Electronic Journal of Statistics
Vol. 8 (2014) 1661–1695
ISSN: 1935-7524
DOI: 10.1214/14-EJS939

Mixture of Gaussian regressions model with logistic weights, a penalized maximum likelihood approach

L. Montuelle

LMO, Université Paris Sud/Inria
e-mail: Lucie.Montuelle@math.u-psud.fr

and

E. Le Pennec

CMAP, École Polytechnique
e-mail: Erwan.Le-Pennec@polytechnique.edu

Abstract: In the framework of conditional density estimation, we use candidates taking the form of mixtures of Gaussian regressions with logistic weights and means depending on the covariate. We aim at estimating the number of components of this mixture, as well as the other parameters, by a penalized maximum likelihood approach. We provide a lower bound on the penalty that ensures an oracle inequality for our estimator. We perform some numerical experiments that support our theoretical analysis.

AMS 2000 subject classifications: 62G08.
Keywords and phrases: Mixture of Gaussian regressions models, mixture of regressions models, penalized likelihood, model selection.

Received November 2013.

Contents

1. Framework

In classical Gaussian mixture models, the density is modeled by

$$s_{K,\upsilon,\Sigma,w}(y) = \sum_{k=1}^{K} \pi_{w,k} \Phi_{\upsilon_k,\Sigma_k}(y),$$

where $K \in \mathbb{N} \setminus \{0\}$ is the number of mixture components, $\Phi_{\upsilon,\Sigma}$ is the Gaussian density with mean υ and covariance matrix Σ,

$$\Phi_{\upsilon,\Sigma}(y) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{\frac{-1}{2}(y-\upsilon)' \Sigma^{-1} (y-\upsilon)}$$

and $\pi_{w,k}$ are the mixture weights, that can always be defined from a K-tuple $w = (w_1, \ldots, w_K)$ with a logistic scheme:

$$\pi_{w,k} = \frac{e^{w_k}}{\sum_{k'=1}^{K} e^{w_{k'}}}.$$

In this article, we consider such a model in which the mixture weights as well as the means can depend on a, possibly multivariate, covariate.

More precisely, we observe n pairs of random variables $((X_i, Y_i))_{1 \le i \le n}$ where the covariates X_is are independent while the Y_is are conditionally independent given the X_is. We assume that the covariates are in some subset $\mathcal{X}$ of $\mathbb{R}^d$ and the Y_is are in $\mathbb{R}^p$. We want to estimate the conditional density $s_0(\cdot|x)$ with respect to the Lebesgue measure of Y given X. We model this conditional density by a mixture of Gaussian regressions with varying logistic weights

$$s_{K,\upsilon,\Sigma,w}(y|x) = \sum_{k=1}^{K} \pi_{w(x),k} \Phi_{\upsilon_k(x),\Sigma_k}(y),$$

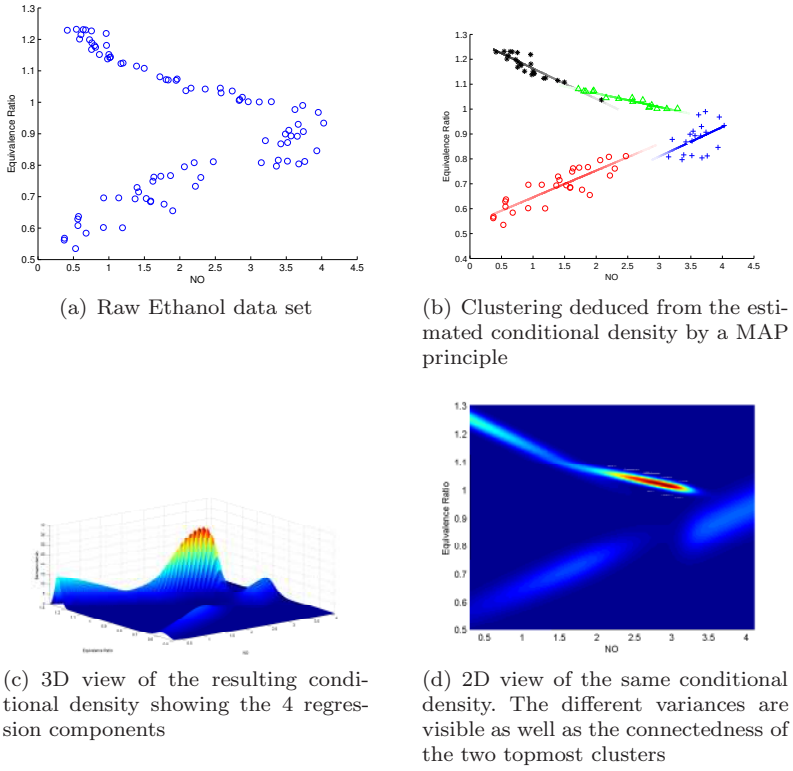
where $\upsilon = (\upsilon_1, \ldots, \upsilon_K)$ and $w = (w_1, \ldots, w_K)$ are now K-tuples of functions chosen, respectively, in a set Υ_K and W_K. Our aim is then to estimate those functions υ_k and w_k, the covariance matrices Σ_k as well as the number of classes K so that the *error* between the estimated conditional density and the true conditional density is *as small as possible*.

The classical Gaussian mixture case has been extensively studied (McLachlan and Peel, 2000). Nevertheless, theoretical properties of such model have been less considered. In a Bayesian framework, asymptotic properties of the posterior distribution are obtained by Choi (2008), Genovese and Wasserman (2000), Van der Vaart and Wellner (1996) when the true density is assumed to be a Gaussian mixture. AIC/BIC penalization scheme are often used to select a number of clusters (see Burnham and Anderson (2002) for instance). Non asymptotic bounds are obtained by Maugis and Michel (2011) even when the true density is not a Gaussian mixture. All these works rely heavily on a *bracketing* entropy analysis of the models, that will also be central in our analysis.

When there is a covariate, the most classical extension of this model is a mixture of Gaussian regressions, in which the means υ_k are now functions. It is well studied as described in McLachlan and Peel (2000). In particular, in a Bayesian framework, Viele and Tong (2002) have used bracketing entropy bounds to prove the consistency of the posterior distribution. Models in which the proportions vary have been considered by Antoniadis et al. (2009). Using an idea of Kolaczyk et al. (2005), they have considered a model in which only proportions depend in a piecewise constant manner from the covariate. Their theoretical results are nevertheless obtained under the strong assumption they exactly know the Gaussian components. This assumption can be removed as shown by Cohen and Le Pennec (2013). Models in which both mixture weights and means depend on the covariate are considered by Ge and Jiang (2006), but in a mixture of logistic regressions framework. They give conditions on the number of components (experts) to obtain consistency of the posterior with logistic weights. Note that similar properties are studied by Lee (2000) for neural networks.

Although natural, mixture of Gaussian regressions with varying logistic weights seems to be mentioned first by Jordan and Jacobs (1994). They provide an algorithm similar to ours, based on EM and Iteratively Reweighted Least Squares, for hierarchical mixtures of experts but no theoretical analysis. Young and Hunter (2010) choose a non-parametric approach to estimate the weights, which are not supposed logistic anymore, using kernels and cross-validation. They also provide an EM-like algorithm and some convincing simulations. This work has an extension in a series of papers (Hunter and Young, 2012), (Huang and Yao, 2012). Young (2014) considers mixture of regressions with changepoints but constant proportions. More recently, Huang et al. (2013) have considered a non-parametric modeling for the means, the proportions as well as the variance for which they give asymptotic properties as well as a numerical algorithm. Closer to our work, Chamroukhi et al. (2010) consider the case of piecewise polynomial regression model with affine logistic weights. In our setting, this corresponds to a specific choice for Υ_K and W_K: a collection of piecewise polynomials and a set of affine functions. They use a variation of the EM algorithm and a BIC criterion and provide numerical experiments to support the efficiency of their scheme.

Young (2014) provides a relevant example for our analysis. The ethanol data set of Brinkman (1981) (Figure 1(a)) shows the relationship between the equivalence ratio, a measure of the air-ethanol mix used as a spark-ignition engine



(a) Raw Ethanol data set

(b) Clustering deduced from the estimated conditional density by a MAP principle

(c) 3D view of the resulting conditional density showing the 4 regression components

(d) 2D view of the same conditional density. The different variances are visible as well as the connectedness of the two topmost clusters

FIG 1. *Estimated density with* 4 *components based upon the NO data set.*

fuel in a single-cylinder automobile test, and the engine's concentration of nitrogen oxide (NO) emissions for 88 tests. Using the methodology described in this paper, we obtain a conditional density modeled by a mixture of four Gaussian regressions. Using a classical maximum likelihood approach, each point of the data set can be assigned to one the four class yielding the clustering of Figure 1(b). The use of logistic weight allows a soft partitioning along the NO axis while still allowing more than one regression for the same NO value. The two topmost classes seem to correspond to a single population whose behavior changes around 1.7 while the two bottom-most classes appear to correspond to two different populations with a gap around 2.6–2.9. Such a result could not have been obtained with non varying weights.

The main contribution of our paper is a theoretical result: an oracle inequality, a non asymptotic bound on the risk, that holds for penalty slightly different from the one used by Chamroukhi et al. (2010).

In Section 2, we recall the penalized maximum likelihood framework, introduce the losses considered and explain the meaning of such an oracle inequality. In Section 3, we specify the models considered and their collections, state our theorem under mild assumptions on the sets Υ_K and W_K and apply this result

to polynomial sets. Those results are then illustrated by some numerical experiments in Section 4. Our analysis is based on an abstract theoretical analysis of penalized maximum likelihood approach for conditional densities conducted in Cohen and Le Pennec (2011) that relies on bracketing entropy bounds. Appendix A summarizes those results while Appendix B contains the proofs specific to this paper, the ones concerning bracketing entropies.

2. A model selection approach

2.1. Penalized maximum likelihood estimator

We will use a model selection approach and define some conditional density models S_m by specifying sets of conditional densities, taking the shape of mixtures of Gaussian regressions, through their number of classes K, a structure on the covariance matrices Σ_k and two function sets Υ_K and W_K to which belong respectively the K-tuple of means $(\upsilon_1, \ldots, \upsilon_K)$ and the K-tuple of logistic weights $(w_1, \ldots, w_K)$. Typically those sets are compact subsets of polynomials of low degree. Within such a conditional density set S_m, we estimate s_0 by the maximizer $\widehat{s}_m$ of the likelihood

$$\widehat{s}_m = \operatorname*{argmax}_{s_{K,\upsilon,\Sigma,w} \in S_m} \sum_{i=1}^{n} \ln s_{K,\upsilon,\Sigma,w}(Y_i|X_i),$$

or more precisely, to avoid any existence issue since the infimum may not be unique or even not be reached, by any η-minimizer of the negative log-likelihood:

$$\sum_{i=1}^{n} -\ln \widehat{s}_m(Y_i|X_i) \leq \inf_{s_{K,\upsilon,\Sigma,w} \in S_m} \sum_{i=1}^{n} -\ln s_{K,\upsilon,\Sigma,w}(Y_i|X_i) + \eta.$$

Assume now we have a collection $\{S_m\}_{m \in \mathcal{M}}$ of models, for instance with different number of classes K or different maximum degree for the polynomials defining Υ_K and W_K, we should choose the best model within this collection. Using only the log-likelihood is not sufficient since this favors models with large complexity. To balance this issue, we will define a penalty pen(m) and select the model $\widehat{m}$ that minimizes (or rather η'-almost minimizes) the sum of the negative log-likelihood and this penalty:

$$\sum_{k=1}^{K} -\ln \widehat{s}_{\widehat{m}}(Y_i|X_i) + \operatorname{pen}(\widehat{m}) \leq \inf_{m \in \mathcal{M}} \sum_{k=1}^{K} -\ln \widehat{s}_m(Y_i|X_i) + \operatorname{pen}(m) + \eta'.$$

2.2. Losses

Classically in maximum likelihood context, the estimator loss is measured with the Kullback-Leibler divergence KL. Since we work in a conditional density framework, we use a *tensorized* version of it. We define the tensorized Kullback-

Leibler divergence $\mathrm{KL}^{\otimes n}$ by

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

which appears naturally in this setting. Replacing t by a convex combination between s and t and dividing by ρ yields the so-called tensorized Jensen-Kullback-Leibler divergence, denoted $\mathrm{JKL}_\rho^{\otimes n}$,

$$\mathrm{JKL}_\rho^{\otimes n}(s,t) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\rho}\mathrm{KL}(s(.|X_i), (1-\rho)s(.|X_i) + \rho t(.|X_i))\right]$$

with $\rho \in (0,1)$. This loss is always bounded by $\frac{1}{\rho}\ln\frac{1}{1-\rho}$ but behaves as KL when t is close to s. This boundedness turns out to be crucial to control the loss of the penalized maximum likelihood estimate under mild assumptions on the complexity of the model and their collection.

Furthermore $\mathrm{JKL}_\rho^{\otimes n}(s,t) \leq \mathrm{KL}_\rho^{\otimes n}(s,t)$. If we let $d^{2\otimes n}$ be the tensorized extension of the squared Hellinger distance d^2, Cohen and Le Pennec (2011) prove that there is a constant C_ρ such that $C_\rho d^{2\otimes n}(s,t) \leq \mathrm{JKL}_\rho^{\otimes n}(s,t)$. Moreover, if we assume that for any $m \in \mathcal{M}$ and any $s_m \in S_m$, $s_0 d\lambda \ll s_m d\lambda$, then

$$\frac{C_\rho}{2 + \ln\|s_0/s_m\|_\infty}\mathrm{KL}^{\otimes n}(s_0, s_m) \leq \mathrm{JKL}_\rho^{\otimes n}(s_0, s_m)$$

with $C_\rho = \frac{1}{\rho}\min(\frac{1-\rho}{\rho}, 1)(\ln(1+\frac{\rho}{1-\rho}) - \rho)$ (see Cohen and Le Pennec (2011)).

