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

EMILIE DEVIJVER

Abstract. We consider a multivariate finite mixture of Gaussian regression models for high-dimensional data, where the number of covariates and the size of the response may be much larger than the sample size. We provide an ℓ_1-oracle inequality satisfied by the Lasso estimator according to the Kullback-Leibler loss. This result is an extension of the ℓ_1-oracle inequality established by Meynet in [8] in the multivariate case. We focus on the Lasso for its ℓ_1-regularization properties rather than for the variable selection procedure, as it was done in Städler in [11].

Contents

1. Introduction 1
2. Notations and framework 3
2.1. Finite mixture regression model 3
2.2. Boundedness assumption on the mixture and component parameters 3
2.3. Maximum likelihood estimator and penalization 3
3. Oracle inequality 4
4. Proof of the oracle inequality 5
4.1. Main propositions used in this proof 5
4.2. Notations 7
4.3. Proof of the Theorem 4.1 thanks to the Propositions 4.2 and 4.3 8
4.4. Proof of the Theorem 3.1 8
5. Proof of the theorem according to $\mathcal{T}$ or $\mathcal{T}^c$ 9
5.1. Proof of the Proposition 4.2 9
5.2. Proof of the Proposition 4.3 11
6. Some details 13
6.1. Proof of the Lemma 5.1 13
6.2. Lemma 6.5 and Lemma 6.7 16
References 19

1. Introduction

Finite mixture regression models are useful for modeling the relationship between response and predictors, arising from different subpopulations. Due to recent improvements, we are faced with high-dimensional data where the number of covariables can be much larger than the sample size. We have to reduce the dimension to avoid identifiability problems. Considering a mixture of linear models, an assumption widely used is to say that only a few covariates explain the response. Among various methods, we focus on the ℓ_1-penalized least squares estimator of parameters to lead to sparse regression matrix. Indeed, it is a convex surrogate for the non-convex ℓ_0-penalization, and it produces sparse solutions. First introduced by Tibshirani in [12] in a linear model $Y = X\beta + \epsilon$, where $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, the Lasso estimator is defined in the

1991 *Mathematics Subject Classification.* 62H30.
Key words and phrases. Finite mixture of multivariate regression model, Lasso, ℓ_1-oracle inequality.

linear model by

$$\hat{\beta}^{\text{Lasso}}(\lambda) = \underset{\beta\in\mathbb{R}^p}{\text{argmin}} \left\{ ||Y - X\beta||_2^2 + \lambda||\beta||_1 \right\}, \qquad \lambda > 0.$$

Many results have been proved to study the performance of this estimator. For example, cite [1] and [4], for studying this estimator as a variable selection procedure in the linear model. Note that those results need strong assumptions on the Gram matrix X^tX, as the restrictive eigenvalue condition, that can be not fulfilled in practice. A summary of assumptions and results is given by Bühlmann and van de Geer in [13]. One can also cite van de Geer in [14] and discussions, who precises a chaining argument to perform rate, even in a non linear case.

If we assume that $(x_i, y_i)_{1\leq i\leq n}$ arise from different subpopulations, we could work with finite mixture regression models. Indeed, the homogeneity assumption of the linear model is often inadequate and restrictive. This model was introduced by Städler et al., in [10]. They assume that, for $i \in \{1, \ldots, n\}$, the observation y_i, conditionally to $X_i = x_i$, comes from a conditional density $s_{\xi^0}(.|x_i)$ which is a finite mixture of K Gaussian conditional densities with proportion vector $\boldsymbol{\pi}$, where

$$Y_i|X_i = x_i \sim s_{\xi^0}(y_i|x_i) = \sum_{k=1}^{K} \frac{\pi_k^0}{\sqrt{2\pi}\sigma_k^0} \exp\left(-\frac{(y_i - \beta_k^0 x_i)^2}{2(\sigma_k^0)^2} \right)$$

for some parameters $\xi^0 = (\pi_k^0, \beta_k^0, \sigma_k^0)_{1\leq k\leq K}$. They extend the Lasso estimator by

$$\hat{s}^{\text{Lasso}}(\lambda) = \underset{s_\xi^K}{\text{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log(s_\xi^K(y_i|x_i)) + \lambda \sum_{k=1}^{K} \pi_k \sum_{j=1}^{p} |\sigma_k^{-1}[\beta_k]_j| \right\}, \qquad \lambda > 0 \tag{1}$$

For this estimator, they provide an ℓ_0-oracle inequality satisfied by $\hat{s}^{\text{Lasso}}(\lambda)$, according to the restricted eigenvalue condition also, and margin conditions, which leads to link the Kullback-Leibler loss function to the ℓ_2-norm of the parameters.

Another way to study this estimator is to look after the Lasso for its ℓ_1-regularization properties. For example, cite [6], [8], and [9]. Contrary to the ℓ_0-results, some ℓ_1-results are valid with no assumptions, neither on the Gram matrix, nor on the margin. This can be achieved due to the fact that they are looking for rate of convergence of order $1/\sqrt{n}$ rather than $1/n$. For finite mixture Gaussian regression models, we could cite Meynet in [8] who gives an ℓ_1-oracle inequality for another extension of the Lasso estimator, defined by

$$\hat{s}^{\text{Lasso}}(\lambda) = \underset{s_\xi^K}{\text{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log(s_\xi^K(y_i|x_i)) + \lambda \sum_{k=1}^{K} \sum_{j=1}^{p} |[\beta_k]_j| \right\}, \qquad \lambda > 0. \tag{2}$$

In this article, we extend this result to finite mixture of multivariate Gaussian regression models. We will work with random multivariate variables $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^q$. As in [8], we shall restrict to the fixed design case, that is to say non-random regressors. We observe $(x_i)_{1\leq i\leq n}$. Without any restriction, we could assume that the regressors $x_i \in [0, 1]^p$ for all $i \in \{1, \ldots, n\}$. Under only bounded parameters assumption, we provide a lower bound on the Lasso regularization parameter λ which guarantees an oracle inequality.

This result is non-asymptotic: the number of observations is fixed, and the number p of covariates can grow. Remark that the number K of clusters in the mixture is supposed to be known. Our result is deduced from a finite mixture multivariate Gaussian regression model selection theorem for ℓ_1-penalized maximum likelihood conditional density estimation. We establish the general theorem following the one of Meynet in [8], which combines Vapnik's structural risk minimization method (see Vapnik in [16]) and theory around model selection (see Le Pennec and Cohen in [3] and Massart in [5]). As in Massart and Meynet in [6], our oracle inequality is deduced from this general theorem, the Lasso estimator being viewed 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 model 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 of multivariate Gaussian regression models. Section 4 is devoted to the proof of this result and of the general theorem, deriving from two easier propositions. Those propositions are proved in Section 5, whereas details of lemma states in Section 6.

2. Notations and framework

2.1. **Finite mixture regression model.** We observe n independent couples $(\mathbf{x},\mathbf{y}) = (x_i,y_i)_{1\leq i\leq n} \in ([0,1]^p \times \mathbb{R}^q)^n$, with $y_i \in \mathbb{R}^q$ a random observation, realization of variable Y_i, and $x_i \in [0,1]^p$ fixed for all $i \in \{1,\dots,n\}$. We assume that, conditionally to the x_is, the Y_is are independent and identically distributed with conditional density $s_{\xi^0}(.|x_i)$, which is a finite mixture of K Gaussian regressions with unknown parameters ξ^0. In this article, K is fixed, then we do not precise it with unknown parameters. We will estimate the unknown conditional density by a finite mixture of K Gaussian regressions. Each subpopulation is then estimated by a multivariate linear model. Detail the conditional density s_ξ. For all $y \in \mathbb{R}^q$, for all $x \in [0,1]^p$,

$$
\begin{aligned}
(3) \qquad & s_\xi(y|x) = \sum_{k=1}^{K} \frac{\pi_k}{(2\pi)^{q/2}\det(\Sigma_k)^{1/2}} \exp\left(-\frac{(y-\beta_k x)^t \Sigma_k^{-1}(y-\beta_k x)}{2}\right) \\
& \xi = (\pi_1,\dots,\pi_K,\beta_1,\dots,\beta_K,\Sigma_1,\dots,\Sigma_K) \in \Xi = \left(\Pi_K \times (\mathbb{R}^{q\times p})^K \times (\mathbb{S}_q^{++})^K\right) \\
& \Pi_K = \left\{(\pi_1,\dots,\pi_K); \pi_k > 0 \text{ for all } k \in \{1,\dots,K\} \text{ and } \sum_{k=1}^{K}\pi_k = 1\right\} \\
& \mathbb{S}_q^{++} \text{ is the set of symmetric positive definite matrices on } \mathbb{R}^q.
\end{aligned}
$$

We want to estimate ξ^0 from the observations. For all cluster $k \in \{1,\dots,K\}$, β_k is the matrix of regression coefficients, and Σ_k is the covariance matrix in the mixture component k, whereas the π_ks are the mixture proportions. For $x \in [0,1]^p$, we define the parameter $\xi(x)$ of the conditional density $s_\xi(.|x)$ by

$$\xi(x) = (\pi_1,\dots,\pi_K,\beta_1 x,\dots,\beta_K x,\Sigma_1,\dots,\Sigma_K) \in]0,1[^K \times (\mathbb{R}^q)^K \times (\mathbb{S}_q^{++})^K.$$

For all $k \in \{1,\dots,K\}$, for all $x \in [0,1]^p$, for all $z \in \{1,\dots,q\}$, $[\beta_k x]_z = \sum_{j=1}^{p}[\beta_k]_{z,j}[x]_j$, and then $\beta_k x \in \mathbb{R}^q$ is the mean vector of the mixture component k for the conditional density $s_\xi(.|x)$.

