Partition-Based Conditional Density Estimation

S. X. Cohen (IPANEMA / Synchrotron Soleil) and E. Le Pennec (SELECT / Inria Saclay - Île de

July 9, 2012

Abstract

We propose a general partition-based strategy to estimate conditional density with candidate densities that are piecewise constant with respect to the covariate. Capitalizing on a general penalized maximum likelihood model selection result, we prove, on two specific examples, that the penalty of each model can be chosen roughly proportional to its dimension. We first study a *classical* strategy in which the densities are chosen piecewise conditional according to the variable. We then consider Gaussian mixture models with mixing proportion that vary according to the covariate but with common mixture components. This model proves to be interesting for an unsupervised segmentation application that was our original motivation for this work.

1 Introduction

Assume we observe n pairs $((X_i, Y_i))_{1\leq i\leq n}$ of random variables, we are interested in estimating the law of the second one $Y_i \in \mathcal{Y}$, called variable, conditionally to the first one $X_i \in \mathcal{X}$, called covariate. In this paper, we assume that the pairs (X_i, Y_i) are independent while Y_i depends on X_i through its law. More precisely, we assume that the covariates X_i's are independent but not necessarily identically distributed. Assumption on the Y_i's is stronger: we assume that, conditionally to the X_i's, they are independent and each variable Y_i follows a law of density $s_0(\cdot|X_i)$ with respect to a common known measure $\mathrm{d}\lambda$. Our goal is to estimate this two-variable conditional density function $s_0(\cdot|\cdot)$ from the observations. In this paper, we apply a penalized maximum likelihood model selection result of [12] to partition-based collection in which the conditional densities depend on covariate in a piecewise constant manner.

The original conditional density estimation problem has been introduced by Rosenblatt [37] in the late 60's. In a stationary framework, he used a link between $s_0(y|x)$ and the supposed existing densities $s_{0'}(x)$ and $s_{0''}(x, y)$ of respectively X_i and (X_i, Y_i),

$$s_0(y|x) = \frac{s_{0''}(x, y)}{s_{0'}(x)},$$

and proposed a plugin estimate based on kernel estimation of both $s_{0''}(x, y)$ and $s_{0'}(x)$. Few other references on this subject seem to exist before the mid 90's with a study of a spline tensor based maximum likelihood estimator proposed by Stone [39] and a bias correction of Rosenblatt's estimator due to Hyndman, Bashtannyk, and Grunwald [26]. Kernel based method have been much studied since as stressed by Li and Racine [33]. To name a few, Fan, Yao, and Tong [18] and Gooijer and Zerom [21] consider local polynomial estimator, Hall, Wolff, and Yao [23] study a locally logistic estimator later extended by Hyndman and Yao [27]. Pointwise convergence properties are considered, and extensions to dependent data are often obtained. Those results are however non adaptive: their performances depend on a critical bandwidth choice that should be chosen according to the regularity of the unknown conditional density. Its practical choice is rarely discussed with the notable exception of Bashtannyk and Hyndman [5]. Extensions to censored cases have also been discussed for instance by Keilegom and Veraverbeke [29]. In the approach of Stone [39], the conditional density is estimated using a representation, a parametrized modelization. This idea has been reused by Györfi and Kohler [22] with a histogram based approach, by Efromovich [16, 17] with a Fourier basis, and by Brunel, Comte, and Lacour [11] and Akakpo and Lacour [2] with piecewise polynomial representation. Risks of those estimators are controlled with a total variation loss for the first one and a quadratic distance for the others. Furthermore within the quadratic framework, almost minimax adaptive estimators are constructed using respectively a blockwise attenuation principle and a penalized model selection approach. Kullback-Leibler type loss, and thus maximum likelihood approach, has only been considered by Stone [39] as mentioned before and by Blanchard et al. [10] in a classification setting with histogram type estimators.

In [12], we propose a penalized maximum likelihood model selection approach to estimate s_0. Given a collection of models $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ comprising conditional densities and their maximum likelihood estimates

$$\widehat{s}_m = \underset{s_m\in S_m}{\operatorname{argmin}} - \sum_{i=1}^{n} \ln s_m(Y_i|X_i),$$

we define, for a given penalty $\operatorname{pen}(m)$, the *best* model $S_{\widehat{m}}$ as the one that minimizes a penalized likelihood:

$$\widehat{m} = \underset{m\in\mathcal{M}}{\operatorname{argmin}} - \sum_{i=1}^{n} \ln \widehat{s}_m(Y_i|X_i) + \operatorname{pen}(m).$$

The main result of [12] is a sufficient condition on the penalty $\operatorname{pen}(m)$ such that an oracle type inequality holds for the conditional density estimation error. In this paper, we show how this theorem can be used to derive results for two interesting partition-based conditional density models, inspired by Kolaczyk and Nowak [31], Kolaczyk, Ju, and Gopal [30] and Antoniadis, Bigot, and Sachs [3].

Both are based on a recursive partitioning of space $\mathcal{X}$, assumed for sake of simplicity to be equal to $[0,1]^{d_X}$, they differ by the choice of the density used, once conditioned by covariates: in the first case, we consider traditional piecewise polynomial models, while, in the second case, we use Gaussian mixture models with common mixture components. The first case is motivated by the work of Willett and Nowak [42] where they propose a similar model for Poissonian intensities. The second one is drived by an application to unsupervised segmentation, which was our original motivation for this work. For both examples, we prove that the penalty can be chosen roughly proportional to the dimension of the model.

In Section 2, we summarize the setting and the results of [12]. We describe the loss considered, explain the penalty structure and present a general penalized maximum likelihood theorem we have proved. This will be a key tool for the study of the partition-based strategy conducted in Section 3. We describe first our general partition based approach in Section 3.1 and exemplify it with piecewise polynomial density with respect to the variable in Section 3.2 and with Gaussian mixture with varying proportion in Section 3.3. Main proofs are given in Appendix while proofs of the most technical lemmas are relegated to our technical report [13].

2 A general penalized maximum likelihood theorem

2.1 Framework and notation

As in [12], we observe n independent pairs $((X_i, Y_i))_{1\leq i \leq n} \in (\mathcal{X}, \mathcal{Y})^n$ where the X_i's are independent, but not necessarily of same law, and, conditionally to X_i, each Y_i is a random variable of unknown conditional density $s_0(\cdot|X_i)$ with respect to a known reference measure $\mathrm{d}\lambda$. For any model S_m, a set of candidate conditional densities, we estimate s_0 by the conditional density $\widehat{s}_m$ that maximizes the likelihood (conditionally to $(X_i)_{1\leq i\leq n}$) or equivalently that minimizes the opposite of the log-likelihood, denoted -log-likelihood from now on:

$$\widehat{s}_m = \underset{s_m \in S_m}{\operatorname{argmin}} \left(\sum_{i=1}^{n} - \ln(s_m(Y_i|X_i)) \right).$$

To avoid existence issue, we should work with almost minimizer of this quantity and define a η -log-likelihood minimizer as any $\widehat{s}_m$ that satisfies

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

Given a collection $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ of models, we construct a penalty $\mathrm{pen}(m)$ and select the best model $\widehat{m}$ as the one that minimizes

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

In [12], we give conditions on penalties ensuring that the resulting estimate $\widehat{s}_{\widehat{m}}$ is a *good* estimate of the true conditional density.

We should now specify our *goodness* criterion. As we are working in a maximum likelihood approach, the most natural quality measure is the Kullback-Leibler divergence *KL*. As we consider law with densities with respect to a known measure $\mathrm{d}\lambda$, we use the following notation

$$KL_\lambda(s,t) = KL(s\mathrm{d}\lambda, t\mathrm{d}\lambda) = \begin{cases} \int_\Omega \frac{s}{t} \ln \frac{s}{t} t\mathrm{d}\lambda & \text{if } s\mathrm{d}\lambda \ll t\mathrm{d}\lambda \\ +\infty & \text{otherwise.} \end{cases}$$

where $s\mathrm{d}\lambda \ll t\mathrm{d}\lambda$ means $\forall \Omega' \subset \Omega, \int_{\Omega'} t\mathrm{d}\lambda = 0 \implies \int_{\Omega'} s\mathrm{d}\lambda = 0$. Remark that, contrary to the quadratic loss, this divergence is an intrinsic quality measure between probability laws: it does not depend on the reference measure $\mathrm{d}\lambda$. However, the densities depend on this reference measure, this is stressed by the index λ when we work with the non intrinsic densities instead of the probability measures. As we study conditional densities and not classical densities, the previous divergence should be further adapted. To take into account the structure of conditional densities and the design of $(X_i)_{1\leq i\leq n}$, we use the following *tensorized* divergence:

$$KL_\lambda^{\otimes_n}(s,t) = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} KL_\lambda(s(\cdot|X_i), t(\cdot|X_i)) \right].$$

This divergence appears as the natural one in this setting and reduces to classical ones in specific settings:

- If the law of Y_i is independent of X_i, that is $s(\cdot|X_i) = s(\cdot)$ and $t(\cdot|X_i) = t(\cdot)$ do not depend on X_i, this divergence reduces to the classical $KL_\lambda(s,t)$.

- If the X_i's are not random but fixed, that is we consider a fixed design case, this divergence is the classical fixed design type divergence in which there is no expectation.
- If the X_i's are i.i.d., this divergence can be rewritten as $KL_\lambda^{\otimes_n}(s,t) = \mathbb{E}\left[KL_\lambda(s(\cdot|X_1), t(\cdot|X_1))\right]$.

Note that this divergence is an *integrated* divergence as it is the average over the index i of the mean with respect to the law of X_i of the divergence between the conditional densities for a given covariate value. Remark that more weight is given to regions of high density of the covariates than to regions of low density and, in particular, divergence values outside the supports of the X_i's are not used. When $\hat{s}$ is an estimator, or any function that depends on the observations, $KL_\lambda^{\otimes_n}(s,\hat{s})$ measures this (random) integrated divergence between s and $\hat{s}$ conditionally to the observations while $\mathbb{E}\left[KL_\lambda^{\otimes_n}(s,\hat{s})\right]$ is the average of this random quantity with respect to the observations.

As often in density estimation, we are not able to control this loss but only a smaller one. Namely, we use the Jensen-Kullback-Leibler divergence JKL_ρ with $\rho \in (0,1)$ defined by

$$JKL_\rho(s\mathrm{d}\lambda, t\mathrm{d}\lambda) = JKL_{\rho,\lambda}(s,t) = \frac{1}{\rho} KL_\lambda\left(s, (1-\rho)s + \rho t\right).$$

Note that this divergence appears explicitly with $\rho = \frac{1}{2}$ in Massart [34], but can also be found implicitly in Birgé and Massart [9] and Geer [19]. We use the name Jensen-Kullback-Leibler divergence in the same way Lin [32] use the name Jensen-Shannon divergence for a sibling in an information theory work. This divergence is smaller than the Kullback-Leibler one but larger, up to a constant factor, than the squared Hellinger one, $d_\lambda^2(s,t) = \int_\Omega |\sqrt{s} - \sqrt{t}|^2 \mathrm{d}\lambda$, and the squared L_1 distance, $\|s-t\|_{\lambda,1}^2 = \left(\int_\Omega |s-t|\mathrm{d}\lambda\right)^2$, as proved in our technical report [13]. More precisely, we use their tensorized counterparts:

$$d_\lambda^{2\otimes_n}(s,t) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n d_\lambda^2(s(\cdot|X'_i), t(\cdot|X'_i))\right] \quad \text{and} \quad JKL_{\rho,\lambda}^{\otimes_n}(s,t) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n JKL_{\rho,\lambda}(s(\cdot|X'_i), t(\cdot|X'_i))\right].$$

2.2 Penalty, bracketing entropy and Kraft inequality

Our condition on the penalty is given as a lower bound on its value:

$$\mathrm{pen}(m) \geq \kappa_0 \left(\mathfrak{D}_m + x_m\right)$$

where κ_0 is an absolute constant, $\mathfrak{D}_m$ is a quantity, depending only on the model S_m, that measures its complexity (and is often almost proportional to its dimension) while x_m is a non intrinsic coding term that depends on the structure of the whole model collection.

The complexity term $\mathfrak{D}_m$ is related to the bracketing entropy of the model S_m with respect to the Hellinger type divergence $d_\lambda^{\otimes_n}(s,t) = \sqrt{d_\lambda^{2\otimes_n}(s,t)}$, or more precisely to the bracketing entropies of its subsets $S_m(\widetilde{s},\sigma) = \left\{s_m \in S_m \middle| d_\lambda^{\otimes_n}(\widetilde{s}, s_m) \leq \sigma\right\}$. We recall that a bracket $[t^-, t^+]$ is a pair of functions such that $\forall (x,y) \in \mathcal{X} \times \mathcal{Y}, t^-(y|x) \leq t^+(y|x)$ and that a conditional density function s is said to belong to the bracket $[t^-, t^+]$ if $\forall (x,y) \in \mathcal{X} \times \mathcal{Y}, t^-(y|x) \leq s(y|x) \leq t^+(y|x)$. The bracketing entropy $H_{[\cdot],d_\lambda^{\otimes_n}}(\delta, S)$ of a set S is defined as the logarithm of the minimum number $N_{[\cdot],d_\lambda^{\otimes_n}}(\delta, S)$ of brackets $[t^-, t^+]$ of width $d_\lambda^{\otimes_n}(t^-, t^+)$ smaller than δ such that every function of S belongs to one of these brackets. To define $\mathfrak{D}_m$, we first impose a structural assumption:

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

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

Note that the function $\sigma \mapsto \int_0^\sigma \sqrt{H_{[\cdot],d_\lambda^{\otimes n}}(\delta, S_m)}\,\mathrm{d}\delta$ does always satisfy this assumption. $\mathfrak{D}_m$ is then defined as $n\sigma_m^2$ with σ_m^2 the unique root of $\dfrac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$. A good choice of ϕ_m is one which leads to a small upper bound of $\mathfrak{D}_m$. The bracketing entropy integral appearing in the assumption, often call Dudley integral, plays an important role in empirical processes theory, as stressed for instance in Vaart and Wellner [41]. The equation defining σ_m corresponds to an approximate optimization of a supremum bound as shown explicitly in the proof. This definition is obviously far from being very explicit but it turns out that it can be related to an entropic dimension of the model. Recall that the classical entropic dimension of a compact set S with respect to a metric d can be defined as the smallest real $\mathcal{D}$ such that there is a $\mathcal{C}$ such

$$\forall \delta > 0, H_d(\delta, S) \leq \mathcal{D}(\log\left(\frac{1}{\delta}\right) + \mathcal{C})$$

where H_d is the classical entropy with respect to metric d. Replacing the classical entropy by a bracketing one, we define the bracketing dimension $\mathcal{D}_m$ of a compact set as the smallest real $\mathcal{D}$ such that there is a $\mathcal{C}$ such

$$\forall \delta > 0, H_{[\cdot],d}(\delta, S) \leq \mathcal{D}(\log\left(\frac{1}{\delta}\right) + \mathcal{C}).$$

As hinted by the notation, for parametric model, under mild assumption on the parametrization, this bracketing dimension coincides with the usual one. It turns out that if this bracketing dimension exists then $\mathfrak{D}_m$ can be thought as roughly proportional to $\mathcal{D}_m$. More precisely, in our technical report [13], we obtain

Proposition 1.