2.3. *Oracle inequality*

Our goal is now to define a penalty $\mathrm{pen}(m)$ which ensures that the maximum likelihood estimate in the selected model performs almost as well as the maximum likelihood estimate in the best model. More precisely, we will prove an oracle type inequality

$$\mathbb{E}\left[\mathrm{JKL}_\rho^{\otimes n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{m\in\mathcal{M}}\left(\inf_{s_m\in S_m}\mathrm{KL}^{\otimes n}(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} + \frac{\eta+\eta'}{n}\right) + \frac{C_2}{n}$$

with a $\mathrm{pen}(m)$ chosen of the same order as the variance of the corresponding single model maximum likelihood estimate.

The name oracle type inequality means that the right-hand side is a proxy for the estimation risk of the best model within the collection. The Kullback-Leibler term $\inf_{s_m\in S_m}\mathrm{KL}_\lambda^{\otimes n}(s_0, s_m)$ is a typical bias term while $\frac{pen(m)}{n}$ plays the role of the variance term. We have three sources of loss here: the constant C_1 can not be taken equal to 1, we use a different divergence on the left and on the right and $\frac{pen(m)}{n}$ is not directly related to the variance. Under a strong assumption, namely a finite upper bound on $\sup_{m\in\mathcal{M}}\sup_{s_m\in S_m}\|s_0/s_m\|_\infty$, the two divergences are *equivalent* for the conditional densities considered and thus the second issue disappears.

The first issue has a consequence as soon as s_0 does not belong to the best model, i.e. when the model is misspecified. Indeed, in that case, the corresponding modeling bias $\inf_{s_m \in S_m} \mathrm{KL}^{\otimes n}(s_0, s_m)$ may be large and the error bound does not converge to this bias when n goes to infinity but to C_1 times this bias. Proving such an oracle inequality with $C_1 = 1$ would thus be a real improvement.

To our knowledge, those two first issues have not been solved in penalized density estimation with Kullback-Leibler loss but only with L^2 norm or aggregation of a finite number of densities as in Rigollet (2012).

Concerning the third issue, if S_m is parametric, whenever $pen(m)$ can be chosen approximately proportional to the dimension $\dim(S_m)$ of the model, which will be the case in our setting, $\frac{pen(m)}{n}$ is approximately proportional to $\frac{\dim(S_m)}{n}$, which is the asymptotic variance in the parametric case. The right-hand side matches nevertheless the best known bound obtained for a single model within such a general framework.

3. Mixtures of Gaussian regressions and penalized conditional density estimation

3.1. Models of mixtures of Gaussian regressions

As explained in introduction, we are using candidate conditional densities of type

$$s_{K,\upsilon,\Sigma,w}(y|x) = \sum_{k=1}^{K} \pi_{w,k}(x)\Phi_{\upsilon_k(x),\Sigma_k}(y),$$

to estimate s_0, where $K \in \mathbb{N} \setminus \{0\}$ is the number of mixture components, $\Phi_{\upsilon,\Sigma}$ is the density of a Gaussian of mean υ and covariance matrix Σ, υ_k is a function specifying the mean given x of the k-th component while Σ_k is its covariance matrix and the mixture weights $\pi_{w,k}$ are defined from a collection of K functions $w_1, \ldots, w_K$ by a logistic scheme:

$$\pi_{w,k}(x) = \frac{e^{w_k(x)}}{\sum_{k'=1}^{K} e^{w_{k'}(x)}}.$$

We will estimate s_0 by conditional densities belonging to some model S_m defined by

$$S_m = \Bigg\{ (x,y) \mapsto \sum_{k=1}^{K} \pi_{w,k}(x)\Phi_{\upsilon_k(x),\Sigma_k}(y) \Big| (w_1, \ldots, w_K) \in W_K, \\ (\upsilon_1, \ldots, \upsilon_K) \in \Upsilon_K, (\Sigma_1, \ldots, \Sigma_K) \in V_K \Bigg\}$$

where W_K is a compact set of K-tuples of functions from $\mathcal{X}$ to $\mathbb{R}$, Υ_K a compact set of K-tuples of functions from $\mathcal{X}$ to $\mathbb{R}^p$ and V_K a compact set of K-tuples of covariance matrices of size $p \times p$. From now on, we will assume that those

sets are parametric subsets of dimensions respectively $\dim(W_K)$, $\dim(\Upsilon_K)$ and $\dim(V_K)$. The dimension $\dim(S_m)$ of the now parametric model S_m is thus nothing but $\dim(S_m) = \dim(W_K) + \dim(\Upsilon_K) + \dim(V_K)$.

Before describing more precisely those sets, we recall that S_m will be taken in a model collection $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$, where $m \in \mathcal{M}$ specifies a choice for each of those parameters. Within this collection, the number of components K will be chosen smaller than an arbitrary $K_{\max}$, which may depend on the sample size n. The sets W_K and Υ_K will be typically chosen as a tensor product of a same compact set of moderate dimension, for instance a set of polynomial of degree smaller than respectively d'_W and d'_Υ whose coefficients are smaller in absolute values than respectively T_W and T_Υ.

The structure of the set V_K depends on the *noise* model chosen: we can assume, for instance, it is common to all regressions, that they share a similar volume or diagonalization matrix or they are all different. More precisely, we decompose any covariance matrix Σ into $LPAP'$, where $L = |\Sigma|^{1/p}$ is a positive scalar corresponding to the volume, P is the matrix of eigenvectors of Σ and A the diagonal matrix of normalized eigenvalues of Σ. Let L_-, L_+ be positive values and λ_-, λ_+ real values. We define the set $\mathcal{A}(\lambda_-, \lambda_+)$ of diagonal matrices A such that $|A| = 1$ and $\forall i \in \{1, \ldots, p\}, \lambda_- \leq A_{i,i} \leq \lambda_+$. A set V_K is defined by

$$V_K = \{(L_1P_1A_1P_1', \ldots, L_KP_KA_KP_K')|\forall k, L_- \leq L_k \leq L_+, P_k \in SO(p), \\ A_k \in \mathcal{A}(\lambda_-, \lambda_+)\},$$

where $SO(p)$ is the special orthogonal group. Those sets V_K correspond to the classical covariance matrix sets described by Celeux and Govaert (1995).

3.2. A conditional density model selection theorem

The penalty should be chosen of the same order as the estimator's complexity, which depends on an intrinsic model complexity and, also, a collection complexity.

We will bound the model complexity term using the *dimension* of S_m: we prove that those two terms are roughly proportional under some structural assumptions on the sets W_K and Υ_K. To obtain this result, we rely on an entropy measure of the complexity of those sets. More precisely, for any K-tuples of functions $(s_1, \ldots, s_K)$ and $(t_1, \ldots, t_K)$, we let

$$d_{\|\sup\|_\infty}((s_1, \ldots, s_K), (t_1, \ldots, t_K)) = \sup_{x\in\mathcal{X}} \sup_{1\leq k\leq K} \|s_k(x) - t_k(x)\|_2,$$

and define the metric entropy of a set F_K, $H_{d_{\|\sup\|_\infty}}(\sigma, F_K)$, as the logarithm of the minimal number of balls of radius at most σ, in the sense of $d_{\|\sup\|_\infty}$, needed to cover F_K. We will assume that the parametric dimension D of the set considered coincides with an entropy based definition, namely there exists a

constant C such that for $\sigma \in (0, \sqrt{2}]$

$$H_{d_{\|\sup\|_\infty}}(\sigma, F_K) \leq D\left(C + \ln\frac{1}{\sigma}\right).$$

Assumption (DIM) There exist two constants C_W and C_Υ such that, for every sets W_K and Υ_K of the models S_m in the collection $\mathcal{S}$, $\forall \sigma \in (0, \sqrt{2}]$,

$$H_{d_{\|\sup\|_\infty}}(\sigma, W_K) \leq \dim(W_K)\left(C_W + \ln\frac{1}{\sigma}\right)$$

and

$$H_{d_{\|\sup\|_\infty}}(\sigma, \Upsilon_K) \leq \dim(\Upsilon_K)\left(C_\Upsilon + \ln\frac{1}{\sigma}\right)$$

Note that one can extend our result to any compact sets for which those assumptions hold for *dimensions* that could be different from the usual ones.

The complexity of the estimator depends also on the complexity of the collection. That is why one needs further to control the complexity of the collection as a whole through a coding type (Kraft) assumption (Barron et al., 2008).

Assumption (K) There is a family $(x_m)_{m\in\mathcal{M}}$ of non-negative numbers and a real number Ξ such that

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

We can now state our main result, a weak oracle inequality:

Theorem 1. *For any collection of mixtures of Gaussian regressions model $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ satisfying (K) and (DIM), there is a constant C such that for any $\rho \in (0,1)$ and any $C_1 > 1$, there is a constant κ_0 depending only on ρ and C_1 such that, as soon as for every index $m \in \mathcal{M}$, $pen(m) = \kappa((C + \ln n)\dim(S_m) + x_m)$ with $\kappa > \kappa_0$, the penalized likelihood estimate $\widehat{s}_{\widehat{m}}$ with $\widehat{m}$ such that*

$$\sum_{i=1}^n -\ln(\widehat{s}_{\widehat{m}}(Y_i|X_i)) + pen(\widehat{m}) \leq \inf_{m\in\mathcal{M}}\left(\sum_{i=1}^n -\ln(\widehat{s}_m(Y_i|X_i)) + pen(m)\right) + \eta'$$

satisfies

$$\mathbb{E}\left[\mathrm{JKL}_\rho^{\otimes n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{m\in\mathcal{M}}\left(\inf_{s_m\in S_m} \mathrm{KL}^{\otimes n}(s_0, s_m) + \frac{pen(m)}{n} + \frac{\kappa_0\Xi + \eta + \eta'}{n}\right).$$

Remind that under the assumption that $\sup_{m\in\mathcal{M}}\sup_{s_m\in S_m}\|s_0/s_m\|_\infty$ is finite, $\mathrm{JKL}_\rho^{\otimes n}$ can be replaced by $\mathrm{KL}^{\otimes n}$ up to a multiplication by a constant depending on ρ and the upper bound. Note that this strong assumption is nevertheless satisfied if we assume that $\mathcal{X}$ is compact, s_0 is compactly supported, the regression functions are uniformly bounded and there is a uniform lower bound on the eigenvalues of the covariance matrices.

As shown in the proof, in the previous theorem, the assumption on $\text{pen}(m)$ could be replaced by the milder one

$$\text{pen}(m) \geq \kappa \left(2\dim(S_m)C^2 + \dim(S_m) \left(\ln \frac{n}{C^2 \dim(S_m)} \right)_+ + x_m \right).$$

It may be noticed that if $(x_m)_m$ satisfies Assumption (K), then for any permutation τ $(x_{\tau(m)})_m$ satisfies this assumption too. In practice, x_m should be chosen such that $\frac{2\kappa x_m}{\text{pen}(m)}$ is as small as possible so that the penalty can be seen as proportional to the two first terms. Notice that the constant C only depends on the model collection parameters, in particular on the maximal number of components $K_{\max}$. As often in model selection, the collection may depends on the sample size n. If the constant C grows no faster than $\ln(n)$, the penalty shape can be kept intact and a similar result holds uniformly in n up to a slightly larger κ_0. In particular, the apparent dependency in $K_{\max}$ is not an issue: $K_{\max}$ only appears in C through a logarithmic term and $K_{\max}$ should be taken smaller than n for identifiability issues. Finally, it should be noted that the $\ln n$ term in the penalty of Theorem 1 may not be necessary as hinted by a result of Gassiat and van Handel (2014) for one dimensional mixtures of Gaussian distribution with the same variance.

3.3. *Linear combination of bounded functions for the means and the weights*

We postpone the proof of this theorem to the Appendix and focus on Assumption (DIM). This assumption is easily verified when the function sets W_K and Υ_K are defined as the linear combination of a finite set of bounded functions whose coefficients belong to a compact set. This quite general setting includes the polynomial basis when the covariable are bounded, the Fourier basis on an interval as well as suitably renormalized wavelet dictionaries. Let d_W and d_Υ be two positive integers, let $(\psi_{W,i})_{1\leq i\leq d_W}$ and $(\psi_{\Upsilon,i})_{1\leq i\leq d_\Upsilon}$ two collections of functions bounded functions from $\mathcal{X} \to [-1,1]$ and define

$$W = \left\{ w : [0,1]^d \to \mathbb{R} \middle| w(x) = \sum_{i=0}^{d_W} \alpha_i \psi_{W,i}(x) \text{ and } \|\alpha\|_\infty \leq T_W \right\}$$

$$\Upsilon = \left\{ \upsilon : [0,1]^d \to \mathbb{R}^p \middle| \forall j \in \{1,\dots,p\}, \right.$$
$$\left. \forall x, \upsilon_j(x) = \sum_{i=0}^{d_\Upsilon} \alpha_i^{(j)} \psi_{\Upsilon,i}(x) \text{ and } \|\alpha\|_\infty \leq T_\Upsilon \right\}$$

where the (j) in $\alpha_r^{(j)}$ is a notation to indicate the link with υ_j. We will be interested in tensorial construction from those sets, namely $W_K = \{0\} \times W^{K-1}$ and $\Upsilon_K = \Upsilon^K$, for which we prove in Appendix that

Lemma 1. *W_K and Υ_K satisfy Assumption (DIM), with $C_W = \ln(\sqrt{2}+T_W d_W)$ and $C_\Upsilon = \ln(\sqrt{2}+\sqrt{p}d_\Upsilon T_\Upsilon)$, not depending on K.*

Note that in this general case, only the functions $\psi_{W,i}$ and $\psi_{\Upsilon,i}$ need to be bounded and not the covariate X itself.

