Electronic Journal of Statistics
Vol. 9 (2015) 2642–2674
ISSN: 1935-7524
DOI: 10.1214/15-EJS1082

Finite mixture regression: A sparse variable selection by model selection for clustering*

Emilie Devijver

Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France
e-mail: emilie.devijver@math.u-psud.fr

Abstract: We consider a finite mixture of Gaussian regression models for high-dimensional data, where the number of covariates may be much larger than the sample size. We propose to estimate the unknown conditional mixture density by a maximum likelihood estimator, restricted on relevant variables selected by an ℓ_1-penalized maximum likelihood estimator. We get an oracle inequality satisfied by this estimator with a Jensen-Kullback-Leibler type loss. Our oracle inequality is deduced from a general model selection theorem for maximum likelihood estimators on a random model subcollection. We can derive the penalty shape of the criterion, which depends on the complexity of the random model collection.

MSC 2010 subject classifications: 62H30.
Keywords and phrases: Variable selection, finite mixture regression, non-asymptotic penalized criterion, ℓ_1-regularized method.

Received September 2014.

1. Introduction

The goal of clustering methods is to discover a structure among individuals described by several variables. Specifically, in regression case, given n observations $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n))$ which are realizations of random variables (X, Y) with $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, one aims at grouping the data into clusters such that the observations Y conditionally to X in the same cluster are more similar to each other than those from the other clusters. Different methods could be considered, more geometric or more statistical. We are dealing with model-based clustering, in order to have a rigorous statistical framework to assess the number of clusters and the role of each variable. This method is known to have good empirical performance relative to its competitors, see for instance [20].

Often, datasets are described by a lot of explicative variables, sometimes much more than the sample size. All the information should not be relevant for the clustering. To solve this problem, we propose a procedure which provides a data clustering from variable selection. In a density estimation way, we could refer to Pan and Shen, in [15], who focus on mean variable selection, Zhou and Pan, in [25], who use the Lasso estimator to regularize Gaussian mixture

*This is an original survey paper

model with general covariance matrices, Sun and Wand, in [19], who propose to regularize the k-means algorithm to deal with high-dimensional data, Guo et al. in [10], who propose a pairwise variable selection method. All those methods deal with penalized model-based clustering.

In a regression framework, the Lasso estimator, introduced by Tibshirani in [21], is a classical tool. Working well in practice, many efforts have been made recently on this estimator to get some theoretical results. Under a variety of different assumptions on the design matrix, we could get oracle inequalities for the Lasso estimator. For example, we can state the restricted eigenvalue condition, introduced by Bickel, Ritov and Tsybakov in [4], who get an oracle inequality with this assumption. For an overview of existing results, refer for example to [22].

Whereas focus on the estimation, the Lasso estimator could be used to select variables, and, for this goal, many results without strong assumptions are proved. The first result in this way is from Meinshausen and Bühlmann, in [13], who prove that, for neighborhood selection in Gaussian graphical models, under a neighborhood stability condition, the Lasso estimator is consistent. Under different assumptions, as the irrepresentable condition, described in [24], one get the same kind of result: true variables are selected consistently.

Thanks to those results, one could refit the estimation, after the variable selection, with an estimator with better properties. In this article, we focus on the maximum likelihood estimator on the estimated relevant set. In a linear regression framework, we could refer to Massart and Meynet, in [12], or Belloni and Chernozhukov, in [3], or also Sun and Zhang, [18] for using this idea.

In our case of finite mixture regression, we propose a procedure which is based on a modeling that recasts variable selection and clustering problems into a model selection problem. In mixture models, the choice of the number of components is often solved by a model selection criterion. In this paper, we select also the relevant variables by this criterion. Indeed, in practice, in high-dimension, we do not have access to the whole model collection. It is then a well-used procedure to restrict ourselves to a random subcollection of the whole collection. This procedure is developed in [7], with methodology, computational issues, simulations and data analysis. First, for some data-driven regularization parameters, we construct a relevant variables set. Then, restricted on those sets, we compute the maximum likelihood estimator. Considering the model collection with various number of components and various sparsities, we select a model thanks to the slope heuristic. Then, we get a clustering of the data thanks to the maximum a posteriori principle. This procedure could be used to cluster heterogeneous multivariate regression data and understand which variables explain the clustering, in high-dimension. Consider a regression clustering could refine a clustering, and it could be more adapted for instance for prediction. In this article, we focus on the theoretical point of view. We define a penalized criterion which allows to select a model as good as possible, from a non-asymptotic point of view. Penalizing the empirical contrast is an idea emerging from the seventies. Akaike, in [1], proposed the Akaike's Information Criterion (AIC) in 1973, and Schwarz in 1978 in [16] suggested the Bayesian

Information Criterion (BIC). Those criteria are based on asymptotic heuristics. To deal with non-asymptotic observations, Birgé and Massart in [5] and Barron et al. in [23], define a penalized data-driven criterion, which leads to oracle inequalities for model selection. In our context of regression, Cohen and Le Pennec, in [6], proposed a general model selection theorem for maximum likelihood estimation, adapted from Massart's Theorem in [11]. Nevertheless, we can not use it directly, because it is stated for a deterministic model collection, whereas our data-driven model collection is random, constructed by the Lasso estimator. It is important to consider a random subcollection model rather than the whole collection because the whole collection is not tractable in practice, due to the high-dimension. As Maugis-Rabusseau and Meynet have done in [14] to generalize Massart's Theorem, we extend the theorem to cope with the randomness of our model collection. By applying this general theorem to the finite mixture regression random model collection constructed by our procedure, we derive a convenient theoretical penalty as well as an associated non-asymptotic penalized criteria and an oracle inequality fulfilled by our Lasso-MLE estimator. The advantage of this procedure is that it does not need any restrictive assumption.

To obtain the oracle inequality, we use a general theorem proposed by Massart in [11], which gives the form of the penalty and associated oracle inequality in term of the Kullback-Leibler and Hellinger loss. In our case of regression, Cohen and Le Pennec, in [6], generalize this theorem in term of Kullback-Leibler and Jensen-Kullback-Leibler loss. Those theorems are based on the centred process control with the bracketing entropy, allowing to evaluate the size of the models. Our setting is more general, because we work with a random family. We have to control the centred process thanks to Bernstein's inequality.

The rest of this article is organized as follows. In the Section 2, we define the multivariate Gaussian mixture regression model, and we describe the main steps of the procedure we propose. We also illustrate the requirement of refitting by some simulations. We present our oracle inequality in the Section 3. In Section 4, we illustrate the procedure on simulated dataset and benchmark dataset. Finally, in Section 5, we give some tools to understand the proof of the oracle inequality, with a global theorem of model selection with a random collection in Section 5.1 and sketch of proofs after. All the technical details are given in Appendix.

2. The Lasso-MLE procedure

In order to cluster high-dimensional regression data, we work with the multivariate Gaussian mixture regression model. This model is developed in [17] in the scalar response case. We generalize it in Section 2.1. Moreover, we want to construct a model collection, with more or less components, and which is more or less sparse, to solve the estimation issue. We propose, in Section 2.2, a procedure called Lasso-MLE which constructs a model collection, with various sparsities and various number of components, of Gaussian mixture regression models. The different sparsities solve the high-dimensional problem. We conclude this section with simulations, which illustrate the advantage of refitting.

2.1. Gaussian mixture regression model

We observe n independent couples $(x_i, y_i)_{1\leq i\leq n}$ realizing the random variables (X, Y), where $X \in \mathbb{R}^p$, and $Y \in \mathbb{R}^q$ comes from a probability distribution with unknown conditional density denoted by s^*. To solve the clustering problem, we use a finite mixture regression model. In particular, we approximate the density of Y conditionally to X with a mixture of K multivariate Gaussian regression models. If the observation i belongs to the cluster k, we are looking for $\beta_k \in \mathbb{R}^{q\times p}$ such that $y_i = \beta_k x_i + \epsilon$, where $\epsilon \sim \mathcal{N}_q(0, \Sigma_k)$. Remark that we also have to estimate the number of clusters K.

Thus, the random response variable $Y \in \mathbb{R}^q$ depends on a set of random explanatory variables, written $X \in \mathbb{R}^p$, through a regression-type model. Give more precisions on the assumptions on the model we use.

- The variables Y_i, conditionally to X_i, are independent for all $i \in \{1, \dots, n\}$;
- $Y_i|X_i = x_i \sim s_\xi^K(y|x_i)dy$, with

$$s_\xi^K(y|x) = \sum_{k=1}^K \frac{\pi_k}{(2\pi)^{q/2}\det(\Sigma_k)^{1/2}} \exp\left(-\frac{(y-\beta_k x)^t \Sigma_k^{-1}(y-\beta_k x)}{2}\right) \quad (1)$$

$$\xi = (\pi_1, \dots, \pi_K, \beta_1, \dots, \beta_K, \Sigma_1, \dots, \Sigma_K) \in \Xi_K$$

$$\Xi_K = \left(\Pi_K \times (\mathbb{R}^{q\times p})^K \times (\mathbb{S}_q^{++})^K\right)$$

$$\Pi_K = \left\{(\pi_1, \dots, \pi_K); \pi_k > 0 \text{ for } k \in \{1, \dots, K\} \text{ and } \sum_{k=1}^K \pi_k = 1\right\}$$

$\mathbb{S}_q^{++}$ is the set of symmetric positive definite matrices on $\mathbb{R}^q$.

We want to estimate the conditional density function s_ξ^K from the observations. For all $k \in \{1, \dots, K\}$, β_k is the matrix of regression coefficients, and Σ_k is the covariance matrix in the mixture component k. The π_ks are the mixture proportions. In fact, for a regressor x, for all $k \in \{1, \dots, K\}$, for all $z \in \{1, \dots, q\}$, $[\beta_k x]_z = \sum_{j=1}^p [\beta_k]_{z,j} x_j$ is the zth component of the mean of the mixture component k. To deal with high-dimensional data, we select relevant variables.

Definition 2.1. *A variable $(z, j) \in \{1, \dots, q\} \times \{1, \dots, p\}$ is said to be* irrelevant *if, for all $k \in \{1, \dots, K\}$, $[\beta_k]_{z,j} = 0$. A variable is* relevant *if it is not irrelevant.*
A model is said to be sparse *if there are a few of relevant variables.*

We denote by $A^{[J]}$ the matrix A with 0 on the set ${}^cJ$, and $\mathcal{S}_{(K,J)}$ the model with K components and with J for relevant variables set:

$$\mathcal{S}_{(K,J)} = \left\{y \in \mathbb{R}^q \mapsto s_\xi^{(K,J)}(y|x)\right\} \quad (2)$$

where

$$s_\xi^{(K,J)}(y|x) = \sum_{k=1}^K \frac{\pi_k}{(2\pi)^{q/2}\det(\Sigma_k)^{1/2}} \exp\left(-\frac{(y-\beta_k^{[J]} x)^t \Sigma_k^{-1}(y-\beta_k^{[J]} x)}{2}\right).$$

This is the main model used in this article. To construct the set of relevant variables J, we use the Lasso estimator. Rather than select only one regularization parameter, we consider a grid of regularization parameter, then it leads to a model collection. Detail the procedure.

2.2. The Lasso-MLE procedure

The procedure we propose, which is particularly interesting in high-dimension, could be decomposed into three main steps. First, we construct a model collection, with models more or less sparse and with more or less components. Then, we refit estimations with the maximum likelihood estimator. Finally, we select a model thanks to the slope heuristic. It leads to a clustering according to the MAP principle on the selected model.

Model collection construction The first step consists of constructing a collection of models $\{\mathcal{S}_{(K,J)}\}_{(K,J)\in\mathcal{M}}$ in which the model $\mathcal{S}_{(K,J)}$ is defined by equation (2), and the model collection is indexed by $\mathcal{M} = \mathcal{K} \times \mathcal{J}$. We denote by $\mathcal{K} \subset \mathbb{N}^*$ the possible number of components, and by $\mathcal{J}$ a collection of subsets of $\{1, \dots, q\} \times \{1, \dots, p\}$.

To detect the relevant variables, and construct the set J for each model, we generalize the Lasso estimator. Indeed, we penalize the empirical contrast by an ℓ_1-penalty on the mean parameters proportional to

$$||P_k\beta_k||_1 = \sum_{j=1}^{p}\sum_{z=1}^{q} |[P_k\beta_k]_{z,j}|,$$

where $P_k^t P_k = \Sigma_k^{-1}$ for all $k \in \{1, \dots, K\}$. Then, we consider

$$\hat{\xi}_K^{\text{Lasso}}(\lambda) = \underset{\xi=(\boldsymbol{\pi},\boldsymbol{\beta},\boldsymbol{\Sigma})\in\Xi_K}{\operatorname{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log(s_\xi^K(y_i|x_i)) + \lambda \sum_{k=1}^{K} \pi_k ||P_k\beta_k||_1 \right\}.$$

This leads to penalize simultaneously the ℓ_1-norm of the mean coefficients and small variances. Computing those estimators lead to construct the relevant variables set. For a fixed number of mixture components $K \in \mathcal{K}$, denote by G_K a candidate of regularization parameters. Fixing a parameter $\lambda \in G_K$, we could then use an EM algorithm to compute the Lasso estimator, and construct the set of relevant variables $J_{(\lambda,K)}$, saying the non-zero coefficients. We denote by $\mathcal{J}$ the random collection of all these sets, $\mathcal{J} = \bigcup_{K\in\mathcal{K}} \bigcup_{\lambda\in G_K} J_{(\lambda,K)}$.