- *if $\exists \mathcal{D}_m \geq 0, \exists \mathcal{C}_m \geq 0, \forall \delta \in (0, \sqrt{2}], H_{[\cdot],d_\lambda^{\otimes n}}(\delta, S_m) \leq \mathcal{V}_m + \mathcal{D}_m \ln \dfrac{1}{\delta}$ then*
  - *if $\mathcal{D}_m > 0$, (H_m) holds with a function ϕ_m such that $\mathfrak{D}_m \leq \left(2C_{\star,m} + 1 + \left(\ln \dfrac{n}{eC_{\star,m}\mathcal{D}_m}\right)_+\right)\mathcal{D}_m$ with $C_{\star,m} = \left(\sqrt{\frac{\mathcal{V}_m}{\mathcal{D}_m}} + \sqrt{\pi}\right)^2$,*
  - *if $\mathcal{D}_m = 0$, (H_m) holds with the function $\phi_m(\sigma) = \sigma\sqrt{\mathcal{V}_m}$ which is such $\mathfrak{D}_m = \mathcal{V}_m$,*
- *if $\exists \mathcal{D}_m \geq 0, \exists \mathcal{V}_m \geq 0, \forall \sigma \in (0, \sqrt{2}], \forall \delta \in (0, \sigma], H_{[\cdot],d_\lambda^{\otimes n}}(\delta, S_m(s_m, \sigma)) \leq \mathcal{V}_m + \mathcal{D}_m \ln \dfrac{\sigma}{\delta}$ then*
  - *if $\mathcal{D}_m > 0$, (H_m) holds with a function ϕ_m such that $\mathfrak{D}_m = C_{\star,m}\mathcal{D}_m$ with $C_{\star,m} = \left(\sqrt{\frac{\mathcal{V}_m}{\mathcal{D}_m}} + \sqrt{\pi}\right)^2$,*
  - *if $\mathcal{D}_m = 0$, (H_m) holds with the function $\phi_m(\sigma) = \sigma\sqrt{\mathcal{V}_m}$ which is such $\mathfrak{D}_m = \mathcal{V}_m$.*

We assume bounds on the entropy only for δ and σ smaller than $\sqrt{2}$, but, as for any conditional density pair (s, t) $d_\lambda^{\otimes n}(s, t) \leq \sqrt{2}$,

$$H_{[\cdot],d_\lambda^{\otimes n}}(\delta, S_m(s_m, \sigma)) = H_{[\cdot],d_\lambda^{\otimes n}}(\delta \wedge \sqrt{2}, S_m(s_m, \sigma \wedge \sqrt{2}))$$

which implies that those bounds are still useful when δ and σ are large.

The coding term x_m is constrained by a Kraft type assumption:

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

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

This condition is an information theory type condition and thus can be interpreted as a coding condition as stressed by Barron et al. [4].

2.3 A penalized maximum likelihood theorem

For technical reason, we also have to assume a separability condition on our models:

Assumption (Sep$_m$). *There exist a 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 a sequence $(t_k)_{k\geq 1}$ of elements of S'_m such that for every x and for every $y\in\mathcal{Y}'_m$, $\ln(t_k(y|x))$ goes to $\ln(t(y|x))$ as k goes to infinity.*

The main result of [12] is

Theorem 1. *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 model collection. Assume Assumption (K) holds while Assumptions (H$_m$) and (Sep$_m$) hold for every model $S_m\in\mathcal{S}$. Let $\widehat{s}_m$ be a η -log-likelihood minimizer 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 are two constants κ_0 and C_2 depending only on ρ and C_1 such that, as soon as for every index $m\in\mathcal{M}$

$$\mathrm{pen}(m) \geq \kappa\left(\mathfrak{D}_m + x_m\right) \quad \text{with } \kappa > \kappa_0$$

where $\mathfrak{D}_m = n\sigma_m^2$ with σ_m the unique root of $\frac{1}{\sigma}\phi_m(\sigma)=\sqrt{n}\sigma$, the penalized likelihood estimate $\widehat{s}_{\widehat{m}}$ with $\widehat{m}$ such that

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

satisfies

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

This theorem extends Theorem 7.11 of Massart [34], which handles only density estimation, and reduces to it if all conditional densites considered do not depend on the covariate. The cost of model selection with respect to the choice of the best single model is proved to be very mild. Indeed, let $\mathrm{pen}(m)=\kappa(\mathfrak{D}_m+x_m)$ then one obtains

$$\begin{aligned}
\mathbb{E}\left[JKL^{\otimes_n}_{\rho,\lambda}(s_0,\widehat{s}_{\widehat{m}})\right] &\leq C_1 \inf_{m\in\mathcal{M}}\left(\inf_{s_m\in S_m} KL^{\otimes_n}_{\lambda}(s_0,s_m) + \frac{\kappa}{n}(\mathfrak{D}_m+x_m)\right) + C_2\frac{\Sigma}{n} + \frac{\eta+\eta'}{n}\\
&\leq C_1\frac{\kappa}{\kappa_0}\left(\max_{m\in\mathcal{M}}\frac{\mathfrak{D}_m+x_m}{\mathfrak{D}_m}\right)\inf_{m\in\mathcal{M}}\left(\inf_{s_m\in S_m} KL^{\otimes_n}_{\lambda}(s_0,s_m) + \frac{\kappa_0}{n}\mathfrak{D}_m\right) + C_2\frac{\Sigma}{n} + \frac{\eta+\eta'}{n}.
\end{aligned}$$

where

$$\inf_{m\in\mathcal{M}}\left(\inf_{s_m\in S_m} KL_\lambda^{\otimes_n}(s_0,s_m)+\frac{\kappa_0}{n}\mathfrak{D}_m\right)+C_2\frac{\Sigma}{n}+\frac{\eta}{n}$$

is the best known bound for a generic single model, as explained in [12]: As soon as the term x_m remains small relatively to $\mathfrak{D}_m$, we have thus an oracle inequality: the penalized estimate satisfies up to a small factor the same bound as the estimate in the best model. The price to pay for the use of a collection of model is thus small. The gain is on the contrary huge: we do not have to know the best model within a collection to almost achieve its performance. Note that as there exists a constant $c_\rho > 0$ such that $c_\rho\|s-t\|_{\lambda,1}^{\otimes_n,2} \leq JKL_{\rho,\lambda}^{\otimes_n}(s,t)$, as proved in our technical report [13], this theorem implies a bound for the squared L_1 loss of the estimator.

For sake of generality, this theorem is relatively abstract. A natural question is the existence of interesting model collections that satisfy these assumptions. Motivated by an application to unsupervised hyperspectral image segmentation, already mentioned in [12], we consider the case where the covariate X belongs to $[0,1]^{d_X}$ and use collections for which the conditional densities depend on the covariate only in a piecewise constant manner.

3 Partition-based conditional density models

3.1 Covariate partitioning and conditional density estimation

Following an idea developed by Kolaczyk, Ju, and Gopal [30], we partition the covariate domain and consider candidate conditional density estimates that depend on the covariate only through the region it belongs. We are thus interested in conditional densities that can be written as

$$s(y|x)=\sum_{\mathcal{R}_l\in\mathcal{P}} s(y|\mathcal{R}_l)\mathbf{1}_{\{x\in\mathcal{R}_l\}}$$

where $\mathcal{P}$ is partition of $\mathcal{X}$, $\mathcal{R}_l$ denotes a generic region in this partition, $\mathbf{1}$ denotes the characteristic function of a set and $s(y|\mathcal{R}_l)$ is a density for any $\mathcal{R}_l\in\mathcal{P}$. Note that this strategy, called as in Willett and Nowak [42] partition-based, shares a lot with the CART-type strategy proposed by Donoho [15] in an image processing setting.

Denoting $\|\mathcal{P}\|$ the number of regions in this partition, the model we consider are thus specified by a partition $\mathcal{P}$ and a set $\mathcal{F}$ of $\|\mathcal{P}\|$-tuples of densities into which $(s(\cdot|\mathcal{R}_l))_{\mathcal{R}_l\in\mathcal{P}}$ is chosen. This set $\mathcal{F}$ can be a product of density sets, yielding an independent choice on each region of the partition, or have a more complex structure. We study two examples: in the first one, $\mathcal{F}$ is indeed a product of piecewise polynomial density sets, while in the second one $\mathcal{F}$ is a set of $\|\mathcal{P}\|$-tuples of Gaussian mixtures sharing the same mixture components. Nevertheless, denoting with a slight abuse of notation $S_{\mathcal{P},\mathcal{F}}$ such a model, our η-log-likelihood estimate in this model is any conditional density $\widehat{s}_{\mathcal{P},\mathcal{F}}$ such that

$$\left(\sum_{i=1}^n -\ln(\widehat{s}_{\mathcal{P},\mathcal{F}}(Y_i|X_i))\right)\leq \min_{s_{\mathcal{P},\mathcal{F}}\in S_{\mathcal{P},\mathcal{F}}}\left(\sum_{i=1}^n -\ln(s_{\mathcal{P},\mathcal{F}}(Y_i|X_i))\right)+\eta.$$

We first specify the partition collection we consider. For the sake of simplicity we restrict our description to the case where the covariate space $\mathcal{X}$ is simply $[0,1]^{d_X}$. We stress that the proposed strategy can easily be adapted to more general settings including discrete variable ordered or not. We impose a strong structural assumption on the partition collection considered that allows to control their *complexity*. We only consider five specific hyperrectangle based collections of partitions of $[0,1]^{d_X}$:

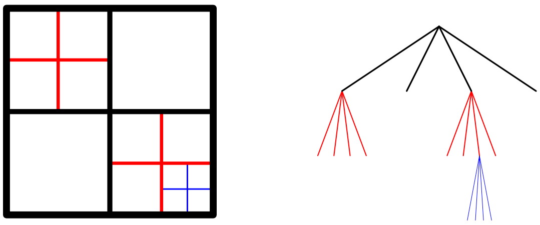

Figure 1: Example of a recursive dyadic partition with its associated dyadic tree.

- Two are recursive dyadic partition collections.
  - The uniform dyadic partition collection ($\mathrm{UDP}(\mathcal{X})$) in which all hypercubes are subdivided in 2^{d_X} hypercubes of equal size at each step. In this collection, in the partition obtained after J step, all the $2^{d_X J}$ hyperrectangles $\{\mathcal{R}_l\}_{1\leq l\leq\|\mathcal{P}\|}$ are thus hypercubes whose measure $|\mathcal{R}_l|$ satisfies $|\mathcal{R}_l| = 2^{-d_X J}$. We stop the recursion as soon as the number of steps J satisfies $\frac{2^{d_X}}{n} \geq |\mathcal{R}_l| \geq \frac{1}{n}$.
  - The recursive dyadic partition collection ($\mathrm{RDP}(\mathcal{X})$) in which at each step a hypercube of measure $|\mathcal{R}_l| \geq \frac{2^{d_X}}{n}$ is subdivided in 2^{d_X} hypercubes of equal size.
- Two are recursive split partition collections.
  - The recursive dyadic split partition ($\mathrm{RDSP}(\mathcal{X})$) in which at each step a hyperrectangle of measure $|\mathcal{R}_l| \geq \frac{2}{n}$ can be subdivided in 2 hyperrectangles of equal size by an even split along one of the d_X possible directions.
  - The recursive split partition ($\mathrm{RSP}(\mathcal{X})$) in which at each step a hyperrectangle of measure $|\mathcal{R}_l| \geq \frac{2}{n}$ can be subdivided in 2 hyperrectangles of measure larger than $\frac{1}{n}$ by a split along one a point of the grid $\frac{1}{n}\mathbb{Z}$ in one the d_X possible directions.
- The last one does not possess a hierarchical structure. The hyperrectangle partition collection ($\mathrm{HRP}(\mathcal{X})$) is the full collection of all partitions into hyperrectangles whose corners are located on the grid $\frac{1}{n}\mathbb{Z}^{d_X}$ and whose volume is larger than $\frac{1}{n}$.

We denote by $\mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}$ the corresponding partition collection where $\star(\mathcal{X})$ is either $\mathrm{UDP}(\mathcal{X})$, $\mathrm{RDP}(\mathcal{X})$, $\mathrm{RDSP}(\mathcal{X})$, $\mathrm{RSP}(\mathcal{X})$ or $\mathrm{HRP}(\mathcal{X})$.

As noticed by Kolaczyk and Nowak [31], Huang et al. [25] or Willett and Nowak [42], the first four partition collections, ($\mathcal{S}_{\mathcal{P}}^{\mathrm{UDP}(\mathcal{X})}$, $\mathcal{S}_{\mathcal{P}}^{\mathrm{RDP}(\mathcal{X})}$, $\mathcal{S}_{\mathcal{P}}^{\mathrm{RDSP}(\mathcal{X})}$, $\mathcal{S}_{\mathcal{P}}^{\mathrm{RSP}(\mathcal{X})}$), have a tree structure. Figure 1 illustrates this structure for a $\mathrm{RDP}(\mathcal{X})$ partition. This specific structure is mainly used to obtain an efficient numerical algorithm performing the model selection. For sake of completeness, we have also added the much more complex to deal with collection $\mathcal{S}_{\mathcal{P}}^{\mathrm{HRP}(\mathcal{X})}$, for which only exhaustive search algorithms exist.

As proved in our technical report [13], those partition collections satisfy Kraft type inequalities with weights constant for the $\mathrm{UDP}(\mathcal{X})$ partition collection and proportional to the number $\|\mathcal{P}\|$ of hyperrectangles for the other collections. Indeed,

Proposition 2. *For any of the five described partition collections $\mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}$, $\exists A_0^\star, B_0^\star, c_0^\star$ and Σ_0 such that for all $c \geq c_0^{\star(\mathcal{X})}$:*

$$\sum_{\mathcal{P}\in\mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}} e^{-c\left(A_0^{\star(\mathcal{X})}+B_0^{\star(\mathcal{X})}\|\mathcal{P}\|\right)} \leq \Sigma_0^{\star(\mathcal{X})} e^{-c\max\left(A_0^{\star(\mathcal{X})},B_0^{\star(\mathcal{X})}\right)}.$$

Those constants can be chosen as follow:

	$\star = \mathrm{UDP}(\mathcal{X})$	$\star = \mathrm{RDP}(\mathcal{X})$	$\star = \mathrm{RDSP}(\mathcal{X})$	$\star = \mathrm{RSP}(\mathcal{X})$	$\star = \mathrm{HRP}(\mathcal{X})$
$A_0^\star$	$\ln\left(\max\left(2, 1 + \frac{\ln n}{d_X \ln 2}\right)\right)$	0	0	0	0
$B_0^\star$	0	$\ln 2$	$\lceil \ln(1+d_X) \rceil_{\ln 2}$	$\lceil \ln(1+d_X) \rceil_{\ln 2}$ $+\lceil \ln n \rceil_{\ln 2}$	$d_X \lceil \ln n \rceil_{\ln 2}$
$c_0^\star$	0	$\frac{2^{d}_{X}}{2^{d}_{X}-1}$	2	2	1
$\Sigma_0^\star$	$1 + \frac{\ln n}{d_X \ln 2}$	2	$2(1+d_X)$	$4(1+d_X)n$	$(2n)^{d_X}$

where $\lceil x \rceil_{\ln 2}$ is the smallest multiple of $\ln 2$ larger than x. Furthermore, as soon as $c \geq 2\ln 2$ the right hand term of the bound is smaller than 1. This will prove useful to verify Assumption (K) for the model collections of the next sections.