For sake of simplicity, we focus on the bounded case and assume $\mathcal{X} = [0,1]^d$. In that case, we can use a polynomial modeling: $\psi_{W,i}$ and $\psi_{\Upsilon,i}$ can be chosen as monomials $x^r = x_1^{r_1} \dots x_d^{r_d}$. If we let d'_W and d'_Υ be two maximum (non negative) degrees for those monomials and define the sets of W_K and Υ_K accordingly, the previous Lemma becomes

Lemma 2. *W_K and Υ_K satisfy Assumption (DIM), with $C_W = \ln(\sqrt{2} + T_W\binom{d'_W+d}{d})$ and $C_\Upsilon = \ln(\sqrt{2}+\sqrt{p}\binom{d'_\Upsilon+d}{d}T_\Upsilon)$, not depending on K.*

To apply Theorem 1, it remains to describe a collection $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ and a suitable choice for $(x_m)_{m\in\mathcal{M}}$. Assume, for instance, that the models in our collection are defined by an arbitrary maximal number of components $K_{\max}$, a common free structure for the covariance matrix K-tuple and a common maximal degree for the sets W_K and Υ_K. Then one can verify that $\dim(S_m) = (K-1+Kp)\binom{d'_W+d}{d} + Kp\frac{p+1}{2}$ and that the weight family $(x_m = K)_{m\in\mathcal{M}}$ satisfy Assumption (K) with $\Xi \le 1/(e-1)$. Theorem 1 yields then an oracle inequality with $\mathrm{pen}(m) = \kappa((C+\ln(n))\dim(S_m) + x_m)$. Note that as $x_m \ll (C+\ln(n))\dim(S_m)$, one can obtain a similar oracle inequality with $\mathrm{pen}(m) = \kappa(C+\ln(n))\dim(S_m)$ for a slightly larger κ. Finally, as explained in the proof, choosing a covariance structure from the finite collection of Celeux and Govaert (1995) or choosing the maximal degree for the sets W_K and Υ_K among a finite family can be obtained with the same penalty but with a larger constant Ξ in Assumption (K).

4. Numerical scheme and numerical experiment

We illustrate our theoretical result in a setting similar to the one considered by Chamroukhi et al. (2010) and on two real data sets. We observe n pairs (X_i, Y_i) with X_i in a compact interval, namely $[0,1]$ for simulated data and respectively $[0,5]$ and $[0,17]$ for the first and second real data set, and $Y_i \in \mathbb{R}$ and look for the best estimate of the conditional density $s_0(y|x)$ that can be written

$$s_{K,\upsilon,\Sigma,w}(y|x) = \sum_{k=1}^{K} \pi_{w,k}(x)\Phi_{\upsilon_k(x),\Sigma_k}(y),$$

with $w \in W_K$ and $\upsilon \in \Upsilon_K$. We consider the simple case where W_K and Υ_K contain linear functions. We do not impose any structure on the covariance matrices. Our aim is to estimate the *best* number of components K as well as the model parameters. As described with more details later, we use an EM type algorithm to estimate the model parameters for each K and select one using the penalized approach described previously.

4.1. The procedure

As often in model selection approach, the first step is to compute the maximum likelihood estimate for each number of components K. To this purpose, we use a numerical scheme based on the EM algorithm (Dempster et al., 1977) similar to the one used by Chamroukhi et al. (2010). The only difference with a classical EM is in the Maximization step since there is no closed formula for the weights optimization. We use instead a Newton type algorithm. Note that we only perform a few Newton steps (5 at most were enough in our experiments) and ensure that the likelihood does not decrease. We have noticed that there is no need to fully optimize at each step: we did not observe a better convergence and the algorithmic cost is high. We denote from now on this algorithm *Newton-EM*. Notice that the lower bound on the variance required in our theorem appears to be necessary in practice. It avoids the spurious local maximizer issue of EM algorithm, in which a class degenerates to a minimal number of points allowing a perfect Gaussian regression fit. We use a lower bound shape of $\frac{C}{n}$. Biernacki and Castellan (2011) provide a precise data-driven bound for mixture of Gaussian regressions: $\frac{\min_{1\leq i<j\leq n}(Y_i-Y_j)^2}{2\chi^2_{n-2K+1}((1-\alpha)^{1/K})}$, with χ^2_{n-2K+1} the chi-squared quantile function, which is of the same order as $\frac{1}{n}$ in our case. In practice, the constant 10 gave good results for the simulated data.

An even more important issue with EM algorithms is initialization, since the local minimizer obtained depends heavily on it. We observe that, while the weights w do not require a special care and can be simply initialized uniformly equal to 0, the means require much more attention in order to obtain a good minimizer. We propose an initialization strategy based on short runs of *Newton-EM* with random initialization.

We draw randomly K lines, each defined as the line going through two points (X_i, Y_i) drawn at random among the observations. We perform then a K-means clustering using the distance along the Y axis. Our *Newton-EM* algorithm is initialized by the regression parameters as well as the empirical variance on each of the K clusters. We perform then 3 steps of our minimization algorithm and keep among 50 trials the one with the largest likelihood. This winner is used as the initialization of a final *Newton-EM* algorithm using 10 steps.

We consider two other strategies: a *naive* one in which the initial lines chosen at random and a common variance are used directly to initialize the *Newton-EM* algorithm and a *clever* one in which observations are first normalized in order to have a similar variance along both the X and the Y axis, a K-means on both X and Y with 5 times the number of components is then performed and the initial lines are drawn among the regression lines of the resulting cluster containing more than 2 points.

The complexity of those procedures differs and as stressed by Celeux and Govaert (1995) the fairest comparison is to perform them for the same amount of time (5 seconds, 30 seconds, 1 minute...) and compare the obtained likelihoods. The difference between the 3 strategies is not dramatic: they yield very similar likelihoods. We nevertheless observe that the *naive* strategy has an important

dispersion and fails sometime to give a satisfactory answer. Comparison between the *clever* strategy and the regular one is more complex since the difference is much smaller. Following Celeux and Govaert (1995), we have chosen the regular one which corresponds to more random initializations and thus may explore more local maxima.

Once the parameters' estimates have been computed for each K, we select the model that minimizes

$$\sum_{i=1}^{n} -\ln(\widehat{s}_m(Y_i|X_i)) + \text{pen}(m)$$

with $\text{pen}(m) = \kappa \dim(S_m)$. Note that our theorem ensures that there exists a κ large enough for which the estimate has good properties, but does not give an explicit value for κ. In practice, κ has to be chosen. The two most classical choices are $\kappa = 1$ and $\kappa = \frac{\ln n}{2}$ which correspond to the AIC and BIC approach, motivated by asymptotic arguments. We have used here the slope heuristic proposed by Birgé and Massart (2007) and described for instance in Baudry et al. (2011). This heuristic comes with two possible criterions: the jump criterion and the slope criterion. The first one consists in representing the dimension of the selected model according to κ (Fig. 3), and finding $\hat{\kappa}$ such that if $\kappa < \hat{\kappa}$, the dimension of the selected model is large, and reasonable otherwise. The slope heuristic prescribes then the use of $\kappa = 2\hat{\kappa}$. In the second one, one computes the *asymptotic* slope of the log-likelihood drawn according to the model dimension, and penalizes the log-likelihood by twice the slope times the model dimension. With our simulated data sets, we are in the not so common situation in which the jump is strong enough so that the first heuristic can be used.

4.2. Simulated data sets

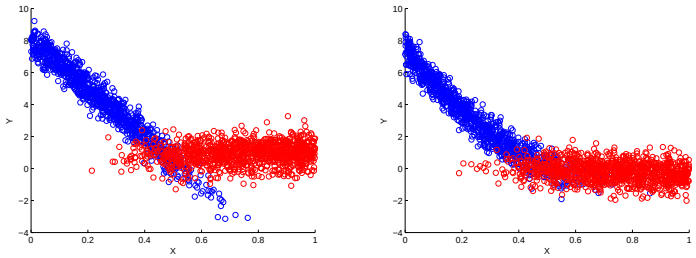
The previous procedure has been applied to two simulated data sets: one in which true conditional density belongs to one of our models, a *well-specified* case, and one in which this is not true, a *misspecified* case. In the first situation, we expect to perform almost as well as the maximum likelihood estimation in the true model. In the second situation, we expect our algorithm to automatically balance the model bias and its variance. More precisely, we let

$$s_0(y|x) = \frac{1}{1+\exp(15x-7)}\Phi_{-15x+8,0.3}(y) + \frac{\exp(15x-7)}{1+\exp(15x-7)}\Phi_{0.4x+0.6,0.4}(y)$$

in the first example, denoted example WS, and

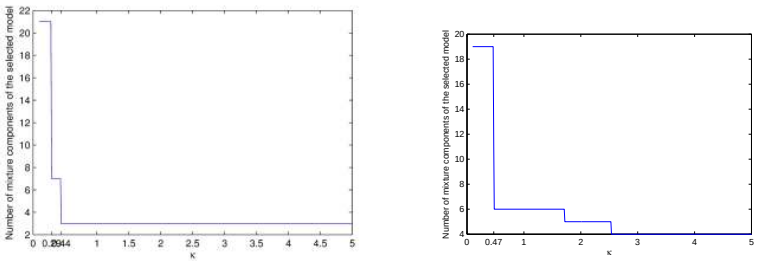
$$s_0(y|x) = \frac{1}{1+\exp(15x-7)}\Phi_{15x^2-22x+7.4,0.3}(y)+\frac{\exp(15x-7)}{1+\exp(15x-7)}\Phi_{-0.4x^2,0.4}(y)$$

in the second example, denoted example MS. For both experiments, we let X be uniformly distributed over $[0,1]$. Figure 2 shows a typical realization.



(a) 2 000 data points of example WS (b) 2 000 data points of example MS

FIG 2. *Typical realizations.*



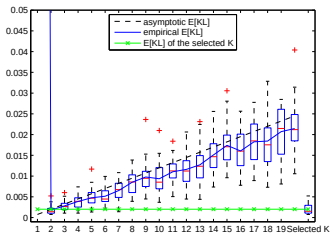
(a) Example WS with 2 000 points (b) Example MS with 2 000 points

FIG 3. *Slope heuristic: plot of the selected model dimension with respect to the penalty coefficient* κ. *In both examples,* $\widehat{\kappa}$ *is of order* $1/2$.

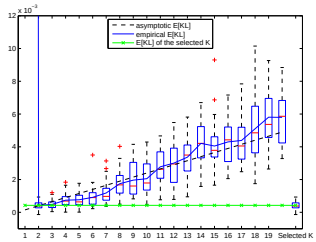
In both examples, we have noticed that the sample's size had no significant influence on the choice of κ, and that very often 1 was in the range of possible values indicated by the jump criterion of the slope heuristic. According to this observation, we have chosen in both examples $\kappa = 1$.

We measure performances in term of tensorized Kullback-Leibler divergence. Since there is no known formula for tensorized Kullback-Leibler divergence in the case of Gaussian mixtures, and since we know the true density, we evaluate the divergence using Monte Carlo method. The variability of this randomized approximation has been verified to be negligible in practice.

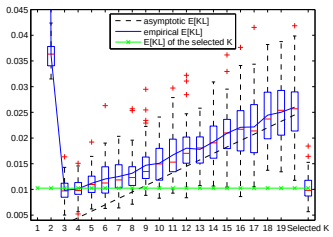
For several numbers of mixture components and for the selected K, we draw in Figure 4 the box plots and the mean of tensorized Kullback-Leibler divergence over 55 trials. The first observation is that the mean of tensorized Kullback-Leibler divergence between the penalized estimator $\hat{s}_{\hat{K}}$ and s_0 is smaller than the mean of tensorized Kullback-Leibler divergence between $\hat{s}_K$ and s_0 over $K \in \{1, \ldots, 20\}$. This is in line with the oracle type inequality of Theorem 1. Our numerical results hint that our theoretical analysis may be pessimistic. A close inspection shows that the bias-variance trade-off differs between the two examples. Indeed, since in the first one the true density belongs to the model, the best choice is $K = 2$ even for large n. As shown on the histogram of Figure 5,



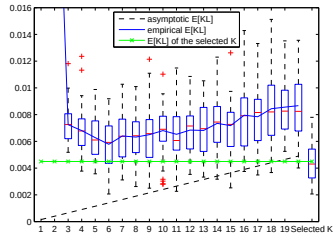
(a) Example WS with 2 000 data points



(b) Example WS with 10 000 data points



(c) Example MS with 2 000 data points

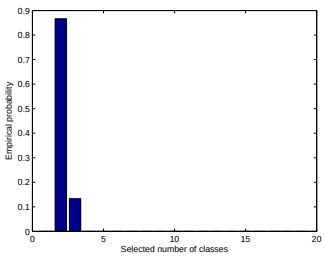


(d) Example MS with 10 000 data points

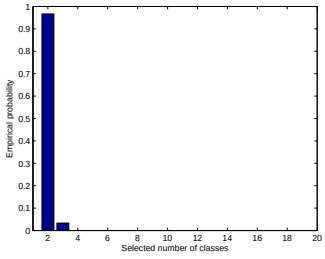
FIG 4. *Box-plot of the Kullback-Leibler divergence according to the number of mixture components. On each graph, the right-most box-plot shows this Kullback-Leibler divergence for the penalized estimator* $\hat{s}_{\widehat{K}}$.

this is almost always the model chosen by our algorithm. Observe also that the mean of Kullback-Leibler divergence seems to behave like $\frac{\dim(S_m)}{2n}$ (shown by a dotted line). This is indeed the expected behavior when the true model belongs to a nested collection and corresponds to the classical AIC heuristic. In the second example, the misspecified one, the true model does not belong to the collection. The best choice for K should thus balance a model approximation error term and a variance one. We observe in Figure 5 such a behavior: the larger n the more complex the model and thus K. Note that the slope of the mean error seems also to grow like $\frac{\dim(S_m)}{2n}$ even though there is no theoretical guarantee of such a behavior.