Refitting The second step consists of approximating the maximum likelihood estimator

$$\hat{s}^{(K,J)} = \underset{t\in\mathcal{S}_{(K,J)}}{\operatorname{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log(t(y_i|x_i)) \right\}$$

using an EM algorithm for each model $(K, J) \in \mathcal{K}\times\mathcal{J}$. Remark that we estimate all parameters, to reduce bias induced by the Lasso estimator.

Model selection The third step is devoted to model selection. We get a model collection, and we need to select the best one. Because we do not have access to s^*, we can not take the one which minimizes the risk. The Theorem 3.2 solves this problem: we get a penalty achieving to an oracle inequality. Then, even if we do not have access to s^*, we know that we can do almost like the oracle.

2.3. Why refit the Lasso estimator?

In order to illustrate the refitting, we compute multivariate data, the restricted eigenvalue condition being not satisfied, and run our procedure. We consider an extension of the model studied in Giraud et al. article [2] in the Section 6.3. Indeed, this model is a linear regression with a scalar response which does not satisfy the restricted eigenvalues condition. Then, we define different classes, to get a finite mixture regression model, which does not satisfied the restricted eigenvalues condition, and extend the dimension for multivariate response. We could compare the result of our procedure with the Lasso estimator, to illustrate the oracle inequality we get. Let precise the model.

Let $[\mathbf{x}]_1, [\mathbf{x}]_2, [\mathbf{x}]_3$ be three vectors of $\mathbb{R}^n$ defined by

$$\begin{aligned}
[\mathbf{x}]_1 &= (1, -1, 0, \dots, 0)^t/\sqrt{2} \\
[\mathbf{x}]_2 &= (-1, 1.001, 0, \dots, 0)^t/\sqrt{1+0.001^2} \\
[\mathbf{x}]_3 &= (1/\sqrt{2}, 1/\sqrt{2}, 1/n, \dots, 1/n)^t/\sqrt{1+(n-2)/n^2}
\end{aligned}$$

and for $4 \leq j \leq n$, let $[\mathbf{x}]_j$ be the j^{th} vector of the canonical basis of $\mathbb{R}^n$. We take a sample of size $n = 20$, and vectors of size $p = q = 10$. We consider two classes, each of them defined by $[\beta_1]_{z,j} = 10$ and $[\beta_2]_{z,j} = -10$ for $j \in \{1, \dots, 2\}$, $z \in \{1, \dots, 10\}$. Moreover, we define the covariance matrix of the noise by a diagonal matrix with 0.01 for diagonal coefficients in each class.

We run our procedure on this model, and compare it with the Lasso estimator, without refitting. We compute the model selected by the slope heuristic over the model collection constructed by the Lasso estimator. In Figure 1 stand the boxplots of each procedure, running 20 times. The Kullback-Leibler divergence is computed over a sample of size 5000.

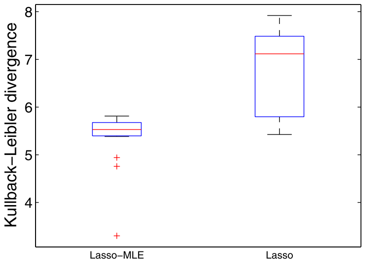

Fig 1. *Boxplot of the Kullback-Leibler divergence between the true model and the one constructed by each procedure, the Lasso-MLE procedure and the Lasso estimator.*

We could see that a refitting after variable selection by the Lasso estimator leads to a better estimation, according to the Kullback-Leibler loss.

3. An oracle inequality for the Lasso-MLE model

Before state the main theorem of this article, we need to precise some definitions and notations.

3.1. Notations and framework

We assume that the observations $(x_i, y_i)_{1\leq i\leq n}$ are i.i.d. realizations of random variables (X, Y), where $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$.

For $(K, J) \in \mathcal{K} \times \mathcal{J}$, for a model $\mathcal{S}_{(K,J)}$, we denote by $\hat{s}^{(K,J)}$ the maximum likelihood estimator

$$\hat{s}^{(K,J)} = \underset{s_\xi^{(K,J)} \in \mathcal{S}_{(K,J)}}{\operatorname{argmin}} \left(-\sum_{i=1}^{n} \log s_\xi^{(K,J)}(y_i|x_i) \right).$$

The best model in this collection is the one with the smallest risk. However, because we do not have access to the true density s^*, we can not select the best model, which we call the oracle. Thereby, there is a trade-off between a bias term measuring the closeness of s^* to the set $\mathcal{S}_{(K,J)}$, and a variance term depending on the complexity of the set $\mathcal{S}_{(K,J)}$ and on the sample size. A good set $\mathcal{S}_{(K,J)}$ is one for which this trade-off leads to a small risk bound. Because we are working with a maximum likelihood approach, the most natural quality measure is thus the Kullback-Leibler divergence denoted by KL.

$$\mathrm{KL}(s,t) = \begin{cases} \displaystyle\int_{\mathbb{R}} \log\left(\frac{s(y)}{t(y)}\right) s(y)dy \text{ if } sdy << tdy; \\ +\infty \text{ otherwise;} \end{cases} \tag{3}$$

for s and t two densities.

As we deal with conditional densities, the previous divergence should be adapted. We define the tensorized Kullback-Leibler divergence by

$$\mathrm{KL}^{\otimes_n}(s,t) = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}(s(.|x_i), t(.|x_i))\right].$$

Namely, we use the Jensen-Kullback-Leibler divergence JKL_ρ with $\rho \in (0,1)$, which is defined by

$$\mathrm{JKL}_\rho(s,t) = \frac{1}{\rho}\,\mathrm{KL}(s, (1-\rho)s + \rho t);$$

and the tensorized one

$$\mathrm{JKL}_\rho^{\otimes_n}(s,t) = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{JKL}_\rho(s(.|x_i), t(.|x_i))\right].$$

This divergence is studied in [6]. We use this divergence rather than the Kullback-Leibler one because we need a boundedness assumption on the controlled functions that is not satisfied by the log-likelihood differences $-\log(s_\xi^{(K,J)}/s^*)$. When considering the Jensen-Kullback-Leibler divergence, those ratios are replaced by ratios

$$-\frac{1}{\rho}\log\left(\frac{(1-\rho)s^* + \rho s_\xi^{(K,J)}}{s^*}\right)$$

that are close to the log-likelihood differences when $s_\xi^{(K,J)}$ are close to s^* and always upper bounded by $-\log(1-\rho)/\rho$. Indeed, this bound is needed to use deviation inequalities for sum of random variables and its supremum, which is the key of the proof of oracle type inequality.

3.2. Oracle inequality

We denote by $(\mathcal{S}_{(K,J)})_{(K,J)\in\mathcal{K}\times\mathcal{J}^L}$ the model collection constructed by the Lasso-MLE procedure, with $\mathcal{J}^L$ a random subcollection of $\mathcal{P}(\{1,\dots,q\}\times\{1,\dots,p\})$ constructed by the Lasso estimator. The grid of regularization parameter considered is data-driven, then random. Because we work in high-dimension, we could not look at all subsets of $\mathcal{P}(\{1,\dots,q\}\times\{1,\dots,p\})$. Considering the Lasso estimator through its regularization path is the solution chosen here, but it needs more control because of the random family. To get theoretical results, we need to work with restricted parameters. Assume that Σ_k is diagonal, with $\Sigma_k = \mathrm{diag}([\Sigma_k]_{1,1},\dots,[\Sigma_k]_{q,q})$, for all $k\in\{1,\dots,K\}$. We define

$$\mathcal{S}^{\mathcal{B}}_{(K,J)} = \Bigg\{ s_\xi^{(K,J)} \in \mathcal{S}_{(K,J)} \Bigg| \text{ for all } k\in\{1,\dots,K\}, \beta_k^{[J]} \in [-A_\beta, A_\beta]^{q\times p}, \\ a_\Sigma \leq [\Sigma_k]_{z,z} \leq A_\Sigma \text{ for all } z\in\{1,\dots,q\}, \text{ for all } k\in\{1,\dots,K\}\Bigg\}. \quad (4)$$

Moreover, we assume that the covariates X belong to an hypercube. Without any restriction, we could assume that $X\in[0,1]^p$.

Remark 3.1. *We have to denote that in this article, the relevant variables set is designed by the Lasso estimator. Nevertheless, any tool could be used to construct this set, and we obtain similar results. We could work with any random subcollection of $\mathcal{P}(\{1,\dots,q\}\times\{1,\dots,p\})$, the controlled size being required in high-dimension case.*

Theorem 3.2. *Let $(x_i,y_i)_{1\leq i\leq n}$ the observations, with unknown conditional density s^*. Let $\mathcal{S}_{(K,J)}$ defined by (2). For $(K,J)\in\mathcal{K}\times\mathcal{J}^L$, $\mathcal{J}^L$ being a random subcollection of $\mathcal{P}(\{1,\dots,q\}\times\{1,\dots,p\})$ constructed by the Lasso estimator, denote by $\mathcal{S}^{\mathcal{B}}_{(K,J)}$ the model defined by (4).*

Consider the maximum likelihood estimator

$$\hat{s}^{(K,J)} = \underset{s_\xi^{(K,J)}\in\mathcal{S}^{\mathcal{B}}_{(K,J)}}{\mathrm{argmin}} \left\{-\frac{1}{n}\sum_{i=1}^n \log s_\xi^{(K,J)}(y_i|x_i)\right\}.$$

Denote by $D_{(K,J)}$ the dimension of the model $\mathcal{S}^{\mathcal{B}}_{(K,J)}$, $D_{(K,J)} = K(|J|+q+1)-1$. Let $\bar{s}^{(K,J)} \in \mathcal{S}^{\mathcal{B}}_{(K,J)}$ such that

$$\mathrm{KL}^{\otimes_n}(s^*, \bar{s}^{(K,J)}) \leq \inf_{t \in \mathcal{S}^{\mathcal{B}}_{(K,J)}} \mathrm{KL}^{\otimes_n}(s^*, t) + \frac{\delta_{\mathrm{KL}}}{n};$$

and let $\tau > 0$ such that

$$\bar{s}^{(K,J)} \geq e^{-\tau} s^*. \tag{5}$$

Let $\mathrm{pen} : \mathcal{K} \times \mathcal{J} \to \mathbb{R}_+$, and suppose that there exists an absolute constant $\kappa > 0$ and an absolute constant $B(A_\beta, A_\Sigma, a_\Sigma)$ such that, for all $(K,J) \in \mathcal{K} \times \mathcal{J}$,

$$\mathrm{pen}(K,J) \geq \kappa \frac{D_{(K,J)}}{n} \Bigg[B^2(A_\beta, A_\Sigma, a_\Sigma) - \log\left(\frac{D_{(K,J)}}{n} B^2(A_\beta, A_\Sigma, a_\Sigma) \wedge 1 \right) \\ + (1 \vee \tau) \log\left(\frac{4epq}{(D_{(K,J)} - q) \wedge pq} \right) \Bigg].$$

Then, the estimator $\hat{s}^{(\hat{K},\hat{J})}$, with

$$(\hat{K}, \hat{J}) = \underset{(K,J) \in \mathcal{K} \times \mathcal{J}^L}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(\hat{s}^{(K,J)}(y_i|x_i)) + \mathrm{pen}(K,J) \right\},$$

satisfies

$$\mathrm{E}\left[\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}^{(\hat{K},\hat{J})}) \right] \\ \leq C_1 \, \mathrm{E}\left(\inf_{(K,J) \in \mathcal{K} \times \mathcal{J}^L} \left(\inf_{t \in S^{\mathcal{B}}_{(K,J)}} \mathrm{KL}^{\otimes_n}(s^*, t) + \mathrm{pen}(K,J) \right) \right) + C_2 \frac{(1 \vee \tau)}{n};$$

for some absolute positive constants C_1 and C_2.

This oracle inequality compares performances of our estimator with the best model in the collection. Nevertheless, as we consider mixture of Gaussian, if we take enough clusters, we could approximate well a lot of densities, then the term in the right-hand side is small for $\mathcal{K}$ well-chosen. This result could be compared with the oracle inequality get in [17], Theorem 4. Indeed, under restricted eigenvalues condition and fixed design, they get an oracle inequality for the Lasso estimator in finite mixture regression model, with scalar response and high-dimension regressors. Note that they control the divergence with the true parameters. We get a similar result for the Lasso-MLE estimator. Moreover, our procedure work in a more general framework, the only assumption needed is to be bounded.