In those sections, we study the two different choices proposed above for the set $\mathcal{F}$. We first consider a piecewise polynomial strategy similar to the one proposed by Willett and Nowak [42] defined for $\mathcal{Y} = [0,1]^{d_Y}$ in which the set $\mathcal{F}$ is a product of sets. We then consider a Gaussian mixture strategy with varying mixing proportion but common mixture components that extends the work of Maugis and Michel [35] and has been the original motivation of this work. In both cases, we prove that the penalty can be chosen roughly proportional to the dimension.

3.2 Piecewise polynomial conditional density estimation

In this section, we let $\mathcal{X} = [0,1]^{d_X}$, $\mathcal{Y} = [0,1]^{d_Y}$ and λ be the Lebesgue measure $\mathrm{d}y$. Note that, in this case, λ is a probability measure on $\mathcal{Y}$. Our candidate density $s(y|x \in \mathcal{R}_l)$ is then chosen among piecewise polynomial densities. More precisely, we reuse a hyperrectangle partitioning strategy this time for $\mathcal{Y} = [0,1]^{d_Y}$ and impose that our candidate conditional density $s(y|x \in \mathcal{R}_l)$ is a square of polynomial on each hyperrectangle $\mathcal{R}^{\mathcal{Y}}_{l,k}$ of the partition $\mathcal{Q}_l$. This differs from the choice of Willett and Nowak [42] in which the candidate density is simply a polynomial. The two choices coincide however when the polynomial is chosen among the constant ones. Although our choice of using squares of polynomial is less natural, it already ensures the positiveness of the candidates so that we only have to impose that the integrals of the piecewise polynomials are equal to 1 to obtain conditional densities. It turns out to be also crucial to obtain a control of the local bracketing entropy of our models. Note that this setting differs from the one of Blanchard et al. [10] in which $\mathcal{Y}$ is a finite discrete set.

We should now define the sets $\mathcal{F}$ we consider for a given partition $\mathcal{P} = \{\mathcal{R}_l\}_{1 \leq l \leq \|\mathcal{P}\|}$ of $\mathcal{X} = [0,1]^{d_X}$. Let $\mathbf{D} = (\mathbf{D}_1, \ldots, \mathbf{D}_{d_Y})$, we first define for any partition $\mathcal{Q} = \{\mathcal{R}^{\mathcal{Y}}_k\}_{1 \leq k \leq \|\mathcal{Q}\|}$ of $\mathcal{Y} = [0,1]^{d_Y}$ the set $\mathcal{F}_{\mathcal{Q},\mathbf{D}}$ of squares of piecewise polynomial densities of maximum degree $\mathbf{D}$ defined in the partition $\mathcal{Q}$:

$$\mathcal{F}_{\mathcal{Q},\mathbf{D}} = \left\{ s(y) = \sum_{\mathcal{R}^{\mathcal{Y}}_k \in \mathcal{Q}} P^2_{\mathcal{R}^{\mathcal{Y}}_k}(y) \mathbf{1}_{\{y \in \mathcal{R}^{\mathcal{Y}}_k\}} \middle| \begin{array}{l} \forall \mathcal{R}^{\mathcal{Y}}_k \in \mathcal{Q}, P_{\mathcal{R}^{\mathcal{Y}}_k} \text{ polynomial of degree at most } \mathbf{D}, \\ \sum_{\mathcal{R}^{\mathcal{Y}}_k \in \mathcal{Q}} \int_{\mathcal{R}^{\mathcal{Y}}_k} P^2_{\mathcal{R}^{\mathcal{Y}}_k}(y) = 1 \end{array} \right\}$$

For any partition collection $\mathcal{Q}^{\mathcal{P}} = (\mathcal{Q}_l)_{1 \leq l \leq \|\mathcal{P}\|} = \left(\{\mathcal{R}^{\mathcal{Y}}_{l,k}\}_{1 \leq k \leq \|\mathcal{Q}_l\|} \right)_{1 \leq l \leq \|\mathcal{P}\|}$ of $\mathcal{Y} = [0,1]^{d_Y}$, we can thus defined the set $\mathcal{F}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$ of $\|\mathcal{P}\|$-tuples of piecewise polynomial densities as

$$\mathcal{F}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} = \left\{ (s(\cdot|\mathcal{R}_l))_{\mathcal{R}_l \in \mathcal{P}} \middle| \forall \mathcal{R}_l \in \mathcal{P}, s(\cdot|\mathcal{R}_l) \in \mathcal{F}_{\mathcal{Q}_l,\mathbf{D}} \right\}.$$

The model $S_{\mathcal{P},\mathcal{F}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}}$, that is denoted $S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$ with a slight abuse of notation, is thus the set

$$S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} = \left\{ s(y|x) = \sum_{\mathcal{R}_l \in \mathcal{P}} s(y|\mathcal{R}_l) \mathbf{1}_{\{x \in \mathcal{R}_l\}} \middle| (s(y|\mathcal{R}_l))_{\mathcal{R}_l \in \mathcal{P}} \in \mathcal{F}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \right\}$$

$$= \left\{ s(y|x) = \sum_{\mathcal{R}_l \in \mathcal{P}} \sum_{\mathcal{R}^{\mathcal{Y}}_{l,k} \in \mathcal{Q}_l} P^2_{\mathcal{R}_l \times \mathcal{R}^{\mathcal{Y}}_{l,k}}(y) \mathbf{1}_{\left\{y \in \mathcal{R}^{\mathcal{Y}}_{l,k}\right\}} \mathbf{1}_{\{x \in \mathcal{R}_l\}} \;\middle|\; \begin{array}{c} \forall \mathcal{R}_l \in \mathcal{P}, \forall \mathcal{R}^{\mathcal{Y}}_{l,k} \in \mathcal{Q}_l, \\ P_{\mathcal{R}_l \times \mathcal{R}^{\mathcal{Y}}_{l,k}} \text{ polynomial of degree at most } \mathbf{D}, \\ \forall \mathcal{R}_l \in \mathcal{P}, \sum_{\mathcal{R}^{\mathcal{Y}}_{l,k} \in \mathcal{Q}_l} \int_{\mathcal{R}^{\mathcal{Y}}_{l,k}} P^2_{\mathcal{R}_l \times \mathcal{R}^{\mathcal{Y}}_{l,k}}(y) = 1 \end{array} \right\}$$

Denoting $\mathcal{R}^{\times}_{l,k}$ the product $\mathcal{R}_l \times \mathcal{R}^{\mathcal{Y}}_{l,k}$, the conditional densities of the previous set can be advantageously rewritten as

$$s(y|x) = \sum_{\mathcal{R}_l \in \mathcal{P}} \sum_{\mathcal{R}^{\mathcal{Y}}_{l,k} \in \mathcal{Q}_l} P^2_{\mathcal{R}^{\times}_{l,k}}(y) \mathbf{1}_{\left\{(x,y) \in \mathcal{R}^{\times}_{l,k}\right\}}$$

As shown by Willett and Nowak [42], the maximum likelihood estimate in this model can be obtained by an independent computation on each subset $\mathcal{R}^{\times}_{l,k}$:

$$\widehat{P}_{\mathcal{R}^{\times}_{l,k}} = \frac{\sum_{i=1}^n \mathbf{1}_{\left\{(X_i,Y_i) \in \mathcal{R}^{\times}_{l,k}\right\}}}{\sum_{i=1}^n \mathbf{1}_{\{X_i \in \mathcal{R}_l\}}} \operatorname*{argmin}_{P, \deg(P) \leq \mathbf{D}, \int_{\mathcal{R}^{\mathcal{Y}}_{l,k}} P^2(y) \mathrm{d}y = 1} \sum_{i=1}^n \mathbf{1}_{\left\{(X_i,Y_i) \in \mathcal{R}^{\times}_{l,k}\right\}} \ln\left(P^2(Y_i)\right).$$

This property is important to be able to use the efficient optimization algorithms of Willett and Nowak [42] and Huang et al. [25].

Our model collection is obtained by considering all partitions $\mathcal{P}$ within one of the UDP($\mathcal{X}$), RDP($\mathcal{X}$), RDSP($\mathcal{X}$), RSP($\mathcal{X}$) or HRP($\mathcal{X}$) partition collections with respect to $[0,1]^{d_X}$ and, for a fixed $\mathcal{P}$, all partitions $\mathcal{Q}_l$ within one of the UDP($\mathcal{Y}$), RDP($\mathcal{Y}$), RDSP($\mathcal{Y}$), RSP($\mathcal{Y}$) or HRP($\mathcal{Y}$) partition collections with respect to $[0,1]^{d_Y}$. By construction, in any cases,

$$\dim(S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}) = \sum_{\mathcal{R}_l \in \mathcal{P}} \left(\|\mathcal{Q}_l\| \prod_{d=1}^{d_Y} (\mathbf{D}_d + 1) - 1 \right).$$

To define the penalty, we use a slight upper bound of this dimension

$$\mathcal{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} = \sum_{\mathcal{R}_l \in \mathcal{P}} \|\mathcal{Q}_l\| \prod_{d=1}^{d_Y} (\mathbf{D}_d + 1) = \|\mathcal{Q}^{\mathcal{P}}\| \prod_{d=1}^{d_Y} (\mathbf{D}_d + 1)$$

where $\|\mathcal{Q}^{\mathcal{P}}\| = \sum_{\mathcal{R}_l \in \mathcal{P}} \|\mathcal{Q}_l\|$. is the total number of hyperrectangles in all the partitions:

Theorem 2. *Fix a collection $\star(\mathcal{X})$ among* UDP($\mathcal{X}$), RDP($\mathcal{X}$), RDSP($\mathcal{X}$), RSP($\mathcal{X}$) *or* HRP($\mathcal{X}$) *for $\mathcal{X} = [0,1]^{d_X}$, a collection $\star(\mathcal{Y})$ among* UDP($\mathcal{Y}$), RDP($\mathcal{Y}$), RDSP($\mathcal{Y}$), RSP($\mathcal{Y}$) *or* HRP($\mathcal{Y}$) *and a maximal degree for the polynomials $\mathbf{D} \in \mathbb{N}^{d_Y}$.*

Let

$$\mathcal{S} = \left\{ S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \middle| \mathcal{P} = \{\mathcal{R}_l\} \in \mathcal{S}^{\star(\mathcal{X})}_{\mathcal{P}} \text{ and } \forall \mathcal{R}_l \in \mathcal{P}, \mathcal{Q}_l \in \mathcal{S}^{\star(\mathcal{Y})}_{\mathcal{P}} \right\}.$$

Then there exist a $C_\star > 0$ and a $c_\star > 0$ independent of n, such that for any ρ and for any $C_1 > 1$, the penalized estimator of Theorem 1 satisfies

$$\mathbb{E}\left[JKL^{\otimes_n}_{\rho,\lambda}(s_0, \widehat{s}_{\widehat{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}}) \right] \leq C_1 \inf_{S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \in \mathcal{S}} \left(\inf_{s_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \in S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}} KL^{\otimes_n}_{\lambda}(s_0, s_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}) + \frac{\mathrm{pen}(\mathcal{Q}^{\mathcal{P}},\mathbf{D})}{n} \right)$$
$$+ C_2 \frac{1}{n} + \frac{\eta + \eta'}{n}$$

as soon as

$$\mathrm{pen}(\mathcal{Q}^{\mathcal{P}},\mathbf{D}) \geq \widetilde{\kappa}\,\mathcal{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$$

for

$$\widetilde{\kappa} > \kappa_0\left(C_\star + c_\star\left(A_0^{\star(\mathcal{X})} + B_0^{\star(\mathcal{X})} + A_0^{\star(\mathcal{Y})} + B_0^{\star(\mathcal{Y})}\right) + 2\ln n\right).$$

where κ_0 and C_2 are the constants of Theorem 1 that depend only on ρ and C_1. Furthermore $C_\star \leq \frac{1}{2}\ln(8\pi e) + \sum_{d=1}^{d_Y}\ln\left(\sqrt{2}(\mathbf{D}_d+1)\right)$ and $c_\star \leq 2\ln 2$.

A penalty chosen proportional to the dimension of the model, the multiplicative factor $\widetilde{\kappa}$ being constant over n up to a logarithmic factor, is thus sufficient to guaranty the estimator performance. Furthermore, one can use a penalty which is a sum of penalties for each hyperrectangle of the partition:

$$\mathrm{pen}(\mathcal{Q}^{\mathcal{P}},\mathbf{D}) = \sum_{\mathcal{R}^{\times}_{l,k}\in\mathcal{Q}^{\mathcal{P}}} \widetilde{\kappa}\left(\prod_{d=1}^{d_Y}(\mathbf{D}_d+1)\right).$$

This additive structure of the penalty allows to use the fast partition optimization algorithm of Donoho [15] and Huang et al. [25] as soon as the partition collection is tree structured.

In Appendix, we obtain a weaker requirement on the penalty

$$\begin{aligned}\mathrm{pen}(\mathcal{Q}^{\mathcal{P}},\mathbf{D}) \geq \kappa\Bigg(&\left(C_\star + 2\ln\frac{n}{\sqrt{\|\mathcal{Q}^{\mathcal{P}}\|}}\right)\mathcal{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}\\ &+ c_\star\left(A_0^{\star(\mathcal{X})} + \left(B_0^{\star(\mathcal{X})} + A_0^{\star(\mathcal{Y})}\right)\|\mathcal{P}\| + B_0^{\star(\mathcal{Y})}\sum_{\mathcal{R}_l\in\mathcal{P}}\|\mathcal{Q}_l\|\right)\Bigg)\end{aligned}$$

in which the complexity part and the coding part appear more explicitly. This smaller penalty is no longer proportional to the dimension but still sufficient to guaranty the estimator performance. Using the crude bound $\|\mathcal{Q}^{\mathcal{P}}\| \geq 1$, one sees that such a penalty penalty can still be upper bounded by a sum of penalties over each hyperrectangle. The loss with respect to the original penalty is of order $\kappa\log\|\mathcal{Q}^{\mathcal{P}}\|\mathcal{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$, which is negligible as long as the number of hyperrectangle remains small with respect to n^2.

Some variations around this Theorem can be obtained through simple modifications of its proof as explained in Appendix. For example, the term $2\ln(n/\sqrt{\|\mathcal{Q}^{\mathcal{P}}\|})$ disappears if $\mathcal{P}$ belongs to $\mathcal{S}_{\mathcal{P}}^{\mathrm{UDP}(\mathcal{X})}$ while $\mathcal{Q}_l$ is independent of $\mathcal{R}_l$ and belongs to $\mathcal{S}_{\mathcal{P}}^{\mathrm{UDP}(\mathcal{X})}$. Choosing the degrees $\mathbf{D}$ of the polynomial among a family $\mathcal{D}^M$ either globally or locally as proposed by Willett and Nowak [42] is also possible. The constant $C_\star$ is replaced by its maximum over the family considered, while the coding part is modified by replacing respectively $A_0^{\star(\mathcal{X})}$ by $A_0^{\star(\mathcal{X})} + \ln|\mathcal{D}^M|$ for a global optimization and $B_0^{\star(\mathcal{Y})}$ by $B_0^{\star(\mathcal{Y})} + \ln|\mathcal{D}^M|$ a the local optimization. Such a penalty can be further modified into an additive one with only minor loss. Note that even if the family and its maximal degree grows with n, the *constant* $C_\star$ grows at a logarithic rate in n as long as the maximal degree grows at most polynomially with n.

