ESAIM: PS 17 (2013) 650–671
DOI: 10.1051/ps/2012016

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

AN ℓ_1-ORACLE INEQUALITY FOR THE LASSO IN FINITE MIXTURE GAUSSIAN REGRESSION MODELS

Caroline Meynet[1]

Abstract. We consider a finite mixture of Gaussian regression models for high-dimensional heterogeneous data where the number of covariates may be much larger than the sample size. We propose to estimate the unknown conditional mixture density by an ℓ_1-penalized maximum likelihood estimator. We shall provide an ℓ_1-oracle inequality satisfied by this Lasso estimator with the Kullback–Leibler loss. In particular, we give a condition on the regularization parameter of the Lasso to obtain such an oracle inequality. Our aim is twofold: to extend the ℓ_1-oracle inequality established by Massart and Meynet [12] in the homogeneous Gaussian linear regression case, and to present a complementary result to Städler *et al.* [18], by studying the Lasso for its ℓ_1-regularization properties rather than considering it as a variable selection procedure. Our oracle inequality shall be deduced from a finite mixture Gaussian regression model selection theorem for ℓ_1-penalized maximum likelihood conditional density estimation, which is inspired from Vapnik's method of structural risk minimization [23] and from the theory on model selection for maximum likelihood estimators developed by Massart in [11].

Mathematics Subject Classification. 62G08, 62H30.

Received January 7, 2012. Revised July 17, 2012.

1. Introduction

In applied statistics, tremendous number of applications deal with relating a random response variable Y to a set of explanatory variables or covariates X through a regression-type model. As a consequence, linear regression $Y = X\beta + \epsilon$ is one of the most studied fields in statistics. Due to computer progress and development of state of the art technologies such as DNA microarrays, we are faced with high-dimensional data where the number of variables can be much larger than the sample size. To solve this problem, the sparsity scenario – which consists in assuming that the coefficients of the high-dimensional vector of covariates are mostly 0 – has been widely studied (see [6, 15] among others). These last years, a great deal of attention [19, 20, 25] has been focused on the ℓ_1-penalized least squares estimator of parameters,

$$\widehat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \quad \lambda > 0,$$

Keywords and phrases. Finite mixture of Gaussian regressions model, Lasso, ℓ_1-oracle inequalities, model selection by penalization, ℓ_1-balls.

[1] Laboratoire de Mathématiques, Faculté des Sciences d'Orsay, Université Paris-Sud, 91405 Orsay, France.
caroline.meynet@math.u-psud.fr

Article published by EDP Sciences

which is called the Lasso according to the terminology of Tibshirani [19] who first introduced this estimator in such a context. This interest has been motivated by the geometric properties of the ℓ_1-norm: ℓ_1-penalization tends to produce sparse solutions and can be thus used as a convex surrogate for the non-convex ℓ_0-penalization. Thus, the Lasso has essentially been developed for sparse recovery based on convex optimization. In this sparsity approach, many results, such as ℓ_0-oracle inequalities, have been proved to study the performance of this estimator as a variable selection procedure ([3, 8, 9, 15, 21] among others). Nonetheless, all these results need strong restrictive eigenvalue assumptions on the Gram matrix X^TX that can be far from being fulfilled in practice (see [5] for an overview of these assumptions). In parallel, a few results on the performance of the Lasso for its ℓ_1-regularization properties have been established [1, 10, 12, 17]. In particular, Massart and Meynet [12] have provided an ℓ_1-oracle inequality for the Lasso in the framework of fixed design Gaussian regression. Contrary to the ℓ_0-results that require strong assumptions on the regressors, their ℓ_1-result is valid with no assumption at all.

In linear regression, the homogeneity assumption that the regression coefficients are the same for different observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ is often inadequate and restrictive. It seems all the more true for the case of high-dimensional data: at least a fraction of covariates may exhibit a different influence on the response among various observations (*i.e.* sub-populations) and parameters may change for different subgroups of observations. Thus, addressing the issue of heterogeneity in high-dimensional data is important in many practical applications. In particular, Städler *et al.* [18] have proved that substantial prediction improvements are possible by incorporating a heterogeneity structure to the model. Such heterogeneity can be modeled by a finite mixture of regressions model. Considering the important case of Gaussian models, we can then assume that, for all $i = 1, \ldots, n$, Y_i follows a law with density $s_\psi(\cdot|x_i)$ which is a finite mixture of K Gaussian densities with proportion vector π,

$$Y_i|X_i = x_i \;\sim\; s_\psi(Y_i|x_i) = \sum_{k=1}^{K} \frac{\pi_k}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\left(Y_i - \mu_k^T x_i\right)^2}{2\sigma_k^2}\right)$$

for some parameter $\psi = (\pi_k, \mu_{kj}, \sigma_k)_{k,j}$.

In spite of the possible advantage of considering finite mixture regression models in high-dimensional data, very few studies have been made on these models. Yet, one can mention Städler *et al.* [18] who propose an ℓ_1-penalized maximum likelihood estimator,

$$\widehat{s}(\lambda) = \underset{s_\psi}{\operatorname{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \ln\left(s_\psi(Y_i|x_i)\right) + \lambda \sum_{k=1}^{K}\sum_{j=1}^{p} |\mu_{kj}| \right\}, \tag{1.1}$$

and provide an ℓ_0-oracle inequality satisfied by this Lasso estimator. Since they work in a sparsity approach, their oracle inequality is based on the same restricted eigenvalue conditions used in the homogeneous linear regression described above. Moreover, the negative ln-likelihood function used for maximum likelihood estimation requires additional mathematical arguments in comparison to the quadratic loss used in the homogeneous linear regression case. In particular, Städler *et al.* [18] have to introduce some margin assumptions so as to link the Kullback–Leibler loss function to the ℓ_2-norm of the parameters and get optimal rates of convergence of order $\|s_\psi\|_0/n$.

In this paper, we propose another approach that does not take into account sparsity. We shall rather study the performance of the Lasso estimator in the framework of finite mixture Gaussian regression models for its ℓ_1-regularization properties, thus extending the results presented in [12] for homogeneous Gaussian linear regression models. As in [12], we shall restrict to the fixed design case, that is to say non-random regressors. We aim at providing an ℓ_1-oracle inequality satisfied by the Lasso with no assumption neither on the Gram matrix nor on the margin. This can be achieved due to the fact that we are only looking for rates of convergence of order $\|s_\psi\|_1/\sqrt{n}$ rather than $\|s_\psi\|_0/n$. We give a lower bound on the regularization parameter λ of the Lasso

in (1.1) to guarantee such an oracle inequality,

$$\lambda \geq CK\left(\ln n\right)^2 \sqrt{\frac{\ln(2p+1)}{n}}, \tag{1.2}$$

where C is a positive quantity depending on the parameters of the mixture and on the regressors whose value is specified in (3.1). Our result is non-asymptotic: the number n of observations is fixed while the number p of covariates can grow with respect to n and can be much larger than n. The numbers K of clusters in the mixture is fixed. A great attention has been paid to obtain a lower bound (1.2) of λ with optimal dependence on p, that is to say $\sqrt{\ln(2p+1)}$ just as in the case of homogeneous Gaussian linear regression in [12].

Our oracle inequality shall be deduced from a finite mixture Gaussian regression model selection theorem for ℓ_1-penalized maximum likelihood conditional density estimation that we establish by following both Vapnik's method of structural risk minimization [23] and the theory [7, 11] around model selection. Just as in [12], the key idea that enables us to deduce our ℓ_1-oracle inequality from such a model selection theorem is to view the Lasso as the solution of a penalized maximum likelihood model selection procedure over a countable collection of ℓ_1-ball models.

The article is organized as follows. The notations and the framework are introduced in Section 2. In Section 3, we state the main result of the article, which is an ℓ_1-oracle inequality satisfied by the Lasso in finite mixture Gaussian regression models. Section 4 is devoted to the proof of this result: in particular, we state and prove the model selection theorem from which it is derived. Finally, some lemmas are proved in Section 5.

2. Notations and framework

2.1. The models

Our statistical framework is a finite mixture of Gaussian regressions model for high-dimensional data where the number of covariates can be much larger than the sample size. We observe n couples $((x_i, Y_i))_{1\leq i\leq n}$ of variables. We are interested in estimating the law of the random variable $Y_i \in \mathbb{R}$ conditionally to the fixed one $x_i \in \mathbb{R}^p$. We assume that the couples (x_i, Y_i) are independent while Y_i depends on x_i through its law. More precisely, we assume that the covariates x_is are independent but not necessarily identically distributed. The assumption on the Y_is are stronger: we assume that, conditionally to the x_is, they are independent and each variable Y_i follows a law with density $s_0(\cdot|x_i)$ which is a finite mixture of K Gaussian densities. Our goal is to estimate this two-variables conditional density function s_0 from the observations.

The model under consideration can be written as follows:

$$\begin{aligned}
&Y_i|x_i \text{ independent}\\
&Y_i|x_i = x \sim s_\psi(y|x)\mathrm{d}y\\
&s_\psi(y|x) = \sum_{k=1}^{K} \frac{\pi_k}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\left(y-\mu_k^T x\right)^2}{2\sigma_k^2}\right),\\
&\psi = \left(\mu_1^T, \ldots, \mu_K^T, \sigma_1, \ldots, \sigma_K, \pi_1, \ldots, \pi_K\right) \in \left(\mathbb{R}^{pK} \times \mathbb{R}_{>0}^K \times \mathit{\Pi}\right),\\
&\mathit{\Pi} = \left\{\pi = (\pi_1, \ldots, \pi_K) :\ \pi_k > 0 \text{ for } k = 1, \ldots, K \text{ and } \sum_{k=1}^{K} \pi_k = 1\right\}.
\end{aligned}$$

The μ_ks are the vectors of regression coefficients, the σ_ks are the standard deviations in mixture component k while the π_ks are the mixture coefficients.

For all $x \in \mathbb{R}^p$, we define the parameter $\psi(x)$ of the conditional density $s_\psi(.|x)$ by

$$\psi(x) = \left(\mu_1^T x, \ldots, \mu_K^T x, \sigma_1, \ldots, \sigma_K, \pi_1, \ldots, \pi_K\right) \in \mathbb{R}^{3K}.$$

For all $k = 1, \ldots, K$, $\mu_k^T x$ is the mean coefficient of the mixture component k for the conditional density $s_\psi(.|x)$.

Since we are working conditionally to the covariates $(x_i)_{1 \leq i \leq n}$, our results shall be expressed with quantities depending on them. In particular, we shall consider the following notation:

$$\|x\|_{\max,n} := \sqrt{\frac{1}{n} \sum_{i=1}^{n} \max_{j=1,\ldots,p} x_{ij}^2}.$$

2.2. Boundedness assumption on the mixture and component parameters

For technical reasons, we shall restrict our study to bounded parameter vectors $\psi = (\mu_k^T, \sigma_k, \pi_k)_{k=1,\ldots,K}$. Specifically, we shall assume that there exist deterministic positive quantities a_μ, A_μ, a_σ, A_σ and a_π such that the parameter vectors belong to the bounded space

$$\Psi = \left\{ \psi : \forall k = 1, \ldots, K,\ a_\mu \leq \inf_{x \in \mathbb{R}^p} \left|\mu_k^T x\right| \leq \sup_{x \in \mathbb{R}^p} \left|\mu_k^T x\right| \leq A_\mu,\ a_\sigma \leq \sigma_k \leq A_\sigma,\ a_\pi \leq \pi_k \right\}. \tag{2.1}$$

We denote by S the set of conditional densities s_ψ in this model:

$$S = \left\{ s_\psi,\ \psi \in \left(\mathbb{R}^{pK} \times \mathbb{R}_{>0}^K \times \Pi\right) \cap \Psi \right\}.$$

To simplify the proofs, we shall also assume that the true density s_0 belongs to S, that is to say there exists ψ_0 such that

$$s_0 = s_{\psi_0}, \quad \psi_0 = \left(\mu_{0,k}^T, \sigma_{0,k}, \pi_{0,k}\right)_{k=1,\ldots,K} \in \left(\mathbb{R}^{pK} \times \mathbb{R}_{>0}^K \times \Pi\right) \cap \Psi.$$

2.3. The Lasso estimator

In a maximum likelihood approach, the loss function taken into consideration is the Kullback–Leibler information, which is defined for two densities s and t by

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

if sdy is absolutely continuous with respect to tdy and $+\infty$ otherwise.