Remark that the penalty is proportional to the dimension of the model, up to a logarithm. This term is needed for two reasons. First, for the proof, we compute the bracketing entropy of the whole model rather than the local bracketing entropy, which is not optimal. Moreover, this penalty takes into account the model collection complexity. Perhaps, because of the high-dimension, a large number of models have the same dimension. Remark that the penalty is the same as if we have considered the whole collection, but it is not tractable in practice. The only assumption needed for the random subcollection is the Assumption (5). Nevertheless, it is not really restrictive, see Section 5.1 for more discussion about this Assumption.

4. Numerical experiments

To illustrate this procedure, we study some simulations and real data. The main algorithm is a generalized version of the EM algorithm, which is used many times for the procedure. We first use it to compute maximum likelihood estimator, to construct the regularization parameter grid. Then, we use it to compute the Lasso estimator for each regularization parameter belonging to the grid, and we are able to construct the relevant variables set. Finally, we could compute the maximum likelihood estimator, restricted to those relevant variables in each model. Among this model collection, we select one using the Capushe package. More details, as initialization rule, stopping rule, and more numerical experiments, are available in [7].

4.1. Simulation illustration

We illustrate the procedure on a simulated dataset, adapted from [17].

Let $\mathbf{x}$ be a sample of size $n = 100$ distributed according to multivariate standard Gaussian. We consider a mixture of two components, and we fix the dimension of the regressor and of the response variables to $p = q = 10$. Besides, we fix the number of relevant variables to 4 in each cluster. More precisely, the first four variables of Y are explained respectively by the four first variables of X. Fix $\pi = (\frac{1}{2}, \frac{1}{2})$, $\beta_1^{[J]} = 3$, $\beta_2^{[J]} = -2$ and $P_k = 3I_q$ for all $k \in \{1, 2\}$.

The difficulty of the clustering is partially controlled by the signal-to-noise ratio. In this context, we could extend the natural idea of the SNR with the following definition, where $\mathrm{Tr}(A)$ denotes the trace of the matrix A.

$$\mathrm{SNR} = \frac{\mathrm{Tr}(\mathrm{Var}(Y))}{\mathrm{Tr}(\mathrm{Var}(Y|\beta_k = 0 \text{ for all } k \in \{1, \dots, K\}))} = 1.88.$$

We take a sample of Y knowing $X = x$ according to a Gaussian mixture, centered in $\beta_k x$ and with covariance matrix $\Sigma_k = (P_k^t P_k)^{-1} = \sigma I_q$, for the cluster $k \in \{1, 2\}$. We run our procedure with the number of components varying in $\mathcal{K} = \{2, \dots, 5\}$.

To compare our procedure with others, we compute the Kullback-Leibler divergence with the true density, the ARI (the Adjusted Rand Index measures the similarity between two data clusterings, knowing that the closer to 1 the ARI, the more similar the two partitions), and how many clusters are selected.

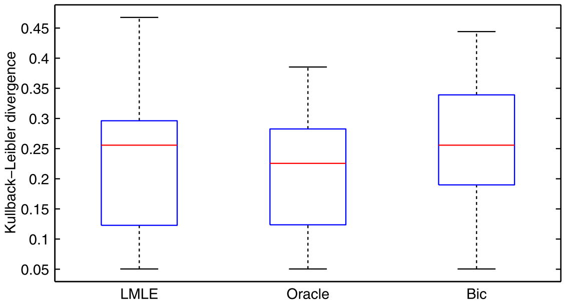

Fig 2. *Boxplots of the Kullback-Leibler divergence between the true model and the one selected by the procedure over the* 20 *simulations, for the Lasso-MLE procedure (LMLE), the oracle (Oracle), and the BIC estimator (BIC).*

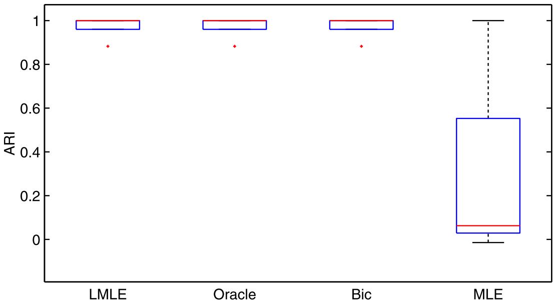

Fig 3. *Boxplots of the ARI over the* 20 *simulations, for the Lasso-MLE procedure (LMLE), the oracle (Oracle), the BIC estimator (BIC) and the MLE (MLE).*

From the Lasso-MLE model collection, we construct two models, to compare our procedure with. We compute the oracle (the model which minimizes the Kullback-Leibler divergence with the true density), and the model which is selected by the BIC criterion instead of the slope heuristic. Thanks to the oracle, we know how good we could be from this model collection for the Kullback-Leibler divergence, and how this model, as good it is possible for the contrast, performs the clustering. The third procedure we compare with is the maximum likelihood estimator, assuming that we know how many clusters there are, fixed to 2. We use this procedure to show that variable selection is necessary.

Results are summarized in Figure 2 and in Figure 3. The Kullback-Leibler divergence is smaller for models coming from our model collection (either by BIC, or by slope heuristic, or the oracle) than for the model constructed by the MLE, which we do not plot in Figure 2 for easier reading. The ARI is closer to 1 in those cases. We could conclude that the model collection is well constructed, selecting relevant variables, and also that the model is well selected among this collection, near the oracle.

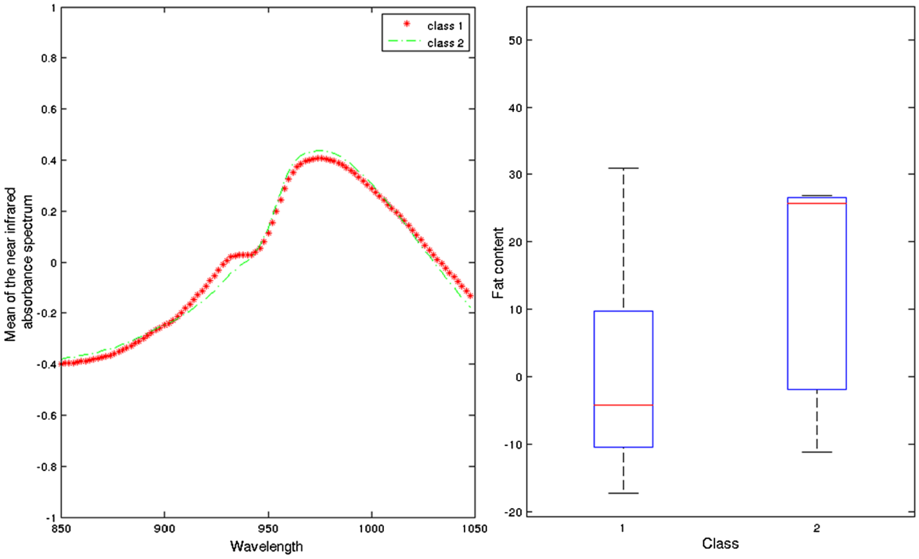

FIG 4. *Summarized results for the model 1. The graph on the left is a candidate for representing each cluster, constructed by the mean of reconstructed spectrum over an a posteriori probability greater than* 0.6*. On the right side, we present the boxplot of the fat values in each class, for observations with an a posteriori probability greater than* 0.6*.*

4.2. Real data

We also illustrate the procedure on the Tecator dataset, which deals with spectrometric data. We summarize here results which are described in [7]. This data has been studied in a lot of articles, refer for example to Ferraty and Vieu's book [8]. The data consists of a 100-channel spectrum of absorbances in the wavelength range $850 - 1050$ nm, and of the percentage of fat. We observe a sample of size 215. In this work, we focus on clustering data according to the reliance between the fat content and the absorbance spectrum. The sample is split into two subsamples, 165 observations for the learning set, and 50 observations for the test set. We split it to have the same marginal distribution of the response in each sample.

The spectrum is a function, which we decompose into the Haar basis, at level 6.

The procedure selects two models, which we describe here. In Figure (4) and Figure (5), we represent clusters done on the training set for the different models.

The graph on the left is a candidate for representing each cluster, constructed by the mean of spectrum over an a posteriori probability greater than 0.6. We plot the curve reconstruction, keeping only relevant variables in the wavelet decomposition. On the right side, we present the boxplot of the fat values in each class, for observations with an a posteriori probability greater than 0.6.

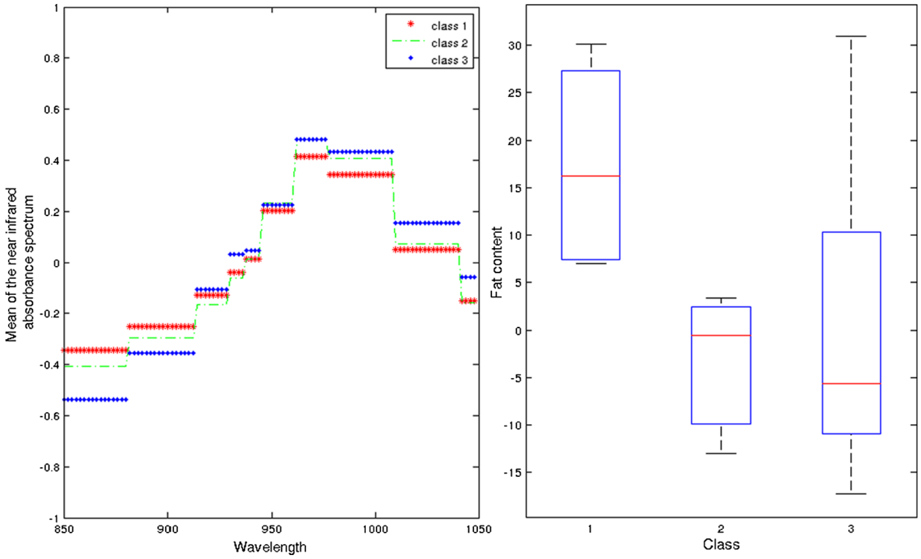

Fig 5. *Summarized results for the model 2. The graph on the left is a candidate for representing each cluster, constructed by the mean of reconstructed spectrum over an a posteriori probability greater than* 0.6*. On the right side, we present the boxplot of the fat values in each class, for observations with an a posteriori probability greater than* 0.6*.*

The first model has two clusters, which could be distinguish in the absorbance spectrum by the bump on wavelength around 940 nm. The first class is dominating, with $\hat{\pi}_1 = 0.95$. The fat content is smaller in the first class than in the second class. According to the signal reconstruction, we could see that almost all variables have been selected. This model seems consistent according to the classification goal.

The second model has 3 clusters, and we could remark several wavelengths which explain the clustering. Around 940 nm, there are some differences between clusters, corresponding to the bump underline in the first model. Moreover, around 970 nm, there are also some differences. The first class is dominating, with $\hat{\pi}_1 = 0.89$. Just a few of variables have been selected, which give to this model the understanding property of which coefficients are discriminating.

We could discuss about those models. The first selects only two clusters, but almost all variables, whereas the second model has more clusters, and less variables: there is a trade-off between clusters and variable selection for the dimension reduction.

5. Tools for proof

In this section, we present the tools needed to understand the proof. First, we present a general theorem for model selection in regression among a random collection. Then, in Subsection 5.2, we present the proof of this theorem, and in the Subsection 5.3 we explain how we could use the main theorem to get the oracle inequality. All details are available in Appendix.

5.1. General theory of model selection with the maximum likelihood estimator.

To get an oracle inequality for our clustering procedure, we have to use a general model selection theorem. Because the model collection constructed by our procedure is random, because of the Lasso estimator which selects variables randomly, we have to generalize Cohen and Le Pennec's theorem. Begin by some general model selection theory.

Before state the general theorem, begin by talk about the assumptions. We work here in a more general context, $(X,Y) \in \mathcal{X} \times \mathcal{Y}$, and $(S_m)_{m\in\mathcal{M}}$ defining a model collection indexed by $\mathcal{M}$. First, we impose a structural assumption on each model indexed by $m \in \mathcal{M}$. It is a bracketing entropy condition on the model S_m with respect to the Hellinger divergence, defined by

$$(d_H^{\otimes_n})^2(s,t) = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} d_H^2(s(.|x_i), t(.|x_i))\right].$$

A bracket $[l,u]$ is a pair of functions such that for all $(x,y) \in \mathcal{X} \times \mathcal{Y}, l(y,x) \leq s(y|x) \leq u(y,x)$. The bracketing entropy $\mathcal{H}_{[.]}(\epsilon, S, d_H^{\otimes_n})$ of a set S is defined as the logarithm of the minimum number of brackets $[l,u]$ of width $d_H^{\otimes_n}(l,u)$ smaller than ϵ such that every function of S belongs to one of these brackets. It leads to the Assumption 1.