2.2. **Boundedness assumption on the mixture and component parameters.** Denote, for a matrix A, $m(A)$ the modulus of the smallest eigenvalue of A, and $M(A)$ the modulus of the largest eigenvalue of A. We shall restrict our study to bounded parameters vector $\xi = (\boldsymbol{\pi},\boldsymbol{\beta},\boldsymbol{\Sigma})$, where $\boldsymbol{\pi} = (\pi_1,\dots,\pi_K), \boldsymbol{\beta} = (\beta_1,\dots,\beta_K), \boldsymbol{\Sigma} = (\Sigma_1,\dots,\Sigma_K)$. Specifically, we assume that there exists deterministic positive constants $A_\beta, a_\Sigma, A_\Sigma, a_\pi$ such that ξ belongs to $\tilde{\Xi}$, with

$$
\begin{aligned}
(4) \quad \tilde{\Xi} = \Bigg\{ & \xi \in \Xi : \text{ for all } k \in \{1,\dots,K\}, \max_{z\in\{1,\dots,q\}} \sup_{x\in[0,1]^p} |[\beta_k x]_z| \leq A_\beta, \\
& a_\Sigma \leq m(\Sigma_k^{-1}) \leq M(\Sigma_k^{-1}) \leq A_\Sigma, a_\pi \leq \pi_k \Bigg\}.
\end{aligned}
$$

Let S the set of conditional densities s_ξ,

$$S = \left\{s_\xi, \xi \in \tilde{\Xi}\right\}.$$

2.3. **Maximum likelihood estimator and penalization.** In a maximum likelihood approach, we consider the Kullback-Leibler information as the loss function, which is defined for two densities s and t by

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

In a regression framework, we have to adapt this definition to take into account the structure of conditional densities. For the fixed covariates $(x_1,\dots,x_n)$, we consider the average loss function

$$\mathrm{KL}_n(s,t) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}(s(.|x_i),t(.|x_i)) = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^q}\log\left(\frac{s(y|x_i)}{t(y|x_i)}\right)s(y|x_i)dy.$$

Using the maximum likelihood approach, we want to estimate s_{ξ^0} by the conditional density s_ξ which maximizes the likelihood conditionally to $(x_i)_{1\leq i\leq n}$. Nevertheless, because we work with high-dimensional

data, we have to regularize the maximum likelihood estimator. We consider the ℓ_1-regularization, and a generalization of the estimator associated, the Lasso estimator, which we define by

$$\hat{s}^{\text{Lasso}}(\lambda) := \underset{\substack{s_\xi \in S \\ \xi=(\boldsymbol{\pi},\boldsymbol{\beta},\boldsymbol{\Sigma})}}{\operatorname{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log(s_\xi(y_i|x_i)) + \lambda \sum_{k=1}^{K}\sum_{z=1}^{q}\sum_{j=1}^{p} |[\beta_k]_{z,j}| \right\};$$

where $\lambda > 0$ is a regularization parameter.
We define also, for s_ξ defined as in (3), and with parameters $\xi = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$,

$$N_1^{[2]}(s_\xi) = ||\boldsymbol{\beta}||_1 = \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{z=1}^{q} |[\beta_k]_{z,j}|. \tag{5}$$

3. Oracle inequality

In this section, we provide an ℓ_1-oracle inequality satisfied by the Lasso estimator in finite mixture multivariate Gaussian regression models, which is the main result of this article.

Theorem 3.1. *We observe n couples $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n)) \in ([0,1]^p \times \mathbb{R}^q)^n$ coming from the conditional density s_{ξ^0}, where $\xi^0 \in \tilde{\Xi}$, where*

$$\tilde{\Xi} = \Bigg\{ \xi \in \Xi : \text{ for all } k \in \{1, \ldots, K\}, \max_{z\in\{1,\ldots,q\}} \sup_{x\in[0,1]^p} |[\beta_k x]_z| \leq A_\beta, \\ a_\Sigma \leq m(\Sigma_k^{-1}) \leq M(\Sigma_k^{-1}) \leq A_\Sigma, a_\pi \leq \pi_k \Bigg\}.$$

Denote by $a \vee b = \max(a, b)$.
We define the Lasso estimator, denoted by $\hat{s}^{Lasso}(\lambda)$, for $\lambda \geq 0$, by

$$\hat{s}^{Lasso}(\lambda) = \underset{s_\xi \in S}{\operatorname{argmin}} \left(-\frac{1}{n}\sum_{i=1}^{n} \log(s_\xi(y_i|x_i)) + \lambda N_1^{[2]}(s_\xi) \right); \tag{6}$$

with

$$S = \left\{ s_\xi, \xi \in \tilde{\Xi} \right\}$$

and where, for $\xi = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$,

$$N_1^{[2]}(s_\xi) = ||\boldsymbol{\beta}||_1 = \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{z=1}^{q} |[\beta_k]_{z,j}|.$$

Then, if

$$\lambda \geq \kappa \left(A_\Sigma \vee \frac{1}{a_\pi} \right) \left(1 + 4(q+1)A_\Sigma \left(A_\beta^2 + \frac{\log(n)}{a_\Sigma} \right) \right) \sqrt{\frac{K}{n}} \left(1 + q\log(n)\sqrt{K\log(2p+1)} \right)$$

with κ an absolute positive constant, the estimator (6) satisfies the following ℓ_1-oracle inequality.

$$\begin{aligned} \mathrm{E}[\mathrm{KL}_n(s_{\xi^0}, \hat{s}^{Lasso}(\lambda))] \leq \quad & (1+\kappa^{-1}) \quad \inf_{s_\xi \in S} \left(\mathrm{KL}_n(s_{\xi^0}, s_\xi) + \lambda N_1^{[2]}(s_\xi) \right) + \lambda \\ + \quad & \kappa' \sqrt{\tfrac{K}{n}} \quad \frac{e^{-\frac{1}{2}} \pi^{q/2} a_\pi}{A_\Sigma^{q/2}} \sqrt{2q} \\ + \quad & \kappa' \sqrt{\tfrac{K}{n}} \quad \left(A_\Sigma \vee \frac{1}{a_\pi} \right) \left(1 + 4(q+1)A_\Sigma \left(A_\beta^2 + \frac{\log(n)}{a_\Sigma} \right) \right) \\ & \qquad \times K \left(1 + A_\beta + \frac{q}{a_\Sigma} \right)^2 \end{aligned}$$

where κ' is a positive constant.

This theorem provides information about the performance of the Lasso as an ℓ_1-regularization algorithm. If 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, which is the solution of the ℓ_1-penalized true risk minimization problem, up to an error term of order λ.
Our result is non-asymptotic: the number n of observations is fixed while the number p of covariates and the size q of the response can grow with respect to n and can be much larger than n. The number K of clusters in the mixture is fixed.
There is no assumption neither on the Gram matrix, nor on the margin, which are classical assumptions for oracle inequality for the Lasso estimator. Moreover, this kind of assumptions involve unknown constants, whereas here, every constants are explicit. We could compare this result with the ℓ_0-oracle inequality established in [10], which needs those assumptions, and is therefore difficult to interpret. Nevertheless, they get faster rate, the error term in the oracle inequality being of order $1/n$ rather than $1/\sqrt{n}$.
The main assumption we make to establish the Theorem 3.1 is the boundedness of the parameters, which is also assumed in [10]. It is needed to tackle the problem of the unboundedness of the likelihood (see [7] for example).
Moreover, we let regressors to belong to $[0,1]^p$. Because we work with fixed covariates, they are finite. To simplify the reading, we choose to rescale $\mathbf{x}$ to get $||\mathbf{x}||_\infty \leq 1$. Nevertheless, if we not rescale the covariates, and the regularization parameter λ bound and the error term of the oracle inequality depend linearly of $||\mathbf{x}||_\infty$.
The regularization parameter λ bound is of order $(q^2+q)/\sqrt{n}\log(n)^2\sqrt{\log(2p+1)}$. For $q=1$, we recognize the same order, as regards to the sample size n and the number of covariates p, to the ℓ_1-oracle inequality in [8]. A great attention has been paid to get a lower bound of λ with optimal dependence on p, which is the number of regressors, but we are aware that dependences in q and K may not be optimal. Indeed, even if roles of p and q are not symmetric, we can wonder if a dependence of order logarithm in q could be expected, which is not achieved here. For the number of components, a dependence in $\sqrt{K}$ could be envisaged, see [8]. Those optimal rates are open problems.
Van de Geer, in [14], gives some tools to improve the bound of the regularization parameter to $\sqrt{\frac{\log(p)}{n}}$. Nevertheless, we have to control eigenvalues of the Gram matrix of some functions $(\psi_j(x_i))_{\substack{1\leq j\leq D\\ 1\leq i\leq n}}$, D being the number of parameters to estimate, where $\psi_j(x_i)$ satisfies

$$|\log(s_\xi(y_i|x_i)) - \log(s_{\tilde{\xi}}(y_i|x_i))| \leq \sum_{j=1}^{D} |\xi_j - \tilde{\xi}_j|\psi_j(x_i).$$

In our case of mixture of regression models, control eigenvalues of the Gram matrix of functions $(\psi_j(x_i))_{\substack{1\leq j\leq D\\ 1\leq i\leq n}}$ corresponds to make some assumptions, as REC, to avoid dimension reliance on n, K and p. Without this kind of assumptions, we could not guarantee that our bound is of order $\sqrt{\frac{\log(p)}{n}}$, because we could not guarantee that eigenvalues does not depend on dimensions. In order to get a result with smaller assumptions, we do not use the chaining argument developed in [14]. Nevertheless, one can easily compute that, under restricted eigenvalue condition, we could perform the order of the regularization parameter to $\lambda \asymp \sqrt{\frac{\log(p)}{n}}\log(n)$.

4. Proof of the oracle inequality

4.1. **Main propositions used in this proof.** The first result we will prove is the next theorem, which is an ℓ_1-ball mixture multivariate regression model selection theorem for ℓ_1-penalized maximum likelihood conditional density estimation in the Gaussian framework.