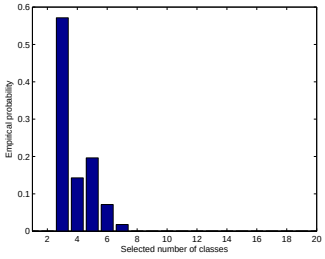
Figure 6 shows the error decay when the sample size n grows. As expected in the well-specified case, example W, we observe the decay in t/n predicted in the theory, with t some constant. The rate in the second case appears to be slower. Indeed, as the true conditional density does not belong to any model, the selected models are more and more complex when n grows which slows the error decay. In our theoretical analysis, this can already be seen in the decay of the *variance* term of the oracle inequality. Indeed, if we let $m_0(n)$ be the optimal oracle model, the one minimizing the right-hand side of the oracle inequality, the variance term is of order $\frac{\dim(S_{m_0(n)})}{n}$ which is larger than $\frac{1}{n}$ as soon as



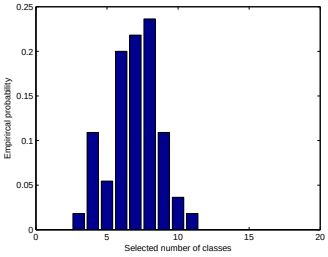
(a) Example WS with 2 000 data points



(b) Example WS with 10 000 data points

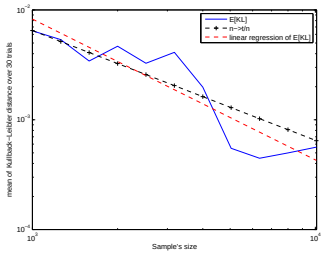


(c) Example MS with 2 000 data points

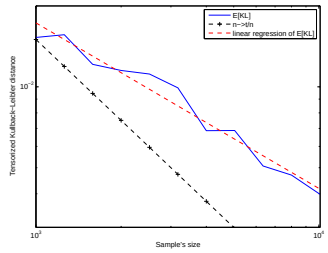


(d) Example MS with 10 000 data points

FIG 5. *Histograms of the selected K.*



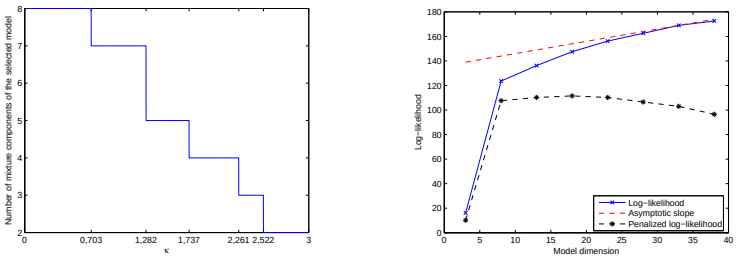
(a) Example WS. The slope of the free regression line is $\simeq -1,3$



(b) Example MS. The slope of the regression line is $\simeq -0,6$

FIG 6. *Kullback-Leibler divergence between the true density and the computed density using* $(X_i, Y_i)_{i \leq N}$ *with respect to the sample size, represented in a log-log scale. For each graph, we added a free linear least-square regression and one with slope* -1 *to stress the two different behavior.*

$\dim(S_{m_0(n)}) \to +\infty$. It is well known that the decay depends on the regularity of the true conditional density. Providing a minimax analysis of the proposed estimator, as have done Maugis and Michel (2012), would be interesting but is beyond the scope of this paper.



(a) Jump criterion (b) Slope criterion

FIG 7. *Slope heuristic for the ethanol data set.*

4.3. Ethanol data set

We explain now with more details the result of Figure 1 for the 88 data point Ethanol data set of Brinkman (1981). Young (2014) proposes to estimate the density of the equivalence ratio R conditioned to the concentration in NO and to use this conditional density to do a clustering of the data set. In our framework, this amounts to estimate the conditional density by

$$\sum_{k=1}^{\widehat{K}} \pi_{\widehat{w}_k(NO)} \Phi_{\widehat{\upsilon_k}(NO),\widehat{\Sigma_k}}(R)$$

with our proposed penalized estimator and to use the classical maximum likelihood approach that associates (NO, R) to the class

$$\arg \max_{1\leq k\leq \widehat{K}} \pi_{\widehat{w}_k(NO)} \Phi_{\widehat{\upsilon_k}(NO),\widehat{\Sigma_k}}(R)$$

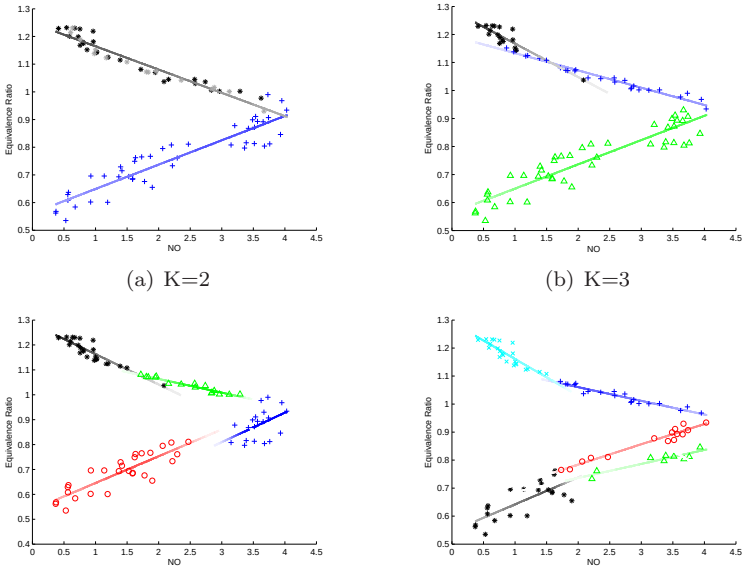
to perform the clustering.

An important parameter of the method is the lower bound of the variance used in the estimation for a given number of class. This is required to avoid spurious maximizers of the likelihood. Here, the value 10^{-4} chosen *by hand* yields satisfactory results.

Since we only have 88 points and roughly 5 parameters per class, the random initialization may yield classes with too few points to have a good estimation. We have slightly modified our K-means procedure in order to ensure than at least 10 points are assigned to each class. In that case, we have verified that the estimated parameters of the conditional density were very stable.

Note that with this strategy, no more than 8 classes can be considered. This prevents the use of the jump criterion to calibrate the penalty because the *big* jump is hard to define. We use instead the slope heuristic. Figure 7 shows that this slope is of order 1 and thus the slope heuristic prescribes a penalty of $2\dim(S_K)$, providing an estimate with 4 components.

It is worth pointing out that the maximum of the penalized likelihood is not sharp, just like in the example MS of simulated data (see figure 5). Indeed, it is quite unlikely that the true density belongs to our model collection. So, there



(a) K=2 (b) K=3

(c) K=4 (d) K=5

FIG 8. *Clustering of NO data set into K classes. The strength of the color of the regression lines corresponds to the mixture proportion.*

may be an uncertainty on the selected number of components between 4, 3 and 5. Note that AIC penalization would have lead to 7 classes while BIC would also have lead to 4 classes. Our estimated penalty is nevertheless in the middle of the zone corresponding to 4 while BIC is nearby the boundary with 3 and thus we expect this choice to be more stable. In Figure 1(b) of the introduction we have shown only this clustering with 4 classes. Figure 8 shows that the choices of 3 or 5 may make sense, even though the choice 5 may seem slightly too complex. A common feature among all those clusterings is the change of slope in the topmost part around 1.7. This phenomena is also visible in Young (2014) in which an explicit change point model is used, ours is only implicit and thus more versatile

To complete our study, in Figure 9, we have considered the more natural regression of NO with respect to the equivalence ratio that has not been studied by Young (2014). Using the same methodology, we have recovered also 4 clusters corresponding to a soft partitioning of the equivalence ratio value. Note that this clustering, which is easily interpretable, is very similar to the one obtained with the previous parameterization.

4.4. ChIP-chip data set

We considere here a second real data set: a Chromatin immunoprecipitation (ChIP) on chip genomic data set. Chromatin immunoprecipitation (ChIP) is

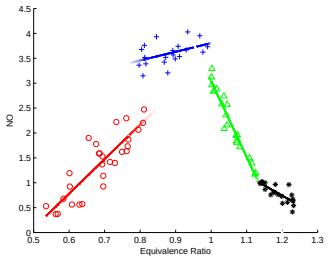
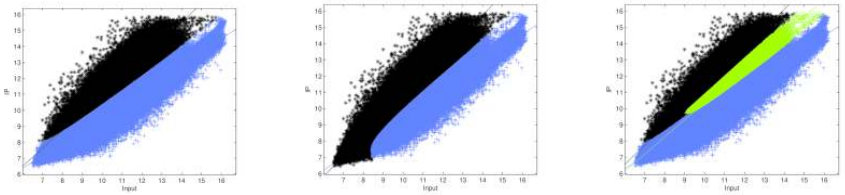


FIG 9. *Clustering of NO data set into 4 classes, considering the regression of NO with respect to the equivalence ratio.*

a procedure used to investigate proteins associated with DNA. The data set considered is the one used by Martin-Magniette et al. (2008). In this experiment, two variables are studied: DNA fragments crosslinked to a protein of interest (IP) and genomic DNA (Input). Martin-Magniette et al. (2008) model the density of log-IP conditioned to log-Input by a mixture of two Gaussian regressions with the same variance. One component corresponds to an enriched one, in which there is more proteins than expected, and the other to a normal one. They use classical proportions that do not depends on the Input. The parameters are estimated using the EM algorithm initialized by values derived from a Principal Component Analysis of the whole data set. The best model between one and two components is selected according to the BIC criterion. For the histone modification in Arabidopsis thaliana data set, they select a two components model similar to the one obtained with logistic weights (Figure 10).

We have first compare the constant proportions model with $K = 2$ to the one proposed in their conclusion in which the proportions depend on the Input. We have used our affine logistic weight model and observed that this model greatly improves the log-likelihood. The dimension of this new model is 8 while the dimension of the original model is 7 so that the log-likelihood increase does not seem to be due to overfitting. We have also compare our solution to the one obtained with a constant weight with $K = 3$ model of dimension 11. The BIC criterion selects the $K = 2$ with affine weight solution.

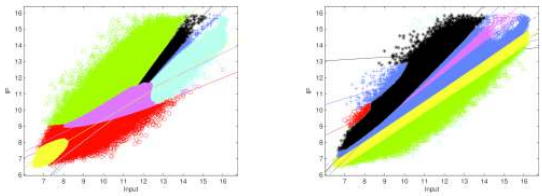
We have then tested more complex models with K up to 20 with a penalty obtained with the slope heuristic. The models chosen are quite complex ($K = 10$ for constant proportions models and $K = 7$ for affine logistic weight models, the later being the overall winner). Although they better explain the data from the statistical point of view, those models become hard to interpret from the biological point of view. We think this is due to the too simple affine models used. Although no conceptual difficulties occur by using more complex function familie (or going to the multivariate setting), the *curse of dimensionality* makes everything more complicated in practice. In particular, initialization becomes harder and harder as the dimension grows and requires probably a more clever



(a) K=2, constant proportions, dimension= 7

(b) K=2, affine logistic weights, dimension= 8

(c) K=3, constant proportions, dimension= 9



(d) K=7, affine logistic weights, dimension= 33

(e) K=10, constant proportions, dimension= 39

FIG 10. *Clustering of ChIP-chip data set into K classes.*

treatment than the one proposed here. In the spirit of Cohen and Le Pennec (2013), we are currently working on a first extension: a numerical algorithm for a bivariate piecewise linear logistic weights model applied to hyperspectral image segmentation.

5. Discussion

We have studied a penalized maximum likelihood estimate for mixtures of Gaussian regressions with logistic weights. Our main contribution is the proof that a penalty proportional, up to a logarithmic factor of the sample size, to the dimension of the model is sufficient to obtain a non asymptotic theoretical control on the estimator loss. This result is illustrated in the simple univariate case in which both the means and the logistic weights are linear. We study a toy model which exhibits the behavior predicted by our theoretical analysis and proposes two simple applications of our methodology. We hope that our contribution helps to popularize those mixtures of Gaussian regressions by giving a theoretical foundation for model selection technique in this area and showing some possible interesting uses even for simple models.

Besides some important theoretical issues on the loss used and the tightness of the bounds, the major future challenge is the extension of the numerical scheme to more complex cases than univariate linear models.

Appendix A: A general conditional density model selection theorem

We summarize in this section the main result of Cohen and Le Pennec (2011) that will be our main tool to obtain the previous oracle inequality.

