of the bracketing entropy for the general Gaussian mixture families established in Maugis and Michel (2011b). The main difficulty lies in the definition of a relevant clustering variable, which depends on the behavior of the mean vectors but also the definition of variance matrices for the studied mixtures. This could have consequences on the penalty shape in particular. This extension will be the topic of a future work.

This use of a ℓ_1-approach to only produce a reasonable collection of relevant variable subsets is an interesting idea which may be beneficial for other variable selection procedures. For instance, Maugis *et al.* (2009b) propose a procedure based on a general variable role modeling but the associated algorithm SELVARCLUSTINDEP is too slow when the number of variables increases. A ℓ_1 variable selection procedure may be used to initialize the forward version of this algorithm to improve its stability and/or give a way of variable subsets to follow in order to dramatically speed up the algorithm.

Acknowledgements

The authors are grateful to Gilles Celeux (INRIA Saclay-Île-de-France) and Pascal Massart (Université Paris-Sud) for helpful discussions and valuable comments.

A A model selection theorem for MLEs in random models

The Hellinger distance between two nonnegative integrable functions t and u is the norm $\|\sqrt{t}-\sqrt{u}\|$, denoted $d_H(t,u)$. Consider $\mathcal{S}$ the set of all densities on $\mathbb{R}^p$. An ε-bracketing for a subset S of $\mathcal{S}$ with respect to d_H is a set of integrable function pairs $(l_1,u_1),\dots,(l_N,u_N)$ such that for each $t\in S$, there exists $j\in\{1,\dots,N\}$ such that $l_j\le t\le u_j$ and $d_H(l_j,u_j)\le\varepsilon$. 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 $\mathcal{H}_{[.]}(\varepsilon,S,d_H)=\ln[\mathcal{N}_{[.]}(\varepsilon,S,d_H)]$.

Let $\{S_m\}_{m\in\mathcal{M}}$ be some at most countable model collection such that $S_m\subset\mathcal{S}$ for all $m\in\mathcal{M}$. We shall say that $\{S_m\}_{m\in\mathcal{M}}$ fulfills Property $(\boldsymbol{\mathcal{P}})$ if, for all $m\in\mathcal{M}$, $\sqrt{\mathcal{H}_{[.]}(\varepsilon,S_m,d_H)}$ is integrable at 0 and if there exists a function Ψ_m on $\mathbb{R}_+$ such that Ψ_m is nondecreasing, $\xi\to\Psi_m(\xi)/\xi$ is nonincreasing on $]0,+\infty[$, and for $\xi\in\mathbb{R}_+$ and $u\in S_m$, denoting $S_m(u,\xi)=\{t\in S_m; d_H(t,u)\le\xi\}$,

$$\int_0^\xi\sqrt{\mathcal{H}_{[.]}(\varepsilon,S_m(u,\xi),d_H)}\,d\varepsilon\le\Psi_m(\xi).\tag{15}$$

Theorem 2. *Let $s\in\mathcal{S}$ be an unknown density to be estimated from a n-sample $(Y_1,\dots,Y_n)$. Consider $\{S_m\}_{m\in\mathcal{M}}$ some at most countable deterministic model collection fulfilling Property $(\boldsymbol{\mathcal{P}})$. Let $\{x_m\}_{m\in\mathcal{M}}$ be some family of nonnegative numbers such that*

$$\sum_{m\in\mathcal{M}}\mathrm{e}^{-x_m}=\Sigma<\infty.\tag{16}$$

For every $m\in\mathcal{M}$, consider Ψ_m defined by $(\boldsymbol{\mathcal{P}})$ and ξ_m such that $\Psi_m(\xi_m)=\sqrt{n}\xi_m^2$.
Let $s_m\in S_m$ such that $KL(s,s_m)\le 2\inf_{t\in S_m}KL(s,t)$ and let $\tau>0$ such that

$$s_m\ge\mathrm{e}^{-\tau}s.\tag{17}$$

Introduce $\{S_m\}_{m\in\widehat{\mathcal{M}}}$ some random subcollection of $\{S_m\}_{m\in\mathcal{M}}$. Let $\rho\ge0$ and consider the collection of ρ-MLEs $\{\hat{s}_m\}_{m\in\widehat{\mathcal{M}}}$:

$$\gamma_n(\hat{s}_m)\le\inf_{t\in S_m}\gamma_n(t)+\rho.$$

Let pen : $\mathcal{M}\mapsto\mathbb{R}_+$. *Suppose that there exists an absolute constant $\kappa>0$ such that, for all $m\in\mathcal{M}$,*

$$\mathrm{pen}(m)\ge\kappa\left(\xi_m^2+(1\vee\tau)\frac{x_m}{n}\right).\tag{18}$$

Let $\rho'\ge0$. Then, any penalized likelihood estimator $\hat{s}_{\hat{m}}$ with $\hat{m}\in\widehat{\mathcal{M}}$ such that

$$\gamma_n(\hat{s}_{\hat{m}})+\mathrm{pen}(\hat{m})\le\inf_{m\in\widehat{\mathcal{M}}}\{\gamma_n(\hat{s}_m)+\mathrm{pen}(m)\}+\rho'\tag{19}$$

satisfies

$$\mathbb{E}\left[d_H^2(s,\hat{s}_{\hat{m}})\right]\le C\left(\mathbb{E}\left[\inf_{m\in\widehat{\mathcal{M}}}\left\{\inf_{t\in S_m}KL(s,t)+\mathrm{pen}(m)\right\}\right]+(1\vee\tau)\frac{\Sigma^2}{n}+\rho+\rho'\right)\tag{20}$$

for some absolute positive constant C.

To prove Theorem 2, we provide an inequality for the moments of order 2 of loglikelihood ratios. This inequality is based on the following Claim 3.

Claim 3. *Let $\tau > 0$. For all $x > 0$, consider $f(x) = x(\ln x)^2$,, $h(x) = x \ln x - x + 1$ and $\phi(x) = e^x - x - 1$. Then, for all $0 < x \leq e^\tau$,*

$$f(x) \leq \frac{\tau^2}{\phi(-\tau)} h(x). \tag{21}$$

Proof. First note that $f(1) = h(1) = 0$, thus we just need to prove (21) for $x \neq 1$. Define

$$\psi : \mathbb{R} \mapsto \mathbb{R},\ y \mapsto \begin{cases} \phi(y)/y^2 & \text{if } y \neq 0, \\ 1/2 & \text{if } y = 0 \end{cases}$$

and

$$\varphi : \mathbb{R} \mapsto \mathbb{R},\ y \mapsto \begin{cases} \phi(y)/y & \text{if } y \neq 0, \\ 0 & \text{if } y = 0. \end{cases}$$