Since we are working with conditional densities and not with classical densities, we define the following adapted Kullback–Leibler information that takes into account the structure of conditional densities. For fixed covariates $x_1, \ldots, x_n$, we consider the average loss function

$$\mathrm{KL}_{\mathrm{n}}(s,t) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{KL}(s(\cdot|x_i), t(\cdot|x_i)) = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} \ln\left(\frac{s(y|x_i)}{t(y|x_i)}\right) s(y|x_i)\ \mathrm{d}y. \tag{2.2}$$

The maximum likelihood approach suggests to estimate s_0 by the conditional density s_ψ that maximizes the likelihood conditionally to $(x_i)_{1 \leq i \leq n}$,

$$\ln\left(\prod_{i=1}^{n} s_\psi(Y_i|x_i)\right) = \sum_{i=1}^{n} \ln\left(s_\psi(Y_i|x_i)\right),$$

or equivalently that minimizes the empirical contrast which is $-\sum_{i=1}^{n} \ln(s_\psi(Y_i|x_i))/n$. But since we want to deal with high-dimensional data, we have to regularize the maximum likelihood estimator in order to obtain reasonably accurate estimates. Here, we shall consider ℓ_1-regularization and its associated so-called Lasso estimator which is the following ℓ_1-norm penalized maximum likelihood estimator:

$$\widehat{s}(\lambda) := \operatorname*{argmin}_{s_\psi \in S} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln\left(s_\psi(Y_i|x_i)\right) + \lambda |s_\psi|_1 \right\}, \tag{2.3}$$

where $\lambda > 0$ is a regularization parameter to be tuned and

$$|s_\psi|_1 := \sum_{k=1}^{K} \|\mu_k\|_1 = \sum_{k=1}^{K}\sum_{j=1}^{p} |\mu_{kj}|$$

for $\psi = (\mu_k^T, \sigma_k, \pi_k)_{k=1,\dots,K}$ and $\mu_k = (\mu_{kj})_{j=1,\dots,p}$ for all $k = 1, \dots, K$.

3. An ℓ_1-ball regression mixture model selection theorem

3.1. An ℓ_1-oracle inequality for the Lasso in mixture Gaussian regression models

We state here the main result of the article: Theorem 3.1 provides an ℓ_1-oracle inequality satisfied by the Lasso estimator in finite mixture Gaussian regression models.

Theorem 3.1. *Denote $a \wedge b = \min(a, b)$. Assume that*

$$\lambda \geq \frac{\kappa}{a_\sigma \wedge a_\pi} \left(1 + \frac{\left(A_\mu^2 + A_\sigma^2 \ln(n)\right)}{a_\sigma^2}\right) \frac{\sqrt{K}}{\sqrt{n}} \left(1 + \|x\|_{\max,n} \ln(n) \sqrt{K \ln(2p+1)}\right) \tag{3.1}$$

for some absolute constant $\kappa \geq 360$. Then, the Lasso estimator $\widehat{s}(\lambda)$ defined by (2.3) satisfies the following ℓ_1-oracle inequality:

$$\mathbb{E}\left[\mathrm{KL}_{\mathrm{n}}\left(s_0, \widehat{s}(\lambda)\right)\right] \leq \left(1 + \kappa^{-1}\right) \inf_{s_\psi \in S} \left(\mathrm{KL}_{\mathrm{n}}(s_0, s_\psi) + \lambda |s_\psi|_1\right) + \lambda$$
$$+ \frac{\kappa' \sqrt{K}}{\sqrt{n}} \left[K \frac{\left(1 + \left(A_\mu + A_\sigma\right)^2\right)}{a_\sigma \wedge a_\pi} \left(1 + \frac{\left(A_\mu^2 + A_\sigma^2 \ln(n)\right)}{a_\sigma^2}\right) + a_\sigma \mathrm{e}^{-\frac{1}{2}\left(1 + \frac{a_\mu^2}{2A_\sigma^2}\right)}\right],$$

where κ' is an absolute positive constant.

Remark 3.2. We have not looked for optimizing the constants in Theorem 3.1. Thus, we do not explicit the value of κ' and the lower bound on κ is sufficient but not optimal.

Theorem 3.1 provides information about the performance of the Lasso as an ℓ_1-regularization algorithm. It highlights the fact that, provided that the regularization parameter λ is properly chosen, the Lasso estimator, which is the solution of the ℓ_1-penalized empirical risk minimization problem, behaves as well as the deterministic Lasso, that is to say the solution of the ℓ_1-penalized true risk minimization problem, up to an error term of order λ. This ℓ_1-result is complementary to the ℓ_0-oracle inequality in [18] whose is rather stated in a sparsity approach looking at the Lasso as a variable selection procedure.

Let us stress that we present here an ℓ_1-oracle inequality with no assumption neither on the Gram matrix nor on the margin. This represents a great advantage compared to the ℓ_0-oracle inequality in [18] which requires some restricted eigenvalue conditions as well as margin assumptions involving unknown constants. Indeed, if one may prove that these assumptions are actually fulfilled for some constants in the case of finite mixture regression models thanks to theoretical arguments such as continuity or differentiability of the functions into consideration, it seems nonetheless very hard to calculate explicit values of the constants for which these assumptions are fulfilled. One has thus no idea of the concrete values of these quantities. Yet, the ℓ_0-oracle inequality established in [18] strongly depends on these unknown quantities. So, it is difficult to interpret the precision of this result and it makes it hardly interpretable. On the contrary, the only assumption used to establish Theorem 3.1 is the boundedness of the parameters of the mixture, which is anyway also assumed in [18] and which is quite usual when working with maximum likelihood estimation [2, 13], at least to tackle the problem of the unboundedness of the likelihood at the boundary of the parameter space [14, 16] and to prevent it from divergence. In fact,

Städler *et al.* [18] must make their eigenvalue condition so as to bound the ℓ_2-norm of the parameter vector on its support and they add assumptions on the margin in order to link the loss function to the ℓ_2-norm of the parameters and get optimal rates of convergence $\|s_\psi\|_0/n$ in a sparsity viewpoint. On the opposite, since we are interested in an ℓ_1-regularization approach, we are just looking for rates of convergence of order $\|s_\psi\|_1/\sqrt{n}$ and we can avoid such restrictive vague assumptions.

Both our ℓ_1-oracle inequality and the ℓ_0-oracle inequality in [18] are valid for regularization parameters of the same order as regards the sample size n and the number of covariates p, that is $(\ln n)^2\sqrt{\ln(2p+1)/n}$. This means that if one considers a Lasso estimator with such a regularization parameter, then, even if one can not be sure that the Lasso indeed performs well as regards variable selection (because one can not have precise idea of the unknown constants present in Städler *et al.* [18]), one is at least guaranteed that the Lasso will act as a good ℓ_1-regularizator.

Our result is non-asymptotic: the number n of observations is fixed while the number p of covariates can grow with respect to n and can be much larger than n. The numbers K of clusters in the mixture is fixed. A great attention has been paid to obtain a lower bound (1.2) of λ with optimal dependence on p, which is the only parameter not to be fixed and which can grow with possibly $p \gg n$. We thus recover the same dependence $\sqrt{\ln(2p+1)}$ as for the homogeneous linear regression in [12]. On the contrary, the dependence on n for the homogeneous linear regression in [12] was $1/\sqrt{n}$ while we have an extra-$(\ln n)^2$ factor here. In fact, the linearity arguments developed in [12] with the quadratic loss function can not be exploited here with the non-linear Kullback–Leibler information. Entropy arguments are instead envisaged, leading to an extra-$\ln n$ factor. Contrary to Städler *et al.* [18], we have paid attention to giving an explicit dependence not only on n and p, but also on the number of clusters K in the mixture as well as on the regressors and all the quantities bounding the mixture parameters of the model. Nonetheless, we are aware of the fact that these dependences may not be optimal. In particular, we get a linear dependence on K in (3.1), while we might think that the true minimal dependence is only $\sqrt{K}$ (see Rem. 5.8 for more details).

4. Proof of Theorem 3.1

4.1. Statement of the main results

To prove Theorem 3.1, we look at the Lasso as the solution of a penalized maximum likelihood model selection procedure over a countable collection of ℓ_1-ball models. Using this basic idea, Theorem 3.1 is an immediate consequence of Theorem 4.1 stated below, which is an ℓ_1-ball mixture regression model selection theorem for ℓ_1-penalized maximum likelihood conditional density estimation in the Gaussian framework.

Theorem 4.1. *Assume we observe $((x_i,Y_i))_{1\le i\le n}$ with unknown conditional Gaussian mixture density s_0. For all $m\in\mathbb{N}^*$, consider the ℓ_1-ball*

$$S_m=\{s_\psi\in S,|s_\psi|_1\le m\} \tag{4.1}$$

and let $\widehat{s}_m$ be a η_m-ln-likelihood minimizer in S_m for some $\eta_m\ge 0$:

$$-\frac{1}{n}\sum_{i=1}^n\ln(\widehat{s}_m(Y_i|x_i))\le\inf_{s_m\in S_m}\left(-\frac{1}{n}\sum_{i=1}^n\ln(s_m(Y_i|x_i))\right)+\eta_m. \tag{4.2}$$

Assume that for all $m\in\mathbb{N}^*$, the penalty function satisfies $\mathrm{pen}(m)=\lambda m$ with

$$\lambda\ge\frac{\kappa}{a_\sigma\wedge a_\pi}\left(1+\frac{\left(A_\mu^2+A_\sigma^2\ln(n)\right)}{a_\sigma^2}\right)\frac{\sqrt{K}}{\sqrt{n}}\left(1+\|x\|_{\max,n}\ln(n)\sqrt{K\ln(2p+1)}\right) \tag{4.3}$$

for some absolute constant $\kappa\ge 360$. Then, any penalized likelihood estimate $\widehat{s}_{\widehat{m}}$ with $\widehat{m}$ such that

$$-\frac{1}{n}\sum_{i=1}^n\ln(\widehat{s}_{\widehat{m}}(Y_i|x_i))+\mathrm{pen}(\widehat{m})\le\inf_{m\in\mathbb{N}^*}\left(-\frac{1}{n}\sum_{i=1}^n\ln(\widehat{s}_m(Y_i|x_i))+\mathrm{pen}(m)\right)+\eta \tag{4.4}$$

for some $\eta \geq 0$ satisfies

$$\mathbb{E}\left[\mathrm{KL}_{\mathrm{n}}(s_0, \widehat{s}_{\widehat{m}})\right] \leq (1+\kappa^{-1}) \inf_{m\in\mathbb{N}^*} \left(\inf_{s_m \in S_m} \mathrm{KL}_{\mathrm{n}}(s_0, s_m) + \mathrm{pen}(m) + \eta_m \right) + \frac{\kappa' \sqrt{K}}{\sqrt{n}} \left[K \frac{\left(1+(A_\mu + A_\sigma)^2\right)}{a_\sigma \wedge a_\pi} \left(1 + \frac{\left(A_\mu^2 + A_\sigma^2 \ln(n)\right)}{a_\sigma^2} \right) + a_\sigma \, e^{-\frac{1}{2}\left(1+\frac{a_\mu^2}{2A_\sigma^2}\right)} \right] + \eta, \tag{4.5}$$

where κ' is an absolute positive constant.

Theorem 4.1 can be deduced from the two following propositions.