To any model S_m, a set of conditional densities, we associate a complexity defined in term of a specific entropy, the bracketing entropy with respect to the square root of the tensorized square of the Hellinger distance $d^{2\otimes n}$. Recall that a bracket $[t^-, t^+]$ is a pair of real functions such that $\forall (x,y) \in \mathcal{X}\times\mathcal{Y}, t^-(x,y) \leq t^+(x,y)$ and a function s is said to belong to the bracket $[t^-,t^+]$ if $\forall (x,y) \in \mathcal{X}\times\mathcal{Y}, t^-(x,y) \leq s(x,y) \leq t^+(x,y)$. The bracketing entropy $H_{[],d}(\delta, S)$ of a set S is defined as the logarithm of the minimal number $N_{[],d}(\delta,S)$ of brackets $[t^-,t^+]$ covering S, such that $d(t^-,t^+) \leq \delta$. The main assumption on models is a property that should satisfies the bracketing entropy:

Assumption (H) For every model S_m in the collection $\mathcal{S}$, there is a non-decreasing function ϕ_m such that $\delta \mapsto \frac{1}{\delta}\phi_m(\delta)$ is non-increasing on $(0,+\infty)$ and for every $\sigma \in \mathbb{R}^+$,

$$\int_0^\sigma \sqrt{H_{[.],d^{\otimes n}}(\delta, S_m)}d\delta \leq \phi_m(\sigma).$$

Such an integral is ofter called a Dudley type integral of these bracketing entropies and is commonly used in empirical process theory (Van der Vaart and Wellner, 1996). The complexity of S_m is then defined as $n\sigma_m^2$ where σ_m is the unique square root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$.

For technical reason, a separability assumption, always satisfied in the setting of this paper, is also required. It is a mild condition, classical in empirical process theory (see for instance Van der Vaart and Wellner (1996)).

Assumption (Sep) For every model S_m in the collection $\mathcal{S}$, there exists some countable subset S'_m of S_m and a set $\mathcal{Y}'_m$ with $\lambda(\mathcal{Y}\backslash\mathcal{Y}'_m) = 0$ such that for every t in S_m, there exists some sequence $(t_k)_{k\geq 1}$ of elements of S'_m such that for every $x \in \mathcal{X}$ and every $y \in \mathcal{Y}'_m, \ln(t_k(y|x)) \xrightarrow[k\to+\infty]{} \ln(t(y|x))$.

The main result of Cohen and Le Pennec (2011) is a condition on the penalty $\text{pen}(m)$ which ensures an oracle type inequality:

Theorem 2. *Assume we observe (X_i, Y_i) with unknown conditional density s_0. Let $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ an at most countable conditional density model collection. Assume Assumptions (H), (Sep) and (K) hold. Let $\widehat{s}_m$ be a η minimizer of the negative log-likelihood in S_m*

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

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

$$pen(m) \geq \kappa(n\sigma_m^2 + x_m)$$

with $\kappa > \kappa_0$ and σ_m the unique square root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$, the penalized likelihood estimate $\widehat{s}_{\widehat{m}}$ with $\widehat{m}$ such that

$$\sum_{i=1}^n -\ln(\widehat{s}_{\widehat{m}}(Y_i|X_i)) + pen(\widehat{m}) \leq \inf_{m\in\mathcal{M}}\left(\sum_{i=1}^n -\ln(\widehat{s}_m(Y_i|X_i)) + pen(m)\right) + \eta'$$

satisfies

$$\mathbb{E}\left[\mathrm{JKL}_\rho^{\otimes n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{m\in\mathcal{M}} \left(\inf_{s_m \in S_m} \mathrm{KL}^{\otimes n}(s_0, s_m) + \frac{pen(m)}{n} \right) + C_1 \frac{\kappa_0 \Xi + \eta + \eta'}{n}.$$

In the next section, we show how to apply this result in our mixture of Gaussian regressions setting and prove that the penalty can be chosen roughly proportional to the intrinsic dimension of the model, and thus of the order of the variance.

Appendix B: Proofs

In Appendix B.1, we give a proof of Theorem 1 relying on several bracketing entropy controls proved in Appendix B.2.

B.1. Proof of Theorem 1

We will show that Assumption (DIM) ensures that for all $\delta \in (0, \sqrt{2}]$, $H_{[.],d^{\otimes n}}(\delta, S_m) \leq \dim(S_m)(\mathfrak{C} + \ln(\frac{1}{\delta}))$ with a common $\mathfrak{C}$.

We show in Appendix that if

Assumption (DIM) There exist two constants C_W and C_Υ such that, for every model S_m in the collection $\mathcal{S}$,

$$H_{d_{\|\sup\|_\infty}}(\sigma, W_K) \leq \dim(W_K)\left(C_W + \ln\frac{1}{\sigma}\right)$$

and

$$H_{d_{\|\sup\|_\infty}}(\sigma, \Upsilon_K) \leq \dim(\Upsilon_K)\left(C_\Upsilon + \ln\frac{1}{\sigma}\right)$$

then, if $n \geq 1$, the complexity of the corresponding model S_m satisfies for any $\delta \in (0, \sqrt{2}]$,

$$H_{[.],d^{\otimes n}}(\delta, S_m) \leq \dim(S_m)\left(\mathfrak{C} + \ln\left(\frac{1}{\delta}\right)\right)$$

with $\dim(S_m) = \dim(W_K) + \dim(\Upsilon_K) + \dim(V_K)$ and $\mathfrak{C}$ that depends only on the constants defining V_K and the constants C_W and C_Υ.

If this happens, Proposition 1 yields the results.

Proposition 1. *If for any $\delta \in (0, \sqrt{2}]$, $H_{[.],d^{\otimes n}}(\delta, S_m) \leq \dim(S_m)(C_m + \ln(\frac{1}{\delta}))$, then the function $\phi_m(\sigma) = \sigma\sqrt{\dim(S_m)}(\sqrt{C_m} + \sqrt{\pi} + \sqrt{\ln(\frac{1}{\min(\sigma,1)})})$ satisfies Assumption (H). Furthermore, the unique square root σ_m of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$ satisfies*

$$n\sigma_m^2 \leq \dim(S_m)\left(2(\sqrt{C_m} + \sqrt{\pi})^2 + \left(\ln\frac{n}{(\sqrt{C_m} + \sqrt{\pi})^2 \dim(S_m)}\right)_+\right).$$

In other words, if we can control models' bracketing entropy with a uniform constant $\mathfrak{C}$, we get a suitable bound on the complexity. This result will be obtain by first decomposing the entropy term between the weights and the Gaussian components. Therefore we use the following distance over conditional densities:

$$\sup_x d_y(s,t) = \sup_{x\in\mathcal{X}} \left(\int_y \left(\sqrt{s(y|x)} - \sqrt{t(y|x)} \right)^2 dy \right)^{\frac{1}{2}}.$$

Notice that $d^{2\otimes n}(s,t) \leq \sup_x d_y^2(s,t)$.

For all weights π and π', we define

$$\sup_x d_k(\pi,\pi') = \sup_{x\in\mathcal{X}} \left(\sum_{k=1}^{K} \left(\sqrt{\pi_k(x)} - \sqrt{\pi'_k(x)} \right)^2 \right)^{\frac{1}{2}}.$$

Finally, for all densities s and t over $\mathcal{Y}$, depending on x, we set

$$\begin{aligned}\sup_x \max_k d_y(s,t) &= \sup_{x\in\mathcal{X}} \max_{1\leq k\leq K} d_y(s_k(x,.),t_k(x,.)) \\ &= \sup_{x\in\mathcal{X}} \max_{1\leq k\leq K} \left(\int_y \left(\sqrt{s_k(x,y)} - \sqrt{t_k(x,y)} \right)^2 dy \right)^{\frac{1}{2}}.\end{aligned}$$

Lemma 3. *Let* $\mathcal{P} = \{(\pi_{w,k})_{1\leq k\leq K} \big| w \in W_K,$ *and* $\forall (k,x), \pi_{w,k}(x) = \frac{e^{w_k(x)}}{\sum_{l=1}^{K} e^{w_l(x)}}\}$ *and* $\mathcal{G} = \{(\Phi_{\upsilon_k,\Sigma_k})_{1\leq k\leq K} | \upsilon \in \Upsilon_K, \Sigma \in V_K\}$. *Then for all* δ *in (0,*$\sqrt{2}$*], for all* m *in* $\mathcal{M}$*,*

$$H_{[.],\sup_x d_y}(\delta, S_m) \leq H_{[.],\sup_x d_k}\left(\frac{\delta}{5}, \mathcal{P}\right) + H_{[.],\sup_x \max_k d_y}\left(\frac{\delta}{5}, \mathcal{G}\right).$$

One can then relate the bracketing entropy of $\mathcal{P}$ to the entropy of W_K

Lemma 4. *For all* $\delta \in (0, \sqrt{2}]$*,*

$$H_{[.],\sup_x d_k}\left(\frac{\delta}{5}, \mathcal{P}\right) \leq H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{20\sqrt{K}}, W_K\right)$$

Since $\mathcal{P}$ is a set of weights, $\frac{3\sqrt{3}\delta}{20\sqrt{K}}$ could be replaced by $\frac{3\sqrt{3}\delta}{20\sqrt{K-1}}$ with an identifiability condition. For example, $W'_K = \{(0, w_2 - w_1, \ldots, w_K - w_1) | w \in W_K\}$ can be covered using brackets of null size on the first coordinate, lowering squared Hellinger distance between the brackets' bounds to a sum of $K-1$ terms. Therefore, $H_{[.],\sup_x d_k}(\frac{\delta}{5}, \mathcal{P}) \leq H_{d_{\|\sup\|_\infty}}(\frac{3\sqrt{3}\delta}{20\sqrt{K-1}}, W'_K)$.

Since we have assumed that $\exists C_W$ s.t $\forall \delta \in (0, \sqrt{2}]$,

$$H_{d_{\|\sup\|_\infty}}(\delta, W_K) \leq \dim(W_K)\left(C_W + \ln\left(\frac{1}{\delta}\right)\right)$$

Then

$$H_{[.],\sup_x d_k}\left(\frac{\delta}{5},\mathcal{P}\right)\leq \dim(W_K)\left(C_W+\ln\left(\frac{20\sqrt{K}}{3\sqrt{3}\delta}\right)\right)$$

To tackle the Gaussian regression part, we rely heavily on the following proposition,

Proposition 2. *Let* $\kappa\geq\frac{17}{29}$, $\gamma_\kappa=\frac{25(\kappa-\frac{1}{2})}{49(1+\frac{2\kappa}{5})}$. *For any* $0<\delta\leq\sqrt{2}$ *and any* $\delta_\Sigma\leq\frac{1}{5\sqrt{\kappa^2\cosh(\frac{2\kappa}{5})+\frac{1}{2}}}\frac{\delta}{p}$, $(\upsilon,L,A,P)\in\Upsilon\times[L_-,L_+]\times\mathcal{A}(\lambda_-,\lambda_+)\times SO(p)$ *and* $(\tilde{\upsilon},\tilde{L},\tilde{A},\tilde{P})\in\Upsilon\times[L_-,L_+]\times\mathcal{A}(\lambda_-,+\infty)\times SO(p),\Sigma=LPAP'$ *and* $\tilde{\Sigma}=\tilde{L}\tilde{P}\tilde{A}\tilde{P}'$, *assume that* $t^-(x,y)=(1+\kappa\delta_\Sigma)^{-p}\Phi_{\tilde{\upsilon}(x),(1+\delta_\Sigma)^{-1}\tilde{\Sigma}}(y)$ *and* $t^+(x,y)=(1+\kappa\delta_\Sigma)^{p}\Phi_{\tilde{\upsilon}(x),(1+\delta_\Sigma)\tilde{\Sigma}}(y)$.

If

$$\begin{cases}\forall x\in\mathbb{R}^d,\|\upsilon(x)-\tilde{\upsilon}(x)\|^2\leq p\gamma_\kappa L_-\lambda_-\frac{\lambda_-}{\lambda_+}{\delta_\Sigma}^2\\ (1+\frac{2}{25}\delta_\Sigma)^{-1}\tilde{L}\leq L\leq\tilde{L}\\ \forall 1\leq i\leq p,|A_{i,i}^{-1}-\tilde{A}_{i,i}^{-1}|\leq\frac{1}{10}\frac{\delta_\Sigma}{\lambda_+}\\ \forall y\in\mathbb{R}^p,\|Py-\tilde{P}y\|\leq\frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\|y\|\end{cases}$$

then $[t^-,t^+]$ *is a* $\frac{\delta}{5}$ *Hellinger bracket such that* $t^-(x,y)\leq\Phi_{\upsilon(x),\Sigma}(y)\leq t^+(x,y)$.