Finally, if we assume that the true conditional density is lower bounded, then

$$KL_\lambda^{\otimes_n}(s,t) \leq \left\|\frac{1}{t}\right\|_\infty \|s-t\|_{\lambda,2}^{\otimes_n,2}$$

as shown by Kolaczyk and Nowak [31]. We can thus reuse ideas from Willett and Nowak [42], Akakpo [1] or Akakpo and Lacour [2] to infer the quasi optimal minimaxity of this estimator for anisotropic Besov spaces (see for instance in Karaivanov and Petrushev [28] for a definition) whose regularity indices are smaller than 1 along the axes of $\mathcal{X}$ and smaller than $\mathbf{D}+1$ along the axes of $\mathcal{Y}$.

3.3 Spatial Gaussian mixtures, models, bracketing entropy and penalties

In this section, we consider an extension of Gaussian mixture that takes account into the covariate into the mixing proportion. This model has been motivated by the unsupervised hyperspectral image segmentation problem mentioned in the introduction. We recall first some basic facts about Gaussian mixtures and their uses in unsupervised classification.

In a classical Gaussian mixture model, the observations are assuming to be drawn from several different classes, each class having a Gaussian law. Let K be the number of different Gaussians, often call the number of clusters, the density s_0 of Y_i with respect to the Lebesgue measure is thus modeled as

$$s_{K,\theta,\pi}(\cdot) = \sum_{k=1}^{K} \pi_k \Phi_{\theta_k}(\cdot)$$

where

$$\Phi_{\theta_k}(y) = \frac{1}{(2\pi \det \Sigma_k)^{p/2}} e^{-\frac{1}{2}(y-\mu_k)'\Sigma_k^{-1}(y-\mu_k)}$$

with μ_k the mean of the kth component, Σ_k its covariance matrix, $\theta_k = (\mu_k, \Sigma_k)$ and π_k its mixing proportion. A model $S_{K,\mathcal{G}}$ is obtained by specifying the number of component K as well as a set $\mathcal{G}$ to which should belong the K-tuple of Gaussian $(\Phi_{\theta_1}, \ldots, \Phi_{\theta_K})$. Those Gaussians can share for instance the same shape, the same volume or the same diagonalization basis. The classical choices are described for instance in Biernacki et al. [7]. Using the EM algorithm, or one of its extension, one can efficiently obtain the proportions $\widehat{\pi}_k$ and the Gaussian parameters $\widehat{\theta}_k$ of the maximum likelihood estimate within such a model. Using tools also derived from Massart [34], Maugis and Michel [35] show how to choose the number of classes by a penalized maximum likelihood principle. These Gaussian mixture models are often used in unsupervised classification application: one observes a collection of Y_i and tries to split them into homogeneous classes. Those classes are chosen as the Gaussian components of an estimated Gaussian mixture close to the density of the observations. Each observation can then be assigned to a class by a simple maximum likelihood principle:

$$\widehat{k}(y) = \underset{1 \leq k \leq \widehat{K}}{\operatorname{argmax}} \, \widehat{\pi}_k \Phi_{\widehat{\theta}_k}(y).$$

This methodology can be applied directly to an hyperspectral image and yields a segmentation method, often called spectral method in the image processing communit. This method however fails to exploit the spatial organization of the pixels.

To overcome this issue, Kolaczyk, Ju, and Gopal [30] and Antoniadis, Bigot, and Sachs [3] propose to use mixture model in which the mixing proportions depend on the covariate X_i while the mixture components remain constant. We propose to estimate simultaneously those mixing proportions and the mixture components with our partition-based strategy. In a semantic analysis context, in which documents replace pixels, a similar Gaussian mixture with varying weight, but without the partition structure, has been proposed by Si and Jin [38] as an extension of a general mixture based semantic analysis model introduced by Hofmann [24] under the name *Probabilistic Latent Semantic Analysis.* A similar model has also been considered in the work of Young and Hunter [43]. In our approach, for a given partition $\mathcal{P}$, the conditional density $s(\cdot|x)$ are modeled as

$$s_{\mathcal{P},K,\theta,\pi}(\cdot|x) = \sum_{\mathcal{R}_l \in \mathcal{P}} \left(\sum_{k=1}^{K} \pi_k[\mathcal{R}_l] \Phi_{\theta_k}(\cdot) \right) \mathbf{1}_{\{x \in \mathcal{R}_l\}}$$

which, denoting $\pi[\mathcal{R}(x)] = \sum_{\mathcal{R}_l \in \mathcal{P}} \pi[\mathcal{R}_l] \, \mathbf{1}_{\{x \in \mathcal{R}_l\}}$, can advantageously be rewritten

$$= \sum_{k=1}^{K} \pi_k[\mathcal{R}(x)] \Phi_{\theta_k}(\cdot) \, .$$

The K-tuples of Gaussian can be chosen is the same way as in the classical Gaussian mixture case. Using a penalized maximum likelihood strategy, a partition $\widehat{\mathcal{P}}$, a number of Gaussian components $\widehat{K}$, their parameters $\widehat{\theta}_k$ and all the mixing proportions $\widehat{\pi}[\widehat{\mathcal{R}}_l]$ can be estimated. Each pair of pixel position and spectrum (x, y) can then be assigned to one of the estimated mixture components by a maximum likelihood principle:

$$\widehat{k}(x, y) = \operatorname*{argmax}_{1 \leq k \leq \widehat{K}} \widehat{\pi}_k[\widehat{\mathcal{R}_l}(x)] \Phi_{\widehat{\theta_k}}(y).$$

This is the strategy we have used at IPANEMA [6] to segment, in an unsupervised manner, hyperspectral images. In these images, a spectrum Y_i, with around 1000 frequency bands, is measured at each pixel location X_i and our aim was to derive a partition in *homogeneous* regions without any human intervention. This is a precious help for users of this imaging technique as this allows to focus the study on a few representative spectrums. Combining the classical EM strategy for the Gaussian parameter estimation (see for instance Biernacki et al. [7]) and dynamic programming strategies for the partition, as described for instance by Kolaczyk, Ju, and Gopal [30], we have been able to implement this penalized estimator and to test it on real datasets. Figure 2 illustrates this methodology. The studied sample is a thin cross-section of maple with a single layer of hide glue on top of it, prepared recently using materials and processes from the Cité de la Musique, using materials of the same type and quality that is used for lutherie. We present here the result for a low signal to noise ratio acquisition requiring only two minutes of scan. Using piecewise constant mixing proportions instead of constant mixing proportions leads to a better geometry of the segmentation, with less isolated points and more structured boundaries. As described in a more applied study [14], this methodology permits to work with a much lower signal to noise ratio and thus allows to reduce significantly the acquisition time.

We should now specify the models we consider. As we follow the construction of Section 3.1, for a given segmentation $\mathcal{P}$, this amounts to specify the set $\mathcal{F}$ to which belong the $\|\mathcal{P}\|$-tuples of densities $(s(y|\mathcal{R}_l))_{\mathcal{R}_l \in \mathcal{P}}$. As described above, we assume that $s(y|\mathcal{R}_l) = \sum_{k=1}^{K} \pi_k[\mathcal{R}_l] \Phi_{\theta_k}(y)$. The mixing proportions within the region $\mathcal{R}_l$, $\pi[\mathcal{R}_l]$, are chosen freely among all vectors of the $K-1$ dimensional simplex $\mathcal{S}_{K-1}$:

$$\mathcal{S}_{K-1} = \left\{ \pi = (\pi_1, \dots, \pi_k) \middle| \forall k, 1 \leq k \leq K, \pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1 \right\} .$$

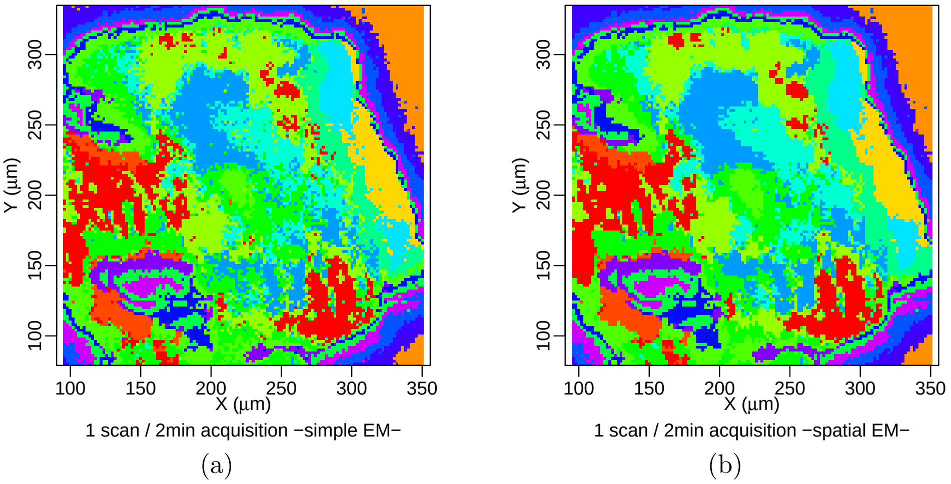

Figure 2: Unsupervised segmentation result: a) with constant mixing proportions b) with piecewise constant mixing proportions.

As we assume the mixture components are the same in each region, for a given number of components K, the set $\mathcal{F}$ is entirely specified by the set $\mathcal{G}$ of K-tuples of Gaussian $(\Phi_{\theta_1}, \ldots, \Phi_{\theta_K})$ (or equivalently by a set Θ for $\theta = (\theta_1, \ldots, \theta_K)$).

To allow variable selection, we follow Maugis and Michel [35] and let E be an arbitrary subspace of $\mathcal{Y} = \mathbb{R}^p$, that is expressed differently for the different classes, and let $E^\perp$ be its orthogonal, in which all classes behave similarly. We assume thus that

$$\Phi_{\theta_k}(y) = \Phi_{\theta_{E,k}}(y_E)\Phi_{\theta_{E^\perp}}(y_{E^\perp})$$

where y_E and $y_{E^\perp}$ denote, respectively, the projection of y on E and $E^\perp$, $\Phi_{\theta_{E,k}}$ is a Gaussian whose parameters depend on k while $\Phi_{\theta_{E^\perp}}$ is independent of k. A model is then specified by the choice of a set $\mathcal{G}_E^K$ for the K-tuples $(\Phi_{\theta_{E,1}}, \ldots, \Phi_{\theta_{E,K}})$ (or equivalently a set Θ_E^K for the K-tuples of parameters $(\theta_{E,1}, \ldots, \theta_{E,K})$) and a set $\mathcal{G}_{E^\perp}$ for the Gaussian $\Phi_{\theta_{E^\perp}}$ (or equivalently a set $\Theta_{E^\perp}$ for its parameter $\theta_{E^\perp}$). The resulting model is denoted $S_{\mathcal{P},K,\mathcal{G}}$

$$S_{\mathcal{P},K,\mathcal{G}} = \left\{ s_{\mathcal{P},K,\theta,\pi}(y|x) = \sum_{k=1}^{K} \pi_k[\mathcal{R}(x)]\, \Phi_{\theta_{E,k}}(y_E)\ \Phi_{\theta_{E^\perp}}(y_{E^\perp}) \,\middle|\, \begin{array}{l} (\Phi_{\theta_{E,1}}, \ldots, \Phi_{\theta_{E,K}}) \in \mathcal{G}_E^K, \\ \Phi_{\theta_{E^\perp}} \in \mathcal{G}_{E^\perp}, \\ \forall \mathcal{R}_l \in \mathcal{P}, \pi[\mathcal{R}_l] \in \mathcal{S}_{K-1} \end{array} \right\}.$$

The sets $\mathcal{G}_E^K$ and $\mathcal{G}_{E^\perp}$ are chosen among the *classical* Gaussian K-tuples, as described for instance in Biernacki et al. [7]. For a space E of dimension p_E and a fixed number K of classes, we specify the set

$$\mathcal{G} = \left\{ (\Phi_{E,\theta_1}, \ldots, \Phi_{E,\theta_K}) \middle| \theta = (\theta_1, \ldots, \theta_K) \in \Theta_{[\cdot]_{p_E}^K} \right\}$$

through a parameter set $\Theta_{[\cdot]_{p_E}^K}$ defined by some (mild) constraints on the means μ_k and some (strong) constraints on the covariance matrices Σ_k.

The K-tuple of means $\mu = (\mu_1, \ldots, \mu_K)$ is either known or unknown without any restriction. A stronger structure is imposed on the K-tuple of covariance matrices $(\Sigma_1, \ldots, \Sigma_K)$. To define it, we need to introduce a decomposition of any covariance matrix Σ into $LDAD'$ where, denoting

$|\Sigma|$ the determinant of Σ, $L = |\Sigma|^{1/p_E}$ is a positive scalar corresponding to the volume, D is the matrix of eigenvectors of Σ and A the diagonal matrix of renormalized eigenvalues of Σ (the eigenvalues of $|\Sigma|^{-1/p_E}\Sigma$). Note that this decomposition is not unique as, for example, D and A are defined up to a permutation. We impose nevertheless a structure on the K-tuple $(\Sigma_1, \ldots, \Sigma_K)$ through structures on the corresponding K-tuples of $(L_1, \ldots, L_K)$, $(D_1, \ldots, D_K)$ and $(A_1, \ldots, A_K)$. They are either known, unknown but with a common value or unknown without any restriction. The corresponding set is indexed by $[\mu_\star\,\mathrm{L}_\star\,\mathrm{D}_\star\,\mathrm{A}_\star]^K_{p_E}$ where $\star = 0$ means that the quantity is known, $\star = K$ that the quantity is unknown without any restriction and possibly different for every class and its lack means that there is a common unknown value over all classes.

To have a set with finite bracketing entropy, we further restrict the values of the means μ_k, the volumes L_k and the renormalized eigenvalue matrix A_k. The means are assumed to satisfy $\forall 1 \leq k \leq K, |\mu_k| \leq a$ for a known a while the volumes satisfy $\forall 1 \leq k \leq K, L_- \leq L_k \leq L_+$ for some known positive values L_- and L_+. To describe the constraints on the renormalized eigenvalue matrix A_k, we define the set $\mathcal{A}(\lambda_-, \lambda_+, p_E)$ of diagonal matrices A such that $|A| = 1$ and $\forall 1 \leq i \leq p_E, \lambda_- \leq A_{i,i} \leq \lambda_+$. Our assumption is that all the A_k belong to $\mathcal{A}(\lambda_-, \lambda_+, p_E)$ for some known values λ_- and λ_+.

Among the $3^4 = 81$ such possible sets, six of them have been already studied by Maugis and Michel [35, 36] in their classical Gaussian mixture model analysis: $[\mu_0\,\mathrm{L}_K\,\mathrm{D}_0\,\mathrm{A}_0]^K_{p_E}$, $[\mu_K\,\mathrm{L}_K\,\mathrm{D}_0\,\mathrm{A}_K]^K_{p_E}$, $[\mu_K\,\mathrm{L}_K\,\mathrm{D}_K\,\mathrm{A}_K]^K_{p_E}$, $[\mu_K\,\mathrm{L}\,\mathrm{D}_0\,\mathrm{A}_K]^K_{p_E}$, $[\mu_K\,\mathrm{L}\,\mathrm{D}_0\,\mathrm{A}]^K_{p_E}$ and $[\mu_K\,\mathrm{L}\,\mathrm{D}\,\mathrm{A}]^K_{p_E}$. All these cases, as well as the others, are covered by our analysis with a single proof.