Theorem 4.1. *We observe $(x_i, y_i)_{1\leq i\leq n}$ with unknown conditional Gaussian mixture density s_{ξ^0}.*

For all $m \in \mathbb{N}^*$, we consider the ℓ_1-ball $S_m = \{s_\xi \in S, N_1^{[2]}(s_\xi) \leq m\}$ for $S = \{s_\xi, \xi \in \tilde{\Xi}\}$, and $\tilde{\Xi}$ defined by

$$\tilde{\Xi} = \Bigg\{ \xi \in \Xi : \text{ for all } k \in \{1, \dots, K\}, \max_{z \in \{1,\dots,q\}} \sup_{x \in [0,1]^p} |[\beta_k x]_z| \leq A_\beta, \\ a_\Sigma \leq m(||\Sigma_k^{-1}||) \leq M(\Sigma_k^{-1}) \leq A_\Sigma, a_\pi \leq \pi_k \Bigg\}.$$

For $\xi = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, let

$$N_1^{[2]}(s_\xi) = ||\boldsymbol{\beta}||_1 = \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{z=1}^{q} |[\beta_k]_{z,j}|.$$

Let $\hat{s}_m$ an η_m-log-likelihood minimizer in S_m, for $\eta_m \geq 0$:

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

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

$$\lambda \geq \kappa \left(A_\Sigma \vee \frac{1}{a_\pi} \right) \left(1 + 4(q+1) A_\Sigma \left(A_\beta^2 + \frac{\log(n)}{a_\Sigma} \right) \right) \sqrt{\frac{K}{n}} \left(1 + q \log(n) \sqrt{K \log(2p+1)} \right)$$

for a constant κ. Then, if $\hat{m}$ is such that

$$-\frac{1}{n} \sum_{i=1}^{n} \log(\hat{s}_{\hat{m}}(y_i|x_i)) + \text{pen}(\hat{m}) \leq \inf_{m \in \mathbb{N}^*} \left(-\frac{1}{n} \sum_{i=1}^{n} \log(\hat{s}_m(y_i|x_i)) + pen(m) \right) + \eta$$

for $\eta \geq 0$, the estimator $\hat{s}_{\hat{m}}$ satisfies

$$\begin{aligned} \mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})) \leq & (1 + \kappa^{-1}) \inf_{m \in \mathbb{N}^*} \left(\inf_{s_m \in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + \text{pen}(m) + \eta_m \right) + \eta \\ & + \kappa' \sqrt{\frac{K}{n}} \frac{e^{-\frac{1}{2}} \pi^{q/2}}{A_\Sigma^{q/2}} \sqrt{2q a_\pi} \\ & + \kappa' \sqrt{\frac{K}{n}} K \left(A_\Sigma \vee \frac{1}{a_\pi} \right) \left(1 + \frac{4(q+1)}{2} A_\Sigma \left(A_\beta^2 + \frac{\log(n)}{a_\sigma} \right) \right) \\ & \times \left(1 + A_\beta + \frac{q}{a_\Sigma} \right)^2; \end{aligned}$$

where κ' is a positive constant.

It is an ℓ_1-ball mixture regression model selection theorem for ℓ_1-penalized maximum likelihood conditional density estimation in the Gaussian framework. Its proof could be deduced from the two following propositions, which split the result if the variable Y is large enough or not.

Proposition 4.2. *We observe $(x_i, y_i)_{1 \leq i \leq n}$, with unknown conditional density denoted by s_{ξ^0}. Let $M_n > 0$, and consider the event*

$$\mathcal{T} := \left\{ \max_{i \in \{1,\dots,n\}} \max_{z \in \{1,\dots,q\}} |[Y_i]_z| \leq M_n \right\}.$$

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

$$S_m = \{s_\xi \in S, N_1^{[2]}(s_\xi) \leq m\}$$

where $S = \{s_\xi, \xi \in \tilde{\Xi}\}$ and

$$N_1^{[2]}(s_\xi) = ||\boldsymbol{\beta}||_1 = \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{z=1}^{q} |[\beta_k]_{z,j}|$$

for $\xi = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$.

Let $\hat{s}_m$ an η_m-log-likelihood minimizer in S_m, for $\eta_m \geq 0$:

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

Let $C_{M_n} = \max\left(\frac{1}{a_\pi}, A_\Sigma + \frac{1}{2}(|M_n|+A_\beta)^2 A_\Sigma^2, \frac{q(|M_n|+A_\beta)A_\Sigma}{2}\right)$. Assume that for all $m \in \mathbb{N}^*$, the penalty function satisfies $\mathrm{pen}(m) = \lambda m$ with

$$\lambda \geq \kappa\frac{4C_{M_n}}{\sqrt{n}}\sqrt{K}\left(1+9q\log(n)\sqrt{K\log(2p+1)}\right)$$

for some absolute constant κ. Then, any estimate $\hat{s}_{\hat{m}}$ with $\hat{m}$ such that

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

for $\eta \geq 0$, satisfies

$$\begin{aligned}\mathrm{E}(\mathrm{KL}_n(s_{\xi^0},\hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}}) \leq & (1+\kappa^{-1})\inf_{m\in\mathbb{N}^*}\left(\inf_{s_m\in S_m}\mathrm{KL}_n(s_{\xi^0},s_m)+\mathrm{pen}(m)+\eta_m\right)\\ &+\frac{\kappa^{'}K^{3/2}qC_{M_n}}{\sqrt{n}}\left(1+\left(A_\beta+\frac{q}{a_\Sigma}\right)^2\right);\end{aligned}$$

where $\kappa^{'}$ is an absolute positive constant.

Proposition 4.3. *Let $s_{\xi^0}, \mathcal{T}$ and $\hat{s}_{\hat{m}}$ defined as in the previous proposition. Then,*

$$\mathrm{E}(\mathrm{KL}_n(s_{\xi^0},\hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}^c}) \leq \frac{e^{-1/2}\pi^{q/2}}{A_\Sigma^{q/2}}\sqrt{2Knqa_\pi}e^{-1/4(M_n^2-2M_nA_\beta)a_\Sigma}.$$

4.2. **Notations.** To prove those two propositions, and the theorem, begin with some notations. For any measurable function $g : \mathbb{R}^q \mapsto \mathbb{R}$, we consider the empirical norm

$$g_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n}g^2(y_i|x_i)};$$

its conditional expectation

$$\mathrm{E}(g(Y|X)) = \int_{\mathbb{R}^q} g(y|x)s_{\xi^0}(y|x)dy;$$

its empirical process

$$P_n(g) := \frac{1}{n}\sum_{i=1}^{n}g(y_i|x_i);$$

and its normalized process

$$\nu_n(g) := P_n(g) - \mathrm{E}_X(P_n(g)) = \frac{1}{n}\sum_{i=1}^{n}\left[g(y_i|x_i) - \int_{\mathbb{R}^q} g(y|x_i)s_{\xi^0}(y|x_i)dy\right].$$

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

$$F_m = \left\{f_m = -\log\left(\frac{s_m}{s_{\xi^0}}\right), s_m \in S_m\right\}.$$

Let $\delta_{\mathrm{KL}} > 0$. For all $m \in \mathbb{N}^*$, let $\eta_m \geq 0$. There exist two functions, denoted by $\hat{s}_{\hat{m}}$ and $\bar{s}_m$, belonging to S_m, such that

$$P_n(-\log(\hat{s}_{\hat{m}})) \leq \inf_{s_m\in S_m} P_n(-\log(s_m))+\eta_m;$$

$$\mathrm{KL}_n(s_{\xi^0},\bar{s}_m) \leq \inf_{s_m\in S_m}\mathrm{KL}_n(s_{\xi^0},s_m)+\delta_{\mathrm{KL}}. \tag{7}$$

Denote by $\hat{f}_m := -\log\left(\frac{\hat{s}_m}{s_{\xi^0}}\right)$ and $\bar{f}_m := -\log\left(\frac{\bar{s}_m}{s_{\xi^0}}\right)$. Let $\eta \geq 0$ and fix $m \in \mathbb{N}^*$. We define the set $M(m)$ by

$$M(m) = \{m' \in \mathbb{N}^* | P_n(-\log(\hat{s}_{m'})) + \text{pen}(m') \leq P_n(-\log(\hat{s}_m)) + \text{pen}(m) + \eta\}. \tag{8}$$

4.3. **Proof of the Theorem 4.1 thanks to the Propositions 4.2 and 4.3.** Let $M_n > 0$ and $\kappa \geq 36$. Let $C_{M_n} = \max\left(\frac{1}{a_\pi}, A_\Sigma + \frac{1}{2}(|M_n| + A_\beta)^2 A_\Sigma^2, q(|M_n| + A_\beta)A_\Sigma/2\right)$. Assume that, for all $m \in \mathbb{N}^*$, $\text{pen}(m) = \lambda m$, with

$$\lambda \geq \kappa C_{M_n} \sqrt{\frac{K}{n}} \left(1 + q\log(n)\sqrt{K\log(2p+1)}\right).$$

We derive from the two propositions that there exists κ' such that, if $\hat{m}$ satisfies

$$-\frac{1}{n}\sum_{i=1}^n \log(\hat{s}_{\hat{m}}(y_i|x_i)) + \text{pen}(\hat{m}) \leq \inf_{m\in\mathbb{N}^*}\left(-\frac{1}{n}\sum_{i=1}^n \log(\hat{s}_m(y_i|x_i)) + \text{pen}(m)\right) + \eta;$$

then $\hat{s}_{\hat{m}}$ satisfies

$$\begin{aligned}
\mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})) =& \mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}}) + \mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}^c}) \\
\leq& (1+\kappa^{-1}) \inf_{m\in\mathbb{N}^*}\left(\inf_{s_m\in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + \text{pen}(m) + \eta_m\right) \\
&+ \kappa' \frac{C_{M_n}}{\sqrt{n}} K^{3/2} q\left(1 + (A_\beta + \frac{q}{a_\Sigma})^2\right) + \eta \\
&+ \kappa' \frac{e^{-1/2}\pi^{q/2}}{A_\Sigma^{q/2}} \sqrt{2Knqa_\pi} e^{-1/4(M_n^2 - 2M_n A_\beta)a_\Sigma}.
\end{aligned}$$