We consider three cases: the parameter (mean, volume, matrix) is known ($\star=0$), unknown but common to all classes ($\star=c$), unknown and possibly different for every class ($\star=K$). For example, $[\nu_K,L_0,P_c,A_0]$ denotes a model in which only means are free and eigenvector matrices are assumed to be equal and unknown. Under our assumption that $\exists C_\Upsilon$ s.t $\forall\delta\in(0,\sqrt{2}]$,

$$H_{d_{\|\sup\|_\infty}}(\delta,\Upsilon_K)\leq\dim(\Upsilon_K)\left(C_\Upsilon+\ln\left(\frac{1}{\delta}\right)\right)$$

we deduce:

$$H_{[.],\max_k\sup_x d_y}\left(\frac{\delta}{5},\mathcal{G}\right)\leq\mathcal{D}\left(\mathcal{C}+\ln\left(\frac{1}{\delta}\right)\right)\tag{1}$$

where $\mathcal{D}=Z_{\upsilon,\star}+Z_{L,\star}+\frac{p(p-1)}{2}Z_{P,\star}+(p-1)Z_{A,\star}$ and

$$\begin{aligned}\mathcal{C}=&\ln\left(5p\sqrt{\kappa^2\cosh\left(\frac{2\kappa}{5}\right)+\frac{1}{2}}\right)+\frac{Z_{\upsilon,\star}C_\Upsilon}{\mathcal{D}}+\frac{Z_{\upsilon,\star}}{2\mathcal{D}}\ln\left(\frac{\lambda_+}{p\gamma_\kappa L_-\lambda_-^2}\right)\\&+\frac{Z_{L,\star}}{\mathcal{D}}\ln\left(\frac{4+129\ln\left(\frac{L_+}{L_-}\right)}{10}\right)+\frac{Z_{P,\star}}{\mathcal{D}}\left(\ln(c_U)+\frac{p(p-1)}{2}\ln\left(\frac{10\lambda_+}{\lambda_-}\right)\right)\\&+\frac{Z_{A,\star}(p-1)}{\mathcal{D}}\ln\left(\frac{4}{5}+\frac{52\lambda_+}{5\lambda_-}\ln\left(\frac{\lambda_+}{\lambda_-}\right)\right)\end{aligned}$$

$$Z_{v,K} = \dim(\Upsilon_K), Z_{v,c} = \dim(\Upsilon_1), Z_{v,0} = 0 \qquad Z_{L,0} = Z_{P,0} = Z_{A,0} = 0,$$
$$Z_{L,c} = Z_{P,c} = Z_{A,c} = 1, \qquad Z_{L,K} = Z_{P,K} = Z_{A,K} = K.$$

We notice that the following upper-bound of $\mathcal{C}$ is independent from the model of the collection, because we have made this hypothesis on C_Υ.

$$\mathcal{C} \leq \ln\left(5p\sqrt{\kappa^2\cosh\left(\frac{2\kappa}{5}\right)+\frac{1}{2}}\right) + C_\Upsilon + \frac{1}{2}\ln\left(\frac{\lambda_+}{p\gamma_\kappa L_-\lambda_-^2}\right) + \ln\left(\frac{4+129\ln\left(\frac{L_+}{L_-}\right)}{10}\right) + \frac{2}{p(p-1)}\ln(c_U) + \ln\left(\frac{10\lambda_+}{\lambda_-}\right) + \ln\left(\frac{4}{5}+\frac{52\lambda_+}{5\lambda_-}\ln\left(\frac{\lambda_+}{\lambda_-}\right)\right) := \mathcal{C}_1.$$

We conclude that $H_{[.],\sup_x d_y}(\delta, S_m) \leq \dim(S_m)(C_m + \ln(\frac{1}{\delta}))$, with

$$\dim(S_m) = \dim(W_K) + \mathcal{D}$$
$$C_m = \frac{\dim(W_K)}{\dim(S_m)}\left(C_W + \ln\left(\frac{20\sqrt{K}}{3\sqrt{3}}\right)\right) + \frac{\mathcal{D}\mathcal{C}_1}{\dim(S_m)} \leq C_W + \ln\left(\frac{20\sqrt{K_{\max}}}{3\sqrt{3}}\right) + \mathcal{C}_1 := \mathfrak{C}$$

Note that the constant $\mathfrak{C}$ does not depend on the dimension $\dim(S_m)$ of the model, thanks to the hypothesis that C_W is common for every model S_m in the collection. Using Proposition 1, we deduce thus that

$$n\sigma_m^2 \leq \dim(S_m)\left(2\left(\sqrt{\mathfrak{C}}+\sqrt{\pi}\right)^2 + \left(\ln\frac{n}{(\sqrt{\mathfrak{C}}+\sqrt{\pi})^2\dim(S_m)}\right)_+\right).$$

Theorem 2 yields then, for a collection $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$, with $\mathcal{M} = \{(K, W_K, \Upsilon_K, V_K) | K \in \mathbb{N}\setminus\{0\}, W_K, \Upsilon_K, V_K$ as previously defined $\}$ for which Assumption (K) holds, the oracle inequality of Theorem 1 as soon as

$$\text{pen}(m) \geq \kappa\left(\dim(S_m)\left(2\left(\sqrt{\mathfrak{C}}+\sqrt{\pi}\right)^2 + \left(\ln\frac{n}{(\sqrt{\mathfrak{C}}+\sqrt{\pi})^2\dim(S_m)}\right)_+\right) + x_m\right).$$

B.2. Lemma proofs

For sake of brevity, some technical proofs are omitted here. They can be found in an extended version.

B.2.1. Bracketing entropy's decomposition

We prove here a slightly more general Lemma than Lemma 3

Lemma 5. *Let*

$$\mathcal{P} = \left\{ \pi = (\pi_k)_{1\leq k\leq K} \middle| \forall k, \pi_k : \mathcal{X} \to \mathbb{R}^+ \text{ and } \forall x \in \mathcal{X}, \sum_{k=1}^{K} \pi_k(x) = 1 \right\},$$

$$\Psi = \left\{ (\psi_1, \ldots, \psi_K) \middle| \forall k, \psi_k : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+, \text{ and } \forall x, \forall k, \int \psi_k(x,y)dy = 1 \right\},$$

$$\mathcal{C} = \left\{ (x,y) \mapsto \sum_{k=1}^{K} \pi_k(x)\psi_k(x,y) \middle| \pi \in \mathcal{P}, \psi \in \Psi \right\}.$$

Then for all δ *in* $(0, \sqrt{2}]$,

$$H_{[.],\sup_x d_y}(\delta, \mathcal{C}) \leq H_{[.],\sup_x d_k}\left(\frac{\delta}{5}, \mathcal{P}\right) + H_{[.],\sup_x \max_k d_y}\left(\frac{\delta}{5}, \Psi\right).$$

The proof mimics the one of Lemma 7 from Cohen and Le Pennec (2011). It is possible to obtain such an inequality if the covariate X is not bounded, using the smaller distance $d^{\otimes n}$ for the entropy with bracketing of $\mathcal{C}$. More precisely,

Lemma 6. *For all* δ *in* $(0, \sqrt{2}]$, $H_{[.],d^{\otimes n}}(\delta, \mathcal{C}) \leq H_{[.],d_{\mathcal{P}}}(\frac{\delta}{2}, \mathcal{P}) + H_{[.],d_\Psi}(\frac{\delta}{2}, \Psi)$, *with* $d^2_{\mathcal{P}}(\pi^+, \pi^-) = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n d_k^2(\pi^+(X_i), \pi^-(X_i))]$ *and* $d^2_\Psi(\psi^+, \psi^-) = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K d_y^2(\psi_k^+(X_i), \psi_k^-(X_i))]$. *But bounding such bracketing entropies for* $\mathcal{P}$ *and* Ψ *becomes much more challenging.*

Proof. First we will exhibit a covering of bracket of $\mathcal{C}$.

Let $([\pi^{i,-}, \pi^{i,+}])_{1\leq i\leq N_{\mathcal{P}}}$ be a minimal covering of δ bracket for $\sup_x d_k$ of $\mathcal{P}$:

$$\forall i \in \{1, \ldots, N_{\mathcal{P}}\}, \forall x \in \mathcal{X}, d_k(\pi^{i,-}(x), \pi^{i,+}(x)) \leq \delta.$$

Let $([\psi^{i,-}, \psi^{i,+}])_{1\leq i\leq N_\Psi}$ be a minimal covering of δ bracket for $\sup_x \max_k d_y$ of Ψ:

$$\forall i \in \{1, \ldots, N_\Psi\}, \forall x \in \mathcal{X}, \forall k \in \{1, \ldots, K\}, d_y(\psi_k^{i,-}(x,.), \psi_k^{i,+}(x,.)) \leq \delta.$$

Let s be a density in $\mathcal{C}$. By definition, there is π in $\mathcal{P}$ and ψ in Ψ such that for all (x,y) in $\mathcal{X} \times \mathcal{Y}$, $s(y|x) = \sum_{k=1}^K \pi_k(x)\psi_k(x,y)$.

Due to the covering, there is i in $\{1, \ldots, N_{\mathcal{P}}\}$ such that

$$\forall x \in \mathcal{X}, \forall k \in \{1, \ldots, K\}, \pi_k^{i,-}(x) \leq \pi_k(x) \leq \pi_k^{i,+}(x).$$

There is also j in $\{1, \ldots, N_\Psi\}$ such that

$$\forall x \in \mathcal{X}, \forall k \in \{1, \ldots, K\}, \forall y \in \mathcal{Y}, \psi_k^{j,-}(x,y) \leq \psi_k(x,y) \leq \psi_k^{j,+}(x,y).$$

Since for all x, for all k and for all y, $\pi_k(x)$ and $\psi_k(x,y)$ are non-negatives, we may multiply term-by-term and sum these inequalities over k to obtain:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \sum_{k=1}^{K} \left(\pi_k^{i,-}(x)\right)_+ \left(\psi_k^{j,-}(x,y)\right)_+ \leq s(y|x) \leq \sum_{k=1}^{K} \pi_k^{i,+}(x)\psi_k^{j,+}(x,y).$$

$$\left(\left[\sum_{k=1}^{K}(\pi_k^{i,-})_+(\psi_k^{j,-})_+, \sum_{k=1}^{K}\pi_k^{i,+}\psi_k^{j,+}\right]\right)_{\substack{1\leq i\leq N_{\mathcal{P}}\\ 1\leq j\leq N_\Psi}}$$

is thus a bracket covering of $\mathcal{C}$.

Now, we focus on brackets' size using lemmas from Cohen and Le Pennec (2011) (namely Lemma 11, 12, 13), To lighten the notations, π_k^- and ψ_k^- are supposed non-negatives for all k. Following their Lemma 12, only using Cauchy-Schwarz inequality, we prove that

$$\sup_x d_y^2\left(\sum_{k=1}^{K}\pi_k^-(x)\psi_k^-(x,.), \sum_{k=1}^{K}\pi_k^+(x)\psi_k^+(x,.)\right) \leq \sup_x d_{y,k}^2(\pi^-(x)\psi^-(x,.), \pi^+(x)\psi^+(x,.))$$

Then, using Cauchy-Schwarz inequality again, we get by their Lemma 11:

$$\sup_x d_{y,k}^2(\pi^-(x)\psi^-(x,.), \pi^+(x)\psi^+(x,.)) \leq \sup_x \left(\max_k d_y(\psi_k^+(x,.), \psi_k^-(x,.))\sqrt{\sum_{k=1}^{K}\pi_k^+(x)} + d_k(\pi^+(x), \pi^-(x))\max_k \sqrt{\int \psi_k^-(x,y)dy}\right)^2$$

According to their Lemma 13, $\forall x, \sum_{k=1}^{K}\pi_k^+(x) \leq 1 + 2(\sqrt{2}+\sqrt{3})\delta$.

$$\sup_x \left(\max_k d_y(\psi_k^+(x,.), \psi_k^-(x,.))\sqrt{\sum_{k=1}^{K}\pi_k^+(x)} + d_k(\pi^+(x), \pi^-(x))\max_k \sqrt{\int \psi_k^-(x,y)dy}\right)^2$$

$$\leq \left(\sqrt{1+2(\sqrt{2}+\sqrt{3})\delta}+1\right)^2\delta^2 \leq (5\delta)^2$$

The result follows from the fact we exhibited a 5δ covering of brackets of $\mathcal{C}$, with cardinality $N_{\mathcal{P}}N_\Psi$. □

B.2.2. Bracketing entropy of weight's families

General case We prove

Lemma 4. *For any* $\delta \in (0, \sqrt{2}]$,

$$H_{[.],\sup_x d_k}\left(\frac{\delta}{5}, \mathcal{P}\right) \leq H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{20\sqrt{K}}, W_K\right).$$

Proof. We show that $\forall (w,z) \in (W_K)^2, \forall k \in \{1,\dots,K\}, \forall x \in \mathcal{X}, |\sqrt{\pi_{w,k}(x)} - \sqrt{\pi_{z,k}(x)}| \leq F(k,x)d(w,z)$, with F a function and d some distance. We define $\forall k, \forall u \in \mathbb{R}^K, A_k(u) = \frac{\exp(u_k)}{\sum_{k=1}^K \exp(u_k)}$, so $\pi_{w,k}(x) = A_k(w(x))$.

$\forall (u,v) \in (\mathbb{R}^K)^2$,

$$\left|\sqrt{A_k(v)} - \sqrt{A_k(u)}\right| = \left|\int_0^1 \nabla\left(\sqrt{A_k}\right)(u + t(v-u)).(v-u)dt\right|$$

Besides,

$$\begin{aligned}\nabla\left(\sqrt{A_k}\right)(u) &= \left(\frac{1}{2}\sqrt{A_k(u)}\frac{\partial}{\partial u_l}(\ln(A_k(u)))\right)_{1\leq l\leq K} \\ &= \left(\frac{1}{2}\sqrt{A_k(u)}\left(\delta_{k,l} - A_l(u)\right)\right)_{1\leq l\leq K}\end{aligned}$$

$$\begin{aligned}&\left|\sqrt{A_k(v)} - \sqrt{A_k(u)}\right| \\ &= \frac{1}{2}\left|\int_0^1 \sqrt{A_k(u+t(v-u))}\sum_{l=1}^K \left(\delta_{k,l} - A_l(u+t(v-u))\right)(v_l - u_l)dt\right| \\ &\leq \frac{\|v-u\|_\infty}{2}\int_0^1 \sqrt{A_k(u+t(v-u))}\sum_{l=1}^K \left|\delta_{k,l} - A_l(u+t(v-u))\right| dt\end{aligned}$$

Since $\forall u \in \mathbb{R}^K, \sum_{k=1}^K A_k(u) = 1$, $\sum_{l=1}^K |\delta_{k,l} - A_l(u)| = 2(1 - A_k(u))$

$$\begin{aligned}&\left|\sqrt{A_k(v)} - \sqrt{A_k(u)}\right| \\ &\leq \|v-u\|_\infty \int_0^1 \sqrt{A_k(u+t(v-u))}\left(1 - A_k(u+t(v-u))\right)dt \\ &\leq \frac{2}{3\sqrt{3}}\|v-u\|_\infty\end{aligned}$$

since $x \mapsto \sqrt{x}(1-x)$ is maximal over [0,1] for $x = \frac{1}{3}$. We deduce that for any (w,z) in $(W_K)^2$, for all k in $\{1,\dots,K\}$, for any x in $\mathcal{X}$, $|\sqrt{\pi_{w,k}(x)} - \sqrt{\pi_{z,k}(x)}| \leq \frac{2}{3\sqrt{3}}\max_l \|w_l - z_l\|_\infty$.