Assumption 1. *There exists a non-decreasing function ϕ_m such that $\varpi \mapsto \frac{1}{\varpi}\phi_m(\varpi)$ is non-increasing on $(0,+\infty)$ and for every $\varpi \in \mathbb{R}^+$ and every $s_m \in S_m$,*

$$\int_0^{\varpi} \sqrt{\mathcal{H}_{[.]}(\epsilon, S_m(s_m,\varpi), d_H^{\otimes_n})}d\epsilon \leq \phi_m(\varpi);$$

where $S_m(s_m,\varpi) = \{t \in S_m, d_H^{\otimes_n}(t,s_m) \leq \varpi\}$. The model complexity $\mathcal{D}_m$ is then defined as $n\varpi_m^2$ with ϖ_m the unique root of

$$\frac{1}{\varpi}\phi_m(\varpi) = \sqrt{n}\varpi. \tag{6}$$

Remark that the model complexity depends on the bracketing entropies not of the global models S_m but of the ones of smaller localized sets. This is a weaker assumption.

For technical reason, a separability assumption is also required.

Assumption 2. *There exists a countable subset $S_m^{'}$ of S_m and a set $\mathcal{Y}_m^{'}$ with $\lambda(\mathcal{Y} \setminus \mathcal{Y}_m^{'}) = 0$ such that for every $t \in S_m$, there exists a sequence $(t_l)_{l\geq 1}$ of elements of $S_m^{'}$ such that for every x and every $y \in \mathcal{Y}_m^{'}$, $\log(t_l(y|x))$ goes to $\log(t(y|x))$ as l goes to infinity.*

This assumption leads to work with a countable family, which allows to cope with the randomness of $\hat{s}_m$. We also need an information theory type assumption on our collection. We assume the existence of a Kraft-type inequality for the collection.

Assumption 3. *There is a family $(w_m)_{m\in\mathcal{M}}$ of non-negative numbers such that*

$$\sum_{m\in\mathcal{M}} e^{-w_m} \leq \Omega < +\infty.$$

The difference with Cohen and Le Pennec's Theorem is that we consider a random collection of models $\check{\mathcal{M}}$, included in the whole collection $\mathcal{M}$. In our procedure, we deal with high-dimensional models, and we cannot test all the models: we have to restrict ourselves to a smaller subcollection of models, which is then random. In the proof of the theorem, we have to be careful with the recentred process of $-\log(\bar{s}_m/s^*)$. Because we conclude by taking the expectation, if $\mathcal{M}$ is fixed, this term is equal to zero, but if we consider a random family, we have to use the Bernstein inequality to control this quantity, and then we have to make the assumption (7).

Let state our main global theorem.

Theorem 5.1. *Assume we observe $(x_i, y_i)_{1\leq i\leq n}$ with unknown conditional density s^*. Let the model collection $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ be at most countable collection of conditional density sets. Assume Assumption 3 holds, while Assumption 1 and Assumption 2 hold for every $m\in\mathcal{M}$. Let $\delta_{\mathrm{KL}} > 0$, and $\bar{s}_m \in S_m$ such that*

$$\mathrm{KL}^{\otimes_n}(s^*, \bar{s}_m) \leq \inf_{t\in S_m} \mathrm{KL}^{\otimes_n}(s^*, t) + \frac{\delta_{\mathrm{KL}}}{n};$$

and let $\tau > 0$ such that

$$\bar{s}_m \geq e^{-\tau} s^*. \tag{7}$$

Introduce $(S_m)_{m\in\check{\mathcal{M}}}$ some random subcollection of $(S_m)_{m\in\mathcal{M}}$. Consider the collection $(\hat{s}_m)_{m\in\check{\mathcal{M}}}$ of η-log-likelihood minimizer in S_m, satisfying, for all $m\in\check{\mathcal{M}}$,

$$\sum_{i=1}^n -\log(\hat{s}_m(y_i|x_i)) \leq \inf_{s_m\in S_m}\left(\sum_{i=1}^n -\log(s_m(y_i|x_i))\right) + \eta.$$

Then, for any $\rho\in(0,1)$ and any $C_1 > 1$, there are two constants κ_0 and C_2 depending only on ρ and C_1 such that, as soon as for every index $m\in\mathcal{M}$,

$$\mathrm{pen}(m) \geq \kappa(\mathcal{D}_m + (1\vee\tau)w_m) \tag{8}$$

with $\kappa > \kappa_0$, and where the model complexity $\mathcal{D}_m$ is defined in (6), the penalized likelihood estimate $\hat{s}_{\hat{m}}$ with $\hat{m}\in\check{\mathcal{M}}$ such that

$$-\sum_{i=1}^n \log(\hat{s}_{\hat{m}}(y_i|x_i)) + \mathrm{pen}(\hat{m}) \leq \inf_{m\in\check{\mathcal{M}}}\left(-\sum_{i=1}^n \log(\hat{s}_m(y_i|x_i)) + \mathrm{pen}(m)\right) + \eta^{'}$$

satisfies

$$\begin{aligned}\mathrm{E}(\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{\hat{m}})) \leq & C_1\,\mathrm{E}\left(\inf_{m\in\check{\mathcal{M}}}\left(\inf_{t\in S_m}\mathrm{KL}^{\otimes_n}(s^*, t)\right) + 2\frac{\mathrm{pen}(m)}{n}\right) \\ & + C_2(1\vee\tau)\frac{\Omega^2}{n} + \frac{\eta' + \eta}{n}.\end{aligned} \tag{9}$$

Obviously, one of the models minimizes the right hand side. Unfortunately, there is no way to know which one without knowing s^*. Hence, this oracle model can not be used to estimate s^*. We nevertheless propose a data-driven strategy to select an estimate among the collection of estimates $\{\hat{s}_m\}_{m \in \check{\mathcal{M}}}$ according to a selection rule that performs almost as well as if we had known this oracle, according to the absolute constant C_1. Using simply the log-likelihood in each model as a criterion is not sufficient. It is an underestimation of the true risk of the estimate and this leads to select models too complex. By adding an adapted penalty $\text{pen}(m)$, one hopes to compensate for both the variance term and the bias term between $-1/n \sum_{i=1}^{n} \log(\hat{s}_{\hat{m}}(y_i|x_i)/s^*(y_i|x_i))$ and $\inf_{s_m \in S_m} \text{KL}^{\otimes_n}(s^*, s_m)$. For a given choice of $\text{pen}(m)$, the best model $S_{\hat{m}}$ is chosen as the one whose index is a minimizer of the penalized η-log-likelihood.

Talk about the assumption (7). If s is bounded, with a compact support, this assumption is satisfied. It is also satisfied in other cases, more particular. Then it is not a strong assumption, but it is needed to control the random family.

This theorem is available for whatever model collection constructed, whereas Assumption 1, Assumption 2 and Assumption 3 are satisfied. In the following, we use this theorem for the procedure we propose to cluster high-dimensional data. Nevertheless, this theorem is not specific for our context, and could be used whatever the problem.

Remark that the constant associated to the Assumption 3 appears squared in the bound. It is due to the random subcollection $\check{\mathcal{M}}$ of $\mathcal{M}$, if the model collection is fixed, we get a linear bound. Moreover, the weights w_m appear linearly in the penalty bound.

5.2. *Proof of the general theorem*

For any model S_m, we have denoted by $\bar{s}_m$ a function such that

$$\text{KL}^{\otimes_n}(s^*, \bar{s}_m) \leq \inf_{s_m \in S_m} \text{KL}^{\otimes_n}(s^*, s_m) + \frac{\delta_{\text{KL}}}{n}.$$

Fix $m \in \mathcal{M}$ such that $\text{KL}^{\otimes_n}(s^*, \bar{s}_m) < +\infty$. Introduce

$$\mathcal{M}(m) = \left\{ m' \in \mathcal{M} \,\middle|\, P_n(-\log \hat{s}_{m'}) + \frac{\text{pen}(m')}{n} \right.$$
$$\left. \leq P_n(-\log \hat{s}_m) + \frac{\text{pen}(m)}{n} + \frac{\eta'}{n} \right\};$$

where $P_n(g) = 1/n \sum_{i=1}^{n} g(y_i|x_i)$. We define the functions $kl(\bar{s}_m), kl(\hat{s}_m)$ and $jkl_\rho(\hat{s}_m)$ by

$$kl(\bar{s}_m) = -\log\left(\frac{\bar{s}_m}{s^*}\right); \qquad kl(\hat{s}_m) = -\log\left(\frac{\hat{s}_m}{s^*}\right);$$
$$jkl_\rho(\hat{s}_m) = -\frac{1}{\rho} \log\left(\frac{(1-\rho)s^* + \rho \hat{s}_m}{s^*}\right).$$

For every $m^{'} \in \mathcal{M}(m)$, by definition,

$$\begin{aligned} P_n(kl(\hat{s}_{m'})) + \frac{\mathrm{pen}(m^{'})}{n} \leq & P_n(kl(\hat{s}_m)) + \frac{\mathrm{pen}(m) + \eta^{'}}{n} \\ \leq & P_n(kl(\bar{s}_m)) + \frac{\mathrm{pen}(m) + \eta^{'} + \eta}{n}. \end{aligned}$$

Let $\nu_n^{\otimes_n}(g)$ denotes the recentred process $P_n(g) - P^{\otimes_n}(g)$. By concavity of the logarithm,

$$kl(\hat{s}_{m'}) \ \geq \ jkl_\rho(\hat{s}_{m'}),$$

and then

$$\begin{aligned} & P^{\otimes_n}(jkl_\rho(\hat{s}_{m'})) - \nu_n^{\otimes_n}(kl(\bar{s}_m)) \\ \leq & P^{\otimes_n}(kl(\bar{s}_m)) + \frac{\mathrm{pen}(m)}{n} - \nu_n^{\otimes_n}(jkl_\rho(\hat{s}_{m'})) + \frac{\eta^{'} + \eta}{n} - \frac{\mathrm{pen}(m^{'})}{n}, \end{aligned}$$

which is equivalent to

$$\begin{aligned} \mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{m'}) - \nu_n^{\otimes_n}(kl(\bar{s}_m)) \leq & \mathrm{KL}^{\otimes_n}(s^*, \bar{s}_m) + \frac{\mathrm{pen}(m)}{n} - \nu_n^{\otimes_n}(jkl_\rho(\hat{s}_{m'})) \\ & + \frac{\eta^{'} + \eta}{n} - \frac{\mathrm{pen}(m^{'})}{n}. \end{aligned} \tag{10}$$

Mimic the proof as done in Cohen and Le Pennec in [6], we could obtain that except on a set of probability less than $e^{-w_{m'} - w}$, for all w, for all $\mathfrak{z}_{m'} > \sigma_{m'}$, there exist absolute constants $\kappa_0^{'}, \kappa_1^{'}, \kappa_2^{'}$ such that

$$\frac{-\nu_n^{\otimes_n}(jkl_\rho(\hat{s}_{m'}))}{\mathfrak{z}_{m'}^2 + \kappa_0^{'}(d_H^{\otimes_n})^2(s^*, \hat{s}_{m'})} \leq \frac{\kappa_1^{'}\sigma_{m'}}{\mathfrak{z}_{m'}} + \kappa_2^{'}\sqrt{\frac{w_{m'} + w}{n\mathfrak{z}_{m'}^2}} + \frac{18}{\rho}\frac{w_{m'} + w}{n\mathfrak{z}_{m'}^2}. \tag{11}$$

To obtain this inequality we use Assumption 1 and Assumption 2. This control is derived from maximal inequalities, described in [11].

Our purpose is now to control $\nu_n^{\otimes_n}(kl(\bar{s}_m))$. This is the difference with the Theorem of Cohen and Le Pennec: we work with a random subcollection $\check{\mathcal{M}}$ of $\mathcal{M}$.

By definition of kl and $\nu_n^{\otimes_n}$,

$$\nu_n^{\otimes_n}(kl(\bar{s}_m)) = -\frac{1}{n}\sum_{i=1}^n \log\left(\frac{\bar{s}_m(y_i|x_i)}{s^*(y_i|x_i)}\right) + \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n \log\left(\frac{\bar{s}_m(Y_i|X_i)}{s^*(Y_i|X_i)}\right)\right].$$

We want to apply Bernstein's inequality, which is recalled in Appendix.

If we denote by Z_i the random variable $Z_i = -\frac{1}{n}\log(\frac{\bar{s}_m(Y_i|X_i)}{s^*(Y_i|X_i)})$, we get

$$\nu_n^{\otimes_n}(kl(\bar{s}_m)) = \sum_{i=1}^n (Z_i - \mathrm{E}(Z_i)).$$

We need to control the moments of Z_i to apply Bernstein's inequality.

Lemma 5.2. *Let s^* and $\bar{s}_m$ two conditional densities with respect to the Lebesgue measure. Assume that there exists $\tau > 0$ such that $\log(||\frac{s^*}{\bar{s}_m}||_\infty) \leq \tau$. Then,*

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^q}\left(\log\left(\frac{s^*(y|x_i)}{\bar{s}_m(y|x_i)}\right)\right)^2 s^*(y|x_i)dy\right) \leq \frac{\tau^2}{e^{-\tau}+\tau-1}\,\mathrm{KL}^{\otimes_n}(s^*,\bar{s}_m).$$

We prove this lemma in Appendix A.2.