Let us first check that ψ is nondecreasing on $\mathbb{R}$. Since $e^y = 1 + y + y^2/2 + o_{y\to 0}(y^2)$, the functions ψ and φ are continuous on $\mathbb{R}$ and, for $y \neq 0$, $\psi(y) = (\varphi(y) - \varphi(0))/(y - 0)$ is the difference quotient at 0 for φ. Thus, we just need to prove that φ is a convex function to derive that ψ is nondecreasing. By differentiating twice φ, we get that $\varphi''(y) = 2e^y g(y)/y^3$ with $g(y) = 1 - y + y^2/2 - e^{-y}$. The function g is nondecreasing because $g'(y) = -1 + y + e^{-y} \geq 0$. But $g(0) = 0$. So, $g(y) \leq 0$ for all $y \leq 0$ and $g(y) \geq 0$ for all $y \geq 0$. It implies that $\varphi''(y) \geq 0$ for all $y \in \mathbb{R}$ and φ is convex.
Now, let $0 < x \leq e^\tau$, $x \neq 1$. Put $y = -\ln x$. Then, $y \geq -\tau$ and, since ψ is nondecreasing, $\psi(y) \geq \psi(-\tau)$. Moreover, $x \neq 1$, so $y \neq 0$ and $\psi(y) = \phi(y)/y^2$. Thus,

$$\frac{\phi(y)}{y^2} \geq \frac{\phi(-\tau)}{\tau^2}. \tag{22}$$

Taking the definition of ϕ and $y = -\ln x$ into account, (22) leads to

$$\ln x - 1 + \frac{1}{x} \geq \frac{\phi(-\tau)}{\tau^2} (\ln x)^2.$$

We get (21) by multiplying the last inequality by $x > 0$. □

Lemma 1. *Let P and Q be two probability measures with $P \ll Q$. Assume that there exists $\tau > 0$ such that $\ln(\|dP/dQ\|_\infty) \leq \tau$. Then,*

$$\int \left(\ln \frac{dP}{dQ}\right)^2 dP \leq \frac{\tau^2}{e^{-\tau} + \tau - 1}\ KL(P, Q). \tag{23}$$

Proof. Since $\ln(dP/dQ) \leq \tau$, we can apply Claim 3 to $x = dP/dQ$:

$$f\left(\frac{dP}{dQ}\right) \leq \frac{\tau^2}{\phi(-\tau)} h\left(\frac{dP}{dQ}\right).$$

Integrating with respect to Q and taking the definition of f, ϕ and h into account, we get

$$\int \frac{dP}{dQ} \left(\ln \frac{dP}{dQ}\right)^2 dQ \leq \frac{\tau^2}{e^{-\tau} + \tau - 1} \left[\int \ln\left(\frac{dP}{dQ}\right) dP - \int dP + \int dQ\right]$$

thus

$$\int \left(\ln \frac{dP}{dQ}\right)^2 dP \leq \frac{\tau^2}{e^{-\tau}+\tau-1}\ \mathrm{KL}(P,Q).$$

□

Proof of Theorem 2. For the sake of simplicity, we assume that $\rho = \rho' = 0$. For any measurable function g, denote by ν_n the recentred process defined by

$$\nu_n(g) = \frac{1}{n}\sum_{i=1}^{n} \left\{g(Y_i) - \mathbb{E}[g(Y_i)]\right\}. \tag{24}$$

For all $m \in \mathcal{M}$, consider s_m such that $\mathrm{KL}(s, s_m) \leq 2 \inf_{t \in S_m} \mathrm{KL}(s,t)$. Define the functions

$$g_m = -\frac{1}{2}\ln\left(\frac{s_m}{s}\right), \quad \hat{g}_m = -\frac{1}{2}\ln\left(\frac{\hat{s}_m}{s}\right), \quad \hat{f}_m = -\ln\left(\frac{s+\hat{s}_m}{2s}\right). \tag{25}$$

Fix $m \in \mathcal{M}$. Introduce

$$\mathcal{M}(m) = \left\{m' \in \mathcal{M},\ \gamma_n\left(\hat{s}_{m'}\right) + \mathrm{pen}\left(m'\right) \leq \gamma_n\left(\hat{s}_m\right) + \mathrm{pen}\left(m\right)\right\}$$

and let $m' \in \mathcal{M}(m)$. By definition of $\hat{s}_m$,

$$\gamma_n\left(\hat{s}_{m'}\right) + \mathrm{pen}\left(m'\right) \leq \gamma_n\left(\hat{s}_m\right) + \mathrm{pen}\left(m\right) \leq \gamma_n\left(s_m\right) + \mathrm{pen}\left(m\right),$$

and according to the definition of g_m and $\hat{g}_{m'}$,

$$\frac{2}{n}\sum_{i=1}^{n} \hat{g}_{m'}(Y_i) + \mathrm{pen}\left(m'\right) \leq \frac{2}{n}\sum_{i=1}^{n} g_m(Y_i) + \mathrm{pen}\left(m\right). \tag{26}$$

Then, by concavity of the logarithm, we have $\hat{f}_{m'} \leq \hat{g}_{m'}$, and (26) gives

$$\frac{2}{n}\sum_{i=1}^{n} \hat{f}_{m'}(Y_i) \leq \frac{2}{n}\sum_{i=1}^{n} g_m(Y_i) + \mathrm{pen}\left(m\right) - \mathrm{pen}\left(m'\right). \tag{27}$$

By taking (24) and (25) into account, Inequality (27) implies

$$2\,\mathrm{KL}\left(s, \frac{s+\hat{s}_{m'}}{2}\right) \leq \mathrm{KL}\left(s, s_m\right) + \mathrm{pen}\left(m\right) - \mathrm{pen}\left(m'\right) + 2\left[\nu_n\left(g_m\right) - \nu_n\left(\hat{f}_{m'}\right)\right]. \tag{28}$$

Our purpose is now to control both $\nu_n\left(g_m\right)$ and $-\nu_n(\hat{f}_{m'})$.

To bound $-\nu_n(\hat{f}_{m'})$, we refer to the proof of Theorem 7.11 in Massart (2007). It is proved that there exists $\kappa'' > 0$ such that for all $u > 0$, for all $m' \in \mathcal{M}(m)$, for all $y > \xi_{m'}$, the following inequality holds except on a set with probability less than $2e^{-u}$:

$$\frac{-\nu_n(\hat{f}_{m'})}{y^2 + \|\sqrt{s} - \sqrt{\hat{s}_{m'}}\|^2} \leq \kappa'' \left(\frac{\xi_{m'} + \sqrt{u/n}}{y} + \frac{u}{ny^2}\right). \tag{29}$$

Let us now focus on controlling $\nu_n\left(g_m\right)$. From (24) and (25), we have

$$\nu_n(g_m) = \sum_{i=1}^{n} X_i - \mathbb{E}[X_i], \quad X_i := \frac{1}{2n}\ln\left(\frac{s(Y_i)}{s_m(Y_i)}\right). \tag{30}$$

To get an upper bound of $\nu_n(g_m)$, we apply Bernstein's Inequality (Massart, 2007). This inequality requires to control the moments of order k for all $k \geq 2$ of X_i defined by (30). Such a control is provided by Lemma 1 on condition that $\ln(\|s/s_m\|_\infty) \leq \tau$.
Assume that (17) is fulfilled. Then, $\ln(\|s/s_m\|_\infty) \leq \tau$ and we deduce from Lemma 1 that

$$\int_{\mathbb{R}^p} \left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)^2 s(y)\,dy \leq \frac{\tau^2}{e^{-\tau}+\tau-1}\ \mathrm{KL}(s, s_m).$$