Proposition 4.2. *Assume we observe $((x_i, Y_i))_{1\leq i\leq n}$ with unknown conditional density s_0. Let $M_n > 0$ and consider the event*

$$\mathcal{T} := \left\{ \max_{i=1,\dots,n} |Y_i| \leq M_n \right\}.$$

For all $m \in \mathbb{N}^*$, consider the ℓ_1-ball

$$S_m = \{ s_\psi \in S, |s_\psi|_1 \leq m \} \tag{4.6}$$

and let $\widehat{s}_m$ be a η_m-ln-likelihood minimizer in S_m for some $\eta_m \geq 0$:

$$-\frac{1}{n}\sum_{i=1}^n \ln(\widehat{s}_m(Y_i|x_i)) \leq \inf_{s_m\in S_m} \left(-\frac{1}{n}\sum_{i=1}^n \ln(s_m(Y_i|x_i)) \right) + \eta_m.$$

Assume that for all $m \in \mathbb{N}^*$, the penalty function satisfies $\mathrm{pen}(m) = \lambda m$ with

$$\lambda \geq \frac{\kappa}{a_\sigma \wedge a_\pi} \left(1 + \frac{(M_n + A_\mu)^2}{a_\sigma^2} \right) \frac{\sqrt{K}}{\sqrt{n}} \left(1 + \|x\|_{\max,n} \ln(n) \sqrt{K \ln(2p+1)} \right) \tag{4.7}$$

for some absolute constant $\kappa \geq 36$. Then, any penalized likelihood estimate $\widehat{s}_{\widehat{m}}$ with $\widehat{m}$ such that

$$-\frac{1}{n}\sum_{i=1}^n \ln(\widehat{s}_{\widehat{m}}(Y_i|x_i)) + \mathrm{pen}(\widehat{m}) \leq \inf_{m\in\mathbb{N}^*} \left(-\frac{1}{n}\sum_{i=1}^n \ln(\widehat{s}_m(Y_i|x_i)) + \mathrm{pen}(m) \right) + \eta \tag{4.8}$$

for some $\eta \geq 0$ satisfies

$$\mathbb{E}\left[\mathrm{KL}_{\mathrm{n}}(s_0, \widehat{s}_{\widehat{m}}) \mathbb{1}_{\mathcal{T}}\right] \leq (1+\kappa^{-1}) \inf_{m\in\mathbb{N}^*} \left(\inf_{s_m \in S_m} \mathrm{KL}_{\mathrm{n}}(s_0, s_m) + \mathrm{pen}(m) + \eta_m \right) + \kappa' K^{3/2} \frac{\left(1+(A_\mu + A_\sigma)^2\right)}{(a_\sigma \wedge a_\pi)\sqrt{n}} \left(1 + \frac{(M_n + A_\mu)^2}{a_\sigma^2} \right) + \eta, \tag{4.9}$$

where κ' is an absolute positive constant.

Proposition 4.3. *Consider s_0, $\mathcal{T}$ and $\widehat{s}_{\widehat{m}}$ defined in Proposition 4.2. Denote by $\mathcal{T}^C$ the complementary event of $\mathcal{T}$,*

$$\mathcal{T}^C = \left\{ \max_{i=1,\dots,n} |Y_i| > M_n \right\}.$$

Assume that the unknown conditional density s_0 is a mixture of Gaussian densities. Then,

$$\mathbb{E}\left[\mathrm{KL}_{\mathrm{n}}(s_0, \widehat{s}_{\widehat{m}}) \mathbb{1}_{\mathcal{T}^C}\right] \leq \sqrt{2\pi K}\, a_\sigma \, e^{-\frac{1}{2}\left(1+\frac{a_\mu^2}{2A_\sigma^2}\right)} \, e^{-\frac{M_n(M_n - 2A_\mu)}{4A_\sigma^2}} \sqrt{n}.$$

Theorem 4.1, Propositions 4.2 and 4.3 are proved below.

4.2. Proofs

The main result is Proposition 4.2. Its proof follows the arguments developed in the proof of a more general model selection theorem for maximum likelihood estimators (Thm. 7.11) in [11]. Nonetheless, these arguments are here lightened. In particular, in the Proof of Theorem 7.11 [11], in addition to the relative expected loss function, another way of measuring the closeness between the elements of the model is required. It is directly connected to the variance of the increments of the empirical process. The main tool used is Bousquet's version of Talagrand's inequality for empirical processes to concentrate the oscillations of the empirical process by the modulus of uniform continuity of the empirical process in expectation. Then, the main task is to compute this modulus of uniform continuity. To evaluate it, some margin conditions (such as the ones in [18]) are necessary. On the contrary, we do not need such conditions to prove Proposition 4.2 because we are just looking for low rates of convergence. Therefore, the Proof of Proposition 4.2 is rather in the spirit of Vapnik's method of structural risk minimization (initiated in [23], further developed in [24] and briefly summarized in Sect. 8.2 in [11]) that provides a less refined – yet sufficient for our study – analysis of the risk of an empirical risk minimizer than Theorem 7.11 [11]. To obtain an upper bound of the empirical process in expectation, we shall use concentration inequalities combined with symmetrization arguments.

4.2.1. Proof of Proposition 4.2

Let us first introduce some definitions and notations that we shall use throughout the proof.

For any measurable function $g : \mathbb{R} \mapsto \mathbb{R}$, consider its empirical norm

$$\|g\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n} g^2(Y_i|x_i)}, \tag{4.10}$$

its conditional expectation

$$\mathbb{E}_X\left[g\right] = \mathbb{E}\left[g(.|X)|X = x\right] = \int_{\mathbb{R}} g(y|x)s_0(y|x)\,\mathrm{d}y,$$

as well as its empirical process

$$P_n(g) := \frac{1}{n}\sum_{i=1}^{n} g(Y_i|x_i), \tag{4.11}$$

and the recentred process

$$\nu_n(g) := P_n(g) - \mathbb{E}_X\left[P_n\left(g\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\left[g(Y_i|x_i) - \int_{\mathbb{R}} g(y|x_i)s_0(y|x_i)\,\mathrm{d}y\right]. \tag{4.12}$$

For all $m \in \mathbb{N}^*$, for all model S_m, define

$$F_m = \left\{f_m = -\ln\left(\frac{s_m}{s_0}\right),\ s_m \in S_m\right\}. \tag{4.13}$$

Let $\delta_{\mathrm{KL}} > 0$. For all $m \in \mathbb{N}^*$, let $\eta_m \geq 0$. Then, there exist two functions $\widehat{s}_m$ and $\overline{s}_m$ in S_m such that

$$P_n(-\ln \widehat{s}_m) \leq \inf_{s_m \in S_m} P_n(-\ln s_m) + \eta_m \tag{4.14}$$

$$\mathrm{KL_n}(s_0, \overline{s}_m) \leq \inf_{s_m \in S_m} \mathrm{KL_n}(s_0, s_m) + \delta_{\mathrm{KL}}. \tag{4.15}$$

Put

$$\widehat{f}_m := -\ln\left(\frac{\widehat{s}_m}{s_0}\right), \quad \overline{f}_m := -\ln\left(\frac{\overline{s}_m}{s_0}\right). \tag{4.16}$$

Let $\eta \geq 0$ and fix $m \in \mathbb{N}^*$. Define

$$\mathcal{M}(m) = \{m' \in \mathbb{N}^* | P_n(-\ln \widehat{s}_{m'}) + \text{pen}(m') \leq P_n(-\ln \widehat{s}_m) + \text{pen}(m) + \eta\}. \tag{4.17}$$

For every $m' \in \mathcal{M}(m)$, we get from (4.17), (4.16) and (4.14) that

$$P_n\left(\widehat{f}_{m'}\right) + \text{pen}(m') \leq P_n\left(\widehat{f}_m\right) + \text{pen}(m) + \eta \leq P_n\left(\overline{f}_m\right) + \text{pen}(m) + \eta_m + \eta,$$

which implies by (4.12) that

$$\mathbb{E}_X\left[P_n\left(\widehat{f}_{m'}\right)\right] + \text{pen}(m') \leq \mathbb{E}_X\left[P_n\left(\overline{f}_m\right)\right] + \text{pen}(m) + \nu_n\left(\overline{f}_m\right) - \nu_n\left(\widehat{f}_{m'}\right) + \eta_m + \eta.$$

Taking into account (2.2), (4.11) and (4.15), we get

$$\text{KL}_n(s_0, \widehat{s}_{m'}) + \text{pen}(m') \leq \inf_{s_m \in S_m} \text{KL}_n(s_0, s_m) + \text{pen}(m) + \nu_n\left(\overline{f}_m\right) - \nu_n\left(\widehat{f}_{m'}\right) + \eta_m + \eta + \delta_{\text{KL}}. \tag{4.18}$$

Thus, all the matter is to control the deviation of $-\nu_n(\widehat{f}_{m'}) = \nu_n(-\widehat{f}_{m'})$. To cope with the randomness of $\widehat{f}_{m'}$, we shall control the deviation of $\sup_{f_{m'} \in F_{m'}} \nu_n\left(-f_{m'}\right)$. Such a control is provided by the following Lemma 4.4.

Lemma 4.4. *Let $M_n > 0$. Consider the event*

$$\mathcal{T} := \left\{\max_{i=1,\ldots,n} |Y_i| \leq M_n\right\}.$$

Put

$$B_n = \frac{1}{a_\sigma \wedge a_\pi}\left(1 + \frac{(M_n + A_\mu)^2}{a_\sigma^2}\right) \tag{4.19}$$

and

$$\Delta_{m'} := m' \|x\|_{\max,n} \ln n \sqrt{K \ln(2p+1)} + 6\left(1 + K(A_\mu + A_\sigma)\right). \tag{4.20}$$

Then, on the event $\mathcal{T}$, for all $m' \in \mathbb{N}^*$, for all $t > 0$, with $\mathbb{P}_X$-probability greater than $1 - \mathrm{e}^{-t}$,

$$\sup_{f_{m'} \in F_{m'}} |\nu_n\left(-f_{m'}\right)| \leq \frac{4B_n}{\sqrt{n}}\left[9\sqrt{K}\,\Delta_{m'} + \sqrt{2}(1 + K(A_\mu + A_\sigma))\sqrt{t}\right]. \tag{4.21}$$

Proof. (See Sect. 4.2.4) □

We derive from (4.18) and (4.21) that on the event $\mathcal{T}$, for all $m \in \mathbb{N}^*$, for all $m' \in \mathcal{M}(m)$, for all $t > 0$, with $\mathbb{P}_X$-probability larger than $1 - \mathrm{e}^{-t}$,

$$\begin{aligned}
\text{KL}_n(s_0, \widehat{s}_{m'}) + \text{pen}(m') &\leq \inf_{s_m \in S_m} \text{KL}_n(s_0, s_m) + \text{pen}(m) + \nu_n\left(\overline{f}_m\right) \\
&\quad + \frac{4B_n}{\sqrt{n}}\left[9\sqrt{K}\,\Delta_{m'} + \sqrt{2}(1 + K(A_\mu + A_\sigma))\sqrt{t}\right] + \eta_m + \eta + \delta_{\text{KL}} \\
&\leq \inf_{s_m \in S_m} \text{KL}_n(s_0, s_m) + \text{pen}(m) + \nu_n\left(\overline{f}_m\right) \\
&\quad + \frac{4B_n}{\sqrt{n}}\left(9\sqrt{K}\,\Delta_{m'} + \frac{1}{2\sqrt{K}}\left(1 + K\left(A_\mu + A_\sigma\right)\right)^2 + \sqrt{K}t\right) \\
&\quad + \eta_m + \eta + \delta_{\text{KL}},
\end{aligned} \tag{4.22}$$

where we get the last inequality by using $2ab \leq \theta a^2 + \theta^{-1} b^2$ for $\theta = 1/\sqrt{K}$.