Because $\frac{\tau^2}{e^{-\tau}+\tau-1} \underset{\tau\to\infty}{\sim} \tau$, there exists A such that $\frac{\tau^2}{e^{-\tau}+\tau-1} \leq 2\tau$ for all $\tau \geq A$. For $\tau \in (0, A]$, because this function is continuous and equivalent to 2 in 0, there exists $B > 0$ such that $\frac{\tau^2}{e^{-\tau}+\tau-1} \leq B$. We obtain that $\sum_{i=1}^{n}\mathrm{E}(Z_i^2) \leq \frac{1}{n}\delta(1\vee\tau)\,\mathrm{KL}^{\otimes_n}(s^*,\bar{s}_m)$, where $\delta = 2 \vee B$.

Moreover, for all integers $K \geq 3$,

$$\begin{aligned}\sum_{i=1}^{n}\mathrm{E}((Z_i)_+^K) &\leq \sum_{i=1}^{n}\frac{1}{n^K}\int_{\mathbb{R}^q}\left(\log\left(\frac{s^*(y|x_i)}{\bar{s}_m(y|x_i)}\right)\right)_+^K s^*(y|x_i)dy\\ &\leq \frac{n}{n^K}\int_{\mathbb{R}^q}\log\left(\frac{s^*(y|x)}{\bar{s}_m(y|x)}\right)^{K-2}\log\left(\frac{s^*(y|x)}{\bar{s}_m(y|x)}\right)^2\\ &\qquad\qquad \times \mathbb{1}_{s^*(y|x)\geq\bar{s}_m(y|x)}s^*(y|x)dy\\ &\leq \frac{n}{n^K}\tau^{K-2}\delta(1\vee\tau)\,\mathrm{KL}^{\otimes_n}(s^*,\bar{s}_m).\end{aligned}$$

Assumptions of Bernstein's inequality are then satisfied, with

$$v = \frac{\delta(1\vee\tau)\,\mathrm{KL}^{\otimes_n}(s^*,\bar{s}_m)}{n}, \qquad c = \frac{\tau}{n},$$

thus, for all $u > 0$, except on a set with probability less than e^{-u},

$$\nu_n^{\otimes_n}(kl(\bar{s}_m)) \leq \sqrt{2vu} + cu.$$

Thus, for all $z > 0$, for all $u > 0$, except on a set with probability less than e^{-u},

$$\frac{\nu_n^{\otimes_n}(kl(\bar{s}_m))}{z^2+\mathrm{KL}^{\otimes_n}(s^*,\bar{s}_m)} \leq \frac{\sqrt{2vu}+cu}{z^2+\mathrm{KL}^{\otimes_n}(s^*,\bar{s}_m)} \leq \frac{\sqrt{vu}}{z\sqrt{2\,\mathrm{KL}^{\otimes_n}(s^*,\bar{s}_m)}} + \frac{cu}{z^2}. \qquad (12)$$

We apply this bound to $u = w + w_m + w_{m'}$. We get that, except on a set with probability less than $e^{-(w+w_m+w_{m'})}$, using that $a^2 + b^2 \geq a^2$, from the inequality (11),

$$-\nu_n^{\otimes_n}(jkl_\rho(\hat{s}_{m'})) \leq \left(\mathfrak{z}_{m'}^2 + \kappa_0'(d_H^{\otimes_n})^2(s^*,\hat{s}_{m'})\right)\left(\frac{\kappa_1'+\kappa_2'}{\theta}+\frac{18}{\theta^2\rho}\right),$$

and, from the inequality (12),

$$\nu_n^{\otimes_n}(kl(\bar{s}_m)) \leq (\beta+\beta^2)\left(z_{m,m'}^2 + \mathrm{KL}^{\otimes_n}(s,s_m)\right),$$

where we have chosen

$$\mathfrak{z}_{m'} = \theta\sqrt{\sigma_{m'}^2 + \frac{w_{m'} + w}{n}},$$

with $\theta > 1$ to fix later, and

$$z_{m,m'} = \beta^{-1}\sqrt{\left(\frac{v}{2\,\mathrm{KL}^{\otimes_n}(s^*, \bar{s}_m)} + c\right)(w + w_m + w_{m'})},$$

with $\beta > 0$ to fix later.

Coming back to the inequality (10),

$$\begin{aligned}\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{m'}) \leq{}& \mathrm{KL}^{\otimes_n}(s^*, \bar{s}_m) + \frac{\mathrm{pen}(m)}{n}\\ &+ (\mathfrak{z}_{m'}^2 + \kappa_0'(d_H^{\otimes_n})^2(s^*, \hat{s}_{m'}))\left(\frac{\kappa_1' + \kappa_2'}{\theta} + \frac{18}{\theta^2\rho}\right)\\ &+ \frac{\eta' + \eta}{n} - \frac{\mathrm{pen}(m')}{n} + (\beta + \beta^2)(z_{m,m'}^2 + \mathrm{KL}^{\otimes_n}(s^*, \bar{s}_m)).\end{aligned}$$

Recall that $\bar{s}_m$ is chosen such that

$$\mathrm{KL}^{\otimes_n}(s^*, \bar{s}_m) \leq \inf_{s_m \in S_m} \mathrm{KL}^{\otimes_n}(s^*, s_m) + \frac{\delta_{\mathrm{KL}}}{n}.$$

Put $\kappa(\beta) = 1 + (\beta + \beta^2)$, and let $\epsilon_1 > 0$, we define θ_1 by $\kappa_0'(\frac{\kappa_1' + \kappa_2'}{\theta_1} + \frac{18}{\theta_1^2\rho}) = C_\rho\epsilon_1$ where C_ρ is defined by $C_\rho(d_H^{\otimes_n})^2(s^*, \hat{s}_{m'}) \leq \mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{m'})$, and put $\kappa_2 = \frac{C_\rho\epsilon_1}{\kappa_0}$. We get that

$$\begin{aligned}(1 - \epsilon_1)\,\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{m'}) \leq{}& \kappa(\beta)\,\mathrm{KL}^{\otimes_n}(s^*, s_m) + \frac{\mathrm{pen}(m)}{n} - \frac{\mathrm{pen}(m')}{n}\\ &+ \kappa(\beta)\frac{\delta_{\mathrm{KL}}}{n} + \frac{\eta' + \eta}{n} + \mathfrak{z}_{m'}^2\kappa_2 + (\beta + \beta^2)z_{m,m'}^2.\end{aligned}$$

Since $\tau \leq 1 \vee \tau$, if we choose β such that $(\beta + \beta^2)(\delta/2 + 1) = \alpha\theta_1^{-2}\beta^{-2}$, and if we put $\kappa_1 = \alpha\gamma^{-2}(\beta^{-2} + 1)$, since $1 \leq 1 \vee \tau$, using the expressions of $\mathfrak{z}_{m'}$ and $z_{m,m'}$, we get that

$$\begin{aligned}(1 - \epsilon_1)\,\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{m'}) \leq{}& \kappa(\beta)\,\mathrm{KL}^{\otimes_n}(s^*, s_m) + \frac{\mathrm{pen}(m)}{n} - \frac{\mathrm{pen}(m')}{n}\\ &+ \kappa(\beta)\frac{\delta_{\mathrm{KL}}}{n} + \frac{\eta' + \eta}{n}\\ &+ \kappa_2\theta_1^2\left(\sigma_{m'}^2 + \frac{w + w_{m'}}{n}\right)\\ &+ \kappa_1(1 \vee \tau)\frac{w + w_m + w_{m'}}{n}\\ \leq{}& \kappa(\beta)\,\mathrm{KL}^{\otimes n}(s^*, s_m) + \left(\frac{\mathrm{pen}(m)}{n} + \kappa_1(1 \vee \tau)\frac{w_m}{n}\right)\end{aligned}$$

$$- \frac{\mathrm{pen}(m')}{n} + \kappa_2 \theta_1^2 \left(\sigma_{m'}^2 + \frac{w_{m'}}{n} \right) + \kappa_1 (1 \vee \tau) \frac{w_{m'}}{n} + \frac{\delta_{\mathrm{KL}}}{n} + \frac{\eta' + \eta}{n} + (\kappa_2 \theta_1^2 + \kappa_1 (1 \vee \tau)) \frac{w}{n}.$$

Now, assume that $\kappa_1 \geq \kappa$ in inequality (8), we get

$$(1 - \epsilon_1)\, \mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{m'}) \leq \kappa(\beta)\, \mathrm{KL}^{\otimes_n}(s^*, s_m) + 2\frac{\mathrm{pen}(m)}{n} + \frac{\delta_{\mathrm{KL}}}{n} + \frac{\eta + \eta'}{n} + (\kappa_2 \theta_1^2 + \kappa_1 (1 \vee \tau)) \frac{w}{n}.$$

It only remains to sum up the tail bounds 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 $e^{-(w + w_m + w_{m'})}$,

$$\sum_{\substack{m \in \mathcal{M} \\ m' \in \mathcal{M}(m)}} e^{-(w + w_m + w_{m'})} \leq e^{-w} \sum_{(m,m') \in \mathcal{M} \times \mathcal{M}} e^{-(w_m + w_{m'})} = e^{-w} \left(\sum_{m \in \mathcal{M}} e^{-w_m} \right)^2 = \Omega^2 e^{-w}$$

from the Assumption 3.

We then have simultaneously for all $m \in \mathcal{M}$, for all $m' \in \mathcal{M}(m)$, except on a set with probability less than $\Omega^2 e^{-w}$,

$$(1 - \epsilon_1)\, \mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{m'}) \leq \kappa(\beta)\, \mathrm{KL}^{\otimes_n}(s^*, s_m) + 2\frac{\mathrm{pen}(m)}{n} + \frac{\delta_{\mathrm{KL}}}{n} + \frac{\eta + \eta'}{n} + \left(\kappa_2 \theta_1^2 + \kappa_1 (1 \vee \tau)\right) \frac{w}{n}.$$

It is in particular satisfied for all $m \in \check{\mathcal{M}}$ and $m' \in \check{\mathcal{M}}(m)$, and, since $\hat{m} \in \check{\mathcal{M}}(m)$ for all $m \in \check{\mathcal{M}}$, we deduce that except on a set with probability less than $\Omega^2 e^{-w}$,

$$\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{\hat{m}}) \leq \frac{1}{(1 - \epsilon_1)} \times \left(\inf_{m \in \check{\mathcal{M}}} \left\{ \kappa(\beta)\, \mathrm{KL}^{\otimes_n}(s^*, s_m) + 2\frac{\mathrm{pen}(m)}{n} \right\} + \frac{\delta_{\mathrm{KL}}}{n} + \frac{\eta + \eta'}{n} + \left(\kappa_2 \theta_1^2 + \kappa_1 (1 \vee \tau)\right) \frac{w}{n} \right).$$

By integrating over all $w > 0$, because for any non negative random variable Z and any $a > 0$, $\mathrm{E}(Z) = a \int_{z \geq 0} P(Z > az) dz$, we obtain that

$$\mathrm{E}\left(\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}_{\hat{m}}) - \frac{1}{(1 - \epsilon_1)} \inf_{m \in \check{\mathcal{M}}} \left(\kappa(\beta)\, \mathrm{KL}^{\otimes n}(s^*, s_m) + 2\frac{\mathrm{pen}(m)}{n} \right) - \frac{1}{(1 - \epsilon_1)} \frac{\delta_{\mathrm{KL}} + \eta + \eta'}{n} \kappa_0 \theta^2 \right) \leq \left(\kappa_2 \theta_1^2 + \kappa_1 (1 \vee \tau)\right) \frac{\Omega^2}{n}.$$

As δ_{KL} can be chosen arbitrary small, this implies that

$$\begin{aligned}
\mathrm{E}(\mathrm{JKL}^{\otimes_n}(s^*, \hat{s}_{\hat{m}})) \leq & \frac{1}{1-\epsilon_1} \mathrm{E}\left(\inf_{m \in \check{\mathcal{M}}} \kappa(\beta) \, \mathrm{KL}^{\otimes_n}(s^*, s_m) + \frac{\mathrm{pen}(m)}{n} \right) \\
& + \frac{\eta + \eta'}{n} + (\kappa_2 \theta_1^2 + \kappa_1(1 \vee \tau)) \frac{\Omega^2}{n} \\
\leq & C_1 \, \mathrm{E}\left(\inf_{m \in \check{\mathcal{M}}} \left(\inf_{t \in S_m} \mathrm{KL}^{\otimes_n}(s^*, t) \right) + \frac{\mathrm{pen}(m)}{n} \right) \\
& + C_2(1 \vee \tau) \frac{\Omega^2}{n} + \frac{\eta' + \eta}{n}
\end{aligned}$$

with $C_1 = \frac{2}{1-\epsilon_1}$ and $C_2 = \kappa_2 \theta_1^2 + \kappa_1$.