On the one hand, $\tau^2/(e^{-\tau}+\tau-1) \sim_{\tau\to\infty} \tau$, so there exists $A > 0$ such that $\tau^2/(e^{-\tau}+\tau-1) \leq 2\tau$ for all $\tau \geq A$. On the other hand, $\tau \mapsto \tau^2/(e^{-\tau}+\tau-1)$ is continuous on $]0, A]$ and $\tau^2/(e^{-\tau}+\tau-1) \sim_{\tau\to 0} 2$, so there exists $B > 0$ such that $\tau^2/(e^{-\tau}+\tau-1) \leq B$ for all $\tau \in]0, A]$. Thus, for all $\tau > 0$, $\tau^2/(e^{-\tau}+\tau-1) \leq \delta(1 \vee \tau)$ with $\delta = 2 \vee B$, and

$$\int_{\mathbb{R}^p} \left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)^2 s(y)\,dy \leq \delta(1 \vee \tau)\,\mathrm{KL}(s, s_m). \tag{31}$$

From (30), (31) and the assumption $\ln(\|s/s_m\|_\infty) \leq \tau$, we derive that

$$\sum_{i=1}^n \mathbb{E}\left[X_i^2\right] \leq \frac{n}{(2n)^2}\int_{\mathbb{R}^p} \left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)^2 s(y)\,dy \leq \frac{\delta(1 \vee \tau)\,\mathrm{KL}(s, s_m)}{4n} \tag{32}$$

and that for all integers ≥ 3,

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}\left[(X_i)_+^k\right] &\leq \frac{n}{(2n)^k}\int_{\mathbb{R}^p} \left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)_+^k s(y)\,dy \\
&\leq \frac{n}{(2n)^k}\int_{\mathbb{R}^p} \left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)^k \mathbb{1}_{\{s(y)\geq s_m(y)\}}\, s(y)\,dy \\
&\leq \frac{n}{(2n)^k}\int_{\mathbb{R}^p} \left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)^{k-2} \left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)^2 \mathbb{1}_{\{s(y)\geq s_m(y)\}}\, s(y)\,dy \\
&\leq \frac{n}{(2n)^k}\,\tau^{k-2}\int_{\mathbb{R}^p} \left(\ln\left(\frac{s(y)}{s_m(y)}\right)\right)^2 \mathbb{1}_{\{s(y)\geq s_m(y)\}}\, s(y)\,dy \\
&\leq \frac{n}{(2n)^k}\,\tau^{k-2}\,\delta(1 \vee \tau)\,\mathrm{KL}(s, s_m) \\
&\leq \frac{1}{2}\left(\frac{\tau}{2n}\right)^{k-2} \frac{\delta(1 \vee \tau)\,\mathrm{KL}(s, s_m)}{2n}.
\end{aligned} \tag{33}$$

From (32) and (33), we can apply Bernstein's Inequality with

$$v := \frac{\delta(1 \vee \tau)\,\mathrm{KL}(s, s_m)}{2n}, \quad c := \frac{\tau}{2n}. \tag{34}$$

It gives that, for every positive u, except on a set with probability less than e^{-u},

$$\nu_n(g_m) \leq \sqrt{2vu} + cu. \tag{35}$$

Let $z > 0$ to be chosen later. Using that $z^2 + \mathrm{KL}(s, s_m) \geq 2z\sqrt{\mathrm{KL}(s, s_m)}$ and $z^2 + \mathrm{KL}(s, s_m) \geq z^2$, we get from (35) that except on a set with probability less than e^{-u},

$$\frac{\nu_n(g_m)}{z^2+\mathrm{KL}(s,s_m)} \leq \frac{\sqrt{2vu}+cu}{z^2+\mathrm{KL}(s,s_m)} \leq \frac{\sqrt{vu}}{z\sqrt{2\,\mathrm{KL}(s,s_m)}}+\frac{cu}{z^2}. \tag{36}$$

Let us gather (29) and (36): There exists $\kappa''>0$ such that, for every positive u, for all $m'\in\mathcal{M}(m)$, for all $z>0$ and for all $y\geq\xi_{m'}$, except on a set with probability less than $3e^{-u}$,

$$\frac{-\nu_n\left(\hat{f}_{m'}\right)}{y^2+\|\sqrt{s}-\sqrt{\hat{s}_{m'}}\|^2} \leq \kappa''\left(\frac{\xi_{m'}}{y}+\frac{\sqrt{u/n}}{y}+\frac{u}{ny^2}\right) \tag{37}$$

and

$$\frac{\nu_n(g_m)}{z^2+\mathrm{KL}(s,s_m)} \leq \frac{\sqrt{vu}}{z\sqrt{2\,\mathrm{KL}(s,s_m)}}+\frac{cu}{z^2}. \tag{38}$$

Now, let $x>0$. Let x_m and $x_{m'}$ be defined by (16). We apply (37) and (38) to $u=x+x_m+x_{m'}$ and we choose adequately y and z by defining for some constants γ and β to be specified later,

$$y_{m,m'} := \gamma^{-1}\sqrt{\xi_{m'}^2+\frac{x+x_m+x_{m'}}{n}}\ . \tag{39}$$

and

$$z_{m,m'} := \beta^{-1}\sqrt{\left(\frac{v}{2\,\mathrm{KL}(s,s_m)}+c\right)(x+x_m+x_{m'})}. \tag{40}$$

Using that $a^2+b^2\geq a^2$, we get that except on a set with probability less than $3e^{-(x+x_m+x_{m'})}$,

$$-\nu_n(\hat{f}_{m'}) \leq \kappa''(2\gamma+\gamma^2)\left(y_{m,m'}^2+\left\|\sqrt{s}-\sqrt{\hat{s}_{m'}}\right\|^2\right) \tag{41}$$

and

$$\nu_n(g_m) \leq (\beta+\beta^2)(z_{m,m'}^2+\mathrm{KL}(s,s_m)). \tag{42}$$

We can now come back to Inequality (28). Injecting (41) and (42) into (28) yields

$$\begin{aligned} 2\,\mathrm{KL}\left(s,\frac{s+\hat{s}_{m'}}{2}\right) \quad \leq \quad & \mathrm{KL}\left(s,s_m\right)+\mathrm{pen}\left(m\right)-\mathrm{pen}\left(m'\right) \\ & +2(\beta+\beta^2)[z_{m,m'}^2+\mathrm{KL}(s,s_m)] \\ & +2\kappa''(2\gamma+\gamma^2)\left(y_{m,m'}^2+\left\|\sqrt{s}-\sqrt{\hat{s}_{m'}}\right\|^2\right). \end{aligned}$$

Putting $\kappa(\beta) := (1+2(\beta+\beta^2))$, using the inequality $\mathrm{KL}(s,(s+\hat{s}_{m'})/2)\geq(2\ln 2-1)\|\sqrt{s}-\sqrt{\hat{s}_{m'}}\|^2$ (Massart, 2007, Lemma 7.23) and choosing γ such that $2\kappa''(2\gamma+\gamma^2)=2\ln 2-1:=\alpha$, we get

$$\alpha\left\|\sqrt{s}-\sqrt{\hat{s}_{m'}}\right\|^2 \leq \kappa(\beta)\,\mathrm{KL}\left(s,s_m\right)+\mathrm{pen}\left(m\right)-\mathrm{pen}\left(m'\right)+2(\beta+\beta^2)z_{m,m'}^2+\alpha y_{m,m'}^2.$$

From (40), (39) and (34), we deduce that

$$\begin{aligned}\alpha\left\|\sqrt{s}-\sqrt{\hat{s}_{m'}}\right\|^2 \leq\ & \kappa(\beta)\,\mathrm{KL}(s,s_m)+\mathrm{pen}(m)-\mathrm{pen}(m')\\ &+(\beta+\beta^2)\beta^{-2}\left(\frac{\delta(1\vee\tau)}{2}+\tau\right)\frac{x+x_m+x_{m'}}{n}\\ &+\alpha\gamma^{-2}\left(\xi_{m'}^2+\frac{x+x_m+x_{m'}}{n}\right).\end{aligned}$$