It remains to sum up the tail bounds (4.22) over all the possible values of $m \in \mathbb{N}^*$ and $m' \in \mathcal{M}(m)$. To get an inequality valid on a great probability set, we need to choose adequately the value of the parameter t depending on $m \in \mathbb{N}^*$ and $m' \in \mathcal{M}(m)$. Let $z > 0$. For all $m \in \mathbb{N}^*$ and $m' \in \mathcal{M}(m)$, apply (4.22) to $t = z + m + m'$. Then, on the event $\mathcal{T}$, for all $m \in \mathbb{N}^*$, for all $m' \in \mathcal{M}(m)$, with $\mathbb{P}_X$-probability larger than $1 - \mathrm{e}^{-(z+m+m')}$,

$$\begin{aligned}\mathrm{KL_n}(s_0, \widehat{s}_{m'}) + \mathrm{pen}(m') \leq \inf_{s_m \in S_m} \mathrm{KL_n}(s_0, s_m) + \mathrm{pen}(m) + \nu_n\left(\overline{f}_m\right) \\ + \frac{4B_n}{\sqrt{n}}\left(9\sqrt{K}\,\Delta_{m'} + \frac{1}{2\sqrt{K}}\left(1 + K\left(A_\mu + A_\sigma\right)\right)^2 + \sqrt{K}(z + m + m')\right) \\ + \eta_m + \eta + \delta_{\mathrm{KL}},\end{aligned} \tag{4.23}$$

and on the event $\mathcal{T}$, with $\mathbb{P}_X$-probability larger than

$$1 - \sum_{(m,m') \in \mathbb{N}^* \times \mathcal{M}(m)} \mathrm{e}^{-(z+m+m')} \geq 1 - \sum_{(m,m') \in \mathbb{N}^* \times \mathbb{N}^*} \mathrm{e}^{-(z+m+m')} = 1 - \mathrm{e}^{-z}\left(\sum_{m \in \mathbb{N}^*} \mathrm{e}^{-m}\right)^2 \geq 1 - \mathrm{e}^{-z},$$

(4.23) holds simultaneously for all $m \in \mathbb{N}^*$ and $m' \in \mathcal{M}(m)$.
Inequality (4.23) can also be written

$$\begin{aligned}\mathrm{KL_n}(s_0, \widehat{s}_{m'}) - \nu_n\left(\overline{f}_m\right) \leq \inf_{s_m \in S_m} \mathrm{KL_n}(s_0, s_m) + \left[\mathrm{pen}(m) + \frac{4B_n}{\sqrt{n}}\sqrt{K}m\right] \\ + \left[\frac{4B_n}{\sqrt{n}}\sqrt{K}\left(9\Delta_{m'} + m'\right) - \mathrm{pen}(m')\right] \\ + \frac{4B_n}{\sqrt{n}}\left(\frac{1}{2\sqrt{K}}\left(1 + K\left(A_\mu + A_\sigma\right)\right)^2 + \sqrt{K}z\right) + \eta_m + \eta + \delta_{\mathrm{KL}}.\end{aligned}$$

Taking into account Definition (4.20) of $\Delta_{m'}$, it gives

$$\begin{aligned}\mathrm{KL_n}(s_0, \widehat{s}_{m'}) - \nu_n\left(\overline{f}_m\right) \leq \inf_{s_m \in S_m} \mathrm{KL_n}(s_0, s_m) + \left[\mathrm{pen}(m) + \frac{4B_n}{\sqrt{n}}\sqrt{K}m\right] \\ + \left[\frac{4B_n}{\sqrt{n}}\sqrt{K}\left(9\|x\|_{\max,n}\ln n\sqrt{K\ln(2p+1)} + 1\right)m' - \mathrm{pen}(m')\right] \\ + \frac{4B_n}{\sqrt{n}}\left(\frac{1}{2\sqrt{K}}\left(1 + K\left(A_\mu + A_\sigma\right)\right)^2 + 54\sqrt{K}\left(1 + K(A_\mu + A_\sigma)\right) + \sqrt{K}z\right) \\ + \eta_m + \eta + \delta_{\mathrm{KL}}.\end{aligned} \tag{4.24}$$

Now, let $\kappa \geq 1$ and assume that $\mathrm{pen}(m)$ satisfies $\mathrm{pen}(m) = \lambda m$ with

$$\lambda \geq 4\kappa\frac{B_n}{\sqrt{n}}\sqrt{K}\left(9\|x\|_{\max,n}\ln n\sqrt{K\ln(2p+1)} + 1\right).$$

Then, (4.24) implies

$$\begin{aligned}\mathrm{KL_n}(s_0, \widehat{s}_{m'}) - \nu_n\left(\overline{f}_m\right) \leq \inf_{s_m \in S_m} \mathrm{KL_n}(s_0, s_m) + \left(1 + \kappa^{-1}\right)\mathrm{pen}(m) \\ + \frac{4B_n}{\sqrt{n}}\left(\frac{1}{2\sqrt{K}}\left(1 + K\left(A_\mu + A_\sigma\right)\right)^2 + 54\sqrt{K}\left(1 + K(A_\mu + A_\sigma)\right) + \sqrt{K}z\right) \\ + \eta_m + \eta + \delta_{\mathrm{KL}}.\end{aligned}$$

Then, using the inequality $2ab \leq \beta a^2 + \beta^{-1} b^2$ for $\beta = \sqrt{K}$,

$$\begin{aligned}\mathrm{KL_n}(s_0, \widehat{s}_{m'}) - \nu_n\left(\overline{f}_m\right) \leq & \inf_{s_m \in S_m} \mathrm{KL_n}(s_0, s_m) + \left(1 + \kappa^{-1}\right) \mathrm{pen}(m) \\ & + \frac{4B_n}{\sqrt{n}} \left(27K^{3/2} + \frac{55}{2\sqrt{K}} \left(1 + K\left(A_\mu + A_\sigma\right)\right)^2 + \sqrt{K} z\right) \\ & + \eta_m + \eta + \delta_{\mathrm{KL}}. \end{aligned} \tag{4.25}$$

Now consider $\widehat{m}$ defined by (4.8). By Definitions (4.8) and (4.17), $\widehat{m}$ belongs to $\mathcal{M}(m)$ for all $m \in \mathbb{N}^*$, so we deduce from (4.25) that on the event $\mathcal{T}$, for all $z > 0$, with $\mathbb{P}_X$-probability greater than $1 - \mathrm{e}^{-z}$,

$$\begin{aligned}\mathrm{KL_n}(s_0, \widehat{s}_{\widehat{m}}) - \nu_n\left(\overline{f}_m\right) \leq & \inf_{m \in \mathbb{N}^*} \left(\inf_{s_m \in S_m} \mathrm{KL_n}(s_0, s_m) + \left(1 + \kappa^{-1}\right) \mathrm{pen}(m) + \eta_m \right) \\ & + \frac{4B_n}{\sqrt{n}} \left(27K^{3/2} + \frac{55}{2\sqrt{K}} \left(1 + K\left(A_\mu + A_\sigma\right)\right)^2 + \sqrt{K} z\right) + \eta + \delta_{\mathrm{KL}}. \end{aligned} \tag{4.26}$$

We end the proof by integrating (4.26) with respect to z. Noticing that $\mathbb{E}\left(\nu_n\left(\overline{f}_m\right)\right) = 0$ and that δ_{KL} can be chosen arbitrary small, we get

$$\begin{aligned}\mathbb{E}\left[\mathrm{KL_n}(s_0, \widehat{s}_{\widehat{m}})\, \mathbb{1}_{\mathcal{T}}\right] \leq & \inf_{m \in \mathbb{N}^*} \left(\inf_{s_m \in S_m} \mathrm{KL_n}(s_0, s_m) + \left(1 + \kappa^{-1}\right) \mathrm{pen}(m) + \eta_m \right) \\ & + \frac{4B_n}{\sqrt{n}} \left(27K^{3/2} + \frac{55}{2\sqrt{K}} \left(1 + K\left(A_\mu + A_\sigma\right)\right)^2 + \sqrt{K}\right) + \eta \\ \leq & \inf_{m \in \mathbb{N}^*} \left(\inf_{s_m \in S_m} \mathrm{KL_n}(s_0, s_m) + \left(1 + \kappa^{-1}\right) \mathrm{pen}(m) + \eta_m \right) \\ & + \frac{112B_n}{\sqrt{n}} K^{3/2} \left(1 + \left(A_\mu + A_\sigma\right)^2\right) + \eta, \end{aligned}$$

hence (4.9) taking into account the value (4.19) of B_n.

4.2.2. Proof of Proposition 4.3

By Cauchy–Schwarz Inequality,

$$\mathbb{E}\left[\mathrm{KL_n}(s_0, \widehat{s}_{\widehat{m}}) \mathbb{1}_{\mathcal{T}^C}\right] \leq \sqrt{\mathbb{E}\left[\mathrm{KL_n^2}(s_0, \widehat{s}_{\widehat{m}})\right]} \sqrt{\mathbb{P}\left(\mathcal{T}^C\right)}. \tag{4.27}$$

Let us bound the two terms of the right-hand side of (4.27).

For the first term, let us bound $\mathrm{KL}(s_0(.|x), s_\psi(.|x))$ for all $s_\psi \in S$ and $x \in \mathbb{R}^p$.

Let $s_\psi \in S$ and $x \in \mathbb{R}^p$. Since s_0 is a density, s_0 is bounded by 1 and thus

$$\begin{aligned}\mathrm{KL}(s_0(.|x), s_\psi(.|x)) &= \int_{\mathbb{R}} \ln\left(\frac{s_0(y|x)}{s_\psi(y|x)}\right) s_0(y|x)\, \mathrm{d}y \\ &= \int_{\mathbb{R}} \ln(s_0(y|x)) s_0(y|x)\, \mathrm{d}y - \int_{\mathbb{R}} \ln(s_\psi(y|x)) s_0(y|x)\, \mathrm{d}y \\ &\leq -\int_{\mathbb{R}} \ln(s_\psi(y|x)) s_0(y|x)\, \mathrm{d}y. \end{aligned} \tag{4.28}$$

Denote $\psi = (\mu_k^T, \sigma_k, \pi_k)_{k=1,\dots,K}$. The parameters ψ and ψ_0 are assumed to belong to the bounded space Ψ defined by (2.1), so for all $y \in \mathbb{R}$,

$$\begin{aligned}
&\ln(s_\psi(y|x))s_0(y|x)\\
&= \ln\left[\sum_{k=1}^K \frac{\pi_k}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2}\left(\frac{y-\mu_k^T x}{\sigma_k}\right)^2\right)\right]\sum_{k=1}^K \frac{\pi_{0,k}}{\sqrt{2\pi}\sigma_{0,k}}\exp\left(-\frac{1}{2}\left(\frac{y-\mu_{0,k}^T x}{\sigma_{0,k}}\right)^2\right)\\
&\geq \ln\left[\sum_{k=1}^K \frac{\pi_k}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{y^2+\left(\mu_k^T x\right)^2}{\sigma_k^2}\right)\right]\sum_{k=1}^K \frac{\pi_{0,k}}{\sqrt{2\pi}\sigma_{0,k}}\exp\left(-\frac{y^2+\left(\mu_{0,k}^T x\right)^2}{\sigma_{0,k}^2}\right)\\
&\geq \ln\left[K\frac{a_\pi}{\sqrt{2\pi}A_\sigma}\exp\left(-\frac{y^2+A_\mu^2}{a_\sigma^2}\right)\right]K\frac{a_\pi}{\sqrt{2\pi}A_\sigma}\exp\left(-\frac{y^2+A_\mu^2}{a_\sigma^2}\right)\\
&= \frac{Ka_\pi}{\sqrt{2\pi}A_\sigma}\exp\left(-\frac{A_\mu^2}{a_\sigma^2}\right)\left[\ln\left(\frac{Ka_\pi}{\sqrt{2\pi}A_\sigma}\right)-\frac{A_\mu^2+y^2}{a_\sigma^2}\right]\exp\left(-\frac{y^2}{a_\sigma^2}\right). \qquad (4.29)
\end{aligned}$$

Therefore, putting $u = \sqrt{2}y/a_\sigma$ and $h(t) = t\ln t$ for all $t \in \mathbb{R}$ and noticing that $h(t) \geq h(\mathrm{e}^{-1}) = -e^{-1}$ for all $t \in \mathbb{R}$, we get from (4.28) and (4.29) that

$$\begin{aligned}
\mathrm{KL}(s_0(.|x), s_\psi(.|x)) &\leq -\frac{Ka_\pi a_\sigma\,\mathrm{e}^{-(A_\mu/a_\sigma)^2}}{\sqrt{2}A_\sigma}\int_{\mathbb{R}}\left[\ln\left(\frac{Ka_\pi}{\sqrt{2\pi}A_\sigma}\right)-\frac{A_\mu^2}{a_\sigma^2}-\frac{u^2}{2}\right]\frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}}\,du\\
&\leq -\frac{Ka_\pi a_\sigma\,\mathrm{e}^{-(A_\mu/a_\sigma)^2}}{\sqrt{2}A_\sigma}\,\mathbb{E}\left[\ln\left(\frac{Ka_\pi}{\sqrt{2\pi}A_\sigma}\right)-\frac{A_\mu^2}{a_\sigma^2}-\frac{U^2}{2}\right] \quad \text{with } U\sim\mathcal{N}(0,1)\\
&\leq -\frac{Ka_\pi a_\sigma\,\mathrm{e}^{-(A_\mu/a_\sigma)^2}}{\sqrt{2}A_\sigma}\left[\ln\left(\frac{Ka_\pi}{\sqrt{2\pi}A_\sigma}\right)-\frac{A_\mu^2}{a_\sigma^2}-\frac{1}{2}\right]\\
&\leq -\sqrt{\pi}\,e^{1/2}a_\sigma\,h\left(\frac{Ka_\pi\,\mathrm{e}^{-\left[(A_\mu/a_\sigma)^2+1/2\right]}}{\sqrt{2\pi}A_\sigma}\right)\\
&\leq \sqrt{\pi}\,\mathrm{e}^{-1/2}a_\sigma. \qquad (4.30)
\end{aligned}$$