5.3. *Sketch of the proof of the oracle inequality 3.2*

To prove the Theorem 3.2, we have to apply the Theorem 5.1. Then, our model collection has to satisfy all the assumptions. Here, the model is defined by $m = (K, J)$. The Assumption 2 is true when we consider Gaussian densities. If s^* is bounded, with compact support, the assumption defined by (7) is satisfied. It is also true in other particular cases. Our model has o satisfy Assumption 1 and Assumption 3. Here we present only the main steps to prove these assumptions. All the technical details stand in Appendix.

5.3.1. *Assumption 1*

We could take $\phi_m(\varpi) = \int_0^{\varpi} \sqrt{\mathcal{H}_{[.]}(\epsilon, S_m, d_H^{\otimes_n})} d\epsilon$ for all $\varpi > 0$. It could be better to consider more local version of the integrated square root entropy, but the global one is enough in this case to define the penalty. As done in Cohen and Le Pennec in [6], we could decompose the entropy by

$$\mathcal{H}_{[.]}(\epsilon, \mathcal{S}^{\mathcal{B}}_{(K,J)}, d_H^{\otimes_n}) \leq \mathcal{H}_{[.]}(\epsilon, \Pi_K, d_H^{\otimes_n}) + K \mathcal{H}_{[.]}(\epsilon, \mathcal{F}_J, d_H^{\otimes_n})$$

where

$$\mathcal{S}^{\mathcal{B}}_{(K,J)} = \left\{ \begin{array}{c} y \in \mathbb{R}^q | x \in \mathbb{R}^p \mapsto s_{\xi}^{(K,J)}(y|x) = \sum_{k=1}^K \pi_k \varphi(y|\beta_k^{[J]} x, \Sigma_k) \\ \xi = \left\{ \pi_1, \ldots, \pi_K, \beta_1^{[J]}, \ldots, \beta_K^{[J]}, \Sigma_1, \ldots, \Sigma_K \right\} \in \tilde{\Xi}_{(K,J)} \\ \tilde{\Xi}_{(K,J)} = \Pi_K \times ([-A_\beta, A_\beta]^{q \times p})^K \times ([a_\Sigma, A_\Sigma]^q)^K \end{array} \right\}$$

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

$$\begin{aligned}
\mathcal{F}_J = \Big\{ & \varphi(.|\beta^{[J]} X, \Sigma); \beta^{[J]} \in [-A_\beta, A_\beta]^{q \times p}, \\
& \Sigma = \mathrm{diag}([\Sigma]_{1,1}, \ldots, [\Sigma]_{q,q}) \in [a_\Sigma, A_\Sigma]^q \Big\}
\end{aligned}$$

where φ denote the Gaussian density, and $A_\beta, a_\Sigma, A_\Sigma$ are absolute constants.

Calculus for the proportions We could apply a result proved by Wasserman and Genovese in [9] to bound the entropy for the proportions. We get that

$$\mathcal{H}_{[.]}(\epsilon, \Pi_K, d_H^{\otimes_n}) \leq \log\left(K(2\pi e)^{K/2}\left(\frac{3}{\epsilon}\right)^{K-1}\right).$$

Calculus for the Gaussian The family

$$B_\epsilon(\mathcal{F}_J) = \left\{ \begin{array}{l} l(y,x) = (1+\delta)^{-p^2q-3q/4}\varphi(y|\nu_J x, (1+\delta)^{-1/4}B^{[1]}) \\ u(y,x) = (1+\delta)^{p^2q+3q/4}\varphi(y|\nu_J x, (1+\delta)B^{[2]}) \\ B^{[a]} = \operatorname{diag}(b_{i(1)}, \dots, b_{i(q)}), \\ \text{ with } i \text{ a permutation, for } a \in \{1,2\}, \\ \quad\text{ and } \left\{ \begin{array}{l} b_l = (1+\delta)^{1-l/2}A_\Sigma, l \in \{2, \dots, N\} \\ \forall (z,j) \in J^c, \nu_{z,j} = 0 \\ \forall (z,j) \in J, \nu_{z,j} = \sqrt{c}\delta A_\Sigma u_{z,j} \end{array} \right. \end{array} \right\} \tag{13}$$

is an ϵ-bracket covering for $\mathcal{F}_J$, where $u_{z,j}$ is a net for the mean, N is the number of parameters needed to recover all the variance set, $\delta = \frac{1}{\sqrt{2}(p^2q+3/4q)}\epsilon$, and $c = \frac{5(1-2^{-1/4})}{8}$.

We obtain that

$$|B_\epsilon(\mathcal{F}_J)| \leq 2\left(\frac{2A_\beta}{\sqrt{c}A_\Sigma}\right)^{|J|}\left(\frac{A_\Sigma}{a_\Sigma}+\frac{1}{2}\right)\delta^{-1-|J|};$$

and then we get

$$\mathcal{H}_{[.]}(\epsilon, \mathcal{F}_J, d_H^{\otimes_n}) \leq \log\left(2\left(\frac{2A_\beta}{\sqrt{c}A_\Sigma}\right)^{|J|}\left(\frac{A_\Sigma}{a_\Sigma}+\frac{1}{2}\right)\delta^{-1-|J|}\right).$$

Proposition 5.3. *Put $D_{(K,J)} = K(1+|J|)$. For all $\epsilon \in (0,1)$,*

$$\mathcal{H}_{[.]}(\epsilon, \mathcal{S}^{\mathcal{B}}_{(K,J)}, d_H^{\otimes_n}) \leq \log(C) + D_{(K,J)}\log\left(\frac{1}{\epsilon}\right);$$

with

$$C = 2K(2\pi e)^{K/2}\left(\frac{2A_\beta}{\sqrt{c}A_\Sigma}\right)^{K|J|}3^{K-1}\left(\frac{A_\Sigma}{a_\Sigma}+\frac{1}{2}\right)^K.$$

Determination of a function ϕ We could take

$$\phi_{(K,J)}(\varpi) = \sqrt{D_{(K,J)}}\varpi\left[B(A_\beta, A_\Sigma, a_\Sigma) + \sqrt{\log\left(\frac{1}{\varpi \wedge 1}\right)}\right].$$

This function is non-decreasing, and $\varpi \mapsto \frac{\phi_{(K,J)}(\varpi)}{\varpi}$ is non-increasing.

The root $\varpi_{(K,J)}$ is the solution of $\phi_{(K,J)}(\varpi_{(K,J)}) = \sqrt{n}\varpi^2_{(K,J)}$. With the expression of $\phi_{(K,J)}$, we get

$$\varpi^2_{(K,J)} = \sqrt{\frac{D_{(K,J)}}{n}}\varpi \left[B(A_\beta, A_\Sigma, a_\Sigma) + \sqrt{\log\left(\frac{1}{\varpi_{(K,J)} \wedge 1}\right)} \right].$$

Nevertheless, we know that $\varpi^* = \sqrt{\frac{D_{(K,J)}}{n}} B(A_\beta, A_\Sigma, a_\Sigma)$ minimizes $\varpi_{(K,J)}$: we get

$$\varpi^2_{(K,J)} \leq \frac{D_{(K,J)}}{n} \left[2B^2(A_\beta, A_\Sigma, a_\Sigma) + \log\left(\frac{1}{\frac{D_{(K,J)}}{n} B^2(A_\beta, A_\Sigma, a_\Sigma) \wedge 1} \right) \right].$$

5.3.2. Assumption 3

We want to group models by their dimension.

Lemma 5.4. *The quantity*

$$card\{(K,J) \in \mathbb{N}^* \times \mathcal{P}(\{1,\dots,q\} \times \{1,\dots,p\}), D(K,J) = D\}$$

is upper bounded by

$$\begin{cases} 2^{pq} \text{ if } pq \leq D - q \\ \left(\frac{epq}{D-q}\right)^{D-q} \text{ otherwise.} \end{cases}$$

Proposition 5.5. *Consider the weight family $\{w_{(K,J)}\}_{(K,J) \in \mathcal{K} \times \mathcal{J}}$ defined by*

$$w_{(K,J)} = D_{(K,J)} \log\left(\frac{4epq}{(D_{(K,J)} - q) \wedge pq} \right).$$

Then we have $\sum_{(K,J) \in \mathcal{K} \times \mathcal{J}} e^{-w_{(K,J)}} \leq 2$.

Appendix: Technical results

In this appendix, we give more details for the proofs.

A.1. Bernstein's lemma

Lemma A.6 (Bernstein's inequality). *Let $(X_1, \dots, X_n)$ be independent real valued random variables. Assume that there exists some positive numbers v and c such that $\sum_{i=1}^n \mathrm{E}(X_i^2) \leq v$, and, for all integers $K \geq 3$, $\sum_{i=1}^n \mathrm{E}((X_i)_+^K) \leq \frac{K!}{2} v c^{K-2}$. Let $S = \sum_{i=1}^n (X_i - \mathrm{E}(X_i))$. Then, for every positive x,*

$$P(S \geq \sqrt{2vx} + cx) \leq \exp(-x).$$

A.2. Proof of Lemma 5.2

This proof is adapted from the one of Maugis-Rabusseau and Meynet in [14]. Begin by a lemma.

Lemma A.7. *Let $\tau > 0$. For all $x > 0$, consider*

$$f(x) = x\log(x)^2, \qquad h(x) = x\log(x) - x + 1, \qquad \phi(x) = e^x - x - 1.$$

Then, for all $0 < x < e^\tau$, we get

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

To prove this, we have to show that $y \mapsto \frac{\phi(y)}{y^2}$ is non-decreasing. We omit the proof here.

We want to apply this inequality, in order to derive the Lemma 5.2. As $\log(||\frac{s^*}{\bar{s}_m}||_\infty) \leq \tau$,

$$\left|\left|\frac{s^*}{\bar{s}_m}\right|\right|_\infty \leq e^\tau;$$

and we could apply the previous inequality to $s^*/\bar{s}_m$. We get, for all x, for all y,

$$f\left(\frac{s^*(y|x)}{\bar{s}_m(y|x)}\right) \leq \frac{\tau^2}{\phi(-\tau)} h\left(\frac{s^*(y|x)}{\bar{s}_m(y|x)}\right).$$

Integrating with respect to the density $\bar{s}_m$, we get that

$$\begin{aligned}
&\int_{\mathbb{R}^q} \frac{s^*(y|.)}{\bar{s}_m(y|.)} \log\left(\frac{s^*(y|.)}{\bar{s}_m(y|.)}\right)^2 \bar{s}_m(y|.)dy \\
\leq &\int_{\mathbb{R}^q} \frac{\tau^2}{e^{-\tau} - \tau - 1} \left(\frac{s^*(y|.)}{\bar{s}_m(y|.)} \log \frac{s^*(y|.)}{\bar{s}_m(y|.)} - \frac{s^*(y|.)}{\bar{s}_m(y|.)} + 1\right) \bar{s}_m(y|.)dy \\
\Longleftrightarrow &\frac{1}{n}\sum_{i=1}^n \int s^*(y|x_i) \log\left(\frac{s^*(y|x_i)}{\bar{s}_m(y|x_i)}\right)^2 dy \\
\leq &\frac{\tau^2}{e^{-\tau} - \tau - 1} \frac{1}{n}\sum_{i=1}^n \int s^*(y|x_i) \log \frac{s^*(y|x_i)}{\bar{s}_m(y|x_i)} dy.
\end{aligned}$$

It concludes the proof.

A.3. Determination of a net for the mean and the variance

In this subsection, we work with a Gaussian density. We denote by $\beta \in \mathbb{R}^{q\times p}$ the conditional mean and by Σ the diagonal covariance.

- **Step 1: construction of a net for the variance**
  Let $\epsilon \in (0,1]$, and $\delta = \frac{1}{\sqrt{2}(p^2q+\frac{3}{4}q)}\epsilon$. Let $b_j = (1+\delta)^{1-\frac{j}{2}}A_\Sigma$. For $2 \leq j \leq N$, we have $[a_\Sigma, A_\Sigma] = [b_N, b_{N-1}]\bigcup \dots \bigcup [b_3, b_2]$, where N is chosen to recover everything. We want that

$$\begin{aligned} & a_\Sigma = (1+\delta)^{1-N/2}A_\Sigma \\ \Leftrightarrow \qquad & \log \frac{a_\Sigma}{A_\Sigma} = \left(1 - \frac{N}{2}\right)\log(1+\delta) \\ \Leftrightarrow \qquad & N = \frac{2\log(\frac{A_\Sigma}{a_\Sigma}\sqrt{1+\delta})}{\log(1+\delta)}. \end{aligned}$$

  We want N to be an integer, then $N = \lceil \frac{2\log(\frac{A_\Sigma}{a_\Sigma}\sqrt{1+\delta})}{\log(1+\delta)} \rceil$. We get a net for the variance. We could let $B = \text{diag}(b_{i(1)}, \dots, b_{i(q)})$, close to Σ (and deterministic, independent of the values of Σ), where i is a permutation such that $b_{i(z)+1} \leq [\Sigma]_{z,z} \leq b_{i(z)}$ for all $z \in \{1, \dots, q\}$. Remember that $\frac{b_{j+1}}{b_j} = \frac{1}{\sqrt{1+\delta}}$.
- **Step 2: construction of a net for the mean vectors**
  We select only the relevant variables detected by the Lasso estimator. For $\lambda \geq 0$,

$$J_\lambda = \left\{(z,j) \in \{1,\dots,q\} \times \{1,\dots,p\} | \hat{\beta}^{\text{Lasso}}_{z,j}(\lambda) \neq 0\right\}.$$

  Let $f = \varphi(.|\beta x, \Sigma) \in \mathcal{F}_{J_\lambda}$.
  - **Definition of the brackets**
    Define the bracket by the functions l and u:

$$\begin{aligned} l(y,x) &= (1+\delta)^{-p^2q-3q/4}\varphi\left(y|\nu_J x, (1+\delta)^{-1/4}B^{[1]}\right); \\ u(y,x) &= (1+\delta)^{p^2q+3q/4}\varphi\left(y|\nu_J x, (1+\delta)B^{[2]}\right). \end{aligned}$$

    We have chosen i such that $[B^{[1]}]_{z,z} \leq \Sigma_{z,z} \leq [B^{[2]}]_{z,z}$ for all $z \in \{1, \dots, q\}$.
    We need to define ν such that $[l,u]$ is an ϵ-bracket for f.
  - **Proof that $[l,u]$ is a bracket for f**
    We are looking for a condition on ν_J to have $\frac{f}{u} \leq 1$ and $\frac{l}{f} \leq 1$.
    We use the following lemma to compute these ratios.