Since $\tau\leq 1\vee\tau$, if we choose β such that $(\beta+\beta^2)(\delta/2+1)=\alpha\gamma^{-2}$, we get

$$\begin{aligned}\alpha\left\|\sqrt{s}-\sqrt{\hat{s}_{m'}}\right\|^2 \leq\ & \kappa(\beta)\,\mathrm{KL}(s,s_m)+\mathrm{pen}(m)-\mathrm{pen}(m')\\ &+\alpha\gamma^{-2}\xi_{m'}^2+\alpha\gamma^{-2}\left[\beta^{-2}(1\vee\tau)+1\right]\frac{x+x_m+x_{m'}}{n}.\end{aligned}$$

Put $\kappa=\alpha\gamma^{-2}(\beta^{-2}+1)$. Then, since $1\leq 1\vee\tau$,

$$\begin{aligned}\alpha\left\|\sqrt{s}-\sqrt{\hat{s}_{m'}}\right\|^2 \leq\ & \kappa(\beta)\,\mathrm{KL}(s,s_m)+\mathrm{pen}(m)-\mathrm{pen}(m')\\ &+\alpha\gamma^{-2}\xi_{m'}^2+\kappa\,(1\vee\tau)\frac{x+x_m+x_{m'}}{n}\\ \leq\ & \kappa(\beta)\,\mathrm{KL}(s,s_m)+\left[\mathrm{pen}(m)+\kappa\ (1\vee\tau)\frac{x_m}{n}\right]\\ &+\left[\alpha\gamma^{-2}\xi_{m'}^2+\kappa(1\vee\tau)\frac{x_{m'}}{n}-\mathrm{pen}(m')\right]+\kappa(1\vee\tau)\frac{x}{n}\\ \leq\ & \kappa(\beta)\,\mathrm{KL}(s,s_m)+\left[\mathrm{pen}(m)+\kappa\,(1\vee\tau)\frac{x_m}{n}\right]\\ &+\left[\kappa\left(\xi_{m'}^2+(1\vee\tau)\frac{x_{m'}}{n}\right)-\mathrm{pen}(m')\right]+\kappa(1\vee\tau)\frac{x}{n}.\end{aligned}$$

Now, assume that Condition (18) on the penalty function is fulfilled for this value of κ. Then, for all $x>0$, for every $m\in\mathcal{M}$ and $m'\in\mathcal{M}(m)$, except on a set with probability less than $3e^{-(x+x_m+x_{m'})}$,

$$\alpha\left\|\sqrt{s}-\sqrt{\hat{s}_{m'}}\right\|^2\leq\kappa(\beta)\,\mathrm{KL}(s,s_m)+2\,\mathrm{pen}(m)+\kappa(1\vee\tau)\frac{x}{n}. \tag{43}$$

It only remains to sum up the tail bounds (43) over all the possible values of $m\in\mathcal{M}$ and $m'\in\mathcal{M}(m)$ by taking the union of the different sets of probability less than $3e^{-(x+x_m+x_{m'})}$. For all $x>0$, except on a set with probability less than

$$\begin{aligned}3\sum_{m\in\mathcal{M},m'\in\mathcal{M}(m)}e^{-(x+x_m+x_{m'})} &\leq 3e^{-x}\sum_{(m,m')\in\mathcal{M}\times\mathcal{M}}e^{-(x_m+x_{m'})}\\ &= 3e^{-x}\left(\sum_{m\in\mathcal{M}}e^{-x_m}\right)^2=3\Sigma^2e^{-x},\end{aligned}$$

we have simultaneously for all $m\in\mathcal{M}$ and $m'\in\mathcal{M}(m)$,

$$\alpha\left\|\sqrt{s}-\sqrt{\hat{s}_{m'}}\right\|^2\leq\kappa(\beta)\,\mathrm{KL}(s,s_m)+2\,\mathrm{pen}(m)+\kappa(1\vee\tau)\frac{x}{n}. \tag{44}$$

Inequality (44) is in particular satisfied for all $m \in \widehat{\mathcal{M}}$ and $m' \in \widehat{\mathcal{M}}(m)$ and, since $\hat{m}$ defined by (19) belongs to $\widehat{\mathcal{M}}(m)$ for all $m \in \widehat{\mathcal{M}}$, we deduce from (44) that for all $x > 0$, except on a set with probability less than $3\Sigma^2 e^{-x}$,

$$\begin{aligned} \alpha \left\| \sqrt{s} - \sqrt{\hat{s}_{\hat{m}}} \right\|^2 &\leq \inf_{m \in \widehat{\mathcal{M}}} \{\kappa(\beta)\, \mathrm{KL}(s, s_m) + 2\,\mathrm{pen}(m)\} + \kappa(1 \vee \tau)\frac{x}{n} \\ &\leq \inf_{m \in \widehat{\mathcal{M}}} \left\{ 2\kappa(\beta) \inf_{t \in S_m} \mathrm{KL}(s,t) + 2\,\mathrm{pen}(m) \right\} + \kappa(1 \vee \tau)\frac{x}{n}\ . \end{aligned} \tag{45}$$

By integrating (45) over $x > 0$, we finally get that there exists an absolute constant $C > 0$ such that

$$\mathbb{E}\left[\left\| \sqrt{s} - \sqrt{\hat{s}_{\hat{m}}} \right\|^2 \right] \leq C \left(\mathbb{E}\left[\inf_{m \in \widehat{\mathcal{M}}} \left\{ \inf_{t \in S_m} \mathrm{KL}(s,t) + \mathrm{pen}(m) \right\} \right] + (1 \vee \tau)\frac{\Sigma^2}{n} \right).$$

□

B Sketch of the proof of Theorem 1

To deduce Theorem 1 from Theorem 2, the control of the entropy bracketing of $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}$ for the Hellinger distance is required. For that, the proof of Maugis and Michel (2011b) is adapted for our specific mixtures. Next, a function $\psi_{(K,J)}$ fulfilling Property ($\boldsymbol{\mathcal{P}}$) is deduced. We just give a sketch of the proof of Theorem 1 and we refer to Maugis and Michel (2011b) for more details.

B.1 Control of the entropy bracketing $\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}, d_H)$

To apply Theorem 2, the first step is to control the bracketing entropy of the Gaussian mixture families $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}$. Note that Theorem 2 only requires to control the local bracketing entropy $\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}(u,\xi), d_H)$. Nevertheless, it is difficult to characterize the subset $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}(u,\xi)$ in function of the parameters of its mixtures. Thus, we rather control the global entropy bracketing $\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}, d_H)$, which is sufficient since the local bracketing entropy is upper bounded by the global bracketing entropy.