Then, for all $s_\psi \in S$,

$$\mathrm{KL}_\mathrm{n}(s_0, s_\psi) = \frac{1}{n}\sum_{i=1}^n \mathrm{KL}\left(s_0(.|x_i), s_\psi(.|x_i)\right) \leq \sqrt{\pi}\,\mathrm{e}^{-1/2}a_\sigma$$

and thus

$$\sqrt{\mathbb{E}\left[\mathrm{KL}_\mathrm{n}^2(s_0, \widehat{s}_{\widehat{m}})\right]} \leq \sqrt{\pi}\,\mathrm{e}^{-1/2}a_\sigma. \qquad (4.31)$$

Let us now provide an upper bound of $\mathbb{P}(\mathcal{T}^C)$.

$$\mathbb{P}\left(\mathcal{T}^C\right) = \mathbb{E}\left(\mathbb{1}_{\mathcal{T}^C}\right) = \mathbb{E}\left[\mathbb{E}_X\left(\mathbb{1}_{\mathcal{T}^C}\right)\right] = \mathbb{E}\left[\mathbb{P}_X\left(\mathcal{T}^C\right)\right] \qquad (4.32)$$

with

$$\mathbb{P}_X\left(\mathcal{T}^C\right) = \mathbb{P}_X\left(\bigcup_{i=1}^n\{|Y_i| > M_n\}\right) \leq \sum_{i=1}^n \mathbb{P}_X(|Y_i| > M_n). \qquad (4.33)$$

For all $i = 1,\dots,n$, $Y_i|x_i \sim \sum_{k=1}^K \pi_k\mathcal{N}\left(\mu_k^T x_i, \sigma_k^2\right)$, so we see from (4.33) that we just need to provide an upper bound of $\mathbb{P}\left(|Y_x| > M_n\right)$ with $Y_x \sim \sum_{k=1}^K \pi_k\mathcal{N}\left(\mu_k^T x, \sigma_k^2\right)$ for $x \in \mathbb{R}^p$. First using Chernoff's inequality for a

centered Gaussian variable (see [11]), and then the fact that ψ belongs to the bounded space Ψ defined by (2.1) and that $\sum_{k=1}^{K} \pi_k = 1$, we get

$$\begin{aligned}
\mathbb{P}\left(|Y_x| > M_n\right) &= \int_{\mathbb{R}} \mathbb{1}_{\{|y|>M_n\}} \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \mathrm{e}^{-\frac{1}{2}\left(\frac{y-\mu_k^T x}{\sigma_k}\right)^2} \mathrm{d}y \\
&= \sum_{k=1}^{K} \pi_k \int_{\mathbb{R}} \mathbb{1}_{\{|y|>M_n\}} \frac{1}{\sqrt{2\pi}\sigma_k} \mathrm{e}^{-\frac{1}{2}\left(\frac{y-\mu_k^T x}{\sigma_k}\right)^2} \mathrm{d}y \qquad (4.34) \\
&= \sum_{k=1}^{K} \pi_k \, \mathbb{P}\left(|Y_{x,k}| > M_n\right) \quad \text{with } Y_{x,k} \sim \mathcal{N}\left(\mu_k^T x, \sigma_k^2\right) \qquad (4.35) \\
&= \sum_{k=1}^{K} \pi_k \left[\mathbb{P}\left(U > \frac{M_n - \mu_k^T x}{\sigma_k}\right) + \mathbb{P}\left(U > \frac{M_n + \mu_k^T x}{\sigma_k}\right)\right] \quad \text{with } U \sim \mathcal{N}(0,1) \\
&\leq \sum_{k=1}^{K} \pi_k \left[\mathrm{e}^{-\frac{1}{2}\left(\frac{M_n - \mu_k^T x}{\sigma_k}\right)^2} + \mathrm{e}^{-\frac{1}{2}\left(\frac{M_n + \mu_k^T x}{\sigma_k}\right)^2}\right] \\
&\leq 2\sum_{k=1}^{K} \pi_k \, \mathrm{e}^{-\frac{1}{2}\left(\frac{M_n - \left|\mu_k^T x\right|}{\sigma_k}\right)^2} \\
&\leq 2\sum_{k=1}^{K} \pi_k \, \mathrm{e}^{-\frac{M_n^2 + \left(\mu_k^T x\right)^2 - 2M_n\left|\mu_k^T x\right|}{2\sigma_k^2}} \\
&\leq 2K\mathrm{e}^{-\frac{M_n^2 + a_\mu^2 - 2M_n A_\mu}{2A_\sigma^2}}. \qquad (4.36)
\end{aligned}$$

We derive from (4.32), (4.33) and (4.36) that

$$\mathbb{P}(\mathcal{T}^C) \leq 2K\mathrm{e}^{-\frac{M_n^2 + a_\mu^2 - 2M_n A_\mu}{2A_\sigma^2}} n, \qquad (4.37)$$

and we finally get from (4.27), (4.31) and (4.37) that

$$\mathbb{E}\left[\mathrm{KL_n}(s_0, \widehat{s}_{\widehat{m}})\mathbb{1}_{\mathcal{T}^C}\right] \leq \sqrt{2\pi K}\, a_\sigma \, \mathrm{e}^{-\frac{1}{2}\left(1+\frac{a_\mu^2}{2A_\sigma^2}\right)} \mathrm{e}^{-\frac{M_n(M_n - 2A_\mu)}{4A_\sigma^2}} \sqrt{n}. \qquad (4.38)$$

4.2.3. Proof of Theorem 4.1

Let $M_n > 0$ and $\kappa \geq 36$. Assume that, for all $m \in \mathbb{N}^*$, the penalty function satisfies $\mathrm{pen}(m) = \lambda m$ with

$$\lambda \geq \frac{\kappa}{a_\sigma \wedge a_\pi} \left(1 + \frac{(M_n + A_\mu)^2}{a_\sigma^2}\right) \frac{\sqrt{K}}{\sqrt{n}} \left(1 + \|x\|_{\max,n} \ln(n) \sqrt{K \ln(2p+1)}\right). \qquad (4.39)$$

We derive from Propositions 4.2 and 4.3 that there exists an absolute positive constant κ' such that any penalized likelihood estimate $\widehat{s}_{\widehat{m}}$ with $\widehat{m}$ such that

$$-\frac{1}{n}\sum_{i=1}^{n} \ln(\widehat{s}_{\widehat{m}}(Y_i|x_i)) + \mathrm{pen}(\widehat{m}) \leq \inf_{m \in \mathbb{N}^*} \left(-\frac{1}{n}\sum_{i=1}^{n} \ln(\widehat{s}_m(Y_i|x_i)) + \mathrm{pen}(m)\right) + \eta$$

satisfies

$$\begin{aligned}\mathbb{E}\left[\mathrm{KL}_{\mathrm{n}}(s_0,\widehat{s}_{\widehat{m}})\right] &= \mathbb{E}\left[\mathrm{KL}_{\mathrm{n}}(s_0,\widehat{s}_{\widehat{m}})\mathbb{1}_{\mathcal{T}}\right] + \mathbb{E}\left[\mathrm{KL}_{\mathrm{n}}(s_0,\widehat{s}_{\widehat{m}})\mathbb{1}_{\mathcal{T}^C}\right] \\ &\leq (1+\kappa^{-1}) \inf_{m\in\mathbb{N}^*}\left(\inf_{s_m\in S_m} \mathrm{KL}_{\mathrm{n}}(s_0,s_m) + \mathrm{pen}(m) + \eta_m\right) \\ &\quad + \kappa' K^{3/2}\frac{\left(1+(A_\mu+A_\sigma)^2\right)}{\sqrt{n}\,(a_\sigma\wedge a_\pi)}\left(1+\frac{(M_n+A_\mu)^2}{a_\sigma^2}\right)+\eta \\ &\quad + \sqrt{2\pi K}\,a_\sigma\,\mathrm{e}^{-\frac{1}{2}\left(1+\frac{a_\mu^2}{2A_\sigma^2}\right)}\,\mathrm{e}^{-\frac{M_n(M_n-2A_\mu)}{4A_\sigma^2}}\,\sqrt{n}. \end{aligned} \tag{4.40}$$

To get inequality (4.5), it only remains to optimize Inequality (4.40) with respect to M_n. Since the two terms depending on M_n in (4.40) have opposite monotony with respect to M_n, we are looking for a value of M_n such that these two terms are of the same order with respect to n. Consider $M_n = A_\mu + \sqrt{A_\mu^2 + 4A_\sigma^2\ln n}$ the positive solution of the equation $X(X-2A_\mu) - 4A_\sigma^2\ln n = 0$. Then, on the one hand,

$$\mathrm{e}^{-\frac{M_n(M_n-2A_\mu)}{4A_\sigma^2}}\sqrt{n} = \mathrm{e}^{-\ln n}\sqrt{n} = \frac{1}{\sqrt{n}}.$$

On the other hand, using the inequality $(a+b)^2 \leq 2(a^2+b^2)$, we have

$$\frac{1}{\sqrt{n}}\left(1+\frac{(M_n+A_\mu)^2}{a_\sigma^2}\right) = \frac{1}{\sqrt{n}}\left(1+\frac{\left(2A_\mu+\sqrt{A_\mu^2+4A_\sigma^2\ln n}\right)^2}{a_\sigma^2}\right) \leq \frac{1}{\sqrt{n}}\left(1+\frac{2\left(5A_\mu^2+4A_\sigma^2\ln n\right)}{a_\sigma^2}\right),$$

hence (4.5).

The upper bound (4.3) of the tuning parameter λ is obtained from the upper bound (4.39) and the fact that $(M_n+A_\mu)^2 \leq 2(5A_\mu^2+4A_\sigma^2\ln n)$ for $M_n = A_\mu + \sqrt{A_\mu^2+4A_\sigma^2\ln n}$.

4.2.4. Proof of Theorem 3.1

Let $\lambda > 0$. Define $\widehat{m}$ as the smallest integer such that $\widehat{s}(\lambda)$ belongs to $S_{\widehat{m}}$, *i.e.* $\widehat{m} = \lceil |\widehat{s}(\lambda)|_1 \rceil$. Then, using the definition of $\widehat{m}$, the definition (2.3) of $\widehat{s}(\lambda)$ and (4.1), we get

$$\begin{aligned} -\frac{1}{n}\sum_{i=1}^n \ln\left(\widehat{s}(\lambda)(Y_i|x_i)\right) + \lambda\widehat{m} &\leq -\frac{1}{n}\sum_{i=1}^n \ln\left(\widehat{s}(\lambda)(Y_i|x_i)\right) + \lambda|\widehat{s}(\lambda)|_1 + \lambda \\ &= \inf_{s_\psi\in\Psi}\left(-\frac{1}{n}\sum_{i=1}^n \ln\left(s_\psi(Y_i|x_i)\right) + \lambda|s_\psi|_1\right) + \lambda \\ &= \inf_{m\in\mathbb{N}^*}\inf_{s_\psi\in\Psi,|\psi|_1\leq m}\left(-\frac{1}{n}\sum_{i=1}^n \ln\left(s_\psi(Y_i|x_i)\right) + \lambda|s_\psi|_1\right) + \lambda \\ &\leq \inf_{m\in\mathbb{N}^*}\left(\inf_{s_m\in S_m}\left(-\frac{1}{n}\sum_{i=1}^n \ln\left(s_m(Y_i|x_i)\right)\right) + \lambda m\right) + \lambda, \end{aligned}$$

which implies

$$-\frac{1}{n}\sum_{i=1}^n \ln\left(\widehat{s}(\lambda)(Y_i|x_i)\right) + \mathrm{pen}(\widehat{m}) \leq \inf_{m\in\mathbb{N}^*}\left(-\frac{1}{n}\sum_{i=1}^n \ln\left(\widehat{s}_m(Y_i|x_i)\right) + \mathrm{pen}(m)\right) + \eta,$$

with $\mathrm{pen}(m) = \lambda m$, $\eta = \lambda$ and $\widehat{s}_m$ defined by (4.2) with $\eta_m = 0$. Thus, $\widehat{s}(\lambda)$ satisfies (4.4) and Theorem 3.1 follows from Theorem 4.1.

5. Proofs of the lemmas

5.1. Proof of Lemma 4.4

Let $m \in \mathbb{N}^*$. From (4.12), we have

$$\sup_{f_m \in F_m} |\nu_n(-f_m)| = \sup_{f_m \in F_m} \left| \frac{1}{n} \sum_{i=1}^{n} (f_m(Y_i|x_i) - \mathbb{E}_X[f_m(Y_i|x_i)]) \right|. \tag{5.1}$$

To control the deviation of such a quantity, we shall combine concentration with symmetrization arguments. We shall first use the following concentration inequality which can be found in [4].