To summarize, our models $S_{\mathcal{P},K,\mathcal{G}}$ are parametrized by a partition $\mathcal{P}$, a number of components K, a set $\mathcal{G}$ of K-tuples of Gaussian specified by a space E and two parameter sets, a set $\Theta_{[\mu_\star\,\mathrm{L}_\star\,\mathrm{D}_\star\,\mathrm{A}_\star]^K_{p_E}}$ of K-tuples of Gaussian parameters for the differentiated space E and a set $\Theta_{[\mu_\star\,\mathrm{L}_\star\,\mathrm{D}_\star\,\mathrm{A}_\star]_{p_{E^\perp}}}$ of Gaussian parameters for its orthogonal $E^\perp$. Those two sets are chosen among the ones described above with the same constants a, L_-, L_+, λ_- and λ_+. One verifies that

$$\dim(S_{\mathcal{P},K,\mathcal{G}}) = \|\mathcal{P}\|(K-1) + \dim\left(\Theta_{[\mu_\star\,\mathrm{L}_\star\,\mathrm{D}_\star\,\mathrm{A}_\star]^K_{p_E}}\right) + \dim\left(\Theta_{[\mu_\star\,\mathrm{L}_\star\,\mathrm{D}_\star\,\mathrm{A}_\star]_{p_{E^\perp}}}\right).$$

Before stating a model selection theorem, we should specify the collections $\mathcal{S}$ considered. We consider sets of model $S_{\mathcal{P},K,\mathcal{G}}$ with $\mathcal{P}$ chosen among one of the partition collections $\mathcal{S}^\star_{\mathcal{P}}$, K smaller than K_M, which can be theoretically chosen equal to $+\infty$, a space E chosen as $\mathrm{span}\{e_i\}_{i \in I}$ where e_i is the canonical basis of $\mathbb{R}^p$ and I a subset of $\{1, \ldots, p\}$ is either known, equal to $\{1, \ldots, p_E\}$ or free and the indices $[\mu_\star\,\mathrm{L}_\star\,\mathrm{D}_\star\,\mathrm{A}_\star]$ of Θ_E and $\Theta_{E^\perp}$ are chosen freely among a subset of the possible combinations.

Without any assumptions on the design, we obtain

Theorem 3. *Assume the collection $\mathcal{S}$ is one of the collections of the previous paragraph.*

Then, there exist a $C_\star > \pi$ and a $c_\star > 0$, such that, for any ρ and for any $C_1 > 1$, the penalized estimator of Theorem 1 satisfies

$$\mathbb{E}\left[JKL^{\otimes_n}_{\rho,\lambda}(s_0, \widehat{s}_{\widehat{\mathcal{P},K,\mathcal{G}}})\right] \leq C_1 \inf_{S_{\mathcal{P},K,\mathcal{G}} \in \mathcal{S}} \left(\inf_{s_{\mathcal{P},K,\mathcal{G}} \in S_{\mathcal{P},K,\mathcal{G}}} KL^{\otimes_n}_{\lambda}(s_0, s_{\mathcal{P},K,\mathcal{G}}) + \frac{\mathrm{pen}(\mathcal{P},K,\mathcal{G})}{n} \right) + \frac{C_2}{n} + \frac{\eta + \eta'}{n}$$

as soon as

$$\mathrm{pen}(\mathcal{P},K,\mathcal{G}) \geq \widetilde{\kappa}_1 \dim(S_{\mathcal{P},K,\mathcal{G}}) + \widetilde{\kappa}_2 \mathcal{D}_E$$

for

$$\widetilde{\kappa}_1 \geq \kappa \left(\left(2C_\star + 1 + \left(\ln \frac{n}{eC_\star} \right)_+ + c_\star \left(A_0^{\star(\mathcal{X})} + B_0^{\star(\mathcal{X})} + 1 \right) \right) \right) \qquad \textit{and} \qquad \widetilde{\kappa}_2 \geq \kappa c_\star$$

with $\kappa > \kappa_0$ where κ_0 and C_2 are the constants of Theorem 1 that depend only on ρ and C_1 and

$$\mathcal{D}_E = \begin{cases} 0 & \textit{if } E \textit{ is known,} \\ p_E & \textit{if } E \textit{ is chosen among spaces spanned by} \\ & \textit{the first coordinates,} \\ (1 + \ln 2 + \ln \frac{p}{p_E}) p_E & \textit{if } E \textit{ is free.} \end{cases}$$

As in the previous section, the penalty term can thus be chosen, up to the variable selection term $\mathcal{D}_E$, proportional to the dimension of the model, with a proportionality factor constant up to a logarithmic term with n. A penalty proportional to the dimension of the model is thus sufficient to ensure that the model selected performs almost as well as the best possible model in term of conditional density estimation. As in the proof of Antoniadis, Bigot, and Sachs [3], we can also obtain that our proposed estimator yields a minimax estimate for spatial Gaussian mixture with mixture proportions having a geometrical regularity even without knowing the number of classes.

Moreover, again as in the previous section, the penalty can have an additive structure, it can be chosen as a sum of penalties over each hyperrectangle plus one corresponding to K and the set $\mathcal{G}$. Indeed

$$\operatorname{pen}(\mathcal{P}, K, \mathcal{G}) = \sum_{\mathcal{R}_l \in \mathcal{P}} \widetilde{\kappa}_1 (K-1) + \widetilde{\kappa}_1 \left(\dim \left(\Theta_{[\mu_\star \, \mathrm{L}_\star \, \mathrm{D}_\star \, \mathrm{A}_\star]^K_{p_E}} \right) + \dim \left(\Theta_{[\mu_\star \, \mathrm{L}_\star \, \mathrm{D}_\star \, \mathrm{A}_\star]_{p_{E^\perp}}} \right) \right) + \widetilde{\kappa}_2 \mathcal{D}_E$$

satisfies the requirement of Theorem 3. This structure is the key for our numerical minimization algorithm in which one optimizes alternately the Gaussian parameters with an EM algorithm and the partition with the same fast optimization strategy as in the previous section.

In Appendix, we obtain a weaker requirement

$$\begin{aligned} \operatorname{pen}(\mathcal{P}, K, \mathcal{G}) \geq \kappa \Bigg(& \left(2C_\star + 1 + \left(\ln \frac{n}{e C_\star \dim(S_{\mathcal{P},K,\mathcal{G}})} \right)_+ \right) \dim(S_{\mathcal{P},K,\mathcal{G}}) \\ & + c_\star \left(A_0^{\star(\mathcal{X})} + B_0^{\star(\mathcal{X})} \|\mathcal{P}\| + (K-1) + \mathcal{D}_E \right) \Bigg) \end{aligned}$$

in which the complexity and the coding terms are more explicit. Again up to a logarithmic term in $\dim(S_{\mathcal{P},K,\mathcal{G}})$, this requirement can be satisfied by a penalty having the same additive structure as in the previous paragraph.

Our theoretical result on the conditional density estimation does not guaranty good segmentation performance. If data are generated according to a Gaussian mixture with varying mixing proportions, one could nevertheless obtain the asymptotic convergence of our class estimator to the optimal Bayes one. We have nevertheless observed in our numerical experiments at IPANEMA that the proposed methodology allow to reduce the signal to noise ratio while keeping meaningful segmentations.

Two major questions remain nevertheless open. Can we calibrate the penalty (choosing the constants) in a datadriven way while guaranteeing the theoretical performance in this specific setting? Can we derive a non asymptotic classification result from this conditional density result? The *slope heuristic*, proposed by Birgé and Massart [8], we have used in our numerical experiments, seems a promising direction. Deriving a theoretical justification in this conditional estimation setting would be much better. Linking the non asymptotic estimation behavior to a non asymptotic classification behavior appears even more challenging.

A Proof for Section 3.2 (Piecewise polynomial conditional density estimation)

Theorem 2 is obtained by proving that Assumption ($\mathrm{H}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$) and ($\mathrm{S}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$) hold for any model $S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$ while Assumption (K) holds for any model collection. Theorem 2 is then a consequence of Theorem 1.

One easily verifies that Assumption ($\mathrm{S}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$) holds whatever the partition choice. Concerning the first assumption,

Proposition 3. *Under the assumptions of Theorem 2, there exists a $D_\star$ such that for any model $S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$ Assumption ($H_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$) is satisfied with a function ϕ such that*

$$\mathfrak{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \leq \left(C_\star + \ln \frac{n^2}{\|\mathcal{Q}^{\mathcal{P}}\|} \right) \mathcal{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$$

with $C_\star = 2D_\star + 2\pi$.

The proof relies on the combination of Proposition 1 and

Proposition 4. $\forall S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}, \forall s_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \in S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$,

$$H_{[\cdot],d^{\otimes_n}}\left(\delta, S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}(s_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}, \sigma)\right) \leq \mathcal{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \left(\frac{1}{2} \ln \frac{n^2}{\|\mathcal{Q}^{\mathcal{P}}\|} + D_\star + \ln \frac{\sigma}{\delta} \right).$$

By using Proposition 2 for both $\mathcal{P}$ and $\mathcal{Q}$, we obtain the Kraft type assumption:

Proposition 5. *Under the assumptions of Theorem 2, for any collection $\mathcal{S}$, there exists a $c_\star > 0$ such that for*

$$x_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} = c_\star \left(A_0^{\star(\mathcal{X})} + \left(B_0^{\star(\mathcal{X})} + A_0^{\star(\mathcal{Y})} \right) \|\mathcal{P}\| + B_0^{\star(\mathcal{Y})} \sum_{\mathcal{R}_l \in \mathcal{P}} \|\mathcal{Q}_l\| \right)$$

Assumption (K) is satisfied with $\displaystyle\sum_{S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \in \mathcal{S}} e^{-x_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}} \leq 1.$

Its complete proof can be found in the technical report [13].

A.1 Proof of Proposition 4

We rely on a link between $\|\cdot\|_2$ and $\|\cdot\|_\infty$ structures of the square roots of the models and a relationship between bracketing entropy and metric entropy for $\|\cdot\|_\infty$ norms.

Following Massart [34], we define the following tensorial *norm* on functions $u(y|x)$

$$\|u\|_2^{2\otimes_n} = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|u(\cdot|X_i)\|_2^2 \right] \quad \text{and} \quad \|u\|_\infty^{2,\otimes_n} = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \|u(\cdot|X_i)\|_\infty^2 \right].$$

As the reference measure is the Lebesgue measure on $[0,1]_Y^d$, $\|u\|_\infty^{2\otimes_n} \geq \|u\|_2^{2\otimes_n}$. By definition $d^{\otimes_n}(s,t) = \|\sqrt{s} - \sqrt{t}\|_2^{\otimes_n}$ and thus for any model S_m and any function $s_m \in S_m$

$$H_{[\cdot],d^{\otimes_n}}(\delta, S_m(s_m,\sigma)) = H_{[\cdot],\|\cdot\|_2^{\otimes_n}}\left(\delta, \left\{ u \in \sqrt{S_m} \middle| \|u - \sqrt{s_m}\|_2^{\otimes_n} \leq \sigma \right\}\right)$$

If $\sqrt{S_m}$ is a subset of a linear space $\overline{\sqrt{S_m}}$ of dimension $\mathcal{D}_m$, as in our model,

$$H_{[\cdot],d^{\otimes_n}}(\delta, S_m(s_m,\sigma)) \leq H_{[\cdot],\|\cdot\|_2^{\otimes_n}}\left(\delta, \left\{u \in \overline{\sqrt{S_m}} \middle| \|u - \sqrt{s_m}\|_2^{\otimes_n} \leq \sigma\right\}\right)$$

so that one can replace, without loss of generality, $\sqrt{s_m}$ by 0 and use

$$H_{[\cdot],d^{\otimes_n}}(\delta, S_m(s_m,\sigma)) \leq H_{[\cdot],\|\cdot\|_2^{\otimes_n}}\left(\delta, \left\{u \in \overline{\sqrt{S_m}} \middle| \|u\|_2^{\otimes_n} \leq \sigma\right\}\right).$$

Using now $\|\cdot\|_\infty^{\otimes_n} \geq \|\cdot\|_2^{\otimes_n}$, one deduces

$$H_{[\cdot],d^{\otimes_n}}(\delta, S_m(s_m,\sigma)) \leq H_{[\cdot],\|\cdot\|_\infty^{\otimes_n}}\left(\delta, \left\{u \in \overline{\sqrt{S_m}} \middle| \|u\|_2^{\otimes_n} \leq \sigma\right\}\right).$$

As for any u, $[u-\delta/2, u+\delta/2]$ is a δ-bracket for the $\|\cdot\|_\infty^{\otimes_n}$ norm, any covering of $\left\{u \in \overline{\sqrt{S_m}} \middle| \|u\|_2^{\otimes_n} \leq \sigma\right\}$ by $\|\cdot\|_\infty^{\otimes_n}$ ball of radius $\delta/2$ yields a covering by the corresponding brackets. This implies

$$H_{[\cdot],d^{\otimes_n}}(\delta, S_m(s_m,\sigma)) \leq H_{\|\cdot\|_\infty^{\otimes_n}}\left(\frac{\delta}{2}, \left\{u \in \overline{\sqrt{S_m}} \middle| \|u\|_2^{\otimes_n} \leq \sigma\right\}\right)$$

where $H_d(\delta, S)$, the classical entropy, is defined as the logarithm of the minimum number of ball of radius δ with respect to norm d covering the set S.

The following proposition, proved in next section, is similar to a proposition of Massart [34]. It provides a bound for this last entropy term under an assumption on a link between $\|\cdot\|_\infty^{2\otimes_n}$ and $\|\cdot\|_2^{2\otimes_n}$ structures:

Proposition 6. *For any basis* $\{\phi_k\}_{1\leq k\leq\mathcal{D}_m}$ *of* $\overline{\sqrt{S_m}}$ *such that*

$$\forall\beta \in \mathbb{R}^{\mathcal{D}_m}, \quad \|\sum_{k=1}^{\mathcal{D}_m}\beta_k\phi_k\|_2^{2\otimes_n} \geq \|\beta\|_2^2,$$

let

$$\overline{r}_m(\{\phi_k\}) = \sup_{\sum_{k=1}^{\mathcal{D}_m}\beta_k\phi_k\neq 0} \frac{1}{\sqrt{\mathcal{D}_m}} \frac{\|\sum_{k=1}^{\mathcal{D}_m}\beta_k\phi_k\|_\infty^{\otimes_n}}{\|\beta\|_\infty}.$$

and let $\overline{r}_m$ *be the infimum over all suitable bases.*

Then $\overline{r}_m \geq 1$ *and*

$$H_{\|\cdot\|_\infty^{\otimes_n}}\left(\frac{\delta}{2}, \left\{u \in \overline{\sqrt{S_m}} \middle| \|u\|_2^{\otimes_n} \leq \sigma\right\}\right) \leq \mathcal{D}_m\left(\mathcal{C}_m + \ln\frac{\sigma}{\delta}\right)$$

with $\mathcal{C}_m = \ln(\kappa_\infty\overline{r}_m)$ *and* $\kappa_\infty \leq 2\sqrt{2\pi e}$.

In our setting, using a basis of Legendre polynomials, we are able to derive from Proposition 6

Proposition 7. *For any model of Section 3.2,*

$$\overline{r}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \leq \prod_{d=1}^{d_Y}\left(\sqrt{\mathbf{D}_d+1}\sqrt{2\mathbf{D}_d+1}\right) \sup_{\mathcal{R}_{l,k}^{\times}\in\mathcal{Q}^{\mathcal{P}}} \frac{1}{\sqrt{\|\mathcal{Q}^{\mathcal{P}}\|}\sqrt{|\mathcal{R}_{l,k}^{\times}|}}$$

so that $\forall s_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \in S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$,

$$H_{[\cdot],d^{\otimes_n}}\left(\delta, S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}(s_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}},\sigma)\right) \leq \mathcal{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}\left(\mathcal{C}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} + \ln\frac{\sigma}{\delta}\right)$$

with $\mathcal{C}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} = \ln\left(\kappa_\infty\overline{r}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}\right)$ *and* $\kappa_\infty \leq 2\sqrt{2\pi e}$.