Proposition 1. *Put $D_{(K,\boldsymbol{J_r})} = K(1 + |\boldsymbol{J_r}|)$. For all $\varepsilon \in (0, 1]$,*

$$\mathcal{N}_{[.]}\left(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}, d_H\right) \leq C(A_\mu, A_\sigma, a_\sigma, K, \boldsymbol{J_r}, p) \left(\frac{1}{\varepsilon}\right)^{D_{(K,\boldsymbol{J_r})}}$$

with

$$C(A_\mu, A_\sigma, a_\sigma, K, \boldsymbol{J_r}, p) := 4\,(2\pi\mathrm{e})^{\frac{K}{2}}\, 3^{K-1} \left(\frac{A_\sigma}{a_\sigma} + \frac{1}{2}\right) \left(\frac{2^{\frac{5}{4}} A_\mu}{\sqrt{c'}a_\sigma}\right)^{K|\boldsymbol{J_r}|} K(3\sqrt{c}p)^{D_{(K,\boldsymbol{J_r})}}, \tag{46}$$

$c = sh(1) + 49/128$ and $c' = 5(1 - 2^{-1/4})/8$.
Hence,

$$\mathcal{H}_{[.]}\left(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}, d_H\right) \leq \ln\left[C(A_\mu, A_\sigma, a_\sigma, K, \boldsymbol{J_r}, p)\right] + D_{(K,\boldsymbol{J_r})} \ln\left(\frac{1}{\varepsilon}\right).$$

Proof. The key idea is that the control of the bracketing entropy of $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}$ can be recast into the control of the bracketing entropies of the associated mixture component density families. Specifically, from (8), each mean vector $\boldsymbol{\mu_k}$ of a p-dimensional Gaussian mixture density in $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}$ can be decomposed into a $|\boldsymbol{J^c_r}|$-dimensional null mean vector and a $|\boldsymbol{J_r}|$-dimensional free mean vector:

$$\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})} = \left\{ \begin{array}{l} \sum_{k=1}^{K} \pi_k \, \Phi\left(\cdot \mid \boldsymbol{\mu_k}, \sigma^2 I\right); \\ \forall\, k : \boldsymbol{\mu_{k[J_r]}} \in [-A_\mu, A_\mu]^{|\boldsymbol{J_r}|}, \quad \boldsymbol{\mu_{k[J^c_r]}} = \boldsymbol{0}, \\ \forall\, k : \pi_k > 0, \sum_{k=1}^{K} \pi_k = 1, \quad \sigma \in [a_\sigma, A_\sigma] \end{array} \right\}. \tag{47}$$

Consider the $(K-1)$-dimensional simplex Π_K defined by

$$\Pi_K := \left\{ (\pi_1, \ldots, \pi_K) \in (0,1)^K ; \sum_{k=1}^{K} \pi_k = 1 \right\}$$

and the family of K-tuples of p-dimensional Gaussian densities

$$\mathcal{F}_{(K,\boldsymbol{J_r})} = \left\{ \begin{array}{l} \left(\Phi\left(\cdot \mid \boldsymbol{\mu_1}, \sigma^2 I\right), \ldots, \Phi\left(\cdot \mid \boldsymbol{\mu_K}, \sigma^2 I\right)\right); \\ \forall\, k : \boldsymbol{\mu_{k[J_r]}} \in [-A_\mu, A_\mu]^{|\boldsymbol{J_r}|}, \quad \boldsymbol{\mu_{k[J^c_r]}} = \boldsymbol{0}, \\ \sigma \in [a_\sigma, A_\sigma] \end{array} \right\}.$$

Following the arguments developed by Maugis (2008) (proof of Proposition 7.A.2), it is easy to show that the study of the bracketing entropy of $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}$ can be recast into the study of the bracketing entropy of Π_K and $\mathcal{F}_{(K,\boldsymbol{J_r})}$:

Lemma 2. *For all* $\varepsilon \in (0,1]$,

$$\mathcal{N}_{[.]}\left(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}, d_H\right) \leq \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \Pi_K, d_H\right) \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{F}_{(K,\boldsymbol{J_r})}, d_H\right)$$

where

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

From Lemma 2, all the matter is to calculate an upper bound of the bracketing entropy of $\mathcal{F}_{(K,\boldsymbol{J_r})}$.

Let $\boldsymbol{f} = (f_1, \ldots, f_K) := \left(\Phi\left(\cdot \mid \boldsymbol{\mu_1}, \sigma^2 I\right), \ldots, \Phi\left(\cdot \mid \boldsymbol{\mu_K}, \sigma^2 I\right)\right) \in \mathcal{F}_{(K,\boldsymbol{J_r})}$. We want to find an ε-bracket for $\boldsymbol{f}$. We shall consider shrunk and dilated Gaussian densities.

Step 1. Construction of a net for the variance

Let $\delta \in (0,1]$ to be chosen later. Let $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. We construct a regular net for the variance $\sigma^2 \in [a^2_\sigma, A^2_\sigma]$. For $l \in \{2, \ldots, r\}$, we define $\sigma^2_l = (1+\delta)^{1-\frac{l}{2}} A^2_\sigma$ where

$$r = \left\lceil 4 \, \frac{\ln\left(\frac{A_\sigma}{a_\sigma}\sqrt{1+\delta}\right)}{\ln(1+\delta)} \right\rceil \tag{48}$$

is chosen so that $\sigma^2_r < a^2_\sigma < \sigma^2_{r-1} \leq \ldots \leq \sigma^2_2 = A^2_\sigma$.

Step 2. Construction of a net for the mean vectors
Let l be the unique integer in $\{2,\dots,r\}$ such that $\sigma_{l+1}^2 < \sigma^2 \leq \sigma_l^2$. For all $k \in \{1,\dots,K\}$, let $\boldsymbol{\nu_k} \in \mathbb{R}^p$ to be specified later. Consider the functions defined on $\mathbb{R}^p$ by

$$\begin{cases} l_k(\boldsymbol{y}) = (1+\delta)^{-p}\,\Phi\left(\boldsymbol{y} \mid \boldsymbol{\nu_k}, (1+\delta)^{-\frac{1}{4}}\sigma_{l+1}^2 I\right) \\ u_k(\boldsymbol{y}) = (1+\delta)^{p}\,\Phi\left(\boldsymbol{y} \mid \boldsymbol{\nu_k}, (1+\delta)\sigma_l^2 I\right). \end{cases}$$

Put $\boldsymbol{l} = (l_1,\dots,l_K)$ and $\boldsymbol{u} = (u_1,\dots,u_K)$. We now determine δ and $(\boldsymbol{\nu_1},\dots,\boldsymbol{\nu_K})$ so that $\boldsymbol{l}$ and $\boldsymbol{u}$ form an ε-bracket for $\boldsymbol{f}$. On the one hand, by using the calculation of the Hellinger distance between two multivariate Gaussian densities (Maugis and Michel, 2011b, Corollary 3) and by upper bounding some usual functions, we get that, for all $k \in \{1,\dots,K\}$, $d_H^2(l_k,u_k) \leq cp^2\delta^2$ where $c = \mathrm{sh}(1) + 49/128$. Thus, we take $\delta = \varepsilon/(\sqrt{c}p)$ so that $d_H(l_k,u_k) \leq \varepsilon$. On the other hand, by using the ratio of two multivariate Gaussian densities (Maugis and Michel, 2011b, Corollary 2), the definition of σ_l and σ_{l+1}, the inequality $\ln(1+\delta) \geq \delta/2$ for all $\delta \in (0,1]$ and the concavity of $\delta \mapsto 1-(1+\delta)^{-1/4}$, we get that a sufficient condition for $l_k \leq f_k \leq u_k$ for all $k \in \{1,\dots,K\}$ is

$$\|\boldsymbol{\mu_k} - \boldsymbol{\nu_k}\|_2^2 \leq c'p\delta^2(1+\delta)^{\frac{2-l}{2}}A_\sigma^2 \tag{49}$$

where $c' = 5(1-2^{-1/4})/8$. Put

$$U_l := \mathbb{Z} \cap \left[\left\lfloor \frac{-A_\mu}{\sqrt{c'}\delta(1+\delta)^{\frac{2-l}{4}}A_\sigma} \right\rfloor, \left\lfloor \frac{A_\mu}{\sqrt{c'}\delta(1+\delta)^{\frac{2-l}{4}}A_\sigma} \right\rfloor\right]. \tag{50}$$