Lemma 5.1 (see [4]). *Let $Z_1, \dots, Z_n$ be independent random variables with values in some space $\mathcal{Z}$ and let Γ be a class of real-valued functions on $\mathcal{Z}$. Assume that there exists R_n a non-random constant such that $\sup_{\gamma \in \Gamma} \|\gamma\|_n \leq R_n$. Then, for all $t > 0$,*

$$\mathbb{P}\left(\sup_{\gamma \in \Gamma} \left| \frac{1}{n} \sum_{i=1}^{n} \gamma(Z_i) - \mathbb{E}(\gamma(Z_i)) \right| > \mathbb{E}\left[\sup_{\gamma \in \Gamma} \left| \frac{1}{n} \sum_{i=1}^{n} \gamma(Z_i) - \mathbb{E}(\gamma(Z_i)) \right| \right] + 2\sqrt{2} R_n \sqrt{\frac{t}{n}} \right) \leq e^{-t}. \tag{5.2}$$

Then, we propose to bound $\mathbb{E}\left[\sup_{\gamma \in \Gamma} \left| \frac{1}{n} \sum_{i=1}^{n} \gamma(Z_i) - \mathbb{E}(\gamma(Z_i)) \right| \right]$ thanks to the following symmetrization argument. The proof of this result can be found in [22].

Lemma 5.2 (see Lem. 2.3.6 in [22]). *Let $Z_1, \dots, Z_n$ be independent random variables with values in some space $\mathcal{Z}$ and let Γ be a class of real-valued functions on $\mathcal{Z}$. Let $(\varepsilon_1, \dots, \varepsilon_n)$ be a Rademacher sequence independent of $(Z_1, \dots, Z_n)$. Then,*

$$\mathbb{E}\left[\sup_{\gamma \in \Gamma} \left| \frac{1}{n} \sum_{i=1}^{n} \gamma(Z_i) - \mathbb{E}(\gamma(Z_i)) \right| \right] \leq 2\mathbb{E}\left[\sup_{\gamma \in \Gamma} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \gamma(Z_i) \right| \right]. \tag{5.3}$$

From (5.3), the problem boils down to providing an upper bound of $\mathbb{E}\left[\sup_{\gamma \in \Gamma} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \gamma(Z_i) \right| \right]$. To do so, we shall apply the following lemma which is adapted from Lemma 6.1 in [11].

Lemma 5.3 (see Lem. 6.1 in [11]). *Let $Z_1, \dots, Z_n$ be independent random variables with values in some space $\mathcal{Z}$ and let Γ be a class of real-valued functions on $\mathcal{Z}$. Let $(\varepsilon_1, \dots, \varepsilon_n)$ be a Rademacher sequence independent of $(Z_1, \dots, Z_n)$. Define R_n a non-random constant such that*

$$\sup_{\gamma \in \Gamma} \|\gamma\|_n \leq R_n. \tag{5.4}$$

Then, for all $S \in \mathbb{N}^*$,

$$\mathbb{E}\left[\sup_{\gamma \in \Gamma} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \gamma(Z_i) \right| \right] \leq R_n \left(\frac{6}{\sqrt{n}} \sum_{s=1}^{S} 2^{-s} \sqrt{\ln\left[1 + N\left(2^{-s} R_n, \Gamma, \|.\|_n\right)\right]} + 2^{-S} \right), \tag{5.5}$$

where $N(\delta, \Gamma, \|.\|_n)$ stands for the δ-packing number of the set of functions Γ equipped with the metric induced by the norm $\|.\|_n$.

From (5.1), we propose to apply a conditional version of Lemma 5.1, Lemma 5.2 and Lemma 5.3 to $\Gamma = F_m$, $(Z_1, \dots, Z_n) = (Y_1|x_1, \dots, Y_n|x_n)$ and $\gamma(Z_i) = f_m(Y_i|x_i)$ so as to control $\sup_{f_m \in F_m} |\nu_n(-f_m)|$. On the one hand, we see from (5.4) that we need an upper bound of $\sup_{f_m \in F_m} \|f_m\|_n$. On the other hand, we see from (5.5) that we need to bound the entropy of the set of functions F_m equipped with the metric induced by the norm $\|.\|_n$. Such bounds are provided by the two following lemmas.

Let $M_n > 0$. Consider the event

$$\mathcal{T} := \left\{ \max_{i=1,\dots,n} |Y_i| \leq M_n \right\}$$

and put

$$B_n = \frac{1}{a_\sigma \wedge a_\pi} \left(1 + \frac{(M_n + A_\mu)^2}{a_\sigma^2} \right).$$

Lemma 5.4. *On the event $\mathcal{T}$, for all $m \in \mathbb{N}^*$,*

$$\sup_{f_m \in F_m} \|f_m\|_n \leq R_n := 2B_n \left(1 + K(A_\mu + A_\sigma)\right). \tag{5.6}$$

Proof. (See Sect. 5.2) □

Lemma 5.5. *Let $\delta > 0$ and $m \in \mathbb{N}^*$. On the event $\mathcal{T}$, we have the following upper bound of the δ-packing number of the set of functions F_m equipped with the metric induced by the norm $\|.\|_n$:*

$$N\left(\delta, F_m, \|.\|_n\right) \leq (2p+1)^{\frac{4B_n^2 K^2 m^2 \|x\|^2_{\max,n}}{\delta^2}} \left(1 + \frac{8B_n K A_\sigma}{\delta}\right)^K \left(1 + \frac{8B_n}{\delta}\right)^K.$$

Proof. (See Sect. 5.2) □

By using the upper bounds provided in Lemmas 5.4 and 5.5, we can apply Lemma 5.3 to get an upper bound of $\mathbb{E}_X\left[\sup_{f_m \in F_m} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_m(Y_i|x_i)\right|\right]$. It gives the following result.

Lemma 5.6. *Let $m \in \mathbb{N}^*$. Consider $(\varepsilon_1, \dots, \varepsilon_n)$ a Rademacher sequence independent of $(Y_1, \dots, Y_n)$. Then, on the event $\mathcal{T}$,*

$$\mathbb{E}_X\left[\sup_{f_m \in F_m} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_m(Y_i|x_i)\right|\right] \leq 18\sqrt{K}\frac{B_n}{\sqrt{n}}\,\Delta_m, \tag{5.7}$$

where

$$\Delta_m := m\|x\|_{\max,n} \ln n \sqrt{K \ln(2p+1)} + 6\left(1 + K(A_\mu + A_\sigma)\right).$$

Proof. (See proof of Lem. 5.9) □

Now, by using (5.7) and applying both Lemma 5.1 and Lemma 5.2 to $\Gamma = F_m$, $(Z_1, \dots, Z_n) = (Y_1|x_1, \dots, Y_n|x_n)$ and $\gamma(Z_i) = f_m(Y_i|x_i)$, we get that for all $m \in \mathbb{N}^*$, for all $t > 0$, with $\mathbb{P}_X$-probability greater than $1 - \mathrm{e}^{-t}$,

$$\begin{aligned}
\sup_{f_m \in F_m} |\nu_n(-f_m)| &= \sup_{f_m \in F_m} \left|\frac{1}{n}\sum_{i=1}^n \left(f_m(Y_i|x_i) - \mathbb{E}_X[f_m(Y_i|x_i)]\right)\right| \\
&\leq \mathbb{E}\left[\sup_{f_m \in F_m} \left|\frac{1}{n}\sum_{i=1}^n f_m(Y_i|x_i) - \mathbb{E}\left(f_m(Y_i|x_i)\right)\right|\right] + 2\sqrt{2}R_n\sqrt{\frac{t}{n}} \\
&\leq 2\mathbb{E}\left[\sup_{f_m \in F_m} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_m(Y_i|x_i)\right|\right] + 2\sqrt{2}R_n\sqrt{\frac{t}{n}} \\
&\leq \frac{4B_n}{\sqrt{n}}\left[9\sqrt{K}\Delta_m + \sqrt{2}(1 + K(A_\mu + A_\sigma))\sqrt{t}\right],
\end{aligned}$$

taking into account Definition (5.6) of R_n.

5.2. Proof of Lemmas 5.4–5.6

Proofs of both Lemmas 5.4 and 5.5 need an upper bound of the uniform norm of the gradient of $\ln s_\psi$ for all $s_\psi \in S$. Let us thus begin by providing such an upper bound.

Lemma 5.7. *For finite mixture regression models as described in Section 2.1,*

$$\sup_{x\in\mathbb{R}^p}\sup_{\psi\in\Psi}\left\|\frac{\partial \ln\left(s_\psi(.|x)\right)}{\partial\psi}\right\|_\infty \leq G(\cdot),$$

with

$$G:\mathbb{R}\mapsto\mathbb{R},\quad y\mapsto\frac{1}{a_\sigma\wedge a_\pi}\left(1+\frac{(|y|+A_\mu)^2}{a_\sigma^2}\right). \tag{5.8}$$

Proof. Let $s_\psi \in S$ with $\psi=(\mu_k^T,\sigma_k,\pi_k)_{k=1,\dots,K}$. For all $x\in\mathbb{R}^p$ and $y\in\mathbb{R}$,

$$\ln\left(s_\psi(y|x)\right)=\ln\left(\sum_{k=1}^K\frac{\pi_k}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2}\frac{\left(y-\mu_k^Tx\right)^2}{\sigma_k^2}\right)\right)=\ln\left(\sum_{k=1}^K f_k(x,y)\right)$$

where we put

$$f_k(x,y)=\frac{\pi_k}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2}\frac{\left(y-\mu_k^Tx\right)^2}{\sigma_k^2}\right)\geq 0.$$

For all $l=1,\dots,K$, by using the fact that $f_l(x,y)/(\sum_{k=1}^K f_k(x,y))\leq 1$ and the fact that ψ belongs to the bounded space Ψ defined by (2.1), we have

$$\begin{aligned}
\left|\frac{\partial \ln\left(s_\psi(y|x)\right)}{\partial(\mu_l^Tx)}\right| &= \left|\frac{f_l(x,y)}{\sum_{k=1}^K f_k(x,y)}\frac{y-\mu_l^Tx}{\sigma_l^2}\right| \leq \frac{|y|+A_\mu}{a_\sigma^2},\quad \left|\frac{\partial \ln\left(s_\psi(y|x)\right)}{\partial\sigma_l}\right| \\
&= \left|\frac{1}{\sum_{k=1}^K f_k(x,y)}\left(-\frac{f_l(x,y)}{\sigma_l}+f_l(x,y)\frac{\left(y-\mu_l^Tx\right)^2}{\sigma_l^3}\right)\right| \\
&\leq \frac{1}{a_\sigma}\left(1+\frac{(|y|+A_\mu)^2}{a_\sigma^2}\right), \\
\left|\frac{\partial \ln\left(s_\psi(y|x)\right)}{\partial\pi_l}\right| &= \frac{f_l(x,y)}{\pi_l\sum_{k=1}^K f_k(x,y)}\leq\frac{1}{a_\pi}.
\end{aligned}$$

Thus, for all $y\in\mathbb{R}$,

$$\begin{aligned}
\sup_{x\in\mathbb{R}^p}\sup_{\psi\in\Psi}\left|\frac{\partial \ln\left(s_\psi(y|x)\right)}{\partial\psi}\right| &\leq \max\left(\frac{|y|+A_\mu}{a_\sigma^2},\frac{1}{a_\sigma}\left(1+\frac{(|y|+A_\mu)^2}{a_\sigma^2}\right),\frac{1}{a_\pi}\right) \\
&\leq \max\left(\frac{1}{a_\sigma}\left(1+\frac{(|y|+A_\mu)^2}{a_\sigma^2}\right),\frac{1}{a_\pi}\right) \\
&\leq \frac{1}{a_\pi\wedge a_\sigma}\left(1+\frac{(|y|+A_\mu)^2}{a_\sigma^2}\right),
\end{aligned}$$

where we used the fact that $1+\theta^2\geq\theta$ for all $\theta\in\mathbb{R}$. □