A proof, essentially computational, can be found in our technical report [13]. One easily verifies that

$$\sup_{\mathcal{R}^{\times}_{l,k}\in\mathcal{Q}^{\mathcal{P}}} \frac{1}{\sqrt{\|\mathcal{Q}^{\mathcal{P}}\|}\sqrt{|\mathcal{R}^{\times}_{l,k}|}} \leq \begin{cases} 1 & \text{if all hyperrectangles have same sizes} \\ \sqrt{\frac{n^2}{\|\mathcal{Q}^{\mathcal{P}}\|}} & \text{otherwise.} \end{cases}$$

Remark that when $\star(\mathcal{X}) = \mathrm{UDP}(\mathcal{X})$, $\star(\mathcal{Y}) = \mathrm{UDP}(\mathcal{Y})$ and $\mathcal{Q}_l$ is independent of $\mathcal{R}_l$, all the hyperrectangles have same sizes and that the n^2 corresponds to the arbitrary limitation imposed on the minimal size of the segmentations. If we limit this minimal size to $\frac{1}{\sqrt{n}}$ instead of $\frac{1}{n}$ this factor becomes n.

Let

$$D_\star = \ln\left(\kappa_\infty \prod_{k=1}^{d_Y}\left(\sqrt{\mathbf{D}_k+1}\sqrt{2\mathbf{D}_k+1}\right)\right)$$

we have slightly more than Proposition 4 as $\forall s_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \in S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$,

$$H_{[\cdot],d^{\otimes_n}}\left(\delta, S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}(s_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}},\sigma)\right) \leq \mathcal{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \begin{cases} \left(D_\star + \ln\frac{\sigma}{\delta}\right) & \text{for the same size case} \\ \left(\frac{1}{2}\ln\frac{n^2}{\|\mathcal{Q}^{\mathcal{P}}\|} + D_\star + \ln\frac{\sigma}{\delta}\right) & \text{otherwise} \end{cases}$$

A.2 Proofs of Proposition 6 and Proposition 7

Proof of Proposition 6. Let $(\phi_k)_{1\leq k\leq \mathcal{D}_m}$ be a basis of $\sqrt{S_m}$ satisfying

$$\forall\beta\in\mathbb{R}^{\mathcal{D}_m}, \quad \left\|\sum_{k=1}^{\mathcal{D}_m}\beta_k\phi_k\right\|_2^{2,\otimes_n} \geq \|\beta\|_2^2.$$

Note that for β defined by $\forall 1\leq k\leq\mathcal{D}_m, \beta_k = 1$

$$\left\|\sum_{k=1}^{\mathcal{D}_m}\beta_k\phi_k\right\|_\infty^{2,\otimes_n} \geq \left\|\sum_{k=1}^{\mathcal{D}_m}\beta_k\phi_k\right\|_2^{2,\otimes_n} \geq \|\beta\|_2^2 = \mathcal{D}_m = \mathcal{D}_m\|\beta\|_\infty^2$$

so that $\overline{r}_m(\phi) \geq 1$.

Let the grid $\mathcal{G}_m(\delta,\sigma)$:

$$\left\{\beta\in\mathbb{R}^{\mathcal{D}_m}\middle| \forall 1\leq k\leq\mathcal{D}_m, \beta_k\in\frac{\delta}{\sqrt{\mathcal{D}_m}\overline{r}_m(\phi)}\mathbb{Z} \text{ and } \min_{\beta',\|\beta'\|_2\leq\sigma}\|\beta-\beta'\|_\infty \leq \frac{\delta}{2\sqrt{\mathcal{D}_m}\overline{r}_m(\phi)}\right\}.$$

By definition, for any $u'\in\sqrt{S_m}$ such that $\|u'\|_2^{\otimes_n}\leq\sigma$ there is a β' such that $u' = \sum_{k=1}^{\mathcal{D}_m}\beta'_k\phi_k$ and $\|\beta'\|_2\leq\sigma$. By construction, there is a $\beta\in\mathcal{G}_m(\delta,\sigma)$ such that

$$\|\beta-\beta'\|_\infty \leq \frac{\delta}{2\sqrt{\mathcal{D}_m}\overline{r}_m(\phi)}.$$

Definition of $\overline{r}_m$ implies then that

$$\left\|\sum_{k=1}^{\mathcal{D}_m}\beta_k\phi_k - \sum_{k=1}^{\mathcal{D}_m}\beta'_k\phi_k\right\|_\infty^{\otimes_n} \leq \overline{r}_m(\phi)\sqrt{\mathcal{D}_m}\|\beta-\beta'\|_\infty$$

$$\leq \frac{\delta}{2}.$$

The set $\left\{\sum_{k=1}^{\mathcal{D}_m} \beta_k \phi_k \middle| \beta \in \mathcal{G}_m(\delta,\sigma)\right\}$ is thus a $\frac{\delta}{2}$ covering of $\left\{u \in \overline{\sqrt{S_m}} \middle| \|u\|_2^{\otimes_n} \leq \sigma\right\}$ for the $\|\cdot\|_\infty^{\otimes_n}$ norm. It remains thus only to bound the cardinality of $\mathcal{G}_m(\delta,\sigma)$.

Let $\overline{\mathcal{G}_m(\delta,\sigma)}$ be the union of all hypercubes of width $\frac{\delta}{\sqrt{\mathcal{D}_m}\overline{r}_m(\phi)}$ centered on the grid $\mathcal{G}_m(\delta,\sigma)$, by construction, for any $\beta \in \overline{\mathcal{G}_m(\delta,\sigma)}$ there is a β' with $\|\beta'\|_2 \leq \sigma$ such that $\|\beta' - \beta\|_\infty \leq \frac{\delta}{\sqrt{\mathcal{D}_m}\overline{r}_m(\phi)}$. As $\|\beta' - \beta\|_2 \leq \sqrt{\mathcal{D}_m}\|\beta' - \beta\|_\infty$, this implies $\|\beta\|_2 \leq \sigma + \frac{\delta}{\overline{r}_m(\phi)}$. We then deduce

$$\begin{aligned}
\mathrm{Vol}\left(\overline{\mathcal{G}_m(\delta,\sigma)}\right) = |\mathcal{G}_m(\delta,\sigma)| \left(\frac{\delta}{\sqrt{\mathcal{D}_m}\overline{r}_m(\phi)}\right)^{\mathcal{D}_m} &\leq \mathrm{Vol}\left(\left\{\beta \in \mathbb{R}^{\mathcal{D}_m} \middle| \|\beta\|_2 \leq \sigma + \frac{\delta}{\overline{r}_m(\phi)}\right\}\right) \\
&\leq \left(\sigma + \frac{\delta}{\overline{r}_m(\phi)}\right)^{\mathcal{D}_m} \mathrm{Vol}\left(\left\{\beta \in \mathbb{R}^{\mathcal{D}_m} \middle| \|\beta\|_2 \leq 1\right\}\right)
\end{aligned}$$

and thus

$$|\mathcal{G}_m(\delta,\sigma)| \leq \left(1 + \frac{\sigma\overline{r}_m(\phi)}{\delta}\right)^{\mathcal{D}_m} \mathcal{D}_m^{\mathcal{D}_m/2} \mathrm{Vol}\left(\left\{\beta \in \mathbb{R}^{\mathcal{D}_m} \middle| \|\beta\|_2 \leq 1\right\}\right)$$

and as $\frac{\sigma\overline{r}_m(\phi)}{\delta} \geq 1$ and $\mathrm{Vol}\left(\left\{\beta \in \mathbb{R}^{\mathcal{D}_m} \middle| \|\beta\|_2 \leq 1\right\}\right) \leq \left(\frac{2\pi e}{\mathcal{D}_m}\right)^{\mathcal{D}_m/2}$

$$|\mathcal{G}_m(\delta,\sigma)| \leq \left(\frac{2\sqrt{2\pi e}\overline{r}_m(\phi)\sigma}{\delta}\right)^{\mathcal{D}_m}$$

which concludes the proof. □

Instead of Proposition 7, by mimicking a proof of Massart [34], we prove in our technical report [13] an extended version of it in which the degree of the conditional densities may depend on the hyperrectangle. More precisely, we reuse the partition $\mathcal{P} \in \mathcal{S}_{\mathcal{P}}^{\star(\mathcal{X})}$ and the partitions $\mathcal{Q}_l \in \mathcal{S}_{\mathcal{P}}^{\star(\mathcal{Y})}$ for $\mathcal{R}_l \in \mathcal{P}$ and define now the model $S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$ as the set of conditional densities such that

$$s(y|x) = \sum_{\mathcal{R}_{l,k}^{\times} \in \mathcal{Q}^{\mathcal{P}}} P_{\mathcal{R}_{l,k}^{\times}}^2(y) \mathbf{1}_{\left\{(x,y) \in \mathcal{R}_{l,k}^{\times}\right\}}$$

where $P_{\mathcal{R}_{l,k}^{\times}}$ is a polynomial of degree at most $\mathbf{D}(\mathcal{R}_{l,k}^{\times}) = \left(\mathbf{D}_1(\mathcal{R}_{l,k}^{\times}), \ldots, \mathbf{D}_{d_Y}(\mathcal{R}_{l,k}^{\times})\right)$ which depends on the leaf.

Instead of the true dimension, we use a slight upper bound

$$\mathcal{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} = \sum_{\mathcal{R}_l \in \mathcal{P}} \sum_{\mathcal{R}_{l,k}^{\mathcal{Y}} \in \mathcal{Q}_l} \prod_{d=1}^{d_Y} \left(\mathbf{D}_d(\mathcal{R}_{l,k}^{\times}) + 1\right) = \sum_{\mathcal{R}_{l,k}^{\times} \in \mathcal{Q}^{\mathcal{P}}} \prod_{d=1}^{d_Y} \left(\mathbf{D}_d(\mathcal{R}_{l,k}^{\times}) + 1\right)$$

Note that the space $S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$ introduced in the main part of the paper corresponds to the case where the degree $\mathbf{D}(\mathcal{R}_{l,k}^{\times})$ does not depend on the hyperrectangle $\mathcal{R}_{l,k}^{\times}$.

Proposition 8. *There exists*

$$\overline{r}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \leq \frac{\sup_{\mathcal{R}^{\times}_{l,k}\in\mathcal{Q}^{\mathcal{P}}} \prod_{d=1}^{d_Y} \left(\sum_{D_d \leq \mathbf{D}_d(\mathcal{R}^{\times}_{l,k})} \sqrt{2D_d+1} \right)}{\inf_{\mathcal{R}^{\times}_{l,k}\in\mathcal{Q}^{\mathcal{P}}} \prod_{d=1}^{d_Y} \sqrt{\mathbf{D}_d(\mathcal{R}^{\times}_{l,k})+1}} \sup_{\mathcal{R}^{\times}_{l,k}\in\mathcal{Q}^{\mathcal{P}}} \frac{1}{\sqrt{\|\mathcal{P}\|}\sqrt{|\mathcal{R}^{\times}_{l,k}|}}$$

such that $\forall s_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \in S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$,

$$H_{[\cdot],d^{\otimes_n}}\left(\delta, S_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}(s_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}},\sigma)\right) \leq \mathcal{D}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} \left(\mathcal{C}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} + \ln \frac{\sigma}{\delta} \right)$$

with $\mathcal{C}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}} = \ln\left(\kappa_\infty \overline{r}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}\right)$ *and* $\kappa_\infty \leq 2\sqrt{2\pi e}$.

Proposition 7 is deduced from this proposition with the help of the simple upper bound

$$\sum_{D_d \leq \mathbf{D}_d(\mathcal{R}^{\times}_{l,k})} \sqrt{2D_d+1} \leq (\mathbf{D}_d(\mathcal{R}^{\times}_{l,k})+1)\sqrt{2\mathbf{D}_d(\mathcal{R}^{\times}_{l,k})+1}.$$

As

$$\frac{\sup_{\mathcal{R}^{\times}_{l,k}\in\mathcal{Q}^{\mathcal{P}}} \prod_{d=1}^{d_Y} \left(\sum_{D_d \leq \mathbf{D}_d(\mathcal{R}^{\times}_{l,k})} \sqrt{2D_d+1} \right)}{\inf_{\mathcal{R}^{\times}_{l,k}\in\mathcal{Q}^{\mathcal{P}}} \prod_{d=1}^{d_Y} \sqrt{\mathbf{D}_d(\mathcal{R}^{\times}_{l,k})+1}} \leq \prod_{d=1}^{d_Y} \max \sqrt{2}(\mathbf{D}_d+1),$$

once a maximal degree is chosen along each axis, the equivalent of constant $C_\star$ of 2 depends only on this maximal degrees. Assumption $\mathrm{H}_{\mathcal{Q}^{\mathcal{P}},\mathbf{D}}$ holds then, with the same constants, simultaneaously for all models of both global choice and local choice strategies. Obtaining the Kraft type assumption, Assumption (K) is only a matter of taking into account the augmentation of the number of models within the collection. Replacing respectively $A_0^{\star(\mathcal{X})}$ by $A_0^{\star(\mathcal{X})} + \ln|\mathcal{D}^M|$ for global optimization and $B_0^{\star(\mathcal{Y})}$ by $B_0^{\star(\mathcal{Y})} + \ln|\mathcal{D}^M|$ for local optimization, where $|\mathcal{D}^M|$ denotes the size of the family of possible degrees, turns out to be sufficient as mentioned earlier.

The proof of Proposition 8 is essentially computational and thus relegated to our extended technical report.

B Proofs for Section 3.3 (Spatial Gaussian mixtures, models, bracketing entropy and penalties)

As in the piecewise polynomial density case, Theorem 3 is obtained by showing that Assumptions ($\mathrm{H}_{\mathcal{P},K,\mathcal{G}}$), ($\mathrm{S}_{\mathcal{P},K,\mathcal{G}}$) and (K) hold for any collection.

Again, one easily verifies that Assumption ($\mathrm{S}_{\mathcal{P},K,\mathcal{G}}$) holds. For the complexity assumption, combining 1 with a bound on the bracketing entropy of the models of type

$$H_{[\cdot],d^{\sup}}(\delta, S_{\mathcal{P},K,\mathcal{G}}) \leq \dim(S_{\mathcal{P},K,\mathcal{G}}) \left(C + \ln\frac{1}{\delta} \right),$$

one obtains

Proposition 9. *There exists a constant C depending only on a, L_-, L_+, λ_- and λ_+ such that for any model $S_{\mathcal{P},K,\mathcal{G}}$ of Theorem 3 Assumption ($H_{\mathcal{P},K,\mathcal{G}}$) is satisfied with a function ϕ such that*

$$\mathfrak{D}_{\mathcal{P},K,\mathcal{G}} \leq \left(2\left(\sqrt{C}+\sqrt{\pi}\right)^2 + 1 + \left(\ln \frac{n}{e\left(\sqrt{C}+\sqrt{\pi}\right)^2 \dim(S_{\mathcal{P},K,\mathcal{G}})} \right)_+ \right) \dim(S_{\mathcal{P},K,\mathcal{G}}).$$

For the Kraft assumption, one can verify that

Proposition 10. *For any collections $\mathcal{S}$ of Theorem 3, there is a $c_\star$ such that for the choice*

$$x_{\mathcal{P},K,\mathcal{G}} = c_\star \left(A_0^{\star(\mathcal{X})} + B_0^{\star(\mathcal{X})} \|\mathcal{P}\| + (K-1) + \mathcal{D}_E \right),$$

Assumption (K) holds with $\displaystyle\sum_{S_{\mathcal{P},K,\mathcal{G}} \in \mathcal{S}} e^{-x_{\mathcal{P},K,\mathcal{G}}} \leq 1.$

As for the piecewise polynomial case section, the main difficulty lies in controlling the bracketing entropy of the models. A proof of Proposition 10 can be found in our technical report [13].