For all $k \in \{1,\dots,K\}$, for all $j \in \boldsymbol{J_r}$, choose

$$u_{kj}^{(l)} = \mathrm{argmin}_{v_{kj} \in U_l} |\mu_{kj} - \sqrt{c'}\delta(1+\delta)^{\frac{2-l}{4}}A_\sigma v_{kj}|.$$

Define $\boldsymbol{\nu_k}^{(l)} := \left(\nu_{k1}^{(l)},\dots,\nu_{kp}^{(l)}\right) \in [-A_\mu, A_\mu]^p$ by

$$\begin{aligned} &\forall\, j \in \boldsymbol{J_r^c}, \quad \nu_{kj}^{(l)} = 0, \\ &\forall\, j \in \boldsymbol{J_r}, \quad \nu_{kj}^{(l)} = \sqrt{c'}\delta(1+\delta)^{\frac{2-l}{4}}A_\sigma u_{kj}^{(l)}. \end{aligned}$$

Then, $\boldsymbol{\nu_k}^{(l)}$ fulfills (49) and we get a net for the mean vectors.

Step 3. Upper bound of the number of ε-brackets for $\mathcal{F}_{(K,\boldsymbol{J_r})}$
From Step 1 and Step 2, the family

$$\mathcal{B}_\varepsilon\left(\mathcal{F}_{(K,\boldsymbol{J_r})}\right) = \left\{ \begin{array}{l} [\boldsymbol{l},\boldsymbol{u}] := \{[l_1,u_1],\dots,[l_K,u_K]\}\,;\ \forall\, k \in \{1,\dots,K\}: \\ l_k = (1+\delta)^{-p}\,\Phi\left(\cdot \mid \left(\nu_{k1}^{(l)},\dots,\nu_{kp}^{(l)}\right), (1+\delta)^{-\frac{1}{4}}\sigma_{l+1}^2 I\right) \\ u_k = (1+\delta)^{p}\,\Phi\left(\cdot \mid \left(\nu_{k1}^{(l)},\dots,\nu_{kp}^{(l)}\right), (1+\delta)\sigma_l^2 I\right) \\ \text{with} \begin{cases} \sigma_l^2 = (1+\delta)^{1-\frac{l}{2}}A_\sigma^2, \quad l \in \{2,\dots,r\}, \\ \forall\, j \in \boldsymbol{J_r},\ \nu_{kj}^{(l)} = \sqrt{c'}\delta(1+\delta)^{\frac{2-l}{4}}A_\sigma u_{kj}^{(l)}, \quad u_{kj}^{(l)} \in U_l \\ \forall\, j \in \boldsymbol{J_r^c},\ \nu_{kj}^{(l)} = 0 \end{cases} \end{array} \right\}$$

is an ε-bracket covering for $\mathcal{F}_{(K,\boldsymbol{J_r})}$. Therefore, an upper bound of the number of ε-brackets necessary to cover $\mathcal{F}_{(K,\boldsymbol{J_r})}$ is deduced from an upper bound of the cardinal of $\mathcal{B}_\varepsilon(\mathcal{F}_{(K,\boldsymbol{J_r})})$. From (48) and (50), we have

$$\begin{aligned}
\left|\mathcal{B}_\varepsilon\left(\mathcal{F}_{(K,\boldsymbol{J_r})}\right)\right| &\leq \sum_{l=2}^{r} \prod_{(k,j)\in\{1,\ldots,K\}\times\boldsymbol{J_r}} \left(\frac{A_\mu}{\sqrt{c'}\delta(1+\delta)^{\frac{2-l}{4}}A_\sigma}\right)\\
&\leq \left(\frac{2A_\mu}{\sqrt{c'}\delta A_\sigma}\right)^{K|\boldsymbol{J_r}|}\sum_{l=2}^{r}(1+\delta)^{\frac{(l-2)K|\boldsymbol{J_r}|}{4}}\\
&\leq \left(\frac{2A_\mu}{\sqrt{c'}\delta A_\sigma}\right)^{K|\boldsymbol{J_r}|}(r-1)(1+\delta)^{\frac{(r-2)K|\boldsymbol{J_r}|}{4}}.
\end{aligned}$$

From (48), $(1+\delta)^{(r-2)/4} \leq (1+\delta)^{1/4}A_\sigma/a_\sigma \leq 2^{1/4}A_\sigma/a_\sigma$ and $r-1 \leq 4(A_\sigma/a_\sigma + 1/2)/\delta$, so

$$\begin{aligned}
\left|\mathcal{B}_\varepsilon\left(\mathcal{F}_{(K,\boldsymbol{J_r})}\right)\right| &\leq 4\left(\frac{2^{5/4}A_\mu}{\sqrt{c'}a_\sigma}\right)^{K|\boldsymbol{J_r}|}\left(\frac{A_\sigma}{a_\sigma}+\frac{1}{2}\right)\delta^{-(1+K|\boldsymbol{J_r}|)}\\
&\leq 4\left(\frac{2^{5/4}A_\mu}{\sqrt{c'}a_\sigma}\right)^{K|\boldsymbol{J_r}|}\left(\frac{A_\sigma}{a_\sigma}+\frac{1}{2}\right)\left(\frac{\sqrt{c}p}{\varepsilon}\right)^{1+K|\boldsymbol{J_r}|}. \qquad (51)
\end{aligned}$$

Finally, Proposition 1 is derived from Lemma 2 and (51). □

B.2 Determination of a function $\Psi_{(K,\boldsymbol{J_r})}$

This section is devoted to the determination of a function $\Psi_{(K,\boldsymbol{J_r})}$ defined by Property **($\boldsymbol{\mathcal{P}}$)**. From Proposition 1, for all $\xi > 0$,

$$\begin{aligned}
\int_0^\xi \sqrt{\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}, d_H)}\,d\varepsilon \quad &\leq \quad \xi\sqrt{\ln\left(C(A_\mu, A_\sigma, a_\sigma, K, \boldsymbol{J_r}, p)\right)}\\
&\quad +\sqrt{D_{(K,\boldsymbol{J_r})}}\int_0^{\xi\wedge 1}\sqrt{\ln\left(\frac{1}{\varepsilon}\right)}\,d\varepsilon\ .
\end{aligned}$$

In order to control the last term of the right-hand side of the last inequality, we apply the following technical result taken from Maugis and Michel (2011b):

Lemma 3. (Maugis and Michel, 2011b) For all $\xi \in]0,1]$,

$$\int_0^\xi \sqrt{\ln\left(\frac{1}{\varepsilon}\right)}\,d\varepsilon \leq \xi\left[\sqrt{\pi}+\sqrt{\ln\left(\frac{1}{\xi}\right)}\right].$$