    Lemma A.8. *Let $\varphi(.|\mu_1, \Sigma_1)$ and $\varphi(.|\mu_2, \Sigma_2)$ be two Gaussian densities. If their variance matrices are assumed to be diagonal, with $\Sigma_a = \text{diag}([\Sigma_a]_{1,1}, \dots, [\Sigma_a]_{q,q})$ for $a \in \{1,2\}$, such that $[\Sigma_2]_{z,z} > [\Sigma_1]_{z,z} > 0$ for all $z \in \{1, \dots, q\}$, then, for all $y \in \mathbb{R}^q$,*

$$\frac{\varphi(y|\mu_1, \Sigma_1)}{\varphi(y|\mu_2, \Sigma_2)}$$

$$\leq \prod_{z=1}^{q} \frac{\sqrt{[\Sigma_2]_{z,z}}}{\sqrt{[\Sigma_1]_{z,z}}} e^{\frac{1}{2}(\mu_1-\mu_2)^t \operatorname{diag}\left(\frac{1}{[\Sigma_2]_{1,1}-[\Sigma_1]_{1,1}},\dots,\frac{1}{[\Sigma_2]_{q,q}-[\Sigma_1]_{q,q}}\right)(\mu_1-\mu_2)}.$$

For the ratio $\frac{f}{u}$ we get:

$$\begin{aligned}\frac{f(y|x)}{u(y,x)} =& \frac{1}{(1+\delta)^{p^2q+3q/4}} \frac{\varphi(y|\beta x,\Sigma)}{\varphi(y|\nu_J x,(1+\delta)B^{[2]})} \\ \leq& \frac{1}{(1+\delta)^{p^2q+3q/4}} \prod_{z=1}^{q} \frac{b_z}{[\Sigma]_{z,z}} \\ & \times (1+\delta)^{q/2} e^{\frac{1}{2}(\beta x-\nu_J x)^t((1+\delta)B^{[2]}-\Sigma)^{-1}(\beta x-\nu_J x)} \\ \leq& (1+\delta)^{p^2q-q/4}(1+\delta)^{q/4} e^{\frac{1}{2}(\beta x-\nu_J x)^t(\delta B^{[2]})^{-1}(\beta x-\nu_J x)} \\ \leq& (1+\delta)^{p^2q} e^{\frac{1}{2\delta}(\beta x-\nu_J x)^t[B^{[2]}]^{-1}(\beta x-\nu_J x)}. \end{aligned} \tag{14}$$

For the ratio $\frac{l}{f}$ we get:

$$\begin{aligned}\frac{l(y,x)}{f(y|x)} =& \frac{1}{(1+\delta)^{p^2q+3q/4}} \frac{\varphi(y|\nu_J x,(1+\delta)^{-1/4}B^{[1]})}{\varphi(y|\beta x,\Sigma)} \\ \leq& \frac{1}{(1+\delta)^{p^2q+3q/4}} \prod_{z=1}^{q} \frac{\Sigma_{z,z}}{b_z} \\ & \times (1+\delta)^{q/8} e^{\frac{1}{2}(\beta x-\nu_J x)^t(\Sigma-B^{[1]})^{-1}(\beta x-\nu_J x)} \\ \leq& (1+\delta)^{-p^2q-3q/8}(1+\delta)^{q/4} \\ & \times e^{\frac{1}{2}(\beta x-\nu_J x)^t((1-(1+\delta)^{-1/4})B^{[1]})^{-1}(\beta x-\nu_J x)} \\ \leq& (1+\delta)^{-p^2q-3q/8} e^{\frac{1}{2(1-(1+\delta)^{-1/4})}(\beta x-\nu_J x)^t[B^{[1]}]^{-1}(\beta x-\nu_J x)}. \end{aligned} \tag{15}$$

We want to bound the ratios (14) and (15) by 1. Put $c = \frac{5(1-2^{-1/4})}{8}$, and develop these calculus. A necessary condition to obtain those bounds is

$$||\beta x - \nu_J x||_2^2 \leq pq\delta^2(1-2^{-1/4})A_\Sigma^2.$$

Indeed, we want

$$\begin{aligned}(1+\delta)^{-p^2q-3q/8} e^{\frac{1}{2(1-(1+\delta)^{-1/4})}(\beta x-\nu_J x)^t[B^{[2]}]^{-1}(\beta x-\nu_J x)} &\leq 1 \\ (1+\delta)^{-p^2q} e^{\frac{1}{2\delta A_\Sigma}(\beta x-\nu_J x)^t[B^{[1]}]^{-1}(\beta x-\nu_J x)} &\leq 1;\end{aligned}$$

which is equivalent to

$$\begin{aligned}||\beta x - \nu_J x||_2^2 &\leq p^2q\frac{\delta^2}{2}A_\Sigma^2; \\ ||\beta x - \nu_J x||_2^2 &\leq \left(p^2q + \frac{3}{4}q\right)\delta^2(1-2^{-1/4})A_\Sigma.\end{aligned}$$

As $||\beta x - \nu_J x||_2^2 \leq p||\beta - \nu_J||_2^2 ||x||_\infty$, and $x \in [0,1]^p$, we need to get $||\beta - \nu_J||_2^2 \leq pq\delta^2(1-2^{-1/4})A_\Sigma^2$ to have the wanted bound. Put

$$U := \mathbb{Z} \cap \left[\left\lfloor \frac{-A_\beta}{\sqrt{c}\delta A_\Sigma} \right\rfloor, \left\lfloor \frac{A_\beta}{\sqrt{c}\delta A_\Sigma} \right\rfloor \right].$$

For all $(z,j) \in J$, choose

$$u_{z,j} = \operatorname*{argmin}_{v_{z,j} \in U} \left| \beta_{z,j} - \sqrt{c}\delta A_\Sigma v_{z,j} \right|. \tag{16}$$

Define ν by

$$\begin{aligned} &\text{for all } (z,j) \in J^c, \nu_{z,j} = 0; \\ &\text{for all } (z,j) \in J \ , \nu_{z,j} = \sqrt{c}\delta A_\Sigma u_{z,j}. \end{aligned}$$

Then, we get a net for the mean vectors.

- **Proof that** $d_H(l,u) \leq \epsilon$

We work with the Hellinger distance.

$$\begin{aligned} d_H^2(l,u) &= \frac{1}{2} \int_{\mathbb{R}^q} (\sqrt{l} - \sqrt{u})^2 \\ &= \frac{1}{2} \int_{\mathbb{R}^q} l + u - 2\sqrt{lu} \\ &= \frac{1}{2} \left[(1+\delta)^{-p^2 q - 3q/4} + (1+\delta)^{p^2 q + 3q/4} \right] - \int_{\mathbb{R}^q} \sqrt{\varphi_l \varphi_u} \\ &= \frac{1}{2} \left[(1+\delta)^{-p^2 q - 3q/4} + (1+\delta)^{p^2 q + 3q/4} \right] \\ &\quad - \left(\prod_{z=1}^q \frac{\sqrt{2b_{i(z)+1} b_{i(z)}}(1+\delta)^{1/2}(1+\delta)^{-1/8}{}^2}{(1+\delta)b_{i(z)+1} + (1+\delta)^{-1/4} b_{i(z)}} \right)^{1/2} * 1. \end{aligned}$$

We have used the following lemma:

Lemma A.9. *The Hellinger distance of two Gaussian densities with diagonal variance matrices is given by the following expression:*

$$\begin{aligned} &d_H^2(\varphi(.|\mu_1, \Sigma_1), \varphi(.|\mu_2, \Sigma_2)) \\ &= 2 - 2 \left(\prod_{z=1}^q \frac{2\sqrt{[\Sigma_1]_{z,z}[\Sigma_2]_{z,z}}}{[\Sigma_1]_{z,z} + [\Sigma_2]_{z,z}} \right)^{1/2} \\ &\qquad \times e^{-\frac{1}{4}(\mu_1 - \mu_2)^t \operatorname{diag}\left(\left(\frac{1}{[\Sigma_1]_{z,z}^2 + [\Sigma_2]_{z,z}^2} \right)_{z \in \{1,\ldots,q\}} \right)(\mu_1 - \mu_2)} \end{aligned}$$

As $b_{i(z)+1} = (1+\delta)^{-1/2} b_{i(z)}$, we get that

$$2 \frac{(1+\delta)^{3/8} b_{i(z)}}{b_{i(z)+1} \left[(1+\delta)^{-1/4} + (1+\delta)^{1/2}(1+\delta) \right]}$$

$$
\begin{aligned}
&= 2\frac{(1+\delta)^{5/8}}{(1+\delta)^{-1/4}+(1+\delta)^{3/2}}\\
&= \frac{2}{(1+\delta)^{-7/8}+(1+\delta)^{7/8}}.
\end{aligned}
$$

Then

$$
\begin{aligned}
d_H^2(l,u) =& \frac{1}{2}\left[(1+\delta)^{-(p^2q+3q/4)}+(1+\delta)^{p^2q+3q/4}\right]\\
&- \left(\frac{2}{(1+\delta)^{-7/8}+(1+\delta)^{7/8}}\right)^{q/2}\\
=& \cosh((p^2q+3q/4)\log(1+\delta)) - 2\cosh(7/8\log(1+\delta))^{-q/2}\\
=& \cosh((p^2q+3q/4)\log(1+\delta)) - 1 + 1\\
&- 2^{-q/2}\cosh(7/8\log(1+\delta))^{-q/2}.
\end{aligned}
$$

We want to apply the Taylor formula to $f(x)=\cosh(x)-1$ to obtain an upper bound, and to $g(x)=1-2^{-q/2}\cosh(x)^{-q/2}$. Indeed, there exists c such that, on the good interval, $f(x)\leq \cosh(c)\frac{x^2}{2}$ and $g(x)\leq q^2\frac{x^2}{2}$. Then, and because $\log(1+\delta)\leq\delta$,

$$
\begin{aligned}
d_H^2(l,u) &\leq \cosh((p^2q+3q/4)\log(1+\delta)) - 2\cosh(7/8\log(1+\delta))^{-q/2}\\
&\leq (p^2q+3q/4)^2\delta^2\left(\cosh(\alpha)+\frac{49}{128}\right)\\
&\leq 2(p^2q+3q/4)^2\delta^2 \leq \epsilon^2;
\end{aligned}
$$

where $\epsilon \geq \sqrt{2}(p^2q+\frac{3}{4}q)\delta$.