In order to optimize this equation with respect to M_n, we consider M_n the positive solution of the polynomial

$$\log(n) - \frac{1}{4}(X^2 - 2XA_\beta)a_\Sigma = 0;$$

we obtain $M_n = A_\beta + \sqrt{A_\beta^2 + \frac{4\log(n)}{a_\Sigma}}$ and $\sqrt{n}e^{-1/4(M_n^2 - 2M_nA_\beta)a_\Sigma} = \frac{1}{\sqrt{n}}$.
On the other hand,

$$\begin{aligned}
C_{M_n} &\leq \left(A_\Sigma \vee \frac{1}{a_\pi}\right)\left[1 + \frac{q+1}{2}A_\Sigma(M_n + A_\beta)^2\right] \\
&\leq \left(A_\Sigma \vee \frac{1}{a_\pi}\right)\left[1 + 4(q+1)A_\Sigma\left(A_\beta^2 + \frac{\log(n)}{a_\Sigma}\right)\right].
\end{aligned}$$

We get

$$\begin{aligned}
\mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})) = \quad & \mathrm{E}(\mathrm{KL}_n \quad (s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}}) + \mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}^c}) \\
\leq \quad & (1+\kappa^{-1}) \quad \inf_{m\in\mathbb{N}^*}\left(\inf_{s_m\in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + \text{pen}(m) + \eta_m\right) + \eta \\
& +\kappa'\sqrt{\tfrac{K}{n}} \quad \frac{e^{-\frac{1}{2}}\pi^{q/2}}{(qA_\Sigma)^{q/2}}\sqrt{2qa_\pi} \\
& +\kappa'\sqrt{\tfrac{K}{n}} \quad \left(A_\Sigma \vee \frac{1}{a_\pi}\right)\left(1 + 4(q+1)A_\Sigma\left(A_\beta^2 + \frac{\log(n)}{a_\Sigma}\right)\right) \\
& \qquad\qquad \times K\left(1 + \left(A_\beta + \frac{q}{a_\Sigma}\right)^2\right).
\end{aligned}$$

4.4. **Proof of the Theorem 3.1.** We will show that there exists $\eta_m \geq 0$, and $\eta \geq 0$ such that $\hat{s}^{\text{Lasso}}(\lambda)$ satisfies the hypothesis of the Theorem 4.1, which will lead to Theorem 3.1.
First, let show that there exists $m \in \mathbb{N}^*$ and $\eta_m \geq 0$ such that the Lasso estimator is an η_m-log-likelihood minimizer in S_m.

For all $\lambda \geq 0$, if $m_\lambda = \lceil N_1^{[2]}(\hat{s}(\lambda)) \rceil$,

$$\hat{s}^{\text{Lasso}}(\lambda) = \underset{\substack{s \in S \\ N_1^{[2]}(s) \leq m_\lambda}}{\operatorname{argmin}} \left(-\frac{1}{n} \sum_{i=1}^{n} \log(s(y_i|x_i)) \right).$$

We could take $\eta_m = 0$.
Secondly, let show that there exists $\eta \geq 0$ such that

$$-\frac{1}{n} \sum_{i=1}^{n} \log(\hat{s}^{\text{Lasso}}(\lambda)(y_i|x_i)) + \text{pen}(m_\lambda) \leq \inf_{m \in \mathbb{N}^*} \left(-\frac{1}{n} \sum_{i=1}^{n} \log(\hat{s}_m(y_i|x_i)) + \text{pen}(m) \right) + \eta.$$

Taking $\text{pen}(m_\lambda) = \lambda m_\lambda$,

$$\begin{aligned}
-\frac{1}{n} \sum_{i=1}^{n} \log(\hat{s}^{\text{Lasso}}(\lambda)(y_i|x_i)) + \text{pen}(m_\lambda) &= -\frac{1}{n} \sum_{i=1}^{n} \log(\hat{s}^{\text{Lasso}}(\lambda)(y_i|x_i)) + \lambda m_\lambda \\
&\leq -\frac{1}{n} \sum_{i=1}^{n} \log(\hat{s}^{\text{Lasso}}(\lambda)(y_i|x_i)) + \lambda N_1^{[2]}(\hat{s}^{\text{Lasso}}(\lambda)) + \lambda \\
&\leq \inf_{s_\xi \in S} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(s_\xi(y_i|x_i)) + \lambda N_1^{[2]}(s_\xi) \right\} + \lambda \\
&\leq \inf_{m \in \mathbb{N}^*} \inf_{s_\xi \in S_m} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(s_\xi(y_i|x_i)) + \lambda N_1^{[2]}(s_\xi) \right\} + \lambda \\
&\leq \inf_{m \in \mathbb{N}^*} \left(\inf_{s_\xi \in S_m} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(s_\xi(y_i|x_i)) \right\} + \lambda m \right) + \lambda \\
&\leq \inf_{m \in \mathbb{N}^*} \left(-\frac{1}{n} \sum_{i=1}^{n} \log(\hat{s}_m(y_i|x_i)) + \lambda m \right) + \lambda;
\end{aligned}$$

which is exactly the goal, with $\eta = \lambda$. Then, according to the Theorem 4.1, with $\hat{m} = m_\lambda$, and $\hat{s}_{\hat{m}} = \hat{s}^{\text{Lasso}}(\lambda)$, for

$$\lambda \geq \kappa \left(A_\Sigma \vee \frac{1}{a_\pi} \right) \left(1 + 4(q+1) A_\Sigma \left(A_\beta^2 + \frac{\log(n)}{a_\Sigma} \right) \right) \sqrt{\frac{K}{n}} \left(1 + q \log(n) \sqrt{K \log(2p+1)} \right),$$

we get the oracle inequality.

5. Proof of the theorem according to $\mathcal{T}$ or $\mathcal{T}^c$

5.1. **Proof of the Proposition 4.2.** This proposition corresponds to the main theorem according to the event $\mathcal{T}$. To prove it, we need some preliminary results.
From our notations, reminded in Section 4.2, we have, for all $m \in \mathbb{N}^*$ for all $m' \in M(m)$,

$$\begin{aligned}
P_n(\hat{f}_{m'}) + \text{pen}(m') &\leq P_n(\hat{f}_m) + \text{pen}(m) + \eta \leq P_n(\bar{f}_m) + \text{pen}(m) + \eta_m + \eta; \\
\text{E}(P_n(\hat{f}_{m'})) + \text{pen}(m') &\leq \text{E}(P_n(\bar{f}_m)) + \text{pen}(m) + \eta_m + \eta + \nu_n(\bar{f}_m) - \nu_n(\hat{f}_{m'}); \\
\text{KL}_n(s_{\xi^0}, \hat{s}_{m'}) + \text{pen}(m') &\leq \inf_{s_m \in S_m} \text{KL}_n(s_{\xi^0}, s_m) + \delta_{\text{KL}} + \text{pen}(m) + \eta_m + \eta + \nu_n(\bar{f}_m) - \nu_n(\hat{f}_{m'}); \qquad (9)
\end{aligned}$$

thanks to the inequality (7).
The goal is to bound $-\nu_n(\hat{f}_{m'}) = \nu_n(-\hat{f}_{m'})$.
To control this term, we use the following lemma.

Lemme 5.1. *Let $M_n > 0$. Let*

$$\mathcal{T} = \left\{ \max_{i \in \{1,\ldots,n\}} \left(\max_{z \in \{1,\ldots,q\}} |[Y_i]_z| \right) \leq M_n \right\}.$$

Let $C_{M_n} = \max\left(\frac{1}{a_\pi}, A_\Sigma + \frac{1}{2}(|M_n| + A_\beta)^2 A_\Sigma^2, \frac{q(|M_n|+A_\beta)A_\Sigma}{2}\right)$ *and*

$$\Delta_{m'} = m' \log(n)\sqrt{K\log(2p+1)} + 6\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right).$$

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

$$\sup_{f_{m'} \in \mathcal{F}_{m'}} |\nu_n(-f_{m'})| \leq \frac{4C_{M_n}}{\sqrt{n}}\left(9\sqrt{K}q\Delta_{m'} + \sqrt{2}\sqrt{t}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)\right)$$

Proof. Page 13 ▲

From (9), on the event $\mathcal{T}$, for all $m \in \mathbb{N}^*$, for all $m' \in M(m)$, for all $t > 0$, with probability greater than $1 - e^{-t}$,

$$\begin{aligned}
\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{m'}) + \mathrm{pen}(m') &\leq \inf_{s_m \in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + \delta_{\mathrm{KL}} + \mathrm{pen}(m) + \nu_n(\bar{f}_m) \\
&\quad + \frac{4C_{M_n}}{\sqrt{n}}\left(9\sqrt{K}q\Delta_{m'} + \sqrt{2}\sqrt{t}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)\right) + \eta_m + \eta \\
&\leq \inf_{s_m \in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + \mathrm{pen}(m) + \nu_n(\bar{f}_m) \\
&\quad + 4\frac{C_{M_n}}{\sqrt{n}}\left(9\sqrt{K}q\Delta_{m'} + \frac{1}{2\sqrt{K}}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)^2 + \sqrt{K}t\right) \\
&\quad + \eta_m + \eta + \delta_{\mathrm{KL}},
\end{aligned}$$

the last inequality being true because $2ab \leq \frac{1}{\sqrt{K}}a^2 + \sqrt{K}b^2$. Let $z > 0$ such that $t = z + m + m'$. On the event $\mathcal{T}$, for all $m \in \mathbb{N}$, for all $m' \in M(m)$, with probability greater than $1 - e^{-(z+m+m')}$,

$$\begin{aligned}
\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{m'}) + \mathrm{pen}(m') &\leq \inf_{s_m \in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + \mathrm{pen}(m) + \nu_n(\bar{f}_m) \\
&\quad + 4\frac{C_{M_n}}{\sqrt{n}}\left(9\sqrt{K}q\Delta_{m'} + \frac{1}{2\sqrt{K}}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)^2\right) \\
&\quad + 4\frac{C_{M_n}}{\sqrt{n}}\left(\sqrt{K}(z + m + m')\right) \\
&\quad + \eta_m + \eta + \delta_{\mathrm{KL}}.
\end{aligned}$$

$$\begin{aligned}
\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{m'}) - \nu_n(\bar{f}_m) &\leq \inf_{s_m \in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + \mathrm{pen}(m) + 4\frac{C_{M_n}}{\sqrt{n}}\sqrt{K}m \\
&\quad + \left[\frac{4C_{M_n}}{\sqrt{n}}\sqrt{K}(m' + 9q\Delta_{m'}) - \mathrm{pen}(m')\right] \\
&\quad + \frac{4C_{M_n}}{\sqrt{n}}\left(\frac{1}{2\sqrt{K}}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)^2 + \sqrt{K}z\right) + \eta_m + \eta + \delta_{\mathrm{KL}}.
\end{aligned}$$