We obtain

$$\begin{aligned}
&\int_0^\xi \sqrt{\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}, d_H)}\,d\varepsilon\\
&\leq \xi\sqrt{\ln\left(C(A_\mu, A_\sigma, a_\sigma, K, \boldsymbol{J_r}, p)\right)} + \sqrt{D_{(K,\boldsymbol{J_r})}}\ (\xi\wedge 1)\left[\sqrt{\pi}+\sqrt{\ln\left(\frac{1}{\xi\wedge 1}\right)}\right]\\
&\leq \sqrt{D_{(K,\boldsymbol{J_r})}}\ \xi\left[\sqrt{\pi}+\sqrt{\frac{\ln\left(C(A_\mu, A_\sigma, a_\sigma, K, \boldsymbol{J_r}, p)\right)}{D_{(K,\boldsymbol{J_r})}}}+\sqrt{\ln\left(\frac{1}{\xi\wedge 1}\right)}\right].
\end{aligned}$$

But from (46) and the fact that $D_{(K,\boldsymbol{J_r})} = K(1+|\boldsymbol{J_r}|)$, we have

$$\begin{aligned}
&\ln\left(C(A_\mu, A_\sigma, a_\sigma, K, \boldsymbol{J_r}, p)\right)\\
&\le \ln 4 + \frac{K}{2}\ln(2\pi \mathrm{e}) + (K-1)\ln 3 + \ln\left(\frac{A_\sigma}{a_\sigma}+\frac{1}{2}\right) + K|\boldsymbol{J_r}|\ln\left(\frac{2^{5/4}A_\mu}{\sqrt{c'}a_\sigma}\right) + \ln K\\
&\quad + D_{(K,\boldsymbol{J_r})}\ln(3\sqrt{c}p)\\
&\le \left[\ln 4 + \frac{\ln(2\pi\mathrm{e})}{2} + \ln 3 + \ln\left(\frac{A_\sigma}{a_\sigma}+\frac{1}{2}\right) + \ln\left(\frac{2^{5/4}A_\mu}{\sqrt{c'}a_\sigma}\right) + 1 + \ln(3\sqrt{c}p)\right] D_{(K,\boldsymbol{J_r})}\\
&\le \left[\ln\left(\frac{72\sqrt{2\pi\mathrm{e}}\,2^{5/4}\mathrm{e}\sqrt{c}}{\sqrt{c'}}\right) + \ln\left[\frac{A_\sigma}{a_\sigma}\left(1+\frac{A_\mu}{a_\sigma}\right)\right] + \ln p\right] D_{(K,\boldsymbol{J_r})}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\int_0^\xi \sqrt{\mathcal{H}_{[.]}(\varepsilon, \mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}, d_H)}\,d\varepsilon\\
&\le \sqrt{D_{(K,\boldsymbol{J_r})}}\;\xi\left[\sqrt{\pi} + \sqrt{\ln\left(\frac{72\sqrt{2\pi\mathrm{e}}\,2^{5/4}\mathrm{e}\sqrt{c}}{\sqrt{c'}}\right) + \ln\left[\frac{A_\sigma}{a_\sigma}\left(1+\frac{A_\mu}{a_\sigma}\right)\right] + \ln p} + \sqrt{\ln\left(\frac{1}{\xi\wedge 1}\right)}\right]\\
&\le \sqrt{D_{(K,\boldsymbol{J_r})}}\;\xi\left[6 + \sqrt{\ln\left[\frac{A_\sigma}{a_\sigma}\left(1+\frac{A_\mu}{a_\sigma}\right)\right]} + \sqrt{\ln p} + \sqrt{\ln\left(\frac{1}{\xi\wedge 1}\right)}\right].
\end{aligned}$$

Consequently, by putting

$$B(A_\mu, A_\sigma, a_\sigma, p) := 6 + \sqrt{\ln\left[\frac{A_\sigma}{a_\sigma}\left(1+\frac{A_\mu}{a_\sigma}\right)\right]} + \sqrt{\ln p},$$

we get that the function $\Psi_{(K,\boldsymbol{J_r})}$ defined on $\mathbb{R}_+^\star$ by

$$\Psi_{(K,\boldsymbol{J_r})}(\xi) = \sqrt{D_{(K,\boldsymbol{J_r})}}\;\xi\left[B(A_\mu, A_\sigma, a_\sigma, p) + \sqrt{\ln\left(\frac{1}{\xi\wedge 1}\right)}\right]$$

satisfies (15). Besides, $\Psi_{(K,\boldsymbol{J_r})}$ is nondecreasing and $\xi \mapsto \Psi_{(K,\boldsymbol{J_r})}(\xi)/\xi$ is nonincreasing, so $\Psi_{(K,\boldsymbol{J_r})}$ is convenient.

B.3 Lower bound of the penalty function

Finally, according to the lower bound (18) of the penalty function, we need to find an upper bound of ξ_* satisfying $\Psi_{(K,\boldsymbol{J_r})}(\xi_*) = \sqrt{n}\,\xi_*^2$ and to calculate the weights $x_{(K,\boldsymbol{J_r})}$ to take into account the richness of the family $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}$. This can be done along the proofs of Maugis and Michel (2011b) by replacing the dimension of the models considered by Maugis and Michel (2011b) by the dimension $D_{(K,\boldsymbol{J_r})}$ of our models $\mathcal{S}^{\mathcal{B}}_{(K,\boldsymbol{J_r})}$. This leads to the two following lemmas:

Lemma 4. *Consider ξ_* such that $\Psi_{(K,\boldsymbol{J_r})}(\xi_*) = \sqrt{n}\,\xi_*^2$. Then,*

$$\xi_*^2 \le \frac{D_{(K,\boldsymbol{J_r})}}{n}\left[2B^2(A_\mu, A_\sigma, a_\sigma, p) + \ln\left(\frac{1}{1\wedge B^2(A_\mu, A_\sigma, a_\sigma, p)\frac{D_{(K,\boldsymbol{J_r})}}{n}}\right)\right]. \tag{52}$$

Lemma 5. *Consider the weight family* $\left\{x_{(K,\boldsymbol{J_r})}\right\}_{(K,\boldsymbol{J_r})\in\mathcal{M}}$ *defined by*

$$x_{(K,\boldsymbol{J_r})} = D_{(K,\boldsymbol{J_r})} \ln\left(\frac{8\mathrm{e}p}{D_{(K,\boldsymbol{J_r})} \wedge p}\right). \tag{53}$$

Then, we have $\sum_{(K,\boldsymbol{J_r})\in\mathcal{M}} \mathrm{e}^{-x_{(K,\boldsymbol{J_r})}} \leq 1.$

From (18), (52) and (53), we can apply Theorem 2 as soon as there exists $\kappa > 0$ such that $\mathrm{pen}\,(K,\boldsymbol{J_r})$ satisfies for all $(K,\boldsymbol{J_r}) \in \mathcal{M}$:

$$\mathrm{pen}\,(K,\boldsymbol{J_r}) \geq \kappa\frac{D_{(K,\boldsymbol{J_r})}}{n}\left[2B^2(A_\mu, A_\sigma, a_\sigma, p) + \ln\left(\frac{1}{1 \wedge B^2(A_\mu, A_\sigma, a_\sigma, p)\frac{D_{(K,\boldsymbol{J_r})}}{n}}\right)\right.$$
$$\left. +(1\vee\tau)\ln\left(\frac{8\mathrm{e}p}{D_{(K,\boldsymbol{J_r})}\wedge p}\right)\right].$$