We focus thus on the proof of Proposition 9. Due to the complex structure of spatial mixture, we did not manage to bound the bracketing entropy of local model. We derive only an upper bound of the bracketing entropy $H_{[\cdot],d^{\otimes_n}}(\delta, S_{\mathcal{P},K,\mathcal{G}})$, but one that is independent of the distribution law of $(X_i)_{1\leq i\leq n}$: the bracketing entropy with a sup norm Hellinger distance $d^{\sup} = \sqrt{d^{2\,\sup}}$, $H_{[\cdot],d^{\sup}}(\delta, S_{\mathcal{P},K,\mathcal{G}})$, where $d^{2\,\sup}$ is defined by

$$d^{2\,\sup}(s,t) = \sup_x d^2\left(s(\cdot|x), t(\cdot|x)\right).$$

Obviously $d^{2\,\sup} \geq d^{2\otimes_n}$ and thus $H_{[\cdot],d^{\sup}}(\delta, S_{\mathcal{P},K,\mathcal{G}}) \geq H_{[\cdot],d^{\otimes_n}}(\delta, S_{\mathcal{P},K,\mathcal{G}})$. This upper bound is furthermore design independent.

Proposition 9 is a direct consequence of Proposition 1 and

Proposition 11. *There exists a constant C depending only on a, L_-, L_+, λ_- and λ_+ such that for any model $S_{\mathcal{P},K,\mathcal{G}}$ of Theorem 3:*

$$H_{[\cdot],d^{\sup}}(\delta, S_{\mathcal{P},K,\mathcal{G}}) \leq \dim(S_{\mathcal{P},K,\mathcal{G}}) \left(C + \ln \frac{1}{\delta} \right).$$

B.1 Entropy of spatial mixtures

Proof of Proposition 11. While we use classical Hellinger distance to measure the complexity of the simplex $\mathcal{S}_{K-1}$ and the set $\mathcal{G}_{E^\perp}$, we use a sup norm Hellinger distance on $\mathcal{G}_E^K$ defined by

$$d^{2\,\max}\left((s_1,\ldots,s_K),(t_1,\ldots,t_K)\right) = \sup_k d^2(s_k,t_k).$$

We say that $[(s_1,\ldots,s_K),(t_1,\ldots,t_K)]$ is a bracket of $\mathcal{G}_E^K$ if $\forall 1 \leq k \leq K, s_k \leq t_k$.

Using a similar proof than Genovese and Wasserman [20], we decompose the entropy in three parts with:

Lemma 1. *For any $\delta \in (0, \sqrt{2}]$,*

$$H_{[\cdot],d^{\sup}}(\delta, S_{\mathcal{P},K,\mathcal{G}}) \leq \|\mathcal{P}\| H_{[\cdot],d}(\delta/3, \mathcal{S}_{K-1}) + H_{[\cdot],d^{\max}}(\delta/9, \mathcal{G}_E^K) + H_{[\cdot],d}(\delta/9, \mathcal{G}_{E^\perp}).$$

We bound those bracketing entropies with the help of two results. We first use a Lemma proved in Genovese and Wasserman [20] that implies the existence of a universal constant $\mathcal{C}_\mathcal{S}$ such that

$$H_{[\cdot],d}(\delta/3, \mathcal{S}_{K-1}) \leq (K-1)\left(\mathcal{C}_\mathcal{S} + \ln\frac{1}{\delta}\right):$$

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

$$H_{[\cdot],d}(\delta/3, \mathcal{S}_{K-1}) \leq (K-1)\left(\mathcal{C}_{\mathcal{S}_{K-1}} + \ln\frac{1}{\delta}\right)$$

with $\mathcal{C}_{\mathcal{S}_{K-1}} = \dfrac{1}{K-1}\ln K + \dfrac{K}{2(K-1)}\ln(2\pi e) + \ln 3\sqrt{2}$

Furthermore, uniformly on K: $\mathcal{C}_{\mathcal{S}_{K-1}} \leq \ln 2 + \dfrac{1}{2}\ln(2\pi e) + \ln 3\sqrt{2} = \mathcal{C}_{\mathcal{S}}$

We then rely on Proposition 12 to handle the bracketing entropy of Gaussian K-tuples collection. It implies the existence of two constants $\mathcal{C}_{[\star]^\star}$ and $\mathcal{C}_{[\star]}$ depending only on a, L_-, L_+, λ_- and λ_+ such that

$$H_{[\cdot],d^{\max}}\left(\delta/9, \mathcal{G}_E^K\right) \leq \dim(\mathcal{G}_E^K)\left(\mathcal{C}_{[\star]^\star} + \ln\frac{1}{\delta}\right)$$
$$H_{[\cdot],d}(\delta/9, \mathcal{G}_{E^\perp}) \leq \dim(\mathcal{G}_{E^\perp})\left(\mathcal{C}_{[\star]} + \ln\frac{1}{\delta}\right).$$

As $\dim(S_{K,\mathcal{P},\mathcal{G}}) = \|\mathcal{P}\|(K-1) + \dim(\mathcal{G}_E^K) + \dim(\mathcal{G}_{E^\perp})$, we obtain Proposition 11 with $C = \max(\mathcal{C}_{\mathcal{S}}, \mathcal{C}_{[\star]^\star}, \mathcal{C}_{[\star]})$. □

B.2 Entropy of Gaussian families

Proposition 12. *For any $\delta \in (0, \sqrt{2}]$,*

$$H_{[\cdot],d^{\max}}(\delta/9, \mathcal{G}_{[\mu_\star,\mathrm{L}_\star,\mathrm{D}_\star,\mathrm{A}_\star]_E^K}) \leq \mathcal{V}_{[\mu_\star,\mathrm{L}_\star,\mathrm{D}_\star,\mathrm{A}_\star]_{p_E}^K} + \mathcal{D}_{[\mu_\star,\mathrm{L}_\star,\mathrm{D}_\star,\mathrm{A}_\star]_{p_E}^K}\ln\frac{1}{\delta}$$

where $\mathcal{D}_{[\mu_\star,\mathrm{L}_\star,\mathrm{D}_\star,\mathrm{A}_\star]_{p_E}^K} = \dim\left(\Theta_{[\mu_\star,\mathrm{L}_\star,\mathrm{D}_\star,\mathrm{A}_\star]_{p_E}^K}\right) = c_{\mu_\star}\mathcal{D}_{\mu,p_E} + c_{\mathrm{L}_\star}\mathcal{D}_L + c_{\mathrm{D}_\star}\mathcal{D}_{D,p_E} + c_{\mathrm{A}_\star}\mathcal{D}_{A,p_E}$ *and*

$$\mathcal{V}_{[\mu_\star,\mathrm{L}_\star,\mathrm{D}_\star,\mathrm{A}_\star]_{p_E}^K} = c_{\mu_\star}\mathcal{V}_{\mu,p_E} + c_{\mathrm{L}_\star}\mathcal{V}_{L,p_E} + c_{\mathrm{D}_\star}\mathcal{V}_{D,p_E} + c_{\mathrm{A}_\star}\mathcal{V}_{A,p_E} \text{ with } \begin{cases} c_{\mu_0} = c_{\mathrm{L}_0} = c_{\mathrm{D}_0} = c_{\mathrm{A}_0} = 0 \\ c_{\mu_K} = c_{\mathrm{L}_K} = c_{\mathrm{D}_K} = c_{\mathrm{A}_K} = K \\ c_\mu = c_\mathrm{L} = c_\mathrm{D} = c_\mathrm{A} = 1 \end{cases},$$

$$\begin{cases} \mathcal{D}_{\mu,p_E} = p_E \\ \mathcal{D}_\mathrm{L} = 1 \\ \mathcal{D}_{\mathrm{D},p_E} = \frac{p_E(p_E-1)}{2} \\ \mathcal{D}_{\mathrm{A},p_E} = p_E - 1 \end{cases} \quad \text{and} \quad \begin{cases} \mathcal{V}_{\mu,p_E} = p_E\left(\ln\left(1 + 108\frac{a}{\sqrt{L_-\lambda_-\frac{\lambda_-}{\lambda_+}}}p_E\right)\right) \\ \mathcal{V}_{\mathrm{L},p_E} = \ln\left(1 + 39\ln\left(\frac{L_+}{L_-}\right)p_E\right) \\ \mathcal{V}_{\mathrm{D},p_E} = \frac{p_E(p_E-1)}{2}\left(\frac{2\ln c_S}{p_E(p_E-1)} + \left(\ln\left(252\frac{\lambda_+}{\lambda_-}p_E\right)\right)\right) \\ \mathcal{V}_{\mathrm{A},p_E} = (p_E-1)\left(\ln\left(2 + 255\frac{\lambda_+}{\lambda_-}\ln\left(\frac{\lambda_+}{\lambda_-}\right)p_E\right)\right) \end{cases}$$

where c_S is a universal constant.

Proof of Proposition 12. We consider all models $\mathcal{G}_{[\mu_\star\ \mathrm{L}_\star\ \mathrm{A}_\star\ \mathrm{D}_\star]_E^K}$ at once by a "tensorial" construction of a suitable $\delta/9$ bracket collection.

We first define a set of grids for the mean μ, the volume L, the eigenvector matrix D and the renormalized eigenvalue matrix A from which one constructs the bracket collection.

- For any δ_μ, the grid $\mathcal{G}_\mu(a, p_E, \delta_\mu)$ of $[-a, a]^{p_E}$:

$$\mathcal{G}_\mu(a, p_E, \delta_\mu) = \left\{ g\delta_\mu \middle| g \in \mathbb{Z}^{p_E}, \|g\|_\infty \leq \frac{a}{\delta_\mu} \right\}.$$

- For any δ_{L}, the grid $\mathcal{G}_{\mathrm{L}}(L_-, L_+, \delta_{\mathrm{L}})$ of $[L_-, L_+]$:
$$\mathcal{G}_{\mathrm{L}}(L_-, L_+, \delta_{\mathrm{L}}) = \{L_-(1+\delta_{\mathrm{L}})^g | g \in \mathbb{N}, L_-(1+\delta_{\mathrm{L}})^g \leq L_+\}\,.$$
- For any δ_{D}, the grid $\mathcal{G}_{\mathrm{D}}(p_E, \delta_{\mathrm{D}})$ of $SO(p_E)$ made of the elements of a δ_{D}-net with respect to the $\|\cdot\|_2$ operator norm (as described by Szarek [40]).
- For any δ_{A}, the grid $\mathcal{G}_{\mathrm{A}}(\lambda_-, \lambda_+, p_E, \delta_{\mathrm{A}})$ of $\mathcal{A}(\lambda_-, \lambda_+(1+\delta_{\mathrm{A}}), p_E)$:
$$\mathcal{G}_{\mathrm{A}}(\lambda_-, \lambda_+, p_E, \delta_{\mathrm{A}}) = \{A \in \mathcal{A}(\lambda_-, \lambda_+(1+\delta_{\mathrm{A}}), p_E) | \forall 1 \leq i < p_E, \exists g_i \in \mathbb{N}, A_i = \lambda_-(1+\delta_{\mathrm{A}})^{g_i}\}\,.$$

Obviously, for any $\mu \in [-a, a]$, there is a $\tilde{\mu} \in \mathcal{G}_\mu(a, p_E, \delta_\mu)$ such that
$$\|\tilde{\mu} - \mu\|^2 \leq p_E \delta_\mu^2$$
while
$$|\mathcal{G}_\mu(a, p_E, \delta_\mu)| \leq \left(1 + 2\frac{a}{\delta_\mu}\right)^{p_E} \leq \max\left(2^{p_E}, \left(\frac{4a}{\delta_\mu}\right)^{p_E}\right).$$
In the same fashion, for any L in $[L_-, L_+]$, there is a $\tilde{L} \in \mathcal{G}_{\mathrm{L}}(L_-, L_+, \delta_{\mathrm{L}})$ such that $(1+\delta_{\mathrm{L}})^{-1} L_{j_L} < L \leq L_{j_L}$ while
$$|\mathcal{G}_{\mathrm{L}}(L_-, L_+, \delta_{\mathrm{L}})| \leq 1 + \frac{\ln\left(\frac{L_+}{L_-}\right)}{\ln(1+\delta_{\mathrm{L}})}.$$
If we further assume that $\delta_{\mathrm{L}} \leq \frac{1}{12}$ then $\ln(1+\delta_{\mathrm{L}}) \geq \frac{12}{13}\delta_{\mathrm{L}}$ and
$$|\mathcal{G}_{\mathrm{L}}(L_-, L_+, \delta_{\mathrm{L}})| \leq 1 + \frac{13\ln\left(\frac{L_+}{L_-}\right)}{12\delta_{\mathrm{L}}}.$$

By definition on a δ_{D}-net, for any $D \in SO(p_E)$ there is a $\tilde{D} \in \mathcal{G}_{\mathrm{D}}(p_E, \delta_{\mathrm{D}})$ such that
$$\forall x, \|(\tilde{D} - D)x\|_2 \leq \delta_{\mathrm{D}} \|x\|_2.$$
As proved by Szarek [40], it exists a universal constant c_S such that, as soon as $\delta_{\mathrm{D}} \leq 1$
$$|\mathcal{G}_{\mathrm{D}}(p_E, \delta_{\mathrm{D}})| \leq c_S \left(\frac{1}{\delta_{\mathrm{D}}}\right)^{\frac{p_E(p_E-1)}{2}}$$
where $\frac{p_E(p_E-1)}{2}$ is the intrinsic dimension of $SO(p_E)$.