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

$$\lambda \geq \frac{4C_{M_n}}{\sqrt{n}}\sqrt{K}\left(1 + 9q\log(n)\sqrt{K\log(2p+1)}\right)$$

Then, as

$$\Delta_{m'} = m' \log(n)\sqrt{K\log(2p+1)} + 6\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right),$$

with

$$\kappa^{-1} = \frac{4C_{M_n}}{\sqrt{n}}\sqrt{K}\frac{1}{\lambda} \leq \frac{1}{1 + 9q\log(n)\sqrt{K\log(2p+1)}},$$

we get that

$$\begin{aligned}
\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{m'}) - \nu_n(\bar{f}_m) &\leq \inf_{s_m \in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + (1+\kappa^{-1})\,\mathrm{pen}(m) \\
&\quad + \frac{4C_{M_n}}{\sqrt{n}} \frac{1}{2\sqrt{K}} \left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)^2 \\
&\quad + \frac{4C_{M_n}}{\sqrt{n}} \left(54\sqrt{K}q\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right) + \sqrt{K}z\right) \\
&\quad + \eta + \delta_{\mathrm{KL}} + \eta_m \\
&\leq \inf_{s_m \in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + (1+\kappa^{-1})\,\mathrm{pen}(m) \\
&\quad + \frac{4C_{M_n}}{\sqrt{n}} \left(27K^{3/2} + \frac{27+1/2}{\sqrt{K}}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)^2 + \sqrt{K}z\right) \\
&\quad + \eta_m + \eta + \delta_{\mathrm{KL}}.
\end{aligned}$$

Let $\hat{m}$ such that

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

and $M(m) = \{m' \in \mathbb{N}^* | P_n(-\log(\hat{s}_{m'})) + \mathrm{pen}(m') \leq P_n(-\log(\hat{s}_m)) + \mathrm{pen}(m) + \eta\}$. By definition, $\hat{m} \in M(m)$. Because for all $m \in \mathbb{N}^*$, for all $m' \in M(m)$,

$$1 - \sum_{\substack{m \in \mathbb{N}^* \\ m' \in M(m)}} e^{-(z+m+m')} \geq 1 - e^{-z} \sum_{(m,m') \in (\mathbb{N}^*)^2} e^{-m-m'} \geq 1 - e^{-z},$$

we could sum up over all models.
On the event $\mathcal{T}$, for all $z > 0$, with probability greater than $1 - e^{-z}$,

$$\begin{aligned}
\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}}) - \nu_n(\bar{f}_m) &\leq \inf_{m \in \mathbb{N}^*} \left(\inf_{s_m \in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + (1+\kappa^{-1})\,\mathrm{pen}(m) + \eta_m\right) \\
&\quad + \frac{4C_{M_n}}{\sqrt{n}} \left(27K^{3/2} + \frac{55q}{2\sqrt{K}}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)^2 + \sqrt{K}z\right) \\
&\quad + \eta + \delta_{\mathrm{KL}}.
\end{aligned}$$

By integrating over $z > 0$, and noticing that $\mathrm{E}(\nu_n(\bar{f}_m)) = 0$ and that δ_{KL} can be chosen arbitrary small, we get

$$\begin{aligned}
\mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}}) &\leq \inf_{m \in \mathbb{N}^*} \left(\inf_{s_m \in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + (1+\kappa^{-1})\,\mathrm{pen}(m) + \eta_m\right) \\
&\quad + \frac{4C_{M_n}}{\sqrt{n}} \left(27K^{3/2} + \frac{q}{\sqrt{K}}\frac{55}{2}\left(1 + K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)^2 + \sqrt{K}\right) + \eta \\
&\leq \inf_{m \in \mathbb{N}^*} \left(\inf_{s_m \in S_m} \mathrm{KL}_n(s_{\xi^0}, s_m) + (1+\kappa^{-1})\,\mathrm{pen}(m) + \eta_m\right) \\
&\quad + \frac{332K^{3/2}qC_{M_n}}{\sqrt{n}} \left(1 + \left(A_\beta + \frac{q}{a_\Sigma}\right)^2\right) + \eta.
\end{aligned}$$

5.2. **Proof of the Proposition 4.3.** We want an upper bound of $\mathrm{E}\left(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}^c}\right)$. Thanks to the Cauchy-Schwarz inequality,

$$\mathrm{E}\left(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}^c}\right) \leq \sqrt{\mathrm{E}(\mathrm{KL}_n^2(s_{\xi^0}, \hat{s}_{\hat{m}}))}\sqrt{P(\mathcal{T}^c)}.$$

However, for all $s_\xi \in S$,

$$\begin{aligned}
\mathrm{KL}_n(s_{\xi^0}, s_\xi) &= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^q} \log\left(\frac{s_{\xi^0}(y|x_i)}{s_\xi(y|x_i)}\right) s_{\xi^0}(y|x_i)dy \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\int_{\mathbb{R}^q} \log(s_{\xi^0}(y|x_i))s_{\xi^0}(y|x_i)dy - \int_{\mathbb{R}^q} \log(s_\xi(y|x_i))s_{\xi^0}(y|x_i)dy\right) \\
&\leq -\frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^q} \log(s_\xi(y|x_i))s_{\xi^0}(y|x_i)dy.
\end{aligned}$$

Because parameters are assumed to be bounded, according to the set $\tilde{\Xi}$ defined in (4), we get, with $(\boldsymbol{\beta^0}, \boldsymbol{\Sigma^0}, \boldsymbol{\pi^0})$ the parameters of s_{ξ^0} and $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$ the parameters of s_ξ,

$$\begin{aligned}
\log(s_\xi(y|x_i))s_{\xi^0}(y|x_i) =& \log\left(\sum_{k=1}^{K}\frac{\pi_k}{(2\pi)^{q/2}\sqrt{\det(\Sigma_k)}}\exp\left(-\frac{(y-\beta_k x_i)^t\Sigma_k^{-1}(y-\beta_k x_i)}{2}\right)\right) \\
&\times \sum_{k=1}^{K}\frac{\pi_k^0}{(2\pi)^{q/2}\sqrt{\det(\Sigma_k^0)}}\exp\left(-\frac{(y-\beta_k^0 x_i)^t(\Sigma_k^0)^{-1}(y-\beta_k^0 x_i)}{2}\right) \\
\geq& \log\left(K\frac{a_\pi\sqrt{\det(\Sigma_1^{-1})}}{(2\pi)^{q/2}}\exp\left(-(y^t\Sigma_1^{-1}y + x_i^t\beta_1^t\Sigma_1^{-1}\beta_1 x_i)\right)\right) \\
&\times K\frac{a_\pi\sqrt{\det((\Sigma_1^0)^{-1})}}{(2\pi)^{q/2}}\exp\left(-(y^t\Sigma_1^{-1}y + x_i^t\beta_1^t\Sigma_1^{-1}\beta_1 x_i)\right) \\
\geq& \log\left(K\frac{a_\pi a_\Sigma^{q/2}}{(2\pi)^{q/2}}\exp\left(-(y^t y + A_\beta^2)A_\Sigma\right)\right) \\
&\times K\frac{a_\pi a_\Sigma^{q/2}}{(2\pi)^{q/2}}\exp\left(-(y^t y + A_\beta^2)A_\Sigma\right).
\end{aligned}$$

Indeed, for $u \in \mathbb{R}^q$, if we use the eigenvalue decomposition of $\Sigma = P^tDP$,

$$\begin{aligned}
|u^t\Sigma u| = |u^tP^tDPu| &\leq ||Pu||_2||DPU||_2 \leq M(D)||Pu||_2^2 \\
&\leq M(D)||u||_2^2 \leq A_\Sigma||u||_2^2.
\end{aligned}$$

To recognize the expectation of a Gaussian standardized variables, we put $u = \sqrt{2A_\Sigma}y$:

$$\begin{aligned}
\mathrm{KL}(s_{\xi^0}(.|x_i), s_\xi(.|x_i)) &\leq -\frac{Ka_\pi e^{-A_\beta^2 A_\Sigma}a_\Sigma^{q/2}}{(2A_\Sigma)^{q/2}}\int_{\mathbb{R}^q}\left[\log\left(\frac{Ka_\Sigma^{q/2}a_\pi}{(2\pi)^{q/2}}\right) - A_\beta^2 A_\Sigma - \frac{u^tu}{2}\right]\frac{e^{\frac{-u^tu}{2}}}{(2\pi)^{q/2}}du \\
&\leq -\frac{a_\Sigma^{q/2}Ka_\pi e^{-A_\beta^2 A_\Sigma}}{(2A_\Sigma)^{q/2}}\,\mathrm{E}\left[\log\left(\frac{Ka_\pi a_\Sigma^{q/2}}{(2\pi)^{q/2}}\right) - A_\beta^2 A_\Sigma - \frac{U^2}{2}\right] \\
&\leq -\frac{Ka_\Sigma^{q/2}a_\pi e^{-A_\beta^2 A_\Sigma}}{(2A_\Sigma)^{q/2}}\left[\log\left(\frac{Ka_\pi a_\Sigma^{q/2}}{(2\pi)^{q/2}}\right) - A_\beta^2 A_\Sigma - \frac{1}{2}\right] \\
&\leq -\frac{Ka_\Sigma^{q/2}a_\pi e^{-A_\beta^2 A_\Sigma - 1/2}}{(2\pi)^{q/2}A_\Sigma^{q/2}}e^{1/2}\pi^{q/2}\log\left(\frac{Ka_\pi e^{-A_\beta^2 A_\Sigma - 1/2}a_\Sigma^{q/2}}{(2\pi)^{q/2}}\right) \\
&\leq \frac{e^{-1/2}\pi^{q/2}}{A_\Sigma^{q/2}};
\end{aligned}$$

where $U \sim \mathcal{N}_q(0,1)$. We have used that for all $t \in \mathbb{R}$, $t\log(t) \geq -e^{-1}$. Then, we get, for all $s_\xi \in S$,

$$\mathrm{KL}_n(s_{\xi^0}, s_\xi) \leq \frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}(s_{\xi^0}(.|x_i), s_\xi(.|x_i)) \leq \frac{e^{-1/2}\pi^{q/2}}{A_\Sigma^{q/2}}.$$