Applying Theorem 2 leads to Theorem 1. □

C Grid of regularization parameters

Pan and Shen (2007) propose to use a deterministic regular grid, similar for each number K of mixture components. In order to find the maximum value of regularization parameters, they conduct a few Lasso algorithms by increasing the value of λ until obtaining a model whose mean parameters all equal zero but this method can reveal quite time-consuming. Moreover, the choice of the grid step is difficult. To construct a grid of regularization parameters, another approach is here considered. The key idea is to construct a data-driven grid G_K (depending on K) of regularization parameters by using the updating formulas of the mixture parameters estimation in the EM algorithm computing the Lasso solutions. More precisely, first the EM algorithm is used with $\lambda = 0$ to determine the ML estimator $(\hat{\pi}_1^0, \ldots, \hat{\pi}_K^0, \hat{\boldsymbol{\mu}}_1^0, \ldots, \hat{\boldsymbol{\mu}}_K^0, \hat{\sigma}^0)$. Then for all $\lambda > 0$ the parameter estimation at the r-th iteration of the EM algorithm is considered: for all $k \in \{1,\ldots,K\}$, for all $j \in \{1,\ldots,p\}$,

$$\pi_k^{(r+1)} = \frac{1}{n}\sum_{i=1}^n \tau_{ik}^{(r)}, \tag{54}$$

$$\mu_{kj}^{(r+1)} = \mathrm{sign}\left({\nu_{kj}}^{(r+1)}\right)\max\left(\left|{\nu_{kj}}^{(r+1)}\right| - \frac{\lambda\,{\sigma^2}^{(r)}}{\pi_k^{(r+1)}}, 0\right) \tag{55}$$

where $\nu_{kj}^{(r+1)} = \sum_{i=1}^n \tau_{ik}^{(r)} Y_{ij} / \sum_{i=1}^n \tau_{ik}^{(r)}$ and sign(.) is the sign function. From (54) and (55), we remark that

$$\mu_{kj}^{(r+1)} = 0 \Longleftrightarrow \lambda \geq \frac{{\pi_k}^{(r+1)}}{{\sigma^2}^{(r)}}\left|{\nu_{kj}}^{(r+1)}\right|.$$

Moreover, we hope that for all $k \in \{1,\ldots,K\}$ and for all $j \in \{1,\ldots,p\}$, the values $\pi_k^{(r+1)}$, $\sigma^{(r)}$ and $\nu_{kj}^{(r+1)}$ are not too far from the estimates $\hat{\pi}_k^0$, $\hat{\sigma}^0$ and $\hat{\mu}_{kj}^0$ respectively. Thus we propose the grid $G_K = \{0, \lambda_{11}, \ldots, \lambda_{1p}, \ldots, \lambda_{K1}, \ldots, \lambda_{Kp}, \lambda_{\text{extra}}\}$ where

$$\lambda_{kj} := \frac{\hat{\pi}_k^0}{\hat{\sigma}^{2\,0}}\left|\hat{\mu}_{kj}^0\right|,$$

and $\lambda_{\text{extra}} = 2\max_{k,j} \lambda_{kj}$ in order to ensure a mixture with very sparse mean component vectors (the factor 2 is quite arbitrary; in practice the algorithm stops as soon as the null solution is reached, even is $\lambda = 2$ is not reached yet). In practice, this method is time-efficient since it only requires to run one EM algorithm for $\lambda = 0$.

References

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

Arlot, S. and Bach, F. (2010). Data-driven calibration of linear estimators with minimal penalties. *Advances in Neural Information Processing Systems (NIPS)*, **54**, 22–46.

Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, **10**, 245–279 (electronic).

Auder, B. and Fischer, A. (2011). Projection-based curve clustering. *Journal of Statistical Computation and Simulation*. To appear.

Bach, F. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM.

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., Maugis, C., and Michel, B. (2011). Slope heuristics: overview and implementation. *Statistics and Computing*, **22**(2), 455–470.

Bertin, K., Le Pennec, E., and Rivoirard, V. (2011). Adaptive dantzig density estimation. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 47, pages 43–74. Institut Henri Poincaré.

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

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

Brusco, M. J. and Cradit, J. D. (2001). A variable selection heuristic for k-means clustering. *Psychometrika*, **66**(2), 249–270.

Caillerie, C. and Michel, B. (2011). Model selection for simplicial approximation. *Foundations of Computational Mathematics*, **11**(6), 707–731.

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

Connault, P. (2011). *Calibration d'algorithmes de type Lasso et analyse statistique de données métallurgiques en aéronautique*. Ph.D. thesis, Université Paris-Sud 11.

Dash, M., Choi, K., Scheuermann, P., and Liu, H. (2002). Feature selection for clustering - a filter solution. *Proceedings of the Second IEEE International Conference on Data Mining*, pages 115–122.

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

Devaney, M. and Ram, A. (1997). Efficient feature selection in conceptual clustering. *Machine Learning: Proceedings of the Fourteenth International Conference*, pages 92–97.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, **32**(2), 407–499.

Fowlkes, E. B., Gnanadesikan, R., and Kettenring, J. R. (1988). Variable selection in clustering. *Journal of Classification*, **5**(2), 205–228.

Fraley, C. and Raftery, A. (2003). Enhanced software for model-based clustering, density estimation, and discriminant analysis: mclust. *Journal of Classification*, **20**, 263–286.

Jouve, P.-E. and Nicoloyannis, N. (2005). A filter feature selection method for clustering. *Proceedings of International Symposium on Methodologies for Intelligent Systems*, pages 583–593.

Kim, S., Tadesse, M. G., and Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, **93**(4), 877–893.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, **85**(4), 717–736.

Lerasle, M. (2011). Optimal model selection for stationary data under various mixing conditions. *Annals of statistics*, **39**(1), 1852–1877.

Lerasle, M. (2012). Optimal model selection in density estimation. *Annales de l'IHP*, **48**(3), 884–908.

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

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

Maugis, C. and Michel, B. (2011a). Data-driven penalty calibration: a case study for Gaussian mixture model selection. *ESAIM Probability and Statistics*, **15**, 320–339.

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

Maugis, C., Celeux, G., and Martin-Magniette, M. (2009a). Variable selection for clustering with Gaussian mixture models. *Biometrics*, **65**(3), 701–709.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009b). Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics and Data Analysis*, **53**, 3872–3882.

Misiti, M., Misiti, Y., Oppenheim, G., and Poggi, J.-M. (2007a). Clustering signals using wavelets. In *Proceedings of the 9th international work conference on Artificial neural networks*, IWANN'07, pages 514–521. Springer-Verlag.

Misiti, M., Misiti, Y., Oppenheim, G., and Poggi, J. (2007b). *Wavelets and their Applications*. Wiley Online Library.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, **8**, 1145–1164.

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.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, **100**(470), 602–617.

Verzelen, N. (2010). Data-driven neighborhood selection of a Gaussian field. *Computational Statistics and Data Analysis*, **54**(5), 1355–1371.

Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(2), 143–161.

Zhao, P. and Yu, B. (2007). On model selection consistency of lasso. *Journal of Machine Learning Research*, **7**, 2541–2567.

Zhou, H., Pan, W., and Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics*, **3**, 1473–1496.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**(476), 1418–1429.

RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

Parc Orsay Université
4 rue Jacques Monod
91893 Orsay Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399