- **Step 3: Upper bound of the number of ϵ-brackets for $\mathcal{F}_J$.** From Step 1 and Step 2, the family

$$
B_\epsilon(\mathcal{F}_J) = \left\{\begin{array}{l}
l(y,x) = (1+\delta)^{-(p^2q+3q/4)}\varphi(y|\nu_J x,(1+\delta)^{-1/4}B^{[1]})\\
u(y,x) = (1+\delta)^{p^2q+3q/4}\varphi(y|\nu_J x,(1+\delta)B^{[2]})\\
B^{[a]} = \mathrm{diag}(b^{[a]}_{i(1)},\ldots,b^{[a]}_{i(q)})\\
\text{ where } i_a \text{ is a permutation, for } a\in\{1,2\},\\
\text{ with } \left\{\begin{array}{l}
b_l = (1+\delta)^{1-l/2}A_\Sigma \text{ for all } l\in\{1,\ldots,q\}\\
\forall (z,j)\in J^c, \nu_{z,j}=0\\
\forall (z,j)\in J, \nu_{z,j}=\sqrt{c}\delta A_\Sigma u_{z,j}
\end{array}\right.
\end{array}\right\}
\tag{17}
$$

is an ϵ-bracket for $\mathcal{F}_J$, for $u_{z,j}$ defined by (16). Therefore, an upper bound of the number of ϵ-brackets necessary to cover $\mathcal{F}_J$ is deduced from an upper bound of the cardinal of $B_\epsilon(\mathcal{F}_J)$.

$$
|B_\epsilon(\mathcal{F}_J)| \leq \sum_{l=2}^{N}\prod_{(z,j)\in J}\left(\frac{2A_\beta}{\sqrt{c}\delta A_\Sigma}\right)
$$

$$\leq \left(\frac{2A_\beta}{\sqrt{c}\delta A_\Sigma}\right)^{|J|} \sum_{l=2}^{N} 1 \leq \left(\frac{2A_\beta}{\sqrt{c}\delta A_\Sigma}\right)^{|J|} (N-1).$$

As $N \leq \frac{2(A_\Sigma/a_\Sigma+1/2)}{\delta}$, we get

$$|B_\epsilon(\mathcal{F}_J)| \leq 2\left(\frac{2A_\beta}{\sqrt{c}A_\Sigma}\right)^{|J|}\left(\frac{A_\Sigma}{a_\Sigma}+\frac{1}{2}\right)\delta^{-1-|J|}.$$

A.4. Calculus for the function ϕ

From the Proposition 5.3, we obtain, for all $\varpi > 0$,

$$\int_0^{\varpi} \sqrt{\mathcal{H}_{[.]}(\epsilon, \mathcal{S}^{\mathcal{B}}_{(K,J)}, d_H^{\otimes_n})}d\epsilon \leq \varpi\sqrt{\log(C)} + \sqrt{D_{(K,J)}}\int_0^{\varpi\wedge 1}\sqrt{\log\left(\frac{1}{\epsilon}\right)}d\epsilon \quad (18)$$

We need to control $\int_0^{\varpi}\sqrt{\log(\frac{1}{\epsilon})}d\epsilon$, which is done by Maugis-Rabusseau and Meynet in [14].

Lemma A.10. *For all $\varpi > 0$,*

$$\int_0^{\varpi}\sqrt{\log\left(\frac{1}{\epsilon}\right)}d\epsilon \leq \varpi\left[\sqrt{\pi}+\sqrt{\log\left(\frac{1}{\varpi}\right)}\right].$$

Then, according to (18),

$$\begin{aligned}
&\int_0^{\varpi}\sqrt{\mathcal{H}_{[.]}(\epsilon, \mathcal{S}^{\mathcal{B}}_{(K,J)}, d_H^{\otimes_n})}d\epsilon \\
\leq &\varpi\sqrt{\log(C)} + \sqrt{D_{(K,J)}}(\varpi\wedge 1)\left[\sqrt{\pi}+\sqrt{\log\left(\frac{1}{\varpi\wedge 1}\right)}\right] \\
\leq &\varpi\sqrt{D_{(K,J)}}\left[\sqrt{\frac{\log(C)}{D_{(K,J)}}}+\sqrt{\pi}+\sqrt{\log\left(\frac{1}{\varpi\wedge 1}\right)}\right]
\end{aligned}$$

Nevertheless,

$$\begin{aligned}
\log(C) \leq &\log(2)+\log(K)+\frac{K}{2}\log(2\pi e) \\
&+K|J|\log\left(\frac{2A_\beta}{\sqrt{c}A_\Sigma}\right)+K\log\left(\frac{A_\Sigma}{a_\Sigma}+\frac{1}{2}\right)+(K-1)\log(3) \\
\leq &D_{(K,J)}\Bigg[\log(2)+\log(\sqrt{2\pi e})+1+\log(3) \\
&\qquad +\log\left(\frac{A_\Sigma}{a_\Sigma}+\frac{1}{2}\right)+\log\left(\frac{2A_\beta}{\sqrt{c}A_\Sigma}\right)\Bigg]
\end{aligned}$$

$$\leq D_{(K,J)} \left[1 + \log\left(\frac{A_\beta}{A_\Sigma} \left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2} \right) \right) + \log\left(2^{5/2} 3 \sqrt{\frac{\pi e}{c}} \right) \right].$$

Then

$$\begin{aligned}
&\int_0^{\varpi} \sqrt{\mathcal{H}_{[.]}(\epsilon, \mathcal{S}^{\mathcal{B}}_{(K,J)}, d_H^{\otimes_n})} d\epsilon \\
\leq & \varpi \sqrt{D_{(K,J)}} \left[\sqrt{1 + \log\left(\frac{A_\beta}{A_\Sigma} \left(\frac{a_\Sigma}{A_\Sigma} + \frac{1}{2} \right) \right) + \log\left(2^{5/2} 3 \sqrt{\frac{\pi e}{c}} \right)} \right. \\
&\qquad\qquad \left. + \sqrt{\pi} + \sqrt{\log\left(\frac{1}{\varpi \wedge 1} \right)} \right] \\
\leq & \varpi \sqrt{D_{(K,J)}} \left[1 + \sqrt{\log\left(\frac{A_\beta}{A_\Sigma} \left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2} \right) \right)} + a + \sqrt{\log\left(\frac{1}{\varpi \wedge 1} \right)} \right] \\
\leq & \varpi \sqrt{D_{(K,J)}} \left[B(A_\beta, A_\Sigma, a_\Sigma) + \sqrt{\log\left(\frac{1}{\varpi \wedge 1} \right)} \right];
\end{aligned}$$

with

$$B(A_\beta, A_\Sigma, a_\Sigma) = 1 + \sqrt{\log\left(\frac{A_\beta}{A_\Sigma} \left(\frac{A_\Sigma}{a_\Sigma} + \frac{1}{2} \right) \right)} + a;$$

and $a = \sqrt{\pi} + \sqrt{\log(2^{5/2} 3 \sqrt{\frac{\pi e}{c}})}$.

A.5. Proof of the Proposition 5.5

We are interested in $\sum_{(K,J) \in \mathcal{K} \times \mathcal{J}} e^{-w_{(K,J)}}$. Considering

$$w_{(K,J)} = D_{(K,J)} \log\left(\frac{4epq}{(D_{(K,J)} - q^2) \wedge pq} \right),$$

we could group models by their dimensions to compute this sum. Denote by C_D the cardinal of models of dimension D.

$$\begin{aligned}
&\sum_{\substack{K \in \mathbb{N}^* \\ J \in \mathcal{P}(\{1,\dots,q\} \times \{1,\dots,p\})}} e^{-D_{(K,J)} \log\left(\frac{4epq}{(D_{(K,J)} - q^2) \wedge pq} \right)} = \sum_{D \geq 1} C_D e^{-D \log\left(\frac{4epq}{(D - q^2) \wedge pq} \right)} \\
&= \sum_{D=1}^{pq+q^2} e^{-D \log\left(\frac{4epq}{(D-q^2)} \right)} \left(\frac{epq}{D - q^2} \right)^{D - q^2} + \sum_{D = pq+q^2+1}^{+\infty} e^{-D \log\left(\frac{4epq}{pq} \right)} 2^{pq} \\
&= \sum_{D=1}^{pq+q^2} 4^{-D} \left(\frac{epq}{D - q^2} \right)^{-q^2} + \sum_{D = pq+q^2+1}^{+\infty} e^{-D(\log(4)+1) + pq\log(2)} \\
&\leq \sum_{D=1}^{pq+q^2} 2^{-D} + \sum_{D = pq+q^2+1}^{+\infty} 2^{-D} = 2.
\end{aligned}$$

A.6. Proof of the Lemma 5.4

We know that $D_{(K,J)} = K - 1 + |J|K + Kq$. Then,

$$\begin{aligned} C_D &= \text{card}\{(K,J) \in \mathbb{N}^* \times \mathcal{P}(\{1,\ldots,q\} \times \{1,\ldots,p\}), D(K,J) = D\} \\ &\leq \sum_{K \in \mathbb{N}^*} \sum_{\substack{1 \leq z \leq q \\ 1 \leq j \leq p}} \binom{pq}{|J|} \mathbb{1}_{K(|J|+q+1)-1=D} \\ &\leq \sum_{|J| \in \mathbb{N}^*} \binom{pq}{|J|} \mathbb{1}_{|J| \leq pq \wedge (D-q)}. \end{aligned}$$

If $pq < D - q$,

$$\sum_{|J|>0} \binom{pq}{|J|} \mathbb{1}_{|J| \leq pq \wedge (D-q)} = 2^{pq}.$$

Otherwise, according to the Proposition 2.5 in Massart, [11],

$$\sum_{|J|>0} \binom{pq}{|J|} \mathbb{1}_{|J| \leq pq \wedge (D-q)} \leq f(D-q)$$

where $f(x) = (epq/x)^x$ is an increasing function on $\{1,\ldots,pq\}$. As pq is an integer, we get the result.

Acknowledgment

I am grateful to Pascal Massart for suggesting me to study this problem, and for stimulating discussions. I also thank referees for interesting comments and remarks which lead to improve this article.

References

[1] Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. MR0423716

[2] Baraud, Y., Giraud, C., and Huet, S. Gaussian model selection with an unknown variance. *The Annals of Statistics*, 37(2):630–672, 2009. URL http://dx.doi.org/10.1214/07-AOS573. MR2502646

[3] Belloni, A. and Chernozhukov, V. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. ISSN 1350-7265. URL http://dx.doi.org/10.3150/11-BEJ410. MR3037163

[4] Bickel, P., Ritov, Y., and Tsybakov, A. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009. ISSN 0090-5364. URL http://dx.doi.org/10.1214/08-AOS620. MR2533469

[5] Birgé, L. and Massart, P. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1–2), 2007. MR2288064

[6] Cohen, S. and Le Pennec, E. Conditional density estimation by penalized likelihood model selection and applications. Research Report RR-7596, 2011. URL https://hal.inria.fr/inria-00575462.

[7] Devijver, E. Model-based clustering for high-dimensional data. Application to functional data, 2014. arXiv:1409.1333.

[8] Ferraty, F. and Vieu, P. *Nonparametric functional data analysis: Theory and practice*. Springer series in statistics. Springer, New York, 2006. ISBN 0-387-30369-3. URL http://opac.inria.fr/record=b1128550. MR2229687

[9] Genovese, C. and Wasserman, L. Rates of convergence for the Gaussian mixture sieve. *Annals of Statistics*, 28(4):1105–1127, 2000. ISSN 0090-5364. URL http://dx.doi.org/10.1214/aos/1015956709. MR1810921

[10] Guo, J., Levina, E., Michailidis, G., and Zhu, J. Pairwise variable selection for high-dimensional model-based clustering. *Biometrics*, 66(3):793–804, 2010. ISSN 0006-341X; 1541-0420/e. MR2758215

[11] Massart, P. *Concentration inequalities and model selection*. Lecture Notes in Mathematics. Springer, 33, 2003, Saint-Flour, Cantal, 2007. ISBN 978-3-540-48497-4. URL http://opac.inria.fr/record=b1122538. MR2319879

[12] Massart, P. and Meynet, C. The Lasso as an ℓ_1-ball model selection procedure. *Electronic Journal of Statistics*, 5:669–687, 2011. ISSN 1935-7524. URL http://dx.doi.org/10.1214/11-EJS623. MR2820635

[13] Meinshausen, N. and Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010. ISSN 13697412. URL http://dx.doi.org/10.1111/j.1467-9868.2010.00740.x. MR2758523

[14] Meynet, C. and Maugis-Rabusseau, C. A sparse variable selection procedure in model-based clustering. Research report, Sept. 2012. URL https://hal.inria.fr/hal-00734316.

[15] Pan, W. and Shen, X. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8:1145–1164, 2007. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1314498.1314537.

[16] Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. MR0468014

[17] Städler, N., Bühlmann, P., and van de Geer, S. ℓ_1-penalization for mixture regression models. *Test*, 19(2):209–256, 2010. MR2677722

[18] Sun, T. and Zhang, C.-H. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012. ISSN 0006-3444. URL http://dx.doi.org/10.1093/biomet/ass043. MR2999166

[19] Sun, W., Wang, J., and Fang, Y. Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics*, 6:148–167, 2012. URL http://dx.doi.org/10.1214/12-EJS668. MR2879675

[20] Thalamuthu, A., Mukhopadhyay, I., Zheng, X., and Tseng, G. Evaluation and comparison of gene clustering methods in microar-

ray analysis. *Bioinformatics*, 22(19):2405–2412, 2006. URL http://bioinformatics.oxfordjournals.org/content/22/19/2405.abstract.

[21] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B.*, 58(1):267–288, 1996. MR1379242

[22] van de Geer, S. and Bühlmann, P. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009. ISSN 1935-7524. URL http://dx.doi.org/10.1214/09-EJS506. MR2576316

[23] Yang, Y. and Barron, A. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999. ISSN 0090-5364. URL http://dx.doi.org/10.1214/aos/1017939142. MR1742500

[24] Zhao, P. and Yu, B. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1248547.1248637. MR2274449

[25] Zhou, H., Pan, W., and Shen, X. Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics*, 3:1473–1496, 2009. URL http://dx.doi.org/10.1214/09-EJS487. MR2578834