As it is true for all $s_\xi \in S$, it is true for $\hat{s}_{\hat{m}}$, then

$$\sqrt{\mathrm{E}(\mathrm{KL}_n^2(s_{\xi^0}, \hat{s}_{\hat{m}}))} \leq \frac{e^{-1/2}\pi^{q/2}}{A_\Sigma^{q/2}}.$$

For the last step, we need to bound $P(\mathcal{T}^c)$.

$$P(\mathcal{T}^c) = \mathrm{E}(\mathbb{1}_{\mathcal{T}^c}) = \mathrm{E}(\mathrm{E}_X(\mathbb{1}_{\mathcal{T}^c})) = \mathrm{E}(P_X(\mathcal{T}^c)) \leq \mathrm{E}\left(\sum_{i=1}^n P_X(||Y_i||_\infty > M_n)\right).$$

Nevertheless, let $Y_x \sim \sum_{k=1}^K \pi_k \mathcal{N}_q(\beta_k x, \Sigma_k)$, then,

$$\begin{aligned}
P(||Y_x||_\infty > M_n) &= \int_{\mathbb{R}^q} \mathbb{1}_{\{||Y_x||_\infty \geq M_n\}} \sum_{k=1}^K \pi_k \frac{1}{(2\pi)^{q/2}\sqrt{\det(\Sigma_k)}} e^{\left(-\frac{(y-\beta_k x)^t \Sigma_k^{-1}(y-\beta_k x)}{2}\right)} dy \\
&= \sum_{k=1}^K \pi_k \int_{\mathbb{R}^q} \mathbb{1}_{\{||Y_x||_\infty \geq M_n\}} \frac{1}{(2\pi)^{q/2}\sqrt{\det(\Sigma_k)}} e^{\left(-\frac{(y-\beta_k x)^t \Sigma_k^{-1}(y-\beta_k x)}{2}\right)} dy \\
&= \sum_{k=1}^K \pi_k P_X(||Y_x^k||_\infty > M_n) \leq \sum_{k=1}^K \sum_{z=1}^q \pi_k P_X(|[Y_x^k]_z| > M_n)
\end{aligned}$$

with $Y_x^k \sim \mathcal{N}(\beta_k x, \Sigma_k)$ and $[Y_x^k]_z \sim \mathcal{N}([\beta_k x]_z, [\Sigma_k]_{z,z})$.
We need to control $P_X(|[Y_x^k]_z| > M_n)$, for all $z \in \{1, \ldots, q\}$.

$$\begin{aligned}
P_X(|[Y_x^k]_z| > M_n) &= P_X([Y_x^k]_z > M_n) + P_X([Y_{x,k}]_z < -M_n) \\
&= P_X\left(U > \frac{M_n - [\beta_k x]_z}{\sqrt{[\Sigma_k]_{z,z}}}\right) + P_X\left(U < \frac{-M_n - [\beta_k x]_z}{\sqrt{[\Sigma_k]_{z,z}}}\right) \\
&= P_X\left(U > \frac{M_n - [\beta_k x]_z}{\sqrt{[\Sigma_k]_{z,z}}}\right) + P_X\left(U > \frac{M_n + [\beta_k x]_z}{\sqrt{[\Sigma_k]_{z,z}}}\right) \\
&\leq e^{-\frac{1}{2}\left(\frac{M_n - [\beta_k x]_z}{\sqrt{[\Sigma_k]_{z,z}}}\right)^2} + e^{-\frac{1}{2}\left(\frac{M_n + [\beta_k x]_z}{\sqrt{[\Sigma_k]_{z,z}}}\right)^2} \\
&\leq 2e^{-\frac{1}{2}\left(\frac{M_n - |[\beta_k x]_z|}{\sqrt{[\Sigma_k]_{z,z}}}\right)^2} \\
&\leq 2e^{-\frac{1}{2}\frac{M_n^2 - 2M_n|[\beta_k x]_z| + |[\beta_k x]_z|^2}{[\Sigma_k]_{z,z}}}.
\end{aligned}$$

where $U \sim \mathcal{N}(0,1)$. Then,

$$P(||Y_x||_\infty > M_n) \leq 2Kqe^{-\frac{1}{2}(M_n^2 - 2M_n A_\beta)a_\Sigma},$$

and we get $P(\mathcal{T}^c) \leq \mathrm{E}\left(\sum_{i=1}^n 2Kqa_\pi e^{-\frac{1}{2}(M_n^2 - 2M_n A_\beta)a_\Sigma}\right) \leq 2Kna_\pi q e^{-\frac{1}{2}(M_n^2 - 2M_n A_\beta)a_\Sigma}$. We have obtained the wanted bound for $\mathrm{E}(\mathrm{KL}_n(s_{\xi^0}, \hat{s}_{\hat{m}})\mathbb{1}_{\mathcal{T}^c})$.

6. Some details

6.1. **Proof of the Lemma 5.1.** First, give some tools to prove the Lemma 5.1.
We define $||g||_n = \sqrt{\frac{1}{n}\sum_{i=1}^n g^2(y_i|x_i)}$ for any measurable function g.
Let $m \in \mathbb{N}^*$. We have

$$\sup_{f_m \in F_m} |\nu_n(-f_m)| = \sup_{f_m \in F_m} \left|\frac{1}{n}\sum_{i=1}^n (f_m(y_i|x_i) - \mathrm{E}(f_m(Y_i|x_i)))\right|.$$

To control the deviation of such a quantity, we shall combine concentration with symmetrization arguments. We first use the following concentration inequality which can be found in [2].

Lemme 6.1. *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 $t > 0$,*

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

Proof. See [2]. ▲

Then, we propose to bound $\mathrm{E}\left[\sup_{\gamma\in\Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\gamma(Z_i) - \mathrm{E}(\gamma(Z_i))\right|\right]$ thanks to the following symmetrization argument. The proof of this result can be found in [15].

Lemme 6.2. *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 $(\epsilon_1, \dots, \epsilon_n)$ be a Rademacher sequence independent of $(Z_1, \dots, Z_n)$. Then,*

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

Proof. See [15]. ▲

Then, we have to control $\mathrm{E}(\sup_{\gamma\in\Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\gamma(Z_i)\right|)$.

Lemme 6.3. *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 $(\epsilon_1, \dots, \epsilon_n)$ be a Rademacher sequence independent of $(Z_1, \dots, Z_n)$. Define R_n a non-random constant such that*

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

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

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

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 $||.||_n$.

Proof. See [5]. ▲

In our case, we get the following lemma.

Lemme 6.4. *Let $m \in \mathbb{N}^*$. Consider $(\epsilon_1, \dots, \epsilon_n)$ a Rademacher sequence independent of $(Y_1, \dots, Y_n)$. Then, on the event T,*

$$\mathrm{E}\left(\sup_{f_m\in F_m}\left|\sum_{i=1}^{n}\epsilon_i f_m(Y_i|x_i)\right|\right) \leq 18\sqrt{K}\frac{C_{M_n}q}{\sqrt{n}}\Delta_m;$$

where $\Delta_m := m\log(n)\sqrt{K\log(2p+1)} + 6(1 + K(A_\beta + \frac{q}{a_\Sigma}))$.

Proof. Let $m \in \mathbb{N}^*$. According to Lemma 6.5, we get that on the event $\mathcal{T}$,

$$\sup_{f_m\in F_m} ||f_m||_n \leq R_n := 2C_{M_n}(1 + K(A_\beta + \frac{q}{a_\Sigma})).$$

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

$$
\begin{aligned}
&\sum_{s=1}^{S} 2^{-s}\sqrt{\log[1+N(2^{-s}R_n, F_m, ||.||_n)]} \leq \sum_{s=1}^{S} 2^{-s}\sqrt{\log(2N(2^{-s}R_n, F_m, ||.||_n))} \\
\leq &\sum_{s=1}^{S} 2^{-s}\left(\sqrt{\log(2)} + \sqrt{\log(2p+1)}\frac{2^{s+1}C_{M_n}qKm}{R_n}\right) \\
&+ \sum_{s=1}^{S} 2^{-s}\sqrt{K\log\left(1+\frac{2^{s+3}C_{M_n}qK}{R_n a_\Sigma}\right)\left(1+\frac{2^{s+3}C_{M_n}}{R_n}\right)} \text{ according to Lemma 6.7} \\
\leq &\sum_{s=1}^{S} 2^{-s}\left(\sqrt{\log(2)} + \sqrt{\log(2p+1)}\frac{2^{s+1}C_{M_n}qKm}{R_n}\right) \\
&+ \sum_{s=1}^{S} 2^{-s}\sqrt{K\log\left(1+2^{s+3}\frac{C_{M_n}}{R_n}\max(1, qK/a_\Sigma)\right)^2} \\
\leq &\sum_{s=1}^{S} 2^{-s}\left[\sqrt{\log(2)} + \sqrt{\log(2p+1)}\frac{2^{s+1}C_{M_n}qKm}{R_n} + \sqrt{2(s+3)K\log(2)q/a_\Sigma}\right] \\
\leq &\frac{2C_{M_n}Kmq}{R_n}S\sqrt{\log(2p+1)} + \sqrt{\log(2)}\left(1+\frac{\sqrt{q}}{a_\Sigma}\left(\sqrt{6K} + 2\sum_{s=1}^{S}2^{-s}\sqrt{s}\right)\right) \\
\leq &\frac{2C_{M_n}Kmq}{R_n}S\sqrt{\log(2p+1)} + \sqrt{\log(2)}\left(1+\frac{\sqrt{q}}{a_\Sigma}\sqrt{6K} + \sqrt{q}\sqrt{K}\frac{\sqrt{2e}}{2-\sqrt{e}}\right)
\end{aligned}
$$