By hypothesis, for any positive ϵ, an ϵ-net $\mathcal{N}$ of W_K may be exhibited. Let w be an element of W_K. There is a z belonging to the ϵ-net $\mathcal{N}$ such that $\max_l \|z_l - w_l\|_\infty \leq \epsilon$. Since for all k in $\{1, \ldots, K\}$, for any x in $\mathcal{X}$,

$$|\sqrt{\pi_{w,k}(x)} - \sqrt{\pi_{z,k}(x)}| \leq \frac{2}{3\sqrt{3}} \max_l \|w_l - z_l\|_\infty \leq \frac{2}{3\sqrt{3}}\epsilon,$$

and

$$\sum_{k=1}^{K} \left(\sqrt{\pi_{z,k}(x)} + \frac{2}{3\sqrt{3}}\epsilon - \sqrt{\pi_{z,k}(x)} + \frac{2}{3\sqrt{3}}\epsilon \right)^2 = K \left(\frac{4\epsilon}{3\sqrt{3}} \right)^2,$$

$([(\sqrt{\pi_z} - \frac{2}{3\sqrt{3}}\epsilon)^2, (\sqrt{\pi_z} + \frac{2}{3\sqrt{3}}\epsilon)^2])_{z\in\mathcal{N}}$ is a $\frac{4\epsilon\sqrt{K}}{3\sqrt{3}}$-bracketing cover of $\mathcal{P}$. As a result, $H_{[],\sup_x d_k}(\frac{\delta}{5}, \mathcal{P}) \leq H_{d_{\|\sup\|_\infty}}(\frac{3\sqrt{3}}{20\sqrt{K}}\delta, W_K)$. □

Case: $W_K = \{0\} \otimes W^{K-1}$ with W constructed from bounded functions We remind that

$$W = \left\{ w : \mathcal{X} \to \mathbb{R} / w(x) = \sum_{i=0}^{d_W} \alpha_i \psi_{W,i} \text{ and } \|\alpha\|_\infty \leq T_W \right\}$$

with $\|\psi_{W,i}\|_\infty \leq 1$.

Proof of Part 1 of Lemma 1. W_K is a finite dimensional compact set. Thanks to the result in the general case, we get

$$H_{[.],\sup_x d_k}\left(\frac{\delta}{5}, \mathcal{P}\right) \leq H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{20\sqrt{K-1}}, W_K\right)$$

now as for all w, v in W_K, $\max_k \|w_k - v_k\|_\infty \leq \max_k \sum_{i=0}^{d_W} |\alpha_{k,i}^w - \alpha_{k,i}^v| \leq d_W \max_{k,i} |\alpha_{k,i}^w - \alpha_{k,i}^v|$

$$\begin{aligned}
&\leq H_{\|.\|_\infty}\left(\frac{3\sqrt{3}\delta}{20\sqrt{K-1}d_W}, \left\{\alpha \in \mathbb{R}^{(K-1)d_W} / \|\alpha\|_\infty \leq T_W\right\}\right) \\
&\leq (K-1)d_W \ln\left(1 + \frac{20\sqrt{K-1}T_W d_W}{3\sqrt{3}\delta}\right) \\
&\leq (K-1)d_W \left[\ln\left(\sqrt{2} + \frac{20}{3\sqrt{3}}T_W\sqrt{K-1}d_W\right) + \ln\left(\frac{1}{\delta}\right)\right].
\end{aligned}$$

□

The second Lemma is just a consequence of $d_W = \binom{d'_W + d}{d}$.

B.2.3. Bracketing entropy of Gaussian families

General case We rely on a general construction of Gaussian brackets:

Proposition. 2. *Let* $\kappa \geq \frac{17}{29}$, $\gamma_\kappa = \frac{25(\kappa-\frac{1}{2})}{49(1+\frac{2\kappa}{5})}$. *For any* $0 < \delta \leq \sqrt{2}$, *any* $p \geq 1$ *and any* $\delta_\Sigma \leq \frac{1}{5\sqrt{\kappa^2\cosh(\frac{2\kappa}{5})+\frac{1}{2}}}\frac{\delta}{p}$, *let* $(\upsilon, L, A, P) \in \Upsilon \times [L_-, L_+] \times \mathcal{A}(\lambda_-, \lambda_+) \times SO(p)$ *and* $(\tilde{\upsilon}, \tilde{L}, \tilde{A}, \tilde{P}) \in \Upsilon \times [L_-, L_+] \times \mathcal{A}(\lambda_-, +\infty) \times SO(p)$, *define* $\Sigma = LPAP'$ *and* $\tilde{\Sigma} = \tilde{L}\tilde{P}\tilde{A}\tilde{P}'$,

$$t^-(x,y) = (1+\kappa\delta_\Sigma)^{-p}\Phi_{\tilde{\upsilon}(x),(1+\delta_\Sigma)^{-1}\tilde{\Sigma}}(y) \text{ and}$$
$$t^+(x,y) = (1+\kappa\delta_\Sigma)^{p}\Phi_{\tilde{\upsilon}(x),(1+\delta_\Sigma)\tilde{\Sigma}}(y).$$

If

$$\begin{cases} \forall x \in \mathcal{X}, \|\upsilon(x) - \tilde{\upsilon}(x)\|^2 \leq p\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+}\delta_\Sigma^2 \\ \left(1+\frac{2}{25}\delta_\Sigma\right)^{-1}\tilde{L} \leq L \leq \tilde{L} \\ \forall 1 \leq i \leq p, |A_{i,i}^{-1} - \tilde{A}_{i,i}^{-1}| \leq \frac{1}{10}\frac{\delta_\Sigma}{\lambda_+} \\ \forall y \in \mathbb{R}^p, \|Py - \tilde{P}y\| \leq \frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\|y\| \end{cases}$$

then $[t^-, t^+]$ *is a* $\delta/5$ *Hellinger bracket such that* $t^-(x,y) \leq \Phi_{\upsilon(x),\Sigma}(y) \leq t^+(x,y)$.

This statement is similar to Lemma 10 in Cohen and Le Pennec (2011). Admitting this proposition, we are brought to construct nets over the spaces of the means, the volumes, the eigenvector matrices and the normalized eigenvalue matrices. We consider three cases: the parameter (mean, volume, matrix) is known ($\star = 0$), unknown but common to all classes ($\star = c$), unknown and possibly different for every class ($\star = K$). For example, $[\nu_K, L_0, P_c, A_0]$ denotes a model in which only means are free and eigenvector matrices are assumed to be equal and unknown.

If the means are free ($\star = K$), we construct a grid G_{Υ_K} over Υ_K, which is compact. Since

$$H_{d_{\|\sup\|_\infty}}\left(\sqrt{p\gamma_\kappa L_-\lambda_-\frac{\lambda_-}{\lambda_+}}\delta_\Sigma, \Upsilon_K\right) \leq \dim(\Upsilon_K)\left(C_\Upsilon + \ln\left(\frac{1}{\sqrt{p\gamma_\kappa L_-\lambda_-\frac{\lambda_-}{\lambda_+}}\delta_\Sigma}\right)\right),$$
$$\left|G_{\Upsilon_K}\left(\sqrt{p\gamma_\kappa L_-\lambda_-\frac{\lambda_-}{\lambda_+}}\delta_\Sigma\right)\right| \leq \left(C_\Upsilon + \ln\left(\frac{1}{\sqrt{p\gamma_\kappa L_-\lambda_-\frac{\lambda_-}{\lambda_+}}\delta_\Sigma}\right)\right)^{\dim(\Upsilon_K)}.$$

If the means are common and unknown ($\star = c$), belonging to Υ_1, we construct a grid $G_{\Upsilon_c}(\sqrt{p\gamma_\kappa L_-\lambda_-\frac{\lambda_-}{\lambda_+}}\delta_\Sigma)$ over Υ_1 with cardinality at most

$$\left(C_\Upsilon + \ln\left(\frac{1}{\sqrt{p\gamma_\kappa L_-\lambda_-\frac{\lambda_-}{\lambda_+}}\delta_\Sigma}\right)\right)^{D_{\Upsilon_1}}.$$

Finally, if the means are known ($\star = 0$), we do not need to construct a grid. In the end,

$$\left|G_{\Upsilon_\star}\left(\sqrt{p\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+}}\delta_\Sigma\right)\right| \leq \left(C_\Upsilon + \ln\left(\frac{1}{\sqrt{p\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+}}\delta_\Sigma}\right)\right)^{Z_{\upsilon,\star}},$$

with $Z_{\upsilon,K} = \dim(\Upsilon_K)$, $Z_{\upsilon,c} = D_{\Upsilon_1}$ and $Z_{\upsilon,0} = 0$.

Then, we consider the grid G_L over $[L_-, L_+]$:

$$G_L\left(\frac{2}{25}\delta_\Sigma\right) = \left\{L_-\left(1+\frac{2}{25}\delta_\Sigma\right)^g / g \in \mathbb{N}, L_-\left(1+\frac{2}{25}\delta_\Sigma\right)^g \leq L_+\right\}$$

$$\left|G_L\left(\frac{2}{25}\delta_\Sigma\right)\right| \leq 1 + \frac{\ln\left(\frac{L_+}{L_-}\right)}{\ln\left(1+\frac{2}{25}\delta_\Sigma\right)}$$

Since $\delta_\Sigma \leq \frac{2}{5}$, $\ln(1+\frac{2}{25}\delta_\Sigma) \geq \frac{10}{129}\delta_\Sigma$.

$$\left|G_L\left(\frac{2}{25}\delta_\Sigma\right)\right| \leq 1 + \frac{129\ln\left(\frac{L_+}{L_-}\right)}{10\delta_\Sigma} \leq \frac{4 + 129\ln\left(\frac{L_+}{L_-}\right)}{10\delta_\Sigma}$$

By definition of a net, for any $P \in SO(p)$ there is a $\tilde{P} \in G_P(\frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma)$ such that $\forall y \in \mathbb{R}^p, \|Py - \tilde{P}y\| \leq \frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\|y\|$. There exists a universal constant c_U such that $|G_P(\frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma)| \leq c_U(\frac{10\lambda_+}{\lambda_-\delta_\Sigma})^{\frac{p(p-1)}{2}}$.

For the grid G_A, we look at the condition on the $p-1$ first diagonal values and obtain:

$$\left|G_A\left(\frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\right)\right| \leq \left(2 + \frac{\ln\left(\frac{\lambda_+}{\lambda_-}\right)}{\ln\left(1+\frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\right)}\right)^{p-1}$$

Since $\delta_\Sigma \leq \frac{2}{5}$, $\ln(1+\frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma) \geq \frac{5}{52}\frac{\lambda_-}{\lambda_+}\delta_\Sigma$, then

$$\begin{aligned}\left|G_A\left(\frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\right)\right| &\leq \left(2 + \frac{52}{5\delta_\Sigma}\frac{\lambda_+}{\lambda_-}\ln\left(\frac{\lambda_+}{\lambda_-}\right)\right)^{p-1} \\ &\leq \left(4 + 52\frac{\lambda_+}{\lambda_-}\ln\left(\frac{\lambda_+}{\lambda_-}\right)\right)^{p-1}\left(\frac{1}{5\delta_\Sigma}\right)^{p-1}\end{aligned}$$

Let $Z_{L,0} = Z_{P,0} = Z_{A,0} = 0$, $Z_{L,c} = Z_{P,c} = Z_{A,c} = 1$, $Z_{L,K} = Z_{P,K} = Z_{A,K} = K$. We define $f_{\upsilon,\star}$ from $\Upsilon_\star$ to Υ_K by

$$\begin{cases} 0 \mapsto (\upsilon_{0,1}, \ldots, \upsilon_{0,1}) \text{ if } \star = 0 \\ \upsilon \mapsto (\upsilon, \ldots, \upsilon) \text{ if } \star = c \\ (\upsilon_1, \ldots, \upsilon_K) \mapsto (\upsilon_1, \ldots, \upsilon_K) \text{ if } \star = K \end{cases}$$

and similarly $f_{L,\star}$, $f_{P,\star}$ and $f_{A,\star}$, respectively from $(\mathbb{R}_+)^{Z_{L,\star}}$ into $(\mathbb{R}_+)^K$, from $(SO(p))^{Z_{P,\star}}$ into $(SO(p))^K$ and from $\mathcal{A}(\lambda_-,\lambda_+)^{Z_{A,\star}}$ into $\mathcal{A}(\lambda_-,\lambda_+)^K$.