5.2.1. Proof of Lemma 5.4

Let $m \in \mathbb{N}^*$ and $f_m \in F_m$. By (4.13), there exists $s_m \in S_m$ such that $f_m = -\ln(s_m/s_0)$. For all $x \in \mathbb{R}^p$, denote by $\psi(x) = (\mu_k^T x, \sigma_k, \pi_k)_{k=1,\dots,K}$ the parameters of the density $s_m(\cdot|x)$. First applying Taylor's inequality and then Lemma 5.7 on the event $\mathcal{T} = \{\max_{i=1,\dots,n} |Y_i| \leq M_n\}$, we get for all $i = 1, \dots, n$,

$$
\begin{aligned}
|f_m(Y_i|x_i)|\,\mathbb{1}_{\mathcal{T}} &= |\ln(s_m(Y_i|x_i)) - \ln(s_0(Y_i|x_i))|\,\mathbb{1}_{\mathcal{T}} \\
&\leq \sup_{x\in\mathbb{R}^p}\sup_{\varphi\in\Psi}\left|\frac{\partial \ln(s_\varphi(Y_i|x))}{\partial\varphi}\right| \|\psi(x_i) - \psi_0(x_i)\|_1\,\mathbb{1}_{\mathcal{T}} \\
&\leq \frac{1}{a_\sigma \wedge a_\pi}\left(1 + \frac{(|Y_i| + A_\mu)^2}{a_\sigma^2}\right)\|\psi(x_i) - \psi_0(x_i)\|_1\,\mathbb{1}_{\mathcal{T}} \\
&\leq \underbrace{\frac{1}{a_\sigma \wedge a_\pi}\left(1 + \frac{(M_n + A_\mu)^2}{a_\sigma^2}\right)}_{:=B_n}\|\psi(x_i) - \psi_0(x_i)\|_1 \\
&\leq B_n \sum_{k=1}^{K}\left(\left|\mu_k^T x_i - \mu_{0,k}^T x_i\right| + |\sigma_k - \sigma_{0,k}| + |\pi_k - \pi_{0,k}|\right).
\end{aligned}
$$

Now, since s_m and s_0 are assumed to belong to the bounded space Ψ defined by (2.1) and $\sum_{k=1}^K \pi_k = 1$, we obtain

$$
|f_m(Y_i|x_i)|\,\mathbb{1}_{\mathcal{T}} \leq B_n\left(2KA_\mu + 2KA_\sigma + 2\right) \leq 2B_n\left(1 + K(A_\mu + A_\sigma)\right),
$$

and thus $\|f_m\|_n\,\mathbb{1}_{\mathcal{T}} \leq 2B_n\left(1 + K(A_\mu + A_\sigma)\right)$.

5.2.2. Proof of Lemma 5.5

Let $m \in \mathbb{N}^*$ and $f_m \in F_m$. By (4.13), there exists $s_m \in S_m$ such that $f_m = -\ln(s_m/s_0)$. Introduce s'_m in S and put $f'_m = -\ln(s'_m/s_0)$. Denote by $(\mu_k^T, \sigma_k, \pi_k)_{k=1,\dots,K}$ and $({\mu'}_k^T, \sigma'_k, \pi'_k)_{k=1,\dots,K}$ the parameters of the densities s_m and s'_m respectively. First applying Taylor's inequality and then Lemma 5.7 on the event $\mathcal{T} = \{\max_{i=1,\dots,n} |Y_i| \leq M_n\}$, we get for all $i = 1, \dots, n$,

$$
\begin{aligned}
|f_m(Y_i|x_i) - f'_m(Y_i|x_i)|\,\mathbb{1}_{\mathcal{T}} &= |\ln(s_m(Y_i|x_i)) - \ln(s'_m(Y_i|x_i))|\,\mathbb{1}_{\mathcal{T}} \\
&\leq \sup_{x\in\mathbb{R}^p}\sup_{\varphi\in\Psi}\left|\frac{\partial \ln(s_\varphi(Y_i|x))}{\partial\varphi}\right| \|\psi(x_i) - \psi'(x_i)\|_1\,\mathbb{1}_{\mathcal{T}} \\
&\leq \frac{1}{a_\sigma \wedge a_\pi}\left(1 + \frac{(|Y_i| + A_\mu)^2}{a_\sigma^2}\right)\|\psi(x_i) - \psi'(x_i)\|_1\,\mathbb{1}_{\mathcal{T}} \\
&\leq \underbrace{\frac{1}{a_\sigma \wedge a_\pi}\left(1 + \frac{(M_n + A_\mu)^2}{a_\sigma^2}\right)}_{:=B_n}\|\psi(x_i) - \psi'(x_i)\|_1 \\
&\leq B_n \sum_{k=1}^{K}\left(\left|\mu_k^T x_i - {\mu'}_k^T x_i\right| + |\sigma_k - \sigma'_k| + |\pi_k - \pi'_k|\right) \\
&\leq B_n\left(\sum_{k=1}^{K}\left|\mu_k^T x_i - {\mu'}_k^T x_i\right| + \|\sigma - \sigma'\|_1 + \|\pi - \pi'\|_1\right).
\end{aligned}
$$

Then, using $(a+b)^2 \leq 2(a^2+b^2)$ and applying Cauchy–Schwarz inequality leads to

$$\begin{aligned}\left(f_m(Y_i|x_i) - f'_m(Y_i|x_i)\right)^2 \mathbb{1}_{\mathcal{T}} &\leq 2B_n^2 \left[\left(\sum_{k=1}^{K} \left|\mu_k^T x_i - {\mu'_k}^T x_i\right|\right)^2 + \left(\|\sigma-\sigma'\|_1 + \|\pi-\pi'\|_1\right)^2\right] \quad (5.9)\\ &\leq 2B_n^2 \left[K \sum_{k=1}^{K} \left(\mu_k^T x_i - {\mu'_k}^T x_i\right)^2 + \left(\|\sigma-\sigma'\|_1 + \|\pi-\pi'\|_1\right)^2\right] \quad (5.10)\\ &\leq 2B_n^2 \left[K \sum_{k=1}^{K} \left(\sum_{j=1}^{p} \mu_{kj} x_{ij} - \sum_{j=1}^{p} \mu'_{kj} x_{ij}\right)^2 + \left(\|\sigma-\sigma'\|_1 + \|\pi-\pi'\|_1\right)^2\right]\end{aligned}$$

and

$$\|f_m - f'_m\|_n^2 \mathbb{1}_{\mathcal{T}} \leq 2B_n^2 \left[\underbrace{K \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \mu_{kj} x_{ij} - \sum_{j=1}^{p} \mu'_{kj} x_{ij}\right)^2}_{(a)} + \left(\|\sigma-\sigma'\|_1 + \|\pi-\pi'\|_1\right)^2\right].$$

So, for all $\delta > 0$, if $(a) \leq \delta^2/(4B_n^2)$, $\|\sigma-\sigma'\|_1 \leq \delta/(4B_n)$ and $\|\pi-\pi'\|_1 \leq \delta/(4B_n)$, then $\|f_m - f'_m\|_n^2 \leq \delta^2$. To bound (a), we write

$$(a) = Km^2 \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \frac{\mu_{kj}}{m} x_{ij} - \sum_{j=1}^{p} \frac{\mu'_{kj}}{m} x_{ij}\right)^2 \quad (5.11)$$

and we apply Lemma 5.9 below to $\mu_k/m = (\mu_{kj}/m)_{j=1,\dots,p}$ for all $k = 1,\dots,K$. Since $s_m \in S_m$, we have

$$\sum_{j=1}^{p} \left|\frac{\mu_{kj}}{m}\right| \leq 1, \quad (5.12)$$

and thus there exists a family $\mathcal{B}$ of $(2p+1)^{4B_n^2K^2m^2\|x\|^2_{\max,n}/\delta^2}$ vectors of $\mathbb{R}^p$ such that for all $k = 1,\dots,K$, for all μ_k, there exists $\mu'_k \in \mathcal{B}$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \frac{\mu_{kj}}{m} x_{ij} - \frac{\mu'_{kj}}{m} x_{ij}\right)^2 \leq \frac{\delta^2}{4B_n^2K^2m^2},$$

so that $(a) \leq \delta^2/(4B_n^2)$. Moreover, since $\|\sigma\|_1 = \sum_{k=1}^{K} |\sigma_k| \leq KA_\sigma$ and $\|\pi\|_1 = \sum_{k=1}^{K} \pi_k = 1$, we get that, on the event $\mathcal{T}$,

$$\begin{aligned}N\left(\delta, F_m, \|.\|_n\right) &\leq \operatorname{card}(\mathcal{B})\ N\left(\frac{\delta}{4B_n}, B_1^K(KA_\sigma), \|.\|_1\right)\ N\left(\frac{\delta}{4B_n}, B_1^K(1), \|.\|_1\right)\\ &\leq (2p+1)^{\frac{4B_n^2K^2m^2\|x\|^2_{\max,n}}{\delta^2}} \left(1 + \frac{8B_nKA_\sigma}{\delta}\right)^K \left(1 + \frac{8B_n}{\delta}\right)^K. \quad (5.13)\end{aligned}$$

Remark 5.8. Let us point out that there may exist some data-sets for which the dependence on K might not be optimal in (5.13). Indeed, notice that S_m is defined as the set of densities whose ℓ_1-norm $\sum_{k=1}^{K}\sum_{j=1}^{p} |\mu_{kj}|$ is less than m, whereas in (5.12) we only use the fact that $\sum_{j=1}^{p} |\mu_{kj}| \leq m$ for all $k = 1,\dots,K$, that is to

say $\max_{k=1,\dots,K}\sum_{j=1}^{p}|\mu_{kj}|\leq m$, which is a weaker assumption than $\sum_{k=1}^{K}\sum_{j=1}^{p}|\mu_{kj}|\leq m$. To use the whole assumption $\sum_{k=1}^{K}\sum_{j=1}^{p}|\mu_{kj}|\leq m$, we should considered $\sum_{k=1}^{K}\sum_{j=1}^{p}(\mu_{kj}-\mu'_{kj})/m\,x_{ij}$ instead of $\sum_{j=1}^{p}(\mu_{kj}-\mu'_{kj})/m\,x_{ij}$ in (5.11). This could be possible if $\sum_{k=1}^{K}|\mu_k^T x_i-{\mu'_k}^T x_i|$ was replaced by $|\sum_{k=1}^{K}\mu_k^T x_i-{\mu'_k}^T x_i|$ in the right-hand side of (5.9). This would require to consider the single parameter $\sum_{k=1}^{K}\mu_k^T x$ in place of the K parameters $(\mu_1^T x,\dots,\mu_K^T x)$ in the parameter $\psi(x)$, but the quantity $\sum_{k=1}^{K}\mu_k^T x$ does not appear naturally in the expression of the density $s_\psi(.|x)$ and it seems thus difficult to differentiate $\ln s_\psi(.|x)$ with respect to it. Yet, if we had managed to do that, this would have avoided to use Cauchy–Schwarz inequality which leads to an extra-K factor in (5.10). So, we might think that the term $(2p+1)^{4B_n^2K^2m^2\|x\|^2_{\max,n}/\delta^2}$ in (5.13) could be improved by $(2p+1)^{4B_n^2Km^2\|x\|^2_{\max,n}/\delta^2}$. Then, taking the square root of the entropy number in (5.5), the term $m\|x\|_{\max,n}\ln n\sqrt{K\ln(2p+1)}$ in (4.20) would be replaced by $m\|x\|_{\max,n}\ln n\sqrt{\ln(2p+1)}$, and the lower bound of the regularization parameter λ in (4.7) would be proportional to $\sqrt{K}$ instead of K.

Lemma 5.9. *Let $\delta>0$ and $(x_{ij})_{i=1,\dots,n\ j=1,\dots,p}\in\mathbb{R}^{np}$. There exists a family $\mathcal{B}$ of $(2p+1)^{\|x\|^2_{\max,n}/\delta^2}$ vectors of $\mathbb{R}^p$ such that for all $\beta\in\mathbb{R}^p$ checking $\|\beta\|_1\leq 1$, there exists $\beta'\in\mathcal{B}$ such that*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}(\beta_j-\beta'_j)\,x_{ij}\right)^2\leq\delta^2. \tag{5.14}$$

Proof. Consider the set of functions $\mathcal{F}=\{f_0,f_1^+,\dots,f_p^+,f_1^-,\dots,f_p^-\}$ defined by

$$\begin{cases} f_0\equiv 0,\\ f_j^+:\mathbb{R}^p\mapsto\mathbb{R},\ x=(x_1,\dots,x_p)\mapsto x_j,\quad j=1,\dots,p,\\ f_j^-:\mathbb{R}^p\mapsto\mathbb{R},\ x=(x_1,\dots,x_p)\mapsto -x_j,\ j=1,\dots,p,\end{cases}$$

and the convex hull $\mathcal{C}_\mathcal{F}$ of $\mathcal{F}$. Let $\delta>0$. Applying Lemma 2.6.11 of [22] to $\mathcal{F}$ which is of cardinal $2p+1$, we deduce that there exists a packing family $\mathcal{G}\subset\mathcal{C}_\mathcal{F}$ of cardinal $(2p+1)^{(\mathrm{diam}\mathcal{F}/\delta)^2}$ for $(\mathcal{C}_\mathcal{F},\|.\|_n)$ where $\mathrm{diam}\mathcal{F}$ is the diameter of $\mathcal{F}$ for $\|.\|_n$. Here, $\mathrm{diam}\mathcal{F}=\|x\|_{\max,n}$.