because $2^{-s}\sqrt{s} \leq \left(\frac{\sqrt{e}}{2}\right)^s$ for all $s \in \mathbb{N}^*$. Then, thanks to the Lemma 6.3,

$$
\begin{aligned}
\mathrm{E}\left(\sup_{f_m \in F_m}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_m(Y_i|x_i)\right|\right) \leq &\quad R_n \quad \left(\frac{6}{\sqrt{n}}\sum_{s=1}^{S}2^{-s}\sqrt{\log[1+N(2^{-s}R_n, F_m, ||.||_n)]} + 2^{-S}\right) \\
\leq &\quad R_n \quad \left[\frac{6}{\sqrt{n}}\left(\frac{2C_{M_n}Kmq}{R_n}S\sqrt{\log(2p+1)}\right.\right. \\
&\left.\left. + \sqrt{\log(2)}\left(1+\frac{q}{a_\Sigma}\sqrt{6K} + \frac{q}{a_\Sigma}\sqrt{K}\frac{2e}{2-\sqrt{e}}\right)\right) + 2^{-S}\right].
\end{aligned}
$$

Taking $S = \frac{\log(n)}{\log(2)}$ to obtain the same order in the both terms depending on S, we could deduce that

$$
\begin{aligned}
&\mathrm{E}\left(\sup_{f_m \in F_m}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_m(Y_i|x_i)\right|\right) \\
\leq &\frac{12C_{M_n}Kmq}{\sqrt{n}}\sqrt{\log(2p+1)}\frac{\log(n)}{\log(2)} + 2C_{M_n}\left(1+K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)\left[\frac{\sqrt{\log(2)}}{\sqrt{n}}\left(1+\sqrt{6K}+\frac{\sqrt{2e}}{2-\sqrt{2e}}\right)+\frac{1}{n}\right] \\
\leq &\frac{18C_{M_n}Kmq}{\sqrt{n}}\sqrt{\log(2p+1)}\log(n) + 2\frac{\sqrt{K}}{\sqrt{n}}C_{M_n}\left(1+K\left(A_\beta + \frac{q}{a_\Sigma}\right)\right)\left[\sqrt{\log(2)}\left(1+\sqrt{6}+\frac{\sqrt{2e}}{2-\sqrt{2e}}\right)+1\right] \\
\leq &18\frac{\sqrt{K}}{\sqrt{n}}C_{M_n}\left[mq\sqrt{K\log(2p+1)}\log(n) + 6\left(1+K\left(A_\beta+\frac{q}{a_\Sigma}\right)\right)\right].
\end{aligned}
$$

It completes the proof. ▲

We are now able to prove the Lemma 5.1.

$$\begin{aligned}
\sup_{f_m \in F_m} |\nu_n(-f_m)| &= \sup_{f_m \in F_m} \left| \frac{1}{n} \sum_{i=1}^{n} (f_m(y_i|x_i) - \mathrm{E}_X(f_m(Y_i|x_i))) \right| \\
&\leq \mathrm{E}\left(\sup_{f_m \in F_m} \left| \sum_{i=1}^{n} f_m(Y_i|x_i) - \mathrm{E}(f_m(Y_i|x_i)) \right| \right) + 2\sqrt{2} R_n \sqrt{\frac{t}{n}} \\
&\text{with probability greater than } 1 - e^{-t} \text{ and where } R_n \\
&\text{is a constant computed from the Lemma 6.5} \\
&\leq 2\,\mathrm{E}\left(\sup_{f_m \in F_m} \left| \sum_{i=1}^{n} \epsilon_i f_m(Y_i|x_i) \right| \right) + 2\sqrt{2} R_n \sqrt{\frac{t}{n}} \\
&\text{with } \epsilon_i \text{ a Rademacher sequence,} \\
&\text{independent of } Z_i \\
&\leq 2\left(18\sqrt{K} \frac{C_{M_n} q}{\sqrt{n}} \Delta_m \right) + 2\sqrt{2} R_n \sqrt{\frac{t}{n}} \\
&\leq 4 C_{M_n} \left(9 \frac{\sqrt{K} q}{\sqrt{n}} \Delta_m + \sqrt{2} \sqrt{\frac{t}{n}} \left(1 + K \left(A_\beta + \frac{q}{a_\Sigma} \right) \right) \right).
\end{aligned}$$

6.2. **Lemma 6.5 and Lemma 6.7.**

Lemme 6.5. *On the event*

$$\mathcal{T} = \left\{ \max_{i \in \{1,\dots,n\}} \max_{z \in \{1,\dots,q\}} |[Y_i]_z| \leq M_n \right\},$$

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

$$\sup_{f_m \in F_m} ||f_m||_n \leq 2 C_{M_n} \left(1 + K \left(A_\beta + \frac{q}{a_\Sigma} \right) \right) := R_n.$$

Proof. Let $m \in \mathbb{N}^*$. Because $f_m \in \mathcal{F}_m = \left\{ f_m = -\log\left(\frac{s_m}{s_{\xi^0}}\right), s_m \in S_m \right\}$, there exists $s_m \in S_m$ such that $f_m = -\log\left(\frac{s_m}{s_{\xi^0}}\right)$. For all $x \in [0,1]^p$, denote $\xi(x) = (\boldsymbol{\pi}, \beta_1 x, \dots, \beta_K x, \boldsymbol{\Sigma})$ the parameters of $s_m(.|x)$. For all $i \in \{1, \dots, n\}$,

$$\begin{aligned}
|f_m(y_i|x_i)| \mathbb{1}_{\mathcal{T}} &= |\log(s_m(y_i|x_i)) - \log(s_{\xi^0}(y_i|x_i))| \mathbb{1}_{\mathcal{T}} \\
&\leq \sup_{x \in [0,1]^p} \sup_{\xi \in \Xi} \left| \frac{\partial \log(s_\xi(y_i|x))}{\partial \xi} \right| ||\xi(x_i) - \xi^0(x_i)||_1 \mathbb{1}_{\mathcal{T}},
\end{aligned}$$

thanks to the Taylor formula. Then, we need an upper bound of the partial derivate.
For all $x \in [0,1]^p$, for all $y \in \mathbb{R}^q$, we could write

$$\log(s_\xi(y|x)) = \log\left(\sum_{k=1}^{K} h_k(x,y) \right)$$

where, for all $k \in \{1, \dots, K\}$,

$$\begin{aligned}
h_k(x,y) =& \frac{\pi_k}{(2\pi)^{q/2} \det \Sigma_k} \\
&\times \exp\left[-\frac{1}{2} \left(\sum_{z_2=1}^{q} \left(\sum_{z_1=1}^{q} y_{z_1} - \sum_{j=1}^{p} x_j [\beta_k]_{z_1,j} \right) [\Sigma_k]^{-1}_{z_1,z_2} \right) \left(y_{z_2} - \sum_{j=1}^{p} [\beta_k]_{z_2,j} x_j \right) \right].
\end{aligned}$$

Then, for all $l \in \{1,\dots,K\}$, for all $z_1 \in \{1,\dots,q\}$, for all $z_2 \in \{1,\dots,q\}$, for all $y \in \mathbb{R}^q$, for all $x \in [0,1]^p$,

$$\begin{aligned}
\left|\frac{\partial \log(s_\xi(y|x))}{\partial([\beta_l x]_{z_1})}\right| &= \left|\frac{h_l(x,y)}{\sum_{k=1}^K h_k(x,y)}\right|\left(-\frac{1}{2}\sum_{z_2=1}^q [\Sigma_l]^{-1}_{z_1,z_2}([\beta_l x]_{z_2} - y_{z_2})\right) \leq \frac{q(|y|+A_\beta)A_\Sigma}{2};\\
\left|\frac{\partial \log(s_\xi(y|x))}{\partial([\Sigma_l]_{z_1,z_2})}\right| &= \frac{1}{\sum_{k=1}^K h_k(x,y)}\\
&\quad \times \left|\frac{-h_l \mathrm{Cof}_{z_1,z_2}(\Sigma_l)}{\det(\Sigma_l)} - \frac{h_l(x,y)(y_{z_1}-[\beta_l x]_{z_1})(y_{z_2}-[\beta_l x]_{z_2})[\Sigma_l]^{-2}_{z_1,z_2}}{2}\right|\\
&\leq \left|\frac{-\mathrm{Cof}_{z_1,z_2}(\Sigma_l)}{\det(\Sigma_l)} + \frac{(y_{z_1}-[\beta_l x]_{z_1})(y_{z_2}-[\beta_l x]_{z_2})[\Sigma_l]^{-2}_{z_1,z_2}}{2}\right|\\
&\leq A_\Sigma + \frac{1}{2}(|y|+A_\beta)^2 A_\Sigma^2,
\end{aligned}$$

where $\mathrm{Cof}_{z_1,z_2}(\Sigma_k)$ is the (z_1,z_2)-cofactor of Σ_k. We also have, for all $l \in \{1,\dots,K\}$, for all $x \in [0,1]^p$, for all $y \in \mathbb{R}^q$,

$$\left|\frac{\partial \log(s_\xi(y,x))}{\partial \pi_l}\right| = \left|\frac{h_l(x,y)}{\pi_l \sum_{k=1}^K h_k(x,y)}\right| \leq \frac{1}{a_\pi}.$$