We define

$$\Gamma : (\upsilon_1,\dots,\upsilon_K,L_1,\dots,L_K,P_1,\dots,P_K,A_1,\dots,A_K) \mapsto (\upsilon_k, L_k P_k A_k P_k')_{1\le k\le K}$$

and $\Psi : (\upsilon_k,\Sigma_k)_{1\le k\le K} \mapsto (\Phi_{\upsilon_k,\Sigma_k})_{1\le k\le K}$. The image of $\Upsilon_\star \times [L_-,L_+]^{Z_{L,\star}} \times SO(p)^{Z_{P,\star}} \times \mathcal{A}(\lambda_-,\lambda_+)^{Z_{A,\star}}$ by $\Psi\circ\Gamma\circ(f_{\upsilon,\star}\otimes f_{L,\star}\otimes f_{P,\star}\otimes f_{A,\star})$ is the set $\mathcal{G}$ of all K-tuples of Gaussian densities of type $[\upsilon_\star, L_\star, P_\star, A_\star]$.

Now, we define B:

$$(\upsilon_k,\Sigma_k)_{1\le k\le K} \mapsto \left((1+\kappa\delta_\Sigma)^{-p}\Phi_{\upsilon_k,(1+\delta_\Sigma)^{-1}\Sigma_k}, (1+\kappa\delta_\Sigma)^{p}\Phi_{\upsilon_k,(1+\delta_\Sigma)\Sigma_k}\right)_{1\le k\le K}.$$

The image of $G_{\Upsilon_\star}\times G_L^{Z_{L,\star}}\times G_P^{Z_{P,\star}}\times G_A^{Z_{A,\star}}$ by $B\circ\Gamma\circ(f_{\upsilon,\star}\otimes f_{L,\star}\otimes f_{P,\star}\otimes f_{A,\star})$ is a $\delta/5$-bracket covering of $\mathcal{G}$, with cardinality bounded by

$$\left(\frac{\sqrt{\lambda_+}\exp(C_\Upsilon)}{\sqrt{p\gamma_\kappa L_-\lambda_-^2}\delta_\Sigma}\right)^{Z_{\Upsilon,\star}} \times \left(\frac{4+129\ln\left(\frac{L_+}{L_-}\right)}{10\delta_\Sigma}\right)^{Z_{L,\star}} \times c_U^{Z_{P,\star}}\left(\frac{10\lambda_+}{\lambda_-\delta_\Sigma}\right)^{\frac{p(p-1)}{2}Z_{P,\star}}$$
$$\times\left(4+52\frac{\lambda_+}{\lambda_-}\ln\left(\frac{\lambda_+}{\lambda_-}\right)\right)^{(p-1)Z_{A,\star}}\left(\frac{1}{5\delta_\Sigma}\right)^{(p-1)Z_{A,\star}}.$$

Taking $\delta_\Sigma = \frac{1}{5\sqrt{\kappa^2\cosh(\frac{2\kappa}{5})+\frac12}}\frac{\delta}{p}$, we obtain

$$H_{[.],\sup_x\max_k d_y}\left(\frac{\delta}{5},\mathcal{G}\right) \le \mathcal{D}\left(\mathcal{C}+\ln\left(\frac{1}{\delta}\right)\right)$$

with $\mathcal{D} = Z_{\upsilon,\star}+Z_{L,\star}+\frac{p(p-1)}{2}Z_{P,\star}+(p-1)Z_{A,\star}$ and

$$\mathcal{C} = \ln\left(5p\sqrt{\kappa^2\cosh\left(\frac{2\kappa}{5}\right)+\frac12}\right) + \frac{Z_{\upsilon,\star}C_\Upsilon}{\mathcal{D}} + \frac{Z_{\upsilon,\star}}{2\mathcal{D}}\ln\left(\frac{\lambda_+}{p\gamma_\kappa L_-\lambda_-^2}\right)$$
$$+\frac{Z_{L,\star}}{\mathcal{D}}\ln\left(\frac{4+129\ln\left(\frac{L_+}{L_-}\right)}{10}\right) + \frac{Z_{P,\star}}{\mathcal{D}}\left(\ln(c_U)+\frac{p(p-1)}{2}\ln\left(\frac{10\lambda_+}{\lambda_-}\right)\right)$$
$$+\frac{Z_{A,\star}(p-1)}{\mathcal{D}}\ln\left(\frac45+\frac{52\lambda_+}{5\lambda_-}\ln\left(\frac{\lambda_+}{\lambda_-}\right)\right)$$

Case: Υ_K generated from bounded functions Using previous work, we only have to handle Υ_K's bracketing entropy. Just like for W_K, we aim at bounding the bracketing entropy by the entropy of the parameters' space

We focus on the case of Lemma 1 where $\Upsilon_K = \Upsilon^K$ and

$$\Upsilon = \left\{\upsilon:\mathcal{X}\to\mathbb{R}^p \middle| \forall j\in\{1,\dots,p\}, \forall x, \upsilon_j(x) = \sum_{i=0}^{d_\Upsilon}\alpha_i^{(j)}\psi_{\Upsilon,i}, \text{ and } \|\alpha\|_\infty \le T_\Upsilon\right\}$$

We consider for any υ, ν in Υ and any x in $[0,1]^d$,

$$\begin{aligned}
\|\upsilon(x)-\nu(x)\|_2^2 &= \sum_{j=1}^{p}\left(\sum_{i=0}^{d_\Upsilon}\left(\alpha_i^{\upsilon,(j)}-\alpha_i^{\nu,(j)}\right)\psi_{\Upsilon,j}(x)\right)^2\\
&\leq \sum_{j=1}^{p}\left(\sum_{i=0}^{d_\Upsilon}\left(\alpha_i^{\upsilon,(j)}-\alpha_i^{\nu,(j)}\right)^2\right)\left(\sum_{i=0}^{d_\Upsilon}|\psi_{\Upsilon,j}(x)|^2\right)\\
&\leq d_\Upsilon\sum_{j=1}^{p}\sum_{i=0}^{d_\Upsilon}\left(\alpha_r^{\upsilon,(j)}-\alpha_r^{\nu,(j)}\right)^2\\
&\leq pd_\Upsilon^2\max_{j,i}\left(\alpha_i^{\upsilon,(j)}-\alpha_r^{\nu,(j)}\right)^2
\end{aligned}$$

So,

$$\begin{aligned}
H_{\max_k\sup_x\|\|_2}(\delta,\Upsilon_K) &\leq H_{\max_{k,j,r}|.|}\left(\frac{\delta}{\sqrt{p}d_\Upsilon},\left\{\left(\alpha_r^{(j,k)}\right)_{\substack{1\leq j\leq p\\ |r|\leq d'_\Upsilon\\ 1\leq k\leq K}}\Big|\|\alpha\|_\infty\leq T_\Upsilon\right\}\right)\\
&\leq pKd_\Upsilon\ln\left(1+\frac{\sqrt{p}d_\Upsilon T_\Upsilon}{\delta}\right)\\
&\leq pKd_\Upsilon\left[\ln\left(\sqrt{2}+\sqrt{p}d_\Upsilon T_\Upsilon\right)+\ln\left(\frac{1}{\delta}\right)\right]\\
&\leq \dim(\Upsilon_K)\left(C_\Upsilon+\ln\left(\frac{1}{\delta}\right)\right)
\end{aligned}$$

with $\dim(\Upsilon_K)=pK\binom{d'_\Upsilon+d}{d}$ and $C_\Upsilon=\ln(\sqrt{2}+\sqrt{p}\binom{d'_\Upsilon+d}{d}T_\Upsilon)$.

The second part of Lemma 2 is deduced from the fact that if $\mathcal{X}=[0,1]^d$ and Υ is the set of linear combination of monomials of degree less that d'_Υ then $d_\Upsilon=\binom{d'_\Upsilon+d}{d}$.

Appendix C: Description of Newton-EM algorithm

In this section, Newton-EM algorithm is detailed. It consists in the classical EM algorithm in which the update of the weights has been replaced by some Newton steps. For further details on EM algorithm, refer to the technical report related to Young and Hunter (2010).

Newton-EM

Initialization Parameters for w, υ and Σ are given.
Newton steps for w Perform at most 5 steps Newton steps for w only while the like likelihood increases.
Maximization Update of υ and Σ with usual formulas in EM algorithm.

Initialization of Newton-EM

1. Draw K couples of points (X_i,Y_i) among data, defining K lines υ_l.
2. Classify the data: $k=\arg\min_l|Y_i-\upsilon_l(X_i)|$.

3. Proceed 3 steps of Newton-EM initialized with $w = 0$ and empirical covariance matrices and means.
4. Repeat 50 times the previous steps and choose the set of parameters with the greatest likelihood among the 50.

References

Antoniadis, A., Bigot, J., and von Sachs, R., A multiscale approach for statistical characterization of functional images. *Journal of Computational and Graphical Statistics*, 18, 2009. MR2649646

Barron, A., Huang, C., Li, J., and Luo, X., The mdl principle, penalized likelihoods, and statistical risk. *Festschrift for Jorma Rissanen. Tampere University Press, Tampere, Finland*, 2008.

Baudry, J.-P., Maugis, C., and Michel, B., Slope heuristics: Overview and implementation. *Statistics and Computing*, 22, 2011. MR2865029

Biernacki, Ch. and Castellan, G., A data-driven bound on variances for avoiding degeneracy in univariate gaussian mixtures. *Pub IRMA Lille*, 71, 2011.

Birgé, L. and Massart, P., Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, 138(1–2):33–73, 2007. ISSN 0178-8051. 10.1007/s00440-006-0011-8. MR2288064

Brinkman, N. D., Ethanol fuel-a single-cylinder engine study of efficiency and exhaust emissions. *SAE Technical Paper*, 810345, 1981.

Burnham, K. P. and Anderson, D. R., *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach*. Springer-Verlag, New-York, 2nd edition, 2002. MR1919620

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

Chamroukhi, F., Samé, A., Govaert, G., and Aknin, P., A hidden process regression model for functional data description. Application to curve discrimination. *Neurocomputing*, 73:1210–1221, March 2010.

Choi, T., Convergence of posterior distribution in the mixture of regressions. *Journal of Nonparametric Statistics*, 20(4):337–351, May 2008. MR2436382

Cohen, S. and Le Pennec, E., Conditional density estimation by penalized likelihood model selection and applications. Technical report, INRIA, 2011.

Cohen, S. X. and Le Pennec, E., Partition-based conditional density estimation. *ESAIM Probab. Stat.*, 17:672–697, 2013. ISSN 1292-8100. 10.1051/ps/2012017. MR3126157

Dempster, A. P., Laird, N. M., and Rubin, D. B., Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B.*, 39(1), 1977. MR0501537

Gassiat, E. and van Handel, R., The local geometry of finite mixtures. *Trans. Amer. Math. Soc.*, 366(2):1047–1072, 2014. MR3130325

Ge, Y. and Jiang, W., On consistency of bayesian inference with mixtures of logistic regression. *Neural Computation*, 18(1):224–243, January 2006. MR2185287

Genovese, C. and Wasserman, L., Rates of convergence for the gaussian mixture sieve. *The Annals of Statistics*, 28(4):1105–1127, August 2000. MR1810921

Huang, M., Li, R., and Wang, S., Nonparametric mixtures of regressions models. *Journal of the American Statistical Association*, 108(503):929–941, 2013. MR3174674

Huang, M. and Yao, W., Mixture of regression models with varying mixing proportions: A semiparametric approach. *J. Amer. Statist. Assoc.*, 107(498):711–724, 2012. ISSN 0162-1459. 10.1080/01621459.2012.682541. MR2980079

Hunter, D. R. and Young, D. S., Semiparametric mixtures of regressions. *J. Nonparametr. Stat.*, 24(1):19–38, 2012. ISSN 1048-5252. 10.1080/10485252.2011.608430. MR2885823

Jordan, M. I. and Jacobs, R. A., Hierarchical mixtures of experts and the em algorithm. In Maria Marinaro and Pietro G. Morasso, editors, *ICANN 94*, pages 479–486. Springer London, 1994. ISBN 978-3-540-19887-1.

Kolaczyk, E. D., Ju, J., and Gopal, S., Multiscale, multigranular statistical image segmentation. *Journal of the American Statistical Association*, 100:1358–1369, 2005. MR2236447

Lee, H. K. H., Consistency of posterior distributions for neural networks. *Neural Networks*, 13:629–642, July 2000.

Martin-Magniette, M. L., Mary-Huard, T., Bérard, C., and Robin, S., Chipmix: Mixture model of regressions for two-color chip-chip analysis. *Bioinformatics*, 24(16):i181–i186, 2008. 10.1093/bioinformatics/btn280.

Maugis, C. and Michel, B., A non asymptotic penalized criterion for gaussian mixture model selection. *ESAIM Probability and Statistics*, 2011. MR2870505

Maugis, C. and Michel, B., Adaptive density estimation using finite gaussian mixtures. *ESAIM Probability and Statistics*, 2012. Accepted for publication.

McLachlan, G. and Peel, D., *Finite Mixture Models*. Wiley, 2000. MR1789474

Rigollet, Ph., Kullback-Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, 40(2):639–665, 2012. 10.1214/11-AOS961. MR2933661

Van der Vaart, A. W. and Wellner, J. A., *Weak Convergence and Empirical Processes*. Springer, 1996. MR1385671

Viele, K. and Tong, B., Modeling with mixtures of linear regressions. *Stat. Comput.*, 12(4):315–330, 2002. ISSN 0960-3174. 10.1023/A:1020779827503. MR1951705

Young, D. S., Mixtures of regressions with changepoints. *Statistics and Computing*, 24(2):265–281, 2014. ISSN 0960-3174. 10.1007/s11222-012-9369-x. MR3165553

Young, D. S. and Hunter, D. R., Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics & Data Analysis*, 54(10):2253–2266, 2010. ISSN 0167-9473. MR2720486