The structure of the grid $\mathcal{G}_{\mathrm{A}}(\lambda_-, \lambda_+, p_E, \delta_{\mathrm{A}})$ is more complex. Although, looking at condition on the $p_E - 1$ first diagonal values,
$$|\mathcal{G}_{\mathrm{A}}(\lambda_-, \lambda_+, p_E, \delta_{\mathrm{A}})| \leq \left(2 + \frac{\ln\left(\frac{\lambda_+}{\lambda_-}\right)}{\ln(1+\delta_{\mathrm{A}})}\right)^{p_E - 1}$$
where $p_E - 1$ is the intrinsic dimension of $\mathcal{A}(\lambda_-, \lambda_+, p_E)$. If we further assume that $\delta_{\mathrm{A}} \leq \frac{1}{84}$ then $\ln(1+\delta_{\mathrm{A}}) \geq \frac{84}{85}\delta_{\mathrm{A}}$ and thus
$$|\mathcal{G}_{\mathrm{A}}(\lambda_-, \lambda_+, p_E, \delta_{\mathrm{A}})| \leq \left(2 + \frac{85\ln\left(\frac{\lambda_+}{\lambda_-}\right)}{84\delta_{\mathrm{A}}}\right)^{p_E - 1}.$$
A key to the succes of this construction is the following approximation property of this grid obtained in our technical report [13] with a calculatory proof:

Lemma 3. *For $A \in \mathcal{A}(\lambda_-, \lambda_+, p_E)$ there is $\tilde{A} \in \mathcal{G}_{\mathrm{A}}(\lambda_-, \lambda_+, p_E, \delta_{\mathrm{A}})$ such that*

$$|\tilde{A}_{i,i}^{-1} - A_{i,i}^{-1}| \leq \delta_{\mathrm{A}} \lambda_-^{-1}.$$

Define $c_{\mu_0} = c_{\mathrm{L}_0} = c_{\mathrm{D}_0} = c_{\mathrm{A}_0} = 0, c_{\mu_K} = c_{\mathrm{L}_K} = c_{\mathrm{D}_K} = c_{\mathrm{A}_K} = K, c_\mu = c_{\mathrm{L}} = c_{\mathrm{D}} = c_{\mathrm{A}} = 1$. Let $f_{K,\mu_\star,p_E}$ be the application from $(\mathbb{R}^{p_E})^{c_{\mu_\star}}$ to $\mathbb{R}^K$ defined by

$$\begin{cases} 0 \mapsto (\mu_{0,1}, \dots, \mu_{0,K}) & \text{if } \mu_\star = \mu_0 \\ (\mu_1, \dots, \mu_K) \mapsto (\mu_1, \dots, \mu_K) & \text{if } \mu_\star = \mu_K \\ \mu \mapsto (\mu, \dots, \mu) & \text{if } \mu_\star = \mu \end{cases},$$

and $f_{K,\mathrm{L}_\star}$ (respectively $f_{K,\mathrm{D}_\star,p_E}$ and $f_{K,\mathrm{A}_\star,p_E}$) be the similar application from $(\mathbb{R}^+)^{c_{\mathrm{L}_\star}}$ into $(\mathbb{R}^+)^K$ (respectively from $(SO(p_E))^{c_{\mathrm{D}_\star}}$ into $(SO(p_E))^K$ and from $(\mathcal{A}(0, +\infty, p_E))^{c_{\mathrm{A}_\star}}$ into $(\mathcal{A}(0, +\infty, p_E))^K$).

By definition, the image of

$$([-a, a]^{p_E})^{c_{\mu_\star}} \times ([L_-, L_+])^{c_{\mathrm{L}_\star}} \times (SO(p_E))^{c_{\mathrm{D}_\star}} \times (\mathcal{A}(\lambda_-, \lambda_+, p_E))^{c_{\mathrm{A}_\star}}$$

by $\left(f_{K,\mu_\star,p_E} \otimes f_{L_{K,\cdot},p_E} \otimes f_{K,\mathrm{D}_\star,p_E} \otimes f_{K,\mathrm{A}_\star}\right)$ is, up to reordering, the set of parameters of all K-tuples of Gaussian densities of type $[\mu_\star \,\mathrm{L}_\star, \mathrm{D}_\star, \mathrm{A}_\star]^K$.

We construct our $\delta/9$ bracket covering with a grid on those parameters. For any K-tuple of Gaussian parameters $((\mu_1, \Sigma_1), \dots, (\mu_K, \Sigma_K))$ and any δ_Σ, we associate the K-tuple of pairs

$$\Bigg(\left((1+\delta_\Sigma)^{-p_E} \Phi_{\mu_1,(1+\delta_\Sigma)^{-1}\Sigma_1}, (1+\delta_\Sigma)^{p_E} \Phi_{\mu_1,(1+\delta_\Sigma)\Sigma_1}\right), \dots, \\ \left((1+\delta_\Sigma)^{-p_E} \Phi_{\mu_K,(1+\delta_\Sigma)^{-1}\Sigma_K}, (1+\delta_\Sigma)^{p_E} \Phi_{\mu_K,(1+\delta_\Sigma)\Sigma_K}\right) \Bigg).$$

We prove in our technical report [13] that, for $\gamma = 18/49$ and $\beta = \sqrt{\cosh(\frac{1}{6}) + \frac{1}{2}}$, the choice

$$\delta_\mu = \frac{\sqrt{\gamma L_- \lambda_- \frac{\lambda_-}{\lambda_+}}}{9\beta} \frac{\delta}{p_E}, \ \delta_{\mathrm{L}} = \frac{1}{18\beta} \frac{\delta}{p_E} \leq \frac{1}{12}, \ \delta_{\mathrm{D}} = \delta_{\mathrm{A}} = \frac{1}{126\beta} \frac{\lambda_-}{\lambda_+} \frac{\delta}{p_E} \leq \frac{1}{84}, \ \delta_\Sigma = \frac{1}{9\beta} \frac{\delta}{p_E} \leq \frac{1}{8}$$

is such that the image of

$$(\mathcal{G}_\mu(a, p_E, \delta_\mu))^{c_{\mu_\star}} \times (\mathcal{G}_{\mathrm{L}}(L_-, L_+, \delta_{\mathrm{L}}))^{c_{\mathrm{L}_\star}} \times (\mathcal{G}_{\mathrm{D}}(p_E, \delta_{\mathrm{D}}))^{c_{\mathrm{D}_\star}} \times (\mathcal{G}_{\mathrm{A}}(\lambda_-, \lambda_+, p_E, \delta_{\mathrm{A}}))^{c_{\mathrm{A}_\star}}$$

by $f_{K,\mu_\star,p_E} \otimes f_{L_{K,\cdot},p_E} \otimes f_{K,\mathrm{D}_\star,p_E} \otimes f_{K,\mathrm{A}_\star}$ is a set of parameters corresponding to a set of pairs that is a $\delta/9$-bracket covering of $\mathcal{G}_{[\mu_\star \,\mathrm{L}_\star \,\mathrm{D}_\star \,\mathrm{A}_\star]_E^K}$ for the $d^{\max}$ norm.

The cardinality of this $\delta/9$-bracket covering is bounded by

$$\left(\left(1 + \frac{18 a \beta p_E}{\sqrt{\gamma L_- \lambda_- \frac{\lambda_-}{\lambda_+}} \delta} \right)^{p_E} \right)^{c_{\mu_\star}} \times \left(\left(1 + \frac{39\beta \ln\left(\frac{L_+}{L_-}\right) p_E}{2\delta} \right) \right)^{c_{\mathrm{L}_\star}} \\ \times \left(c_S \left(\frac{126\beta \frac{\lambda_+}{\lambda_-} p_E}{\delta} \right)^{\frac{p_E(p_E-1)}{2}} \right)^{c_{\mathrm{D}_\star}} \times \left(\left(2 + \left(\frac{255\beta \frac{\lambda_+}{\lambda_-} \ln\left(\frac{\lambda_+}{\lambda_-}\right) p_E}{2\delta} \right) \right)^{p_E - 1} \right)^{c_{\mathrm{A}_\star}}$$

So

$H_{[\cdot],d^{\max}}(\delta/9, \mathcal{G}_{[\mu_\star,\mathrm{L}_\star,\mathrm{D}_\star,\mathrm{A}_\star]_E^K})$

$$\begin{aligned}\leq c_{\mu_\star} p_E \left(\ln \left(1 + \frac{18\beta a p_E}{\sqrt{\gamma L_- \lambda_- \frac{\lambda_-}{\lambda_+}}} \right) + \ln \frac{1}{\delta} \right) + c_{\mathrm{L}_\star} \left(\ln \left(1 + \frac{39}{2} \beta \ln \left(\frac{L_+}{L_-} \right) p_E \right) + \ln \frac{1}{\delta} \right) \\ + c_{\mathrm{D}_\star} \frac{p_E(p_E-1)}{2} \left(\frac{2 \ln c_S}{p_E(p_E-1)} + \ln \left(126\beta \frac{\lambda_+}{\lambda_-} p_E \right) + \ln \frac{1}{\delta} \right) \\ + c_{\mathrm{A}_\star} (p_E - 1) \left(\ln \left(2 + \frac{255}{2} \beta \frac{\lambda_+}{\lambda_-} \ln \left(\frac{\lambda_+}{\lambda_-} \right) p_E \right) + \ln \frac{1}{\delta} \right)\end{aligned}$$

which concludes the proof as soon as one notices that $1/9 \leq \gamma \leq 1/3$ and $1 \leq \beta \leq 2$. □

References

[1] N. Akakpo. “Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection”. Submitted. 2010.

[2] N. Akakpo and C. Lacour. “Inhomogeneous and anisotropic conditional density estimation from dependent data”. Submitted. 2011.

[3] A. Antoniadis, J. Bigot, and R. von Sachs. “A multiscale approach for statistical characterization of functional images”. In: *J. Comput. Graph. Statist.* 18.1 (2008), pp. 216–237.

[4] A. Barron et al. “MDL Principle, Penalized Likelihood, and Statistical Risk”. In: Tampere University Press, 2008. Chap. in Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday.

[5] D. Bashtannyk and R. Hyndman. “Bandwidth selection for kernel conditional density estimation”. In: *Computational Statistics & Data Analysis* 36.3 (2001), pp. 279 –298.

[6] L. Bertrand et al. “European research platform IPANEMA at the SOLEIL synchrotron for ancient and historical materials”. In: *Journal of Synchrotron Radiation* 18.5 (Sept. 2011).

[7] C. Biernacki et al. “Model-based cluster and discriminant analysis with the MIXMOD software”. In: *Comput. Statist. Data Anal.* 51.2 (2006), pp. 587–600.

[8] L. Birgé and P. Massart. “Minimal penalties for Gaussian model selection”. In: *Probability theory and related fields* 138.1-2 (2007), pp. 33–73.

[9] L. Birgé and P. Massart. “Minimum contrast estimators on sieves: exponential bounds and rates of convergence”. In: *Bernoulli* 4.3 (1998), pp. 329–375.

[10] G. Blanchard et al. “Optimal dyadic decision trees”. In: *Machine Learning* 66.2 (2007), pp. 209–241.

[11] E. Brunel, F. Comte, and C. Lacour. “Adaptive Estimation of the Conditional Density in Presence of Censoring”. In: *Sankhyā* 69.4 (2007), pp. 734–763.

[12] S. Cohen and E. Le Pennec. “Conditional Density Estimation by Penalized Likelihood Model Selection”. Submitted. 2011.

[13] S. Cohen and E. Le Pennec. “Conditional Density Estimation by Penalized Likelihood Model Selection and Applications”. ArXiv 1103.2021. 2011.

[14] S. Cohen and E. Le Pennec. “Unsupervised segmentation of hyperspectral images with spatialized Gaussian mixture model and model selection”. ArXiv. 2012.

[15] D. Donoho. “CART and best-ortho-basis: a connection”. In: *Ann. Statist.* 25.5 (1997), pp. 1870–1911.

[16] S. Efromovich. “Conditional density estimation in a regression setting”. In: *Ann. Statist.* 35.6 (2007), pp. 2504–2535.

[17] S. Efromovich. “Oracle inequality for conditional density estimation and an actuarial example”. In: *Annals of the Institute of Statistical Mathematics* 62 (2 2010), pp. 249–275.

[18] J. Fan, Q. Yao, and H. Tong. “Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems”. In: *Biometrika* 83.1 (1996), pp. 189–206.

[19] S. van de Geer. “The method of sieves and minimum contrast estimators”. In: *Math. Methods Statist.* 4 (1995), pp. 20–38.

[20] C. Genovese and L. Wasserman. “Rates of convergence for the Gaussian mixture sieve”. In: *Ann. Statist.* 28.4 (2000), pp. 1105–1127.

[21] J. de Gooijer and D. Zerom. “On conditional density estimation”. In: *Statist. Neerlandica* 57.2 (2003), pp. 159–176.

[22] L. Györfi and M. Kohler. “Nonparametric estimation of conditional distributions”. In: *IEEE Trans. Information Theory* 53 (2007), pp. 1872–1879.

[23] P. Hall, R. Wolff, and Q. Yao. “Methods for estimating a conditional distribution function”. In: *J. Amer. Statist. Assoc.* 94 (1999), pp. 154–163.

[24] T. Hofmann. “Probabilistic latent semantic analysis”. In: *Proc. of Uncertainty in Artificial Intelligence.* 1999.

[25] Y. Huang et al. “Fast search for best representations in multitree dictionaries”. In: *IEEE Transactions on Image Processing* 15.7 (July 2006), pp. 1779 –1793.

[26] R. Hyndman, D. Bashtannyk, and G. Grunwald. “Estimating and visualizing conditional densities”. In: *Journal of Computational and Graphical Statistics* 5 (1996), pp. 315–336.

[27] R. Hyndman and Q. Yao. “Nonparametric estimation and symmetry tests for conditional density functions”. In: *Journal of nonparametric statistics* 14.3 (2002), pp. 259–278.

[28] B. Karaivanov and P. Petrushev. “Nonlinear piecewise polynomial approximation beyond Besov spaces”. In: *Applied and Computational Harmonic Analysis* 15.3 (2003), pp. 177 – 223.

[29] I. van Keilegom and N. Veraverbeke. “Density and hazard estimation in censored regression models”. In: *Bernoulli* 8.5 (2002), pp. 607–625.

[30] E. Kolaczyk, J. Ju, and S. Gopal. “Multiscale, multigranular statistical image segmentation”. In: *J. Amer. Statist. Assoc.* 100.472 (2005), pp. 1358–1369.

[31] E. Kolaczyk and R. Nowak. “Multiscale generalised linear models for nonparametric function estimation”. In: *Biometrika* 92.1 (2005), pp. 119–133.

[32] J. Lin. “Divergence measures based on the Shannon entropy”. In: *Information Theory, IEEE Transactions on* 37.1 (Jan. 1991), pp. 145 –151.

[33] Q. Li and J. Racine. *Nonparametric Econometrics: Theory and Practice.* Princeton University Press, 2007.

[34] P. Massart. *Concentration inequalities and model selection.* Vol. 1896. Lecture Notes in Mathematics. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. Berlin: Springer, 2007, pp. xiv+337.

[35] C. Maugis and B. Michel. “A non asymptotic penalized criterion for Gaussian mixture model selection”. In: *ESAIM: P & S* (2010). To appear.

[36] C. Maugis and B. Michel. “Data-driven penalty calibration: a case study for Gaussian mixture model selection”. In: *ESAIM: P & S* (2011). To appear.

[37] M. Rosenblatt. “Conditional probability density and regression estimators”. In: *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968.* New York: Academic Press, 1969, pp. 25–31.

[38] L. Si and R. Jin. “Adjusting Mixture Weights of Gaussian Mixture Model via Regularized Probabilistic Latent Semantic Analysis”. In: *Advances in Knowledge Discovery and Data Mining.* 2005, pp. 218–252.

[39] C. Stone. “The Use of Polynomial Splines and Their Tensor Products in Multivariate Function Estimation”. In: *Ann. Statist.* 22.1 (1994), pp. 118–171.

[40] S. Szarek. “Metric entropy of homogeneous spaces”. In: *Quantum Probability (Gdansk 1997)* (1998), pp. 395–410.

[41] A. van der Vaart and J. Wellner. *Weak Convergence.* Springer, 1996.

[42] R. Willett and R. Nowak. “Multiscale Poisson Intensity and Density Estimation”. In: *IEEE Transactions on Information Theory* 53.9 (2007), pp. 3171–3187.

[43] D. Young and D. Hunter. “Mixtures of regressions with predictor-dependent mixing proportions”. In: *Computational Statistics & Data Analysis* 54.10 (2010), pp. 2253–2266.