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

$$\sup_{x\in[0,1]^p}\sup_{\xi\in\tilde{\Xi}}\left|\frac{\partial \log(s_\xi(y|x))}{\partial \xi}\right| \leq \max\left(\frac{1}{a_\pi}, A_\Sigma + \frac{1}{2}(|y|+A_\beta)^2 A_\Sigma^2, \frac{q(|y|+A_\beta)A_\Sigma}{2}\right) = C_y.$$

We have $C_y \leq \left(A_\Sigma \wedge \frac{1}{a_\pi}\right)\left[1 + \frac{q+1}{2}A_\Sigma(|y|+A_\beta)^2\right]$. For all $m \in \mathbb{N}^*$,

$$\begin{aligned}
|f_m(y_i|x_i)|\mathbb{1}_{\mathcal{T}} &\leq C_{y_i}||\xi(x_i) - \xi^0(x_i)||_1 \mathbb{1}_{\mathcal{T}}\\
&\leq C_{M_n}\sum_{k=1}^K (||\beta_k x_i - \beta_k^0 x_i||_1 + ||\Sigma_k - \Sigma_k^0||_1 + |\pi_k - \pi_k^0|).
\end{aligned}$$

Since f_m and f_m^0 belong to $\tilde{\Xi}$, we obtain

$$|f_m(y_i|x_i)|\mathbb{1}_{\mathcal{T}} \leq 2C_{M_n}(KA_\beta + K\frac{q}{a_\Sigma} + 1)$$

and then

$$\sup_{f_m\in F_m} ||f_m||_n \mathbb{1}_{\mathcal{T}} \leq 2C_{M_n}(KA_\beta + K\frac{q}{a_\Sigma} + 1).$$

▲

For the next results, we need the following lemma, proved in [8].

Lemme 6.6. *Let $\delta > 0$ and $(A_{i,j})_{\substack{i\in\{1,\dots,n\}\\ j\in\{1,\dots,p\}}} \in [0,1]^{n\times p}$. There exists a family B of $(2p+1)^{1/\delta^2}$ vectors of $\mathbb{R}^p$ such that for all $\mu \in \mathbb{R}^p$ in the ℓ_1-ball, there exists $\mu' \in B$ such that*

$$\frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^p (\mu_j - \mu'_j)A_{i,j}\right)^2 \leq \delta^2.$$

Proof. See [8]. ▲

With this lemma, we can prove the following one:

Lemme 6.7. *Let $\delta > 0$ and $m \in \mathbb{N}^*$. On the event $\mathcal{T}$, we have the upper bound of the δ-packing number of the set of functions F_m equipped with the metric induced by the norm $||.||_n$:*

$$N(\delta, F_m, ||.||_n) \leq (2p+1)^{4C_{M_n}^2 K^2 q^2 m^2/\delta^2} \left(1 + \frac{8C_{M_n} qK}{a_\Sigma \delta}\right)^K \left(1 + \frac{8C_{M_n}}{\delta}\right)^K.$$

Proof. Let $m \in \mathbb{N}^*$ and $f_m \in F_m$. There exists $s_m \in S_m$ such that $f_m = -\log(s_m/s_{\xi^0})$. Introduce s'_m in S and put $f'_m = -\log(s'_m/s_{\xi^0})$. Denote by $(\beta_k, \Sigma_k, \pi_k)_{1\leq k\leq K}$ and $(\beta'_k, \Sigma'_k, \pi'_k)_{1\leq k\leq K}$ the parameters of the densities s_m and s'_m respectively. First, applying Taylor's inequality, on the event

$$\mathcal{T} = \left\{ \max_{i\in\{1,\dots,n\}} \max_{z\in\{1,\dots,q\}} |[Y_i]_z| \leq M_n \right\},$$

we get, for all $i \in \{1, \dots, n\}$,

$$\begin{aligned}
|f_m(y_i|x_i) - f'_m(y_i|x_i)|\mathbb{1}_\mathcal{T} &= |\log(s_m(y_i|x_i)) - \log(s'_m(y_i|x_i))|\mathbb{1}_\mathcal{T} \\
&\leq \sup_{x\in[0,1]^p} \sup_{\xi\in\tilde{\Xi}} \left| \frac{\partial \log(s_\xi(y_i|x))}{\partial \xi} \right| ||\xi(x_i) - \xi'(x_i)||_1 \mathbb{1}_\mathcal{T} \\
&\leq C_{M_n} \sum_{k=1}^K \left(\sum_{z=1}^q |[\beta_k x_i]_z - [\beta'_k x_i]_z| + ||\Sigma_k - \Sigma'_k||_1 + |\pi_k - \pi'_k| \right).
\end{aligned}$$

Thanks to the Cauchy-Schwarz inequality, we get that

$$\begin{aligned}
(f_m(y_i|x_i) - f'_m(y_i|x_i))^2 \mathbb{1}_\mathcal{T} &\leq 2C_{M_n}^2 \left[\left(\sum_{k=1}^K \sum_{z=1}^q |\beta_k x_i - \beta'_k x_i| \right)^2 + (||\boldsymbol{\Sigma} - \boldsymbol{\Sigma'}||_1 + ||\boldsymbol{\pi} - \boldsymbol{\pi'}||)^2 \right] \\
&\leq 2C_{M_n}^2 \left[Kq \sum_{k=1}^K \sum_{z=1}^q \left(\sum_{j=1}^p [\beta_k]_{z,j}[x_i]_j - \sum_{j=1}^p [\beta'_k]_{z,j}[x_i]_j \right)^2 + (||\boldsymbol{\Sigma} - \boldsymbol{\Sigma'}||_1 + ||\boldsymbol{\pi} - \boldsymbol{\pi'}||)^2 \right],
\end{aligned}$$

and

$$\begin{aligned}
||f_m - f'_m||_n^2 \mathbb{1}_\mathcal{T} \leq & 2C_{M_n}^2 \left[Kq \sum_{k=1}^K \sum_{z=1}^q \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p [\beta_k]_{z,j}[x_i]_j - \sum_{j=1}^p [\beta'_k]_{z,j}[x_i]_j \right)^2 \right. \\
& \left. + (||\boldsymbol{\Sigma} - \boldsymbol{\Sigma'}||_1 + ||\boldsymbol{\pi} - \boldsymbol{\pi'}||)^2 \right].
\end{aligned}$$

Denote by

$$a = Kq \sum_{k=1}^K \sum_{z=1}^q \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p [\beta_k]_{z,j}[x_i]_j - \sum_{j=1}^p [\beta'_k]_{z,j}[x_i]_j \right)^2.$$

Then, for all $\delta > 0$, if

$$\begin{aligned}
a &\leq \delta^2/(4C_{M_n}^2) \\
||\boldsymbol{\Sigma} - \boldsymbol{\Sigma'}||_1 &\leq \delta/(4C_{M_n}) \\
||\boldsymbol{\pi} - \boldsymbol{\pi'}|| &\leq \delta/(4C_{M_n})
\end{aligned}$$

then $||f_m - f'_m||_n^2 \leq \delta^2$. To bound a, we write

$$a = Kqm^2 \sum_{=1}^K \sum_{z=1}^q \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \frac{[\beta_k]_{z,j}}{m}[x_i]_j - \sum_{j=1}^p \frac{[\beta'_k]_{z,j}}{m}[x_i]_j \right)^2$$

and we apply Lemma 6.6 to $[\beta_k]_{z,.}/m$ for all $k \in \{1,\dots,K\}$, and for all $z \in \{1,\dots,q\}$. Since $s_m \in S_m$, we have $\sum_{z=1}^{q}\sum_{j=1}^{p}\left|\frac{[\beta_k]_{z,j}}{m}\right| \leq 1$, thus there exists a family $\mathcal{B}$ of $(2p+1)^{4C_{M_n}^2 q^2K^2m^2/\delta^2}$ vectors of $\mathbb{R}^p$ such that for all $k \in \{1,\dots,K\}$, for all $z \in \{1,\dots,q\}$, for all $[\beta_k]_{z,.}$, there exists $[\beta_k^{'}]_{z,.} \in \mathcal{B}$ such that $a \leq \delta^2/(4C_{M_n}^2)$. Moreover, since $||\mathbf{\Sigma}||_1 \leq \frac{qK}{a_\Sigma}$ and $||\boldsymbol{\pi}||_1 \leq 1$, we get that, on the event $\mathcal{T}$,

$$\begin{aligned} N(\delta, F_m, ||.||_n) &\leq \mathrm{card}(\mathcal{B}) N\left(\frac{\delta}{4C_{M_n}}, B_1^K(\frac{qK}{A_\Sigma}), ||.||_1\right) N\left(\frac{\delta}{4C_{M_n}}, B_1^K(1), ||.||_1\right) \\ &\leq (2p+1)^{4C_{M_n}^2 q^2K^2m^2/\delta^2}\left(1+\frac{8C_{M_n}qK}{a_\Sigma\delta}\right)^K\left(1+\frac{8C_{M_n}}{\delta}\right)^K \end{aligned}$$

▲

References

[1] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[2] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[3] S. Cohen and E. Le Pennec. Conditional density estimation by penalized likelihood model selection and applications. Research Report RR-7596, 2011.

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

[5] P. Massart. *Concentration inequalities and model selection*. Lecture Notes in Mathematics. Springer, 33, 2003, Saint-Flour, Cantal, 2007.

[6] P. Massart and C. Meynet. The Lasso as an ℓ_1-ball model selection procedure. *Electronic Journal of Statistics*, 5:669–687, 2011.

[7] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley series in probability and statistics: Applied probability and statistics. Wiley, 2004.

[8] C. Meynet. An ℓ_1-oracle inequality for the lasso in finite mixture gaussian regression models. *ESAIM: Probability and Statistics*, 17:650–671, 2013.

[9] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

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

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

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

[13] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[14] S. van de Geer, P. Bühlmann, and S. Zhou. The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). *Electronic Journal of Statistics*, 5:688–749, 2011.

[15] AW van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996.

[16] V. Vapnik. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1982.

Laboratoire de Mathématiques d'Orsay, Faculté des Sciences d'Orsay, Université Paris-Sud, 91405 Orsay, France