Now, let $\beta\in\mathbb{R}^p$ such that $\|\beta\|_1\leq 1$ and introduce the function

$$f_\beta:\mathbb{R}^p\mapsto\mathbb{R},x\mapsto\sum_{j=1}^{p}\beta_j x_j. \tag{5.15}$$

For all $x\in\mathbb{R}^p$,

$$f_\beta(x)=\left(\sum_{j:\beta_j>0}|\beta_j|f_j^+(x)+\sum_{j:\beta_j<0}|\beta_j|f_j^-(x)+\left(1-\sum_{j:\beta_j\neq 0}|\beta_j|\right)f_0\right),$$

with $\sum_{j:\beta_j>0}|\beta_j|+\sum_{j:\beta_j<0}|\beta_j|+\left(1-\sum_{j:\beta_j\neq0}|\beta_j|\right)=1$, $\left(1-\sum_{j:\beta_j\neq0}|\beta_j|\right)\geq 0$ and $|\beta_j|\geq 0$ for all $j=1,\dots,p$. So, f_β belongs to $\mathcal{C}_\mathcal{F}$ and there exists f'_β in $\mathcal{G}$ such that $\|f_\beta-f'_\beta\|_n\leq\delta$, that is to say

$$\frac{1}{n}\sum_{i=1}^{n}\left(f_\beta(x_i)-f'_\beta(x_i)\right)^2\leq\delta^2. \tag{5.16}$$

Since f'_β belongs to $\mathcal{C}_\mathcal{F}$, there exist coefficients $\left(\alpha_0,\alpha_1^+,\dots,\alpha_p^+,\alpha_1^-,\dots,\alpha_p^-\right)$ such that $f'_\beta=\alpha_0 f_0+\sum_{j=1}^{p}\alpha_j^+f_j^++\alpha_j^-f_j^-$, and for all $x\in\mathbb{R}^p$,

$$f'_\beta(x)=\alpha_0 f_0(x)+\sum_{j=1}^{p}\alpha_j^+f_j^+(x)+\alpha_j^-f_j^-(x)=\sum_{j=1}^{p}(\alpha_j^+-\alpha_j^-)x_j=\sum_{j=1}^{p}\beta'_j x_j \tag{5.17}$$

if we put $\beta'_j := \alpha_j^+ - \alpha_j^-$ for all $j = 1, \dots, p$. For each function f'_β, we thus define a vector $\beta' \in \mathbb{R}^p$ associated to f'_β, which leads to the construction of a family $\mathcal{B}$ of $(2p+1)^{\|x\|^2_{\max,n}/\delta^2}$ vectors of $\mathbb{R}^p$. Inequality (5.14) is obtained from (5.16)–(5.17). □

5.2.3. Proof of Lemma 5.6

Let $m \in \mathbb{N}^*$. From Lemma 5.4, on the event $\mathcal{T}$, $\sup_{f_m \in F_m} \|f_m\|_n$ is bounded by

$$R_n := 2B_n \left(1 + K(A_\mu + A_\sigma)\right). \tag{5.18}$$

Besides, from Lemma 5.5, on the event $\mathcal{T}$, for all $S \in \mathbb{N}^*$,

$$\begin{aligned}\sum_{s=1}^{S} 2^{-s}\sqrt{\ln\left[1 + N\left(2^{-s}R_n, F_m, \|.\|_n\right)\right]} &\le \sum_{s=1}^{S} 2^{-s}\sqrt{\ln\left[2N\left(2^{-s}R_n, F_m, \|.\|_n\right)\right]} \\ &\le \sum_{s=1}^{S} 2^{-s}\left[\sqrt{\ln 2} + \frac{2^{s+1}B_nKm\|x\|_{\max,n}}{R_n}\sqrt{\ln(2p+1)}\right. \\ &\quad \left. + \sqrt{K\ln\left[\left(1 + \frac{2^{s+3}B_nKA_\sigma}{R_n}\right)\left(1 + \frac{2^{s+3}B_n}{R_n}\right)\right]}\right]. \end{aligned} \tag{5.19}$$

Notice from (5.18) that $R_n \ge 2B_n \max(KA_\sigma, 1)$. Moreover $1 \le 2^{s+2}$ and $2^{-s}\sqrt{s} \le (\sqrt{\mathrm{e}}/2)^s$ for all $s \in \mathbb{N}^*$. So, we get from (5.19) that

$$\begin{aligned}&\sum_{s=1}^{S} 2^{-s}\sqrt{\ln\left[2N\left(2^{-s}R_n, F_m, \|.\|_n\right)\right]} \\ &\le \sum_{s=1}^{S} 2^{-s}\left[\sqrt{\ln 2} + \frac{2^{s+1}B_nKm\|x\|_{\max,n}}{R_n}\sqrt{\ln(2p+1)} + \sqrt{K\ln\left[(2^{s+3})\times(2^{s+3})\right]}\right] \\ &\le \sum_{s=1}^{S} 2^{-s}\left[\sqrt{\ln 2} + \frac{2^{s+1}B_nKm\|x\|_{\max,n}}{R_n}\sqrt{\ln(2p+1)} + \sqrt{K}\sqrt{2(s+3)\ln 2}\right] \\ &\le \frac{2B_nKm\|x\|_{\max,n}}{R_n}\sqrt{\ln(2p+1)}S + \sqrt{K}\sqrt{2\ln 2}\sum_{s=1}^{S} 2^{-s}\sqrt{s} + \sqrt{\ln 2}(1+\sqrt{6K}) \\ &\le \frac{2B_nKm\|x\|_{\max,n}}{R_n}\sqrt{\ln(2p+1)}S + \sqrt{\ln 2}\left(1 + \sqrt{K}\left(\sqrt{6} + \frac{\sqrt{2\mathrm{e}}}{2-\sqrt{\mathrm{e}}}\right)\right) \\ &\le \frac{2B_nKm\|x\|_{\max,n}}{R_n}\sqrt{\ln(2p+1)}S + \sqrt{K\ln 2}\left(1 + \sqrt{6} + \frac{\sqrt{2\mathrm{e}}}{2-\sqrt{\mathrm{e}}}\right), \end{aligned} \tag{5.20}$$

and we get from (5.5) and (5.20) that for all $S \in \mathbb{N}^*$,

$$\begin{aligned}&\mathbb{E}_X\left[\sup_{f_m \in F_m}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f_m(Y_i|x_i)\right|\right] \\ &\le R_n\left[\frac{6}{\sqrt{n}}\left(\frac{2B_nKm\|x\|_{\max,n}}{R_n}\sqrt{\ln(2p+1)}S + \sqrt{K\ln 2}\left(1 + \sqrt{6} + \frac{\sqrt{2\mathrm{e}}}{2-\sqrt{\mathrm{e}}}\right)\right) + 2^{-S}\right]. \end{aligned} \tag{5.21}$$

Let us now choose $S = \ln n / \ln 2$ so that the two terms depending on S in (5.21) are of the same order. In particular, for this value of S, $2^{-S} \leq 1/n$, and we deduce from (5.21) and (5.18) that

$$\begin{aligned}
&\mathbb{E}_X \left[\sup_{f_m \in F_m} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_m(Y_i|x_i) \right| \right] \\
&\leq \frac{12 B_n K m \|x\|_{\max,n}}{\sqrt{n}} \sqrt{\ln(2p+1)} \frac{\ln n}{\ln 2} + 2B_n \left(1 + K(A_\mu + A_\sigma)\right) \left(6\sqrt{\ln 2} \left(1 + \sqrt{6} + \frac{\sqrt{2\mathrm{e}}}{2 - \sqrt{\mathrm{e}}} \right) \frac{\sqrt{K}}{\sqrt{n}} + \frac{1}{n} \right) \\
&\leq \frac{18 B_n K m \|x\|_{\max,n}}{\sqrt{n}} \sqrt{\ln(2p+1)}\, \ln n + 2\frac{\sqrt{K}}{\sqrt{n}} B_n \left(1 + K(A_\mu + A_\sigma)\right) \left(6\sqrt{\ln 2} \left(1 + \sqrt{6} + \frac{\sqrt{2\mathrm{e}}}{2 - \sqrt{\mathrm{e}}} \right) + 1 \right) \\
&\leq 18\sqrt{K} \frac{B_n}{\sqrt{n}} \left[m \|x\|_{\max,n} \sqrt{K \ln(2p+1)}\, \ln n + 6 \left(1 + K(A_\mu + A_\sigma)\right) \right].
\end{aligned}$$

Acknowledgements. We thank the associate editor and the referees for their interesting comments on the paper.

References

[1] P.L. Bartlett, S. Mendelson and J. Neeman, ℓ_1-regularized linear regression: persistence and oracle inequalities, Probability and related fields. Springer (2011).

[2] J.P. Baudry, Sélection de Modèle pour la Classification Non Supervisée. Choix du Nombre de Classes. Ph.D. thesis, Université Paris-Sud 11, France (2009).

[3] P.J. Bickel, Y. Ritov and A.B. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.* **37** (2009) 1705–1732.

[4] S. Boucheron, G. Lugosi and P. Massart, *A non Asymptotic Theory of Independence*. Oxford University press (2013).

[5] P. Bühlmann and S. van de Geer, On the conditions used to prove oracle results for the Lasso. *Electr. J. Stat.* **3** (2009) 1360–1392.

[6] E. Candes and T. Tao, The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Stat.* **35** (2007) 2313–2351.

[7] S. Cohen and E. Le Pennec, *Conditional Density Estimation by Penalized Likelihood Model Selection and Applications*, RR-7596. INRIA (2011).

[8] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, Least Angle Regression. *Ann. Stat.* **32** (2004) 407–499.

[9] M. Hebiri, Quelques questions de sélection de variables autour de l'estimateur Lasso. Ph.D. Thesis, Université Paris Diderot, Paris 7, France (2009).

[10] C. Huang, G.H.L. Cheang and A.R. Barron, Risk of penalized least squares, greedy selection and ℓ_1-penalization for flexible function librairies. Submitted to *the Annals of Statistics* (2008).

[11] P. Massart, *Concentration inequalities and model selection*. Ecole d'été de Probabilités de Saint-Flour 2003. *Lect. Notes Math.* Springer, Berlin-Heidelberg (2007).

[12] P. Massart and C. Meynet, The Lasso as an ℓ_1-ball model selection procedure. *Elect. J. Stat.* **5** (2011) 669–687.

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

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

[15] N. Meinshausen and B. Yu, Lasso type recovery of sparse representations for high dimensional data. *Ann. Stat.* **37** (2009) 246–270.

[16] R.A. Redner and H.F. Walker, Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** (1984) 195–239.

[17] P. Rigollet and A. Tsybakov, Exponential screening and optimal rates of sparse estimation. *Ann. Stat.* **39** (2011) 731–771.

[18] N. Städler, B.P. Hlmann and S. van de Geer, ℓ_1-penalization for mixture regression models. *Test* **19** (2010) 209–256.

[19] R. Tibshirani, Regression shrinkage and selection via the Lasso. *J. Roy. Stat. Soc.* Ser. B **58** (1996) 267–288.

[20] M.R. Osborne, B. Presnell and B.A. Turlach, On the Lasso and its dual. *J. Comput. Graph. Stat.* **9** (2000) 319–337.

[21] M.R. Osborne, B. Presnell and B.A Turlach, A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **20** (2000) 389–404.

[22] A. van der Vaart and J. Wellner, *Weak Convergence and Empirical Processes*. Springer, Berlin (1996).

[23] V.N. Vapnik, *Estimation of Dependencies Based on Empirical Data*. Springer, New-York (1982).

[24] V.N. Vapnik, *Statistical Learning Theory*. J. Wiley, New-York (1990).

[25] P. Zhao and B. Yu On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** (2006) 2541–2563.